Základní, výběrový a datový soubor

Základním souborem rozumíme libovolnou neprázdnou množinu E. Prvky množiny E značíme ε a nazýváme je objekty. Libovolnou neprázdnou podmnožinu $\{\varepsilon_1, ..., \varepsilon_n\}$ základního souboru E nazýváme výběrový soubor rozsahu n. Je-li množina $G \subseteq E$, pak symbolem N(G) rozumíme absolutní četnost množiny G ve výběrovém souboru, tj. počet těch objektů množiny G, které patří do výběrového souboru. Relativní četnost množiny G ve výběrovém souboru zavedeme vztahem $p(G) = \frac{N(G)}{n}$.

Příklad: Základním souborem E je množina všech ekonomicky zaměřených studentů 1. ročníku českých vysokých škol. Množina G_1 je tvořena těmi studenty, kteří uspěli v prvním zkušebním termínu z matematiky a množina G_2 obsahuje ty studenty, kteří uspěli v prvním zkušebním termínu z angličtiny. Ze základního souboru bylo náhodně vybráno 20 studentů, kteří tvoří výběrový soubor $\{\varepsilon_1, ..., \varepsilon_{20}\}$. Z těchto 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech. Zapište absolutní a relativní četnosti úspěšných matematiků, angličtinářů a oboustranně úspěšných studentů.

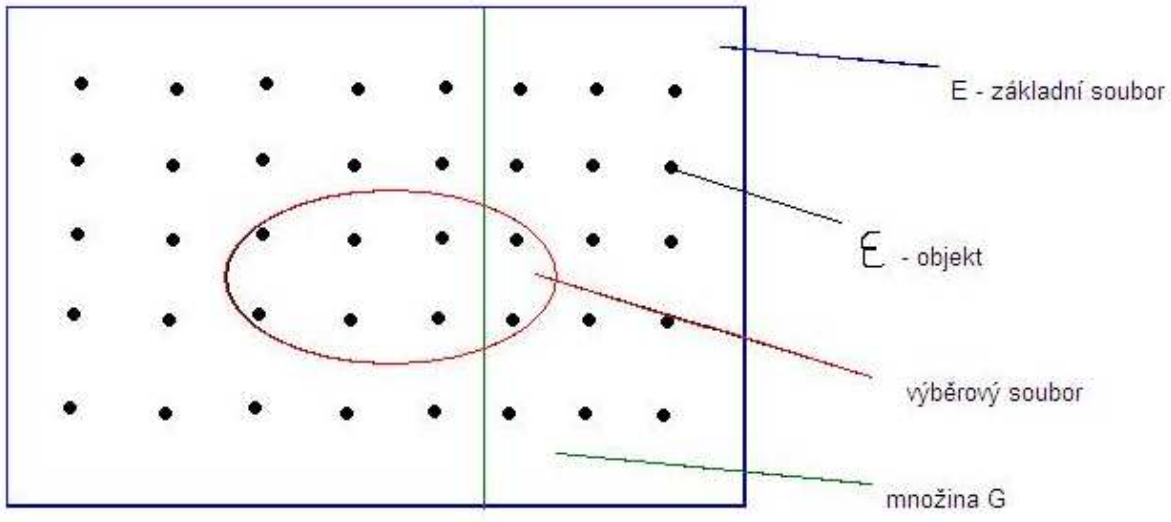
Řešení:

$$N(G_1) = 12, N(G_2) = 15, N(G_1 \cap G_2) = 11, n = 20, p(G_1) = \frac{12}{20} = 0{,}6, p(G_2) = \frac{15}{20} = 0{,}75,$$

$$p(G_1 \cap G_2) = \frac{11}{20} = 0{,}55$$

Vidíme, že úspěšných matematiků je 60%, angličtinářů 75% a oboustranně úspěšných studentů jen 55%.

Ilustrace



Vlastnosti relativní četnosti: Relativní četnost má následujících 12 vlastností, které jsou obdobné vlastnostem procent.

- $p(\varnothing) = 0$
- $p(G) \geq 0$ (nezápornost)
- $p(G) \leq 1$
- $p(G_1 \cup G_2) + p(G_1 \cap G_2) = p(G_1) + p(G_2)$
- $1 + p(G_1 \cap G_2) \geq p(G_1) + p(G_2)$
- $p(G_1 \cup G_2) + 0 \leq p(G_1) + p(G_2)$ (subaditivita)
- $G_1 \cap G_2 = \varnothing \Rightarrow p(G_1 \cup G_2) = p(G_1) + p(G_2)$ (aditivita)
- $p(G_2 \setminus G_1) = p(G_2) - p(G_1 \cap G_2)$
- $G_1 \subseteq G_2 \Rightarrow p(G_2 \setminus G_1) = p(G_2) - p(G_1)$ (subtraktivita)
- $G_1 \subseteq G_2 \Rightarrow p(G_1) \leq p(G_2)$ (monotonie)
- $p(E) = 1$ (normovanost)
- $p(G) + p(\overline{G}) = 1$ (komplementarita)

Pojem podmíněné relativní četnosti: Pokud se v daném základním souboru zajímáme o dvě podmnožiny, můžeme zavést pojem podmíněné relativní četnosti jedné podmnožiny v daném výběrovém souboru za předpokladu, že objekt pochází z druhé podmnožiny.
Nechť E je základní soubor, G_1, G_2 jeho podmnožiny, $\{\varepsilon_1, ..., \varepsilon_n\}$ výběrový soubor. Definujeme podmíněnou relativní četnost množiny G_1 ve výběrovém souboru za předpokladu G_2: $p(G_1/G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{p(G_1 \cap G_2)}{p(G_2)}$ a podmíněnou relativní četnost G_2 ve výběrovém souboru za předpokladu G_1: $p(G_2/G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{p(G_1 \cap G_2)}{p(G_1)}$.

