



Petr Soukup
Ladislav Rabušic
Petr Mareš

Statistická analýza sociálněvědních dat v R

MASARYKOVA
UNIVERZITA

Petr Soukup
Ladislav Rabušic
Petr Mareš

Statistická analýza
sociálněvědních dat v R

MUNI
PRESS

Petr Soukup / Ladislav Rabušic / Petr Mareš

Statistická analýza
sociálněvědních dat v R

Masarykova univerzita

Brno 2023

KATALOGIZACE V KNIZE – NÁRODNÍ KNIHOVNA ČR

Soukup, Petr, 1976-

Statistická analýza sociálněvědních dat v R / Petr Soukup, Ladislav Rabušic, Petr Mareš. --

1. vydání. -- Brno : Masarykova univerzita, 2023. -- 1 online zdroj

Obsahuje bibliografie, bibliografické odkazy a rejstřík

ISBN 978-80-280-0151-3 (online ; pdf)

* 303.7 * 519.23 * 004.438R * 004.9:311 * 30 * (075.8)

– analýza dat

– statistická analýza

– R (programovací jazyk)

– statistický software

– sociální vědy

– učebnice vysokých škol

311 - Statistika [4]

37.016 - Učební osnovy. Vyučovací předměty. Učebnice [22]

Knihu recenzoval

prof. Mgr. Stano Pekár, Ph.D.

© 2023 Petr Soukup, Ladislav Rabušic, Petr Mareš

© 2023 Masarykova univerzita

ISBN 978-80-280-0151-3

ISBN 978-80-280-0150-6 (brožováno)

Obsah

Úvod	11
Kapitola 1	
Než začneme	19
Memento na začátek	19
1.1 Logika kvantitativního výzkumu	26
1.2 Hromadná data	28
1.3 Soubory a způsoby výběru jednotek	30
1.4 Měření	32
1.4.1 Koncepty a jejich operacionalizace – indikátory	33
1.4.2 Proměnná	35
1.4.3 Typy škál – proč jsou důležité	37
1.4.4 Aspekty měření	40
1.5 Hypotézy a modely	42
1.5.1 Od tématu přes problém k výzkumné hypotéze	42
1.5.2 Typy hypotéz	43
1.5.3 Složitější modely	45
1.6 Jak získat data pro analýzu	48
1.6.1 Sekundární analýza dat	49
Literatura	51
Kapitola 2	
Práce s hromadnými daty před analýzou	53
2.1 Prostředí R – instalace a spuštění	53
2.2 R Commander – prostředí pro ovládání R pomocí nabídek	55
2.3 Práce se sociálněvědními daty v R	58
2.3.1 Vytvoření vlastního datového souboru	58
2.3.2 Načtení existujícího datového souboru	64
2.4 Další práce s datovými soubory	77
2.4.1 Slučování souborů (procedura <i>Merge data sets</i>)	80
2.5 Výběr případů z výběrového souboru	85
2.5.1 Výběr případů prostřednictvím pravděpodobnostního (náhodného) výběru ..	85
2.5.2 Výběr případů s ohledem na věcnou otázku	88

Kapitola 3

Základy jednorozměrné analýzy	93
3.1 Rozložení kategorizovaných dat	95
3.1.1 Čištění dat – jak na to	95
3.1.2 Deskripce struktury souboru – explorace pomocí grafů	98
3.2 Popis rozložení proměnných prostřednictvím čísel	105
3.3 Rozložení spojitých proměnných	110
3.3.1 Kontrola nekategorizovaných proměnných	110
3.3.2 Popis rozložení kardinální proměnné	112
3.4 Střední hodnoty a míry variability	112
3.4.1 Nominální proměnné	112
3.4.2 Ordinální proměnné	115
3.4.3 Kardinální proměnné	117
3.5 Výpočty středních hodnot a variability v R	123
3.5.1 Dodatek: Analýza ordinální proměnné s dlouhou stupnicí	127
Literatura	131

Kapitola 4

Normální a standardizované normální rozdělení	133
4.1 Normální rozdělení	133
4.1.1 Jak zjistit, zdali je rozdělení normální?	136
4.1.2 Co dělat, když zjistíme, že rozdělení není normální?	145
4.2 Standardizované (normované) normální rozdělení	147
4.2.1 Standardizovaná náhodná veličina neboli z-skóre	148
4.2.2 K čemu může z-skóre být?	152
4.3 Parametrické a neparametrické testy	153
Příloha kapitoly 4	154

