PROBLÉM CHYBĚJÍCÍCH DAT V DOTAZNÍKOVÝCH ŠETŘENÍCH

Iva Pecáková*

Úvod

Sebelepšími metodami realizovaná statistická analýza nemůže poskytnout hodnotné výsledky, je-li založena na nekvalitních datech. Jednou z okolností ovlivňujících kvalitu dat je výskyt chybějících údajů. S tímto problémem se lze setkat prakticky u jakéhokoliv reálného datového souboru, ve značné míře pak především u jevů týkajících se populací a zjišťovaných dotazováním vybraných osob. Problematice chybějících dat je věnována v posledním desetiletí značná pozornost a lze říci, že podíl příspěvků z této oblasti dramaticky roste. Cílem článku je poukázat na skutečnost, že běžně používaný postup vynechávání jednotek s chybějícími údaji z analýzy je riskantní a že i pro oblast dotazníkových šetření existují jiné možnosti.

Teoretické základy moderní statistické analýzy datových souborů s chybějícími údaji položil Rubin (1976). Zaměřil se na mechanismus vzniku chybějících dat a prakticky všechny pozdější příspěvky v této oblasti vycházejí z jím zavedené terminologie a symboliky. Přístupy k souborům s chybějícími daty jsou např. v Little a Rubin, 2002 shrnuty do čtyř následujících skupin:

- vynechání jednotek s chybějícími daty (buď všech jednotek s jakýmkoliv chybějícím údajem, nebo jednotek, u nichž jsou identifikovány chybějící údaje pro konkrétní dvojice proměnných);
- vynechání jednotek s chybějícími daty a následně redukce nepříznivých dopadů jejich eliminace aplikací vah;
- doplnění chybějících údajů, buď jednorázové, nebo opakované s cílem využít získané statistiky (jednorázová a vícenásobná *imputace* dat);
- doplňování údajů založené na modelování pořízených dat.

Chybějící údaje mají různé příčiny. Za relativně nejjednodušší z hlediska řešení lze považovat jejich výskyt u plánovaných experimentů, kdy pro známou kombinaci hodnot proměnných považovaných za vysvětlující chybí z nějakého důvodu hodnota veličiny vysvětlované, kterou měl poskytnout právě realizovaný experiment. Ve výběrových šetřeních uskutečněných dotazováním osob je charakter chybějících údajů poněkud jiný. Mohou být důsledkem nezastižení dotazované jednotky či odmítnutím její účasti

* Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky, katedra statistiky a pravděpodobnosti (e-mail: iva.pecakova@vse.cz).

v šetření (*jednotková nonresponse)*, ale mohou vzniknout také u jednotlivých otázek v průběhu zjišťování (*položková nonresponse*). Příčiny neposkytnutí relevantního údaje ze strany respondenta jsou v zásadě tři: (1) neporozumění otázce, (2) neochota či (3) neschopnost odpovědět na ni. Na straně zjišťovatele pak může jít o administrativní či technickou chybu. Datové soubory z dotazníkových šetření často prostřednictvím použitého kódování příčiny vzniku nezjištěných údajů rozlišují, což může na druhé straně být samo zajímavou informací pro analýzu.

Chybějící údaje v souboru představují problém zejména pro metody vícerozměrné statistiky, zvláště pokud u různých jednotek chybí hodnoty různých veličin, byť v relativně malém procentu. V souhrnu to může znamenat jen relativně malý soubor jednotek s kompletními údaji. Běžná praxe ve výběrových šetřeních je taková, že jednotky, u nichž jsou chybějící hodnoty zjišťovaných proměnných zaznamenány, jsou z analýzy vynechány. Vynechání jednotek vede v první řadě ke snížení rozsahu souboru se všemi nežádoucími důsledky (snížení přesnosti odhadů a síly realizovaných testů). Může však vést až ke zkresleným výsledkům, která neposkytují validní základ pro odpovídající zobecňování na příslušnou populaci.