Příklad: Pro údaje z příkladu o studentech vypočtěte podmíněnou relativní četnost úspěšných matematiků mezi úspěšnými angličtináři a podmíněnou relativní četnost úspěšných angličtinářů mezi úspěšnými matematiky.

Řešení:
$N(G_1) = 12, N(G_2) = 15, N(G_1 \cap G_2) = 11, n = 20,$
$p(G_1/G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{11}{15} = 0{,}73$ (tzn., že 73% těch studentů, kteří byli úspěšní v angličtině, uspělo i v matematice)
$p(G_2/G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{11}{12} = 0{,}92$ (tzn., že 92% těch studentů, kteří byli úspěšní v matematice, uspělo i v angličtině)

Pojem četnostní nezávislosti dvou množin: O četnostní nezávislosti dvou množin v daném výběrovém souboru hovoříme tehdy, když informace o původu objektu z jedné množiny nijak nemění šance, s nimiž soudíme na jeho původ i z druhé množiny.
V příkladě se studenty by množiny úspěšných matematiků a úspěšných angličtinářů byly četnostně nezávislé, pokud podíl úspěšných matematiků mezi úspěšnými angličtináři by byl stejný jako podíl úspěšných matematiků mezi všemi zkoušenými studenty a stejně tak podíl úspěšných angličtinářů mezi úspěšnými matematiky by byl stejný jako podíl úspěšných angličtinářů mezi všemi zkoušenými studenty, tj.

$$\frac{n(G_1 \cap G_2)}{n(G_2)} = \frac{n(G_1)}{n} \wedge \frac{n(G_1 \cap G_2)}{n(G_1)} = \frac{n(G_2)}{n}.$$

Po snadné úpravě dostaneme multiplikativní vztah

$$\frac{n(G_1 \cap G_2)}{n} = \frac{n(G_1)}{n} \cdot \frac{n(G_2)}{n}, \text{ tj. } p(G_1 \cap G_2) = p(G_1)p(G_2)$$

Řekneme tedy, že množiny G_1, G_2 jsou četnostně nezávislé v daném výběrovém souboru, jestliže $p(G_1 \cap G_2) = p(G_1)p(G_2)$.
(V praxi jen zřídka dojde k tomu, že uvedený vztah platí přesně. Většinou je jen naznačena určitá tendence četnostní nezávislosti.)

Příklad: Pro údaje z příkladu o studentech zjistěte, zda úspěchy v matematice a angličtině jsou v daném výběrovém souboru četnostně nezávislé.

Řešení:
$p(G_1 \cap G_2) = 0{,}55$, $p(G_1)p(G_2) = 0{,}6 \times 0{,}75 = 0{,}45$, tedy skutečná relativní četnost oboustranně úspěšných studentů je větší než by odpovídalo četnostní nezávislosti množin G_1, G_2 v daném výběrovém souboru. Znamená to, že úspěch v matematice se zpravidla sdružuje s úspěchem v angličtině a naopak.

Pojem skalárního a vektorového znaku: Vlastnosti objektů vyjadřujeme číselně pomocí znaků.
Nechť E je základní soubor. Funkce $X: E \to R$, $Y: E \to R$, ..., $Z: E \to R$, které každému objektu přiřazují číslo, se nazývají (skalární) znaky. Uspořádaná p-tice (X, Y, ..., Z) se nazývá vektorový znak.

Označení: Nechť je dán výběrový soubor $\{\varepsilon_1, ..., \varepsilon_n\} \subseteq E$. Hodnoty znaků X, Y, ..., Z pro i-tý objekt označíme $x_i = X(\varepsilon_i)$, $y_i = Y(\varepsilon_i)$, ..., $z_i = Z(\varepsilon_i)$, $i = 1, ..., n$.

Pojem datového souboru: Matice $\begin{pmatrix} x_1 & y_1 & \cdots & z_1 \\ x_2 & y_2 & \cdots & z_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_m & y_n & \cdots & z_n \end{pmatrix}$ typu n × p se nazývá datový soubor. Její řádky odpovídají jednotlivým objektům, sloupce znakům. Libovolný sloupec této matice nazýváme jednorozměrným datovým souborem. Jestliže uspořádáme hodnoty některého znaku (např. znaku X) v jednorozměrném datovém souboru vzestupně podle velikosti, dostaneme uspořádaný datový soubor $\begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}$, kde $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Vektor $\begin{pmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{pmatrix}$, kde $x_{[1]} < \ldots < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá vektor variant.

