NON-RESPONSE: RŮZNÉ MECHANISMY NON-RESPONSE, OŠETŘENÍ NON-RESPONSE VÁŽENÍM, SIMULAČNÍ STUDIE PRO RŮZNÉ MECHANISMY NON-RESPONSE

Lucie Janošíková, Veronika Myslilová

Abstrakt

Non-response je velmi častý problém vyskytující se při realizaci výběrových šetření. Tato práce se zaměřuje na představení dané problematiky jako celku, s důrazem na různé způsoby ošetření non-response a jejich praktickou aplikaci. První část definuje non-response, popisuje její různé druhy a příčiny vzniku. Druhá část představuje teoretické vymezení mechanismů ošetření non-response, jimiž jsou vážení a imputace. Třetí část aplikuje již představené mechanismy na výukovém datovém souboru Datového archivu Sociologického ústavu AVČR o mzdách zaměstnanců se simulovanou non-responsí a porovnává jejich výsledky mezi sebou na ilustračních příkladech.

Klíčová slova: non-response, vážení, imputace, výběrová šetření

Obsah

Abstrakt 1
Úvod 3
1. Non-response a její typy 4
1.1. Jednotková non-response 4
1.2. Položková non-response 4
1.3. Typy chybějících pozorování 4
2. Ošetření non-response 5
2.1. Jednotková non-response 6
2.2. Položková non-response 6
3. Praktická část 8
3.1. Jednotková non-response 8
3.2. Položková non-response 9
Závěr 12
Zdroje 13

Úvod

Práce se zabývá problémem non-response ve výběrových šetřeních. Tento častý jev může působit značné problémy při výsledcích jednotlivých šetřeních, jelikož hodnoty pak můžou být zkreslené. Non-response se vyskytuje prakticky v každém šetření a do určité míry se dá tolerovat, avšak nikdy bychom tento problém neměli ignorovat. Cílem této práce je tak ukázat, jak můžeme problematiku non-response řešit a představit příčiny a důvody jejího vzniku. Ačkoliv přístupů k řešení non-response je celá řada a zabývá se jimi spoustu odborníků, cílem této práce je představit hlavně ty metody, které jsou nejvíce rozšířené. Vše bude znázorněno na ukázkovém datovém souboru, který se týká mezd. Jelikož přístup k řešení non-response se liší dle jejího typu, jsou představeny také jednotlivé druhy a rozdíly mezi nimi.

1. Non-response a její typy

Non-response je neochota nebo neschopnost respondentů poskytnout informace pro účely výběrového šetření (1). Jedná se tedy o chybějící pozorování. Mohou chybět buď pouze jednotlivé proměnné, ale i kompletní údaje pro jednotky výběru. Podle toho, jaké informace nám chybí, rozlišujeme základní dva typy: i) non-response jednotkovou (unit non-response); ii) položkovou (item non-response) (2).

1.1. Jednotková non-response

Jednotková non-response vzniká, když nemáme žádnou informaci o vybrané jednotce. Jedná se tedy o neprošetření vybraných jednotek. Příčinou tohoto jevu může být nezastižení respondenta či jeho odmítnutí se participovat na šetření (3).

1.2. Položková non-response

Položková non-response nastává, pokud chybí některé údaje v prošetřených jednotkách nebo u jednotlivých proměnných. Tento problém může vzniknout např. tím, že respondent neporozumí otázce, nezná na ni odpověď či odmítá odpovědět z různých důvodů (např. jedná se o citlivé téma) nebo nebyl v daný čas zastižen (šetření v domácnostech, telefonická šetření). Příčinou non-response může být i špatně položená otázka nebo otázka, která nemá pro respondenta smysl.

1.3. Typy chybějících pozorování

Rozlišujeme různé druhy chybějících pozorování podle toho, jakým způsobem vznikly.

