JEDNOROZMĚRNÁ POPISNÁ STATISTIKA – 2. ČÁST

Statistické charakteristiky (míry)

- shrnují informaci, obsaženou v datech, vyjadřují ji v koncentrované formě
- charakterizují základní rysy zkoumaného souboru dat
- umožňují porovnávání více souborů.

Skupiny statistických charakteristik:

1. charakteristiky polohy
2. charakteristiky variability
3. charakteristiky šikmosti
4. charakteristiky špičatosti.

Obecné principy konstrukce statistických charakteristik:

a) Charakteristiky, které jsou funkcí všech hodnot dané proměnné:
- jsou ovlivněny případnými extrémy
- výpočet je prováděn podle určitého funkčního předpisu.

b) Charakteristiky, které nejsou funkcí všech hodnot dané proměnné:
- nejsou ovlivněny případnými extrémy
- jsou to konkrétní hodnoty proměnné, vybrané podle určitého kritéria.

Charakteristiky polohy

- charakterizují střed, kolem něhož hodnoty kolísají
- charakterizují úroveň (velikost, hladinu) proměnné
- používá se pro ně rovněž pojem *střední hodnoty*.

a) Charakteristiky, které jsou funkcí všech hodnot – průměry

Aritmetický průměr

prostý: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ *vážený:* $\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$

Použití: Používá se tam, kde má informační smysl součet hodnot proměnné.

Harmonický průměr

prostý: $\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$ *vážený:* $\bar{x}_H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Použití: Používá se tam, kde má smysl součet převrácených hodnot proměnné. Např. k výpočtu průměrné doby potřebné ke splnění úkolu, kdy jednotky plní úkoly současně.

Geometrický průměr

prostý: $\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot ..... \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

vážený: $\bar{x}_G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot ... \cdot x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$

Použití: Používá se tam, kde má smysl součin hodnot proměnné. Např. k výpočtu průměrného koeficientu růstu v časových řadách.

Kvadratický průměr

prostý: $\bar{x}_K = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$ *vážený:* $\bar{x}_K = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{\sum_{i=1}^{k} n_i}}$

Použití: Používá se tam, kde má smysl součet čtverců hodnot proměnné. Např. tehdy, jestliže jednotlivé hodnoty jsou odchylkami původních hodnot od aritmetického průměru, odchylkami od normy apod.

Vztahy mezi průměry

Jsou -li výše uvedené 4 typy průměrů vypočítány z týchž kladných hodnot proměnné, platí pro ně vztah $\bar{x}_H \le \bar{x}_G \le \bar{x} \le \bar{x}_K$.

b) Charakteristiky, které nejsou funkcí všech hodnot

- patří sem především modus a kvantily
- jejich výhodou je, že nejsou ovlivněny odlehlými pozorováními.

Modus

- varianta s největší četností (tzv. typická hodnota)
- vrchol rozdělení
- označujeme symbolem $\hat{x}$.

Kvantily

- hodnoty, které rozdělují uspořádaný statistický soubor na určitý počet stejně obsazených částí
- hodnoty menší event. stejné tvoří určitou stanovenou část rozsahu souboru (určitý podíl, určité procento).

Uspořádaný statistický soubor: hodnoty proměnné jsou seřazeny do neklesající řady.

Obecné označení kvantilů:

x_p, kde *p* je relativní četnost

$\tilde{x}_{100p}$, kde 100*p* je relativní četnost vyjádřená v %.

Druhy kvantilů:

- **Medián –** $\tilde{x}$, $\tilde{x}_{50}$, $x_{0,5}$ – je prostřední hodnota uspořádaného statistického souboru. Člení uspořádaný statistický soubor na dvě stejně četné části, existuje tedy 50 % hodnot menších nebo stejných a 50 % hodnot větších nebo stejných.

  Výpočet mediánu:

  - *rozsah souboru n je liché číslo*

    $\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$, kde výraz $\frac{n+1}{2}$ udává pořadí mediánu v neklesající řadě hodnot.

    Při lichém rozsahu souboru je mediánem konkrétní prvek.

  - *rozsah souboru n je sudé číslo*

    $$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}}{2}$$

    Při sudém rozsahu souboru existují 2 prostřední hodnoty a medián je jejich aritmetickým průměrem.

- **Tercily –** $\tilde{x}_{33,\bar{3}}\left(x_{0,\bar{3}}\right)$, $\tilde{x}_{66,\bar{6}}\left(x_{0,\bar{6}}\right)$ – jsou 2 kvantily, které rozdělují uspořádaný statistický soubor na 3 stejně četné části.

- **Kvartily –** $\tilde{x}_{25}\left(x_{0,25}\right)$, $\tilde{x}$, $\tilde{x}_{75}\left(x_{0,75}\right)$ – jsou 3 kvantily, které rozdělují uspořádaný statistický soubor na 4 stejně četné části.