Kapitola 5

Inferenční statistika a testování hypotéz	155
5.1 Populace a výběry	158
5.2 Centrální limitní věta	161
5.3 Inference ze statistiky (výběru) na hodnotu parametru v základním souboru	165
5.3.1 Výběrová chyba	165
5.4 Statistická hypotéza a základy jejího testování	175
5.4.1 Nulová hypotéza	176
5.4.2 Dvoustranné a jednostranné alternativní hypotézy, resp. testy	178
5.4.3 Postup testování	180
5.4.4 Statisticky významné nemusí být věcně významným	183
Literatura	185

Kapitola 6

Úpravy proměnných a příbuzné procedury	187
6.1 Procedura <i>Recode variables</i> (změna kódovacího schématu proměnné)	188
6.1.1 Proměnné s mnoha kategoriemi	190
6.1.2 Změna pořadí kódů	193
6.1.3 Přetočení stupnice (obrácené pořadí kódů)	194
6.2 Vytvoření nové proměnné načítáním hodnot (procedura <i>Count</i>)	195
6.3 Vytvoření nové proměnné početními operacemi (procedura <i>Compute</i>)	196
6.4 Vytvoření nové proměnné prostřednictvím logických podmínek – vytváření typů	199
6.5 Vychýlený výběr a co s ním	202
6.5.1 Vážení souboru podle jedné proměnné	203
6.5.2 Vážení souboru podle více proměnných	206
6.5.3 Typy vah pro data	207
6.5.4 Manipulace s datovým souborem	208
Literatura	210

Kapitola 7

Srovnávání středních hodnot spojitých znaků a testování jejich shody v základním souboru	211
7.1 Porovnání průměrů – procedura <i>Means</i>	212
7.2 T-test neboli Testování hypotézy o shodě dvou populačních průměrů	219
7.2.1 T-test pro jediný výběr – One-Sample T Test	220
7.2.2 T-test pro dva nezávislé výběry – Independent-Samples T Test	222
7.3 Parametrické a neparametrické testy pro střední hodnoty.	227
7.3.1 Jednostranný a dvoustranný test (hypotézy)	229
7.3.2 Obecné pravidlo o nulové hypotéze	230
7.4 Testování shody několika populačních průměrů – analýza rozptylu (ANOVA).	231
7.5 Kruskalův–Wallisův test aneb Neparametrický „bratranec“ jednofaktorové analýzy rozptylu	239
7.6 Exkurz o chybě prvního a druhého druhu (Statistika jako analogie trestního soudnictví).	242
Literatura	244

Kapitola 8

Základy dvourozměrné (bivariační) analýzy kategoriálních proměnných	245
8.1 Test nezávislosti chí-kvadrát (χ^2).	252
8.2 Poměr šancí (<i>odds ratio</i>)	258
8.3 Analýza kontingenčních tabulek bez nutnosti získání originálních dat	261
Literatura	263

Kapitola 9

Měření vztahů mezi dvěma proměnnými (analýza závislostí, korelační analýza)	265
9.1 Asociace a korelace	265

9.2	Míry kontingence pro nominální znaky	267
9.2.1	Míry založené na chí-kvadrátu	267
9.2.2	Další koeficienty pro nominální znaky	269
9.3	Míry souvislosti pro ordinální znaky	270
9.4	Míra souhlasu	275
9.5	Míra souvislosti pro intervalové znaky	277
9.6	Souvislost nominálního znaku s kardinální proměnnou	285
9.7	Shrnutí	285
	Literatura	291

Kapitola 10

Jak odhalit vliv třetí proměnné (elaborace)	293
10.1 Co je elaborace	293
10.2 Podmíněné kontingenční tabulky	295
10.3 Podmíněné korelační koeficienty	303
10.4 Využití dílčích (parciálních) koeficientů	307
10.5 Příklad výpočtu parciální korelace v R	309
Literatura	316