Prvním typem je MCAR (Missing Completely At Random – chybějící zcela náhodně), kdy pravděpodobnost chybějícího pozorování je stejná pro všechny jednotky a nezávisí na ostatních jednotkách, sledované proměnné či způsobu měření (3, 4, 5). Tudíž příčina chybějících dat nesouvisí s daty samotnými (6).

Druhým typem je MAR (Missing At Random – chybějící náhodně). Zde chybějící údaje závisí na vysvětlující proměnné (vyjma cílové proměnné) či již zjištěných hodnotách (3, 5). Pravděpodobnost chybějícího pozorování je stejná pouze pro skupiny definované již zjištěnými daty (6).

Posledním typem je NMAR (Not Missing At Random – chybějící nenáhodně). V tomto případě chybějící proměnné závisí na cílové proměnné (5). Jsou obecně dané postupy, kterými lze rozpoznat typ chybějících pozorování. Např. ověření shody středních hodnot veličiny u dvou skupin, které jsou rozděleny podle toho, zda mají chybějící údaje u nějaké jiné veličiny (3).

2. Ošetření non-response

Problému non-response se chceme zbavit hlavně ze dvou důvodů. Prvním je, že non-response způsobuje zkreslení bodových odhadů a druhým, že non-response zvyšuje rozptyl bodových odhadů, jelikož rozsah výběru je menší než plánovaný (2). Cílem je tedy snížit zkreslení vzniklé díky non-response, které je způsobeno převážně tím, že respondenti a non-respondenti se mezi sebou liší ve zjišťovaných proměnných (2).

Abychom předešli non-response, můžeme udělat řadu opatření již před provedením šetření. Při návrhu šetření je nutné promyslet veškeré jeho složky jako např. délku šetření, způsob dotazování, finanční aspekty, proveditelnost atd. a nastavit je tak, aby se non-response minimalizovala (4, 5). Za stejným účelem je možné také provést pilotní šetření (5). Lze využít také zkušeností z předešlých podobných šetření. Je nutné dopředu zvážit zatížení respondenta (časová dotace, náročnost dotazníku, typy otázek) (4).

Během šetření může být prospěšné vysvětlit respondentům jeho důležitost a smysl, což také vede ke snížení non-response. Dále lze respondenty motivovat k účastí pomocí různých odměn a benefitů (4, 5). V případě, že respondent v prvním dotazování neodpoví, lze využít urgence (callbacks), tedy opětovného oslovení respondenta, a to například telefonicky nebo osobně.

Pokud jsme provedli všechny výše zmíněné kroky, stejně musíme vyřešit non-response při zpracování dat, neboť ignorovat nonresponse není akceptovatelné. Prvním řešením může být vynechání chybějících údajů. „Listwise“ metoda vynechá celou prošetřenou jednotku. Je však vhodný pouze pokud podíl chybějících údajů je malý a předpokládáme MCAR, zároveň tento postup zvětšuje jednotkovou non-response (3, 5). Existuje také metoda „pairwise”, kdy se použijí pouze dosažitelné hodnoty, avšak toto může způsobit značné problémy při výpočtu směrodatných chyb, jelikož jsou zde různě velké rozsahy u jednotlivých proměnných (3).

Pro jednotlivé typy non-response se postupy liší. Pokud jde o jednotkovou, lze ji ošetřit došetřením z non-respondentů (tedy snaha došetřit údaje o konkrétních jednotkách) či pomocí vážení zohlednit chybějící jednotky. U položkové non-response se aplikují statistické postupy, které vytvoří hodnoty, které doplníme za chybějící údaje – jde tedy o imputaci hodnot. Alternativou pak může být dohledání chybějících údajů pomocí administrativních zdrojů a externích databází (5).

