Zpracování náhodného výběru – popisná statistika

Základní pojmy

- Úkolem statistiky je na základě vlastností výběrového souboru usuzovat o vlastnostech celé populace.
- **Populace** (**základní soubor**) je souhrn všech existujících prvků, které sledujeme při statistickém šetření (např. při volebních průzkumech je populace tvořena všemi občany ČR s právem volit).

Základní pojmy

- Jelikož je počet prvků populace zpravidla vysoký, je proto z časových, ekonomických a jiných důvodů provedení **vyčerpávajícího šetření** (tedy šetření celé populace) nereálné.
- Proto se zpravidla prování **výběrové šetření**, tj. šetření na vybrané části populace – výběr. Možností, jak výběr z populace provést, je více.

Základní pojmy

- Zpravidla provádíme **náhodný výběr** (každý prvek populace má stejnou šanci být do výběru zařazen). Údajům, které u souboru pozorujeme, říkáme **proměnné** (např. věk apod., značí se zpravidla velkými písmeny), jednotlivým hodnotám, kterých proměnná nabývá (nebo může nabývat), říkáme **varianty proměnné**.

Základní pojmy

- Proměnné můžeme rozdělit na proměnné:
  1) **Kvalitativní** – varianty proměnné jsou vyjádřeny slovně (např. pohlaví, národnost apod.).
  2) **Kvantitativní** – varianty proměnné jsou vyjádřeny číselně (např. věk, hmotnost apod.).
- Podle rozsahu výběru *n* zpravidla rozlišujeme:
  1) **Výběr malého rozsahu** – $n < 30$.
  2) **Výběr velkého rozsahu** – $n \geq 30$.

Základní pojmy

- Při zpracování náhodného výběru zavádíme pojem četnosti, přičemž rozeznáváme:
  1) **Absolutní četnosti** n_i,
  2) **Relativní četnosti** p_i,
  3) **Kumulativní četnosti** m_i,
  4) **Relativní kumulativní četnosti** F_i.

Základní pojmy

- Absolutní četnost n_i vyjadřuje, kolikrát se konkrétní varianta proměnné v_i v souboru objevila. Označíme-li k počet variant proměnné, které se v souboru vyskytly, pak musí platit:
$$\sum_{i=1}^{k} n_i = n.$$
- Varianty proměnné v_i seřazené podle velikosti a jejich absolutní četnosti tvoří **variační řadu**.

Základní pojmy

- Relativní četnost p_i je definována jako podíl četnosti n_i a rozsahu souboru n, tedy:

$$p_i = \frac{n_i}{n}.$$

- Je zřejmé, že dále musí platit:

$$\sum_{i=1}^{k} p_i = 1.$$

Základní pojmy

- Kumulativní četnost m_i je definována jako součet absolutních četností variant proměnné menší nebo rovno variantě v_i, tedy:

  $$m_i = \sum_{v \le v_i} n_i .$$

- Je zřejmé, že dále musí platit:

  $$m_k = n ,$$

  kumulativní četnost nejvyšší varianty proměnné je tedy rovna rozsahu souboru.

Základní pojmy

- Relativní kumulativní četnost F_i je definována jako podíl kumulativní četnosti m_i a rozsahu souboru n, tedy:

$$F_i = \frac{m_i}{n}.$$

- Je zřejmé, že dále musí platit:

$$F_k = 1.$$

Základní pojmy

- Grafické nebo tabulkové znázornění seřazených variant proměnné a jejich kumulativních četností se nazývá **distribuční funkce kumulativní četnosti**, příp. **empirická distribuční funkce**.

Základní pojmy

- Nechť je v_{min} minimální varianta proměnné, v_{max} maximální varianta proměnné. Potom interval $\langle v_{\min}; v_{\max} \rangle$ bývá označován jako **variační obor** proměnné.
- Rozdíl maximální a minimální varianty proměnné bývá označován jako **variační rozpětí** R:

$$R = v_{\max} - v_{\min} .$$

Výběrové charakteristiky

- Datový soubor získaný náhodným výběrem lze znázornit pomocí číselných charakteristik, které nazýváme výběrové charakteristiky, které zpravidla dělíme na:
  1) **Míry polohy** – určují typické rozložení hodnot souboru.
  2) **Míry variability** – určují variabilitu (rozptyl) hodnot kolem své typické hodnoty.