1. Mechanismus vzniku chybějících dat

Samotné hodnoty (kategorie) proměnných, které jsou v datové matici **X** (typu $n \times p$; n je počet jednotek v souboru a p je počet proměnných) vynechávány, jejich velikost či typ, jsou totiž jen jedním aspektem problému chybějících dat. Pro nalezení adekvátního postupu je totiž podstatný především *mechanismus* vzniku chybějících údajů, kterým se pro účely analýzy rozumí jejich vztah k hodnotám dalších proměnných v datové matici. Vysvětlení vzniku chybějících údajů a argumenty pro nakládání s datovými soubory chybějící údaje obsahujícími je třeba hledat už v procesu sběru dat. Pomoci mohou důkladné znalosti předmětné oblasti zkoumání, ale i tak lze detailní představu o mechanismu vzniku chybějících dat sotva očekávat.

To, zda údaj chybí či nikoliv, lze pro analýzu mechanismu vzniku chybějících údajů indikovat jedničkou (zastupující chybějící údaje) a nulou (zastupující údaje zjištěné). Výsledkem takového záznamu je další matice typu $n \times p$, označme ji **M**.

Podle Rubinovy klasifikace podle mechanismu vzniku (např. Little a Rubin, 2002) mohou pak být chybějící údaje:

a) zcela náhodné (*missing completely at random* − MCAR), kdy rozdělení **M** nezávisí na **X**, neboli pravděpodobnost, že údaj bude u jednotky chybět, nezávisí na veličině samé ani na jakékoliv jiné veličině obsažené v matici **X** (například chybějící údaje o příjmu nezávisejí na tom, zda jde o příjmy malé či velké, příjmy mužů či žen atd.). Nejvíce žádoucí případ nezpůsobuje zkreslení prováděných odhadů;

b) náhodné (*missing at random* − MAR), kdy rozdělení **M** závisí pouze na *zjištěných* hodnotách **X** (například chybějí spíše údaje o příjmu mužů, nezávisle na tom, zda jde o příjmy menší či větší). Žádoucí případ za určitých podmínek, MCAR je zřejmě zvláštním případem MAR;

c) nenáhodné (*not missing at random* − NMAR), kdy rozdělení **M** závisí na *chybějících* hodnotách **X**. Například chybějí spíše údaje o příjmu mužů, které jsou spíše vyšší; nejproblematičtější případ.

Lze ověřit mechanismus vzniku chybějících údajů? V případě MCAR se nabízejí i některé obecně známé postupy, jako například ověření shody středních hodnot některé veličiny ve skupinách vzniklých tříděním podle toho, zda u jiné veličiny je údaj k dispozici či chybí. Na této myšlence je v zásadě založena i svého času poměrně populární Cohenova metoda (Allison, 2001), kdy v regresním modelu každou proměnnou s chybějícími údaji doprovází indikátor rozlišující disponibilní a chybějící údaje. Významné regresní parametry u těchto indikátorů pak upozorňují na pravděpodobné porušení MCAR. Vzhledem k nadhodnocení reziduálního rozptylu však tato metoda nevede vždy k odpovídajícím závěrům (Pigott, 2001).

Zda chybějící údaje veličiny závisejí na jejich velikosti, však nelze ověřit. Z dat nijak nevyplývá, zda skutečnost, že údaj není k dispozici, souvisí s jeho velikostí či nikoliv. Allison, 2009 doporučuje zahrnout do datové matice co nejvíce proměnných, které by mohly přispět k vysvětlení veličin s potenciálně chybějícími hodnotami, a ponechat tak jen malý prostor pro zpochybnění předpokladu dat MAR. Nenáhodný charakter chybějících údajů neumožňuje mechanismus jejich vzniku jednoduše ignorovat; vyplývá z toho, že mnohé konvenční metody používané při práci s neúplnými daty mají vážné nedostatky a nemohou vést k uspokojivým výsledkům.

