UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Statistická analýza intervalových dat v symbolic data analysis

Vedoucí bakalářské práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Aneta Andrášiková**
Studijní program: B1103 Aplikovaná matematika
Studijní obor: Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2015

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Aneta Andrášiková

Název práce: Statistická analýza intervalových dat v symbolic data analysis

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2015

Abstrakt: Ve statistické praxi se v dnešní době můžeme poměrně často setkat s rozsáhlými datovými soubory, které obsahují až stovky tisíců pozorování. Práce s takovými soubory je umožněna vyspělou úrovní počítačů, především pak statistických softwarů, které analýzu s nimi usnadňují. Otázkou však zůstává, jak hromadná data z těchto souborů analyzovat. Nové pojetí, které přináší užitečný nástroj analýzy datových souborů, neboť shrnuje datové soubory do „rozumných“ rozměrů, je označováno jako analýza symbolických dat, neboli symbolic data analysis (SDA). Nejdříve vysvětlíme základní myšlenky přístupu SDA. Dále tento koncept zúžíme na případ intervalových dat, jež představují jeden z typů proměnných v SDA a nacházejí své využití v praxi. Teoretické poznatky využijeme pro zpracování datového souboru týkajícího se počasí.

Klíčová slova: Symbolická data, Intervalová data, Reprezentace pomocí distribuční funkce, Parametrická reprezentace

Počet stran: 64

Počet příloh: 9

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Aneta Andrášiková

Title: Statistical analysis of interval data in symbolic data analysis

Type of thesis: Bachelor's

Department:
Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2015

Abstract: In statistical practice nowadays we quite often meet up with large data sets that contain hundreds of thousands of observations. Dealing with these files is possible due to advanced level of computers, especially of statistical software to analyze them easier. However, the question remains, how to analyze mass data from these data sets. The new approach that delivers a useful tool for analyzing data sets, because it summarizes the data set to „reasonable" values, is known as symbolic data analysis (SDA). First, we explain the basic ideas SDA approach. Furthermore, for this concept we turn our attention to the case of interval data, which represent one of variable types, and demonstrate their use in practice. Theoretical knowledge will be used for processing data sets related to the weather.

Key words: Symbolic data, Interval data, Representation using the distribution function, Parametric distribution

Number of pages: 64

Number of appendices: 9

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením doc. RNDr. Karla Hrona, Ph.D. s použitím uvedené literatury.

V Olomouci dne 19. března 2015

Obsah

Úvod **7**

1 Od klasických dat k symbolickým **9**

2 Základní myšlenka symbolic data analysis **13**

3 Představení nových typů proměnných **15**
3.1 Vícehodnotová proměnná 17
3.2 Modálně hodnotová proměnná 19
3.3 Intervalová proměnná 22

4 Intervalová proměnná **23**
4.1 Intervalová aritmetika 23
4.2 Reprezentace intervalových dat pomocí distribuční funkce 24
4.3 Parametrická reprezentace intervalových dat 36

5 Řešení reálného datového souboru **41**

Závěr **52**

Literatura **53**

Přílohy **55**

A Datový soubor „Krevní tlak“ **55**
A.1 Reprezentace intervalových dat pomocí distribuční funkce – empirická funkce hustoty 55
A.2 Reprezentace intervalových dat pomocí distribuční funkce – histogram 56
A.3 Parametrická reprezentace intervalových dat 57

B Datový soubor „Okresy“ **58**
B.1 Výpočet mezí intervalů pro nadmořské výšky 0 – 200 m a 701 – 800 m 61
B.2 Upravený datový soubor „Okresy“ pro proměnnou Y_1 61
B.3 Upravený datový soubor „Okresy“ pro proměnnou Y_2 62
B.4 Reprezentace intervalových dat pomocí distribuční funkce – empirická funkce hustoty 62
B.5 Reprezentace intervalových dat pomocí distribuční funkce – histogram 63
B.6 Parametrická reprezentace intervalových dat 64

Poděkování

Ráda bych poděkovala vedoucímu bakalářské práce doc. RNDr. Karlu Hronovi, Ph.D. za obětavost a čas, který mi věnoval při konzultacích. Dále děkuji Mgr. Janě Vrbkové, Ph.D. za konzultace při práci se statistickým softwarem R. Taktéž bych chtěla poděkovat Centru dopravního výzkumu, v. v. i. za poskytnutá data. Děkuji také své rodině za podporu během celého studia.

Úvod

Statistická teorie nám přináší řešení a postupy, jak pracovat s problematicky velkými datovými soubory, označovanými jako „databáze“ či „datasety“. Běžná praxe však nabízených metodik ve většině případů nepoužívá. Jedním ze zásadních důvodů je nejistota, zda datový soubor, se kterým pracujeme, je opravdu výběrem z celé populace a nikoli pouze z její části (viz [3]). Tato nejistota může být (dle [10]) ve značné míře odstraněna, pokud prvky datového souboru získáme náhodným výběrem. Řekněme, že by cílem našeho experimentu bylo určit průměrnou výšku obyvatelstva. Pokud bychom potřebná měření výšky prováděli například v šatně basketbalistů, nebo naopak při tréninku baletek, náš závěr z provedené studie by byl pravděpodobně nesprávný, ať už nadhodnocený nebo podhodnocený. Tímto jsme narazili na jednu z potřebných vlastností datového souboru, jedná se o reprezentativnost. Nehledě na to, že použitelnost některých statistických metod jednoznačně závisí na velikosti souboru.

Nesnáze při zpracování souborů s velkými rozsahy výběru nás přivádí k novému pojetí práce s daty nesoucí označení SDA. V současnosti se tímto pojetím zatím zabývá jen několik málo zahraničních odborníků v [3], [4], [6] a [11].

Představme si situaci, kdy máme k dispozici n pozorování a u každého z nich p proměnných. V malém měřítku může jít například o třídu třiceti dětí, přičemž každého z nich se zeptáme na jeho věk, výšku, váhu, počet sourozenců atd. a sběr dat je tímto u konce. Při takto malých hodnotách počtu jednotek n (rozsahu výběru) a proměnných p můžeme provést klasickou analýzu.

V praxi se ovšem často jedná řádově o stovky tisíců jednotek, v případě n, a stovky či více proměnných, v případě p, tedy $n \gg p$. Tato situace již pro klasické pojetí statistické analýzy přináší svá úskalí ve formě komplikovanosti výpočtů. Hlavním důvodem, proč potřebujeme vypracovat nové postupy pro analýzu dat, jsou velké soubory, které dnes oproti minulosti převládají.

Cílem této práce je popis vybraných přístupů symbolic data analysis (dále SDA) ke statistické analýze intervalových dat. Nejprve se zaměříme na koncept SDA. Zjistíme, proč vznikl a v čem se liší od klasického přístupu, popíšeme si

také nové typy proměnných, bez kterých se dále neobejdeme. Následně se budeme zabývat intervalovými daty, zmíníme se mimo jiné o intervalové aritmetice. Především si popíšeme dva z možných přístupů reprezentace intervalových dat – pomocí distribuční funkce a parametricky. V obou případech si teoretické poznatky vysvětlíme na ilustrativním příkladu. Poslední kapitola je věnována aplikacím těchto dvou metod na reálný datový soubor.

1. Od klasických dat k symbolickým

V této kapitole se budeme věnovat důvodům, proč jsou pro analýzu některých datových souborů standardní metody nepostačující a je potřeba hledat vhodnější alternativu, která datový soubor shrne do „rozumných“ rozměrů a zároveň zohlední vnitřní variabilitu dat.

Představme si skupinu lidí, kteří spolu jedou autobusem. Skupina zde představuje statistický soubor, který obsahuje statistické jednotky (lidi). Ačkoliv se nemusí striktně jednat o konkrétní osobu, použijeme ji jako názornou pomůcku. Konkrétní jednotlivec se může lišit od ostatních například svým věkem, výškou, bydlištěm, zaměstnáním. Podle [11] záleží především na prováděné analýze to, jakými proměnnými se budeme zabývat. Pokud by tedy ve zmíněném autobuse jel konkrétní pacient, tak by nás kromě jeho věku měl zajímat také jeho životní styl, prodělané nemoci, genetické predispozice. Naopak u spolusedícího klienta žádajícího o půjčku je prioritním příjem, úspory, v některých situacích také trestní rejstřík. Nejprve si tedy musíme celou studii důkladně naplánovat a teprve pak začít se sběrem dat.

Získané údaje (stejně jako v [11]) zapisujeme v klasickém případě do podoby datového pole, kde každá buňka v i-tém řádku a j-tém sloupci obsahuje hodnotu proměnné j pro jednotlivce i.

i	krevní skupina	váha v kg	počet operací
1	B	65	2
2	B	70	0
3	AB	87	1
4	A	59	0
5	0	62	1

Tabulka 1: Klasické datové pole.

Tabulka 1 ukazuje příklad datového souboru, kde pro každého z pěti pacientů známe hodnoty jedné kvalitativní proměnné (krevní skupina) a dvou kvantitativních proměnných (váha, počet operací)[1]. Sloupec i označuje jednotlivce a slouží

[1]Dělení proměnných dat klasického typu je popsáno např. v [2].

pouze k rozlišení pozorování (v tomto případě osob). Podívejme se na interpretaci údajů prvního pacienta. Bezprostředně zjistíme, že pacient, kterého jsme označili číslem 1, má krevní skupinu B, váží 65 kg a podstoupil 2 operace. Obdobně bychom postupovali též u popisu dat dalších pacientů.

Rozvoj počítačové techniky znamená také větší objem datových souborů. Analytik pak řeší dva hlavní problémy, a to dimenzi datové sady a sdružování podle cílových skupin. Rozsáhlost dat ale není neřešitelným problémem. Data můžeme shrnout do tabulky takových rozměrů, aby práce s ní byla méně problematická (viz [3] nebo [6]).

Představme si tento postup na příkladu medicínských dat. Zaznamenávání si všech dat jednotlivých pacientů za jednotku času je, řekněme na krajské nebo celostátní úrovni, příliš podrobné a dle [3] se tedy provádí specifické zjednodušení. Ať už se jedná o údaje, které shrneme za delší časové období (např. měsíc, rok), změnu formátu dostupných dat na interval, jehož krajní meze ohraničují naměřené hodnoty (u výšky např. 145 cm – 210 cm), třídění dle druhu onemocnění (např. TBC, žloutenka, rakovina, zánět mozkových blan, ...), či jejich kombinaci (např. {podvýživa, 40 kg – 50 kg}, {nadváha, 150 kg – 200 kg}).

Vnitřní strukturou se tato data liší od klasických dat, neboť představují rozmezí možných nabývaných hodnot, seznamy, v dalších případech také histogramy. Na taková data nemůžeme dále pohlížet tradičním způsobem, proto je označujeme pojmem symbolická data[2] (dále SD).

Dalším ilustrativním příkladem může být charakteristika barvy očí tří osob, kterým pro rozlišení náhodně přiřadíme číslice 1, 2, 3. Řekněme, že

$$\begin{aligned} &\text{osoba } 1 = \{\text{modrá, zelená}\}, \\ &\text{osoba } 2 = \{\text{černá}\}, \\ &\text{osoba } 3 = \{0{,}25 \text{ hnědá, } 0{,}25 \text{ modrá, } 0{,}50 \text{ šedá}\}. \end{aligned}$$

Všimněme si, že naše zkoumaná proměnná „barva“ nemusí být nutně pro da-

[2]Překlad z anglického „symbolic data“ (viz [3]).

nou jednotku jediná, ale může se jednat o seznam barev, či výčet s odpovídající proporcí.

Při analýze se setkáváme nejen s datovými soubory, které jsou již od svého počátku strukturovány jako symbolické, ale také se soubory, které mají zprvu klasický ráz a ke změně vnitřního uspořádání a přiblížení se k přívlastku „symbolické“ dochází až později (viz [3]), a to v závislosti na konkrétních datech.

Rozdíl mezi analýzou jednotlivce a skupiny je mnohem zásadnější, než může být na první pohled viditelné. Při analýze skupiny je totiž potřeba vzít v úvahu vnitřní variabilitu zkoumaného celku. Doposud běžné klasické pojetí je však příliš omezeno a onu proměnlivost a nejistotu zahrnutou v datech zanedbává (viz [11]).

Představme si výzkum zabývající se pacienty jistých zdravotnických zařízení. K dispozici máme informace týkající se jejich věku, anamnézy a BMI[3]. Mohlo by nás napadnout určit průměry nebo nejčastěji se opakující hodnoty v rámci celku či každé instituce zvlášť. Tímto přístupem však ztratíme důležité informace a nezjistíme v podstatě nic zásadního. Bez uvážení vnitřní variability jednotlivých tříd nebudeme schopni podat souhrnné a obecně platné popisy a závěry. Této nesnázi je potřeba se vyhnout. Proto vznikla myšlenka konceptu SDA, který napozorovanou variabilitu zohledňuje. Než však budeme ve výkladu pokračovat, zmiňme několik důležitých událostí a jmen, které sehrály v rozvoji SDA velikou roli (viz [11]).

Původní myšlenka konceptu SDA měla jasný cíl, a to umožnit analýzu datových souborů, které svou strukturou nevyhovují klasickým modelům. O představení SDA veřejnosti se na konci 80. let minulého století zasloužil E. Diday, jehož publikace inspirují autory dodnes.

Pro další vývoj SDA byl významný především vznik projektu SODAS (Symbolic Objects Data Analysis System). Přispěl výzkumu vůbec první statistickou softwarovou knihovnou pro SDA, díky níž pak mohli běžní uživatelé i odborní analytici tvořit, upravovat a také analyzovat SD. Navíc byla ve stejnou dobu publikována první kniha [5], která se zabývá právě problematikou SDA.

[3]BMI (body mass index) je počítáno jako váha v kilogramech vydělena druhou mocninou výšky v metrech.

V projektu SODAS dále pokračovalo ASSO (Analysis System of Symbolic Official data), publikující druhou knihu [7]. Projekt se zabýval vylepšením a tvorbou dalších statistických knihoven.

V následujících letech vývoj SDA pokračoval a kromě různých publikací v mezinárodních časopisech, byla vydána další kniha [4], která obsahuje úvod do problematiky, zabývá se rozdíly mezi klasickými a symbolickými daty, dále popisuje typy proměnných, metody popisné statistiky pro jednu náhodou veličinu, posléze také pro dvourozměrný případ aj.

Z dříve především evropského centra výzkumu se SDA rozšířila mezi týmy z celého světa a nalezla uplatnění v mnoha odvětvích. Zaměřme se nyní na SDA podrobněji.

2. Základní myšlenka symbolic data analysis

Základním stavebním kamenem pro statistiku jsou statistické jednotky. Zmínili jsme se o nich v kontextu konkrétních jednotlivců. V případě analýzy rozsáhlých datových souborů jsou však častěji za statistické jednotky brány celky vyšší úrovně (viz [11]).

Ve spojení s uvedeným příkladem medicínských dat na krajské úrovni nás pravděpodobně budou více zajímat data jedné nemocnice (resp. nemocnic z jednoho města) než všechny samotné údaje týkající se jednotlivých pacientů. Nezapomínejme, že shrnutím dat, ať už se jedná o konkrétní formu souhrnu pro statistické jednotky či třídění dle určitých kritérií, nesmíme opomenout nanejvýš důležitou variabilitu uvnitř tříd.

Nové typy proměnných, se kterými se v tomto pojetí setkáváme (viz [3], [4] nebo [11]), nezapomínají na onu variabilitu. Nazýváme je symbolické proměnné[4]. Jejich výhodou je bezpochyby to, že pro každý případ dovolují předpokládat mnohonásobné, případně vážené, hodnoty. Tedy každá proměnná může zároveň nabývat několika hodnot, některé s větší váhou, např. podle aktuálnosti pozorování. Stejně jako v klasickém případě se zde setkáváme s náhodnými veličinami, přesto však jiné povahy. Tuto odlišnost zdůrazňuje pojem „symbolické“ (viz [11]).