Kapitola 11

Základy lineární regrese	317
11.1 Základní podstata regresní analýzy – regresní přímka a její rovnice	317
11.2 Regresní diagnostika – predikované hodnoty a rezidua	328
11.2.1 Dílčí shrnutí	339
11.3 Dodatek: Analýza po skupinách a použití jiné než lineární funkce	340
Literatura	347

Kapitola 12

Mnohonásobná lineární regrese	349
12.1 Předpoklady regresní analýzy	350
12.1.1 Jak testovat předpoklady	351
12.1.2 Různé formy mnohonásobné regrese	353
12.2 Provedení regrese a její výstupy v R	359
12.2.1 Jak zadat výpočet	361
12.2.2 Regresní koeficienty	363
12.2.3 Hodnocení výstupu regresní analýzy	367
Literatura	371

Kapitola 13

Binární logistická regrese	373
13.1 Proč pro dichotomickou závisle proměnnou nelze využít lineární regresi?	373

13.1.1 Logit, pravděpodobnost a šance	375
13.2 Předpoklady binární logistické regrese	378
13.3 Realizace logistické regrese	379
Literatura	392
Příloha kapitoly 13:	
Základní popisné statistiky nezávisle proměnných a korelace kardinálních a ordinálních proměnných se závisle proměnnou	393
Kapitola 14	
Multinomiální logistická regrese	395
14.1 Předpoklady multinomiální logistické regrese	396
14.2 Realizace multinomiální logistické regrese	396
Literatura	412
Kapitola 15	
Explorační faktorová analýza	413
15.1 Extrakce (nalezení) faktorů pokračování	420
15.2 Pojmenování faktorů	426
15.2.1 Rotace faktorů	429
15.3 Závěrečné poznámky	444
15.3.1 Exkurz: vnitřní konzistence škál – Cronbachovo alfa a faktorová analýza ...	445
Literatura	452
Kapitola 16	
Seskupovací analýza	455
16.1 Hierarchická seskupovací analýza	456
16.1.1 Způsoby měření vzdálenosti v mnohorozměrném prostoru	458
16.1.2 Seskupování případů – jednotlivé techniky	462
16.1.3 Nalezení „ideálního“ počtu seskupení a práce s nimi	468
16.1.4 Poznámky závěrem k hierarchickému seskupování	474
16.2 Relokační seskupování (K-průměry, <i>K-means</i> nebo <i>quick cluster</i>)	475
16.3 Seskupování proměnných jako alternativa k faktorové analýze	478
16.4 Dvoustupňová seskupovací analýza (<i>two step cluster</i>) a další příbuzné postupy	481
16.5 Stručné shrnutí k seskupovacím metodám	482
16.6 Dodatek o tvorbě agregovaných dat	483
Literatura	486
Rejstřík	487
Písmena řecké abecedy	491

Úvod

Žijeme ve světě, který je prochnut daty. Data se stala tak pevnou součástí života jedince i společnosti, že se dnes hovoří o datové revoluci. Technologický rozvoj počítačů, rozvoj jejich hardwaru i softwaru a jejich stále rostoucí schopnost zpracovávat data nejrůznější povahy (data obrazová, jazyková a samozřejmě i numerická) rozšiřuje možnosti vývoje umělé inteligence. Je zřejmé, že schopnost rozumět datům a umět je analyzovat se stává životní nutností, obzvláště u lidí s vysokoškolským diplomem.

Specifickým druhem dat jsou data statistická. Statistická data nás doprovázejí každodenně při čtení novin, poslechu rozhlasu, sledování televizních pořadů. Citování statistických údajů velmi často sloužilo a dodnes v každodenním životě slouží jako důkaz, který má potvrdit správnost argumentace – bohužel v množství nejrůznějších statistických údajů se často nacházejí i takové, které si navzájem protirečí. To může vést k pochybnostem o jejich pravdivosti, a potažmo také k pochybnostem o samotné statistice jako vědě (tj. o statistické vědě), jak to naznačují dehonestující výroky typu „statistikou lze dokázat cokoli“ nebo „nevěřím žádné statistice, kterou jsem sám nezfalšoval“. I proto napsal už v roce 1954 americký žurnalista Darrell Huff populární útlou knížku s názvem *How to Lie with Statistics* (česky vyšla v roce 2013 pod názvem *Jak lhát se statistikou*) s cílem ukázat, jak se nedopouštět různých druhů chyb (a tedy statisticky nelhat) při interpretaci statistických dat.