2.1. Jednotková non-response

Pokud provedeme libovolné šetření a jeho součástí je jednotková non-response, je potřeba ji ošetřit. Důvodem je hlavně zkreslení, které by vzniklo, pokud bychom zkoumali pouze ty jednotky, u kterých informaci máme. Non-respondenti mohou vykazovat odlišné charakteristiky než respondenti. Jedním z řešení může být následné rozdělení souboru do dvou strat – na respondenty n_1 a non-respondenty n_2 a došetření mezi non-respondenty (6). V praxi toto znamená např. opětovné volání nezastiženým respondentům či jejich další návštěva. Došetřením nám vzniká část výběru o velikost $n_2' \leq n_2$. Tudíž konečný rozsah výběru je dán nasčítáním přes strata: $n' = n_1 + n_2'$. Velikost n' závisí na tom, kolik z původní skupiny non-respondentů n_2. se nám podařilo dovybrat. Odhad průměru poté vzniká jako vážený průměr průměrů obou strat (5,7):

$$\frac{n_1}{n'} \cdot \overline{y_1} + \frac{n_2'}{n'} \cdot \overline{y_2} \quad . \tag{1.1}$$

Jelikož se jedná o stratifikovaný výběr, odhady jsou zkreslené, přičemž zkreslení je úměrné podílu $\frac{n_2}{n}$ a rozdílu průměrů respondentů a non-respondentů (5).

Jestliže se nám nepodaří provést úplné došetření ve skupině non-respondentů, lze využít metodu vážení. Váhy sestrojíme buď podle nějaké vedlejší informace o neprošetřených jednotkách, nebo odhadneme pravděpodobnost zahrnutí jednotky do výběru. Váhy pak určíme jako 1/pravděpodobnost zahrnutí, odpovídající výběru o rozsahu n. Lze také váhy transformovat tak, aby se jejich součet rovnal n_1 (3, 5). Pokud máme vedlejší informaci o poměrném zastoupení různých jednotek, váhy poté sestrojíme jako zlomek, kde v čitateli je hodnota zastoupení v populaci a ve jmenovateli je hodnota zastoupení ve vzorku.

2.2. Položková non-response

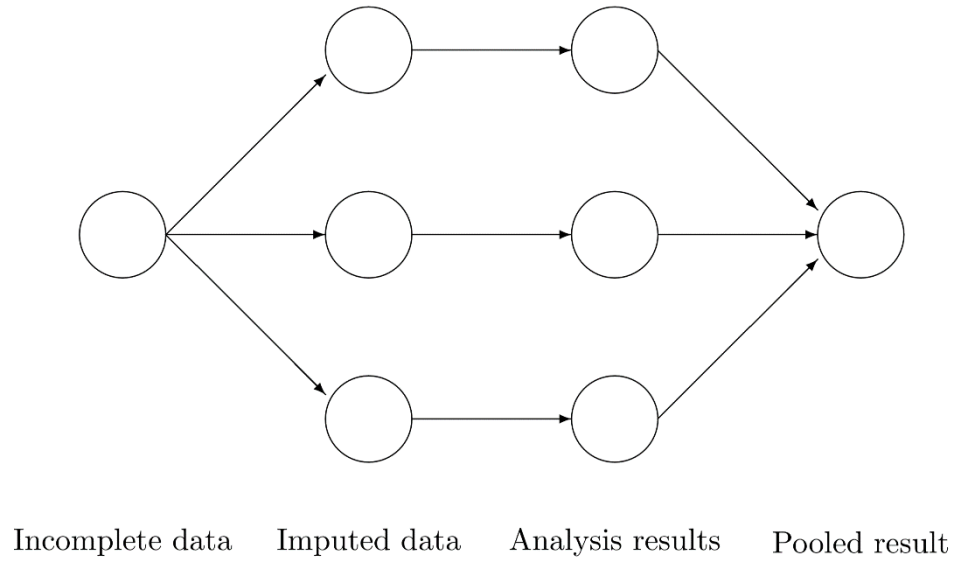
Řešení položkové non-response se provádí pomocí imputace, tedy nahrazením chybějících hodnot pomocí jiných údajů. Imputovanou hodnotu můžeme získat různými metodami, které závisí na tom, jaké jsou hodnoty v souboru. "*Konvenční metody jednoduché imputace, tj. jednorázového doplnění chybějícího údaje hodnotou, kterou lze v dané souvislosti považovat za vhodnou či rozumnou, jsou založeny na použití průměru, deterministického modelu, případně stochastického regresního modelu*" (3, str.70).

