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

Kapitola 1

Než začneme

Jasnost je intelektuální hodnota; ne však přesnost a preciznost.
Absolutní preciznost je nedosažitelná; je neúčelné chtít být přesnější,
než to vyžaduje naše problémová situace.

Karl R. Popper

Aforismus o statistice aneb tři druhy lži: lež prostá, lež sprostá, statistika.

Benjamin Disraeli

Memento na začátek

V tomto učebním textu se budeme pohybovat v diskurzu kvantitativního výzkumu, v tzv. kvantitativním paradigmatu. Připomínáme, že sociální vědy jsou vědami multiparadigmatickými, což znamená, že vedle sebe koexistují různé způsoby a pravidla, jak dělat vědu, jak řešit její hlavolamy. Různost těchto vzorců je v podstatě dána tím, jak jednotlivá paradigmata odpovídají na tři základní otázky: ontologickou, epistemologickou a metodologickou. 1) Ontologická otázka se ptá, jaká je povaha reality, kterou zkoumáme. 2) Epistemologická otázka řeší, jaká je podstata poznání a jaký je vztah mezi tím, kdo poznává, a tím, co je poznáváno. 3) Metodologická otázka se pídí po tom, jakým způsobem se produkuje vědění, porozumění a pochopení. Na tomto základě se dnes definují tři základní skupiny paradigmat: pozitivistické, interpretativní a emancipativní (Mertens, 1998).[9]

Kvantitativní paradigma má svůj vzor v přírodních vědách. Vychází z přesvědčení, že realita je vnější a objektivně poznatelná. Klade velký důraz na měření vlastností, tj. na jejich kvantifikaci. Jelikož převážná většina vlastností lidského chování

[9] V literatuře najdeme i další názvy: synonymicky s interpretativním paradigmatem se objevují výrazy etnografické, fenomenologické, hermeneutické nebo naturalistické paradigma. Vedle označení emancipativní paradigma nacházíme také výrazy jako feministické, participativní nebo kriticky teoretické paradigma.

a lidského světa, jimiž se sociální vědy zabývají, představuje složité konstrukty a entity, musíme se ve výzkumu velmi často spokojit s měřením ne daných vlastností (nejsou totiž přímo pozorovatelné), ale s měřením jejich pozorovatelných indikátorů. Nemůžeme například přímo změřit vzdělanost jedinců, ale můžeme na jejich vzdělanost usuzovat z výše dosaženého vzdělání. Vzdělanost je v tomto případě vlastností, úroveň dosaženého vzdělání jejím indikátorem.

Nemožnost přímého měření vlastností sociálního světa je v sociálněvědním výzkumu zdrojem jistých potíží. Mnozí metodologové – a my s nimi – proto zdůrazňují, že jedním z klíčových momentů kvantifikace a měření v sociálních vědách je **operacionalizace**, tedy převod abstraktních konstruktů do měřitelných znaků. S operacionalizací je spojen důležitý prvek, a to otázka **validity** těchto operací, což je posouzení, zda je námi vytvořený měřitelný znak (indikátor) dobrým a skutečným reprezentantem vlastnosti, kterou chceme změřit. Proto se také validita definuje jako schopnost měřit to, co skutečně měřit chceme.

Operacionalizace je náročným tvůrčím procesem, v němž postupujeme od sociálněvědních konceptů a jejich nominálních definic přes odhalování jejich dimenzí a subdimenzí ke konkrétním operacím (operacionálním definicím), které nám říkají, co vlastně máme ve výzkumu zjišťovat a měřit. Názornou ukázkou nechť je schéma operacionalizace, které použil de Vaus (1990) pro koncept (pojem) deprivace – viz obrázek 1.1. Co z něj rozhodně stojí za zapamatování, je to, že při zjišťování „míry deprivace" nevystačíme pouze s jedním indikátorem. De Vaus jich navrhuje zjišťovat devět, a to ještě použití škály introverze/extroverze a škály asertivity zahrnuje zjišťování řady dalších údajů.[10]

Operacionalizaci a měření vnímáme jako ústřední metodologické téma kvantitativního výzkumu. Jak konstruovat dobré, tedy validní a samozřejmě i dostatečně spolehlivé (reliabilní) měřicí nástroje, a jakým způsobem měřit sociální vlastnosti, to jsou kardinální otázky kvantitativního paradigmatu. Podle zastánců paradigmatu kvalitativního jsou ale také zásadními – a z jejich pohledu jen obtížně překonatelnými – překážkami vědecké práce.[11]

Problém měření se navíc umocňuje tím, že velkou část našich kvantitativních dat získáváme na základě standardizovaných výpovědí, tj. na základě standardizovaných rozhovorů tazatele se subjekty výzkumu. Standardizovanými výpověďmi rozumíme (i v dalším textu) výpovědi, které lze vyjádřit čísly (věk, příjem apod.) nebo číslicemi (přiřazenými jednotlivým variantám možných odpovědí), umožňujícími jejich

---

[10] Zájemce o detailnější vysvětlení problematiky konceptualizace a operacionalizace odkazujeme např. na pasáže v učebnici E. Babbieho (Babbie, 2001, s. 119–145).

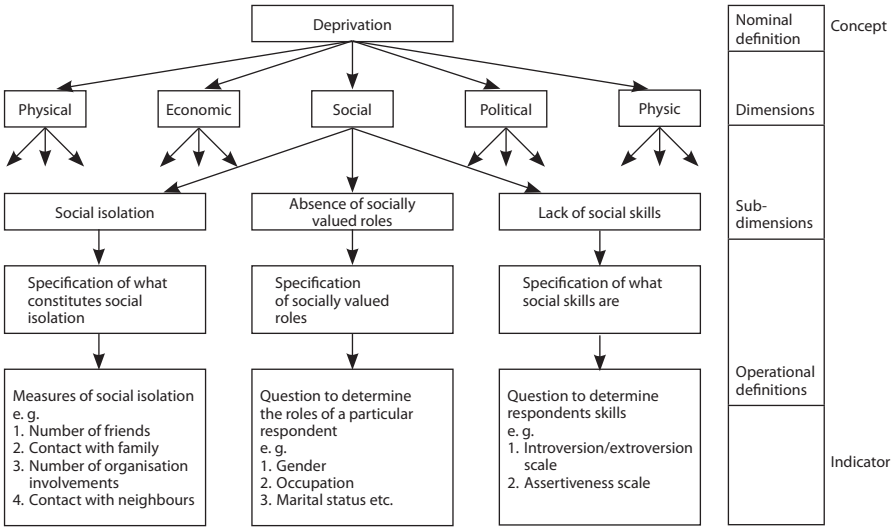
[11] Není cílem tohoto textu tento spor posuzovat, dodejme pouze, že po letech původně nesmiřitelných diskuzí v poslední čtvrtině minulého století došlo nyní mezi oběma tábory ke smíru a ke koexistenci především prostřednictvím tzv. *mixed methods research* neboli metod smíšeného výzkumu, který kombinuje (směšuje) relevantní metody kvantitativního a kvalitativního výzkumu. Pro zájemce o metody smíšeného výzkumu může být dobrým úvodem kniha *Advances in Mixed Methods Research: Theories and Applications*, editovaná Manfredem Maxem Bergmanem (2008).

statistické zpracování. Klasickým příkladem standardizovaného rozhovoru je použití dotazníku, v němž jsou všem zkoumaným subjektům kladeny stejné otázky ve stejném pořadí, přičemž je na ně možno odpovědět (většinou) jen volbou jedné z předložených variant možných odpovědí, označených číslicemi (např. otázka „jaké je vaše nejvyšší dosažené vzdělání" s možností zatrhnout jednu z nabízených variant: 1 = základní, 2 = středoškolské, 3 = vysokoškolské).[12] Zde je dobré si připomenout, že při takovém způsobu měření může docházet k celé řadě poruch, neboť zaznamenáváme pouze tzv. verbální (neboli symbolické) chování, z něhož usuzujeme na chování skutečné. Při práci s takto získanými daty je proto nutné si neustále uvědomovat, že mezi skutečným předmětem výpovědi (tj. jeho vlastnostmi) a tím, co je obsahem výpovědi (co se sděluje), může být obrovský rozdíl. Platí zde proto základní poučka, kterou si dobře pamatujme: Výpovědi o realitě zdaleka nemusejí být realitou samotnou! Mějme tedy skutečnost, že naše závěry mohou být zatíženy nedokonalostmi naší operacionalizace a našeho měření, při formulaci závěrů z kvantitativního výzkumu neustále na paměti a permanentně ji reflektujme!

Permanentní reflexe však není úplně jednoduchá. Ve hře je totiž naše psychika. Po převodu vlastností do znaků (proměnných) a po záznamu jejich hodnot – tedy po vytvoření čísel a jejich nahrání do statistického softwaru počítače – dochází postupně k tomu, že těmto číslům začneme bezmezně věřit. Nastává jev, který označil Petrusek (1993, s. 92) jako tzv. **durifikaci** (ztvrzení) **dat**. To znamená, že postupně začneme s daty pracovat jako s naprosto přesnými čísly, bezmezně jim věříme a výsledky neproblematizujeme. Může to vést dokonce až k jakési hyperpřesnosti: začneme uvádět (nereflektovaně!) výsledky na několik desetinných míst. Např.: „58,36 % respondentů nesouhlasilo s přijímáním imigrantů do zaměstnání"; „průměrná míra pocitu anomie byla 2,856"; „souvislost mezi úrovní dosaženého vzdělání a příjmem respondenta měřená Spearmanovým pořadovým koeficientem je 0,4681". Osvícený výzkumník by samozřejmě hovořil o 58 %, průměr anomie by uvedl na jedno desetinné místo (2,9) a hodnotu koeficientu asociace by zaokrouhlil na 0,47. Nedopustil by se přitom žádné redukce informace, spíše naopak.

Z metodologického hlediska je základem pro kvantitativní výzkum metodologie přírodních věd. Organickou součástí sociálněvědního kvantitativního výzkumu je statistická analýza dat. Jsme přesvědčeni o tom, že právě ve statistické analýze spočívá značná síla tohoto přístupu: umožňuje víceméně exaktně (samozřejmě při vědomí všech možných omezení, která v sobě kvantitativní paradigma skrývá) popisovat zkoumané fenomény, navíc často v jejich vývoji, prostřednictvím analýzy časových řad; skýtá možnosti explorovat fenomény nové a umí ověřovat teorie – to vše za pomoci exaktního matematicko-statistického aparátu. Jelikož často pracuje s výběrovými soubory, dokáže s relativně malými náklady zobecňovat (generalizovat) své závěry na velké populace, a přinášet tudíž zobecnitelné poznatky, jež je možné využívat v praxi. Rozvoj sofistikovaných postupů statistické analýzy, jejich přepis do počítačových programů a obecná dostupnost

[12] Blíže viz oddíl 1.4.2, v němž hovoříme o proměnných.



Podle: de Vaus (1990), s. 52.

Obr. 1.1 Schéma procesu operacionalizace pojmu „deprivace"

osobních počítačů vedou k tomu, že i ty nejmodernější postupy jsou dnes přístupny prakticky každému výzkumníkovi včetně studentů.

Právě zde leží podle našeho názoru obrovská potence, neboť moderní postupy statistické analýzy pomáhají odkrývat vztahy a souvislosti, které bychom jinak v datech nebyli schopni vidět. Jdou do latentních datových struktur, odhalují efekty působení nezávislé proměnné na proměnnou závislou, očištěnou od efektů dalších proměnných, umí prostřednictvím binární logistické regrese nebo loglineární analýzy smysluplně pracovat s kategorizovanými daty a to vše ve velmi krátkém čase.[13] Tyto postupy pak vedou k výsledkům, které dříve nebylo možné nalézt a formulovat.

Většina těchto postupů je součástí běžně dostupných statistických programových balíků (*statistical packages*), takže je zřejmé, že možnosti statistické analýzy jsou dnes obrovské. Otázkou ovšem je, zdali je umíme dobře používat. Právě v tom spočívá, domníváme se, kardinální limit české kvantitativní sociální vědy, sociologii nevyjímaje.