Alespoň trochu rozumět statistice, a být tak statisticky gramotný je pro každého nesmírně užitečné. Ostatně již v roce 1950 americký statistik Samuel S. Wilks ve své řeči po zvolení předsedou Americké statistické asociace geniálně předpověděl: „Statistické myšlení bude jednou pro skutečné naplňování občanství stejně nezbytné jako schopnost číst a psát“,¹ s čímž my, autoři této učebnice, plně souhlasíme; navíc jsme přesvědčeni, že tato doba již nastala.

Pro studenty sociálních věd je základní znalost statistických operací důležitá obzvláště: nejen proto, že jistě chtějí naplňovat svá práva a povinnosti jako občané této země, ale především proto, že jistě chtějí být i úspěšnými badateli. A jak již v průběhu svého studia zjistili, značná část sociálněvědních závěrů a generalizujících výroků je

¹ Tento výrok je často mylně připisován známému anglickému spisovateli H. G. Wellsovi. Má pocházet z jeho knihy *Mankind in the Making* z roku 1903. Wells se sice v podobném duchu vyjádřil, ale autorem skvělé parafráze jeho myšlenky je skutečně Wilks.

založena právě na statistických analýzách. Studenti proto musejí být připraveni na to, že je nutné se statistiku naučit, neboť statistické operace budou organickou součástí jejich výzkumné práce. Proto musejí vědět, že statistické operace jsou založeny na určitých předpokladech, které, pokud nejsou naplněny, vedou – eufemisticky řečeno – k produkci statistických artefaktů, tj. – řečeno lapidárně – k produkci mylných výsledků.

Ale i ti studenti, kteří se chtějí pohybovat především v prostředí tzv. kvalitativní metodologie, jež je založena na „práci bez čísel“, by měli zvládnutí základních statistických dovedností považovat za užitečné – přinejmenším proto, aby rozuměli tomu, jak statistické údaje vznikají a jaká čertova kopýtka se ve statistických analýzách mohou skrývat.

V této učebnici se budeme zabývat problematikou analýzy statistických dat v prostředí, které je označováno jako R, někdy též jako R projekt.² Považujeme za důležité hned v úvodu zdůraznit, že čtenářům nepředkládáme klasickou učebnici statistiky (proto také popisujeme základní statistické pojmy, aniž bychom věnovali větší pozornost tomu, jak jsou matematicky definovány), ale soubor návodu, jak statisticky analyzovat datové soubory obsahující hromadné kvantitativní údaje. Učebnice je primárně určena pro studenty společenskovědních oborů, kteří chtějí proniknout do světa volně šířitelného a stále rozvíjeného prostředí R. Tento produkt je stále rozšířenější, a když nahlédneme do konferenčních příspěvků či programu prestižní letní školy, stává se dobrým standardem i v oblasti sociálních věd. Je tedy namístě, aby se s ním seznámil i český čtenář. I při poměrně technicistním přístupu, který R nabízí, se budeme snažit využít naše učitelské zkušenosti. Jako dlouholetí učitelé kurzů „analýza dat“ pro studenty sociálních věd máme totiž opakovanou zkušenost, že naučit naše studenty statistické analýze vyžaduje poněkud jiný přístup, než jaký se uplatňuje ve standardní výuce statistiky. Proto je naše učebnice napsána tak, že od čtenáře nevyžaduje více než jen základní znalosti z aritmetiky a elementární algebry. Výklad každé problematiky podáváme podle následujícího vzorce: předestřeme čtenáři analytický problém (například jaká je souvislost mezi mírou religiozity respondentů a jejich postojem k možnosti zavedení eutanazie), poté popíšeme, jakým způsobem lze naložit s výpočty v R (většinou využijeme prostředí, které nám kroky usnadní pomocí nabídek, pro úplnost ale vysvětlíme i příkazy v pozadí), a nakonec ukážeme, jak je možné výsledek, který R vyprodukuje, vyložit a interpretovat. Jelikož jsou všechny analytické úlohy řešeny prostřednictvím výpočtů na počítači, nemusí se nic počítat ručně. Aby ovšem čtenáři pracovali „statisticky poučeně“, výkladům některých principů statistiky se samozřejmě nevyhneme. Snažili jsme se ovšem, aby byl tento výklad maximálně srozumitelný, proto jsme v mnoha případech museli výrazně zjednodušovat (někdy jsme se přitom dostali, jak nás ve svých posudcích upozorňovali recenzenti předchozích textů, až na samou hranici ještě přijatelného zjednodušení). Pevně věříme, že čtenáři též pomohou mnohé obrázky, které v knize i výuce hojně užíváme.