- **Kvintily –** $\tilde{x}_{20}\left(x_{0,2}\right)$, $\tilde{x}_{40}\left(x_{0,4}\right)$, $\tilde{x}_{60}\left(x_{0,6}\right)$, $\tilde{x}_{80}\left(x_{0,8}\right)$ – jsou 4 kvantily, které rozdělují uspořádaný statistický soubor na 5 stejně četných částí.

- **Sextily** – 5 kvantilů, 6 částí.

- **Septily** – 6 kvantilů, 7 částí.

- **Oktávily** – 7 kvantilů, 8 částí.

- **Nonily** – 8 kvantilů, 9 částí.

- **Decily** – 9 kvantilů, 10 částí.

- **Percentily** – 99 kvantilů, 100 částí, atd.

Pozn.: Kvantily menší než $\tilde{x}$ bývají označovány pojmem *dolní kvantily*, kvantily větší než $\tilde{x}$ pojmem *horní kvantily*.

Charakteristiky variability

- udávají rozptýlení (kolísání) hodnot kolem zvoleného středu, např. kolem nějaké střední hodnoty
- variabilita = měnlivost = kolísavost = odlišnost.

Míry absolutní variability

Variační rozpětí: $R = x_{\max} - x_{\min}$

Kvantilová rozpětí

kvartilové rozpětí: $R_q = \tilde{x}_{75} - \tilde{x}_{25}$ *decilové rozpětí:* $R_d = \tilde{x}_{90} - \tilde{x}_{10}$, atd.

Kvantilové odchylky

kvartilová odchylka: $Q = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2}$ *decilová odchylka:* $D = \frac{\tilde{x}_{90} - \tilde{x}_{10}}{8}$, atd.

Průměrná absolutní odchylka

prostá: $\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$ *vážená:* $\bar{d} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}|\, n_i}{\sum_{i=1}^{k} n_i}$

Rozptyl

prostý (klasický): $s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ *vážený (klasický):* $s_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}{\sum_{i=1}^{k} n_i}$

Výpočtový tvar rozptylu

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

vážený: $s_x^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{\sum_{i=1}^{k} n_i} - \left(\frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} \right)^2 = \overline{x^2} - \bar{x}^2$

Směrodatná odchylka

- kladná odmocnina z rozptylu, tj. $s_x = \sqrt{s_x^2}$; vhodná pro interpretaci
- udává, jak se v průměru liší jednotlivé hodnoty znaku od aritmetického průměru v obou směrech (±).

Pokud pracujeme s výběrovým souborem, počítáme ***výběrový rozptyl*** a ***výběrovou směrodatnou odchylku***:

prostý: $s_x'^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ *vážený:* $s_x'^2 = \dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1}$

Míry relativní variability

Variační koeficient

- bezrozměrné číslo; umožňuje porovnávat variabilitu souborů s různou úrovní či různými měrnými jednotkami
- obecně může nabývat hodnot z intervalu $(-\infty, \infty)$, pro kardinální proměnnou z intervalu $\langle 0, \infty)$.

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

Variabilita nominální proměnné (mutabilita)

Míra mutability

- bezrozměrné číslo, lze vyjádřit v %.
- podíl dvojic jednotek s různou obměnou z celkového počtu všech možných dvojic.

$$M = \frac{n^2 - \sum_{i=1}^{k} n_i^2}{n(n-1)}; \quad M \in \langle 0,1 \rangle$$

Nominální variance

- používá se v případě, že známe pouze relativní četnosti a není znám rozsah souboru
- skutečný stupeň variability podhodnocuje.

$$NOMVAR = 1 - \sum_{i=1}^{k} p_i^2; \quad NOMVAR \in \langle 0,1)$$

Charakteristiky šikmosti (asymetrie)

- šikmost je v podstatě rozdílná koncentrace malých hodnot znaku ve srovnání s koncentrací velkých hodnot znaku.

Míry šikmosti

prostá: $\alpha = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n s_x^3}$ *vážená:* $\alpha = \dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i}{n s_x^3}$

jednoduchá charakteristika: $\alpha' = \frac{n' - n''}{n}$

kde n' je počet podprůměrných hodnot
n'' je počet nadprůměrných hodnot.

Interpretace charakteristik šikmosti:

- v symetrickém rozdělení = 0; počet podprůměrných hodnot je stejný jako počet hodnot nadprůměrných
- v kladně sešikmeném rozdělení > 0; existuje více hodnot podprůměrných než nadprůměrných
- v záporně sešikmeném rozdělení < 0; existuje více hodnot nadprůměrných než podprůměrných.

Charakteristiky špičatosti (excesu)

- špičatost spočívá ve větší nahuštěnosti hodnot prostřední velikosti ve srovnání se stupněm nahuštěnosti ostatních hodnot resp. všech hodnot proměnné.
- špičatější rozdělení má výraznější vrchol.

Míry špičatosti

prostá: $\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n s_x^4} - 3$ *vážená:* $\beta = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^4 n_i}{n s_x^4} - 3$

Interpretace charakteristik špičatosti:

- špičatější je to rozdělení, které má hodnotu β vyšší
- základem pro srovnání je normované normální rozdělení.