Uvažujme nyní dobu potřebnou k přípravě středoškolského studenta k testu. Patrně nás nepřekvapí, že kromě různě dlouhého času stráveného učením je významným činitelem pro změnu délky času zvláště charakteristika prostředí. Pokud se student bude učit v tichém prostředí, hodnota proměnné bude náležet intervalu, řekněme $\langle 60\text{ min}, 180\text{ min}\rangle$. V případě rušivého prostředí se hodnoty intervalu můžou změnit, např. $\langle 150\text{ min}, 250\text{ min}\rangle$.

Opačně, může se stát, že je pro nás potřebný spíše rozbor uvádějící rozložení četností použitých výukových materiálů, tedy například (učebnice 90 %, zvuková nahrávka 10 %). Zde opět uveďme, že se jedná o sumarizaci dat vztahující se k více dnům, týdnům. Pro příklad takovýchto dat se podívejme na tabulku 2.

Zdůrazněme, že dané hodnoty mají původ v několika provedených měřeních.

[4]Překlad z anglického „symbolic variables“ (viz [11]).

u	doba přípravy na test	výukové materiály
Gymnázium 1	⟨20 min, 50 min⟩	U (75 %), ZN (25 %)
Gymnázium 2	⟨35 min, 65 min⟩	U (35 %), ZN (65 %)
Gymnázium 3	⟨15 min, 70 min⟩	D (10 %), U (50 %), ZN (40 %)
Gymnázium 4	⟨50 min, 90 min⟩	D (20 %), U (80 %)

Tabulka 2: Symbolická datová tabulka, zde označme U (učebnice), ZN (zvukové nahrávky), D (dokumenty).

V prvním případě jde o opakovaná měření vykonaná v tichém prostředí, řekněme po dobu několika dnů, týdnů. Ze zjištěných údajů byla vybrána minimální a maximální hodnota, které pak utvořily daný interval. Pokud bychom shromažďovali informace vztahující se k jednotlivým dnům, týdnům pro každý typ prostředí, výsledná tabulka dat by obsahovala přílišné množství hodnot, a proto jsme provedli shrnutí do intervalů.

Strukturou zápisu se data podobají klasickému případu, neboť jsou zobrazena v matici (viz tabulka 1). Označujeme ji však pojmem symbolická datová tabulka (viz [11]), protože každá buňka obsahuje SD. Řádky odpovídají jednotlivým celkům a informují o symbolickém popisu, sloupce pak obsahují údaje o symbolických proměnných (blíže viz kapitola 3).

Ve sloupci „doba přípravy na test“ můžeme vidět intervaly vypovídající o časovém rozmezí potřebného k přípravě na test v rámci jednotlivých vzdělávacích institucí. Další proměnnou jsou využívané „výukové materiály“. Sloupec u slouží opět k rozlišení pozorování – v tomto případě institucí. Rozdílné označení identifikačního sloupce zdůrazňuje „symbolické“ pojetí dat. Z uvedených údajů je zřejmé, že nejdéle připravující se studenti navštěvují Gymnázium 4 a ti, kteří studují na Gymnáziu 3, používají nejvíce forem výukových materiálů.

3. Představení nových typů proměnných

Nyní si přiblížíme různé druhy nově zavedených proměnných, se kterými se při pojetí SDA můžeme setkat (viz [3], [4] nebo [11]). Tato kapitola je strukturována tak, aby definice a charakteristiky proměnných byly doplněny i vlastními příklady, které napomůžou lepšímu pochopení tématu. Definice a poznámky čerpají z [4]. Nejprve si ve stručnosti připomeňme dělení proměnných v klasickém pojetí, které je popsáno např. v [2] nebo [8].

V klasickém případě proměnné dělíme na kvalitativní nebo kvantitativní. Kvalitativní (neboli kategoriální) proměnné, jejichž hodnoty vyjadřujeme slovně, dále rozlišujeme na nominální a ordinální - podle toho, zda je lze jednoznačně seřadit či nikoli. Kvantitativní (neboli numerické) proměnné vyjadřujeme číselně, jejich hodnoty vybíráme z podmnožin reálné přímky $\mathbb{R}$ a můžeme se s nimi setkat v podobě měřitelné anebo pořadové. Podrobnějším výkladem se v této práci však zabývat nebudeme, místo toho se nyní zaměřme na proměnné SD.

Proměnné SD se vyznačují tím, že jednotlivá pozorování zpravidla nabývají několika hodnot zároveň. Pokud by nějaké pozorování nabývalo pouze jediné hodnoty, jedná se o speciální případ, stejný jako u klasických dat. Pro ostatní situace, které v konceptu SDA nastávají častěji, je ale (dle [11]) potřeba zavést nové proměnné, které nabývání více hodnot současně umožňují. Dále si uvedeme některé z těchto specifických proměnných. Podívejme se na následující vlastní příklad.

Příklad 3.1. (Obchodní cestující)
Představme si evropský sjezd obchodních cestujících (dále OC) jedné nadnárodní společnosti. Každý z OC má ve své správě jedno město, které navštěvuje, a se kterým zprostředkovává obchodní styk. Povinností všech OC je mimo jiné každý první den v měsíci podat zprávu o tom, kolik za minulý měsíc vydělali a jaké množství zakázek se jim podařilo provést.

Ředitel této nadnárodní společnosti si získané informace každý měsíc sepisuje do tabulky, aby měl o provedené práci svých podřízených přehled a mohl se

na základě jejich práce rozhodnout o tom, komu udělí prémii a koho naopak pokárá. Konkrétním příkladem takových dat je tabulka 3.

i	stát	město	věk	doprava	výdělek	počet zakázek
1	Polsko	Poznaň	26	A	15 000 Kč	10
2	ČR	Brno	35	A	19 000 Kč	12
3	ČR	Praha	31	A	45 000 Kč	16
4	Polsko	Katovice	22	V	29 000 Kč	11
5	Francie	Lyon	54	L	31 000 Kč	8
6	Itálie	Milán	39	A	27 000 Kč	14
7	Německo	Berlín	32	V	38 000 Kč	18
8	ČR	Olomouc	28	V	24 000 Kč	9
9	Polsko	Varšava	43	A	33 000 Kč	13
10	Francie	Paříž	38	L	49 000 Kč	21
11	Itálie	Řím	36	L	48 000 Kč	19
12	Německo	Frankfurt	42	V	36 000 Kč	15
13	ČR	Ostrava	37	A	16 000 Kč	6
14	Itálie	Neapol	45	A	21 000 Kč	11
15	Francie	Nantes	28	V	31 000 Kč	13
16	Německo	Mnichov	29	L	29 000 Kč	10
17	Polsko	Lešno	31	A	18 000 Kč	7
18	Itálie	Benátky	35	V	25 000 Kč	12
19	Německo	Hamburk	51	L	33 000 Kč	15
20	ČR	Pardubice	43	V	19 000 Kč	8
21	Polsko	Gdaňsk	39	A	27 000 Kč	14
22	Německo	Drážďany	27	A	34 000 Kč	16
23	Francie	Rennes	32	L	23 000 Kč	9
24	Itálie	Bologna	34	V	28 000 Kč	11
25	Francie	Orléans	44	A	22 000 Kč	10
26	Polsko	Štětín	31	V	11 000 Kč	5
27	Itálie	Padova	38	A	25 000 Kč	10
28	Německo	Ulm	24	L	19 000 Kč	8
29	Francie	Dijon	30	L	24 000 Kč	9
30	ČR	Jihlava	47	V	20 000 Kč	7

Tabulka 3: OC, zde označme A (osobní automobil), L (letadlo), V (vlak).

Každý z řádků představuje konkrétního jedince, přesněji statistickou jednotku. Použitá čísla 1, ..., 30 ve sloupci i jsou přiřazena zcela náhodně a slouží pouze k identifikaci. O každém z třiceti OC máme k dispozici tři kategorické proměnné (stát, město, doprava) a tři numerické proměnné (věk, výdělek, počet zakázek).

Tabulka 3 má v této podobě klasický ráz. Pokud se však ředitel rozhodne zaměřit zvlášť na jednotlivé státy, dojde ke změně konceptu. Místo tabulky klasických dat dostaneme tabulku SD, obsahující jiné typy proměnných než doposud. Podívejme se na ně blíže.

Poznámka 3.1. Dále uvažujme i-tého jednotlivce, kde $i \in \Omega = \{1, \ldots, n\}$, přičemž $n \in \mathbb{N}$ je počet jednotlivců v datovém souboru; a dále u-tý koncept/kategorii ω_u nabývající hodnot z množiny $E = \{w_1, \ldots, w_m\}$, kde E představuje množinu m symbolických konceptů/kategorií.

Poznámka 3.2. Označme klasickou hodnotu či realizaci j-té proměnné Y_j, kde $j = 1, \ldots, p$, na jednotce $i = 1, \ldots, n$, jako x_{ij}, tedy $Y_j(i) = x_{ij}$, zatímco symbolickou hodnotu či realizaci j-té symbolické proměnné označíme jako ξ_{ij}, tedy $Y_j(i) = \xi_{ij}$.

Poznámka 3.3. Hodnoty x_{ij}, které označují konkrétní hodnoty pro jednotlivce i na j-té náhodné veličině Y_j, tvoří matici $\mathbf{X}$ o rozměrech $n \times p$. Obdobně pro symbolický případ, hodnota realizace proměnné Y_j, kterou měříme na u-té kategorii w_u, odpovídá zápisu $Y_j(w_u) = \xi_{uj}$ a hodnoty ξ_{uj} tvoří matici $\boldsymbol{\xi}$. Dále, jestliže obor hodnot náhodné veličiny Y_j označíme $\mathcal{Y}_j$, $j = 1, \ldots, p$, matice $\mathbf{X}$ nabývá hodnot z $\mathcal{X} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_p = X_{i=1}^{p} \mathcal{Y}_j$ a matice $\boldsymbol{\xi}$ nabývá hodnot z $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_p$.

Neopomeňme, že v klasickém pojetí každé $x_{ij} \in \mathcal{X}$ nabývá přesně jedné možné hodnoty, oproti tomu v symbolickém pojetí každé ξ_{uj} nabývá několika hodnot zároveň.

3.1. Vícehodnotová proměnná

Definice 3.1. Symbolická vícehodnotová náhodná veličina Y je taková veličina, která pro každou realizaci nabývá jedné nebo více hodnot ze svého oboru hodnot $\mathcal{Y}$. Výčet možných hodnot obsažených v $\mathcal{Y}$ je konečný. V případě kvalitativní

proměnné se jedná o množinu kategorií, v případě kvantitativní proměnné o množinu reálných čísel.

Příklad 3.2. Předpokládejme, že se ředitel nadnárodní společnosti rozhodl, že s některým státem ukončí spolupráci. Potřebuje tedy spolehlivým způsobem posoudit, který ze států prosperuje nejméně. Modifikací předchozí tabulky získáme tabulky 4 a 5.

u	město
ČR	{Brno, Jihlava, Olomouc, Ostrava, Pardubice, Praha}
Francie	{Dijon, Lyon, Nantes, Orléans, Paříž, Rennes}
Itálie	{Benátky, Bologna, Milán, Neapol, Padova, Řím}
Německo	{Berlín, Drážďany, Frankfurt, Hamburk, Mnichov, Ulm}
Polsko	{Gdaňsk, Katovice, Lešno, Poznaň, Štětín, Varšava}

Tabulka 4: Vícehodnotová kvalitativní proměnná.

u	počet zakázek
ČR	{6, 7, 8, 9, 12, 16}
Francie	{8, 9, 10, 13, 21}
Itálie	{10, 11, 12, 14, 19}
Německo	{8, 10, 15, 16, 18}
Polsko	{5, 7, 10, 11, 13, 14}

Tabulka 5: Vícehodnotová kvantitativní proměnná.

Jak můžeme z údajů tabulek 4 a 5 vidět, místo údajů týkajících se konkrétních jednotlivců jsme data roztřídili podle příslušnosti k určitému státu. První proměnnou je „město“. Jedná se o vícehodnotovou kvalitativní proměnnou, protože množina obsahuje kategorie z definičního oboru. Proměnná „počet zakázek“ je však vícehodnotová kvantitativní, neboť jde o množinu čísel.

3.2. Modálně hodnotová proměnná

Definice 3.2. Nechť náhodná veličina Y nabývá hodnot z množiny $\{\eta_k;\, k \in \mathbb{N}\}$ v rámci oboru hodnot $\mathcal{Y}$. Potom se jedná o modální proměnnou, pokud má tvar:

$$Y(w_u) = \xi_u = \{\eta_k,\, \pi_k;\, k = 1, \ldots, s_u\}\,,$$

pro pozorování u, kde π_k je nezáporná míra spojená s η_k a kde s_u je počet hodnot skutečně vzatých z oboru hodnot $\mathcal{Y}$. Možné η_k můžou být konečné či nekonečné, kvalitativní či kvantitativní.

Míry $\{\pi_k\}$ jsou obvykle váhy, pravděpodobnosti, nebo relativní četnosti, které odpovídají příslušným výstupům η_k. Množina výstupů $\{\eta_k\}$ může být složena z kategoriálních hodnot (viz definice 3.3), nebo podmnožinou reálné osy $\mathbb{R}$ (viz definice 3.4).

Definice 3.3. Nechť $\mathcal{Y}_{cat}$ je oborem hodnot možných výsledků vícehodnotové náhodné proměnné Y_{cat}, přičemž $\mathcal{Y}_{cat} = \{\eta_1,\, \eta_2,\, \ldots\}$. Potom modální vícehodnotovou proměnnou nazveme takovou proměnnou, která nabývá hodnot patřících do podmnožiny oboru hodnotu $\mathcal{Y}_{cat}$ s nezápornými mírami připojenými ke každé hodnotě v oné podmnožině. Jednotlivé pozorování pro kategorii w_u nabývá tvaru:

$$Y\,(w_u) = \xi_u = \{\eta_{u1},\, p_{u1}; \ldots; \eta_{us_u},\, p_{us_u}\}\,,$$

kde $\{\eta_{u1},\, \ldots,\, \eta_{us_u}\} \subseteq \mathcal{Y}_{cat}$. Výsledek η_{uk} nastane s váhou p_{uk}, kde $k = 1,\, \ldots,\, s_u$ a s_u je počet hodnot skutečně vzatých z oboru hodnot $\mathcal{Y}_{cat}$. Dále platí, že $\sum_{k=1}^{s_u} p_{uk} = 1$, $\forall u = 1,\, \ldots,\, m$.

Příklad 3.3. Uvažme v příkladu OC proměnnou „doprava“. V tabulce 6 vidíme, že jediná proměnná je rozdělena do několika podkategorií, v tomto případě do tří. Jsou to podkategorie „osobní automobil“, „letadlo“ a „vlak“. Zaměřme se na OC

u	doprava		
	os. automobil	letadlo	vlak
ČR	1/2	0	1/2
Francie	1/6	2/3	1/6
Itálie	1/2	1/6	1/3
Německo	1/6	1/2	1/3
Polsko	2/3	0	1/3

Tabulka 6: Modálně hodnotová proměnná.

spravující města ČR. Z tabulky 3 jsme zjistili, že tito OC ve třech ze šesti případů používají osobní automobil a zbývající tři z nich využívají jako dopravu vlak. Z dat tedy můžeme udělat závěr, že 50 % z českých OC používá osobní automobil a dalších 50 % vlak, neboť $\frac{3}{6} = \frac{1}{2} = 0{,}5 = 50$ %. Údaje shrneme do této podoby:

$$Y_{\text{doprava}}(\text{ČR}) = \{\text{osobní automobil, 0,50; letadlo, 0,00; vlak, 0,50}\}.$$

Proměnná „doprava“ je zde modálně hodnotová.