V Čechách, na Moravě a ve Slezsku působí zatím jen nemnoho badatelů, kteří umějí možnosti současné statistické analýzy plně využívat. Jednou z příčin je to, že statistika bývá při studiu sociálních věd často obávaným předmětem, kterým je potřeba „nějak

[13] V dalších pasážích této úvodní kapitoly čtenář možná nalezne pojmy, které mu nebudou zcela jasné. Není to jeho chyba, prosíme o strpení, postupně budou všechny vysvětleny v dalších kapitolách. Věříme, že to není ani chyba didaktická, jak by se v takovém případě u učebního textu mohlo zdát. Považovali jsme prostě za nutné na úvod sdělit naše zásadní stanovisko k problematice kvantitativní analýzy dat, byť s vědomím, že mnohé bude jasnější až po přečtení celého textu. Doporučujeme vrátit se k první kapitole po přečtení celé učebnice.

projít“, bez ambice pochopit její smysl a kouzlo. Studenti se statistiky obávají – čísla jsou pro humanitně orientované osoby často strašákem – a mnozí učitelé statistiky studenty tohoto strachu nejenom nezbavují, ale ještě jej zvyšují. A přitom, jak věříme, je možné naučit statistickou analýzu i studenty, kteří šli studovat sociální vědy právě z toho důvodu, že se báli čísel a – jak oni říkají – „matematiky“. Možná právě způsob, jakým se v Česku učí kvantitativní metody výzkumu a analýzy, vede k tomu, že se dnes mezi studenty sociálních věd stal do značné míry módou výzkum kvalitativní (kvali-výzkum). Studenti se domnívají, že je to výzkum lehčí, neboť je založen na analýze slov namísto analýzy čísel. Hluboce se však mýlí – takový výzkum není lehčí, neboť stejně jako kvanti-výzkum klade obrovské nároky na schopnost sociologické imaginace. Jsme pevně přesvědčeni o tom, že student, který nezvládne metodologii výzkumu kvantitativního a analýzu jeho dat, nebude ani dobrým výzkumníkem kvalitativním.

Analýza studentských prací by nám jistě odhalila i to, že v českých sociálních vědách panují některé obsese, jichž je potřeba se urychleně zbavit. Uvádíme je níže – a jelikož jsme je s úpravami převzali od anglického kolegy, který má podobné pocity a zkušenosti (viz Blaikie 2003, s. 6–7), s jistou škodolibou útěchou konstatujeme, že v tom asi nejsme v Česku tak úplně sami.

Obsese českých sociálních věd:

a) Sociálněvědní výzkum musí vždy začínat hypotézami.
b) Testy statistické významnosti (statistické signifikance) jsou esenciálním rysem analýzy dat.
c) Zjištěná míra asociace nebo korelace mezi dvěma znaky (proměnnými) postačuje k vysvětlení zkoumaného jevu.
   Tyto poněkud radikální teze nyní rozviňme.

Ad a) Musí sociálněvědní výzkum začínat hypotézami?
Jednoduchá odpověď zní: ne vždy. Argumenty jsou následující:

– Každý sociálněvědní výzkum musí mít na svém počátku nějaký problém, který je přeložen do zkoumatelné podoby – má formu výzkumné otázky (nebo několika výzkumných otázek).
– Existuje několik typů výzkumných otázek: otázky „co“, „kolik“, „do jaké míry“ popisují věci, a jsou tedy typické pro deskripce. Otázka „proč“ je otázkou na příčiny a je typická pro explanační (vysvětlující) výzkum. Otázka „jak“ je otázkou na sociální mechanismy a je typická pro akční výzkum, pro intervenci.
– Pouze otázka „proč“ vede badatele k výzkumu, který je založen (deduktivně) na teorii, a proto vyžaduje hypotézy.
– Existují dva druhy hypotéz: **teoretické hypotézy**, které jsou odvozovány z teorie a které nabízejí předběžné vysvětlení otázky typu „proč“. Vedle toho jsou zde

hypotézy statistické, které se používají k **zobecňování výsledků z reprezentativního výběrového souboru na cílovou populaci**,[14] z níž byl výběrový soubor získán.

– Tento rozdíl mezi statistickou a teoretickou (výzkumnou) hypotézou není často akceptován a vede u začínajících výzkumníků ke zmatku.

– Teoretické hypotézy jsou relevantní pouze v případě, kdy hledáme odpovědi na otázku „proč"; statistické hypotézy jsou relevantní, když data pocházejí z pravděpodobnostního (náhodného) či randomizovaného[15] výběru, při kterém mají všechny osoby v dané populaci na začátku výběru stejnou pravděpodobnost, že budou do zkoumaného výběrového souboru vybrány.[16] Některý druh výzkumu může vyžadovat oba druhy hypotéz, některý může vyžadovat pouze jeden z typů; značná část výzkumu ovšem nevyžaduje ani jeden z nich. Co však každý výzkum musí mít, je výzkumná otázka!

– Určitý druh výzkumu hypotézami končí, místo aby jimi začínal (tzv. explorační výzkum).

Ad b) Jsou testy významnosti esenciálním rysem analýzy dat?

Nejdříve opět jednoduchá odpověď: ne, nejsou. Avšak v českých sociálních vědách (a nejen v nich) bohužel velmi často bývají.

Testy signifikance jsou pravděpodobně nejhůře pochopeným aspektem statistické analýzy dat. Jsou součástí tzv. statistické indukce (nebo také inferenční analýzy či statistického usuzování) a používají se tehdy a jen tehdy, když pracujeme s pravděpodobnostním výběrovým souborem. Slouží k tomu, abychom z charakteristik výběru odhadli charakteristiky populace.

Inferenční analýza se využívá pro dva typy úloh: a) k odhadu charakteristik populace z dat výběrového souboru (odhadujeme např. průměrný příjem); b) ke zjištění, zdali vztah (či rozdíl) nalezený ve výběru je možné očekávat také v populaci, z níž byl vybírán (toto je důležité). Např. odhadujeme, zdali i v populaci platí, že existuje vztah mezi vzděláním respondenta a jeho intencí účastnit se kurzů celoživotního vzdělávání.

První typ inferenčních úloh není v sociologii příliš častý; výzkumníci málokdy počítají pravděpodobnou hodnotu populační charakteristiky. Často se používá tzv. bodového odhadu, kdy se prostě předpokládá, že hodnota výběrové charakteristiky bude stejná i v populaci (což ovšem není úplně v pořádku, měly by se počítat intervaly

[14] Je zvykem hovořit často spíše o základním souboru než o cílové populaci, ale spolu s řadou dalších autorů rozlišujeme mezi cílovou populací jakožto souborem jednotek, na které chceme zobecnit své závěry, a základním souborem tuto cílovou populaci zastupujícím (ne všechny jednotky této cílové populace jsou totiž ve všech případech v daném okamžiku výzkumu dostupné).

[15] Randomizace jako náhodné přiřazování osob do zkoumaných skupin (experimentální a kontrolní) se často používá v klinickém výzkumu v medicíně či v psychologii, ale i v ekonomii, méně často v sociologii.

[16] Nikoliv tedy nahodilého!

spolehlivosti). Druhému způsobu inferenční analýzy, testům statistické významnosti (jimiž jsou, jak uvidíme později, např. test chí-kvadrát pro nominální znaky, test významnosti pořadových koeficientů, t-test pro rozdíl dvou průměrů, analýza rozptylu pro rozdíl více průměrů), je naopak věnována velká pozornost a podle našich zkušeností jsou tyto postupy dokonce v české sociální vědě nadužívány, zneužívány a používány špatně (blíže k tomu viz Soukup & Rabušic, 2007). Děje se tak proto, že je špatně pochopen jejich účel a smysl, takže se používají k operacím, pro které nejsou vhodné.

Badatelé a badatelky se zkrátka setrvačně domnívají, že jim testy významnosti řeknou, co je v datech důležitého, a že jim pomohou odhalit těsnost vztahu dvou proměnných. Dále jsou přesvědčeni, že tyto testy musejí být aplikovány na všechny výsledky bez ohledu na to, zdali data pocházejí z vyčerpávajícího zjišťování (z censu), z pravděpodobnostního (náhodného) výběru nebo z výběru nenáhodného (kvótního, záměrného, samovýběru).

Nic z toho ovšem statistická inference neumí. Z toho tedy vyplývá, že testy významnosti:

– Nemohou v žádném případě automaticky sloužit k rozhodnutí, zdali je zjištěný výsledek vědecky nebo prakticky důležitý.
– Nejsou míry asociace.
– Jsou použitelné pouze tehdy, testujeme-li statistické hypotézy. Ty používáme pouze tehdy, když z dat pravděpodobnostního (náhodného) výběrového souboru odhadujeme charakteristiky populace.
– Mohou být aplikovány pouze tehdy, pracujeme-li s výběrovým souborem, který byl vybrán z populace za pomoci postupů pravděpodobnostního (náhodného) výběru a návratnost (např. dotazníků) je relativně vysoká – měla by se, jak nabádá Blaikie (2003, s. 167), pohybovat kolem 85 %.[17]
– Je mylné a chybné používat je v případech, kdy výběr není pravděpodobnostní. A už žádný smysl nemá jejich použití tehdy, když nemáme výběr, ale náš soubor je populací (např. když provedeme výzkum na všech žácích gymnázia ve městě X).
– Nemohou být použity k testování teoretických hypotéz.
– Nemohou sloužit ke generalizujícím výpovědím za populaci, z níž nebyl soubor vybrán. Pokud například provedeme pravděpodobnostní (náhodný) výběr z populace všech studentů gymnázií v Brně (ze seznamu všech brněnských gymnázií bychom si udělali seznam všech tříd a z tohoto seznamu bychom vylosovali takový počet tříd, aby náš výběrový soubor měl dostatečný počet jednotek), naše závěry

[17] V současnosti se bohužel tak vysoké návratnosti dosahuje poměrně obtížně – k tomu viz např. článek J. Krejčího Problém nízké návratnosti výběrových dotazovacích šetření (*Data a výzkum – SDA Info 8*(2), 1–3). nebo obsáhlejší text Krejčí, J. (2007). Non-Response in Probability Sample Surveys in the Czech Republic. *Sociologický časopis / Czech Sociological Review*, *43*(3), 561–587.

nemůžeme zobecňovat na všechny studenty brněnských středních škol (my jsme totiž dělali výzkum pouze na gymnáziích, přičemž víme, že existují ještě další typy středních škol), a už vůbec ne na studenty v jiných městech.

Ad c) Je zjištěná dvourozměrná míra asociace nebo korelace postačující pro vysvětlení?
Ne, není. Důvody jsou následující:

– Vysvětlením říkáme, proč něco existuje. Nalezení těsnosti vztahu mezi dvěma proměnnými je součástí deskriptivního výzkumu, neboť cílem deskripce je určit charakteristiky nějakého sociálního fenoménu, popsat jeho vývoj v čase a typické vzorce vazby na jiné fenomény.
– Nalezení vztahu mezi proměnnými a jeho změření prostřednictvím měr asociace a korelace je formou vyšší míry deskripce.
– Ačkoliv tato deskripce může posloužit k pochopení fenoménu (a někteří i tvrdí, že může sloužit jako základ k predikci), míry asociace nemohou přinést odpověď na otázku „proč“, neboť otázka „proč“ se ptá na příčiny. Asociace vyjadřuje pouze stochastickou (pravděpodobnostní) souvislost, nikoliv však souvislost příčinnou, kauzální.
– Nicméně, chceme-li vysvětlit nějaký jev, musíme nejdříve najít asociaci – neboť tam, kde není žádná stochastická asociace, nemůže být ani kauzální spojení.

Jak je vidět, dobře pochopená a osvojená statistická analýza dat může být klíčem k badatelsky platným výsledkům. Může být mocným nástrojem, ale samozřejmě pouze v rukou, které vědí, jak s ní pracovat. Při provádění kvantitativního výzkumu proto mějme stále na paměti zásady, na nichž musí být každý dobrý kvantitativní výzkum založen (viz Blahuš, 2000):

1) Vědecká průkaznost výsledků výzkumu spočívá v logicky správném a metodologicky čistém designu výzkumu.
2) Design výzkumu nemohou nahradit žádné dodatečné statistické, byť téměř akrobatické cviky s daty.
3) Vědecká průkaznost výsledků výzkumu nespočívá v jejich „statistické významnosti“.

