Marie Budíková
Maria Králová
Bohumil Maroš

Průvodce základními statistickými metodami

- Jak se orientovat v datových souborech pomocí tabulek a grafů
- Jak pochopit zákonitosti matematiky náhody
- Posuzování rozdílnosti několika souborů
- Modelování závislostí pomocí regrese
- Jak získat z dat nápady ke zlepšení chodu firmy
- Řešení praktických úloh s využitím softwaru STATISTICA a MINITAB

GRADA

Upozornění pro čtenáře a uživatele této knihy

GRADA®

RNDr. Marie Budíková, Dr.,

Mgr. Maria Králová, Ph.D.,

Doc. RNDr. Bohumil Maroš, CSc.

Průvodce základními statistickými metodami

Vydala Grada Publishing, a.s.
U Průhonu 22, 170 00 Praha 7
tel.: +420 234 264 401, fax: +420 234 264 400
www.grada.cz
jako svou 4147. publikaci

Autorský kolektiv:
RNDr. Marie Budíková, Dr. – kapitoly 1, 2, 3, 8, 13, 14, 15, 16, 17, 18
Mgr. Maria Králová, Ph.D. – kapitoly 4, 5, 6, 7, 9, 10
Doc. RNDr. Bohumil Maroš, CSc. – kapitoly 11, 12, 19, 20, 21

Odborná recenzentka:
Doc. Ing. Eva Jarošová, CSc.

Vydání odborné knihy schválila Vědecká redakce nakladatelství Grada Publishing, a.s.

Odpovědný redaktor Petr Somogyi
Sazba Mgr. David Hampel, Ph.D.
Počet stran 272
První vydání, Praha 2010
Vytiskly Tiskárny Havlíčkův Brod, a.s.
Husova ulice 1881, Havlíčkův Brod

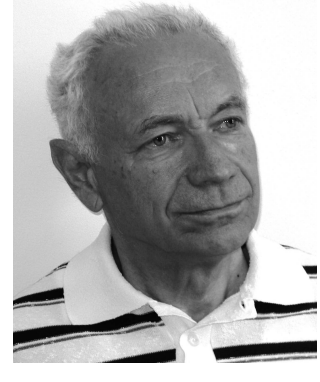
ISBN 978-80-247-3243-5 (tištěná verze)
ISBN 978-80-247-7511-1 (elektronická verze ve formátu PDF) © Grada Publishing, a.s. 2012

Obsah

O autorech

RNDr. Marie Budíková, Dr.
Působí jako lektorka Ústavu matematiky a statistiky Přírodovědecké fakulty Masarykovy univerzity v Brně. Má dlouholeté zkušenosti s výukou předmětů zaměřených na pravděpodobnost a statistiku, a to nejen pro studenty Přírodovědecké fakulty, ale též Ekonomicko-správní fakulty, Fakulty informatiky a Fakulty strojního inženýrství Vysokého učení technického v Brně. Od 90. let využívá při výuce statistické programové systémy SPSS a STATISTICA. Specializuje se na využití statistických metod v praxi, především v klimatologii, hydrologii, medicíně a psychologii. V současnosti úzce spolupracuje s brněnskými klimatology, kteří zkoumají polární klima na stanici J. G. Mendela v Antarktidě. Je autorkou či spoluautorkou více než 70 vědeckých a odborných článků, publikací a učebních textů. Řadu let působí ve výboru České statistické společnosti.

Mgr. Maria Králová, Ph.D.
V roce 1996 absolvovala Přírodovědeckou fakultu MU v Brně, obor matematika-biologie. Postgraduální studium ukončila v roce 2003 disertační prací Markovské modely pozornosti. Ve své výzkumné činnosti se zaměřuje zejména na stochastické modelování v psychologii, je spoluřešitelkou Výzkumného záměru MŠMT ČR CEZ: J22/98:261100009. V pedagogické praxi se věnuje především základním kurzům pravděpodobnosti a matematické statistiky, působí na Ekonomicko-správní fakultě Masarykovy univerzity v Brně. Je členkou České statistické společnosti.



