Základy spracovania a vizualizácie dát

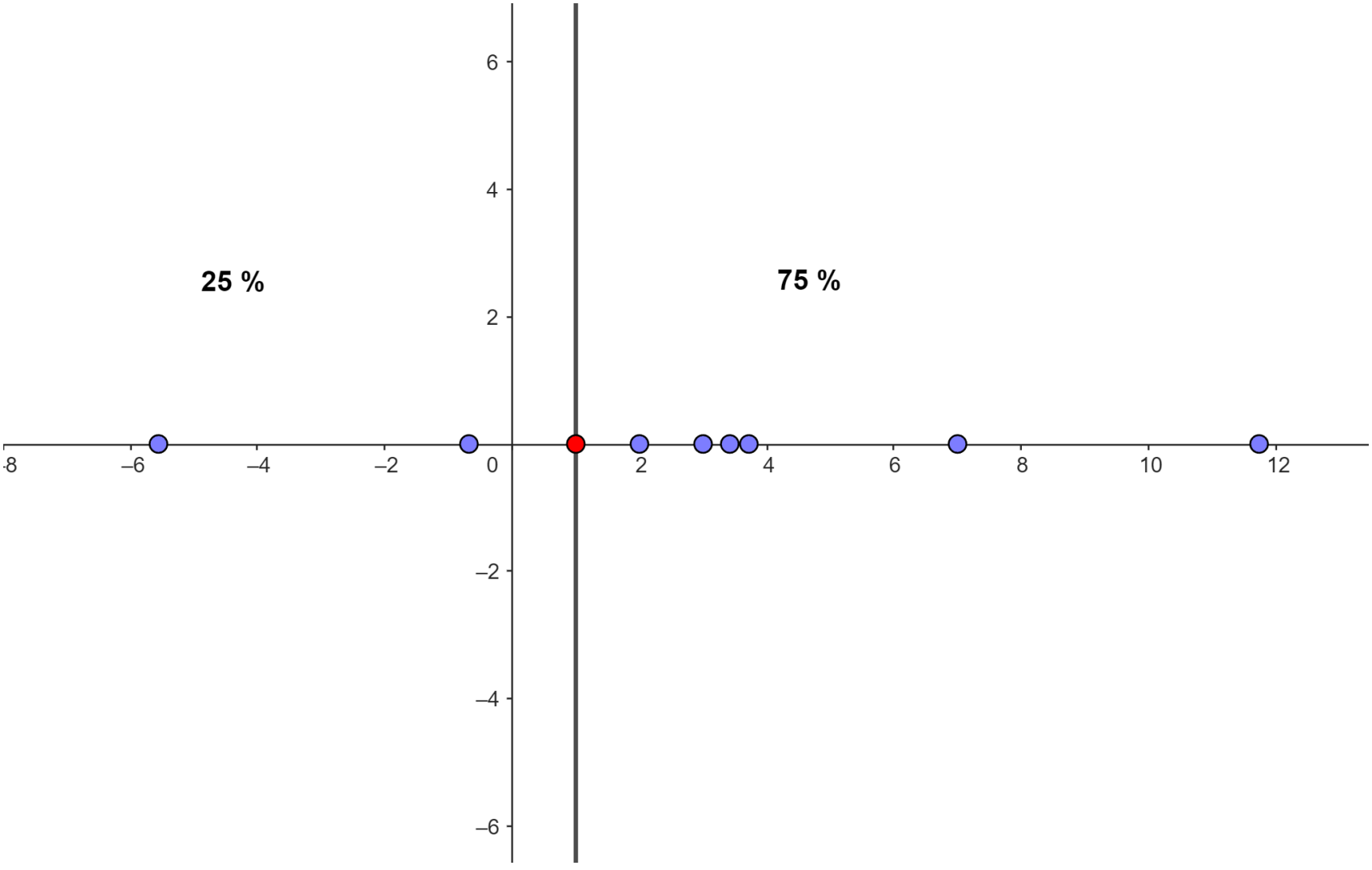
11. marca 2024

charakteristiky polohy

- minimum, maximum
- priemer - priemerná hodnota
  medián - prostredná hodnota
  modus - najpočetnejšia hodnota
- kvantily
  - značenie q_α
  - $\alpha 100$ % kvantil = naľavo od tejto hodnoty je $\alpha 100$ % dát, napravo je $(1-\alpha)100$ % dát
  - počítanie interpoláciou

kvantily

- najznámejšie sú kvartily, decily a percentily (využitie napr. monitor, maturita)
- $q_{0,25}$ = 1. kvartil = 25 % kvantil



charakteristiky variability

- rozptyl
  - celá populácia

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

  - výberový súbor

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

- smerodajná odchýlka

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

- variančné rozpätie

$$max - min$$

- medzikvartilové rozpätie

$$IQR = q_{0,75} - q_{0,25}$$

cvičenie

- načítajte dáta *alk_faj_data.txt*
- spočítajte pre *Percento denných fajčiarov*
  - minimum, maximum, priemer, medián, modus (zistite, či je jediný) a kvartily
  - výberový rozptyl, smerodajnú odchýlku, variančné rozpätie a medzikvartilové rozpätie

outlieri

- extremálne hodnoty - hodnoty veľmi sa líšiace od zvyšku súboru
- rôzne metódy určovania outlierov (závisí od typu dát, zamýšľanej metódy a pod.), napr.
  - mimo intervalu $(q_{0,05}; q_{0,95})$, $(q_{0,10}; q_{0,90})$, ...
  - mimo intervalu $(q_{0,25} - 1{,}5IQR; q_{0,75} + 1{,}5IQR)$
- IQR = medzikvartilové rozpätie

cvičenie

- identifikujte outlierov na základe metódy využívajúcej *IQR*
- vykreslite graf zobrazujúci veľké množstvo číselných charakteristík súboru (boxplot - škatuľový graf)
- pridajte do tabuľky ďalší stĺpec, v ktorom bude vypísané „outlier" (ak je daná krajina outlierom) alebo tam bude pôvodná hodnota *Percento denných fajčiarov*

samostatné cvičenie

- spočítajte číselné charakteristiky pre *Percento týždenných konzumentov alkoholu*
- identifikujte outlierov na základe metódy využívajúcej *IQR*
- vykreslite boxplot