TECHNICKÁ UNIVERZITA V LIBERCI

Nové možnosti rozvoje vzdělávání na Technické univerzitě v Liberci

Specifický cíl A3: Tvorba nových profesně zaměřených studijních programů

NPO_TUL_MSMT-16598/2022

Popisné charakteristiky

Ing. Vladimíra Hovorková Valentová, Ph.D.

Financováno Evropskou unií NextGenerationEU

Národní plán obnovy

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Popisné charakteristiky

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

4 skupiny statistických charakteristik:

1. charakteristiky polohy (úrovně),
2. charakteristiky variability,
3. charakteristiky šikmosti,
4. charakteristiky špičatosti.

Dva způsoby konstrukce statistických charakteristik:

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

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

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

! 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: $\bar{x}_H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

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

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

! 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: $\bar{x}_K = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{\sum_{i=1}^{k} n_i}}$

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

Vztah mezi průměry

Jsou -li výše uvedené 4 typy průměrů vypočítány z týchž kladných hodnot proměnné, platí pro ně vztah:

$\bar{x}_H \leq \bar{x}_G \leq \bar{x} \leq \bar{x}_K$

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 (typická hodnota);
- vrchol rozdělení četností;
- označení symbolem $\hat{x}$.

Kvantily

- hodnoty, které rozdělují uspořádaný statistický soubor na určitý počet stejně obsazených částí;

- hodnoty menší resp. 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 *100p* je relativní četnost vyjádřená v %

Druhy kvantilů:

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

  Výpočet:

  a) 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 dané neklesající řadě hodnot.

  Při lichém rozsahu souboru je mediánem konkrétní prvek.

  b) rozsah souboru *n* je sudé číslo

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

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

Pozn.: Kvantily < $\tilde{x}$ se nazývají *dolní kvantily*, kvantily > $\tilde{x}$ se nazývají *horní kvantily*.

- **tercily** - $\tilde{x}_{33,\bar{3}}\ (x_{0,\bar{3}})$, $\tilde{x}_{66,\bar{6}}\ (x_{0,\bar{6}})$

  - jsou to 2 kvantily, které rozdělují uspořádaný statistický soubor na 3 stejně četné části.

- **kvartily** - $\tilde{x}_{25}\ (x_{0,25})$, $\tilde{x}$, $\tilde{x}_{75}\ (x_{0,75})$

  - jsou to 3 kvantily, které rozdělují uspořádaný statistický soubor na 4 stejně četné části.

- **kvintily** - $\tilde{x}_{20}\ (x_{0,20})$, $\tilde{x}_{40}\ (x_{0,40})$, $\tilde{x}_{60}\ (x_{0,60})$, $\tilde{x}_{80}\ (x_{0,80})$

  - jsou to 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.

Výpočet pořadového čísla kvantilu:

$$n \cdot p < m_p < n \cdot p + 1$$

n ..... rozsah statistického souboru

p ..... relativní četnost

m_p ..... pořadové číslo příslušného kvantilu

2. Charakteristiky variability

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

a) Míry absolutní variability

Variační rozpětí

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

Kvantilová rozpětí

- kvartilové rozpětí: $R_q = \tilde{x}_{75} - \tilde{x}_{25}$

- decilové rozpětí: $R_d = \tilde{x}_{90} - \tilde{x}_{10}$

Kvantilové odchylky

- kvartilová odchylka: $Q = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2}$

- decilová odchylka: $D = \frac{\tilde{x}_{90} - \tilde{x}_{10}}{8}$

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

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

Směrodatná odchylka

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

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

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

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

Vlastnosti rozptylu:

1) Rozptyl konstanty je roven 0.
2) Přičteme-li ke všem hodnotám znaku stejnou konstantu, rozptyl se nezmění.
3) Násobíme-li všechny hodnoty znaku stejnou nenulovou konstantou, je původní rozptyl násoben čtvercem této konstanty.
4) Rozklad rozptylu – Skládá-li se statistický soubor z *k* dílčích podsouborů, v nichž známe jednotlivé dílčí rozptyly s_i^2, dílčí průměry $\bar{x}_i$ a četnosti n_i, pak rozptyl celého souboru s_x^2 můžeme rozložit na součet 2 rozptylů, z nichž jeden charakterizuje variabilitu mezi skupinami a druhý variabilitu uvnitř skupin: $s_x^2 = s_{\bar{x}_i}^2 + \overline{s_i^2}$.

   Rozptyl skupinových průměrů: $s_{\bar{x}_i}^2 = \frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k}\bar{x}_i^2 n_i}{\sum_{i=1}^{k} n_i} - \left(\frac{\sum_{i=1}^{k}\bar{x}_i n_i}{\sum_{i=1}^{k} n_i}\right)^2$

   ***Průměr skupinových rozptylů*:** $\overline{s_i^2} = \frac{\sum_{i=1}^{k} s_i^2 n_i}{\sum_{i=1}^{k} n_i}$

b) Míry relativní variability

Variační koeficient

- je to 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 $\langle 0, \infty)$.

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

3. Charakteristiky šikmosti

- šikmost = asymetrie.

Momentová míra šikmosti α

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

Jednoduchá míra šikmosti α´ (Cyhelského míra šikmosti)

$$\alpha' = \frac{n' - n''}{n}$$

n´ ….. počet podprůměrných hodnot;

n´´ …. Počet nadprůměrných hodnot.

Interpretace:

- v **symetrickém** rozdělení α = 0; obvykle platí vztah: $\bar{x} = \hat{x} = \tilde{x}$; počet podprůměrných hodnot je stejný jako počet hodnot nadprůměrných;
- v **kladně sešikmeném** rozdělení α > 0; obvykle platí vztah: $\hat{x} < \tilde{x} < \bar{\bar{x}}$; více hodnot podprůměrných než nadprůměrných;
- v **záporně sešikmeném** rozdělení α < 0; obvykle platí vztah: $\bar{\bar{x}} < \tilde{x} < \hat{x}$; více hodnot nadprůměrných než podprůměrných.

4. Charakteristiky špičatosti

- špičatost = exces;
- špičatost spočívá ve větší nahuštěnosti hodnot prostřední velikosti ve srovnání se stupněm nahuštěnosti ostatních hodnot resp. všech hodnot proměnné;
- špičatější rozdělení má výraznější vrchol (tzn. že vrchol více vystupuje).

Momentová míra špičatosti β

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

Interpretace:

- vyšší hodnota míry znamená větší špičatost, tzn. špičatější je to rozdělení, které má β vyšší;
- základem pro srovnání je normované normální rozdělení (viz další výklad).

Charakteristiky nominální proměnné

1. **polohy**:

Modus $\hat{x}$

2. **variability**:

Míra mutability M

- Mutabilita = variabilita hodnot nominální proměnné;
- udává podíl dvojic jednotek s různou obměnou z celkového počtu všech možných dvojic jednotek;
- lze ji vyjádřit v %.

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

Nominální variance

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