# Úvod

[…] in this world nothing can be said to be certain, except death and taxes.

Benjamin Franklin, 1789

Tento známý citát, který pochází z korespondence Benjamina Franklina, vtipně vystihuje jeden z aspektů současného světa. Svět společenských vědců pracujících s daty je však ve srovnání s běžnými smrtelníky platícími daně ještě o trochu pestřejší. Jejich svět totiž disponuje i třetí jistotou. Tu třetí jistotu představují chybějící hodnoty v datech, které společenskovědní výzkumníci analyzují. Nezáleží na tom, jestli pracují s individuálními daty nebo s daty za nějaké vyšší agregované celky. Chybějící hodnoty jsou všudypřítomným fenoménem (nejen) společenskovědních dat.

Chybějící hodnoty představují komplikaci, se kterou je při věcné analýze potřebné se vypořádat. Jinými slovy, výzkumník musí zvolit metodu, která tento problém řeší. Volba konkrétní strategie přitom může mít různé následky. V nejkrajnějším případě mohou scházející údaje vést až k nesprávným závěrům realizovaných věcných analýz. Na druhé straně může mít jiná strategie minimální vliv na závěry vědeckého zkoumání. Předkládaná publikace prezentuje několik metod řešících problém chybějících hodnot v datech. Zároveň jsou zde definovány takzvané mechanismy chybějících hodnot, které do značné míry předurčují vhodnost aplikování jednotlivých metod.

V rámci statistiky představují chybějící hodnoty už od sedmdesátých let minulého století svébytnou oblast vědeckého zájmu, která je charakterizována vlastní terminologií, taxonomií, notací a také širokým okruhem publikací (Molenberghs 2007: 861). Byli to právě statistici, kteří navrhli množství metod pro řešení problému chybějících hodnot, přičemž mnohé z nich se postupem času staly součástí programů pro analýzu dat. Oblast statistiky nadále zůstává v čele výzkumu o chybějících hodnotách, jelikož vývoj nových a vylepšování stávajících technik neustále pokračuje.

Anglosaské společenskovědní prostředí pohotově zareagovalo na tento vývoj v oblasti statistiky a problematice chybějících hodnot je v tomto prostředí věnována významná pozornost. V angličtině existuje hned několik knih (např. Graham 2012; Enders 2010; Allison 2002), které „překládají“ matematicko-statistický jazyk do přístupné podoby pro méně technicky zdatné společenské vědce. Dále jednotlivé společenskovědní společnosti (např. American Political Science Association nebo American Psychological Association) formulují přesná doporučení, jak při publikování reportovat rozsah chybějících hodnot v analyzovaných datech a explicitně uvádět použité metody pro řešení tohoto problému. Někteří společenští vědci s techničtějším vzděláním (např. Stef van Buuren nebo Gary King) dokonce implementovali navržené metody do vlastních programů, které jsou následně volně využívány širší vědeckou komunitou.

V českém společenskovědním prostředí dosud neexistovala žádná publikace, která by se komplexně věnovala problematice scházejících údajů v datech. Vzhledem

k všudypřítomnosti chybějících hodnot v společenskovědních datech proto předkládaná kniha představuje jednoznačný průlom. Kniha totiž na stále relativně malém prostoru představuje všechny klíčové aspekty této problematiky a zároveň na praktických příkladech názorně demonstruje rizika spojené s výskytem chybějících hodnot v datech.

Pro úplnost dodejme, že v českém jazyce byl problematice chybějících hodnot do vydání této monografie věnován pouze jeden vědecký článek (Pejčoch 2011) a čtyři univerzitní závěrečné práce (Kupčák 2012, Nárožná 2013, Robotková 2011, Rychlá 2011). Pejčoch prezentuje různé metody práce s chybějícími hodnotami na prostoru menším než deset stran. Uvedené závěrečné práce byly vytvořeny výhradně na přírodovědeckých fakultách, což dále poukazuje na zanedbání tohoto tématu komunitou společenských vědců. Představená monografie je ve srovnání s dosavadními českými publikacemi mnohem rozsáhlejší.

Tato kniha je primárně určená českým a slovenským společenským vědcům, kteří pracují s daty z výběrových šetření. Zároveň je kniha určená i pro studenty společenskovědních oborů, jako např. sociologie, politologie a psychologie, kteří se při analyzování kvantitativních dat s chybějícími hodnotami rovněž nevyhnutelně setkávají. Co se týče způsobů řešení tohoto problému, kniha pojednává o všech metodách určených k práci s chybějícími hodnotami implementovanými v SPSS. Společenští vědci se tak seznámí s možnostmi, které poskytuje v českém prostředí tento stále nejvíce rozšířený program pro analýzu dat.

Pomocí realizovaných vlastních simulací kniha detailně vysvětluje výhody a nevýhody různých metod. Čtenáři jsou tak v praktických výzkumných situacích demonstrovány rozličné aspekty statistického usuzování při analyzování dat se scházejícími údaji. Společenskovědní výzkumník si tak může udělat představu o potenciálních „hrozbách" v analogických situacích, a případně tak i zvolit vhodnou metodu práce s chybějícími hodnotami. Zároveň kniha prezentuje jednu z možností, jak ve společenskovědní praxi vyžít simulace jako analytický nástroj. Námi realizované simulace jsou přitom plně reprodukovatelné.