Kromě průměru můžeme nahrazovat i dalšími mírami polohy jako jsou medián, minimum, maximum atd. U těchto měr poloh ale může docházet ke zkreslení (pokud neplatí MCAR) a podhodnocení variability a kovariance. Chybějící hodnoty lze také nahradit pomocí hodnot

vypočítaných regresním modelem, který je vytvořen z jednotek, kterým hodnoty nechybí. Pomocí metody nejmenších čtverců jsou tak chybějící hodnoty odhadovány jako podmíněné průměry. Podmínkou je však dodržení předpokladů MCAR (3).

Lze také využít metody hot-deck, kdy je chybějící hodnota nahrazena hodnotou již zjištěnou u jiného pozorování, které se tomu nahrazovanému nějakým způsobem podobá (8). Tento postup je vhodný, pokud množství chybějících hodnot není příliš velké (9). Podobná je metoda cold-deck, u které také k nahrazení použijeme hodnoty podobných jednotek, ale podobné jednotky hledáme u šetření již provedených (např. šetření z předchozích období či podobných šetření) (10).

Složitější metodou řešení položkové non-response je vícenásobná imputace. Tento postup má 3 fáze. V první fázi imputujeme chybějící hodnoty, pomocí náhodně vygenerovaných dat z rozdělení, která vzniknou odhadem z dostupných pozorování (5). Toto provedeme m-krát, kdy $m > 1$. Jelikož jde o náhodně vygenerovaná data, získáme tak m-kompletních datasetů, které jsou identické až na imputované hodnoty (6, 11). Druhou fází je analýza doplněného souboru, kde každý dataset vede k jiným parametrům populace (6, 11). Ve třetí fázi se zprůměrováním těchto parametrů získá výsledný průměr a odchylky nalezených odhadů (3, 5). Toto kombinování výsledků se v angličtině nazývá pooling – viz obrázek 1.



Obrázek 1 Vícenásobná imputace,zdroj: (12)

3. Praktická část

V této části práce jsou aplikovány předešle vysvětlené metody vypořádání se s non-response. Metody jsou předvedeny na výukovém souboru Mzdy, který byl získán z Datového archivu Sociologického ústavu AV ČR. V souboru jsou údaje 1134 zaměstnanců s proměnnými Mzda a Vzdělání. Mzda se pohybuje v rozmezí od 6 059 Kč až do 48 487 Kč, s průměrnou hodnotou 20 274 Kč. Proměnná vzdělání nabývá následujících hodnot: 1 – pro základní vzdělání, 2 - pro středoškolské vzdělání bez maturity, 3 – pro středoškolské vzdělání s maturitou a 4 – pro vysokoškolské vzdělání. Celý soubor je kompletní, non-response pro účely této práce tak byla simulována. Jednotlivé příklady ošetření non-response jsou provedeny na ilustračních příkladech, aby byly představené metody přehledně znázorněny. Všechny aplikace jsou provedeny v programu **R** (13).

3.1. Jednotková non-response

Pro jednotkovou non-response byla zvolena míra non-response 5 %, šlo tedy o 57 pozorování. Chybějící hodnoty byly náhodně vybrány, tudíž se jedná o MCAR. Ošetření bylo provedeno dvěma způsoby – vážením a vážením s došetřením (došetřeno 40 z původních 57 non-respondentů). První vážení bylo realizováno s využitím vedlejší informace o zastoupení v populaci, která byla získána z úplného souboru bez chybějících dat. Váhy byly tedy spočteny jako četnost skupiny v populaci (v úplném souboru) vydělená četností skupiny ve výběru (soubor s non-response). Jednotlivé váhy dle vzdělání jsou k dispozici v tabulce 1. Výsledná průměrná mzda po převážení byla 21 309 Kč. Bez převážení byla průměrná mzda 20 197 Kč. V našem ilustračním příkladě tedy převážení nezlepšilo odhad populační hodnoty. Sledujeme tedy, že převážení navýšilo odhad průměrné mzdy. To je dáno tím, že převážení nám poskytuje lepší odhad z hlediska struktury, ale snižuje jeho přesnost (5).