Doc. RNDr. Bohumil Maroš, CSc.
Jako vysokoškolský učitel působí více než 45 let na Ústavu matematiky Fakulty strojního inženýrství Vysokého učení technického v Brně. Vyučuje zde základní i pokročilé metody statistického přístupu k řešení problémů. Více než 30 let se zabývá problematikou statistického zpracování dat a matematického modelování procesů. Úspěšně propaguje metody Design of Experiments nejen ve výuce, ale i v praxi pro výrobní podniky či organizace poskytující služby. Je externím spolupracovníkem společností SC&C Partner, která se specializuje na zlepšování jakosti výroby či služeb pomocí metody Six Sigma. Je autorem či spoluautorem více než 60 vědeckých publikací, 8 výzkumných zpráv, 6 knih, 6 skript. Je členem České statistické společnosti.

Úvodní slovo recenzenta

V knize jsou shrnuty základní vzorce a postupy z popisné statistiky, počtu pravděpodobnosti a matematické statistiky. Moderní pojetí statistiky předpokládá zpracovávání výběrových šetření nebo experimentálních dat, a proto je větší pozornost věnována počtu pravděpodobnosti a matematické statistice, která z něho vychází. Kniha obsahuje 21 kapitol, které je možné rozdělit na tři části. První tři kapitoly se zabývají převážně popisnou statistikou, ve čtvrté až desáté kapitole jsou vysvětleny základy počtu pravděpodobnosti a zbývajících jedenáct kapitol je věnováno matematické statistice. K ilustraci postupů slouží řešené příklady, pro procvičení látky jsou zařazeny další úlohy s uvedeným řešením. Řada příkladů obsahuje návod k obsluze statistického systému STATISTICA, v některých případech také systému MINITAB.

I když lze zjednodušeně říci, že většina metod se vyučuje v základních kurzech statistiky na různých vysokých školách nematematického zaměření, při bližším prostudování je zřejmé, že zvláště počet zařazených statistických testů společný základ přesahuje. Jako příklad testů, které nebývají zcela běžně v základní literatuře, můžeme uvést metody mnohonásobného porovnávání, speciální testy normality nebo shody s exponenciálním či Poissonovým rozdělením, testy pro čtyřpolní tabulky nebo některé testy o korelačních koeficientech.

Věřím, že Průvodce základními statistickými metodami využijí nejen studenti Ekonomicko-správní fakulty Masarykovy univerzity, jimž je kniha určena především, ale i řada dalších uživatelů statistických metod, a to zvláště ti, kteří pracují s některým ze statistických systémů STATISTICA nebo MINITAB.

doc. Ing. Eva Jarošová, CSc.
Katedra statistiky a pravděpodobnosti
Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze

Předmluva

Učebnice „Průvodce základními statistickými metodami“ je primárně určena posluchačům Ekonomicko-správní fakulty Masarykovy univerzity v Brně a má sloužit jako základní studijní literatura předmětů Statistika 1 a Statistika 2. Poučení v ní však samozřejmě najdou i posluchači jiných studijních zaměření či uživatelé statistiky v praxi, kteří potřebují provádět analýzu dat.

Autoři předpokládají, že čtenář je obeznámen se základy maticového počtu, diferenciálního a integrálního počtu a má zkušenosti s používáním tabulkového kalkulátoru.

Knihu tvoří 21 kapitol, z nichž první tři se zabývají popisnou statistikou, dalších sedm počtem pravděpodobnosti a zbývajících jedenáct pak matematickou statistikou. Partie, které se týkají popisné statistiky a počtu pravděpodobnosti, jsou probírány v předmětu Statistika 1, matematická statistika pak tvoří náplň předmětu Statistika 2.

Každá z kapitol zavádí základní pojmy statistické či pravděpodobnostní teorie a objasňuje je na praktických příkladech. Kapitoly z popisné a matematické statistiky rovněž obsahují návody na řešení příkladů pomocí statistického softwaru, konkrétně pomocí systému STATISTICA, v některých kapitolách je rovněž použit systém MINITAB, který se na některých vysokých školách technického zaměření používá jako základní statistický softwarový nástroj. Výsledné tabulky a grafy jsou podrobně komentovány. Své znalosti si může čtenář ověřit na příkladech určených k samostatnému řešení. Datové soubory k jednotlivým příkladům a statistické tabulky najdete na adrese www.math.muni.cz/∼budikova.