² Česky se někdy hovorově užívá výraz R-ko.

Společenské vědy studují, jak známo, sociální jevy (fenomény), tj. lidské kolektivní jednání, které je výsledkem vztahů a interakcí mezi lidmi a které se odehrává v prostředí lidské kultury a jejích organizací a institucí. Přitom se stejně jako ostatní vědy řídí třemi cíli. Studované jevy se: 1) Nejdříve musejí **popsat**. 2) Poté se musejí prostřednictvím nalezení pravděpodobnostních nebo příčinných (kauzálních) vztahů **vysvětlit**. 3) A nakonec je třeba se pokoušet o **predikci** (předpověď) budoucího způsobu (popřípadě variantních způsobů) jejich chování nebo existence. Jelikož sociální vědy potřebují ke splnění těchto cílů data, používají ve svém kvantitativním paradigmatu statistickou vědu. Ta umí prostřednictvím svých postupů, tj. s použitím postupů **deskriptivní statistiky**, především data **sumarizovat**, tedy **popsat**. Řekneme-li například na základě sociologického výzkumu, který byl proveden na výběrovém souboru 1 812 osob, že v roce 2017 bylo v ČR 90 % respondentů ve svém životě šťastných, zatímco nešťastných bylo pouhých 10 %, ³ pak jsme statisticky shrnuli 1 812 individuálních odpovědí na otázku, zda se respondent(ka) cítí celkově šťastný nebo nešťastný. Podobně sumarizující výpovědi bude, když řekneme, že v pocitu štěstí se muži a ženy nelišili nebo že celkově byly v roce 2017 šťastnější spíše mladší věkové skupiny než skupiny starší, neboť ve věku 18–29 let bylo šťastných 95 % respondentů, zatímco ve věkové skupině 60 let a starších bylo šťastných 85 %. Postupům popisné statistiky jsou věnovány především kapitoly 3, 4 a 7.

Zkoumání vztahů mezi jevy s cílem nalézt jejich pravděpodobnostní nebo kauzální **vysvětlení** jsou věnovány kapitoly 8, 9, 10, 11 a 12, 13 a 14. V kapitolách 11–14, jež podávají výklad postupů regresní analýzy, se čtenář navíc seznámí s metodami, které umožňují **predikovat** budoucí vývoj analyzovaných jevů.

Statistická věda má ovšem pro analýzu sociálněvědních problémů ještě jeden – a to podstatný – přínos. Ukazuje, za jakých okolností je z údajů, které sociální vědy získají z výběrových souborů (a je pro sociální vědy charakteristické, že ve svých zkoumáních pracují ne s celou populací, ale pouze s její menší či větší částí), možné zobecňovat prostřednictvím postupů **statistické inference** na celou populaci (více o tom v kapitole 5).

Jak jsme již uvedli výše, naše učebnice se snaží poskytnout čtenářům pouze základní orientaci v procedurách statistické analýzy. Dobře si uvědomujeme, že z žádného čtenáře statistika neudělá. Věříme ale, že pomůže pochopit, co statistika je, k čemu může sloužit, jak se v ní získávají nejen přesné, ale i spolehlivé a relevantní výsledky, jak těmto výsledkům rozumět a jak je interpretovat – pozor ale, interpretace je již ze značné části za hranicemi znalostí statistiky a musí být vždy doprovázena znalostmi příslušné sociálněvědní disciplíny. Naším hlavním cílem je tedy naučit čtenáře, jak statistiku používat pro odpovědi na otázky, které si sociální vědy obecně (a sociologie specificky) kladou, a jak přitom udělat co nejméně chybných kroků a rozhodnutí nebo falešných závěrů.

³ Viz Rabušic, Chromková Manea (2018, s. 29) nebo též <http://evs.fss.muni.cz/aktualne-k-vyzkumu/vysledky-a-publikace>.