1.1 Logika kvantitativního výzkumu

V kvantitativním výzkumu pracujeme s hromadnými daty. Tato data jsme získali na základě designu výzkumu, který se odvíjí od naší výzkumné otázky. Ta vzniká většinou na základě naší touhy vysvětlit nějaký problém. Výzkumná otázka určuje, co sledovat, jaké vlastnosti měřit. V sociologii tato hromadná standardizovaná data (viz dále) získáváme většinou dotazníkovým šetřením na výběrových souborech (viz dále), kdy jednotkou dotazování je většinou jedinec, nebo ze statistických výkazů.

Dotazujeme-li se jedince, chceme:

– zjistit jeho stav, identifikovat jeho měřitelné vlastnosti (například: Je to muž, nebo žena? Jakou stranu volí?…) a míru těchto vlastností (Jaké má vzdělání? Jakou mírou anomie se vyznačuje? Do jaké míry souhlasí s určitým názorem?…);
– klasifikovat jedince podle těchto zjištění do obecnější kategorie jednotek (Volí levici, nebo pravici? Je liberál, nebo etatista?…);[18]
– usuzovat z výskytu nějaké vlastnosti (intenzity vlastnosti) na jinou vlastnost (respektive její intenzitu);
– sledovat vývoj kvantifikovaných vlastností jedinců v čase.

Vlastnostmi neboli charakteristikami, kterými je jedinec popisován, nejsou jen jeho psychické atributy a stavy (rysy osobnosti, inteligence, neurotické stavy, emoční vyladění, frustrační tolerance), demografické charakteristiky (pohlaví, věk, počet dětí) a jeho zařazení do sociální struktury (sociální třída, profese, velikost místa bydliště) či výrazy tohoto zařazení (sociální status, prestiž), ale i jeho postoje a preference (volební preference, hodnotové orientace, míra xenofobie či rasismu, distance od jiných osob či sociálních skupin, obliba určitých TV programů, religiozita či náboženská konfese, etatismus či liberalismus, způsob trávení dovolené), aktivity či jednání (náplň volného času, účast v sociálních hnutích, návštěvy kulturních akcí, účast na stávce), vědomosti či míra informovanosti a stavy či podmínky, jimž je vystaven (nezaměstnanost, anomie, nemoc, dlouhodobý stres, deprivace) atd.

Rámeček 1.1 Vlastnosti jedince

V sociologii jsou ovšem údaje o jedincích jen přechodnou informací a slouží k tomu, abychom získali informace o sociálních útvarech (skupinách, kategoriích, institucích), jejichž členy/členkami tito jedinci jsou. Nezajímají nás tedy ani tak hodnoty naměřené u jednotlivých subjektů, ale tendence, která se v naměřených hodnotách projevuje u – z jistého hlediska – homogenních skupin či kategorií těchto objektů (u žen, u osob s vysokoškolským vzděláním, u osob s vysokou mírou anomie apod.). Míry závislosti či souvislosti mezi vlastnostmi jsou měřitelné jen v souborech objektů.

[18] Na to se lze v prvním případě zeptat přímo (předložením škály se stupnicí a póly levice/pravice se žádostí, aby se na ni respondenti zařadili podle svých politických názorů), ale zařazení lze také opřít o klasifikaci jednotlivých politických stran jako levicových, pravicových či středových a jedince přiřadit k levici, středu či pravici podle jím volené strany. V druhém výše uvedeném případě se přímo ptát nemůžeme nejen proto, že mnozí z dotazovaných by neměli představu, na co se jich vlastně ptáme. Problémem by bylo i to, že ti znalejší by mohli používat vlastní definice liberalismu a etatismu – a my bychom nevěděli, co si vlastně pod těmito pojmy jednotliví respondenti představovali. Výzkumník musí mít vlastní definici obou postojů a z ní odvozené otázky indikující jejich přítomnost u daného jedince. Na tuto skutečnost ovšem nesmí zapomínat ani při interpretaci svých výsledků. Nejde o liberalismus/etatismus obecně, abstraktně, ale o liberalismus/etatismus jím konkrétně definovaný.

Ilustrace

Měření souvislosti mezi pohlavím a vzděláním není na úrovni jednoho subjektu smysluplné. Smysl dostává až na úrovni „z jistých hledisek homogenních souborů" subjektů, kde můžeme konstatovat například: *Mezi nezaměstnanými* (soubor osob homogenní z hlediska jejich postavení na trhu práce) *je více žen než mužů* (soubory osob homogenní z hlediska jejich pohlaví). Nebo z jiného úhlu: *Mezi ženami je více nezaměstnaných než mezi muži*. Popřípadě: *Příslušníci jednoho pohlaví (ženy/muži) mají vyšší/nižší míru nezaměstnanosti než příslušníci druhého pohlaví.* A tak dále.

V případě souboru, skupiny chceme:

– zjistit stav vlastností jednotek, z nichž je soubor složen, identifikovat jejich vlastnosti (míru těchto vlastností), popsat je pomocí vlastností souboru a zjistit rozložení těchto vlastností v souboru: u kategorizovaných proměnných počty a podíly jednotek s určitou vlastností, intenzitou vlastností v souboru (například: *kolik bylo ve sledovaném souboru katolíků a jaký podíl v něm tvoří?*), u spojitých proměnných stanovit střední hodnoty (například: *jaký je průměrný věk, příjem, … v daném souboru?*);

– hledat příčiny variability neboli hledat vztahy či souvislosti mezi vlastnostmi (proměnnými): pohlaví, vzdělání nebo věk, příjem, respektive vztah mezi pohlavím a vzděláním či věkem a příjmem (*jak se mění výše příjmu v závislosti na vzdělání?*);[19]

– sledovat vývoj kvantifikovaných vlastností v souboru. Z hlediska sociologické metodologie zde lze použít opakované výzkumy, kohortní analýzu či panelová šetření (což jsou formy longitudinálních výzkumů). Ke zpracování dat shromážděných těmito technikami má statistika vlastní nástroje, s nimiž se ovšem vzhledem k danému rozsahu našeho textu seznamovat nebudeme.

1.2 Hromadná data

Kvantitativní výzkum, zaměřující se na to, **jak mnoho** (jaký podíl) něčeho ve společnosti existuje a **jak něco s něčím souvisí**, pracuje s hromadnými daty. V hromadnosti se projevují pravidelnosti, které u jednotlivých případů nemůžeme identifikovat (setkáváme-li se s jednotlivými případy, působí na nás dojmem obrovské proměnlivosti a individuálnosti), zatímco v hromadných datech existují jisté tendence. Například individuálně se můžeme setkat s velkou variabilitou příjmů a můžeme nalézt případy, kdy osoby se středoškolským vzděláním mají vyšší příjmy než některé osoby se vzděláním vysokoškolským,[20] přesto v hromadných datech nepochybně zjistíme již

[19] Problém závislosti statistické a kauzální jsme již otevřeli v úvodu tohoto textu. Blíže bude popsán v kapitole o měření statistické souvislosti. Dopředu pouze upozorňujeme, že mluvíme-li v tomto textu o vztahu, nemáme ve většině případů na mysli vztah kauzální (příčinný).

[20] Nezapomeňme ovšem na možnost intervence dalších faktorů do vztahu dvou vlastností (proměnných), tj. mezi výší vzdělání a výší příjmu. K tomuto problému se ještě dostaneme, zde jen připomeňme, že je třeba vzít v úvahu působení přinejmenším třetího faktoru, jímž je věk jedince. To je příklad uváděný Dismanem (1993, s. 24) jako dvojí příčina nebo také nepravá závislost. Více se tomuto fenoménu věnujeme v 10. kapitole.

zmíněnou tendenci: čím vyšší stupeň vzdělání, tím je mezi osobami, které ho dosáhly, vyšší průměrný příjem.

Ilustrace

Lidé uzavírají sňatky v nejrůznějších životních situacích a s partnery s nejrůznějšími osobními i sociálními charakteristikami – v reálném životě se setkáme se všemi kombinacemi podle výše vzdělání obou partnerů. Jakmile ale přestaneme posuzovat jednotlivé případy a začneme brát v úvahu větší počty, začnou se projevovat určité pravidelnosti, respektive nepravidelnosti, například tendence k homogamii (to je tendence brát si partnera s analogickým sociálním statusem, vzděláním apod.). A zjišťovat tyto tendence je jedním z důležitých úkolů statistické analýzy hromadných dat.

Hromadná data se zpracovávají statistickou procedurou, jejíž obsah tvoří:

- **výběr jednotek pro sledování** (zjišťování jevů či charakteristik). Zde se rozhodujeme, jaký druh výběru použijeme, zdali pravděpodobnostní, kvótní, záměrný apod.;
- **zjišťování údajů** (prostřednictvím pozorování a měření), většinou u velkého počtu jednotek (pozor: sama statistika nemůže nahradit konceptualizaci pojmů a konstrukci nástroje sběru dat). Zde měříme vlastnosti jednotek, které nás na základě naší výzkumné otázky zajímají;
- **kontrola údajů**, jak formální, tj. kontrola návratnosti a úplnosti údajů, tak i logická, kdy kontrolujeme nerozpornost údajů. Zde tzv. „čistíme“ data;
- získání informace o struktuře souboru z hlediska jednotlivých sledovaných vlastností jeho jednotek. **Třídění údajů** ústí do informace o rozložení jevů (třídění prvního stupně), popřípadě o rozložení kombinace jevů či charakteristik objektů (neboli třídění vyšších stupňů). Zde se již dostáváme do oblasti statistické analýzy;
- **výpočet statistik** (středních hodnot charakterizujících rozložení hodnot proměnných, koeficientů asociace charakterizujících sílu, respektive směr vztahů mezi proměnnými) a sledování časové (časové řady) i věcné (testování významností rozdílů a shody) proměnlivosti jevů;
- **vyjádření rozložení jevů** (charakteristik) v tabulkách a grafech, časových řadách;
- **relační neboli vztahová analýza** (hledající souvislosti), která může být jednoduchá a vícefaktorová nebo také kauzální (hledající příčinné závislosti);
- **statistická verifikace hypotéz**. Pozor, jde o verifikaci operacionalizovaných hypotéz, které lze dosáhnout jen s určitou přesností – určitou pravděpodobností chyby. Není verifikací věcnou!
- **inferenční statistika** (statistická indukce) neboli zobecnění výsledků našeho výběrového souboru na cílovou populaci, z níž byl vybrán (jde o pravděpodobnost platnosti našich výsledků i v cílové populaci, samozřejmě za předpokladu, že pracujeme s reprezentativním výběrovým souborem).

Pro sběr hromadných dat je v sociologii typický zejména *survey* – nejčastěji používané dotazníkové šetření (vedle například rutinně shromažďovaných statistik nebo panelových šetření) na výběrovém souboru. Blíže k logice *survey* a postupu při něm viz například de Vaus (1990).

1.3 Soubory a způsoby výběru jednotek

Již bylo řečeno, že data pro statistickou analýzu většinou shromažďujeme standardizovaným způsobem. Máme-li je statisticky zpracovat, musejí mít standardní podobu a musejí být zaznamenána jako čísla (neboli kardinální proměnné) nebo číslice (proměnné ordinální a nominální). Navíc bychom si měli uvědomit, že jen málokdy pracujeme s vyčerpávajícími šetřeními, zahrnujícími všechny dostupné členy dané cílové populace, neboli se základními soubory. Každý výzkum vyžaduje určení okruhu zkoumaných jednotek (osob), tedy určení „zkoumané populace". Jestliže říkáme, že se naše „zkoumání zaměří" na určitý soubor jednotek, máme většinou na mysli širší množinu jednotek, než postihne naše výzkumná činnost. Je to tzv. cílová populace (*general universum/population*). **Cílová populace** představuje okruh osob (soubor jednotek) vymezených nějakými sociálními či demografickými vlastnostmi, pro který chceme vyslovit závěry. Toto vymezení cílové populace může být různé a určuje i její velikost (zahrnuje všechny jednotky s vymezenými charakteristikami).