Autoři si jsou vědomi toho, že v jediné nepříliš rozsáhlé učebnici nelze vysvětlit všechny důležité a v praxi často používané statistické metody, ani se podrobněji zabývat jejich zajímavými aspekty. Proto uvažují o vytvoření druhého dílu této učebnice, který by sloužil jako průvodce pokročilejšími statistickými metodami.

Summary

This easy-to-follow publication containing many examples and case studies aims primarily at students of economics and professionals in economics and engineering. The readers will familiarize themselves with essential statistical techniques and practical applications of statistics. The book is divided into three parts. The first part deals with descriptive statistics, the second one treats the basic concepts of the probability theory and the third part looks at the selected methods of inductive statistics.

Each topic is introduced by a brief theoretical introduction followed by sample exercises and a set of tasks for self-study. It is sample tasks from economic and technical practice accompanied by the answer key that are the focal point of this publication. Careful attention is paid to verification of underlying assumptions of statistical techniques and to detailed interpretation of results. This book demonstrates solutions to problems using the STATISTICA (version 9) and MINITAB (version 15) computer software. Task solutions are supplemented by both detailed comments on computer outputs and instructions how to reach them.

Úvod

Teorie pravděpodobnosti
Teorie pravděpodobnosti je matematická disciplína (budovaná axiomaticky), která se zabývá studiem zákonitostí v náhodných pokusech a jejich modelováním matematickými prostředky. Pod pojmem náhoda rozumíme působení faktorů, které se živelně mění při různých provedeních téhož pokusu. (Je to „matematika náhody".)

Popisná statistika
Popisná statistika je disciplína, která popisuje a sumarizuje informace obsažené ve velkém množství dat pomocí tabulek, grafů, funkcionálních a číselných charakteristik. Činí tak pomocí základních matematických operací. Cílem je zpřehlednit informace „ukryté" v datových souborech.

Matematická statistika
Matematická statistika je věda, která buduje metody pro analýzu dat a využívá při tom princip statistické indukce. (Informace získané z náhodného výběru zobecňuje na základní soubor.) Její součástí je teorie odhadu, testování statistických hypotéz, statistická predikce.

Kapitola 1

Základní, výběrový a datový soubor

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

Prvky základního souboru, které vykazují určitou společnou vlastnost, tvoří množinu. Statistik zkoumá absolutní a relativní četnost prvků této množiny v daném výběrovém souboru.

Zajímají-li nás ve výběrovém souboru dvě množiny, můžeme zkoumat výskyt 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, které máme o původu objektu z jedné množiny, nijak nemění naše očekávání o původu objektu z druhé množiny.

Každému objektu základního souboru lze pomocí funkce zvané znak přiřadit číslo (nebo i více čísel). Pod těmito čísly se mohou skrývat i slovní popisy, např. hodnota 1 pro znak „rodinný stav“ znamená „svobodný“, hodnota 2 „ženatý“ apod. 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.

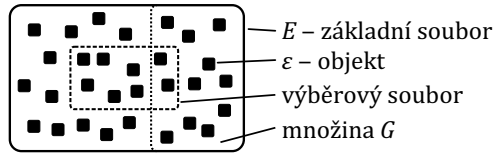
Jevem rozumíme tu skutečnost, že znak nabyl hodnoty z nějaké číselné množiny. Můžeme zkoumat absolutní a relativní četnost jevu v daném výběrovém souboru.

1.1 Základní a výběrový soubor, absolutní a relativní četnost množiny

Základním souborem rozumíme libovolnou neprázdnou množinu E. Prvky množiny E značíme ε a nazýváme je objekty. Libovolnou neprázdnou podmnožinu $\{\varepsilon_1, \ldots, \varepsilon_n\}$ 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 $p(G) = \frac{N(G)}{n}$.

Výběrový soubor by měl být reprezentativním obrazem základního souboru. Toho lze docílit např. tak, že objekty ze základního souboru vybíráme do výběrového souboru losováním. Výběrový soubor je ilustrován na obrázku 1.1.