2. Vynechávání chybějících údajů

Jak jsme již konstatovali, běžně se při zjištění chybějících údajů v datech jednotky s neúplnými informacemi z analýzy vynechávají; například statistický programový systém SPSS (IBM SPSS Statistics) tento postup nabízí jako *Listwise Deletion*. Je to nejstarší, ale i v současnosti zcela převládající přístup. Vychází ovšem z předpokladu, že chybějící údaje jsou zcela náhodné (MCAR). Odstranění některých jednotek ze souboru představuje pak vlastně jen realizaci náhodného výběru menšího rozsahu. Jenom v takovém případě lze získat výsledky, jež nejsou pominutím části souboru ovlivněny, neboť populační charakteristiky veličin pro skupinu reprezentovanou kompletními a nekompletními daty se neliší. Postup lze proto doporučit jen v situaci, kdy výskyt chybějících údajů je jen malý, nebo pokud postupy ověřující charakter chybějících údajů MCAR nezpochybnily.

Jsou-li takto vynechanými jednotkami například většinou muži, neboť neudali příjem (bez ohledu na jeho výši − mechanismus vzniku chybějících údajů je MAR; muži mají vyšší úroveň příjmů než ženy), nebo neudávají-li příjem většinou muži s nejvyššími příjmy (NMAR), dojde k systematickému podhodnocení úrovně příjmů. Jsou-li v takovém případě k dispozici informace o složení populace z hlediska proměnných, jejichž údaje datový soubor obsahuje (zde kromě pohlaví například vzdělání či věková skupina), používají se pro zmírnění zkreslení vyvolaného chybějícími jednotkami váhy.

Tento přístup je v praxi terénních průzkumů poměrně oblíbený. Vychází z jednoduché úvahy (Cochran, 1977): Na základě pravděpodobnostního výběru

z populace (o rozsahu N), již tvoří roztříděním podle K kategorií nějaké proměnné K skupin o velikostech N_k, $k = 1, 2 \ldots, K$, je analogicky roztříděn vzorek o rozsahu n; velikost skupin ve vzorku je n_k, $k = 1, 2 \ldots, K$. Každou jednotku s k-tou kategorií této proměnné ve vzorku lze pak považovat za reprezentanta N_k/n_k takových jednotek v populaci, což lze vyjádřit jako její váhu, například

$$w_k = \frac{N_k}{n_k} \Big/ \sum_k \frac{N_k}{n_k} = \pi_k^{-1} / \sum_k \pi_k^{-1} . \tag{1}$$

Součet vah w_k se rovná jedné; alternativně po vynásobení výrazu (1) velikostí vzorku se součet vah rovná n.

Uvedený přístup lze snadno modifikovat právě pro případ výskytu chybějících údajů ve skupinách (Little a Rubin, 2002). Označíme-li počet disponibilních údajů ve skupině r_k, pak jejich podíl na rozsahu skupiny je $f_k = r_k/n_k$ a váhu (1) lze upravit na

$$w_k = f_k \pi_k^{-1} / \sum_k f_k \pi_k^{-1} = \frac{N_k}{r_k} \Big/ \sum_k \frac{N_k}{r_k} . \tag{2}$$

Stanovení vah z hlediska ne pouze jednorozměrné, ale vícerozměrné struktury populace vyžaduje znalost této struktury; avšak informace o složení populace mají jen vzácně vícerozměrný charakter. Odhad této vícerozměrné struktury na základě jednorozměrných marginálních rozdělení (*rating*) lze realizovat například relativně jednoduchým iteračním postupem IPF (např. Fienberg, 1970; Pecáková, 2011), kdy je v podstatě libovolná výchozí struktura v cyklech iteračních kroků postupně přepočítávána tak, aby odpovídala marginálním rozdělením jednotlivých uvažovaných veličin.

Následný výpočet vážených odhadů je relativně jednoduchý, na rozdíl od komplikovaného stanovení směrodatných chyb takových odhadů, které neusnadňují ani softwarové aplikace vážení nabízející. Příslušné procedury totiž používají váhy způsobem, který vede k podhodnocování směrodatných chyb (Little a Rubin, 2002), a postup lze tedy doporučit pouze v případě, že otázka přesnosti pořízených odhadů je podružná.