Definice 3.4. Nechť je Y kvantitativní náhodná veličina, která nabývá hodnot na konečném počtu nepřekrývajících se intervalů $\{\langle a_k, b_k)\,,\ k \in \mathbb{N}\}$, kde $a_k \leq b_k$. Potom je výsledek pozorování w_u pro proměnnou histogram tvaru:

$$Y\,(w_u) = \xi_u = \{\langle a_{uk}, b_{uk})\,,\ p_{uk};\ k = 1, \ldots, s_u\}\,,$$

kde $s_u < \infty$ je konečný počet intervalů, na kterých je výsledná hodnota $Y\,(w_u)$ odpovídající pozorování w_u nenulová, a p_{uk} je váha pro konkrétní podinterval $\langle a_{uk}, b_{uk})$, kde $k = 1, \ldots, s_u$. Intervaly (a_k, b_k) můžou být otevřené či uzavřené. Dále platí, že $\sum_{k=1}^{s_u} p_{uk} = 1$, $\forall u = 1,\ \ldots,\ m$.

Zásadním krokem při práci s histogramovými proměnnými je omezení hodnot dolní a horní mezí a výpočet četností mezi určenými hranicemi. Tímto postupem si zajistíme zachování dostatku informací, které jsou pro nás bezesporu klíčové.

Přirozeně platí, že pro různá pozorování w_u může být počet a délka podintervalů odlišný. Dále předpokládáme, že hodnoty všech pozorování mají uvnitř podintervalu, do kterého patří, rovnoměrné rozdělení.

Poznamenejme, že intervalové proměnné, jimž bude věnována další kapitola, jsou obecně považovány za speciální případ histogramových proměnných, neboť pro $k = 1$ dochází (dle [11]) k degradaci histogramové proměnné na pouhý interval. Zároveň, histogramové proměnné představují numerickou obdobu modálně hodnotových proměnných, které jsme představili v definici 3.3.

Příklad 3.4. Je známo, že na výkonnost zaměstnanců má vliv mimo jiné i jejich věk – ať už se jedná o jejich vitálnost, nové nápady nebo naopak dlouholeté zkušenosti. Tabulka 7 nám prozradí bližší informace o věkové struktuře společnosti OC. Hodnoty v jednotlivých řádcích označují proporcionální zastoupení.

u	věk		
	< 30	$\langle 30, 40 \rangle$	> 40
ČR	1/6	1/2	1/3
Francie	1/6	1/2	1/3
Itálie	0	5/6	1/6
Německo	1/2	1/6	1/3
Polsko	1/3	1/2	1/6

Tabulka 7: Histogramová proměnná.

Podívejme se opět na údaje patřící OC ČR. Z dříve uvedené tabulky 3 víme, že OC ČR navštěvují Brno, Jihlavu, Olomouc, Ostravu, Pardubice, Prahu a mají 28, 31, 35, 37, 43, 47 let.Všimněme si, že v takto nově upravené tabulce je jediná proměnná „věk“ rozdělena do tří kategorií, a to mladší než třicet let, třicet až čtyřicet let, starší než čtyřicet let. Proměnnou „věk“ v tomto příkladu označujeme jako histogramovou proměnnou. Spočítejme si pravděpodobnost příslušnosti jednotlivce do konkrétní kategorie. První z OC, který má 28 let, patří svým věkem do kategorie < 30. Protože má ČR dohromady šest OC, ale do této kategorie patří pouze jeden z nich, pak p_{uk} odpovídá hodnotě $\frac{1}{6} = 0,167$, tedy 16,7 %. Počítáme obdobně i pro další kategorie $\langle 30, 40 \rangle$ a > 40. Kontrolou nám může býti

to, že podle předchozí definice platí: $\sum_{k=1}^{s_u} p_{uk} = 1,\ \forall u = 1,\ \ldots,\ m$. Provedením příslušných výpočtů jsme zjistili, že 16,7 % z českých OC jsou mladší třiceti let, 50 % patří do intervalu $\langle 30,\ 40\rangle$ a 33,3 % z nich jsou starší čtyřiceti let. Zápis výsledků je možné provést takto:

$$Y_{\text{věk}}\left(\text{ČR}\right) = \{\langle < 30\rangle\,,\ 0{,}167;\ \langle 30,\ 40\rangle\,,\ 0{,}50;\ \langle > 40\rangle\,,\ 0{,}333\}\,.$$

3.3. Intervalová proměnná

Definice 3.5. Symbolickou intervalovou náhodnou veličinou Y nazveme proměnnou, jejímiž hodnotami jsou intervaly. Neboli $Y = \xi = \langle a,\ b\rangle \subset \mathbb{R}$, přičemž platí nerovnost $a \leq b$, kde $a,\ b \in \mathbb{R}$. Interval ξ může být definován jako otevřený, polootevřený (polouzavřený) nebo uzavřený.

Příklad 3.5. Nejvíce vypovídající proměnnou je v našem příkladu týkající se OC rozmezí výdělků (viz tabulka 8). Jak můžeme vidět, nejméně lukrativní se z hlediska výdělku jeví Polsko, neboť horní mez jeho intervalu je nejmenší. Proto se ředitel společnosti rozhodne ukončit spolupráci právě s ním. Proměnná „výdělek“ nabývá hodnoty intervalu, proto se jedná o intervalovou proměnnou.

stát	výdělek
ČR	⟨16 000 Kč, 45 000 Kč⟩
Francie	⟨22 000 Kč, 49 000 Kč⟩
Itálie	⟨21 000 Kč, 48 000 Kč⟩
Německo	⟨19 000 Kč, 38 000 Kč⟩
Polsko	⟨11 000 Kč, 33 000 Kč⟩

Tabulka 8: Intervalová proměnná.

4. Intervalová proměnná

Intervalové proměnné nás ve spojení s SDA zajímají nejvíce, proto jim věnujme následující kapitolu. S intervalovými daty, zkráceně intervaly, se můžeme setkat (viz [6] nebo [11]) v mnoha různých situacích a odvětvích. Nemusí se striktně jednat o data původní. K intervalům vede i agregace rozsáhlých datových souborů jednotlivých pozorování. Z dalších možných zdrojů tak uveďme například původní symbolická data.

Pro práci s intervalovými proměnnými se nabízí možnost využít intervalovou aritmetiku. V následující podkapitole však uvidíme, že se při jejím využití setkáme s jistými problémy.

4.1. Intervalová aritmetika

Intervalovou matematiku, jak intervalovou aritmetiku také nazýváme, představil M. Warmus v druhé polovině dvacátého století, a to konkrétně roku 1956 (více v [12]). Díky ní nyní dokážeme určit meze chyb vznikajících při zaokrouhlování. Své využití má mimo jiné při optimalizačních problémech a také při řešení rovnic. Když srovnáme intervalovou aritmetiku s klasickou, nalezneme zásadní rozdíl. Představme si, že chceme odhadnout maximální denní teplotu v konkrétním místě. Místo toho, abychom klasickou aritmetikou odhadli 25 °C, pomocí intervalové aritmetiky dosáhneme hodnoty z intervalu $\langle 24{,}7\ °\mathrm{C};\ 25{,}3\ °\mathrm{C}\rangle$.

Klasická aritmetika definuje operace na jednotlivých číslech, kdežto intervalová aritmetika definuje soubor operací na intervalech takto:

$$X = T \circ S = \{x \mid \exists y \in T \wedge \exists z \in S : x = y \circ z\},$$

kde T, S, X odpovídají intervalům a x, y, $z \in \mathbb{R}$. Symbol "$\circ$" představuje jednu z operací sčítání, odečítání, násobení a dělení.

V následující definici si uvedeme základní operace intervalové aritmetiky. Uvědomme si, že nepracujeme s konkrétním reálným číslem x, ale s koncovými body intervalu $\langle a, b\rangle$, přičemž $x \in \langle a, b\rangle$.

Definice 4.1. Nechť $\langle a, b\rangle$, $\langle c, d\rangle \subset \mathbb{R}$. Pak jsou operace sčítání, odečítání, násobení a dělení definovány takto:

$\langle a, b\rangle + \langle c, d\rangle = \langle a + c, b + d\rangle$,

$\langle a, b\rangle - \langle c, d\rangle = \langle a - d, b - c\rangle$,

$\langle a, b\rangle \cdot \langle c, d\rangle = \langle \min\{a \cdot c, a \cdot d, b \cdot c, b \cdot d\}, \max\{a \cdot c, a \cdot d, b \cdot c, b \cdot d\}\rangle$,

$\langle a, b\rangle / \langle c, d\rangle = \langle \min\{a/c, a/d, b/c, b/d\}, \max\{a/c, a/d, b/c, b/d\}\rangle$,

přičemž pro poslední vztah platí podmínka: $0 \notin \langle c, d\rangle$.

Kromě běžných vlastností sčítání a násobení, jako jsou komutativita a asociativita, platí také sub-aditivita, podle níž $X(Y + Z) \subset XY + XZ$, kde X, Y, Z označují jednotlivé intervaly. Distributivní zákon jako takový ovšem v intervalové aritmetice neplatí, což její praktické použití pro statistické zpracování intervalových dat poněkud znesnadňuje.

Pro podrobnější informace o intervalové aritmetice odkazuji zájemce na [9] nebo [13], ze kterých jsem v této podkapitole čerpala. Nyní se již zaměřme na konkrétní možnosti přístupů práce s intervalovými daty.

4.2. Reprezentace intervalových dat pomocí distribuční funkce

Tato podkapitola povede na řešení ilustrativního příkladu jedním z možných přístupů analýzy dat. Nejprve si uveďme několik dále potřebných definicí, které vycházejí z [4]. Pro lepší názornost jsou připojeny i vlastní příklady.

Poznámka 4.1. Zdůrazněme, že označení „bod“ předpokládá v symbolickém pojetí hodnoty hyperkostky (či hyperobdélníku), oproti klasickému pojetí osvojujícímu si hodnoty samotného bodu v p-rozměrném prostoru.

Definice 4.2. Nechť má náhodná veličina Y_j, kde $j = 1, \ldots, p$, obor hodnot $\mathcal{X} = X_{j=1}^{p}\mathcal{Y}_j$. Potom se každý bod $\mathbf{x} = (x_1, \ldots, x_p) \in \mathcal{X}$ nazývá popisný vektor.

Příklad 4.1. Pojem popisného vektoru si vysvětlíme na dříve uvedené tabulce 1, která zobrazuje příklad klasických dat. Uvažme některé z proměnných, například Y_1, jež označuje „krevní skupinu“ a Y_3, která přináší informace o „počtu operací“. Hodnota $\mathbf{x}(w_1) = (\text{B}, 2)$ odpovídá pozorování $i = 1$. Popisný vektor tedy obsahuje informace konkrétního jednotlivce.

Obdobně pro popisné vektory $\mathbf{x}(w_2), \ldots, \mathbf{x}(w_5)$, které se vztahují ke zbývajícím pozorováním $i = 2, \ldots, 5$.

Definice 4.3. Nechť má náhodná veličina Y_j, kde $j = 1, \ldots, p$, obor hodnot $\mathcal{X} = X_{j=1}^{p} \mathcal{Y}_j$ a dále nechť $D_j \subseteq \mathcal{Y}_j$. Pak p-rozměrný podprostor D, přičemž $D = (D_1, \ldots, D_p) \subseteq \mathcal{X}$, nazveme popisnou množinou.
Pokud navíc D, pro $D = X_{j=1}^{p} D_j$, je kartézským součinem množin D_j, potom D nazýváme kartézskou popisnou množinou.

Příklad 4.2. Podívejme se na tabulku 2, zobrazující příklad symbolické datové tabulky. Řekněme, že nás při popisu dat budou zajímat údaje vztahující se ke Gymnáziu 1. Pak popisnou množinou rozumíme pozorování:

$$Y(w_1) = (\langle 20 \text{ min}, 50 \text{ min} \rangle, \{\text{U } (75\%), \text{ZN } (25\%)\}).$$

Následující definici uvádíme pouze pro úplnost, už bez konkrétního příkladu.

Definice 4.4. Virtuální popis popisného vektoru $\mathbf{d}$, který označíme jako $vir(\mathbf{d})$, je množinou všech jednotlivých popisných vektorů $\mathbf{x}$ splňujících všechna pravidla logické závislosti v na definičním oboru $\mathcal{X}$ [5].

[5] Obvykle zapisujeme $v : \langle \mathbf{x} \in A \rangle \Rightarrow \langle \mathbf{x} \in B \rangle$, více v [3], [4] nebo [11].

Nyní předpokládejme, že to, co nás při analýze dat zajímá, je konkrétní náhodná veličina $Y_j \equiv Z$, jejíž realizace pro u-té pozorování w_u odpovídá intervalu $Z(w_u) = \langle a_u, b_u \rangle$. Dále předpokládejme, že jednotlivé popisné vektory $x \in vir(d_u)$ mají rovnoměrné rozdělení přes interval $Z(w_u)$. Potom pro každé ξ platí:

$$P\{x \leq \xi | x \in vir(d_u)\} = \begin{cases} 0 & \xi < a_u, \\ \frac{\xi - a_u}{b_u - a_u} & a_u \leq \xi < b_u, \\ 1 & \xi \geq b_u. \end{cases} \tag{1}$$

Předpoklad stejné pravděpodobnosti $\frac{1}{m}$ pro všech m objektů nás přivádí k empirické distribuční funkci $F_z(\xi)$, která je sestavena z m rovnoměrných rozdělení $\{Z(w_u), u = 1, \ldots, m\}$. Tímto způsobem z rovnice (1) dostáváme:

$$F_z(\xi) = \frac{1}{m} \sum_{u \in E} P\{x \leq \xi | x \in vir(d_u)\} = \frac{1}{m} \left\{ \sum_{\xi \in Z_u} \left(\frac{\xi - a_u}{b_u - a_u} \right) + |(u|\xi \geq b_u)| \right\}. \tag{2}$$

Po derivaci výrazu (2) dle proměnné ξ získáme empirickou funkci hustoty (viz definice 4.5).

Definice 4.5. Empirická funkce hustoty je pro intervalovou náhodnou veličinu Z definována jako

$$f(\xi) = \frac{1}{m} \sum_{u:\xi \in Z(u)} \left(\frac{1}{b_u - a_u} \right) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi)}{||Z(u)||}, \xi \in \mathbb{R}, \tag{3}$$

kde m označuje počet objektů (intervalů), $I_u(.)$ je indikátorová funkce a $Z(u)$ daný interval. Tedy pokud $\xi \in Z(u)$, tak $I_u(\xi) = 1$, případně pro $\xi \notin Z(u)$ dostáváme $I_u(\xi) = 0$. $||Z(u)||$ odpovídá délce daného intervalu.

Za účelem konstrukce histogramu označme intervalem I takový interval, který pokrývá všechny napozorované hodnoty $Z \in \mathcal{X}$, tedy $I = \langle \min_{u \in E} a_u, \max_{u \in E} b_u \rangle$. Rozdělme nyní interval I na r podintervalů[6]. $I_g = \langle \zeta_{g-1}, \zeta_g \rangle$, $g = 1, \ldots, r-1$

[6] O volbě čísla $r \in \mathbb{N}$ se zmíníme později.

a $I_r = \langle \zeta_{r-1}, \zeta_r \rangle$. Pro další definici mějme na paměti, že podinterval I_g náleží histogramu.