Obrázek 1.1: *Ilustrace výběrového souboru.*

Příklad 1.1. Základním souborem E je množina všech žáků 9. ročníků základních škol. Množina G_1 je tvořena těmi žáky, kteří v pololetí 9. třídy uspěli v předmětu fyzika a množina G_2 obsahuje ty žáky, kteří v pololetí 9. třídy uspěli v předmětu chemie. Ze základního souboru bylo náhodně vybráno 30 žáků, kteří tvoří výběrový soubor $\{\varepsilon_1, \ldots, \varepsilon_{30}\}$. Z těchto 30 žáků 28 uspělo ve fyzice, 27 v chemii a 25 v obou předmětech. Zapište absolutní a relativní četnosti žáků úspěšných ve fyzice, v chemii a oboustranně úspěšných žáků.

Řešení
Spočetli jsme $N(G_1) = 28$, $N(G_2) = 27$, $N(G_1 \cap G_2) = 25$, $n = 30$, $p(G_1) = \frac{28}{30} = 0{,}9\overline{3}$, $p(G_2) = \frac{27}{30} = 0{,}9$, $p(G_1 \cap G_2) = \frac{25}{30} = 0{,}8\overline{3}$. Vidíme, že ve výběrovém souboru je 93,3 % žáků úspěšných ve fyzice, 90 % v chemii a oboustranně úspěšných žáků je 83,3 %. □

1.2 Vlastnosti relativní četnosti

Relativní četnost má následujících 12 vlastností, které jsou obdobné vlastnostem procent.

1. $p(\emptyset) = 0$,
2. $p(G) \geq 0$ (nezápornost),
3. $p(G_1 \cup G_2) + p(G_1 \cap G_2) = p(G_1) + p(G_2)$,
4. $1 + p(G_1 \cap G_2) \geq p(G_1) + p(G_2)$,
5. $p(G_1 \cup G_2) \leq p(G_1) + p(G_2)$ (subaditivita),
6. $G_1 \cap G_2 = \emptyset \Rightarrow p(G_1 \cup G_2) = p(G_1) + p(G_2)$ (aditivita),
7. $p(G_2 \backslash G_1) = p(G_2) - p(G_1 \cap G_2)$,
8. $G_1 \subseteq G_2 \Rightarrow p(G_2 \backslash G_1) = p(G_2) - p(G_1)$ (subtraktivita),
9. $G_1 \subseteq G_2 \Rightarrow p(G_1) \leq p(G_2)$ (monotonie),
10. $p(E) = 1$ (normovanost),

11. $p(G) + p(G') = 1$ (komplementarita),
12. $p(G) \leq 1$.

Pokud se v daném základním souboru zajímáme o dvě podmnožiny objektů, můžeme zavést pojem podmíněné relativní četnosti jedné podmnožiny v daném výběrovém souboru za předpokladu, že objekt pochází z druhé podmnožiny. V příkladu 1.2 vypočteme podmíněné relativní četnosti žáků úspěšných ve fyzice mezi žáky úspěšnými v chemii a naopak.

1.3 Podmíněná relativní četnost

Nechť E je základní soubor, G_1, G_2 jeho podmnožiny, $\{\varepsilon_1, \ldots, \varepsilon_n\}$ výběrový soubor. Podmíněná relativní četnost množiny G_1 ve výběrovém souboru za podmínky G_2 je zavedena vztahem

$$p(G_1 \mid G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{p(G_1 \cap G_2)}{p(G_2)}$$

a podmíněná relativní četnost G_2 ve výběrovém souboru za podmínky G_1 vztahem

$$p(G_2 \mid G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{p(G_1 \cap G_2)}{p(G_1)}.$$

Příklad 1.2. Pro údaje z příkladu 1.1 vypočtěte podmíněnou relativní četnost žáků úspěšných ve fyzice mezi žáky úspěšnými v chemii a podmíněnou relativní četnost žáků úspěšných v chemii mezi žáky úspěšnými ve fyzice.

Řešení
Spočetli jsme

$$p(G_1 \mid G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{25}{27} = 0{,}926.$$

Znamená to, že 92,6 % těch žáků, kteří byli úspěšní v chemii, uspělo i ve fyzice. Dále jsme spočetli

$$p(G_2 \mid G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{25}{28} = 0{,}893,$$

což znamená, že 89,3 % těch žáků, kteří byli úspěšní ve fyzice, uspělo i v chemii.