Vzdělání	Počet hodnot v původním souboru	Počet hodnot v upraveném souboru	Váhy
1	125	118	1.059322
2	442	422	1.047393
3	408	389	1.048843
4	159	148	1.074324

Tab. 1 Převážení souboru, zpracování: vlastní

Pro došetření byla vybrána míra non-response mezi původními non-respondenty 30 %. Z chybějících 57 hodnot tak bylo dovyšetřeno 40. Vážený průměr ze skupin respondentů a non-respondentů byl spočítán následovně:

$$\frac{1077}{1117} \cdot 20\,196{,}48 + \frac{57}{1117} \cdot 21\,675{,}20 = 20\,579\ K\check{c}.$$

Srovnáme-li hodnotu průměrné mzdy při 57 chybějících jednotkách s hodnotou průměrné mzdy po realizaci došetření, dojdeme k závěru, že tato metoda v našem ilustračním příkladě opět nepřinesla zlepšení odhadu.

3.2. Položková non-response

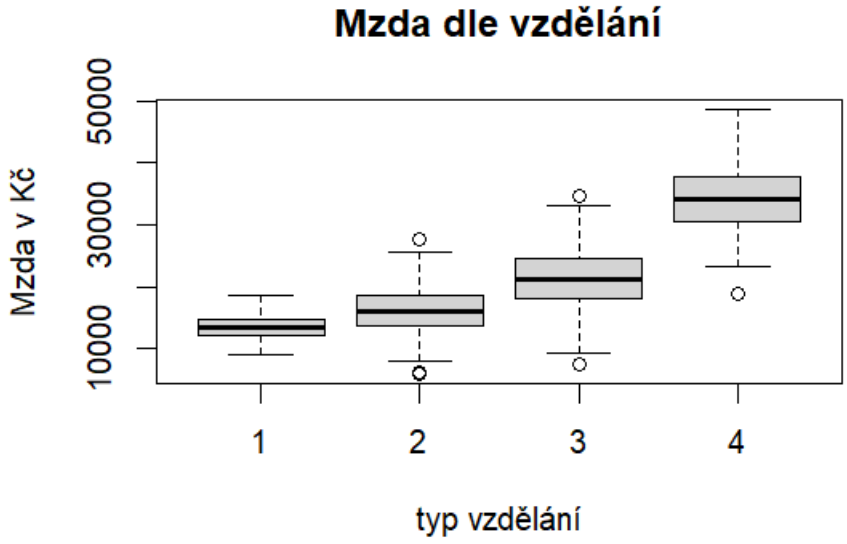
Procentuální míra položkové non-response byla stanovena na 10 %. Chybějících pozorování u proměnné Mzdy bylo tedy 113 a tato pozorování byla opět náhodně zvolena, aby se jednalo o MCAR. Ošetření bylo provedeno pomocí nahrazení průměrem, mediánem a vícenásobnou imputací. Nahrazení regresí nebylo možné vzhledem ke kategoriálnímu charakteru dat. Byl tedy použit lineární regresní model s dummy proměnnými dle vzdělání a pomocí tohoto modelu provedena imputace průměrů dle vzdělání. Jednoduchá imputace byla zajištěna **R** balíčky *simputation* a *naniar* (14, 15). Použití vícenásobné imputace umožnil balíček *mice* (16). Tento balíček náhodně generuje data pomocí iterativní série prediktivních modelů (17).