Míry polohy

- Mezi základní míry polohy se řadí:
  1) **Výběrový průměr** $\bar{x}$,
  2) **Modus** *Mod*,
  3) **Výběrové kvantily** x_p – především **medián** $x_{0,5}$.

Míry polohy

- Mějme náhodný výběr x_1, x_2,…, x_n. Výběrový průměr se nejčastěji stanovuje jako **aritmetický průměr** všech pozorování, tedy:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \, .$$

- Pro aritmetický průměr platí:

$$\sum_{i=1}^{n} \left(x_i - \bar{x} \right) = 0,$$

součet všech odchylek pozorovaných hodnot od jejich aritmetického průměru je roven 0.

Míry polohy

- Ne vždy je ale vhodné použít aritmetický průměr. V případech, kdy pracujeme s proměnnou vyjadřující relativní změny, používáme **geometrický průměr**:

$$\bar{x}_g = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}.$$

Míry polohy

- V případech, kdy pracujeme s proměnnou mající charakter části z celku, potom používáme **harmonický průměr**:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Míry polohy

- Modus je definován jako varianta proměnné s největší četností. Na rozdíl od průměru, který je pouze jeden, může mít statistický soubor více modů. Proměnnou s jedním modem nazýváme unimodální, proměnnou s dvěma mody bimodální.

Míry polohy

- Výběrový kvantil je obecně definován jako hodnota rozdělující výběrový soubor na dvě části – první část obsahuje hodnoty, které jsou menší než daný kvantil, a druhá část obsahuje hodnoty které jsou rovny nebo větší než hodnota daného kvantilu.
- Kvantil x_p nazýváme $100 \cdot p\%$-ní kvantil.

Míry polohy

- Rozeznáváme následují kvantily:
  1) **Percentily** – $x_{0,01}$, $x_{0,02}$,…, $x_{0,99}$.
  2) **Decily** – $x_{0,1}$, $x_{0,2}$,…, $x_{0,9}$.
  3) **Kvartily** – **dolní kvartil** $x_{0,25}$, medián $x_{0,5}$, **horní kvartil** $x_{0,75}$.

Míry polohy

- Postup při určování kvantilů:
  1) Datový soubor uspořádáme vzestupně podle velikosti.
  2) Seřazeným pozorováním přiřadíme pořadí od 1 do n.
  3) $100 \cdot p\%$-ní kvantil je potom roven pozorování s pořadím z_p, kde:

     $$z_p = n \cdot p + 0{,}5.$$

     Není-li z_p celé číslo, potom je příslušný kvantil roven aritmetickému průměru pozorování s pořadím $[z_p]$ a $[z_p] + 1$, kde $[z_p]$ označuje celou část čísla z_p.

Míry variability

- Mezi základní míry variability se řadí:
  1) Výběrový rozptyl s^2.
  2) Výběrová směrodatná odchylka s.
  3) Variační koeficient V_x.
  4) Variační rozpětí R.
  5) Interkvartilové rozpětí IQR.
  6) Medián absolutních odchylek od mediánu MAD.

Míry variability

- Výběrový rozptyl je definován vztahem:

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

- Nevýhodou rozptylu je, že jeho jednotka je druhou mocninou jednotky proměnné. Proto zavádíme výběrovou směrodatnou odchylku definovanou vztahem:

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

Míry variability

- Chceme-li porovnat variabilitu proměnných vyjádřených v různých jednotkách, použijeme k tomu variační koeficient definovaný:

  $$V_x = \frac{s}{\overline{x}}.$$

  Variační koeficient je bezrozměrný a vyjadřuje relativní míru variability proměnné.
- Variační rozpětí jsme již definovali jako:

  $$R = v_{\max} - v_{\min}.$$

Míry variability

- Interkvartilové rozpětí je definováno jako rozdíl horního a dolního kvartilu:

  $IQR = x_{0,75} - x_{0,25}$.