Definice 4.6. Pro intervalovou proměnnou Z definujme napozorovanou četnost podintervalu $I_g = \langle \zeta_{g-1}, \zeta_g \rangle$, $g = 1, \ldots, r$ jako

$$f_g = \sum_{u \in E} \frac{||Z(u) \cap I_g||}{||Z(u)||}, \tag{4}$$

a relativní četnost tvarem

$$p_g = f_g / m. \tag{5}$$

Definice 4.7. Symbolický výběrový průměr je pro intervalovou náhodnou veličinu Z dán vztahem:

$$\bar{Z} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u), \tag{6}$$

a symbolický výběrový rozptyl vztahem:

$$S^2 = \frac{1}{3m} \sum_{u \in E} \left(b_u^2 + b_u a_u + a_u^2 \right) - \frac{1}{4m^2} \left[\sum_{u \in E} (b_u + a_u) \right]^2. \tag{7}$$

Pro odvození uvedených vztahů odkazuji zájemce na [4], s. 79; interpretace autorů je standardní, jak ji známe z popisné statistiky. Nyní již máme zadefinováno vše potřebné pro výpočty k následujícímu příkladu.

Příklad 4.3. (Krevní tlak A)
V tabulce 9 vidíme medicínská data ve formě intervalů. Data se vztahují k hodnotám patnácti pacientů, a to konkrétně k jejich „tepové frekvenci“ (proměnnou dále označíme jako Y_1), „systolickému tlaku“ (dále Y_2) a „diastolickému tlaku“ (dále Y_3). Jednotlivé pacienty rozlišuje sloupec u. Tedy například první pacient

má „tepovou frekvenci“ v rozmezí 44 až 68 úderů za minutu, „systolický tlak“ mezi 90 a 110 mmHg, zatímco „diastolický tlak“ nabývá hodnot z intervalu 50 až 70 mmHg.

u	Y_1 [úderů/min]	Y_2 [mmHg]	Y_3[mmHg]
1	$\langle 44; 68\rangle$	$\langle 90; 110\rangle$	$\langle 50; 70\rangle$
2	$\langle 60; 72\rangle$	$\langle 90; 130\rangle$	$\langle 70; 90\rangle$
3	$\langle 56; 90\rangle$	$\langle 140; 180\rangle$	$\langle 90; 100\rangle$
4	$\langle 70; 112\rangle$	$\langle 110; 142\rangle$	$\langle 80; 108\rangle$
5	$\langle 54; 72\rangle$	$\langle 90; 100\rangle$	$\langle 50; 70\rangle$
6	$\langle 70; 100\rangle$	$\langle 134; 142\rangle$	$\langle 80; 110\rangle$
7	$\langle 72; 100\rangle$	$\langle 130; 160\rangle$	$\langle 76; 90\rangle$
8	$\langle 76; 98\rangle$	$\langle 110; 190\rangle$	$\langle 70; 110\rangle$
9	$\langle 86; 96\rangle$	$\langle 138; 180\rangle$	$\langle 90; 110\rangle$
10	$\langle 86; 100\rangle$	$\langle 110; 150\rangle$	$\langle 78; 100\rangle$
11	$\langle 53; 55\rangle$	$\langle 160; 190\rangle$	$\langle 205; 219\rangle$
12	$\langle 50; 55\rangle$	$\langle 180; 200\rangle$	$\langle 110; 125\rangle$
13	$\langle 73; 81\rangle$	$\langle 125; 138\rangle$	$\langle 78; 99\rangle$
14	$\langle 60; 75\rangle$	$\langle 175; 194\rangle$	$\langle 90; 100\rangle$
15	$\langle 42; 52\rangle$	$\langle 105; 115\rangle$	$\langle 70; 82\rangle$

Tabulka 9: Intervalová data krevního tlaku (viz [4]).

Našim úkolem je vypočítat napozorované a relativní četnosti proměnných Y_1, Y_2 a Y_3, dále symbolický výběrový průměr a symbolický výběrový rozptyl pro každou z proměnných zvlášť.

Nejdříve určíme v rámci všech proměnných minimální a maximální hodnotu, které vymezí pro každou z proměnných interval I. Označme je dále jako I_{Y_1}, I_{Y_2} a I_{Y_3}. Poté z takto vzniklých intervalů utvoříme r podintervalů, přičemž volba r závisí na nás, jedná se pouze o rozdíl v dosažené podrobnosti analýzy zkoumaných dat, tedy pro stejnou proměnnou máme více možností volby r. Nicméně, můžeme podintervaly volit tak, aby hodnota levé meze prvního podintervalu (bráno vzestupně) byla rovna minimu dané proměnné a hodnota pravé meze posledního podintervalu odpovídala maximu[7]. Tím získáme hrubou před-

[7]Volba minimální a maximální hodnoty však nemusí být vždy optimální, neboť se může jednat o odlehlá pozorování.

stavu o tom, jaký úsek reálné osy je potřeba rozdělit. Budeme se řídit obvyklými pravidly pro práci s intervaly (viz [1] nebo [8]) a data rozdělíme do stejně širokých podintervalů s ohledem na to, aby třídění příliš nezjednodušovalo, ale aby nebylo tříd příliš mnoho. Nabízí se nám možnost využití Sturgesova pravidla (viz [1], s. 23), podle kterého pro počet tříd k, v našem případě podintervalů k, platí vztah:

$$k \doteq 1 + 3{,}3 \ \log n \doteq 1 + 1{,}43 \ \ln n,$$

kde n označuje celkový počet pozorování. Výsledná hodnota k přibližně odpovídá hledanému r. V našem příkladu pro $n = 15$ uvažujme $r = 5$, neboť

$$k \doteq 1 + 3{,}3 \log 15 \doteq 1 + 1{,}43 \ln 15 \doteq 4{,}873.$$

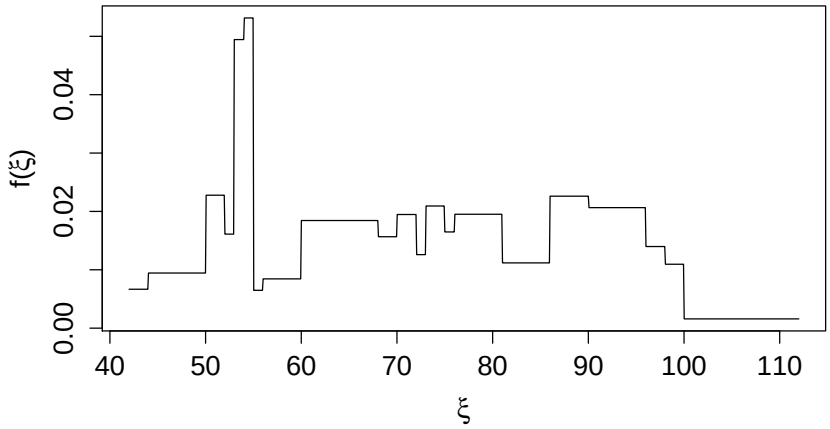
Zabývejme se nejprve proměnnou Y_1. Jak můžeme vidět z příslušného sloupce tabulky, minimální hodnota proměnné Y_1 odpovídá číslu 42, naopak maximální hodnota číslu 112. Interval I má v tomto případě tvar $I_{Y_1} = \langle 42; 112 \rangle$. Řekněme, že první podinterval bude začínat na hodnotě 42, obdobně pro poslední podinterval zvolíme číslo 112. Rozdělením na pět podintervalů získáme intervaly tohoto tvaru: $\langle 42; 56)$, $\langle 56; 70)$, $\langle 70; 84)$, $\langle 84; 98)$, $\langle 98; 112)$. K provedení samotného výpočtu můžeme využít statistického softwaru R.

Datový soubor nejprve upravíme do potřebného tvaru (viz příloha A), dále uložíme s příponou .csv a načteme do softwaru příkazem:

```
data = read.csv2("tlak.csv").
```

Vykreslíme si empirickou funkci hustoty, kdy využijeme vztahu (3). Do softwaru zadáme:

```
fk = function(a,b,x){sapply(x,function(xx)
        {sum(as.numeric(xx>=a & xx<=b)/(b-a))})/length(a)}
k = seq(min(data[,1]),max(data[,2]),length=1000).
plot(k,fk(data[,1],data[,2],k),type="l",
     xlab=expression(xi),ylab="")
mtext(expression(paste("f(",xi,")")), side=2, line=2.5)
```



Obrázek 1: Empirická funkce hustoty proměnné Y_1.

Nyní spočítáme napozorované četnosti $f_1, \ldots, f_5$ dle vztahu (4). Jsou nezbytné pro výpočet relativních četností těch případů, kdy náhodný popisný vektor x leží v podintervalu I_g, tedy výpočtu p_g dle (5). Ručním výpočtem nebo pomocí softwaru dostáváme:

$$f_1 = \frac{||\,\langle 44;\,68\rangle \cap \langle 42;\,56)\,||}{||\,\langle 44;\,68\rangle\,||} + \ldots + \frac{||\,\langle 42;\,52\rangle \cap \langle 42;\,56)\,||}{||\,\langle 42;\,52\rangle\,||} =$$

$$= \frac{56-44}{68-44} + \ldots + \frac{52-42}{52-42} = 0{,}5 + \ldots + 1 = 3{,}611.$$

Z toho $p_1 = \frac{f_1}{m} = \frac{3{,}611}{15} = 0{,}241$. Obdobně počítáme pro napozorované četnosti f_2 až f_5.

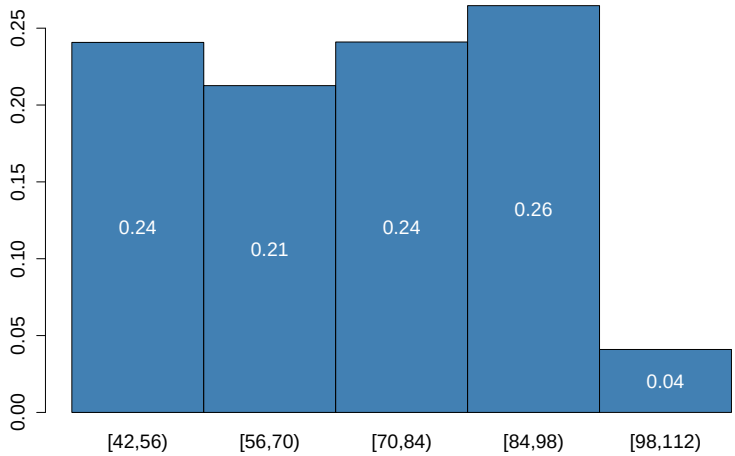
Získané výsledky můžeme shrnout do tabulky 10, respektive vykreslit pomocí softwaru do histogramu. Napozorované četnosti odpovídají sloupci f_g a relativní četnosti sloupci p_g.

Kontrolou nám může býti to, že součet hodnot ve sloupci p_g odpovídá zhruba jedné (s ohledem na mírné zaokrouhlování).

g	f_g	p_g
1	3,611	0,241
2	3,190	0,213
3	3,615	0,241
4	3,970	0,265
5	0,614	0,041

Tabulka 10: Napozorované a relativní četnosti proměnné Y_1.

```
r = nclass.Sturges(1:nrow(data))
histogram = function(a,b){subI=min(a)+(max(b)-min(a))/r*seq(0,r)
    h = apply(sapply(1:nrow(data),function(j)
     {pom = sapply(1:r,function(i){ifelse(b[j]<subI[i+1],b[j],
            subI[i+1])-ifelse(a[j]<subI[i],subI[i],a[j])})
      pom[pom<0] = 0
      pom = pom/(b[j]-a[j])}),1,sum)/length(a)
      b = barplot(h,names=sapply(2:length(subI),function(i)
          {paste0("[",subI[i-1],",",subI[i],")")}),space=0,
          col="steelblue")
      text(b,h/2,round(h,2),col="white")}
histogram(data[,1],data[,2])
```



Obrázek 2: Histogram proměnné Y_1.

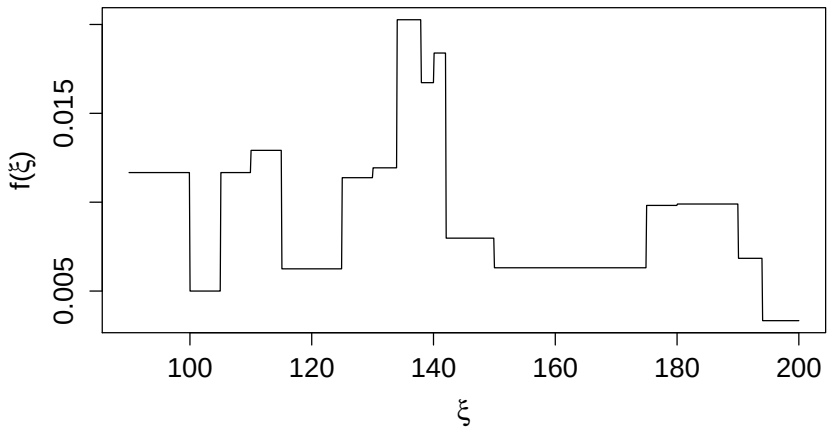
Zbývá nám určit symbolický výběrový průměr a symbolický výběrový rozptyl dle (6) a (7). Ručním výpočtem i pomocí softwaru dostáváme:

$$\bar{Y_1} = \frac{1}{15}\left(\frac{68+44}{2} + \ldots + \frac{52+42}{2}\right) = 72{,}6,$$

$$S_1^2 = \frac{1}{15}\left[\frac{68^2+68\cdot 44+44^2}{3} + \ldots + \frac{52^2+52\cdot 42+42^2}{3}\right] -$$
$$- \frac{1}{15^2}\left[\left(\frac{68+44}{2}\right) + \ldots + \left(\frac{52+42}{2}\right)\right]^2 = 271{,}107.$$

Symbolický výběrový průměr pro proměnnou Y_1 odpovídá hodnotě 72,6 a symbolický výběrový rozptyl hodnotě 271,107.

Pro proměnnou Y_2 obdobně. Začneme s určením minimální a maximální hodnoty, kterých proměnná Y_2 nabyla. Získáme tak interval $I_{Y_2} = \langle 90; 200\rangle$, který rozdělíme na pět podintervalů takto: $\langle 90; 112)$, $\langle 112; 134)$, $\langle 134; 156)$, $\langle 156; 178)$, $\langle 178; 200)$. K vykreslení empirické funkce hustoty proměnné Y_2 opět využijeme vztahu (3) a softwaru R (viz příloha A.1).

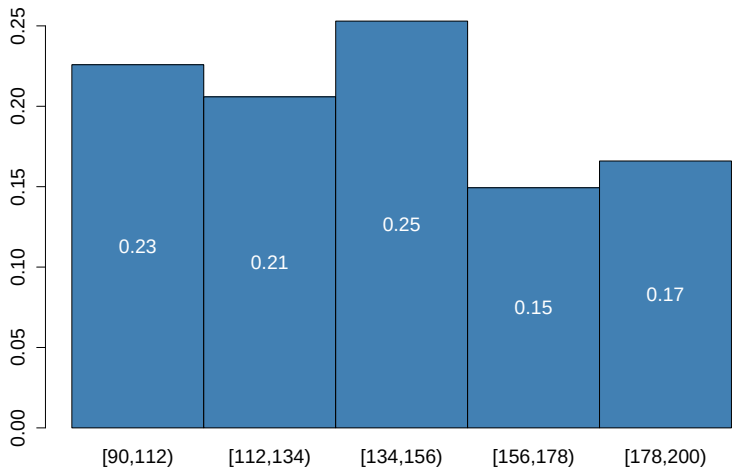


Obrázek 3: Empirická funkce hustoty proměnné Y_2.

V dalším kroku spočítáme dle (4) a (5) napozorované a relativní četnosti, jejichž výsledky přehledně shrneme do tabulky 11. Napozorovaným četnostem opět odpovídá sloupec f_g a relativním četnostem sloupec p_g. Relativní četnosti proměnné Y_2 je možné zobrazit také histogramem, k jehož vykreslení použijeme softwaru R. Podrobný postup výpočtu v softwaru je zaznamenán v příloze A.2.

g	f_g	p_g
1	3,388	0,226
2	3,088	0,206
3	3,795	0,253
4	2,240	0,149
5	2,490	0,166

Tabulka 11: Napozorované a relativní četnosti proměnné Y_2.