	průměr	**medián**	**rozptyl**	**odchylka**
Původní data	20 275	18 512	55253263	7 433
Nahrazeno průměrem	20 230	19 417	50233943	7 088
Nahrazeno mediánem	20 057	18 486	50507174	7 107
Nahrazení skupinovými průměry	20 245	18 526	53619642	7 323
Vícenásobná imputace	20 233	18 526	55814965	7 471
Nahrazení skupinovými mediány	20 231	18 525	53692254	7 328

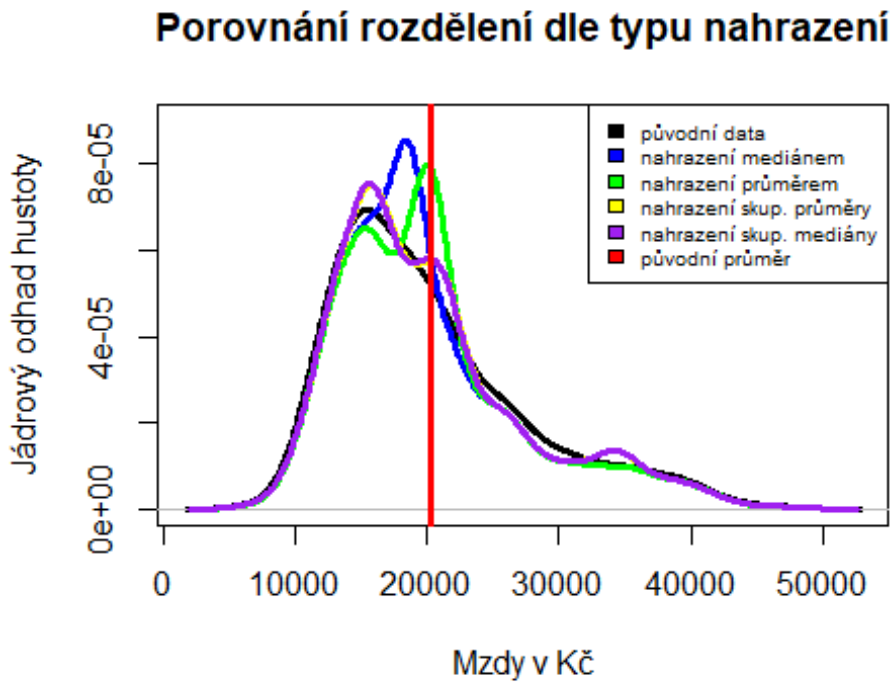
Tab. 2 Imputace, zpracování:

V tabulce č. 2 je souhrn výsledků jednotlivých imputací, a to konktrétně průměr, medián, rozptyl a směrodatná odchylka pro názorné zobrazení rozdílu mezi jednotlivými imputovanými datasety. Pro tento ilustrační příklad nalézáme hodnoty nejblíže těm populačním u složitějších metod jako je nahrazení skupinovým průměrem a vícenásobná imputace. Nahrazení průměrem má podhodnocenou variabilitu a pro odhad těchto mezd, jejichž rozdělení připomíná logaritmicko-normální (viz graf 2), se průměr nejeví jako příliš vhodný. Nejslabších výsledků,

pro tento ukázkový soubor, dosáhlo nahrazení mediánem, a to z důvodu velmi odlišných mediánů v jednotlivých skupinách dle vzdělání (viz graf 1).