- Medián absolutních odchylek od mediánu stanovíme následujícím postupem:
  1) Stanovíme absolutní odchylky jednotlivých pozorování od mediánu, tedy $\left|x_i - x_{0,5}\right|$.
  2) Absolutní odchylky seřadíme vzestupně podle velikosti.
  3) Známým způsobem nalezneme medián absolutních odchylek, čili MAD.

Identifikace odlehlých pozorování

- Odlehlým pozorováním rozumíme pozorování, které se mimořádně liší od ostatních hodnot a tím ovlivňují reprezentativnost výběru. Nyní se tedy zaměříme na způsoby, jak odlehlá pozorování identifikovat. Nejčastěji se uvádí tři způsoby:
  1) Pomocí tzv. vnitřních hradeb.
  2) Pomocí *z*-souřadnice.
  3) Pomocí $x_{0,5}$-souřadnice (mediánová souřadnice).

Identifikace odlehlých pozorování

- ad 1) Za odlehlé pozorování lze považovat hodnotu x_i, která je od dolního, resp. od horního kvartilu vzdálena o více než 1,5 násobek interkvartilového rozpětí. Odlehlá pozorování tedy leží v intervalu:

$$\left(-\infty; x_{0,25} - 1{,}5 \cdot IQR\right) \cup \left(x_{0,75} + 1{,}5 \cdot IQR; \infty\right).$$

Identifikace odlehlých pozorování

- ad 2) Za odlehlé pozorování lze považovat hodnotu x_i, jejíž absolutní hodnota z-souřadnice je větší než 3, přičemž z-souřadnice je definována:

$$z - souř._i = \frac{x_i - \bar{x}}{s},$$

z-souřadnice tedy udává, kolikrát je pozorování x_i vzdáleno o hodnotu směrodatné odchylky od výběrového průměru.

Identifikace odlehlých pozorování

- ad 3) Za odlehlé pozorování lze považovat takovou hodnotu x_i, jejíž absolutní hodnota $x_{0,5}$-souřadnice je větší než 3, přičemž $x_{0,5}$-souřadnice je definována:

$$x_{0,5} - souř._i = \frac{x_i - x_{0,5}}{1{,}483 \cdot MAD}.$$

Zpracování rozsáhlého statistického souboru

- V případě, že máme rozsáhlý statistický soubor, sdružujeme jednotlivá pozorování do **tříd**.
- Zpravidla se volí konstantní šířka třídy (vyjma krajních tříd).
- Doporučuje se volit počet tříd v rozmezí 5 – 20.
- Každé pozorování musí být jednoznačně přiřazeno pouze do jedné třídy!

Zpracování rozsáhlého statistického souboru

- Pro stanovení počtu tříd existuje více pravidel, nejčastěji se setkáváme se **Sturgesovým pravidlem**, kterým stanovíme počet tříd *k* podle vztahu:

$$k \approx 1 + 3{,}3 \cdot \log n .$$

- Šířku třídy *h* potom stanovíme podle vztahu:

$$h \approx \frac{R}{k},$$

kde *R* je variační rozpětí.

Zpracování rozsáhlého statistického souboru

- Všechna pozorování zahrnuta v třídě *i* jsou potom reprezentována jednou zástupnou hodnotou – **třídním znakem** z_i, který je aritmetickým průměrem dolní a horní hranice třídy, tvoří tedy střed třídy.

Zpracování rozsáhlého statistického souboru

- Máme-li statistický soubor zadán pouze pomocí tříd *i* a jejich třídními četnostmi n_i, musíme pro výpočet základních výběrových charakteristik použít vztahy ve vážené formě:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} n_i \cdot z_i, s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{k} n_i \cdot (z_i - \bar{x})^2,$$

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{k} n_i \cdot (z_i - \bar{x})^2}.$$

