Popisné statistiky

Uvažujme naměřené hodnoty délky kojenců (v cm) v 1 roce věku. Tato proměnná má hodnoty:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}
83	72	80	77	78	75	78	81	78	73	82	73	70	68	76	72	80	76	77	78	76	82	78

Počet dat se většinou značí n. Zde tedy $n = 23$.

U každého souboru dat je potřeba charakterizovat jeho polohu (kolem jaké tak hodnoty se data pohybují) a variabilitu (jak jsou data kolem této hodnoty rozptýlena).

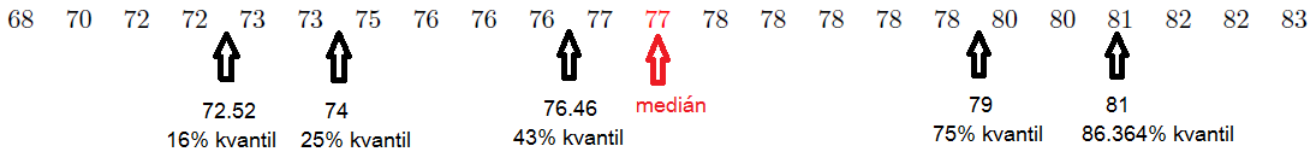
Charakteristiky polohy

Průměr (angl. mean)

$$\bar{x} = \frac{1}{23}(x_1 + x_2 + \ldots + x_{23}) = \frac{1}{n}\sum_{i=1}^{n} x_i = 76.65217$$

Kvantil (angl. quantile)

$p-$procentní kvantil je takové číslo, že p % dat je menších nebo rovno tomuto číslu. Někdy se též nazývá **percentil**. Kvantily určujeme vždy z dat uspořádaných od nejmenšího k největšímu.



Některé speciální kvantily mají své vlastní názvy:

- **medián** = 50% kvantil (prostřední hodnota v uspořádaných datech), značí se $\tilde{x}$
- **dolní kvartil** = 25% kvantil (takové číslo, že čvrtina dat je menších nebo rovno tomuto číslu), značí se Q_1
- **horní kvartil** = 75% kvantil (takové číslo, že 75 % dat je menších nebo rovno tomuto číslu), značí se Q_3
- **první decil** = 10% kvantil

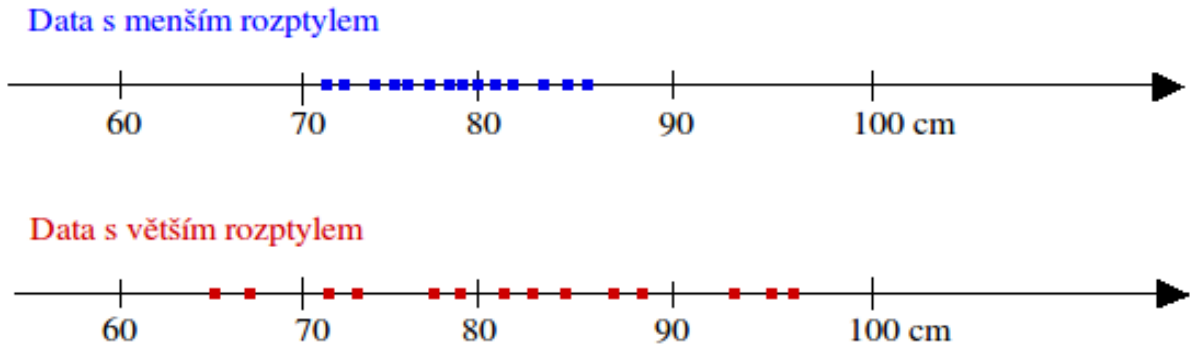
Minimum, maximum

jsou také mírami polohy. Lze je (trošku uměle) vyjádřit i jako kvantily: minimum = 0% kvantil, maximum = 100% kvantil.

Charakteristiky variability

Rozptyl (angl. variance)

Ilustrace:



Výpočet:
Spočteme odchylky $(x_1 - \bar{x}), \ldots, (x_{23} - \bar{x})$, umocníme je (aby byly všechny kladné) a zprůměrujeme, tedy:

$$s_x^2 = \frac{1}{23-1}\Big((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_{23} - \bar{x})^2\Big) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Pozor, ve jmenovateli je opravdu $n - 1$ nikoli n.
Často se též značí "var".

Směrodatná odchylka (angl. standard deviation)

Stejně jako rozptyl charakterizuje rozptýlenost dat. Spočte se jako druhá odmocnina z rozptylu, tedy:

$$s_x = \sqrt{s_x^2}$$

Často se též značí "sd".

Mezikvartilové rozpětí (angl. interquartile range)

Říká, jak jsou od sebe dolní a horní kvartil daleko, a tak vlastně také charakterizuje rozptýlenost dat.

$$\mathrm{IQR} = Q_3 - Q_1$$

Rozpětí (angl. range)

Říká, jak jsou od sebe daleko minimum a maximum.

$$R = x_{max} - x_{min}$$