

5. Charakteristiky variability

Každou charakteristiku polohy chápeme jako číslo, kolem něhož jednotlivé hodnoty znaku kolísají. Velikost tohoto kolísání vyjadřují charakteristiky variability (proměnlivosti) znaku.

Příklady

- ① Při deseti opakovaných měření nějaké fyzikální veličiny jsme získali hodnoty uvedené v tabulce. Určete průměr, směrodatnou odchylku a variační koeficient.

č.	délka (m)	$x_i - \bar{x}$ (m)	$(x_i - \bar{x})^2$ (m ²)
1.	2,11	0,05	0,002 5
2.	2,01	-0,05	0,002 5
3.	2,09	0,03	0,000 9
4.	2,11	0,05	0,002 5
5.	2,02	-0,04	0,001 6
6.	2,03	-0,03	0,000 9
7.	2,03	-0,03	0,000 9
8.	2,10	0,04	0,001 6
9.	2,05	-0,01	0,000 1
10.	2,05	-0,01	0,000 1
součty	20,6	0,17	0,013 6
		-0,17	
Aritmetický průměr			Rozptyl
$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{20,6}{10} = 2,06 \text{ m}$			$s_x^2 = \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})^2 = 0,00136 \text{ m}^2$
Závěr: $x = (2,06 \pm 0,04) \text{ m}$ s přesností 2 %			Směrodatná odchylka (někdy tzv. absolutní chyba Δx , průměrná odchylka - ve škole zaokrouhlujeme na jednu platnou číslici)
Pro školní účely: $x = (\bar{x} \pm \Delta x)$ s relativní chybou δ_x - průměrnou odchylku zaokrouhlujeme na jednu platnou číslici - aritmetický průměr potom na stejný řád jako má průměrná odchylka, relativní chybu na 1-2 platné číslice Laboratorní měření se považuje za dostatečně přesné s relativní chybou pod 1 %, v provozních podmínkách i větší chyba (1,5-5 %)			$s_x = \sqrt{s_x^2} \doteq 0,04 \text{ m}$
			Variační koeficient (charakterizuje přesnost měření, někdy tzv. relativní chyba δ_x)
			$v_x = \frac{s_x}{\bar{x}} \cdot 100 \% \doteq 1,9 \% \doteq 2 \%$

ROZPTYL - zn. s_x^2

- charakteristika variability, kterou volíme zpravidla, je-li charakteristikou polohy aritmetický průměr, vyjádřen v *druhých mocninách jednotek, ve kterých jsou udány hodnoty znaku*
- definován jako **průměr druhých mocnin odchylek od aritmetického průměru**

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ resp. } s_x^2 = \frac{1}{n} \sum_{j=1}^r n_j (x_j^* - \bar{x})^2 \text{ pro rozptyl z tabulky četností}$$

- pro praktické výpočty lze vzorce upravit na výhodnější tvary:

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

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2, \quad \text{resp. } s_x^2 = \frac{1}{n} \sum_{j=1}^r n_j x_j^2 - \bar{x}^2 \text{ pro výpočty z tabulky s četnostmi}$$

– podle těchto vzorců se zpravidla počítá rozptyl při ručních výpočtech

SMĚRODATNÁ ODCHYLKA – zn. s_x

– druhá odmocnina z rozptylu

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{resp. } s_x = \sqrt{\frac{1}{n} \sum_{j=1}^r n_j x_j^2 - \bar{x}^2} \text{ pro výpočty z tabulky s četnostmi}$$

- výhodou je, že charakterizuje variabilitu znaku *v týchž jednotkách* měření (fyzikálních jednotkách), *v jakých* jsou udány *hodnoty znaku*
- **udává, jak daleko jsou čísla x_i rozmístěná kolem aritmetického průměru**
- výpočet rozptylu a směrodatné odchylky si lze usnadnit zavedením pomocného znaku: jsou-li hodnoty znaků x a u vázány vztahem $x = a + bu$, kde a, b jsou konstanty, pak
 - pro rozptyly těchto znaků platí vztah $s_x^2 = b^2 s_u^2$
 - pro směrodatné odchylky platí vztah $s_x = |b| s_u$
 (lze ověřit obdobným způsobem jako u aritmetického průměru)