Ilustrace

Cílovou populaci může tvořit 4 100 studentů, kteří navštěvují Fakultu sociálních studií v Brně; nebo to mohou být všechny ženy, které porodily v roce 2008 v porodnici města X své dítě; příjemci sociálních dávek s trvalým bydlištěm v Ostravě; obyvatelé celé ČR ve věku 15–65 let, ekonomicky aktivní obyvatelé ČR, obyvatelé ČR s volebním právem apod.[21]

Určení zkoumané populace je dáno především naším výzkumným tématem. Většinou přímo nezkoumáme, nemusíme zkoumat a mnohdy ani fyzicky zkoumat nemůžeme celou cílovou populaci – kdybychom tak učinili, provedli bychom tzv. **vyčerpávající zjišťování** neboli **census**. Proto převážně pracujeme s **výběrovými soubory** (*samples*), které jsou tvořené jednotkami vybranými podle určitých kritérií z cílového souboru, provádíme tedy **výběrová šetření.**[22] Jednotky pro náš výzkum vybíráme obvykle tak, aby byl výběrový soubor pro základní soubor **reprezentativní.**

Reprezentativnost znamená dvojí věc:

– Výběrový soubor má strukturu analogickou struktuře cílové populace – z hlediska známých i neznámých charakteristik jeho prvků, což ovšem zajistí jen jejich **pravděpodobnostní (náhodný) výběr** (viz například Babbie, 2001 nebo Disman, 1993).
– Zjištěné výsledky proto mohou být zobecněny na cílovou populaci. Co platí ve výběrovém souboru, platí i v cílové populaci. Takové zobecnění je ovšem možné jen za dodržení určitých podmínek při výběru jednotek – při dodržení pravidel pravděpodobnostního (náhodného) výběru.

[21] Všimněme si, že vymezení cílové populace vyžaduje často definici (co rozumíme ekonomickou aktivitou, jak je vymezeno volební právo, jaké sociální dávky máme na mysli apod.).

[22] Způsoby, jak tyto jednotky vybíráme, aby byl náš soubor reprezentativní, zde nepopisujeme, čtenář se s nimi může seznámit v publikacích věnovaných sociologické metodologii, jako jsou například Babbie (2001), de Vaus (1990) či Disman (1993).

Znovu připomínáme, že zobecnění je možné jedině pro cílovou populaci, z níž byl výběrový soubor vybrán, a pro žádnou jinou. Na to často autoři zapomínají, a proto se lze setkat i se statěmi, které mají například tendenci vypovídat o povaze sexuálního života české populace na základě dat získaných mezi klienty sexuologických ordinací a poraden. Potíž u takovýchto závěrů spočívá jednoduše v tom, že klienti sexuologických ordinací, byť by byli vybráni na základě všech pravidel pro reprezentativní výběr (tj. i kdyby všechny jednotky cílové populace – zde je možná na místě hovořit o základním souboru cílovou populaci zastupujícím – měly stejnou pravděpodobnost být vybrány do souboru výběrového), prostě nejsou reprezentanty pro výpovědi o sexuálním chování české populace, nýbrž pouze pro výpovědi o sexuálním chování souboru klientů poraden a ordinací (se všemi jejich zvláštními charakteristikami).

Proto pozor: **před získáním výběrového souboru musíte vždy vymezit cílovou populaci**, což není vždy tak jednoduché, jak by se na první pohled mohlo zdát.

Ilustrace

Nelze zkoumat hodnotu mateřství jen v souboru žen s dětmi ani jen v souboru žen vdaných. Můžeme se samozřejmě na tuto cílovou populaci omezit, ale pak si musíme být vědomi limitů svých výsledků. Jaký by asi byl příspěvek k poznání toho, „jakou hodnotu má dítě pro ženu", bez dotazování žen, které dítě chtějí, ale mít nemohou (neplodnost, zdravotní problémy), či žen, které mít dítě programově odmítají?

Upozorňujeme také, že z jedné cílové populace, respektive z jednoho základního souboru, lze učinit celou řadu výběrů (k této otázce se dostaneme podrobněji v pasážích věnovaných testování hypotéz a inferenční statistice). Závisí to na velikosti obou, nebo lépe na poměru jejich velikostí.

1.4 Měření

Chceme-li analyzovat výzkumná data, musíme je nejdříve získat. Ačkoliv se kolem nás vyskytuje obrovské množství dat, která jsou již někde uložena a dychtivě čekají, abychom je dále zpracovali,[23] ne vždy jsou po ruce data taková, abychom mohli vyřešit naši výzkumnou otázku. Proto musíme jít často do terénu a vlastní výzkumná data získat, tj. musíme fenomény, s nimiž budeme operovat, změřit.

Již Galileo Galilei formuloval na přelomu 16. a 17. století pro vědu požadavek **měřit všechno, co je měřitelné, a snažit se učinit měřitelným vše, co dosud měřitelné není**.

Tento přístup vedl v 18. a zejména v 19. století k revoluci poznání v přírodních vědách. Není divu, že učaroval i mnoha sociálním vědcům. V pozitivistické empirické sociologii bylo měření od počátku chápáno jako jediná záruka vědeckosti jejích výsledků. Tento radikální požadavek je dnes již minulostí, model poznávání sociální reality pomocí měření ovšem (a to je dobře) mrtvý není. Jen se musel tváří v tvář kvalitativním přístupům zříci monopolu, který byl nejzřetelněji formulován a uplatňován v 30. až 50. letech dvacátého století.[24]

Základním předpokladem měření v sociologii je, že sociální objekt či jev, respektive jeho atributy (vlastnosti), jsou popsatelné pomocí čísel nebo číslic, které jsou jim připsány v procesu měření. Tyto číselné vlastnosti ovšem objekt nemá de facto – jsou mu připisovány teprve v procesu měření.

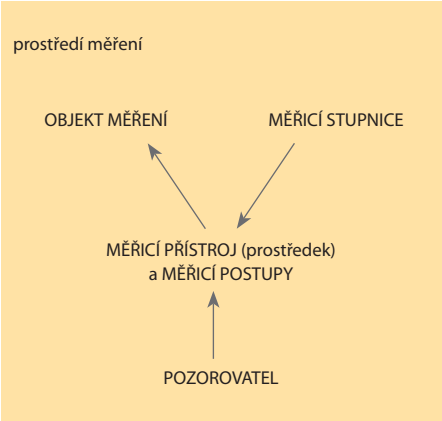
Měření lze jednoduše definovat slovy klasika měření Stanleyho Stevense – který vlastně jen parafrázoval definici jiného klasika měření, Normana Campbella – jako přiřazování čísel objektům a jevům podle pravidel. Český statistik Jan Řehák tuto definici rozvinul a formuloval ji následovně: „Měření je proces realizace homeomorfních zobrazení empirického systému s relacemi do jednoznačně určeného abstraktního systému s relacemi.“ (Řehák, 1971, s. 646)[25]

Celý proces měření lze znázornit jednoduchým schématem (viz obr. 1.2). Popíšeme-li jej slovně (a zde si půjčíme formulaci Řeháka), pak platí: Pozorovatel měří vlastnosti **objektu měření** tak, že mu přiřazuje hodnoty **měřicí stupnice** (škály) pomocí **měřicího přístroje** a měřicích postupů v určitém **prostředí měření**. Z toho vyplývá, že abychom mohli úspěšně měřit, musíme mít: 1) jednoznačně určenou stupnici, jejímž

[23] Data např. leží (v agregované podobě) na statistických úřadech (v Českém statistickém úřadě, v evropské databázi Eurostat, datové databázi OSN atd.) nebo jsou – jako primární data – uložena v datových archivech sociálněvědních výzkumů a čekají na sekundární analýzu, tj. na nové zpracování. Blíže k tomu viz přílohu, která obsahuje internetové linky aktuální v době vydání této publikace.

[24] Upozornění na limity tohoto přístupu naleznete již v klasické práci Charlese Wrighta Millse *Sociologická imaginace* z roku 1959 (česky vydáno v letech 1968 a 2002).

[25] Ostatně jeho stať o měření nazvaná Definice měření ve společenských vědách, kterou publikoval v *Sociologickém časopise* v roce 1971, stojí rozhodně za přečtení.



Obr. 1.2 Obecné schéma procesu měření

Podle: Řehák (1971), s. 639.

prostřednictvím měříme; 2) musí být zkonstruován reliabilní a validní měřicí přístroj a musíme vědět, jaké jsou adekvátní měřicí postupy; 3) musí být maximálně eliminován vliv prostředí a 4) musí být maximálně eliminován vliv subjektu pozorovatele na výsledek měření. Žádná z těchto podmínek není triviální a v sociálních vědách může být vážnou překážkou měření. Pro ilustraci: Jak bychom asi v tomto kontextu změřili míru individuálního blaha (tzv. well-being), která by, jak naznačují někteří ekonomové, mohla sloužit jako obecný ukazatel celkové ekonomické úspěšnosti a ekonomického pokroku společnosti? Na měření blaha přece nemáme ani žádnou reliabilní stupnici, ani validní měřicí prostředek a rovněž by se poněkud obtížně eliminoval vliv prostředí (pokud by v době měření ve společnosti převažovala „blbá nálada", byly by výsledky zřejmě jiné, než kdyby panoval duch veselí a optimismu).

Znovu jsme se tak obloukem dostali k problematice operacionalizace. Teoretický rozbor našeho konceptu a určení jeho dimenzí a subdimenzí po vzoru de Vausova schématu (viz obr. 1.1) by nás jistě přivedl k formulaci adekvátních a validních indikátorů a ke způsobu jejich měření – vždyť přece ve vědě je vše měřitelné, a co není měřitelné, se měřitelným musí učinit![26]

1.4.1 Koncepty a jejich operacionalizace – indikátory

Chceme-li nějaký jev změřit, musíme vědět, jak je vymezen, ohraničen, definován. Musíme tedy mít jeho koncept a jeho definici. Když budeme např. zjišťovat podíl nezaměstnaných v ekonomicky aktivní populaci (míru nezaměstnanosti), musíme mít definici nezaměstnanosti – kdo je za nezaměstnaného považován. Definic konceptu

[26] A skutečně, základní kroky již byly v tomto ohledu učiněny – viz zprávu: Stiglitz, J., Sen, A., & Fitoussi, J.-P. (2008). *Issues paper – Commission on the Measurement of Economic Performance and Social Progress.*

může být ovšem více – ostatně v sociologii jakožto multiparadigmatické vědě je to jev docela častý. Můžeme např. použít definici formulovanou Mezinárodní organizací práce (ILO), která je základem oficiálního určení nezaměstnaných osob v zemích EU, a tedy i v České republice:

Definice ILO považuje za nezaměstnané ty osoby (v metodice Eurostatu ve věku 15–74 let), které: a) v referenčním období neměly zaměstnání, b) neodpracovaly ani jednu hodinu za mzdu nebo odměnu a c) aktivně hledaly práci, d) do které by byly schopny nastoupit nejpozději do dvou týdnů ode dne zjišťování.

Ještě ilustrativnějším příkladem toho, že se neobejdeme bez úvodní definice, může být zkoumání chudoby. Jaké osoby máme do svého výzkumného souboru zahrnout (kdo je vlastně chudý, a tudíž předmětem našeho zájmu)? Osoby s příjmem pod hranicí stanoveného životního minima? Nebo domácnosti vydávající více než 30 % svých příjmů na potraviny? Či osoby s příjmem nedosahujícím 60 % mediánu příjmového rozložení v dané zemi (což je podle definice Eurostatu tzv. příjmová chudoba)? Anebo osoby či domácnosti dosahující určité hodnoty na indexu deprivace? Popřípadě osoby, respektive domácnosti, které se deklarují jako chudé (což je tzv. subjektivní chudoba)? To vše jsou příklady definic chudoby a nejde vůbec o jejich vyčerpávající výčet.[27] Chudobu tedy nikdy nezkoumáme samu o sobě, vždy půjde o její určitý koncept. Budeme-li tedy analyzovat data o chudobě, nebudeme vypovídat o chudobě jako takové, ale o chudobě, jak jsme ji definovali (nebo podle definice chudoby, kterou jsme přijali). Tento aspekt sociálních věd mějme neustále na paměti, když vypovídáme o sociálních jevech, které zkoumáme, to znamená, při interpretaci výsledků našich analýz.

V sociálních vědách se celá věc ještě komplikuje tím, jak jsme již uvedli na začátku této kapitoly, že vlastnosti jednotek často nejsme schopni měřit přímo, takže musíme měřit pouze indikátory (ukazatele) těchto vlastností.[28] Například politickou orientaci jedince z hlediska levice či pravice jsme schopni určit na základě toho, jakou volil stranu v parlamentních volbách, a to ještě pouze na základě jeho výpovědi.
Při měření jevů (konceptů, indikátorů) měříme různé aspekty:

1. **Intenzitu vlastností** zkoumaných jednotek výzkumu a také vlastností objektů vnějšího světa, který je obklopuje (kontextuální vlastnosti – podmínky). V zásadě ale zjišťujeme:
   – Jaká je intenzita nějaké vlastnosti nějakého zkoumaného jevu (objektu) v určitém okamžiku.