Příklad: Pro studenty z výběrového souboru uvedeného výše byly zjišťovány hodnoty znaků X – známka z matematiky v prvním zkušebním termínu, Y – známka z angličtiny v prvním zkušebním termínu, Z – pohlaví studenta (0 ... žena, 1 ... muž). Byl získán datový soubor

$$\begin{pmatrix} 2 & 2 & 0 \\ 1 & 3 & 1 \\ 4 & 3 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 4 & 4 & 1 \\ 3 & 3 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 4 & 2 & 1 \\ 4 & 4 & 0 \\ 2 & 2 & 0 \\ 4 & 3 & 1 \\ 2 & 3 & 1 \\ 4 & 4 & 0 \\ 1 & 1 & 0 \\ 4 & 3 & 1 \\ 4 & 4 & 1 \\ 1 & 3 & 0 \end{pmatrix}$$

Utvořte jednorozměrný uspořádaný i neuspořádaný datový soubor pro známky z matematiky a vektory variant pro známky z matematiky.

Řešení:

$$\begin{pmatrix}2\\1\\4\\1\\1\\4\\3\\3\\1\\1\\4\\4\\2\\4\\2\\4\\1\\4\\4\\1\end{pmatrix}, \begin{pmatrix}1\\1\\1\\1\\1\\1\\1\\2\\2\\2\\3\\3\\4\\4\\4\\4\\4\\4\\4\end{pmatrix}, \begin{pmatrix}1\\2\\3\\4\end{pmatrix}$$

Pojem jevu: Nechť $\{\varepsilon_1, ..., \varepsilon_n\}$ je výběrový soubor, X, Y, ..., Z jsou znaky, B, B_1, B_p jsou číselné množiny.
Zápis $\{X \in B\}$ znamená jev „znak X nabyl hodnoty z množiny B“ .
Zápis $\{X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p\}$ znamená jev „znak X nabyl hodnoty z množiny B_1 a současně znak Y nabyl hodnoty z množiny B_2 atd. až znak Z nabyl hodnoty z množiny B_p“.
Symbol $N(X \in B)$ značí absolutní četnost jevu $\{X \in B\}$ ve výběrovém souboru, tj. počet těch objektů ve výběrovém souboru, pro něž $x_i \in B$.
Symbol $p(X \in B)$ znamená relativní četnost jevu $\{X \in B\}$ ve výběrovém souboru, tj. $p(X \in B) = \frac{N(X \in B)}{n}$.
Analogicky $N(X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p)$ resp. $p(X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p)$ znamená absolutní resp. relativní četnost jevu $\{X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p\}$ ve výběrovém souboru.

Příklad: Pro datový soubor s údaji o známkách najděte relativní četnost
a) matematických jedničkářů
b) úspěšných matematiků
c) oboustranně neúspěšných studentů.

Řešení:

ad a) $p(X = 1) = \frac{7}{20} = 0{,}35;$

ad b) $p(X \leq 3) = \frac{12}{20} = 0{,}60;$

ad c) $p(X = 4 \wedge Y = 4) = \frac{4}{20} = 0{,}20.$

Zjistili jsme, že jedničku z matematiky mělo 35% studentů, zkoušku z matematiky úspěšně složilo 60% studentů a oboustranně neúspěšných bylo 20% studentů.

Shrnutí:

Předmětem statistického zájmu není jednotlivý objekt, nýbrž soubor objektů, tzv. **základní soubor**. Zpravidla není možné vyšetřovat všechny objekty, ale jenom určitý omezený počet objektů, které tvoří **výběrový soubor**.

Ty prvky základního souboru, které vykazují určitou společnou vlastnost, tvoří **množinu**. Statistik zkoumá **absolutní a relativní četnost množiny** v daném výběrovém souboru.

Zajímají-li nás ve výběrovém souboru dvě množiny, můžeme zkoumat výskyty objektů z jedné množiny mezi objekty pocházejícími z druhé množiny. Tím dospíváme k pojmu **podmíněné relativní četnosti**. Rovněž lze ověřovat **četnostní nezávislost** těchto dvou množin v daném výběrovém souboru. Četnostní nezávislost vlastně znamená, že informace o původu objektu z jedné množiny nijak nemění šance, s nimiž soudíme na jeho původ z druhé množiny.

Každému objektu základního souboru lze pomocí funkce zvané **znak** přiřadit číslo (nebo i více čísel). Pokud hodnoty znaků pro objekty daného výběrového souboru uspořádáme do matice tak, že řádky odpovídají jednotlivým objektům a sloupce znakům, dostaneme **datový soubor**. Libovolný sloupec této matice tvoří **jednorozměrný datový soubor**, který můžeme uspořádat podle velikosti a vytvořit tak **uspořádaný datový soubor** nebo z něj lze získat **vektor variant**.

Jevem rozumíme tu skutečnost, že znak nabyl hodnoty z nějaké číselné množiny. Můžeme zkoumat **absolutní a relativní četnost jevu** v daném výběrovém souboru.