Obrázek 4: Histogram proměnné Y_2.

Nyní vypočtěme symbolický výběrový průměr a rozptyl proměnné Y_2, obdobně jako pro proměnnou Y_1. Využijeme vztahů (6) a (7).

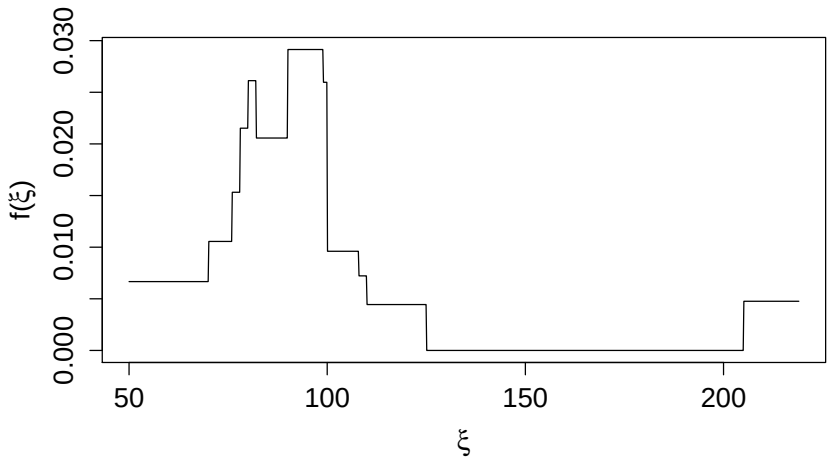
$$\bar{Y_2} = \frac{1}{15}\left(\frac{110+90}{2} + \ldots + \frac{115+105}{2}\right) = 140{,}267,$$

$$S_2^2 = \frac{1}{15}\left[\frac{110^2 + 110 \cdot 90 + 90^2}{3} + \ldots + \frac{115^2 + 115 \cdot 105 + 105^2}{3}\right] -$$
$$- \frac{1}{15^2}\left[\left(\frac{110 + 90}{2}\right) + \ldots + \left(\frac{115 + 105}{2}\right)\right]^2 = 922{,}396.$$

Symbolický výběrový průměr tedy odpovídá hodnotě 140,267 a symbolický výběrový rozptyl hodnotě 922,396.

Ve srovnání s proměnnou Y_1 je symbolický výběrový průměr i symbolický výběrový rozptyl proměnné Y_2 větší, v případě průměru dvojnásobně, v případě rozptylu přibližně trojnásobně. Poznamenejme ovšem, že to samozřejmě odpovídá použitým jednotkám u jednotlivých veličin.

Na závěr této podkapitoly se zaměřme na proměnnou Y_3. Interval I, tvořený minimální a maximální hodnotou nabytou proměnnou Y_3, je tvaru $I_{Y_3} = \langle 50;\ 219 \rangle$, po rozdělení na pět stejně širokých podintervalů můžeme zapsat: $\langle 50;\ 83{,}8)$, $\langle 83{,}8;\ 117{,}6)$, $\langle 117{,}6;\ 151{,}4)$, $\langle 151{,}4;\ 185{,}2)$, $\langle 185{,}2;\ 219)$. Empirickou funkci hustoty proměnné Y_3 získáme s využitím (3) a softwaru R (viz příloha A.1) v této podobě:



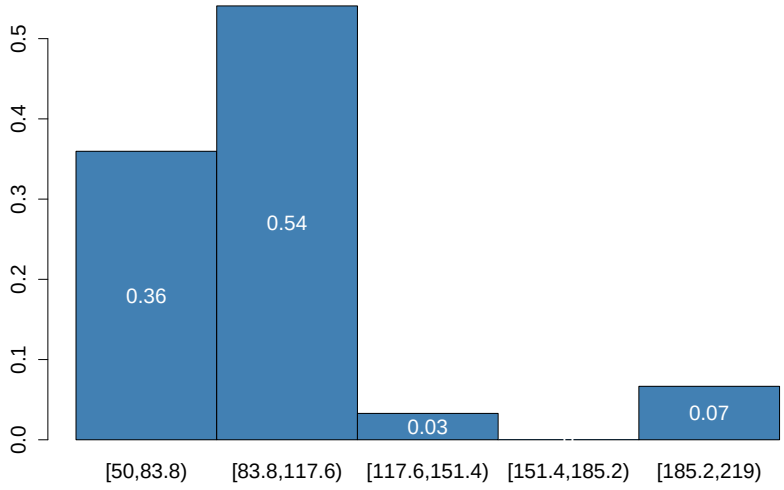
Obrázek 5: Empirická funkce hustoty proměnné Y_3.

Výsledky výpočtu napozorovaných a relativních četností dle (4) a (5) můžeme

vidět v tabulce 12. Sloupec f_g odpovídá opět napozorovaným četnostem a sloupec p_g relativním četnostem. Relativní četnosti proměnné Y_3 znázorňuje histogram, jehož vyobrazení a použité vztahy popisuje, stejně jako v případě proměnné Y_2, příloha A.2.

g	f_g	p_g
1	5,394	0,360
2	8,112	0,541
3	0,493	0,033
4	0	0
5	1	0,067

Tabulka 12: Napozorované a relativní četnosti proměnné Y_3.



Obrázek 6: Histogram proměnné Y_3.

Hodnoty symbolického výběrového průměru a rozptylu proměnné Y_3 získáme obdobně jako pro proměnné Y_1 a Y_2 s využitím (6) a (7).

$$\bar{Y_3} = \frac{1}{15}\left(\frac{70+50}{2} + \ldots + \frac{82+70}{2}\right) = 95{,}667,$$

$$S_3^2 = \frac{1}{15}\left[\frac{70^2 + 70 \cdot 50 + 50^2}{3} + \ldots + \frac{82^2 + 82 \cdot 70 + 70^2}{3}\right] -$$
$$- \frac{1}{15^2}\left[\left(\frac{70+50}{2}\right) + \ldots + \left(\frac{82+70}{2}\right)\right]^2 = 1204{,}133.$$

Symbolický výběrový průměr proměnné Y_3 odpovídá hodnotě 95,667 a symbolický výběrový rozptyl hodnotě 1204,133. Ve srovnání s ostatními proměnnými se jedná o proměnnou s největším symbolickým výběrovým rozptylem.

Dále bychom se mohli věnovat také vztahům mezi jednotlivými proměnnými. Vzhledem k rozsahu práce však nebudeme. Místo toho se zaměříme na jinou možnost výpočtu.

4.3. Parametrická reprezentace intervalových dat

Tato podkapitola přináší odlišný pohled na práci s intervalovými daty (viz [6]), než kterou jsme se doposud zabývali. Pro ilustraci, v čem je tato metoda výpočtu odlišná, využijeme stejného datového souboru jako v příkladu 4.3.

Definice 4.8. Nechť je dána množina n statistických jednotek $S = \{s_1, \ldots, s_n\}$. Intervalová proměnná je pak definována funkčním vztahem:

$$Y : S \to B \text{ je zobrazení takové, že } s_i \to Y(s_i) = \langle a_i, b_i \rangle ,$$

kde B je množina uzavřených intervalů $\langle a_i, b_i \rangle$.

Z definice vidíme, že hodnota intervalové proměnné Y_j je pro každou jednotku $s_i \in S$ popsána mezemi a_{ij} a b_{ij}, kde $i = 1, \ldots, n$ a $j = 1, \ldots, p$. Zamysleme se nyní nad tabulkou 13, která ukazuje univerzální příklad intervalového datového pole.

Jak si můžeme všimnout, datové pole má rozměry $n \times p$ a přestavuje hodnoty n intervalů p proměnných na souboru jednotek S. Pro účely modelování

	Y_1	...	Y_j	...	Y_p
s_1	$\langle a_{11},\, b_{11}\rangle$	...	$\langle a_{1j},\, b_{1j}\rangle$	...	$\langle a_{1p},\, b_{1p}\rangle$
...	...	...	...	...	...
s_i	$\langle a_{i1},\, b_{i1}\rangle$	...	$\langle a_{ij},\, b_{ij}\rangle$	...	$\langle a_{ip},\, b_{ip}\rangle$
...	...	...	...	...	...
s_n	$\langle a_{n1},\, b_{n1}\rangle$	...	$\langle a_{nj},\, b_{nj}\rangle$	...	$\langle a_{np},\, b_{np}\rangle$

Tabulka 13: Intervalové datové pole (viz [6]).

je užitečným popis každého intervalu pomocí středu c_{ij} a rozpětí r_{ij}, respektive zlogaritmovaného rozpětí $\ln r_{ij}$ [8], přičemž

$$c_{ij} = \frac{a_{ij} + b_{ij}}{2}, \tag{8}$$

$$r_{ij} = a_{ij} - b_{ij}. \tag{9}$$

Dle [6] uvažujme, že každý interval $\langle a_{ij},\, b_{ij}\rangle$ je reprezentován svým středem c_{ij} a rozpětím r_{ij}. Můžeme pak také předpokládat, že středy C a logaritmy rozpětí R mají sdružené mnohorozměrné normální rozdělení, tedy pro $R^* = \ln(R)$ můžeme psát $(C,\, R^*) \sim \mathrm{N}_{2p}(\boldsymbol{\mu},\, \boldsymbol{\Sigma})$[9], kde $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_C^T,\, \boldsymbol{\mu}_{R^*}^T\right)^T$ a $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{CC} & \boldsymbol{\Sigma}_{CR^*} \\ \boldsymbol{\Sigma}_{R^*C} & \boldsymbol{\Sigma}_{R^*R^*} \end{pmatrix}$. Přitom $\boldsymbol{\mu}_C$, $\boldsymbol{\mu}_{R^*}$ jsou p-rozměrné sloupcové vektory středních hodnot (ty zde ztotožňujeme s jejich odhady v pojetí matematické statistiky), středů a zlogaritmovaných rozpětí. Prvky matice $\boldsymbol{\Sigma}$, tzn. $\boldsymbol{\Sigma}_{CC}$, $\boldsymbol{\Sigma}_{CR^*}$, $\boldsymbol{\Sigma}_{R^*C}$, $\boldsymbol{\Sigma}_{R^*R^*}$, jsou matice o rozměrech $p \times p$, které obsahují jejich rozptyly a kovariance. Právě kvůli asociaci tohoto přístupu s modelováním pomocí normálního rozdělení hovoříme o parametrické reprezentaci intervalových dat.

Získané teoretické poznatky aplikujme na konkrétním příkladu.

Příklad 4.4. (Krevní tlak B)

Uvažujme dále datový soubor z příkladu 4.3 (viz příloha A). K určení sdružené distribuce středů a zlogaritmovaných rozpětí potřebujeme nejprve znát jejich hod-

[8] Zlogaritmováním předcházíme nesnázím s jeho omezeným definičním oborem (dle [6]).

[9] Speciálně pro $p = 1$ dostáváme dvourozměrné normální rozdělení.

noty, tedy hodnoty středů c_{ij} a zlogaritmovaných rozpětí $\ln r_{ij}$ každého intervalu dané proměnné, kde $i = 1, \ldots, n$ a $j = 1, \ldots, p$. Vzhledem k rozsahu práce uvažujme $p = 1$, budeme tedy proměnné uvažovat jednotlivě, čemuž přizpůsobíme i další značení (zejména C a R^*), tak můžeme z dvojitého indexování přejít k jednoduchým indexům, tedy psát c_i a r_i místo původního c_{ij} a r_{ij}, neboť $j = 1\ \forall i$, $i = 1, \ldots, n$.

Začněme s výpočty vztahujícími se k proměnné Y_1. První interval této proměnné odpovídá intervalu $\langle 44, 68\rangle$. Podle (8) dostáváme:

$$c_1 = \frac{44 + 68}{2} = \frac{112}{2} = 56,$$

Pro výpočet středů $c_2, \ldots, c_{15}$ postupujeme obdobně. Získané hodnoty středů jednotlivých intervalů $\mathbf{c}$ tvoří (dle [6]) náhodný výběr z $\mathrm{N}(\mu_C, \Sigma_{CC})$, přičemž

$$\mathbf{c} = (56;\ 66,\ 73;\ 91;\ 63;\ 85;\ 86;\ 87;\ 91;\ 93;\ 54;\ 52{,}5;\ 77;\ 67{,}5;\ 47)^T.$$

Jelikož se jedná o náhodný výběr z normálního rozdělení, můžeme jeho parametry μ_C a Σ_{CC} odhadnout. Střední hodnotu μ_C odhadneme jako průměr z hodnot náhodného výběru $\mathbf{c}$ a odhadem rozptylu Σ_{CC} je výběrový rozptyl hodnot středů. K výpočtu konkrétních hodnot se vrátíme později.

Dále nás zajímají hodnoty rozpětí daných intervalů, přesněji jejich logaritmy. S využitím (9) dostáváme pro první interval:

$$r_1 = 68 - 44 = 24,$$
$$\ln r_1 = \ln(68 - 44) = \ln 24 = 3{,}178.$$

Postup opakujeme pro výpočet $\ln r_2, \ldots, \ln r_{15}$. Výsledkem je náhodný výběr logaritmů rozpětí $\mathbf{r}^*$ z $\mathrm{N}(\mu_{R^*}, \Sigma_{R^*R^*})$, kde

$$\mathbf{r}^* = (3{,}178;\ 2{,}485;\ 3{,}526;\ 3{,}738;\ 2{,}890;\ 3{,}401;\ 3{,}332;\ 3{,}091;$$
$$2{,}303;\ 2{,}639;\ 0{,}693;\ 1{,}609;\ 2{,}079;\ 2{,}708;\ 2{,}302)^T.$$

Pro určení parametrů normálního rozdělení logaritmů rozpětí R^* využijeme opět odhadů, tedy průměru pro střední hodnotu μ_{R^*} a výběrového rozptylu pro rozptyl $\Sigma_{R^*R^*}$.

K ručním výpočtům konkrétních hodnot můžeme využít běžných vztahů pro výpočet průměru, výběrového rozptylu a kovariance (viz [1] a [8]). Nabízí se však možnost využití softwaru. Postup výpočtu softwarem R pro proměnnou Y_1 vypadá takto:

```
data=read.csv2('tlak.csv',header=TRUE)
p=nrow(data)
C=function(a,b){(a+b)/2}
lnR=function(a,b){log(b-a)}


mu1 = c(C(data[,1],data[,2]),lnR(data[,1],data[,2]))
mu1C = mean(mu1[1:p])
mu1R = mean(mu1[-(1:p)])
```

Nyní již můžeme sestavit dvourozměrný vektor $\boldsymbol{\mu}$. Pro proměnnou Y_1 tak získáváme vektor $\boldsymbol{\mu} = (72{,}6;\ 2{,}665)^T$.

Pro určení dvourozměrného rozdělení středů intervalů a logaritmů rozpětí potřebujeme kromě vektoru středních hodnot $\boldsymbol{\mu}$ znát také prvky varianční matice $\boldsymbol{\Sigma}$. V našem případě se jedná o čtvercovou matici rozměru 2×2, protože její dílčí prvky Σ_{CC}, Σ_{CR^*}, Σ_{R^*C} a $\Sigma_{R^*R^*}$ jsou rozměru 1×1. Diagonálu varianční matice tvoří rozptyly a mimodiagonální prvky jsou kovariance. Hodnota Σ_{CC} odpovídá rozptylu středů jednotlivých intervalů, prvky Σ_{CR^*} a Σ_{R^*C} nesou informaci o kovarianci mezi středy intervalů a odpovídajícími zlogaritmovanými rozpětími, naopak hodnota $\Sigma_{R^*R^*}$ je rozptylem zlogaritmovaných rozpětí.