---

[27] Jen pro ilustraci, „příjmovou chudobou“ bylo v roce 2014 ohroženo 9,7 % české populace, což bylo vůbec nejméně v celé EU, a tento podíl se víceméně udržel až do roku 2017. Tzv. „subjektivní chudobou“, která se měří ve speciálních výzkumech dotazem na to, jak vychází domácnost respondenta s celkovým měsíčním příjem, a ti, kteří odpovídají, že „s velkými obtížemi“, jsou považování za subjektivně chudé, bylo v roce 2014 postiženo 9,3 % českých obyvatel.

[28] Hledáme něco, co indikuje existenci nějaké vlastnosti, která sama o sobě není pozorovatelná. Čtenáři si to mohou sami zkusit, chtějí-li „změřit“ lásku svého partnera, respektive své partnerky. Říkají-li si „miluje mě“, z čeho tak usuzují, z jakých jeho/jejích projevů?

- K jaké změně intenzity vlastnosti došlo v nezměněných podmínkách v určitém čase.
- K jaké změně intenzity určité vlastnosti došlo v podmínkách, které se v určitém čase známým způsobem změnily

2. **Distanci objektů** (vlastností), což je ve svém primárním významu geometrický pojem, konkretizovaný v teorii měření metrickou veličinou délky. Tato distance může být měřena i v určitém konstruovaném prostoru, jak je tomu v případě prostoru znaků postulovaném Lazarsfeldem (Barton, 1955), v sémantickém prostoru (ten zkoumáme prostřednictvím speciální výzkumné techniky zvané „sémantický diferenciál") či v prostoru vytvářeném statistickou technikou faktorové analýzy (více v kapitole 15). Ve svém sekundárním významu jde o vztah mezi dvěma bezprostředně sousedními škálovými hodnotami, který je numericky reprezentován jako jejich rozdíl vyjádřený v absolutních hodnotách.
3. **Závislosti** (asymetrický vztah) či **souvislosti** (symetrický vztah) mezi dvěma vlastnostmi zkoumaných jednotek (tj. subjektů či objektů výzkumu) či mezi dvěma jevy, respektive mezi proměnnými, jež je reprezentují.
4. **Globální vlastnosti souborů** (například průměrný věk občanů jednotlivých okresů ČR, průměrný příjem vzdělanostních skupin, porodnost v zemích EU apod.).

1.4.2 Proměnná

Kvantitativní sociálněvědní výzkum může nalézt řešení jen pro problémy, které je možno popsat v termínech vztahu mezi pozorovatelnými proměnnými. Proměnné představují logicky uspořádané charakteristiky/vlastnosti zkoumaných jednotek (hodnoty proměnných). Dovolují zkoumané jednotky podle jejich vlastností pouze zařadit do kategorií (nominální proměnné – např. pohlaví, umožňující jednotky zařadit mezi muže, nebo ženy) nebo je zařadit do kategorií uspořádaných podle nějaké míry (ordinální proměnné – např. vzdělání, umožňující jednotky zařadit do kategorií seřazených podle stupňující se míry dané vlastnosti, mezi absolventy pouze základního vzdělání, absolventy středních škol a absolventy vysokých škol) nebo určit číselně intenzitu, jaké daná vlastnost nabývá (spojité neboli kardinální proměnné – např. věk, příjem apod.).[29]

Ilustrace

- Muž a žena jsou vlastnosti (attributes) a současně hodnoty (values) proměnné (variable) nazývané „pohlaví". Tyto dvě vlastnosti představují obor možných hodnot této proměnné (její varianty).
- Zaměstnaný a nezaměstnaný představují dvě vlastnosti a současně dvě hodnoty proměnné „postavení na trhu práce".
- Vzdělání základní, středoškolské a vysokoškolské představují tři vlastnosti a současně tři hodnoty proměnné nazývané „nejvyšší dosažené vzdělání". Jak proměnnou konstruujeme, závisí na našich

[29] Někdy se namísto výrazu „proměnná" používá výraz „znak". Pojem znaku se často používá tehdy, jestliže je kladen důraz na to, že něco značí, zastupuje (nějakou vlastnost, stav apod.). Pojem proměnné používáme tehdy, když se naše úvahy koncentrují na to, jak se tento znak mění, jakých hodnot sledovaný objekt nabývá.

výzkumných otázkách. Proměnnou „vzdělání“ si jistě dovedeme představit i s jiným počtem hodnot (variant). Například ve variantě základní vzdělání můžeme odlišit ty, kdo základní vzdělání nedokončili, od osob s dokončeným základním vzděláním; nebo v rámci vysokoškolského vzdělání můžeme rozlišit stupeň bakalářský a magisterský. V případě, že budeme chtít srovnávat vzdělání v různých zemích, musí být mezi nimi dosaženo shody na obsahu jednotlivých kategorií (jejich standardizace). I v rámci pouhé Evropy se totiž školské soustavy jednotlivých zemí vyvíjely různým způsobem, takže je nutná společná metodika, jak absolventy různých typů škol na škále nejvyššího dosaženého vzdělání umisťovat. K tomu slouží Mezinárodní standardní klasifikace vzdělávání neboli ISCED (International Standard Classification of Education). Vydalo ji UNESCO v roce 1976 a od té doby je pravidelně aktualizována. V současnosti se používá verze z roku 1997, která má 7 úrovní vzdělávání, v každé z nich je navíc ještě vnitřní členění A až C.[30] V České republice ji používá například Výběrové šetření pracovních sil, prováděné Českým statistickým úřadem, běžně se používá ve velkých komparativních výzkumech (například European Social Survey nebo European Values Study zahrnujícím většinu evropských zemí).

– Silný souhlas, souhlas, nesouhlas a silný nesouhlas s privatizací představují čtyři vlastnosti a současně hodnoty proměnné nazývané „postoj k privatizaci“ apod.

Složité vlastnosti lidí jsou tímto způsobem převáděny na relativně jednoduchý soubor informací. Člověk se tak jeví jako statistická jednotka (nositel statistické informace), jako nositel určitých proměnných, nebo lépe řečeno, jako nositel jistých vlastností uvnitř daných proměnných. V kvantitativním výzkumu je člověk:

– mužem či ženou (proměnná pohlaví);
– osobou s určitým počtem let (proměnná věk);
– příslušníkem vyšší, střední či nižší třídy (proměnná sociální třída);
– osobou s určitou výší příjmu (proměnná příjem);
– vlastníkem určitého statku (proměnná vlastnictví tohoto statku);
– osobou s určitým dosaženým vzděláním (proměnná vzdělání);
– příslušníkem nějaké profese (proměnná profese);
– rezidentem v určité komunitě (proměnná bydliště);
– potenciálním nebo i reálným voličem jedné z politických stran (proměnná volební preference);
– jedincem, který nějak tráví svůj volný čas (proměnná struktura volného času);
– osobou, která má určité postavení na trhu práce – má či nemá placené zaměstnání (proměnná ekonomická aktivita);
– nositelem určitého postoje či názoru (proměnnou je postoj: například nesouhlas se snahou vlády zvyšovat daně);
– sociálním aktérem (proměnná akce: například účast na antiglobalistické demonstraci, volební účast – zúčastní se/nezúčastní se voleb apod.).

[30] **ISCED** (International Standard Classification of Education) je nástrojem, který vyvinulo UNESCO pro porovnávání indikátorů a statistik vzdělávání ve svých členských zemích a který bere v úvahu rozdíly jejich vzdělávacích systémů. Více informací najdete v knize *Cesta k datům* (Soukup, 2012) nebo přímo na http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx. Českou uživatelskou příručku pro zavádění ISCED 97 v zemích OECD vydal v roce 1999 Ústav pro informace ve vzdělávání, Praha.

1.4.3 Typy škál – proč jsou důležité

Při měření používáme měřicí stupnice (měřicí škály). Tím obvykle rozumímečíselnou stupnici, předem daný jednoznačný systém, jehož hodnoty jednoznačně odpovídají úrovním měřené vlastnosti. Měřicí stupnice bývá určena svým měřítkem nebo jednotkou a způsobem uspořádání svých hodnot. Pod vlivem Stanleyho Stevense rozlišujeme různé úrovně měření, což závisí na tom, jaký je vztah mezi tím, co se měří, a číslem, které reprezentuje výsledek měření. Tyto úrovně (škály) měření jsou: nominální, ordinální a kardinální (ty se ještě dělí na intervalové a podílové). Podle těchto úrovní potom rozeznáváme proměnné nominální, ordinální a kardinální.

Nominální proměnná je taková, jejíž hodnoty jsou kategorie označené číselnými kódy, které jim výzkumník arbitrárně přiřadil. Příkladem nominální proměnné jsou např. barva vlasů (1. černá, 2. kaštanová, 3. rezavá, 4. blond), okresy ČR, druhy zaměstnání. Byť mají jednotlivé kategorie nominální proměnné číselná označení, není možné je uspořádat do pořadí, neboť přiřazené číslo je pouhým symbolem a v žádném případě neoznačuje množství měřené vlastnosti – objekty zde můžeme pouze přiřazovat do číselných kategorií podle předem stanovených pravidel.

Ordinální proměnná je taková, jejíž kategorie lze uspořádat do pořadí. Zatímco u kategorií nominální proměnné pouze zjišťujeme, zdali se jednotlivé kategorie vyskytly, nebo ne, a pokud se vyskytly, pak jak často, o kategoriích ordinální proměnné jsme schopni říci, která je v pořadí výše než jiná. Příkladem ordinální proměnné je míra spokojenosti (stupnice může mít podobu: 1. velmi spokojen, 2. spokojen, 3. nespokojen, 4. velmi nespokojen), stupeň dosaženého vzdělání, výsledky v soutěži krásy. Pozor tedy: ordinální stupnice zobrazují pouze pořadí, nikoliv stupeň odlišnosti – nedokážeme zde totiž určit, o jaké množství spokojenosti se liší „velmi spokojen" od „spokojen", byť číselná řada by nezkušenému výzkumníkovi sugerovala, že jde o rozdíl/množství jednoho stupně.

Kardinální proměnná je taková, jejíž číselné kategorie již vyjadřují skutečné množství sledované vlastnosti (číselné kódy tedy nejsou arbitrární). Jednotlivé kategorie této proměnné jsme proto nejenom schopni seřadit do pořadí, ale umíme i říci, o kolik (o jaké množství) se liší. Kardinální proměnné rozdělujeme na **proměnné intervalové** (jejich stupnice nemá přirozenou smysluplnou nulu) a **proměnné poměrové** (s existencí přirozené nuly). Poměrové proměnné jsou snem každého sociologického kvantitativního výzkumníka. Důvodem je skutečnost, že u nich jsme schopni říci nejenom to, o kolik se kategorie liší, ale také kolikrát je nějaká kategorie vyšší než jiná (je to dáno právě tím, že poměrové škály mají přirozenou nulu). Příkladem kardinální proměnné je věk, příjem respondenta, počet dětí, které žena porodila, apod. V jemnějším pohledu lze kardinální proměnné ještě členit na **diskrétní**, tedy takové, které mohou nabývat pouze určitých hodnot (většinou to jsou celá čísla, např. počet dětí), a na jejich opak, tedy proměnné **spojité** (*continuous*), které mohou, jak naznačuje název, nabývat jakýchkoliv hodnot. Např. věk člověka může být měřen – podle potřeby přesnosti – nejenom na roky, ale i na dny, sekundy, popřípadě mikrosekundy a nanosekundy…

Zvláštním případem je proměnná, která nabývá pouze dvou kategorií. Označuje se jako proměnná **dichotomická** (nebo také binární).[31] Příkladem je muž – žena, zaměstnaný – nezaměstnaný, živý – mrtvý apod. Chování dichotomických proměnných je trochu zvláštní. Z pohledu metodologie jde o nominální proměnnou, ale pokud užíváme zavedených kódovacích schémat (často 0 vs. 1 či 0 vs. 100), lze s ní ve statistice často operovat jako s proměnnou kardinální.

Rozlišovat úrovně měření a druhy proměnných je velmi důležité, neboť na druhu proměnných jsou závislé statistické operace, které můžeme s daty provádět. Ideálem pro statistiku jsou pochopitelně proměnné kardinální, neboť umožňují nejvyšší množství statistických operací. V sociálních vědách jich ovšem příliš mnoho nemáme, velmi často se pohybujeme pouze v oboru proměnných ordinálních a nominálních.