Obvykle se u žáků sdružuje úspěch ve fyzice s úspěchem v chemii a podobně neúspěch ve fyzice s neúspěchem v chemii, neboť osobní nadání a schopnosti působí v těchto dvou přírodovědných předmětech stejným směrem. □

V jaké situaci budeme hovořit o četnostní nezávislosti obou úspěchů? Zřejmě tehdy, když podíl žáků úspěšných ve fyzice (resp. chemii) mezi žáky úspěšnými v chemii (resp. fyzice) bude stejný jako podíl těchto žáků mezi všemi žáky ve výběrovém souboru. Požadujeme tedy, aby $p(G_2 \mid G_1) = p(G_2)$ a současně $p(G_1 \mid G_2) = p(G_1)$. Oba tyto požadavky jsou ekvivalentní s multiplikativním vztahem

$$p(G_1 \cap G_2) = p(G_1) \cdot p(G_2),$$

který slouží k definování četnostní nezávislosti dvou množin v daném výběrovém souboru.

1.4 Četnostní nezávislost dvou množin v daném výběrovém souboru

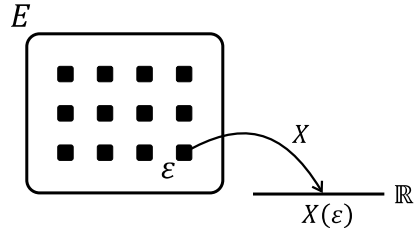
Množiny $G_1 \subseteq E$, $G_2 \subseteq E$ jsou četnostně nezávislé ve výběrovém souboru $\{\varepsilon_1, \ldots, \varepsilon_n\}$, jestliže platí multiplikativní vztah $p(G_1 \cap G_2) = p(G_1) \cdot p(G_2)$.

Informace o původu objektu z jedné množiny nijak neovlivňují naše očekávání, s nímž usuzujeme na jeho původ z druhé množiny. V našem příkladě se žáky tento multiplikativní vztah splněn není, neboť $\frac{25}{30} = 0{,}8\overline{3} \neq \frac{28}{30} \cdot \frac{27}{30} = 0{,}84$.

Nyní každý objekt základního souboru ohodnotíme jedním nebo více čísly pomocí funkce, která se nazývá znak. Čísla, která se vztahují pouze k objektům výběrového souboru, sestavíme do matice zvané datový soubor.

1.5 Skalární a vektorový znak

Funkce $X : E \to \mathbb{R}, Y : E \to \mathbb{R}, \ldots, Z : E \to \mathbb{R}$, které každému objektu přiřazují číslo, se nazývají (skalární) znaky (ilustrace viz obrázek 1.2). Uspořádaná p-tice $(X, Y, \ldots, Z)$ se nazývá vektorový znak.



Obrázek 1.2: *Ilustrace skalárního znaku.*

1.6 Datový soubor

Je-li dán výběrový soubor $\{\varepsilon_1, \ldots, \varepsilon_n\} \subseteq E$, pak hodnoty znaků $X, Y, \ldots, Z$ pro i-tý objekt označíme $x_i = X(\varepsilon_i)$, $y_i = Y(\varepsilon_i)$, …, $z_i = Z(\varepsilon_i)$, $i = 1, \ldots, n$. Matice

$$\begin{pmatrix} x_1 & y_1 & \ldots & z_1 \\ x_2 & y_2 & \ldots & z_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & \ldots & z_n \end{pmatrix}$$

typu $n \times p$ se nazývá datový soubor. Její řádky odpovídají jednotlivým objektům, sloupce znakům. Libovolný sloupec této matice nazýváme jednorozměrný datový soubor. Uspořádané

hodnoty např. znaku X: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ tvoří uspořádaný datový soubor

$$\begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}.$$

Vektor

$$\begin{pmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{pmatrix},$$

kde $x_{[1]} < \ldots < x_{[r]}$, $r \leq n$, jsou navzájem různé hodnoty znaku X, se nazývá vektor variant. Pozor – je důležité rozlišovat, zda je index bez závorky, v kulaté závorce či hranaté závorce!