Pomocí výpočtu v R, konkrétně využitím příkazu:

```
sigma1 = var(cbind(mu1[1:p],mu1[-(1:p)]))
```

dostáváme varianční matici pro proměnnou Y_1 tvaru

$$\boldsymbol{\Sigma} = \begin{pmatrix} 249{,}721 & 6{,}569 \\ 6{,}569 & 0{,}644 \end{pmatrix}.$$

Pro další proměnné Y_2 a Y_3 počítáme obdobně (viz příloha A.3). Můžeme také volit $p = 2$, nebo $p = 3$. Pro $p = 2$ pracujeme s různými dvojicemi proměnných (např. Y_1 a Y_2). Pro situaci $p = 3$ uvažujeme všechny proměnné zároveň a výsledkem je matice $\boldsymbol{\mu}$ o rozměru 2×3 a matice $\boldsymbol{\Sigma}$ o rozměru 6×6, konkrétně

$$\boldsymbol{\mu} = \begin{pmatrix} 72{,}600 & 140{,}267 & 95{,}667 \\ 2{,}665 & 3{,}176 & 2{,}908 \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 249{,}721 & 15{,}989 & -102{,}589 & 6{,}569 & 4{,}386 & 3{,}381 \\ 15{,}989 & 884{,}817 & 666{,}185 & -9{,}113 & 5{,}587 & -4{,}117 \\ -102{,}589 & 666{,}185 & 1249{,}845 & -19{,}984 & 4{,}482 & -2{,}716 \\ 6{,}569 & -9{,}113 & -19{,}984 & 0{,}644 & 0{,}030 & 0{,}089 \\ 4{,}386 & 5{,}587 & 4{,}482 & 0{,}030 & 0{,}426 & 0{,}046 \\ 3{,}381 & -4{,}117 & -2{,}716 & 0{,}089 & 0{,}046 & 0{,}158 \end{pmatrix}.$$

Jak si můžeme všimnout, první sloupec matice $\boldsymbol{\mu}$ zde odpovídá dvourozměrnému vektoru, který jsme získali pro proměnnou Y_1 při $p = 1$. První řádek obsahuje hodnoty totožné s výběrovými průměry, které jsme spočetli v příkladu 4.3. Dále zde však máme druhý řádek matice $\boldsymbol{\mu}$, který oproti příkladu 4.3 rozšiřuje získané poznatky, neboť popisuje i vnitřní strukturu intervalů. Obdobně další sloupce pro proměnné Y_2 a Y_3.

Poznamenejme, že matici $\boldsymbol{\Sigma}$ můžeme pomyslně rozdělit na čtyři bloky, každý o rozměru 3×3. Jak již víme, jednotlivé bloky odpovídají dílčím maticím $\boldsymbol{\Sigma}_{CC}$, $\boldsymbol{\Sigma}_{CR^*}$, $\boldsymbol{\Sigma}_{R^*C}$ a $\boldsymbol{\Sigma}_{R^*R^*}$ (viz výše). Matice $\boldsymbol{\Sigma}_{CC}$ obsahuje informace týkající se pouze středů jednotlivých intervalů, matice $\boldsymbol{\Sigma}_{CR^*}$, $\boldsymbol{\Sigma}_{R^*C}$ vypovídají o středech i zlogaritmovaných rozpětích a matice $\boldsymbol{\Sigma}_{R^*R^*}$ popisuje pouze zlogaritmovaná rozpětí. Diagonální prvky matice $\boldsymbol{\Sigma}$ odpovídají rozptylům a mimodiagonální prvky kovariancím.

5. Řešení reálného datového souboru

Nyní budeme poznatky získané v předchozí kapitole aplikovat na reálný datový soubor, který nám pro účel této práce poskytlo Centrum dopravního výzkumu, v. v. i. Vzhledem k dimenzi datového souboru, který máme k dispozici, je nejprve potřeba data sumarizovat a teprve následně provést jednotlivé výpočty symbolického výběrového průměru a rozptylu. K těmto výpočtům použijeme metody, které jsme si popsali v předchozí kapitole, tedy přístupu využívající distribuční funkci i parametrického přístupu.

Příklad 5.1. (Okresy A)
Mějme k dispozici data týkající se dvaceti dvou okresů. Konkrétně:

ID	název okresu	ID	název okresu
0	Blansko	11	Šumperk
1	Brno – město	12	Kroměříž
2	Brno – venkov	13	Uherské Hradiště
3	Břeclav	14	Vsetín
4	Hodonín	15	Zlín
5	Vyškov	16	Bruntál
6	Znojmo	17	Frýdek – Místek
7	Jeseník	18	Karviná
8	Olomouc	19	Nový Jičín
9	Prostějov	20	Opava
10	Přerov	21	Ostrava – město

Tabulka 14: Okresy.

Studie realizovaná v těchto okresech se zabývala měřením průměrných hodnot nového sněhu v jednotkách [mm/den]. Tuto proměnnou dále označme AVG. Jednotlivá pozorování byla prováděna v letech 2006 až 2010, ve dnech měsíců leden až duben a říjen až prosinec.

Území těchto okresů můžeme rozdělit podle nadmořské výšky. Obecně tak vytvoříme devět klasifikačních tříd, každou o rozsahu 100 m. Není však zaručeno, že každý okres obsahuje všechny třídy nadmořské výšky. Může se stát, že některý

z okresů bude „rovinou“ a jiný např. „horským hřebenem“. Z poskytnutých dat je patrné, že pouze v okresu Uherské Hradiště jsou zastoupeny všechny třídy nadmořské výšky, naopak nadmořská výška Ostravy – města nepřesahuje hodnotu 400 metrů. Skutečnost, že se v daném okresu konkrétní nadmořská výška nevyskytuje, je v datech zaznamenána nesmyslnou zápornou hodnotou. Tato pozorování z původní datové sady vyloučíme.

v	nadmořská výška v [m]
2	0 – 200
3	201 – 300
4	301 – 400
5	401 – 500
6	501 – 600
7	601 – 700
8	701 – 800
9	801 – 900
10	900 a více

Tabulka 15: Klasifikace nadmořské výšky, označení převzato z původních dat.

Pro výpočty je potřeba upravit datovou sadu do potřebného tvaru (ukázka datové sady viz příloha B)[10]. Takto upravený soubor pak můžeme načíst do softwaru R:

```
d = read.csv2("okresy.csv")
attach(d)
```

Dále každé klasifikační třídě přiřadíme odpovídající počet intervalů, podle toho, kolik z celkového počtu okresů patří do dané třídy, viz tabulka 16. Nadmořská výška 0 – 200 m je zastoupena ve čtrnácti okresech z celkového počtu dvaceti dvou, atd.

Jednotlivé intervaly, které daným třídám přiřadíme, budou tvořeny dolním a horním kvartilem hodnot AVG v [mm/den] dosaženými v čase pro každý okres

[10]Jedná se pouze o ukázku datové sady, neboť datový soubor má i po úpravě přes 100 000 řádků (pozorování) a 6 sloupců (proměnných).

nadmořská výška v [m]	četnost intervalů
0 – 200	14
201 – 300	21
301 – 400	22
401 – 500	20
501 – 600	19
601 – 700	18
701 – 800	15
801 – 900	9
900 a více	7

Tabulka 16: Četnost intervalů jednotlivých klasifikačních tříd nadmořské výšky.

zvlášť, neboť uvážením minima a maxima hodnot bychom v tomto případě dostali vždy interval s krajní mezí nula. Jak vidíme v tabulce 16, četnosti žádných dvou klasifikačních tříd nadmořské výšky se sobě nerovnají. Nicméně, pro účely výpočtů symbolického výběrového průměru a rozptylu nám tato skutečnost vadit nebude. Vybereme si dva soubory intervalů a jejich výsledky pak mezi sebou porovnáme. Volme soubory extrémnějších rozdílů, např. soubory odpovídající nadmořské výšce 0 – 200 m (dále proměnná Y_1) a nadmořské výšce 701 – 800 m (dále proměnná Y_2). V prvním případě je potřeba vytvořit čtrnáct intervalů, v druhém případě jich vytvoříme patnáct.

Pro určení mezí intervalů můžeme využít softwaru R. Z datového souboru vybereme ta pozorování, která se vztahují k požadované nadmořské výšce, a přidáme podmínku na označení okresu. Podrobný postup je uveden v příloze B.1.

Čtrnáct výsledných intervalů, které sumarizují původní rozsáhlou datovou sadu, zaznamenáme do tabulky 17. Obdobně pro nadmořskou výšku 701 – 800 m (viz příloha B.1).

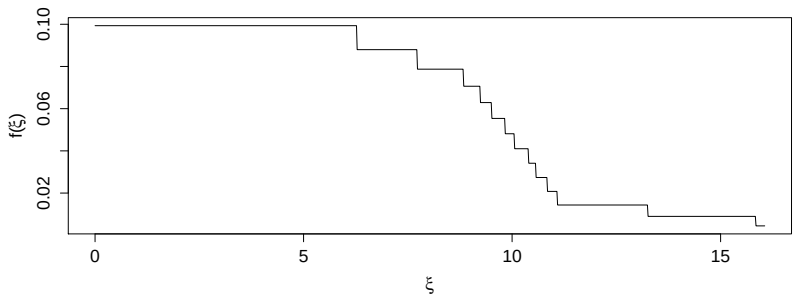
K interpretaci získaných intervalů připomeňme, že hodnoty intervalů v prvním sloupci tabulky 17 odpovídají hodnotám dolního a horního kvartilu proměnné AVG v [mm/den] v průběhu času pro takové okresy, jež obsahují danou nadmořskou výšku, tedy 0 – 200 m. Pro nadmořskou výšku 701 – 800 m obdobně druhý sloupec.

nadmořská výška v [m]	
0 – 200	701 – 800
⟨0; 6,284⟩	⟨4,174; 52,631⟩
⟨0; 9,230⟩	⟨4,274; 55,360⟩
⟨0; 13,249⟩	⟨14,356; 110,345⟩
⟨0; 9,512⟩	⟨14,759; 97,353⟩
⟨0; 7,723⟩	⟨6,536; 60,459⟩
⟨0; 8,828⟩	⟨4,210; 57,220⟩
⟨0; 10,060⟩	⟨12,176; 97,356⟩
⟨0; 10,390⟩	⟨9,423; 89,073⟩
⟨0; 10,564⟩	⟨13,671; 105,861⟩
⟨0; 10,843⟩	⟨24,731; 101,875⟩
⟨0; 9,833⟩	⟨11,570; 92,656⟩
⟨0; 11,085⟩	⟨7,830; 71,455⟩
⟨0; 15,849⟩	⟨33,032; 106,567⟩
⟨0; 16,058⟩	⟨35,804; 119,451⟩
–	⟨6,495; 57,887⟩

Tabulka 17: Modifikace datového souboru Okresy.

Data nyní máme ve formě intervalů. Po úpravě do potřebného formátu (viz přílohy B.2 a B.3) můžeme začít s konkrétními výpočty.

Zabývejme se nyní reprezentací intervalových dat pomocí distribuční funkce. Nejprve určíme interval I_{Y_1}. Bude složen z minimální a maximální hodnoty meze proměnné Y_1. Dostáváme tak interval $I_{Y_1} = \langle 0;\, 16{,}058\rangle$, který dle [1] rozdělíme na pět stejně dlouhých podintervalů: $\langle 0;\, 3{,}212)$, $\langle 3{,}212;\, 6{,}423)$, $\langle 6{,}423;\, 9{,}635)$, $\langle 9{,}635;\, 12{,}846)$, $\langle 12{,}846;\, 16{,}058)$. S využitím vztahu (3) si v softwaru R zadefinujeme empirickou funkci hustoty a následně ji vykreslíme (viz příloha B.4).



Obrázek 7: Empirická funkce hustoty proměnné Y_1.

Dále dle (4) spočítejme napozorované četnosti $f_1, \ldots, f_5$ a z nich podle (5) následně dopočítejme relativní četnosti.

$$f_1 = \frac{|| \langle 0;\, 6{,}284\rangle \cap \langle 0;\, 3{,}212) \,||}{|| \langle 0;\, 6{,}284\rangle \,||} + \ldots + \frac{|| \langle 0;\, 16{,}058\rangle \cap \langle 0;\, 3{,}212) \,||}{|| \langle 0;\, 16{,}058\rangle \,||} =$$

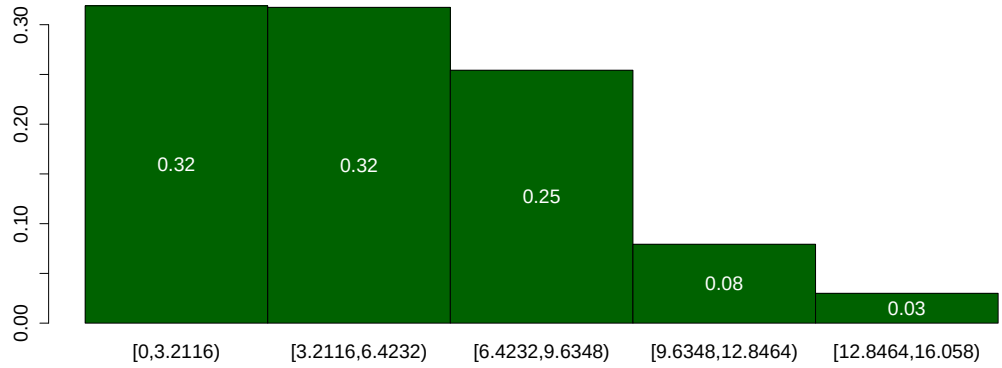
$$= \frac{3{,}212 - 0}{6{,}284 - 0} + \ldots + \frac{3{,}212 - 0}{16{,}058 - 0} = 0{,}511 + \ldots + 0{,}200 = 4{,}48.$$

Z toho $p_1 = \frac{f_1}{m} = \frac{4{,}48}{14} = 0{,}32$. Obdobně pro f_2 až f_5.

Výsledky shrňme pro přehlednost do tabulky 18, respektive pomocí softwaru R vykresleme histogramem (viz příloha B.5).

g	f_g	p_g
1	4,48	0,32
2	4,48	0,32
3	3,5	0,25
4	1,12	0,08
5	0,42	0,03

Tabulka 18: Napozorované a relativní četnosti proměnné Y_1.



Obrázek 8: Histogram pro proměnnou Y_1.

Symbolický výběrový průměr a rozptyl proměnné Y_1 určíme dle (6) a (7) takto:

$$\bar{Y_1} = \frac{1}{14}\left(\frac{6{,}284 + 0}{2} + \ldots + \frac{16{,}058 + 0}{2}\right) = 5{,}340,$$

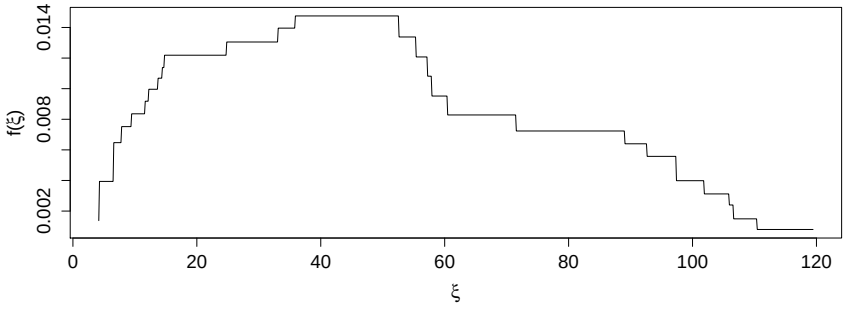
$$S_1^2 = \frac{1}{14}\left[\frac{6{,}284^2 + 6{,}284 \cdot 0 + 0^2}{3} + \ldots + \frac{16{,}058^2 + 16{,}058 \cdot 0 + 0^2}{3}\right] -$$
$$- \frac{1}{14^2}\left[\left(\frac{6{,}284 + 0}{2}\right) + \ldots + \left(\frac{16{,}058 + 0}{2}\right)\right]^2 = 11{,}851.$$

Symbolický výběrový průměr odpovídá hodnotě 5,340 a symbolický výběrový rozptyl hodnotě 11,851.

