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.

ISBN 978-80-280-0151-3
ISBN 978-80-280-0150-6 (brožováno)

Obsah

Kapitola 3

Kapitola 4

Kapitola 5

Kapitola 6

Kapitola 7

Kapitola 8

Kapitola 9

Úvod

Žijeme ve světě, který je prodchnut 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 protiřečí. 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 cokoliv" nebo „nevěřím žádné statistice, kterou jsem sám nezfalšoval". I proto napsal už v roce 1954 americký žurnalista Darrell Huff populární útlou knížečku 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 návodů, jak statisticky analyzovat datové soubory obsahující hromadné kvantitativní údaje. Učebnice je primárně určena pro studenty společenskovědních oborů, kteří chtějí proniknout do světa volně šiřitelného 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 výpovědí 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 statistky 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] Ukázku 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 deskripci 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 soustřeďujeme 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 seskupovací (**kapitola 16**). K šestnácti kapitolám přidáváme několik dodatků, které jsou k dispozici online na webu ke knize.

A nakonec nám dovolte 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 vděčni 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