Příklad 1.3. Pro žáky z výběrového souboru rozsahu 30 uvedeného v příkladu 1.1 byly zjišťovány hodnoty znaků X – známka z fyziky v pololetí, Y – známka z chemie v pololetí, Z – pohlaví žáka ($0 \ldots$ dívka, $1 \ldots$ hoch). Byl získán datový soubor (z důvodu úspory místa je matice typu 30×3 rozdělena na tři části).

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 3 & 1 \\ 2 & 3 & 0 \\ 4 & 5 & 1 \\ 2 & 1 & 1 \\ 4 & 3 & 1 \\ 2 & 2 & 1 \\ 4 & 2 & 0 \\ 3 & 3 & 0 \\ 4 & 5 & 0 \end{pmatrix} \quad \begin{pmatrix} 2 & 3 & 1 \\ 2 & 3 & 0 \\ 2 & 2 & 0 \\ 4 & 5 & 0 \\ 3 & 3 & 1 \\ 3 & 4 & 1 \\ 2 & 3 & 1 \\ 2 & 2 & 0 \\ 5 & 3 & 1 \\ 3 & 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 3 & 4 & 0 \\ 5 & 4 & 0 \\ 2 & 1 & 0 \\ 2 & 2 & 0 \\ 3 & 1 & 1 \\ 3 & 4 & 0 \\ 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix}$$

Utvořte jednorozměrný datový soubor pro znak X, vektory variant pro znaky X a Z a najděte x_2 , $x_{(2)}$, $x_{[2]}$.

Řešení

Jednorozměrný datový soubor pro znak X je sloupcový vektor o 30 řádcích. Opět z důvodu úspory místa ho napíšeme jako transponovaný řádkový vektor

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 & 5 & 5 \end{pmatrix}'.$$

Vektor variant pro znak X je $\begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$, pro znak Z $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, dále $x_2 = 2$, $x_{(2)} = 1$, $x_{[2]} = 2$.

□

1.7 Jev a jeho absolutní a relativní četnost

Jevem rozumíme skutečnost, že znak X nabyl hodnoty z číselné množiny B (formálně píšeme $\{X \in B\}$) resp. znaky $X, Y, \ldots, Z$ současně nabyly hodnot z číselných množin $B_1, \ldots, B_p$ (píšeme $\{X \in B_1 \wedge Y \in B_2 \wedge \ldots \wedge Z \in B_p\}$).

Je-li $\{\varepsilon_1, \ldots, \varepsilon_n\}$ výběrový soubor, pak zavedeme absolutní četnost jevu $\{X \in B\}$ ve výběrovém souboru (značíme ji $N(X \in B)$) jako počet těch objektů ve výběrovém souboru, pro něž znak $X \in B$. Relativní četnost jevu $\{X \in B\}$ ve výběrovém souboru je pak podíl absolutní četnosti a rozsahu souboru, tj.

$$p(X \in B) = \frac{N(X \in B)}{n}.$$

Analogicky $N(X \in B_1 \wedge Y \in B_2 \wedge \ldots \wedge Z \in B_p)$ resp. $p(X \in B_1 \wedge Y \in B_2 \wedge \ldots \wedge Z \in B_p)$ znamená absolutní resp. relativní četnost jevu $\{X \in B_1 \wedge Y \in B_2 \wedge \ldots \wedge Z \in B_p\}$ ve výběrovém souboru.

Např. pro datový soubor z příkladu 1.3 vyjádříme absolutní četnost dívek, které mají z obou sledovaných předmětů jedničku nebo dvojku, následovně: $N(X \leq 2 \wedge Y \leq 2 \wedge Z = 0) = 6$. Nebo relativní četnost žáků, kteří propadají z fyziky, pak můžeme vyjádřit takto: $p(X = 5) = \frac{2}{30} = 0{,}0\overline{6}$.