U proměnné Y_2 postupujeme obdobně. Nezapomeňme však, že počet intervalů proměnné Y_2 se liší od počtu intervalů proměnné Y_1. Nyní pracujeme s patnácti intervaly místo se čtrnácti, neboť právě patnáct okresů z celkového počtu dvaceti dvou obsahuje území náležící do nadmořské výšky 701 – 800 m. Upravený datový soubor, se kterým nyní budeme pracovat je obsažen v příloze B.3.

Nejdříve vytvoříme interval pokrývající hodnoty dolních a horních kvartilů proměnné Y_2, tedy $I_{Y_2} = \langle 4{,}174;\ 119{,}451\rangle$. Podle Sturgesova pravidla rozdělíme interval I_{Y_2} do pěti podintervalů, konkrétně: $\langle 4{,}174;\ 27{,}229)$, $\langle 27{,}229;\ 50{,}7285)$, $\langle 50{,}7285;\ 73{,}340)$, $\langle 73{,}340;\ 96{,}396)$, $\langle 96{,}396;\ 119{,}451)$.

V dalším kroku si za pomocí softwaru R a vztahu (3) zadefinujeme a vykreslíme empirickou funkci hustoty proměnné Y_2 (viz příloha B.4).



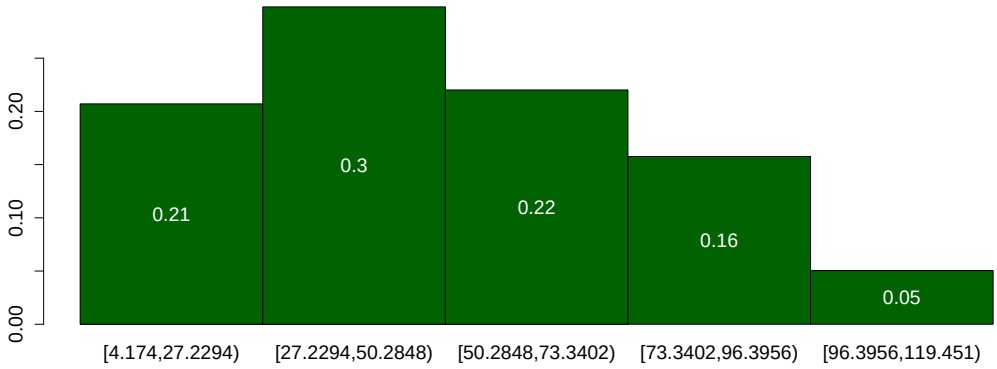
Obrázek 9: Empirická hustota proměnné Y_2.

Výpočet napozorovaných četností $f_1,\ \ldots,\ f_5$ podle (4) a z nich získaných relativních četností $p_1,\ \ldots,\ p_5$ dle (5) provedeme obdobně jako pro proměnnou Y_1.

Výsledky shrneme do tabulky 19. Relativní četnosti proměnné Y_2 zobrazuje také histogram (viz příloha B.5).

g	f_g	p_g
1	3,15	0,21
2	4,5	0,3
3	3,3	0,22
4	2,4	0,16
5	0,75	0,05

Tabulka 19: Napozorované a relativní četnosti proměnné Y_2.



Obrázek 10: Histogram pro proměnnou Y_2.

Nyní zbývá vypočítat symbolický výběrový průměr a rozptyl.

$$\bar{Y_2} = \frac{1}{15}\left(\frac{52{,}631 + 4{,}174}{2} + \ldots + \frac{57{,}887 + 6{,}495}{2}\right) = 49{,}286,$$

$$S_2^2 = \frac{1}{15}\left[\frac{52{,}631^2 + 52{,}631 \cdot 4{,}174 + 4{,}174^2}{3} + \ldots + \frac{57{,}887^2 + 57{,}887 \cdot 6{,}495 + 6{,}495^2}{3}\right] -$$
$$- \frac{1}{15^2}\left[\left(\frac{52{,}631 + 4{,}174}{2}\right) + \ldots + \left(\frac{57{,}887 + 6{,}495}{2}\right)\right]^2 = 685{,}461.$$

Ručním výpočtem i pomocí sofwaru tak získáváme hodnoty 49,286 pro symbolický výběrový průměr a 685,461 pro symbolický výběrový rozptyl.

Výsledné hodnoty příslušící proměnné Y_2 jsou tedy několikanásobně větší než pro proměnnou Y_1.

Příklad 5.2. (Okresy B)
Zabývejme se opět daty z příkladu 5.1 (viz přílohy B.2 a B.3), ale k výpočtům použijme parametrický přístup. Uvažujme nejprve proměnnou vztahující se k nadmořské výšce 0 – 200 m, tzn. Y_1. Protože počet proměnných $p = 1$, dostáváme se k situaci dvourozměrného normálního rozdělení.

Nejdříve spočítáme středy jednotlivých intervalů, přičemž budeme postupovat obdobně jako v předchozí kapitole. První interval proměnné Y_1 odpovídá intervalu $\langle 0;\ 6{,}284\rangle$. Podle (8) dostáváme:

$$c_1 = \frac{0 + 6{,}284}{2} = \frac{6{,}284}{2} = 3{,}142.$$

Pro výpočet středů $c_2,\ \ldots,\ c_{15}$ postupujeme obdobně. Získáme tak náhodný výběr z $\mathrm{N}\,(\mu_C,\ \Sigma_{CC})$, pro

$$\begin{aligned}\mathbf{c} = (&3{,}142;\ 4{,}615;\ 6{,}625;\ 4{,}756;\ 3{,}862;\ 4{,}414;\ 5{,}030;\ 5{,}195;\ 5{,}282;\\ &5{,}422;\ 4{,}917;\ 5{,}543;\ 7{,}925;\ 8{,}029)^T.\end{aligned}$$

Dále počítáme zlogaritmovaná rozpětí, tedy s využitím (9) dostáváme pro první interval této proměnné:

$$\begin{aligned} r_1 &= 6{,}284 - 0 = 6{,}284\\ \ln r_1 &= \ln(6{,}284 - 0) = \ln 6{,}284 = 1{,}838\end{aligned}$$

Postup opakujeme pro výpočet zbývajících logaritmů rozpětí $\ln r_2,\ \ldots,\ \ln r_{15}$. Výsledné hodnoty pak tvoří náhodný výběr z $\mathrm{N}\,(\mu_{R^*},\ \Sigma_{R^*R^*})$ pro

$$\begin{aligned}\mathbf{r}^* = (&1{,}838;\ 2{,}222;\ 2{,}584;\ 2{,}253;\ 2{,}044;\ 2{,}178;\ 2{,}309;\ 2{,}341;\ 2{,}357;\ 2{,}384;\\ &2{,}286;\ 2{,}406;\ 2{,}763;\ 2{,}776)^T.\end{aligned}$$

V dalším kroku potřebujeme určit parametry těchto dvou náhodných výběrů $\boldsymbol{c}$ a $\boldsymbol{r}^*$. Parametry μ_C, resp. μ_{R^*} odhadneme jako průměry hodnot $\boldsymbol{c}$, resp. $\boldsymbol{r}^*$. Odhadem rozptylů C a R^* budou jejich výběrové rozptyly. Za použití softwaru R (viz příloha B.6) zjistíme hodnoty prvků dvourozměrného vektoru $\boldsymbol{\mu}$, tedy

$$\mu_C = 5{,}340, \qquad \mu_{R^*} = 2{,}339.$$

Dále jsme za pomocí softwaru zjistili hodnoty varianční matice $\boldsymbol{\Sigma}$, kterou zapíšeme takto:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1{,}896 & 0{,}34 \\ 0{,}343 & 0{,}064 \end{pmatrix}.$$

Pro zajímavost spočítejme též hodnotu korelačního koeficientu, abychom vyjádřili sílu vztahu mezi středy a logaritmy rozpětí. V případě proměnné Y_1 vychází hodnota korelačního koeficientu rovna 0,987, jedná se tedy o silnou kladnou závislost.

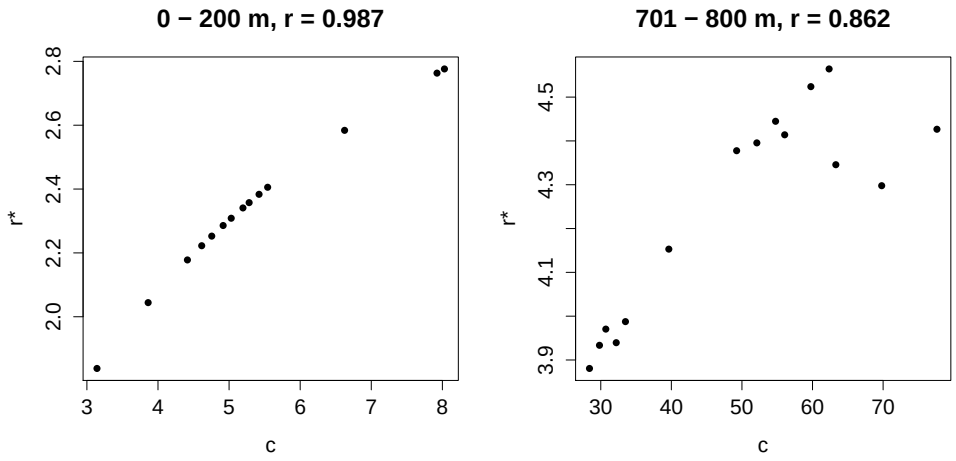
Obdobně pro proměnnou vztahující se k nadmořské výšce 701 – 800 m, tedy Y_2. Postup výpočtu je uveden v příloze B.6. Zde uvedeme pouze výsledky. Nejprve spočítáme hodnoty středů a zlogaritmovaných rozpětí intervalů proměnné Y_2. Ty utvoří náhodné výběry s neznámými parametry, které následně odhadneme pomocí průměru a výběrového rozptylu. Po provedení výpočtů zjistíme, že

$$\mu_C = 49{,}286, \qquad \mu_{R^*} = 4{,}244.$$

Varianční matice $\boldsymbol{\Sigma}$ odpovídá matici tvaru

$$\boldsymbol{\Sigma} = \begin{pmatrix} 255{,}675 & 3{,}310 \\ 3{,}310 & 0{,}058 \end{pmatrix}.$$

Opět spočítáme hodnotu korelační koeficientu. Pro proměnnou Y_2 je tato hodnota rovna 0,862. Opět se jedná o kladnou korelaci, i když slabší než v případě proměnné Y_1, čehož si můžeme všimnout i z obrázku 11.



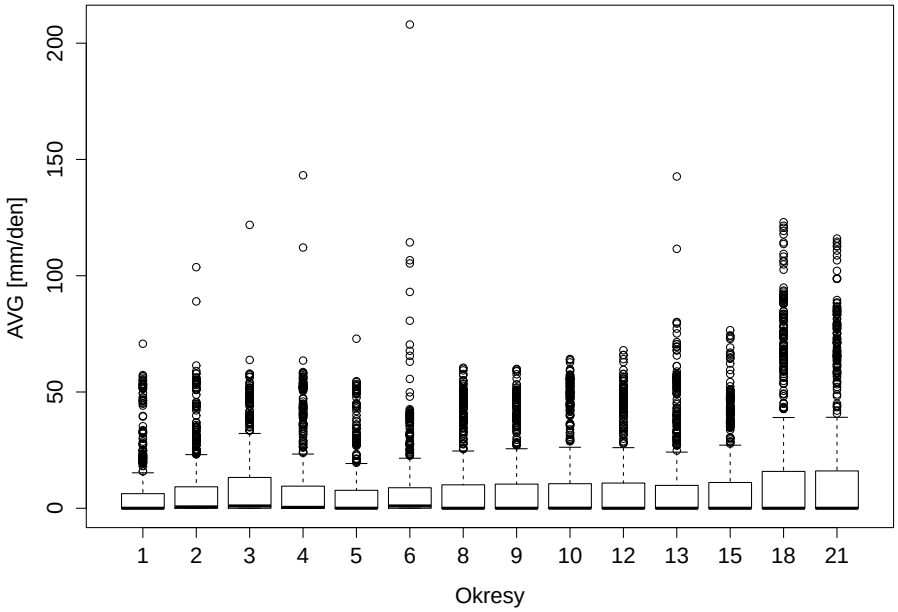
Obrázek 11: Závislost středů a logaritmů rozpětí pro proměnné Y_1 a Y_2.

Z výsledků užitím SDA i parametrického přístupu vidíme, že hodnota μ_C proměnné AVG v [mm/den] pro proměnnou Y_2, která odpovídá hodnotě symbolického výběrového průměru z příkladu 5.1, je větší než hodnota μ_C pro proměnnou Y_1. Této skutečnosti si můžeme všimnout i z následujících boxplotů, pokud v softwaru R zadáme:

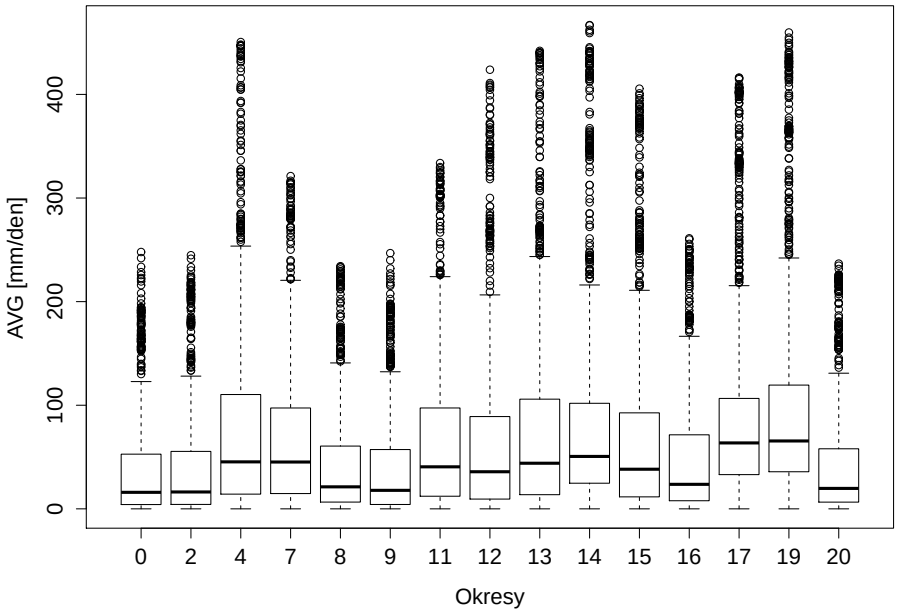
```
d = read.csv2("okresy.csv")
dd2 = d[v==2,]
oo8 = d[v==8,]
boxplot(dd2$AVG~dd2$id,xlab="Okresy",ylab="AVG [mm/den]")
boxplot(oo8$AVG~oo8$id,xlab="Okresy",ylab="AVG [mm/den]")
```

Obrázek 12 zobrazuje situaci pro okresy, které obsahují území o nadmořské výšce 0 – 200 m. Oproti obrázku 13, zobrazujícímu situaci pro okresy obsahující nadmořskou výšku 701 – 800 m, se hodnoty pohybují blíže k nule. Množství nově napadaného sněhu je tedy pro vyšší nadmořskou výšku větší.

Pokud provedeme výpočty vztahující se ke všem nadmořským výškám, zjistíme, že s vyšší nadmořskou výškou roste i AVG, tedy platí přímá úměra.



Obrázek 12: Množství AVG pro jednotlivé okresy proměnné Y_1.



Obrázek 13: Množství AVG pro jednotlivé okresy proměnné Y_2.