Kniha u čtenářů předpokládá aktivní znalost základů statistického usuzování z výběrových dat na populaci (tzn. konceptů, jako jsou intervaly spolehlivosti a testy významnosti). Zároveň se předpokládá familiárnost s některými atributy kvality bodových (např. konzistence, nestrannost a eficience) a intervalových odhadů (např. přesnost). Čtenáři by dále měli ovládat základy korelační analýzy a mnohonásobné lineární regrese. Ačkoliv se tato kniha primárně snaží vyhýbat matematickým zápisům, na některých místech jsme kvůli zjednodušení výkladu a jednoznačnosti tyto zápisy zařadili. To se týká zejména třetí kapitoly, která obsahuje hned několik matematických vzorců. Pro zjednodušení orientace v matematických zápisech slouží přehled používaného značení, který je umístěn na začátku knihy.

Základy pro vznik této monografie byly položeny na Fakultě sociálních věd Univerzity Karlovy v Praze. Kniha totiž představuje rozšířenou a doplněnou verzi magisterské diplomové práce, která byla obhájena na katedře sociologie této fakulty v akademickém roce 2013/2014. Diplomová práce byla napsána pod odborným vede-

ním PhDr. Ing. Petra Soukupa. Závěrečná magisterská práce se skládala ze tří hlavních kapitol, přičemž tyto kapitoly byly v mírně pozměněné podobě převzaty pro potřebu této monografie.

## Struktura knihy

Kniha je rozdělena na teoretickou a analytickou část. Teoretická část se skládá ze tří kapitol, které se zaměřují na základní aspekty problematiky. Jednak jsou v nich představeny klíčové koncepty spojené s chybějícími hodnotami a zároveň i metody práce s chybějícími hodnotami. Analytická část se také skládá ze tří kapitol. Tyto kapitoly je možné označit jako analytické, jelikož prakticky aplikují teoretické koncepty a prostřednictvím dvou vlastních simulačních studií srovnávají různé metody práce s chybějícími hodnotami. Nyní představíme jednotlivé kapitoly ve větším detailu.

První kapitola nejdříve definuje chybějící hodnoty. Následně jsou popsané vzorce chybějících hodnot, které charakterizují různé rozmístění chybějících hodnot v datové matici. Zbytek kapitoly pojednává o mechanismech chybějících hodnot. Tyto mechanismy definují různé druhy vztahů mezi chybějícími a pozorovanými hodnotami v datech.

Druhá kapitola se věnuje metodám práce s chybějícími hodnotami. Jinými slovy, jsou v ní představeny různé techniky, které je možné použít při věcné analýze dat se scházejícími údaji. Nejdříve jsou představeny metody založené na vynechávání případů s chybějícími hodnotami z analýzy. Následují metody založené na nahrazování chybějících hodnot. Nakonec je prostor věnován metodám založeným na maximální věrohodnosti.

Třetí kapitola pojednává o mnohonásobných imputacích chybějících hodnot. Jedná se o metodu, při které je každá chybějící hodnota nahrazena současně několika různými hodnotami. Mnohonásobné imputace bývají v poslední době považovány za jedno z nejvhodnějších řešení problému scházejících údajů v datech. Kvůli narůstajícímu významu této metody a její značné komplexnosti je právě této metodě věnována samostatná kapitola.

Čtvrtá kapitola popisuje metodologii analytické části. Kapitola začíná vymezením výzkumných otázek. Hlavním cílem studie je porovnat fungování představených metod práce s chybějícími hodnotami v různých situacích. Konkrétně nás zajímá vliv mechanismů chybějících hodnot a podílů chybějících hodnot v datech na získané odhady zkoumaných parametrů. Tato metodologická kapitola tedy představuje postup při realizaci simulací potřebných k zodpovězení výzkumných otázek. Krátce řečeno, v rámci simulací jsou nejdříve generovány tisíce výběrů z populace. Následně jsou v jednotlivých výběrech podle různých mechanismů vytvářeny chybějící hodnoty. Na takto redukovaných datech jsou s pomocí metod práce s chybějícími hodnotami realizovány věcné analýzy. Výsledky těchto analýz jsou následně srovnávány se známými hodnotami populačních parametrů. Čtvrtá kapitola tak představuje společný rámec pro simulační studie realizované ve dvou zbývajících kapitolách knihy.

V páté kapitole je realizována první simulační studie, která srovnává čtyři tradiční metody práce s chybějícími hodnotami (analýzu kompletních případů, nahrazování chybějících hodnot aritmetickým průměrem, nahrazování regresí a nahrazování stochastickou regresí). Věcně je tato simulace zasazena do kontextu zkoumání vztahu mezi hrubou měsíční mzdou a vrozenou inteligencí. Výsledky studie ukazují, že v některých situacích mohou mít tradiční metody práce s chybějícími hodnotami drastické důsledky na získané odhady populačních parametrů (zejména v případě nahrazování chybějících hodnot aritmetickým průměrem a prostřednictvím regrese).

Šestá kapitola obsahuje druhou simulační studii, která je ve srovnání s předcházející v několika ohledech komplexnější (např. počet a typ proměnných, vzorec chybějících hodnot). Studie přebírá dvě nejúspěšnější tradiční metody z předchozí simulace (analýza kompletních případů, nahrazování stochastickou regresí) a přidává analýzu dostupných případů jako „sesterskou“ metodu analýzy kompletních případů. Zároveň jsou součástí této simulační studie i mnohonásobné imputace, které jsou kvůli své technické náročnosti ve společenskovědní praxi zatím relativně málo využívané. Věcně tato simulační studie analyzuje vliv chybějících hodnot při zkoumání determinantů politické znalosti. Konkrétně se zajímáme o odhady průměrné politické znalosti v populaci oprávněných voličů České republiky a regresních koeficientů u determinantů politické znalosti (pohlaví, dosažené vzdělání a zájem o politiku). Výsledky simulační studie dokazují, že mnohonásobné imputace představují nejvhodnější techniku ze všech námi zkoumaných metod.