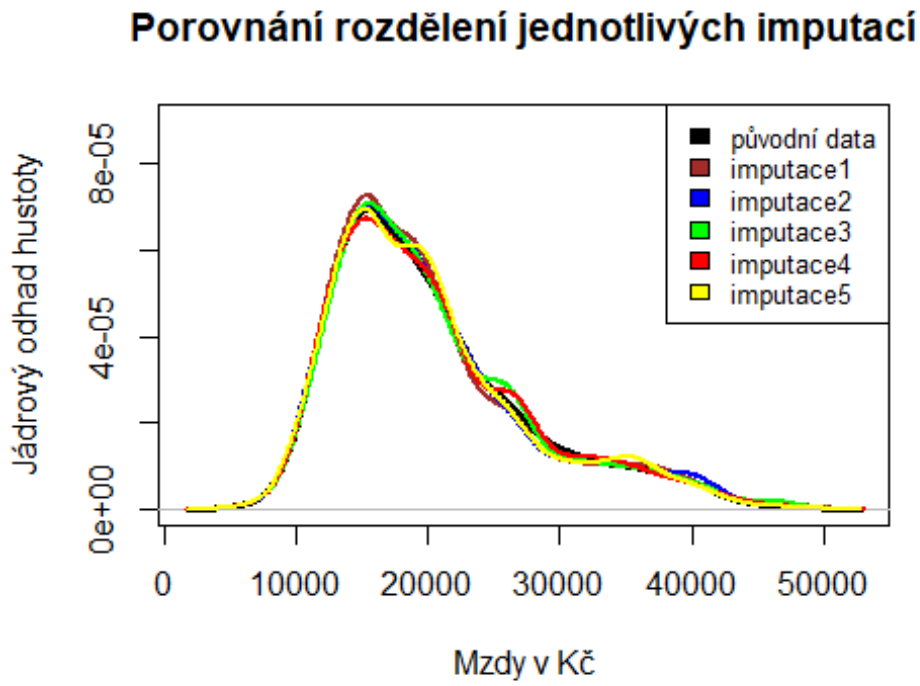


Graf 1 Mzda dle vzdělání, krabičkový graf, zdroj: vlastní



Graf 2 Porovnání rozdělení dle typu nahrazení, zpracování: vlastní

V grafu č. 2 je vidět jádrový odhad hustoty jednotlivých imputací. V tomto ilustračním příkladě zaznamenalo nahrazení mediánem a průměrem větší výkyvy oproti původním hodnotám. Naopak lepších výsledků bylo dosaženo při využití skupinových průměrů a mediánů.



Graf 3 Porovnání rozdělení jednotlivých imputací, zpracování: vlastní

Graf 3 znázorňuje jádrový odhad hustoty jednotlivých imputací z vícenásobné imputace. Jelikož tato metoda vytváří *m* samostatných datasetů (v tomto případě *m*=5), bylo nutno dát jednotlivá rozdělení do samostatného grafu. Lze vidět, že jednotlivé imputace poměrně přesně kopírují průběh původního rozdělení.

Závěr

V seminární práci byly představeny původy vzniku non-response a celkově byla vysvětlena problematika týkající se non-response. Rozlišujeme dva druhy non-response, a to jednotkovou a položkovou. Jednotková non-response byla řešena pomocí vážení a položková pomocí imputace. Ukázalo se, že pro daný ukázkový soubor má při vážení přesnější výsledky metoda s použitím populačních hodnot oproti metodě s došetřením non-response. Při imputaci byly srovnávány metody nahrazení průměrem, mediánem, skupinovými mediány a průměry, a vícenásobné imputace. Nejlepších odhadů populačních hodnot bylo dosaženo při vícenásobné imputaci a při nahrazení skupinovými průměry. Nejslabších naopak při nahrazení mediánem. Představeny byly pouze nejběžnější metody řešení non-response, k dispozici je však celá škála jiných metod, které lze aplikovat také.

Zdroje

(1) NON-RESPONSE. IN: CROS PORTAL: COLLABORATION IN RESEARCH AND METHODOLOGY FOR OFFICIAL STATISTICS [ONLINE]. 8 MAY, 2019 [CIT. 2022-12-04]. DOSTUPNÉ Z: HTTPS://EC.EUROPA.EU/EUROSTAT/CROS/CONTENT/NON-RESPONSE_EN

(2) HANDLING ITEM NONRESPONSE IN SURVEYS. IN: THE STATISTICAL SOCIETY OF CANADA [ONLINE]. 2008 [CIT. 2022-12-04]. DOSTUPNÉ Z: HTTPS://SSC.CA/EN/CASE-STUDY/HANDLING-ITEM-NONRESPONSE-SURVEYS