VARIAČNÍ KOEFICIENT – zn. v_x

- charakterizuje variabilitu znaku bezrozměrným číslem
- má smysl jen tehdy, nabývá-li znak x jen nezáporných hodnot
- definován jako podíl směrodatné odchylky a aritmetického průměru, vyjadřuje se obvykle v procentech

$$v_x = \frac{s_x}{\bar{x}} \cdot 100 \%$$

Příklady

- ② Vypočítejte ještě jednou aritmetický průměr, směrodatnou odchylku a variační koeficient pro naměřené hodnoty z příkladu 1 s využitím tabulky četností a výhodnějších vzorců.

délka (m)	2,01	2,02	2,03	2,05	2,09	2,10	2,11
četnost	1	1	2	2	1	1	2

$$\bar{x} = \frac{2,01 + 2,02 + 2 \cdot 2,03 + 2 \cdot 2,05 + 2,09 + 2,10 + 2 \cdot 2,11}{10} = \frac{20,6}{10} = 2,06 \text{ m}$$

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} \sum_{j=1}^r n_j x_j^2 - \bar{x}^2} =$$

$$= \sqrt{\frac{1}{10} (2,01^2 + 2,02^2 + 2 \cdot 2,03^2 + 2 \cdot 2,05^2 + 2,09^2 + 2,10^2 + 2 \cdot 2,11^2) - 2,06^2}$$

$$s_x = \sqrt{\frac{1}{10} \cdot 42,4496 - 4,2436} \doteq 0,0369 \doteq 0,04 \text{ m}$$

$$v_x = \frac{s_x}{\bar{x}} \cdot 100 \% = \frac{0,04}{2,06} \cdot 100 \% \doteq 1,9417 \% \doteq 1,9 \%$$

- ③ Určete průměrnou dobu a směrodatnou odchylku pro dobu strávenou denně přípravou na vyučování.

doba (v hodinách) x_j^*	1	2	3	4	5
četnost n_j	1	9	10	3	1

$$\bar{x} = \frac{1}{n} \sum_{j=1}^r x_j^* n_j = \frac{1 \cdot 1 + 2 \cdot 9 + 3 \cdot 10 + 4 \cdot 3 + 5 \cdot 1}{24} = \frac{72}{24} = 2,75 \text{ h}$$

$$s_x = \sqrt{\frac{1}{n} \sum_{j=1}^r x_j^{*2} n_j - \bar{x}^2} = \sqrt{\frac{1^2 \cdot 1 + 2^2 \cdot 9 + 3^2 \cdot 10 + 4^2 \cdot 3 + 5^2 \cdot 1}{24} - 2,75^2} \doteq 0,878$$

Mezikvartilová odchylka – zn. $Q(x)$

- pokud volíme místo aritmetického průměru za charakteristiku polohy medián, pak z téhož důvodu nahradíme směrodatnou odchylku jinou charakteristikou variability, zpravidla mezikvartilovou odchylkou $Q(x)$

– mezikvartilová odchylka znaku x $Q(x) = \frac{1}{2}(Q_3 - Q_1)$

- je-li $\text{Med}(x)$ prostřední hodnota znaku (hodnoty $x_1 \leq x_2 \leq \dots \leq x_{n-1} \leq x_n$), pak

- **první kvartil Q_1** je medián z hodnot $x_1 \leq x_2 \leq \dots \leq \text{Med}(x)$, tj. čtvrtinová hodnota znaku

- **třetí kvartil Q_3** je medián z hodnot $\text{Med}(x) \leq \dots \leq x_{n-1} \leq x_n$, tj. tříčtvrtinová hodnota znaku