Naše učebnice je výsledkem určité potřeby a z ní plynoucí poptávky, což určuje jak její obsah a výběr jednotlivých témat, tak i rozsah, který jim věnuje. Určuje ale i způsob jejich výkladu. Jsme si přitom vědomi, že se ocitáme v konkurenci se skvělými úvody do statistiky, jako jsou např. *Analýza kategorizovaných dat v sociologii* (Řehák a Řeháková, 1986) nebo *Přehled statistických metod* (Hendl, 2015). I textů o analýze dat v R existuje v České republice několik. Nejpoužívanější je trojice knih od Pekára a Brabce (2009, 2012 a 2019),⁴ které se zaměřují na analýzu biologických dat. Za pozornost stojí i čtvrtý díl učebnice biomedicínské statistiky od Zváry (2013).

Náš výklad je přizpůsoben nejen požadavkům a logice výuky statistiky v bakalářském programu sociologie, ale i logice prostředí R, které je v knize využíváno.⁵ I když zatím v české sociologii (a příbuzných disciplínách) není R příliš užíváno, věříme, že právě díky naší učebnici se tato situace může změnit. Dodejme, že analytické možnosti R jsou v zásadě neomezené (téměř každá statistická novinka je obratem implementována), naše publikace z těchto možností vybírá jen malý díl. Z didaktického hlediska upozorňujeme čtenáře na fakt, že pouhá četba našeho textu sice poskytne základní orientaci ve statistice, ale nenaučí jejímu praktickému použití, které je pro svou složitost vázáno na počítačové zpracování hromadných dat. K získání skutečné kompetence při práci s hromadnými daty je třeba při čtení učebnice zároveň pracovat se softwarem a uváděné příklady si krok za krokem skutečně samostatně procvičovat. Z tohoto důvodu může čtenář nalézt všechny datové soubory, s nimiž se v učebnici operuje, na webové adrese <https://metody.fsv.cuni.cz/>. Na tento web budou postupně přidávány i další materiály a úlohy pro samostatné procvičování.⁶ Ale nejen to, doporučujeme klást si nad příslušnými daty další samostatné výzkumné otázky a odpovídat si na ně na základě vlastních výpočtů. Jsme si vědomi toho, že číst učebnici a mít současně zapnutý počítač, na němž krok za krokem sledujeme učebnicový výklad tématu a provádíme příslušné počítačové operace, není úplně standardní studijní postup, ale v tomto případě to bohužel jinak nejde. Je to podobné, jako byste se chtěli naučit jezdit na snowboardu. O tom, jak se to dělá, si můžete přečíst horu návodů, ale dokud to podle nich – a nejlépe s instruktorem – sami nezkusíte, nenaučíte se to nikdy. Ke schopnosti správným způsobem analyzovat statistická (hromadná) data

⁴ Ukázkou z první knihy může čtenář nalézt pomocí platformy Researchgate: https://www.researchgate.net/publication/269988678_Moderni_analyza_biologickych_dat_1_Zobecnene_linearni_modely_v_prostredi_R/link/56e0176e08aec4b3333cfcff/download.

⁵ Pro přípravu bylo užito R verze 3.6.0. V této oblasti dochází neustále k vývoji, nicméně procedury použité v této knize by tento vývoj neměl nijak ovlivnit (jsou zpracovatelné i ve starších verzích).

⁶ Naším základním datovým souborem, na němž je předvedena většina postupů, je reprezentativní soubor České republiky z mezinárodního výzkumu *European Values Study* z roku 1999 (viz soubor EVS99-cvicny). Jsme si vědomi, že tato data jsou pro mnohé čtenáře této učebnice zastaralá (někteří možná ještě ani nebyli v době jejich sběru na světě), ale to není z hlediska pochopení smyslu statistické analýzy vůbec důležité. V této knize nám jde primárně o popis a vysvětlení statistických postupů, méně již o věcnou sociologickou analýzu hodnotových orientací a preferencí.

prostřednictvím programu R či jiného softwaru vede pouze jediná cesta: domácí praktické studium a pravidelné cvičení s učitelem (instruktorem) v rámci výuky. S programem je třeba živě a pravidelně pracovat. Platí zde ono okřídlené *learning by doing*, tedy učím se tím, že to sám dělám.