Jako alternativa k vynechávání kompletních jednotek s chybějícími údaji se nabízí analýza všech disponibilních dat, tj. selekce jednotek až v okamžiku, kdy je prováděn výpočet statistik pro dvojice veličin a u některých jednotek jsou identifikovány chybějící údaje (např. v SPSS volba *Pairwise Deletion*). Skutečnost, že statistiky jsou v tomto případě počítány a zobecnění následně realizována ze souborů o *různém* rozsahu, však působí nejrůznější problémy, především opět při stanovení směrodatných chyb. Postup také může vést (zejména v případě silných závislostí mezi veličinami) až k nepřijatelným hodnotám korelačních koeficientů (mimo obor možných hodnot), resp. v případě vícerozměrných úloh k nepřijatelným kovariančním maticím. Protože nelze předvídat, kdy analýza disponibilních dat přinese neadekvátní výsledky, nelze ji obecně doporučit.

3. Doplňování chybějících údajů

Výše uvedené postupy různým způsobem jednotky s chybějícími informacemi *eliminují*, což mimo jiné nežádoucím způsobem ovlivňuje (zmenšuje) velikost vzorku. V tomto směru je výhodnější pokusit se chybějící údaje u jednotek *doplnit* (imputovat). 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.

Použití průměru zjištěných hodnot příslušné veličiny pro doplnění jejích hodnot chybějících může vést ke zkresleným odhadům (neplatí-li MCAR), v každém případě pak vede k podhodnocení jejich variability: do výpočtu jsou zahrnuty jednotky, které (zdánlivě) k variabilitě veličiny nijak nepřispívají. Pokud jde o zkoumání vztahů mezi veličinami, vede použití průměrů pro doplnění chybějících údajů veličin ze stejného důvodu k podhodnocení kovariancí. Zejména v případě vyššího podílu chybějících údajů tak opět nelze postup doporučit.

Variabilita veličiny s chybějícími údaji (např. X_1) může souviset s jinými proměnnými z datové matice, jejichž hodnoty jsou známy (X_2, X_3 …, X_q; $q \leq p$). Chybějící údaje pak lze odvozovat právě od těchto známých hodnot na základě jednoduché lineární regresní funkce

$$\hat{x}_{i1} = b_0 + \sum_{j=2}^{q} b_j x_{ij} \,, \tag{3}$$

kde $\hat{x}_{i1}$ představuje odhadnutou hodnotu veličiny X_1 pro i-tou jednotku se známými hodnotami proměnných X_2, X_3 …, X_q; $q \leq p$, kterou bude chybějící údaj nahrazen. Parametry funkce (3) jsou odhadnuty metodou nejmenších čtverců na základě těch řádků v datové matici, které jsou kompletní. Kategoriální proměnné mohou být do rovnice zahrnuty prostřednictvím indikátorů, např. typu *dummy*.

Na základě regresní funkce jsou chybějící údaje odhadovány jako podmíněné průměry. Postup je bezproblémový, použijeme-li ho pro odhad například náhodně chybějících výdajů na základě známých příjmů (MCAR). Jedná-li se však například o odhad výdajů náhodně chybějících spíše u nižších příjmů (MAR), předpoklad lineárního charakteru vztahu, jak známo, nemusí být oprávněný, představuje-li odhad chybějících hodnot extrapolaci mimo rozpětí disponibilních údajů. A dále, vzhledem k tomu, že tento typ imputace nezohledňuje variabilitu jednotlivých pozorování kolem podmíněného průměru, dochází opět k systematickému podhodnocení směrodatných chyb a až k nesprávným úsudkům o populaci. Zahrnutí stochastické složky z_i do regresního modelu (3)

$$\hat{x}_{i1} = b_0 + \sum_{j=2}^{q} b_j x_{ij} + z_i \,, \tag{4}$$

umožňuje přejít při nahrazování chybějícího údaje od podmíněného průměru k jedné podmíněné hodnotě, náhodně vybrané z normálního rozdělení s variabilitou odhadnutou reziduálním rozptylem plynoucím z regrese X_1 na X_2, X_3 …, X_q; $q \leq p$, založené na kompletních datech (předpokládáme-li normální rozdělení s touto variabilitou

a nulovou střední hodnotou u složky z_i). Posledně uvedený postup, navržený pro jednorozměrnou imputaci chybějících hodnot, je podle Little a Rubin, 2002 z hlediska úsudků o populaci nejvýhodnější, a to nejen v případě, že chybějící údaje jsou MCAR, ale i pro méně omezující předpoklad MAR.