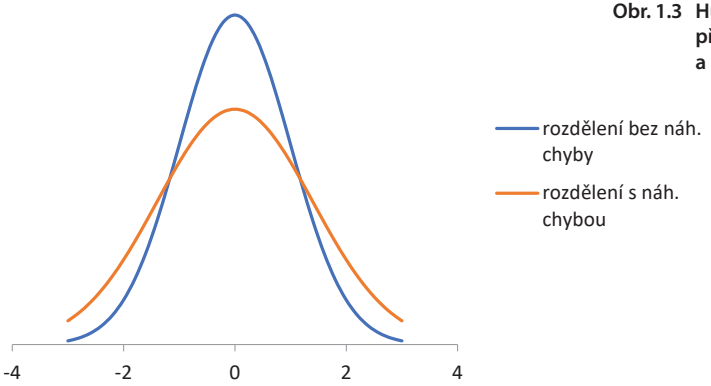
Typy proměnných je třeba doplnit ještě o jednu důležitou klasifikaci, a to o **proměnné nezávislé a závislé**. Základem pro toto dělení již není množství vlastnosti a způsob měření, ale úvaha o tom, co je „příčinou" a co „následkem". Když hledáme souvislosti mezi jevy, začínáme obvykle úvahou o vztahu dvou jevů, dvou proměnných. Přitom předpokládáme, že jeden jev je příčinou, druhý následkem. Proměnná reprezentující v našich úvahách (v naší hypotéze) příčinu je proměnná nezávislá (*independent variable*), proměnná reprezentující uvažovaný následek je proměnná závislá (*dependent variable*). Např. hledáme-li vztah mezi pocitem štěstí a věkem, můžeme formulovat předpoklad, že mladší lidé budou v průměru šťastnější než senioři, tedy že s rostoucím věkem se bude snižovat průměrný pocit štěstí. Rozlišit nezávisle a závisle proměnnou není obvykle problém; pokud si nejsme jisti, vezměme v úvahu časový průběh: v našem případě pocit štěstí prostě nemůže být nezávisle proměnná, pocit štěstí nemůže být příčinou určitého věku.[32] V analýze dat se bez rozlišování nezávisle a závisle proměnných neobejdeme, Hendl dokonce říká, že „výzkum začíná určením nezávisle a závisle proměnných" (Hendl, 2004, s. 40).

Při měření se pochopitelně můžeme dopustit řady chyb. Ostatně **základní teorie měření** (anglicky *true score theory*) s chybou počítá, neboť říká, že každé měření sestává ze dvou (aditivních) složek: ze skutečné správné hodnoty a z chyb měření.

---

[31] Jen pro úplnost: opakem dichotomických proměnných jsou proměnné polytomické (nebo také multinomické, anglicky multinominal), tedy takové, které mají více než dvě kategorie.

[32] Někteří autoři namítají, že tato klasifikace je patřičná pouze pro experimentální výzkum, v němž jsme schopni manipulovat s příčinami: experimentátor vystavuje subjekty experimentu různým podnětům – nezávisle proměnným – a zkoumá jejich reakce (závisle proměnné). Pro analýzu neexperimentálních dat, např. pro data ze sociologických výzkumů založených na dotazníku, navrhují nazývat nezávisle proměnné **prediktorem** (*predictor variable*), závisle proměnné pak **výsledkem** (*outcome variable*). Česká sociologická praxe je taková, že používáme termíny nezávisle a závisle proměnná.



Obr. 1.3 Hustoty pravděpodobnosti při náhodné chybě měření a při měření bez chyby

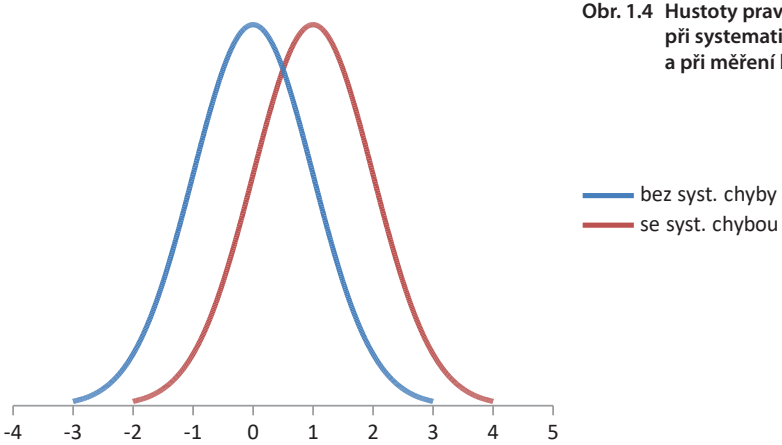
Vyjádřeno jednoduchým vzorcem:

$X = T + e$,

kde X je naměřená hodnota vlastnosti, T je její skutečná, správná (*true*) hodnota, kterou ale neznáme (snažíme se ji totiž zjistit právě našim měřením), a e je celková chyba měření (*error*). Celková chyba měření může být systematická nebo náhodná. **Náhodná chyba** (*random error*) většinou nepředstavuje větší problém, neboť odchylky od měření sice způsobují větší variabilitu v datech, ale jelikož mají tendenci se náhodně vychylovat v kladném i záporném směru (někdy naměříme více, než je skutečná vlastnost, jindy zase méně), vzájemně se vyruší, takže průměr měřené vlastnosti zůstává nedotčen. Proto se někdy o náhodné chybě hovoří jako o šumu (*noise*). Hezky to vyjadřuje obrázek 1.3, v němž modrá křivka ukazuje, jaké bychom měli hodnoty bez náhodné chyby měření, zatímco červená křivka zobrazuje hodnoty s náhodnou chybou. Vidíme, že střední hodnota (průměr) je u obou křivek shodná, rozdíl mezi křivkami je pouze v šířce rozložení hodnot – červená (s chybou) má širší rozložení, hodnoty zde mají větší rozptyl (variabilitu), modrá je sevřenější, rozptyl je menší.

Systematická chyba (*systematic error*) – viz obr. 1.4 – představuje trvalou, nikoliv náhodnou chybu, a pokud se při měření vyskytne, je to skutečný problém. Jelikož jde o systematické zkreslení, jsou hodnoty při měření opakovaně vychýleny buď kladně, nebo záporně, takže se navzájem nemohu vyrušit. Výsledné měření je pak systematicky vychýlené (na obr. 1.4 je to červená křivka), a tudíž zkreslující – průměr hodnot se systematickou chybou (u červené křivky) je odlišný od průměru hodnot, které chybu neobsahují (modrá křivka). Systematickou chybu nazýváme **zkreslením** (*bias*).

Zdroje systematické chyby jsou různé. Chyba může být na straně pozorovatele, pokud je jeho pozorování selektivní (viz schéma měření na obr. 1.2), ale může být způsobena i měřicím nástrojem (pokud je například nevalidní). Jejím zdrojem může být také měřený objekt, například když u výběrových souborů jde o chybný výběr jejich jednotek. Zobecňujeme-li z výběrového souboru na cílovou populaci, může jít o chybu spočívající v neoprávněném zobecnění – hovoříme zde o chybě neoprávněného rozšíření domény



Obr. 1.4 Hustoty pravděpodobnosti při systematické chybě měření a při měření bez chyby

měření, kdy zobecňujeme své výsledky na širší soubor, než je ten, z něhož jsme provedli náš výběr. Může jít například o snahu zobecnit volební preference z výběrového souboru z pražské populace na celou populaci ČR.

Obecně ve statistice **chyba** (*error*) **znamená rozdíl mezi pozorováním a predikcí** (či odhadem) a je spíše nepřesností než skutečnou chybou. Jde většinou o náhodnou chybu. V tomto ohledu se v kapitole o inferenční statistice (viz kapitolu 5) seznámíme se **standardní chybou** (*standard error*) a **výběrovou chybou** (*sampling error*). Zde jde o to, jak (s určitou zvolenou pravděpodobností) jsou výsledky získané výzkumem výběrového souboru blízké příslušným charakteristikám populace.[33]

1.4.4 Aspekty měření

Při měření si musíme všímat několika aspektů. V zásadě platí, že použitý způsob měření by měl být relevantní, validní, reliabilní, senzitivní a přesný. To jsou hlavní aspekty měření (viz též Disman, 1993, s. 62 nebo Babbie, 2001, s. 140–145). **Relevance** znamená vhodnost použité procedury měření ve vztahu k problému. **Validita** se týká rozsahu, v němž měření korespondují se skutečnou vlastností, která má být měřena (měříme skutečně to, co měřit chceme?).

Základní otázkou, kterou si můžeme u každého měřicího nástroje položit, je: „Co měří?“. Validitou se zhruba chápe platnost měřicích procedur, získaných údajů, měřicích nástrojů, prostě všech složek měření a škálování. **Měření je validní, jestliže měří to, co myslíme (očekáváme), že měří.** Ve skutečnosti ovšem není validní nebo nevalidní samo měření, ale jeho použití. Validita tak závisí na tom, jak je definován

[33] V jazyce statistické analýzy označujeme výsledky získané výpočty (ve výběrovém souboru) výrazem „statistiky“ a číselným charakteristikám populace říkáme „parametry“.

měřený pojem.[34] Babbie (2001) konstatuje, že konvenční používání termínu validita se vztahuje k rozsahu, ve kterém empirická míra adekvátně reflektuje skutečný význam uvažovaného pojmu.

Ilustrace

Použijeme-li proměnnou „úroveň dosaženého vzdělání" k měření sociálního statusu, problémem nebude, zda měříme s úspěchem tuto úroveň vzdělání, ale zda tímto způsobem (prostřednictvím tohoto ukazatele) skutečně měříme to, co měřit chceme: tedy sociální status.

Reliabilita označuje rozsah, ve kterém způsob měření dává konzistentní výsledky. U reliability se tedy ptáme, do jaké míry jsou výsledky opakovaného měření shodné s původním měřením.[35] Ilustrativní pro reliabilitu je Segalův zákon, který praví: Když má člověk jen jedny hodinky, ví vždycky, kolik je hodin. Když má člověk dvoje hodinky, nemůže si být nikdy jist.

Reliabilitou se zhruba chápe spolehlivost výsledků měření v závislosti na: a) objektivní spolehlivosti měřicích procedur, technik či nástrojů, b) subjektivní spolehlivosti respondentů a experimentátora.

Reliabilita je také považována za míru stability měřicích nástrojů, s jakou lze při opakovaných měřeních či testech získávat přibližně stejné výsledky. Jde o záruku, že proměnlivé výsledky měření nejsou způsobeny špatným měřicím prostředkem, ale skutečnou variabilitou měřené vlastnosti. Jestliže lidé odpovídají na otázky při opakovaném dotazování stejným způsobem, pak jsou otázky reliabilní. Synonyma reliability jsou: spolehlivost, stabilita, konzistence, prediktabilita neboli předpověditelnost, přesnost (Kerlinger, 1972, s. 421).[36] Zdrojem nereliability mohou být špatné formulace otázek, rozdílné kulturní významy vkládané do použitých termínů v různých sociálně-kulturních prostředích, vliv tazatelů (a to nejen záměrný). Pohlaví tazatele, jeho etnický původ, oblečení, sociální zařazení, to vše může při dotazování (tj. sběru dat) hrát značnou roli. Jinak odpovídá žena ženě, jinak muži a podobně a v tomto odlišném odpovídání se skrývá zdroj chyb.

---

[34] Měříme-li příjmovou chudobu definovanou například poměřením příjmu jednotky s příjmovým rozložením (například za chudé považujeme všechny jednotky s příjmem nižším než 60 % příjmového mediánu), bude dotaz typu „cítíte se být chudou rodinou" nepochybně velmi zajímavý, ale neměří námi definovanou příjmovou chudobu – z tohoto hlediska je nevalidní, i když ve vztahu k subjektivní chudobě jistě validní je.

[35] Požadavek opakovaného měření jakožto kontroly spolehlivosti měřicího nástroje je ovšem v sociálněvědním výzkumu problematický. Stěží si lze např. představit, že půjdeme se stejným dotazníkem ke stejnému respondentovi třeba opakovaně pět dnů po sobě a budeme mu klást stejné otázky.