Doplňme, že tato učebnice je „variací“ na knihu *Statistická analýza sociálněvědních dat (prostřednictvím SPSS)* od stejného autorského kolektivu, samozřejmě maximálně přizpůsobenou logice prostředí R.⁷ To konkrétně znamená, že některé pasáže jsou podrobnější (tam, kde R nabízí více než SPSS), některé naopak stručnější (například příprava datového souboru, označení proměnných a jejich kategorií) a některé části zcela chybí (např. příloha o Custom Tables, které R nenabízí).

S čím konkrétním se čtenář v jednotlivých kapitolách setká? V **první kapitole** nabídneme čtenářům odkazy na užitečný metodologický kontext zpracování hromadných dat a pochopitelně se také zabýváme otázkami o povaze hromadných dat. Připomínané metodologické poznatky jsou sice triviální, ale bez jejich znalosti nelze statistiku ve společenskovědním výzkumu úspěšně používat. Statistika je totiž dobrým nástrojem jen pro toho, kdo umí nejen adekvátně aplikovat její postupy, ale současně zná i teorii a metodologii svého vědního oboru. Jen se širšími metodologickými a teoretickými oborovými znalostmi dokážeme formulovat relevantní výzkumné otázky, jsme schopni nalézat relevantní způsoby, jak se pokusit na tyto otázky odpovědět, a nakonec dokážeme formulovat relevantní interpretace výsledků našich analýz – relevantní v tom smyslu, že jsou zasazeny do existujícího teoretického kontextu naší disciplíny. Značná část první kapitoly je také věnována vysvětlení důvodů, proč je pro správnou volbu statistických procedur a konkrétních statistik tak důležité rozlišovat úroveň měření a typy stupnic použitých proměnných.

Ve **druhé kapitole** seznámíme čtenáře s prostředím R. Naučíme se, jak nainstalovat základní prostředí, jak zajistit instalaci dodatečných balíčků a základy práce s datovým souborem po spuštění. S ohledem na flexibilitu popíšeme možnosti načítání datových souborů z různých datových formátů. Tato kapitola je jednoznačně nejvíce technicistní, nicméně bez jejího zvládnutí není možné R používat. Mnohé knihy obsahují velice detailní popisy prostředí R, my se omezíme jen na kroky nezbytné pro sociálněvědní analýzu dat. Navíc budeme využívat prostředí s nabídkami, tzv. R Commander, které by mělo být pro sociální vědce velice přístupné.

Třetí kapitola je již věnována prvním krokům statistické analýzy, a to analýze jednorozměrné, která nabízí deskriptci rozdělení hodnot jednotlivých proměnných v souboru. Ukazuje, jak zjistit, jaký je podíl jednotek s určitými vlastnostmi ve výběrovém souboru, popřípadě jak jsou v něm jednotlivé vlastnosti rozděleny, což lze vyjadřovat graficky (např. prostřednictvím histogramů) či numericky prostřednictvím percentilů nebo souhrnných statistik, jako jsou střední hodnoty s jejich mírami variability.

⁷ Viz Rabušic, L., Soukup, P., & Mareš, P. (2019). *Statistická analýza sociálněvědních dat (prostřednictvím SPSS)* (2., přepracované vydání). Brno: Masarykova univerzita.