Příklady

- ④ V příkladu 4 minulého článku je medián vhodnější charakteristikou než průměr. Určete mezikvartilovou odchylku.

roční příjem	80	90	100	110	120	890
četnost	1	6	6	5	1	1

- průměrný roční příjem

$$\bar{x} = \frac{1 \cdot 80 + 6 \cdot 90 + 6 \cdot 100 + 5 \cdot 110 + 1 \cdot 120 + 1 \cdot 890}{20} = \frac{2\,780}{20} = 139 \text{ (tisíc) Kč}$$

- (směrodatná odchylka by byla příliš ovlivněna hodnotou 890, vyšlo by $s_x = 173$ tisíc Kč)

- medián (n sudé)

$$\text{Med}(x) = \frac{x_{(10)} + x_{(11)}}{2} = \frac{100 + 100}{2} = 100 \text{ (tisíc) Kč}$$

- mezikvartilová odchylka

$$Q_1 = \frac{x_{(5)} + x_{(6)}}{2} = \frac{90 + 90}{2} = 90 \text{ (tisíc) Kč}$$

$$Q_3 = \frac{x_{(15)} + x_{(16)}}{2} = \frac{110 + 110}{2} = 110 \text{ (tisíc) Kč}$$

$$Q(x) = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(110 - 90) = 10 \text{ (tisíc) Kč}$$

6. Korelace

Zatím jsme se zabývali popisem jediného znaku (nebo pro více znaků jen popisem každého z nich samostatně). Nyní se budeme zabývat popisem dvojice znaků (x, y) , výsledkem šetření v souboru o rozsahu n jsou pak data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Budeme se zabývat vedle charakteristik polohy a variability, počítaných pro každý ze znaků zvlášť, o **míru statistické závislosti dvou znaků**. Volíme-li za charakteristiku polohy aritmetické průměry $\bar{x}$ a $\bar{y}$ a za charakteristiku variability směrodatné odchylky s_x a s_y , pak za charakteristiku závislosti volíme zpravidla koeficient korelace (korelační koeficient).

Korelační koeficient znaků x a y – charakterizuje matematicky závislost dvou znaků

$$r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} \quad \text{lze upravit na vhodnější tvar} \quad r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{s_x \cdot s_y}$$

Příklady

- ① V tabulce je uvedeno prvních 6 dvojic znaků známka z matematiky a známka z fyziky. Určete jejich korelační koeficient.

známka z matematiky x	3	2	2	4	3	2
známka z fyziky y	3	2	1	3	2	2

Pomocné výpočty:

$$\bar{x} = \frac{3 + 2 + 2 + 4 + 3 + 2}{6} = 2,67 \quad \bar{y} = \frac{3 + 2 + 1 + 3 + 2 + 2}{6} = 2,17$$

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{6} (3^2 + 2^2 + 2^2 + 4^2 + 3^2 + 2^2) - 2,67^2} = 0,733$$

$$s_y = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2} = \sqrt{\frac{1}{6} (3^2 + 2^2 + 1^2 + 3^2 + 2^2 + 2^2) - 2,17^2} = 0,677$$

$$r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{s_x \cdot s_y} = \frac{\frac{1}{6} (3 \cdot 3 + 2 \cdot 2 + 2 \cdot 1 + 4 \cdot 3 + 3 \cdot 2 + 2 \cdot 2) - 2,67 \cdot 2,17}{0,733 \cdot 0,677} = 0,751$$

Hodnota korelačního koeficientu $r_{xy} = 0,751$ již znamená značnou míru závislosti.

Korelační koeficient je kladný (většina součinů v čitateli vzorce bude kladných)

⇔ nadprůměrným hodnotám znaku x odpovídají zpravidla nadprůměrné hodnoty y
a podprůměrným podprůměrné (vůči průměru)

Výpočet lze urychlit tím, že sestavíme a využijeme **tabulku četnosti**, ale tentokrát tabulku četnosti **dvojic hodnot znaků x a y** ⇒ každá dvojice hodnot potřebuje svou buňku.

