7. přednáška – POPISNÁ STATISTIKA

Explorační (popisná) statistika bývá prvním krokem k odhalení informací skrytých ve velkém množství proměnných a jejich variant.

STATISTICKÉ CHARAKTERISTIKY NUMERICKÝCH PROMĚNNÝCH

Třídění dat

Nejčastěji postupujeme tak, že data uspořádáme podle velikosti a stanovíme intervaly, které odpovídají jednotlivým třídám. Mluvíme pak o intervalovém třídění.

Četnosti

Absolutní četnost - f_i;

Relativní četnost - $\varphi_i = \dfrac{f_i}{n}$;

Kumulativní četnost - $F_i = \sum\limits_{k=1,\dots,i} f_k$;

Relativní kumulativní četnost - $\Phi_i = \dfrac{F_i}{n}$;

Příklad: Ze 7 možných výsledků jsme dostali datový soubor o 14 datech: 2,1,3,2,5,2,7,1,4,5,4,2,1,5. Určete četnosti.

Grafická znázornění

Histogram - graf kdy na vodorovnou osu znázorníme třídy a na svislou osu četnosti či relativní četnosti.

Kruhový (výsečový) diagram - je znázornění pomocí výsečí kruhu, kde každé třídě odpovídá jedna výseč. Velikosti obsahů výsečí odpovídají četnostem třídy.

Stem-and-Leaf diagram - je uspořádání dat do tabulky

Krabicový graf - znázorňuje význačné a extrémní hodnoty souboru

Příklad: pokračování, graficky znázorněte data

Charakteristiky míry

Průměr (aritmetický)

- datového souboru je definován $\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$
- soubor s četnostmi: $\overline{x} = \frac{1}{n}\sum_{i=1}^{m} z_i \cdot f_i$, kde z_i je zástupce i-té třídy

Medián $x_{0,5}$ - je ta hodnota, která rozděluje soubor na dvě části o stejném počtu prvků

Modus - $Mo(x)$ je ta hodnota, která má největší absolutní četnost

Pokračování *příklad*: určete průměr, medián a modus

Charakteristiky variability

Následující statistické charakteristiky nám umožňují popis variability (rozptýlenosti) výběrového souboru, neboli popis rozptylu jednotlivých hodnot kolem středu proměnné – nazýváme je tedy mírami variability.

Variační rozpětí R - rozdíl mezi největší a nejmenší hodnotou kvantitativního znaku

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

Interkvartilové rozpětí IQR – vzdálenost mezi dolním a horním kvartilem

$$\mathrm{IQR} = x_{0,75} - x_{0,25}$$

Výběrový rozptyl s^2 - nejrozšířenější míra variability výběrového souboru

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \overline{x}\right)^2$$

Pokud jsou data zadána jako soubor s četnostmi, využijeme vzorce:

$$s^2 = \frac{1}{n}\sum_{k=1}^{n} n_k\left(x_k - \overline{x}\right)^2$$

Výběrová směrodatná odchylka s - definována jako kladná odmocnina výběrového rozptylu

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

Pokračování *příklad*: určete charakteristiky variability

Variační koeficient V_x - vyjadřuje relativní míru variability proměnné x. Podle níže uvedeného vztahu jej lze stanovit pouze pro proměnné, které nabývají výhradně kladných hodnot. Variační koeficient je bezrozměrný. Uvádíme-li jej v [%], hodnotu získanou z definičního vzorce vynásobíme 100%.

$$V_x = \frac{s}{\bar{x}}$$

Příklad: V levé skupině máme tři čísla : 7; 8; 9. V pravé skupině máme čísla 1; 10, 13. Určete variační koeficient pro obě skupiny.

Koeficient šikmosti - charakteristiky šikmosti udávají, jsou-li hodnoty kolem zvoleného středu rozloženy souměrně nebo je-li rozdělení hodnot zešikmeno na jednu stranu

$$A = \frac{1}{n}\sum_{k=1}^{n}\frac{(x_k - \bar{x})^3}{s^3}$$

Koeficient špičatosti - charakteristiky špičatosti udávají, jaký průběh má rozdělení hodnot kolem zvoleného středu (rozdělení).

$$e = \frac{1}{n}\sum_{k=1}^{n}\frac{(x_k - \bar{x})^4}{s^4}$$