9.3.4 Charakteristiky variability

Předpoklady: 9303

Př. 1: Při fyzikálním praktiku měřili studenti tloušťku dřevěného kvádříku. Petr měřil pravítkem a naměřil tyto hodnoty (v mm): 50; 50; 51; 51; 52; 51; 50; 50; 52; 51. Jarda použil šupleru a získal tyto výsledky (v mm): 50,8; 50,7; 50,7; 51,0; 50,8; 50,9; 50,9; 50,7; 50,8; 50,7. Urči průměrnou tloušťku kvádříku podle výsledků obou měření. Rozhodni zda jsou jejich výsledky rovnocenné.

Průměr podle Petra:

$$\bar{x} = \frac{50+50+51+51+52+51+50+50+52+51}{10} = 50,8$$

Průměr podle Jardy:

$$\bar{x} = \frac{50,8+50,7+50,7+51,0+50,8+50,9+50,9+50,7+50,8+50,7}{10} = 50,8$$

Ačkoliv oba získají stejnou průměrnou hodnotu, jejich výsledky rovnocenné nejsou, protože jednotlivá měření prováděná Jardou se od výsledného průměru liší podstatně méně než měření prováděná Petrem. Jarda měřil daleko přesněji.

Předchozí příklad je ukázkou situace, ve které není důležité pouze okolo jaké hodnoty jednotlivé výsledky kolísají, ale i to, jak moc kolísají $\Rightarrow$ hledáme charakteristiku **variability (proměnlivosti)** znaku.

Jak spočítat hodnotu, která vyjádří, jak moc se změřené hodnoty liší od průměru?

První nápad: udělat průměr z jednotlivých odchylek.

Př. 2: Urči průměr z odchylek jednotlivých měření od průměru pro sadu hodnot, které naměřil Petr i Jarda.

Petr:

$$\frac{(50-50,8)+(50-50,8)+(51-50,8)+\ldots+(50-50,8)+(52-50,8)+(51-50,8)}{10} = 0$$

Jarda:

$$\bar{x} = \frac{(50,8-50,8)+(50,7-50,8)+\ldots+(50,8-50,8)+(50,7-50,8)}{10} = 0$$

Proč náš pokus selhal?
Část odchylek je kladná, část záporná $\Rightarrow$ společně se odečtou na nulu (už když jsme počítali průměr, počítali jsme ho tak, aby odchylky na obě strany byly stejné)
$\Rightarrow$ musíme zajistit, aby se odchylky navzájem neodečítaly $\Rightarrow$
dvě možnosti:

- absolutní hodnota $|x_i - \bar{x}|$
- druhá mocnina $(x_i - \bar{x})^2$ - to se používá

Jako základní charakteristikou proměnlivosti je rozptyl, který je definovaný jako průměr druhých mocnin odchylek od aritmetického průměru.

$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ pokud použijeme tabulku četností: $s_x^2 = \frac{1}{n}\sum_{j=1}^{r}\left(x_j^* - \overline{x}\right)^2 n_j$

Uvedený vzorec je pro počítání poměrně těžkopádný. Je možné ho upravit takto:

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

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

$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$ pokud použijeme tabulku četností: $s_x^2 = \frac{1}{n}\sum_{j=1}^{r} x_j^{*2} n_j - \overline{x}^2$

Pedagogická poznámka: Předchozí odvození si jenom ukážeme, neztrácíme čas tím, že si ho studenti budou opisovat do sešitů.

Př. 3: Urči rozptyl pro měření Petra a Jardy pomocí obou vzorců.

Petr:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{10}\left[\left(50-50,8\right)^2 + \left(50-50,8\right)^2 + \ldots + \left(52-50,8\right)^2 + \left(51-50,8\right)^2\right] =$$

$$= 0,56$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{10}\left(50^2 + 50^2 + 51^2 + 51^2 + 52^2 + 51^2 + 50^2 + 50^2 + 52^2 + 51^2\right) - 50,8^2 =$$

$$= 0,56$$

Jarda:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{10}\left[\left(50,8-50,8\right)^2 + \left(50,7-50,8\right)^2 + \ldots + \left(50,8-50,8\right)^2 - \left(50,7-50,8\right)^2\right] =$$

$$= 0,01$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{10}\left(50,8^2 + 50,7^2 + 50,7^2 + 51^2 + 50,8^2 + 50,9^2 + 50,9^2 + 50,7^2 +\right.$$

$$\left. + 50,8^2 + 50,7^2\right) - 50,8^2 = 0,01$$

Výsledky jsou stejné podle obou vzorců a navíc zachycují to, co jsme chtěli. Do ideálního stavu ještě něco chybí: měřili jsme délku (v metrech), ale rozptyl je průměr druhých mocnin (a je tedy v m^2).