- ② Zadaná tabulka ukazuje četnosti známek z matematiky a příslušné četnosti týdenního počtu hodin domácí přípravy na matematiku. Určete korelační koeficient.

		Známka z matematiky				
		1	2	3	4	5
Domácí příprava na matematiku (týdně v hodinách)	2				1	
	4		1	4	1	
	6		4	1	1	
	8	6	3	1		
	10	1				

$$\bar{x} = \frac{7 \cdot 1 + 8 \cdot 2 + 6 \cdot 3 + 3 \cdot 4}{24} = 2,21 \quad \bar{y} = \frac{1 \cdot 2 + 6 \cdot 4 + 6 \cdot 6 + 10 \cdot 8 + 1 \cdot 10}{24} = 6,33$$

$$s_x = \sqrt{\frac{1}{n} \sum_{j=0}^r n_j x_j^2 - \bar{x}^2} = \sqrt{\frac{1}{24} (7 \cdot 1^2 + 8 \cdot 2^2 + 6 \cdot 3^2 + 3 \cdot 4^2) - 2,21^2} = 0,995$$

$$s_y = \sqrt{\frac{1}{n} \sum_{j=0}^r n_j y_j^2 - \bar{y}^2} = \sqrt{\frac{1}{24} (1 \cdot 2^2 + 6 \cdot 4^2 + 6 \cdot 6^2 + 10 \cdot 8^2 + 1 \cdot 10^2) - 6,33^2} = 1,98$$

$$\frac{1}{24} \sum_{j=1}^r n_j x_j^* y_j^* = \frac{48 + 10 + 8 + 48 + 48 + 48 + 18 + 24 + 8 + 16 + 24}{24} = \frac{300}{24} = 12,5$$

$$\text{– koeficient korelace } r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{s_x \cdot s_y} = \frac{\frac{1}{24} \cdot 300 - 2,21 \cdot 6,33}{0,995 \cdot 1,98} = \frac{-1,49}{1,9701} = -0,756$$

Korelační koeficient vyjadřuje značnou závislost mezi známkou z M a týdenní dobou přípravy na M, je záporný – více hodin domácí přípravy znamená zpravidla lepší známku z M, tj. nadprůměrným hodnotám doby přípravy odpovídají podprůměrné (nižší) známky.

Korelační koeficient (koeficient korelace)

- **vždy číslo z intervalu $\langle -1, 1 \rangle$** , bezrozměrné číslo – nemá jednotku
- krajních hodnot 1 a -1 nabývá tehdy, je-li mezi znaky x a y funkční závislost (nejenom statistická) a to lineární $y = ax + b$, kde znaménko čísla b určuje, zda jde o hodnotu 1 či -1

Pokud koeficient korelace

- **kladný** (většina součinů v čitateli vzorce bude kladných)
 - nadprůměrným hodnotám znaku x odpovídají zpravidla nadprůměrné hodnoty y a podprůměrným podprůměrné
- **záporný** (většina součinů v čitateli vzorce bude záporných)
 - nadprůměrným hodnotám znaku x odpovídají zpravidla podprůměrné hodnoty y a podprůměrným nadprůměrné
- **blízký nule** (kladné i záporné součiny v čitateli vzorce budou mít tendenci se v součtu rušit)
 - mezi x a y není žádná závislost

Pomocí korelace lze prokázat vzájemnou souvislost dvou znaků, ALE **ze vzájemné souvislosti nevyplývá příčinný vztah**.

Například výška platu koreluje s cenou soukromého automobilu. Tvrdit, že si musíme koupit drahé auto, aby nám zvýšili plat, by bylo velmi odvážné (příčinná souvislost je zřejmě opačná).