Zvláštní úlohu plní **kapitola čtvrtá**, v níž se věnujeme jednomu ze základních konceptů statistiky, jímž je normální rozdělení. V jejím závěru se při výkladu standardizovaného normálního rozdělení dotkneme problematiky inferenční statistiky a testování hypotéz. Toto téma detailněji rozvíjí **kapitola pátá**. Ve statistické analýze nám totiž nejde pouze o popis výběrového souboru, s nímž pracujeme, nebo jen o analýzu vztahů mezi proměnnými v něm. Cílem je zobecnění výsledků získaných z výběrového souboru na populaci, z níž byly jeho jednotky vybrány. Nemohli jsme se proto vyhnout stručnému, a proto snad i poněkud povrchnímu vhledu do počtu pravděpodobnosti, neboť právě na něm zobecňování (inference) stojí a s ním také padá. Jak konstatoval nositel Nobelovy ceny Ragnar Frisch v úvodu ke knize Helmuta Swobody *Moderní statistika*: „(...) u hypotézy v technickém a statistickém smyslu se soustředíme na charakteristický způsob rozdělení pravděpodobnosti určitého jevu.“ (Swoboda, 1977, s. 10). Je důležité uvědomit si také souvislost této kapitoly s kapitolou věnovanou metodologickému kontextu. Inferenční statistika má totiž smysl, jen pokud mají hromadná data určitou povahu. Především musejí představovat takový **výběr** z populace (inference nemá smysl u vyčerpávajících šetření zahrnujících všechny jednotky definované populace), při němž všechny jednotky v populaci mají stejnou šanci, že se do výběru dostanou. A aby naše inference byla korektní, je třeba též vědět, jak určit populaci, z níž náš soubor budeme vybírat. V páté kapitole je čtenář také upozorněn na podstatný prvek práce s výběrovými soubory, na skutečnost, že naše výsledky jsou vždy zatíženy tzv. výběrovou chybou a že se pohybují s určitou pravděpodobností v tzv. intervalu spolehlivosti. Tato kapitola se také snaží nabourat mýtus statistické signifikance.

Šestá kapitola je věnována transformacím proměnných, či jinak řečeno, úpravám jejich stupnic. Při sběru dat je mnohdy výhodné použít stupnice, pokud však nejsou upraveny, mohou analýzu komplikovat. Dobře využitelné stupnice naopak umožňují variabilitu úprav podle různě kladených výzkumných otázek. Úpravy stupnic mohou zahrnovat např. vhodné slučování hodnot (kategorií) stupnice, změnu orientace pořadových stupnic, aby z nich bylo možno počítat součtové indexy, nebo různé výpočty s těmito hodnotami. Transformace nám také umožňují úpravy oboru hodnot jednotlivých proměnných, ale i vytváření typů kombinacemi hodnot (vlastností zkoumaných jednotek) dvou či více proměnných. Například z hodnot proměnné pohlaví (muž a žena) a dichotomické proměnné pocit štěstí (pocit štěstí a absence pocitu štěstí) lze vytvořit čtyři typy (muž, respektive žena, cítící se šťastnými a muž či žena cítící se nešťastnými).

Ve druhé polovině se učebnice zaměřuje na základní statistické procedury dvojrozměrné analýzy, která hledá vztahy mezi proměnnými prostřednictvím kontingenčních tabulek a měří jejich sílu pomocí měr asociace a korelace – to je obsahem **kapitol 7, 8 a 9**. A jelikož víme, že v sociální realitě jsou sociální jevy složitě determinovány více skutečnostmi, ukazujeme v **kapitole 10**, jak do analýzy o vztazích mezi dvěma proměnnými přidat působení další, tedy třetí proměnné. Na tuto pasáž pak navazují **kapitoly 11, 12, 13 a 14**, v nichž ukazujeme, jak je možné od dvourozměrné deskripce

přecházet k vícerozměrné analýze a také k predikci. Tím se dostaneme k tzv. multivariačním (vícerozměrným) analytickým postupům. Poslední dvě kapitoly nás seznámí s tzv. exploračními technikami – s faktorovou analýzou (**kapitola 15**) a analýzou skupovací (**kapitola 16**). K šestnácti kapitolám přidáváme několik dodatků, které jsou k dispozici online na webu ke knize.

A nakonec nám dovoluňte několik poděkování. Náš dík patří především několika generacím studentů, na nichž jsme si každoročně postupně ověřovali nové a nové varianty našich textů – děkujeme jim za to, že to s námi vydrželi a že nám dávali důležité podněty o slabých místech v nich. Dále patří velký dík recenzentovi prof. Mgr. Stanislavu Pekárovi, Ph.D., za detailní a cenné připomínky k našemu rukopisu – pokud však čtenáři naleznou v textu nedokonalosti, případně i chyby, není to v žádném případě vina recenzenta, ale pouze a toliko vina naše. A budeme pochopitelně našim čtenářům vděční za jakékoliv podněty pro další zkvalitňování textu.⁸

Petr Soukup, Ladislav Rabušic a Petr Mareš

V Brně a Praze v březnu 2022

⁸ Připomínky prosíme na e-mailovou adresu: soukup@fsv.cuni.cz