[36] Koncept reliability je pro nás důležitý obecně, ale i specificky – například v případě sumačních indexů (viz kapitolu 6 Úpravy proměnných a příbuzné procedury), kde je nutno změřit reliabilitu jednotlivých položek (jejich vnitřní konzistenci), z nichž je sumační index vytvořen.

Jiným zdrojem chyb může být kódování, neboť různí tazatelé mohou kódovat stejné odpovědi rozdílně, uplatňuje se selektivní slyšení a podobně. Reliabilita je nízká i u otázek, na které lidé nemají názor (mínění, postoj), a otázka tento názor uměle vytváří (může, ale spíše nemusí zůstat stabilní). Reliabilitu se proto snažíme zvyšovat použitím více indikátorů (zajistit reliabilitu je vždy obtížnější při použití jednoho než při použití více indikátorů), pečlivou formulací otázek, ale i určitým výcvikem (instruktážemi) tazatelů a standardizací způsobu kódování.

Senzitivitou se rozumí schopnost testu dávat pozitivní odpověď, jestliže daná osoba má příslušnou vlastnost, **specificitou** pak schopnost testu dávat negativní odpověď, jestliže daná osoba nemá příslušnou vlastnost. Toto rozlišení zná velmi dobře například medicína a medicínský výzkum.

1.5 Hypotézy a modely

1.5.1 Od tématu přes problém k výzkumné hypotéze

Hypotéza je určité očekávání o povaze věcí, odvozené většinou z teorie (je to tvrzení o tom, jaká má povaha věcí být, má-li být teorie, z které je hypotéza odvozena, pravdivá). **Výzkumná hypotéza** je předběžný předpoklad, domněnka o: 1) existenci a 2) příčinách jevů, 3) o vztahu mezi jevy, 4) o průběhu nějakého procesu, 5) o změně apod. Svou povahou leží na myšlenkovém rozhraní mezi teoretickou a empirickou fází výzkumu. Je návodem k výzkumu, který je na jedné straně determinován dosavadním poznáním, na druhé straně je orientován na další poznání. Může být odvozena z kontextu vědy (formulována z teorie nebo z jiných hypotéz), ale i ze zkušenosti. Má charakter výroku, tvrzení, modelu, či dokonce teorie, ale jsou to výroky, tvrzení, modely a teorie, jež nebyly dosud přijaty jako obecně platné. Její podstatnou charakteristikou je, že ji můžeme různým způsobem empiricky ověřovat. Podle etologa Konráda Lorenze (1903–1989) si nejdříve něco myslíme, pak to srovnáváme se zkušeností a s dalšími smyslovými daty, až nakonec – podle toho, shoduje-li se to s nimi, nebo ne – rozhodneme o správnosti či chybnosti toho, co jsme vymysleli. Hypotéza je něco, co dosud nebylo ověřeno, a má proto v rámci vědy dočasný status (Lorenz, 1990, s. 72–77).[37]

Velmi často se setkáváme s nepochopením při rozlišování mezi tématem výzkumu a výzkumnou otázkou, respektive výzkumnou hypotézou.Rozlišuje se:

– **téma**, respektive předmět výzkumu. Jestliže řekneme, že chceme zkoumat sociální nerovnost, tj. jestliže v nerovnosti vidíme problém (sociální či výzkumný) a chceme výzkumem získat odpovědi na některé otázky týkající se „sociální nerovnosti“, vymezili jsme si prozatím jen téma/předmět výzkumu;

– **výzkumný problém**, na který hledáme prostřednictvím výzkumu odpověď a který je zúžením tématu na některý z jeho (pro nás) významných aspektů;

[37] Připomeňme si Popperovo stanovisko, že všechna vědecká tvrzení jsou vlastně jen hypotézy, jež mají dočasný status (Popper, 1974, s. 87–88).

– **výzkumná otázka/hypotéza**, která problém dále specifikuje do takové podoby, abychom z odpovědi na ni mohli tento problém pochopit – teprve výzkumná otázka určuje otázky, s nimiž se obracíme ke zkoumaným jednotkám: například otázky dotazníku.

Ilustrace

Téma (předmět) výzkumu je sociální nerovnost.

Jaké odpovědi a na jaké otázky chceme získat (co je výzkumný problém)?

– Jaký typ nerovnosti budeme zkoumat (ekonomickou, mocenskou, politickou)? Jak ji budeme měřit (rozdíly ve statusech, v příjmech)?
– Jaký výzkumný prostor zvolíme? Celou společnost, nebo pouze její určitý výsek (organizaci, komunitu, sociální skupinu, rodinu)?
– Jaké období bude časovým rámcem výzkumu? Soustředíme se na jednu generaci, nebo budeme zkoumat i mezigenerační vztahy a procesy (například přenos nerovností z generace na generaci)?
– Jaký aspekt nerovnosti budeme studovat? Zajímá nás rozsah, jakého nerovnost ve společnosti nabývá, její rozložení ve společnosti, její příčiny, její důsledky, funkce ve společnosti nebo ještě něco jiného (například extrémní forma nerovnosti)?
– Bude nás zajímat, které instituce, ideologie, hodnoty a normy nerovnost podporují a jaké mechanismy ji zajišťují, mechanismy, jimiž se nerovnost ve společnosti reprodukuje?
– Jde nám o to, jaké zájmy stojí za udržováním nerovnosti či jak je ideologizována, respektive legitimizována, jak ji lidé hodnotí či jak vysvětlují její příčiny?
– Budeme zkoumat prožitek nerovnosti a deprivaci, anomii či marginalizaci, které ji doprovázejí?

Všechny výše zmíněné prvky ovlivňují povahu hypotéz, které v návaznosti na volbu výzkumného problému formulujeme. Například si zvolíme jako prostor rodinu (nerovnosti v rodině) a budeme se zajímat o genderový aspekt nerovnosti, konkrétně v disponování rodinnými zdroji. Můžeme formulovat hypotézu: *Pohlaví bude mít vliv na osobní spotřebu finančních zdrojů rodiny.* Tu bychom mohli ještě zpřesnit (zaměřit) tím způsobem, že budeme předpovídat, že: *Muži pro svou osobní spotřebu využívají více finančních zdrojů rodiny než ženy.* Vyšší výdaje na vlastní spotřebu indikují větší moc nad finančními prostředky rodiny, a tím i nerovnost v rodině.

1.5.2 Typy hypotéz

Hypotézy lze klasifikovat podle různých hledisek. Jaké druhy tedy rozlišujeme?

Hypotézy **teoretické** a **empirické**. Teoretické hypotézy jsou výroky formulované jazykem teorie. Empirické hypotézy jsou výroky, které jsou empiricky testovatelné.

Hypotézy **výchozí** a **pracovní**. Výchozí hypotézy bývají obvykle hypotézami teoretickými nebo empirickými na vyšší úrovni obecnosti a hypotézami komplexními. Statistická analýza slouží k ověřování pracovních hypotéz, které Disman definuje následujícím způsobem: 1) Pracovní hypotéza je tvrzení předpovídající souvislosti mezi dvěma nebo více proměnnými. 2) Všechny proměnné zmíněné v hypotéze musejí mít validní operační definici. 3) Soubor pracovních hypotéz musí zahrnovat nejen proměnné reprezentující zkoumané koncepty, ale i ty proměnné, které mohou významně zkreslit interpretaci testovaných vztahů (Disman, 1993, s. 79).

Hypotézy **kauzální** a **vztahové**.[38] Hypotézy mohou být hypotézami o stavu či struktuře (zde je užitečná jednorozměrná analýza, neboť technicky jde o otázky týkající se hodnot proměnných) – někdy o nich lze hovořit jako o hypotézách popisných. Většinou však ve statistické analýze testujeme hypotézy o vztahu dvou proměnných (provádíme dvojrozměrnou analýzu) nebo několika proměnných (vícerozměrná, multivariační analýza), které mohou být označeny také jako hypotézy vysvětlující (explanační). Zajímá nás, zda mezi určitými sledovanými proměnnými existují nějaké vztahy a jakou povahu tyto vztahy mají. Primárně nás zajímá, zda jde o vztah statistický, nebo kauzální. Považujeme-li vztah za kauzální, uvažujeme o kauzální hypotéze. Od Lazarsfelda pochází návod, jak mezi dvěma proměnnými identifikovat kauzální vztah: 1) Proměnné musejí být empiricky asociovány nebo korelovány, musejí existovat jejich souběžné změny. 2) Kauzální proměnná, rozumějme příčina, musí v čase předcházet proměnné, kterou ovlivňuje, rozumějme důsledku. 3) Pozorovaný důsledek nemůže být vysvětlen působením jiných proměnných. Pamatujme proto na to, že v datech nalezená asociace nebo korelace nepředstavuje kauzalitu. Popravdě řečeno, empirická sociologie ve většině případů pracuje se statistickými vztahy (které nesplňují třetí ze zmíněných kritérií), a její závěry proto mají vždy pravděpodobnostní charakter. Např.: Čím má člověk vyšší příjem, tím je pravděpodobnější, že bude volit pravicové strany (proměnná *X* má vliv na proměnnou *Y*).

Hypotézy **věcné** a **statistické**. Z hlediska statistické analýzy dat má toto rozlišení zásadní význam. Věcnou hypotézou je domněnka o existenci vztahu mezi dvěma nebo více proměnnými. Statistická hypotéza je hypotetické tvrzení o relacích vyvozených ze vztahu ve věcné hypotéze vyjádřené ve statistických termínech. Věcné hypotézy se mohou týkat: 1) existence, výskytu a stavu předmětů, jevů, událostí, lidí, skupin (Existují či proběhly, jak byly početné?); 2) vlastností předmětů, jevů, událostí, lidí, skupin (Jak byly početné jejich určité kategorie, jakou měly intenzitu?); 3) vztahů mezi předměty, jevy, událostmi, lidmi, skupinami, respektive zda mezi jejich vlastnostmi existovaly, jaké povahy byly (příčinné, statistické), kdy nastaly atd.; 4) vývoje předmětů, jevů, událostí, lidí, skupin, jeho etap a stadií a jejich charakteristik (V jakých etapách se vyvíjely, co bylo pro tyto etapy charakteristické?); 5) procesů, jichž se lidé či skupiny účastní nebo jež v nich probíhají; respektive se mohou týkat vztahů mezi těmito skutečnostmi (existencí, výskytem, stavy, vlastnostmi, procesy apod.).

Věcnou hypotézu není možno podle Kerlingera (1972) samu o sobě testovat, proto testujeme tzv. statistickou hypotézu (představuje tedy způsob testování věcné hypotézy) – což je matematický model sociální reality o chování proměnných, který se snažíme zamítnout (viz pasáž o nulové hypotéze níže).[39] Jde de facto o formulaci věcné hypotézy ve statistických termínech. Protože statistická teorie dovoluje testovat jen jednoduché hypotézy, rozkládá se často komplexní sociologická (věcná) hypotéza na řadu dílčích statistických hypotéz. To, že má hypotéza statistický charakter, znamená

[38] Blíže k otázkám kauzalit v sociálním výzkumu Babbie (2001, s. 68–87).

[39] Pozor! Mnohdy se statistické hypotézy neodlišují od věcných a testování statistických hypotéz je neoprávněně chápáno jako testování hypotéz věcných.

podle Nowaka (1975), že jednoznačně nehovoří o vlastnostech nebo chování každého člověka, nýbrž vypovídá o relativních četnostech, statistických závislostech mezi vlastnostmi a chováním lidí. Tyto závislosti pak platí buď pro určitá seskupení lidí, nebo jsou formulovány jako univerzální teze.

Hypotéza v empirickém výzkumu je výrokem o vztahu alespoň dvou jevů a v konečné fázi (v kvantitativním výzkumu) výrokem o vztahu (alespoň) dvou nebo více proměnných. Čistě teoreticky může být hypotézou i výrok: „staří lidé jsou konzervativní", protože obsahuje implicite dvě proměnné – věk a politickou ideologii. Lépe je ovšem formulovat hypotézu „věk je pozitivně vztažen ke konzervativismu" nebo „s růstem věku roste příklon osob ke konzervativismu" či „s růstem věku roste počet osob volících strany s konzervativním programem" nebo také „čím jsou lidé starší, tím vyšší dosahují skóre na škále konzervativismu" (všimněte si, že tyto hypotézy nevyjadřují vždy totéž).