(3) PECÁKOVÁ, I. (2014). PROBLÉM CHYBĚJÍCÍCH DAT V DOTAZNÍKOVÝCH ŠETŘENÍCH. ACTA OECONOMICA PRAGENSIA, 22(6), 66-78

(4) RESPONSE AND NONRESPONSE. IN: STATISTICS CANADA [ONLINE]. 2015 [CIT. 2022-12-04]. DOSTUPNÉ Z: HTTPS://WWW150.STATCAN.GC.CA/N1/PUB/12-539-X/2009001/RESPONSE-REPONSE-ENG.HTM

(5) MALÁ, I. (29.11.2022). TEORIE VÝBĚROVÝCH ŠETŘENÍ (4ST414) - NONRESPONSE. KSTP FIS VŠE.

(6) VAN BUUREN, S. (2018). FLEXIBLE IMPUTATION OF MISSING DATA. CRC PRESS

(7) THOMPSON, S. K. (2012). SAMPLING (VOL. 755). JOHN WILEY & SONS.

(8) ANDRIDGE RR, LITTLE RJ. A REVIEW OF HOT DECK IMPUTATION FOR SURVEY NON-RESPONSE. INT STAT REV. 2010 APR;78(1):40-64. HTTPS://DOI.ORG/10.1111/J.1751-5823.2010.00103.X

(9) NÁROŽNÁ, M. (2013). IMPUTACE CHYBĚJÍCÍCH HODNOT V ROZSÁHLÝCH DATOVÝCH SOUBORECH. OLOMOUC

(10) SHAO, J. (2000). COLD DECK AND RATIO IMPUTATION. SURVEY METHODOLOGY, 26(1), 79-86

(11) ROBOTKOVÁ, A. (2011). METODY ANALÝZY CHYBĚJÍCÍCH ÚDAJŮ VE STATISTICE. MASARYKOVA UNIVERZITA, DIPLOMOVÁ PRÁCE.

(12) FIGURE 5.1: SCHEME OF MAIN STEPS IN MULTIPLE IMPUTATION. IN: FLEXIBLE IMPUTATION OF MISSING DATA: STEF VAN BUUREN [ONLINE]. 2018 [CIT. 2022-12-04]. DOSTUPNÉ Z: HTTPS://STEFVANBUUREN.NAME/FIMD/WORKFLOW.HTML

(13) R CORE TEAM (2021). R: A LANGUAGE AND ENVIRONMENT FOR STATISTICAL COMPUTING. R FOUNDATION FOR STATISTICAL COMPUTING, VIENNA, AUSTRIA. URL HTTPS://WWW.R-PROJECT.ORG/.

(14) VAN DER LOO M (2022). _SIMPUTATION: SIMPLE IMPUTATION_. R PACKAGE VERSION 0.2.8, HTTPS://CRAN.R-PROJECT.ORG/PACKAGE=SIMPUTATION

(15) TIERNEY N, COOK D, MCBAIN M, FAY C (2021). _NANIAR: DATA STRUCTURES, SUMMARIES, AND VISUALISATIONS FOR. MISSING DATA_. R PACKAGE VERSION 0.6.1, HTTPS://CRAN.R-PROJECT.ORG/PACKAGE=NANIAR

(16) STEF VAN BUUREN, KARIN GROOTHUIS-OUDSHOORN (2011). MICE: MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS IN R. JOURNAL OF STATISTICAL SOFTWARE, 45(3), 1-67. HTTPS://DOI.ORG/10.18637/JSS.V045.I03

(17) WILSON, SAM. THE MICE ALGORITHM [ONLINE]. 2021 [CIT. 2022-12-04]. DOSTUPNÉ Z: HTTPS://CRAN.R-PROJECT.ORG/WEB/PACKAGES/MICERANGER/VIGNETTES/MICEALGORITHM.HTML