Závěr

Tato práce pojednává o jednom z nejnovějších přístupů k analýze dat pomocí metod symbolické analýzy dat (SDA). Nastiňuje základní myšlenky konceptu SDA, která je výhodným nástrojem pro práci s rozsáhlými datovými soubory. Datové soubory jsou upraveny do přípustných (snáze interpretovatelných) rozměrů, s nimiž pak dále můžeme pracovat jednodušeji a pohodlněji. Toto pojetí zavádí mimo jiné též nové typy proměnných, které odpovídají výslednému agregovanému souboru. Z nich můžeme zdůraznit intervalové proměnné, kterými jsem se také ve spojení s SDA zabývala podrobněji. Následuje popsání dvou z možných metod práce s intervalovými daty, což bylo také jedním z cílů práce. Jedná se o reprezentaci intervalových dat pomocí distribuční funkce a parametrickou reprezentaci. Obě metody jsou vysvětleny na ilustrativním příkladu a následně aplikovány na příkladu reálného datového souboru. V závěru práce jsem ukázala, že výpočty základních charakteristik, jako jsou symbolický výběrový průměr či rozptyl, nejsou pro intervalové proměnné nikterak složité. Výpočty nám navíc usnadňuje statistický software R, jehož skripty jsou vloženy buď přímo v textu, nebo v přílohách. Největším přínosem pro mě byla práce s reálným datovým souborem, konkrétně aplikace nového přístupu k analýze dat ve spojení s užitím softwaru. Myslím, že analýza reálných dat může být po „školních“ příkladech často překvapující, především z hlediska jistých komplikací, které v průběhu analýzy dat mohou nastat. Věřím, že i přesto, že se jedná o jeden z nejnovějších přístupů k analýze dat, nalezne SDA brzy svá uplatnění v mnoha oborech.

Literatura

[1] Anděl, J., *Statistické metody*, 4. vydání, Praha: Matfyzpress, 2007. ISBN 80-7378-003-8.

[2] Bílková, D., Budinský, P., Vohánka, V., *Pravděpodobnost a statistika*, Plzeň: Aleš Čeněk, 2009. ISBN 978-80-7380-224-0.

[3] Billard, L., Diday, E., *From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis*, Journal of the American Statistical Association **98**(462), 470 – 487 (2003).

[4] Billard, L., Diday, E., *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Chichester: Wiley, 2006. ISBN 978-0-470-09016-9.

[5] Bock, H.– H., Diday, E., *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Berlin–Heidelberg: Springer–Verlag, 2000. ISBN 3-540-66619-2.

[6] Brito, P., Silva, A. P. D., *Modelling interval data with Normal and Skew-Normal distributions*, Journal of Applied Statistics **39**(1), 3 – 20 (2012).

[7] Diday, E., Noirhomme-Fraiture, M., *Symbolic Data Analysis and the Sodas Sofware*, Chichester: Wiley, 2008. ISBN 978-0-470-01883-5.

[8] Hron, K., Kunderová, P., *Základy počtu pravděpodobnosti a metod matematické statistiky*, Olomouc: Univerzita Palackého v Olomouci, 2013. ISBN 978-80-244-3396-7

[9] Hickey, T., Ju, Q., Van Emden, M. H., *Interval Arithmetic: from Principles to Implementation*, Journal of the ACM **48**(5), 1038 – 1068 (2001).

[10] Marshal, M., N., *Sampling for qualitative research*, Family Practise **13**(6), 522– 525 (1996).

[11] Noirhomme-Fraiture, M., Brito, P., *Far Beyond the Classical Data Models: Symbolic Data Analysis*, Statistical Analysis and Data Mining **4**(2), 157 – 170 (2011).

[12] Warmus, M., *Calculus of approximations*, Bulletin de l'Academie Polonaise des Sciences **4**(5), 253 – 257 (1956).

[13] Interval arithmetic [online], dostupné z: `http://en.wikipedia.org/wiki/Interval_arithmetic`, [citováno 10. 1. 2015].

Přílohy

A. Datový soubor „Krevní tlak“

a_{u1}	b_{u1}	a_{u2}	b_{u2}	a_{u3}	b_{u3}
44	68	90	110	50	70
60	72	90	130	70	90
56	90	140	180	90	100
70	112	110	142	80	108
54	72	90	100	50	70
70	100	134	142	80	110
72	100	130	160	76	90
76	98	110	190	70	110
86	96	138	180	90	110
86	100	110	150	78	100
53	55	160	190	205	219
50	55	180	200	110	125
73	81	125	138	78	99
60	75	175	194	90	100
42	52	105	115	70	82

Tabulka 20: tlak.csv

A.1. Reprezentace intervalových dat pomocí distribuční funkce – empirická funkce hustoty

```
data = read.csv2("tlak.csv")
fk = function(a,b,x){sapply(x,function(xx){sum(as.numeric
     (xx>=a & xx<=b)/(b-a))})/length(a)}

# pro proměnnou Y2
k = seq(min(data[,3]),max(data[,4]),length=1000)
plot(k,fk(data[,3],data[,4],k),type="l",
     xlab=expression(xi),ylab="")
mtext(expression(paste("f(",xi,")")),side=2,line=2.5)
```

```
# pro proměnnou Y3
k = seq(min(data[,5]),max(data[,6]),length=1000)
plot(k,fk(data[,5],data[,6],k),type="l",
     xlab=expression(xi),ylab="")
mtext(expression(paste("f(",xi,")")),side=2,line=2.5)
```

A.2. Reprezentace intervalových dat pomocí distribuční funkce – histogram

```
data = read.csv2("tlak.csv")
r = nclass.Sturges(1:nrow(data))
histogram = function(a,b){subI=min(a)+(max(b)-min(a))
            /r*seq(0,r)
   h = apply(sapply(1:nrow(data),function(j)
    {pom = sapply(1:r,function(i){ifelse(b[j]<subI[i+1],
           b[j],subI[i+1])-ifelse(a[j]<subI[i],subI[i],a[j])})
     pom[pom<0] = 0
     pom = pom/(b[j]-a[j])}),1,sum)/length(a)
     b = barplot(h,names=sapply(2:length(subI),
         function(i){paste0("[",subI[i-1],",",subI[i],")")}),
         space=0,col="steelblue")
     text(b,h/2,round(h,2),col="white")}

# pro proměnnou Y2
histogram(data[,3],data[,4])

# pro proměnnou Y3
histogram(data[,5],data[,6])
```

A.3. Parametrická reprezentace intervalových dat

```
data=read.csv2('tlak.csv',header=TRUE)
p=nrow(data)
C=function(a,b){(a+b)/2}
lnR=function(a,b){log(b-a)}

# pro proměnnou Y2
mu2 = c(C(data[,3],data[,4]),lnR(data[,3],data[,4]))
mu2C = mean(mu2[1:p])
mu2R = mean(mu2[-(1:p)])
sigma2 = var(cbind(mu2[1:p],mu2[-(1:p)]))

# pro proměnnou Y3
mu3 = c(C(data[,5],data[,6]),lnR(data[,5],data[,6]))
mu3C = mean(mu3[1:p])
mu3R = mean(mu3[-(1:p)])
sigma3 = var(cbind(mu3[1:p],mu3[-(1:p)]))

# pro p = 3
mu = cbind(rbind(mu1C,mu1R),rbind(mu2C,mu2R),rbind(mu3C,
     mu3R))
Sigma = var(cbind(mu1[1:p],mu2[1:p],mu3[1:p],mu1[-(1:p)],
        mu2[-(1:p)],mu3[-(1:p)]))
```

B. Datový soubor „Okresy“

i	v	AVG	rok	měsíc	den
0	3	14,446	2006	1	1
0	4	27,203	2006	1	1
0	5	44,644	2006	1	1
0	6	66,881	2006	1	1
0	7	94,354	2006	1	1
0	8	113,505	2006	1	1
1	2	15,929	2006	1	1
1	3	17,475	2006	1	1
1	4	20,063	2006	1	1
1	5	24,356	2006	1	1
2	2	20,726	2006	1	1
2	3	15,186	2006	1	1
2	4	17,857	2006	1	1
2	5	31,328	2006	1	1
2	6	48,855	2006	1	1
2	7	82,755	2006	1	1
2	8	95,419	2006	1	1
3	2	26,172	2006	1	1
3	3	21,333	2006	1	1
3	4	17,918	2006	1	1
3	5	9,914	2006	1	1
3	6	9,065	2006	1	1
4	2	28,926	2006	1	1
4	3	32,849	2006	1	1
4	4	53,715	2006	1	1
4	5	78,527	2006	1	1
4	6	99,874	2006	1	1
4	7	121,975	2006	1	1
4	8	145,623	2006	1	1
5	2	16,343	2006	1	1
5	3	25,585	2006	1	1
5	4	38,71	2006	1	1
5	5	52,954	2006	1	1
5	6	68,27	2006	1	1
5	7	91,688	2006	1	1

Tabulka 21: ukázka okresy.csv

i	v	AVG	rok	měsíc	den
6	2	18,21	2006	1	1
6	3	16,203	2006	1	1
6	4	11,888	2006	1	1
6	5	9,509	2006	1	1
6	6	19,629	2006	1	1
7	3	16,872	2006	1	1
7	4	40,276	2006	1	1
7	5	68,665	2006	1	1
7	6	102,208	2006	1	1
7	7	134,15	2006	1	1
7	8	164,208	2006	1	1
7	9	193,088	2006	1	1
7	10	256,534	2006	1	1
8	2	14,215	2006	1	1
8	3	9,801	2006	1	1
8	4	28,033	2006	1	1
8	5	47,201	2006	1	1
8	6	83,861	2006	1	1
8	7	103,599	2006	1	1
8	8	128,435	2006	1	1
9	2	15,266	2006	1	1
9	3	17,264	2006	1	1
9	4	29,767	2006	1	1
9	5	42,872	2006	1	1
9	6	71,136	2006	1	1
9	7	98,457	2006	1	1
9	8	118,135	2006	1	1
10	2	18,381	2006	1	1
10	3	24,705	2006	1	1
10	4	36,004	2006	1	1
10	5	47,206	2006	1	1
10	6	65,773	2006	1	1
10	7	81,557	2006	1	1
11	3	12,327	2006	1	1
11	4	32,593	2006	1	1
11	5	58,108	2006	1	1
11	6	92,989	2006	1	1

Tabulka 22: ukázka okresy.csv – pokračování

i	v	AVG	rok	měsíc	den
11	7	140,759	2006	1	1
11	8	171,319	2006	1	1
11	9	200,345	2006	1	1
11	10	269,865	2006	1	1
12	2	22,464	2006	1	1
12	3	33,941	2006	1	1
12	4	49	2006	1	1
12	5	74,098	2006	1	1
12	6	103,782	2006	1	1
12	7	133,46	2006	1	1
12	8	152,289	2006	1	1
12	9	168,55	2006	1	1
13	2	21,84	2006	1	1
13	3	39,36	2006	1	1
13	4	58,028	2006	1	1
13	5	78,048	2006	1	1
13	6	103,45	2006	1	1
13	7	126,51	2006	1	1
13	8	147,513	2006	1	1
13	9	163,738	2006	1	1
13	10	186,005	2006	1	1
14	3	41,178	2006	1	1
14	4	62,3	2006	1	1
14	5	102,332	2006	1	1
14	6	148,667	2006	1	1
14	7	214,049	2006	1	1
14	8	284,819	2006	1	1
14	9	339,569	2006	1	1

Tabulka 23: ukázka okresy.csv – pokračování

B.1. Výpočet mezí intervalů pro nadmořské výšky 0 – 200 m a 701 – 800 m

```
d= read.csv2("okresy.csv")
attach(d)
library(reshape2)
q1 = melt(tapply(AVG,list(v,id),function(x){quantile(x,.25)}))
q2 = melt(tapply(AVG,list(v,id),function(x){quantile(x,.75)}))
merge(q1,q2,by.x=1:2,by.y=1:2)
```

B.2. Upravený datový soubor „Okresy“ pro proměnnou Y_1

a_{u1}	b_{u1}
0	6,284
0	9,230
0	13,249
0	9,512
0	7,723
0	8,828
0	10,060
0	10,390
0	10,564
0	10,843
0	9,833
0	11,085
0	15,849
0	16,058

Tabulka 24: okresy1.csv

B.3. Upravený datový soubor „Okresy“ pro proměnnou Y_2

a_{u2}	b_{u2}
4,174	52,631
4,274	55,360
14,356	110,345
14,759	97,353
6,536	60,459
4,21	57,220
12,176	97,356
9,423	89,073
13,671	105,861
24,731	101,875
11,570	92,656
7,830	71,455
33,032	106,567
35,804	119,451
6,495	57,887

Tabulka 25: okresy2.csv

B.4. Reprezentace intervalových dat pomocí distribuční funkce – empirická funkce hustoty

```
data1=read.csv2("okresy1.csv")
data2=read.csv2("okresy2.csv")

fk = function(a,b,x){sapply(x,function(xx){sum(as.numeric
     (xx>=a & xx<=b)/(b-a))})/length(a)}

# pro proměnnou Y1
k = seq(min(data1[,1]),max(data1[,2]),length=1000)
plot(k,fk(data1[,1],data1[,2],k),type="l",
     xlab=expression(xi),ylab="")
mtext(expression(paste("f(",xi,")")), side=2, line=2.5)
```

```
# pro proměnnou Y2
k=seq(min(data2[,1]),max(data2[,2]),length=1000)
plot(k,fk(data2[,1],data2[,2],k),type="l",
     xlab=expression(xi),ylab="")
mtext(expression(paste("f(",xi,")")), side=2, line=2.5)
```

B.5. Reprezentace intervalových dat pomocí distribuční funkce – histogram

```
data1=read.csv2("okresy1.csv")
data2=read.csv2("okresy2.csv")

r = nclass.Sturges(1:nrow(data1))
r = nclass.Sturges(1:nrow(data2))

histogram = function(a,b){
  subI = min(a)+(max(b)-min(a))/r*seq(0,r)
  h = apply(sapply(1:nrow(data1),function(j){
      pom = sapply(1:r,function(i){ifelse(b[j]<subI[i+1],
       b[j],subI[i+1])-ifelse(a[j]<subI[i],subI[i],a[j])})
    pom[pom<0] = 0
    pom = pom/(b[j]-a[j])}),1,sum)/length(a)
      b = barplot(h,names=sapply(2:length(subI),function(i)
     {paste0("[",subI[i-1],",",subI[i],")")}),
      space=0,col="darkgreen")
  text(b,h/2,round(h,2),col="white")}

#pro proměnnou Y1
histogram(data1[,1],data1[,2])
```

```
#pro proměnnou Y2
histogram(data2[,1],data2[,2])
```

B.6. Parametrická reprezentace intervalových dat

```
data1=read.csv2("okresy1.csv")
p1=nrow(data1)
C=function(a,b){(a+b)/2}
lnR=function(a,b){log(b-a)}

# pro nadmorskou vysku 0-200 m
mu1 = c(C(data1[,1],data1[,2]),lnR(data1[,1],data1[,2]))
mu1C = mean(mu1[1:p1])
mu1R = mean(mu1[-(1:p1)])
sigma1 = var(cbind(mu1[1:p1],mu1[-(1:p1)]))
korelace1 = cor(cbind(mu1[1:p1],mu1[-(1:p1)]))

# pro nadmorskou vysku 701-800 m
data2=read.csv2("okresy2.csv")
p=nrow(data2)

mu2 = c(C(data2[,1],data2[,2]),lnR(data2[,1],data2[,2]))
mu2C = mean(mu2[1:p])
mu2R = mean(mu2[-(1:p)])
sigma2 = var(cbind(mu2[1:p],mu2[-(1:p)]))
korelace2 = cor(cbind(mu2[1:p],mu2[-(1:p)]))
```