Analýza a vizualizácia dát

4. marca 2024

- rozcvičky na konci hodiny:
  - 11.3.2024
  - 15.4.2024
  - 29.4.2024
  - 13.5.2024 - opravná / náhradná

deskriptívna štatistika

- rozlišujeme dáta celej populácie a dáta výberového súboru (výber z populácie)
- prehľadne reprezentuje dáta
- tabuľky, frekvenčné tabuľky
- číselné charakteristiky súboru
  - polohy
  - variability
  - tvaru
- grafy

pivotky

- frekvenčné tabuľky
- intervalové tabuľky
- kontingenčné tabuľky
- viacúrovňové kontingenčné tabuľky

cvičenie

- načítajte dáta *HR-Employee-Attrition*
- vytvorte pivotku pre znak Attrition
- vytvorte pivotku pre BusinessTravel v percentách
- vytvorte intervalovú tabuľku pre Age v percentách
- skúmajte závislosť znakov Department a Attrition
- skúmajte závislosť znakov Department a BusinessTravel s Attrition
- vložte graf spojený s pivotkou

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

cvičenie

- spočítajte charakteristiky polohy pre znak Age
- *MIN(), MAX(), AVERAGE(), MEDIAN(), MODE.MULT(), MODE.SNGL(), QUARTILE.INC()*

pozor na

- outlierov (extremálne hodnoty)
- interpretáciu (priemer nie je prostredná hodnota)
- heterogenitu dát
- kategoriálne dáta

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

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

cvičenie

- spočítajte charakteristiky variability pre znak Age
- *VAR.P(), VAR.S(), STDEV.P(), STDEV.S()*

charakteristiky tvaru

- často slúžia na porovnanie s Gaussovou krivkou
- šikmosť

$$\frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- špicatosť

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

cvičenie

- spočítajte charakteristiky tvaru pre znak *Age*
- *SKEW(), SKEW.P(), KURT()*

cvičenie

- pre znak *Age* vytvorte histogram a porovnajte ho (vykreslením) s hustotou normálneho rozdelenia

samostatné cvičenie

- v dátovom súbore *Infant_mortality_data_1965* pre znak *Mortal*
  - vytvorte frekvenčnú intervalovú tabuľku
  - spočítajte číselné charakteristiky