Příklady k samostatnému řešení

1.A Nechť množiny G_1, G_2 jsou neslučitelné, relativní četnost množiny G_1 je $0{,}15$, relativní četnost sjednocení množin G_1, G_2 je $0{,}8$. Vypočtěte relativní četnost množiny G_2.

$$[p(G_2) = p(G_1 \cup G_2) - p(G_1) = 0{,}8 - 0{,}15 = 0{,}65]$$

1.B Nechť množina G_1 je podmnožinou množiny G_2 a její relativní četnost je $0{,}33$. Relativní četnost rozdílu množin $G_2 \backslash G_1$ je $0{,}15$. Vypočtěte relativní četnost množiny G_2.

$$[p(G_2) = p(G_2 \backslash G_1) + p(G_1) = 0{,}15 + 0{,}33 = 0{,}48]$$

1.C Nechť relativní četnost rozdílu množin $G_1 \backslash G_2$ je $0{,}36$ a relativní četnost jejich průniku je $0{,}12$. Vypočtěte relativní četnost množiny G_1.

$$[p(G_1) = p(G_1 \backslash G_2) + p(G_1 \cap G_2) = 0{,}36 + 0{,}12 = 0{,}48]$$

1.D Je dán dvourozměrný datový soubor

$$\begin{pmatrix} 2 & 1 \\ 2 & 0 \\ 1 & 0 \\ 4 & 2 \\ 4 & 2 \\ 3 & 2 \\ 3 & 1 \\ 5 & 3 \\ 5 & 2 \\ 2 & 0 \end{pmatrix}.$$

Znak X (levý sloupec) znamená počet členů domácnosti a znak Y (pravý sloupec) počet dětí do 15 let v této domácnosti.

a) Utvořte uspořádané datové soubory pro znaky X a Y.

b) Najděte vektory variant znaků X a Y.

c) Vypočtěte relativní četnost tříčlenných domácností.

d) Vypočtěte relativní četnost nejvýše tříčlenných domácností.

e) Vypočtěte relativní četnost bezdětných domácností.

f) Vypočtěte relativní četnost dvoučlenných bezdětných domácností.

g) Vypočtěte podmíněnou relativní četnost dvoučlenných bezdětných domácností.

[Ad a), uspořádané datové soubory:

$$\begin{aligned}&\text{pro } X\text{: } \begin{pmatrix} 1 & 2 & 2 & 2 & 3 & 3 & 4 & 4 & 5 & 5 \end{pmatrix}',\\ &\text{pro } Y\text{: } \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 2 & 2 & 2 & 2 & 3 \end{pmatrix}'.\end{aligned}$$

Ad b), vektor variant pro znak X: $\begin{pmatrix} 1\\2\\3\\4\\5 \end{pmatrix}$, pro znak Y: $\begin{pmatrix} 0\\1\\2\\3 \end{pmatrix}$.

Ad c), relativní četnost tříčlenných domácností je $0{,}2$. Ad d), relativní četnost nejvýše tříčlenných domácností je $0{,}6$. Ad e), relativní četnost bezdětných domácností je $0{,}3$. Ad f), relativní četnost dvoučlenných domácností je $0{,}3$. Ad g), podmíněná relativní četnost těch dvoučlenných domácností, které jsou bezdětné, je $0{,}\overline{6}$.]

Kapitola 2

Bodové a intervalové rozložení četností

Informace obsažené v datovém souboru můžeme zpřehlednit pomocí tabulek a grafů. V této kapitole si ukážeme způsoby tabulkového a grafického zpracování pro jednorozměrné a dvourozměrné datové soubory. Tyto metody zvolíme podle toho, jak se liší počet variant znaku od rozsahu výběrového souboru.

2.1 Jednorozměrné bodové rozložení četností

Jestliže počet variant znaku X v jednorozměrném datovém souboru není příliš velký ve srovnání s rozsahem výběrového souboru, pak přiřazujeme četnosti jednotlivým variantám a hovoříme o bodovém rozložení četností.

Základní pojmy

Nechť je dán jednorozměrný datový soubor $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, v němž znak X nabývá r variant. Pro $j = 1, \ldots, r$ definujeme absolutní četnost varianty $x_{[j]}$ ve výběrovém souboru

$$n_j = N\left(X = x_{[j]}\right),$$

relativní četnost varianty $x_{[j]}$ ve výběrovém souboru

$$p_j = \frac{n_j}{n},$$

absolutní kumulativní četnost prvních j variant ve výběrovém souboru

$$N_j = N\left(X \leq x_{[j]}\right) = n_1 + \ldots + n_j$$