V posledních dvou případech máme již pracovní hypotézu, neboť obě proměnné jsou operacionalizovány (věk vyjadřujeme v letech a na základě definice konzervativismu jsme schopni určit, které strany můžeme označit jako konzervativní, popřípadě máme k dispozici konstruovanou škálu konzervativismu). Ve zmíněné podobě jde o hypotézu vztahovou a pravděpodobnostní. Netvrdíme, že mezi oběma proměnnými existuje kauzální vztah a že všichni staří lidé jsou konzervativní. Jen předpokládáme, že pravděpodobnost výskytu konzervativismu (podíl osob zastávající konzervativní postoje) mezi starými lidmi je vyšší než mezi lidmi mladými. Je to hypotéza věcná, testovat bychom ji mohli jako statistickou (nulovou) hypotézu „neexistuje vztah mezi věkem jedince a skórem, kterého tento jedinec dosahuje na škále konzervativismu" nebo „koeficient korelace vztahu proměnné věk a proměnné hodnota dosažená jedincem na škále měřící míru konzervativismu je roven nule".[40]

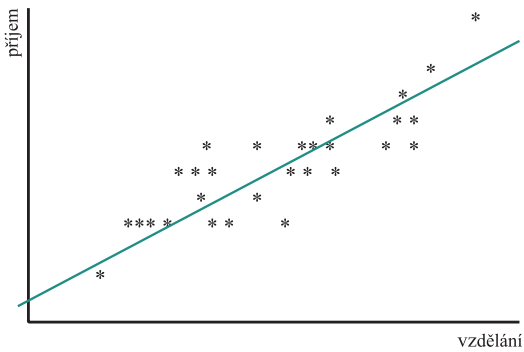
1.5.3 Složitější modely

Modely jsou zvláštním způsobem výkladu reality, který reprodukuje aspekty jevu a umožňuje deduktivní odvození a výroky, jež mohou být přezkoušeny ve zkušenosti. Nejjednoduššími modely jsou **abstraktní popisy systémů** jako modely jejich jistých aspektů. Někdy jsou modely jen synonymem pro teorie – obvykle velmi jednoduché – nebo pro část teorie. Identifikujeme i tzv. **konceptuální modely**. Ty jsou pokusem prezentovat sociální svět v pojmech řady vzájemně vztažených pojmů (blahobyt, moc, prestiž jako odměna za výkon určité role). **Teoretické modely** představují zabudování určité teorie, která vysvětluje výběr a uspořádání prvků modelu (např. teorie racionální volby umožňuje uspořádat jednotlivé kroky jednajícího aktéra). Mohou mít i charakter metaforických analogií. Modely mohou mít verbální či matematickou prezentaci, mohou být také prezentovány jako diagramy.

---

[40] Měli bychom ale být opatrní, protože jak si ukážeme dále, nulová hodnota korelačního koeficientu znamená pouze neexistenci asociace u vztahů lineárních (mohlo by se stát, že vztah existuje, ale není lineární).

Model na obr. 1.5 je matematické vyjádření vztahu mezi výší příjmu a výší vzdělání. Modelem tohoto vztahu je přímka, která říká, že vztah mezi oběma proměnnými je lineární. Model navíc vyjadřuje, že čím vyšší je vzdělání, tím vyšší je také příjem.[41] Pokud bychom tento vztah vyjádřili matematicky ve formě rovnice, mohli bychom přesně predikovat, o kolik jednotek vzroste příjem (závisle proměnná), zvýší-li se výše vzdělání o jednotku.[42]



Obr. 1.5 Vztah mezi vzděláním a příjmem (smyšlená data)

Na dalším obrázku (viz obr. 1.6, což je model Blaua a Duncana, který pochází z teorie sociální stratifikace) je zachycen příklad složitějšího modelu vyjadřujícího vztah mezi jednotlivými proměnnými, které byly předmětem zkoumání (čísla zde představují sílu vztahu mezi jednotlivými prvky vyjádřenou specifickými koeficienty asociace). V tomto případě jde o model dosahování sociálního statusu respondenta/respondentky v závislosti na vzdělání a zaměstnaneckém statusu otce.

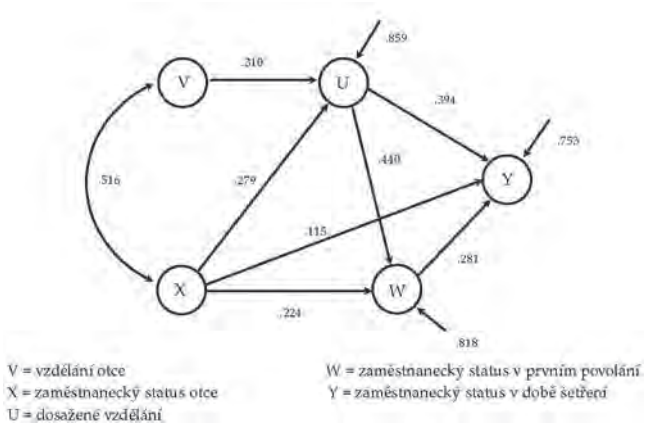
Sociální systémy, které jsou předmětem zkoumání v sociologii, jsou ovšem systémy: a) komplexními a b) otevřenými. Jejich **komplexnost** má za následek, že když je popíšeme jednoduchým modelem, dochází ke značnému zjednodušení, k simplifikaci. Ta je ovšem nezbytná (čím složitější modely konstruujeme, tím více se vymykají

[41] Takový vztah je nejen lineární, ale i přímo úměrný, neboť s růstem hodnoty jedné proměnné roste současně i hodnota druhé proměnné. U nepřímo úměrného vztahu s růstem hodnot jedné proměnné hodnoty druhé proměnné klesají: například bylo empiricky ověřeno, že s růstem příjmu klesá míra anomie měřená na Sroleho škále (viz Rabušic & Mareš, 1996) – tento vztah je nejen nepřímo úměrný, ale také asymetrický, což znamená, že jedna proměnná je jasně nezávislá a druhá závislá. Hypoteticky bychom sice mohli předpokládat, že anomie jako nezávisle proměnná bude ovlivňovat závisle proměnnou, tj. výši příjmu, což bychom spekulativně formulovali tak, že anomičtí lidé jsou natolik ve svých životech frustrovaní, že nepodávají v práci řádný výkon, a tím jsou jejich příjmy nižší, ale logičtější je očekávat příčinu a důsledek v opačném gardu: příjem ovlivňuje míru anomie. K otázce aspektů vztahů mezi proměnnými se ještě dostaneme.

[42] V tomto konkrétním případě bychom ale museli vzdělání vyjádřit ve formě kardinální proměnné, například počtem let strávených ve škole (studiem), a nikoliv, jak je obvyklé, stupněm dosaženého vzdělání (základní, středoškolské, vysokoškolské), což je proměnná ordinální. S ordinálními proměnnými totiž, striktně řečeno, nemá regresní analýza pracovat.

možnosti testovat je) a většinou i postačující, protože se nesnažíme o úplné vysvětlení jevů, ale spíše o odkrytí míry vlivu určitých konkrétních stimulů na ně.

Otevřenost sociálních systémů znamená, že žádný z námi konstruovaných modelů nevystačí jen s proměnnými, které jsou do něho zahrnuty. Vždy musíme počítat s vlivem proměnných, které v našem modelu nejsou obsaženy, neboť nebyly dostupné našemu měření nebo o jejich vlivu, či dokonce o jejich existenci nemáme ani ponětí. Všimněme si, že Blau a Duncan v demonstrovaném modelu s těmito vlivy počítají (viz šipky směřující zvenčí do bodů U, Y a W na obr. 1.6).



Obr. 1.6 Model proměnných ovlivňujících dosažení sociálního statusu

Podle: Blau, P. M., & Duncan, O. D. (1967). *The American Occupational Structure*. New York: Wiley. Převzato z Matějů, P. (2005). Ke kořenům sociálně psychologického modelu sociální stratifikace. *Sociologický časopis*, *41*(1), 7–30.

I když se svými modely pracujeme jako s uzavřenými systémy, statistika nám v některých svých procedurách umožňuje alespoň odhadnout, jaký podíl variance závisle proměnné/proměnných z našeho modelu je vysvětlen nezávislými proměnnými zahrnutými do našeho modelu a jaký je podíl vlivu proměnných nezahrnutých do modelu.[43] Podíl vysvětlené variance závisle proměnné vlivem námi použitých nezávislých proměnných může být někdy i překvapivě vysoký, většinou jsme však nadmíru spokojeni i s 50–60 %. Zbylé procento variance závisle proměnné (dopočítáváme do 100 %) jde pak na vrub proměnných, se kterými v modelu nepočítáme, a to většinou proto, že ani nevíme, které by to mohly být. Můžeme sice experimentovat a propočítávat modely s různými proměnnými (měnit jejich počet) i s různými předpoklady vztahů mezi nimi, jsme však přitom omezeni jen na proměnné, které máme ve svých datech.

Ilustrace

Zjistili jsme, že výše příjmu silně koreluje s věkem, ale koeficient determinace nám říká, že věk jako nezávisle proměnná vysvětluje jen 36 % variance v příjmu, zbylých 64 % jde na vrub vlivu jiných proměnných (tušíme, že ve hře je přinejmenším i vzdělání).

[43] U procedur, se kterými se zde seznamujeme, se s tím setkáme například u lineární regrese, binární logistické regrese (koeficient determinace či jeho obdoby) nebo u faktorové analýzy (viz kapitoly 11–15).

1.6 Jak získat data pro analýzu

Hromadná data pro analýzu lze získat mnoha způsoby. Primární podmínkou je, aby byla získávána jako data standardizovaná. Například používáme-li pro jejich sběr dotazník, musíme vedle takových podmínek, aby všem byly kladeny stejné otázky ve stejném znění a stejném pořadí, zajistit také podmínky plynoucí ze statistických požadavků na proměnné:

– Proměnná musí variovat, musí tedy nabývat alespoň dvou hodnot (hovoříme zde o diskriminabilitě čili **rozlišitelnosti** mezi vlastnostmi objektu uvnitř proměnné: například u proměnné pohlaví lze rozlišit mezi muži a ženami).
– Ke každému stavu vlastnosti existuje příslušná hodnota znaku (**zařaditelnost**). Všechny pozorovatelné vlastnosti objektu musejí být zařaditelné do některé z hodnot proměnné (např. proměnná „volební preference“ má mít tolik hodnot, kolik politických stran postavilo své kandidátky do parlamentních voleb).
– Dvě různé hodnoty znaku nemohou odpovídat jednomu stavu vlastnosti (**jednoznačnost**). U žádné z pozorovatelných vlastností objektu nemůžeme být na rozpacích, jakou hodnotu proměnné jí přiřadit neboli do které kategorie ji zařadit. Například proměnná věk nemůže mít vedle hodnoty 20–30 let hodnotu 30–40 let, neboť by nebylo možné jednoznačně rozhodnout, do které kategorie by měly být zařazeny třicetileté osoby.
– Naše data by měla být reprezentativní (viz předchozí text), aby nám umožnila zobecnění výsledků našich výpočtů z výběrového souboru na soubor základní (za využití statistické inference – viz dále) a do výzkumu musí být zahrnut dostatečný počet výzkumných jednotek.[44]

Poslední podmínku je ovšem třeba mírně korigovat: ne vždy musejí být naše data reprezentativní. Máme-li například malou populaci (studenty jednoho gymnázia), uděláme vyčerpávající šetření (census), při kterém do výzkumu jednoduše zahrneme všechny její členy. V takovém případě ovšem ztrácí smysl tzv. inferenční statistika (viz kapitolu 5), kterou lze aplikovat pouze v souborech, jehož jednotky byly vybrány náhodně.

Zdrojem hromadných dat pro statistickou analýzu může být především vlastní sběr dat, dále data posbíraná jinými výzkumníky nebo statistické výkaznictví a speciální šetření, jako je například sčítání lidu nebo mikrocensus.

[44] Mlhavý výraz „dostatečný“ si blíže specifikujeme, až se dostaneme k inferenční statistice a k otázce výběrové chyby. Co tímto výrazem míníme, závisí na velikosti výběrové chyby, kterou jsme ochotni připustit, a také na hloubce zamýšlené analýzy (do kolika podsouborů budeme soubor, s nímž pracujeme, členit). Může nás uspokojit třeba i 300–400 výzkumných jednotek, ale nemusí nám stačit ani 5 000 výzkumných jednotek. Vše závisí na řadě okolností.