$\Rightarrow$ **směrodatná odchylka:** $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$ (absolutní chyba)

$\Rightarrow$ **variační koeficient:** $v_x = \frac{s_x}{\overline{x}} \cdot 100\%$ (relativní chyba δ_x)

Př. 4: Urči směrodatnou odchylku a variační koeficient obou měření.

Petr:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \sqrt{0,56} = 0,75$$

$$v_x = \frac{s_x}{\overline{x}} \cdot 100\% = \frac{0,75}{50,8} \cdot 100\% \doteq 1,5\%$$

Jarda:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \sqrt{0,01} = 0,1$$

$$v_x = \frac{s_x}{\overline{x}} \cdot 100\% = \frac{0,1}{50,8} \cdot 100\% \doteq 0,2\%$$

Používání rozptylu a směrodatné odchylky je svázáno s používáním průměru. Pokud charakterizujeme polohu pomocí $\mathrm{Mod}(x)$ nebo $\mathrm{Med}(x)$, měl bychom použít místo rozptylu **mezikvartilovou odchylku.**

Co je kvartil?
Podobně jako medián je kvartil případ kvantilu:

- medián – dělí seřazené jednotky na dvě poloviny
- kvartily - dělí seřazené jednotky na čtyři čtvrtiny
- decily - dělí seřazené jednotky na deset desetin
- percentily - dělí seřazené jednotky na sto setin

první kvartil Q_1 je čtvrtinová hodnota (nebo také medián z hodnot $x_1 \leq x_2 \leq \ldots \leq \mathrm{Med}(x)$)

třetí kvartil Q_3 je tříčtvrtinová hodnota (nebo také medián z hodnot $x_n \geq x_{n-1} \geq \ldots \geq \mathrm{Med}(x)$)

Mezikvartilová odchylka $Q(x) = \dfrac{Q_3 - Q_1}{2}$

Př. 5: Urči Q_1, Q_3 a $Q(x)$ pro velikost postavy ze statistického výzkumu.

x_j^*	160	165	170	175	180	185	190
n_j	1	1	4	6	4	2	1

Celkem 19 hodnot $\Rightarrow$ 10 hodnota je medián

Q_1 je prostřední hodnota z prvních deseti hodnot $\Rightarrow$ jde o průměr 5. a 6. hodnoty

$$Q_1 = \frac{170+170}{2} = 170$$

Q_3 je prostřední hodnota z posledních deseti hodnot $\Rightarrow$ jde o průměr 14. a 15. hodnoty

$$Q_3 = \frac{180+180}{2} = 180$$

$$Q(x) = \frac{Q_3 - Q_1}{2} = \frac{180-170}{2} = 5$$

Kvantily se používají pro vyjadřování relativního umístění v nějakém pořadí. Například pokud dosáhnete percentil 92%, znamená to, že 92% účastníků je v pořadí za Vámi.

Př. 6: Přijímacího řízení se zúčastnilo 756 studentů. Petr získal umístění v 85 percentilu. Urči kolik procent studentů uspělo v testu hůře než on. Kolik studentů uspělo lépe než on? Jaké pořadí by mohl celkově zaujímat? Kolik studentů se umístilo v šestém decilu?

85 percentil $\Rightarrow$ 85 % studentů uspělo hůře
85 percentil $\Rightarrow$ 15 % studentů uspělo lépe
100% … 756 studentů
15% … $756 \cdot 0,15 = 113,4$ $\Rightarrow$ 113 studentů se umístilo lépe než Petr
$\Rightarrow$ Petr se umístil na 114 místě
decil dělí skupinu na deset částí $\Rightarrow$ v každém decilu se umístilo 76 (75) studentů

Př. 7: Petáková:
strana 175/cvičení 69
strana 175/cvičení 70

Shrnutí: Charakteristiky vyjadřují míru proměnlivosti znaku (jak moc se liší jednotlivé hodnoty jedna od druhé).