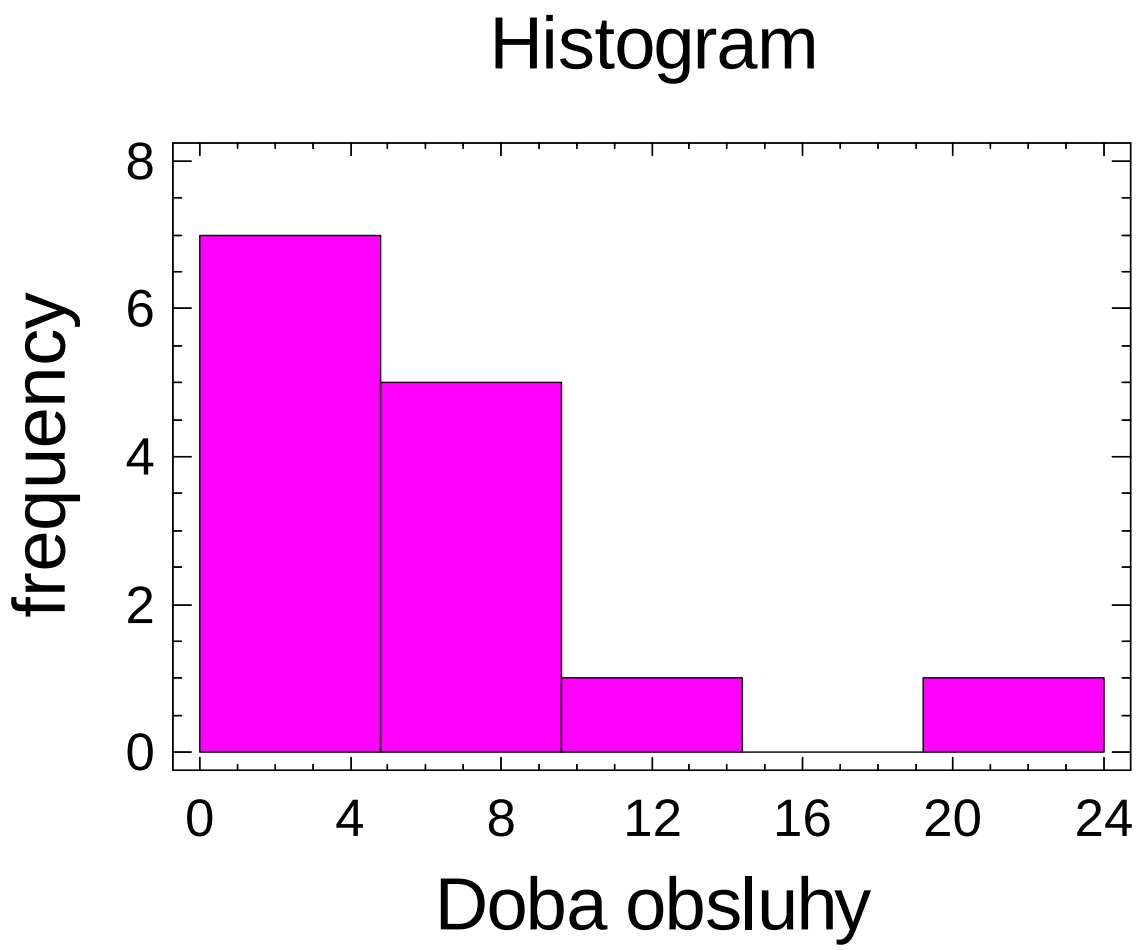
Grafické znázornění statistického souboru

- Základní typy grafů, které se používají:
  1) **Koláčový (výsečový) graf**.
  2) **Histogram**.
- Koláčový graf prezentuje relativní četnosti jednotlivých variant proměnné. Používá se pro menší počet variant proměnné.

Grafické znázornění statistického souboru

- Histogram je sloupcový graf, kde na vodorovnou osu vynášíme jednotlivé varianty proměnné, resp. třídy v případě souboru rozděleného na třídy, jednotlivé četnosti (absolutní nebo relativní) jsou potom zobrazovány jako sloupce.

Grafické znázornění statistického souboru



Grafické znázornění statistického souboru

- **Krabicový graf** je graf, který slouží k zakreslení základních výběrových charakteristik kvantitativní proměnné.