Výtky adresované konvenčním přístupům k chybějícím datům motivují studium alternativních metod řešení této problematiky. Většina z nich je založena na věrohodnosti pořízených dat za předpokladu existence nějakého modelu. Lze říci, že dnes věrohodnostní přístup i v této oblasti analýzy dat převládá jistě i proto, že věrohodnost hraje důležitou roli v bayesovské statistice, jež i v analýze datových souborů s chybějícími údaji nabývá na významu.

Použití metody maximální věrohodnosti vede k odhadům parametrů předpokládaného populačního modelu, nepřináší tedy náhradu za chybějící údaje u jednotlivých proměnných. Maximálně věrohodné odhady parametrů přitom mají, jak známo, žádoucí vlastnosti: jsou konzistentní, asymptoticky vydatné a asymptoticky normální. Existence chybějících dat pro jejich nalezení samozřejmě představuje komplikaci; v závislosti na uvažovaném modelu k tomu účelu slouží tzv. EM algoritmus.

EM algoritmus lze popsat jako iterativní cyklus dvou kroků. První z nich (M) představuje pořízení maximálně věrohodných odhadů parametrů předpokládaného modelu z disponibilních dat. Ve druhém kroku (E) jsou na základě odhadnutých parametrů a disponibilních dat odhadnuty chybějící údaje potřebně pro další provedení kroku M atd. Postupu je věnována v posledních letech velká pozornost, neboť umožňuje dosáhnout uspokojivých výsledků i v případě, že data nejsou MCAR ani MAR − podmínkou je ovšem nalezení odpovídajícího modelu pro mechanismus vzniku chybějících údajů.

Pro spojité proměnné je obvyklým východiskem pro realizaci věrohodnostního postupu předpoklad vícerozměrné normality. Za tohoto předpokladu představují zmíněné dva kroky EM algoritmu

1. stanovení výchozích hodnot výběrových statistik (průměrů, rozptylů, kovariancí) z disponibilních dat;
2. použití těchto statistik k výpočtu parametrů regresních rovnic vyjadřujících závislosti mezi veličinami a odhad chybějících hodnot pro všechny jednotky a proměnné.

V dalším cyklu jsou stanoveny nové hodnoty výběrových statistik ze všech zjištěných i doplněných údajů, nové parametry rovnic atd. a kroky se opakují, dokud změna v odhadovaných parametrech neklesne pod stanovenou mez.

Vraťme se nyní k myšlence nahrazování chybějících údajů a připomeňme výtky, které jsou adresovány metodám jednorázové imputace. S cílem využít její jednoduchost na jedné straně a naopak odstranit nedostatky na straně druhé vznikla metoda imputace vícenásobné (MI). Metoda dosahuje odhadů s vlastnostmi téměř stejnými jako metoda maximální věrohodnosti (obecně nejsou asymptoticky vydatné; Allison, 2009); pokud jde o mechanismus vzniku chybějících dat, předpokladem je alespoň MAR. Výhodou MI je jednoduchost, nevýhodou skutečnost, že imputované hodnoty jsou generovány náhodně, což vede v opakovaných aplikacích k různým výsledkům.

Vícenásobná imputace je obvykle realizována metodou MCMC (Markov Chain Monte Carlo); od algoritmu EM pro vícerozměrné normální rozdělení se postup liší jen zahrnutím náhodné složky do regresní rovnice. Nahrazení chybějících hodnot je přitom realizováno vícekrát, výstupem je pokaždé soubor s jinými doplněnými hodnotami. Každý vytvořený soubor tedy vede k poněkud odlišným odhadům parametrů populace, které jsou následně zprůměrovány. Při určení jejich směrodatných chyb je pak nutno stanovit kromě průměru jejich rozptylů („vnitřní" variability) také jejich rozptyl kolem vypočteného průměrného odhadu vyvolaný opakováním imputace („vnější" variabilitu); detailněji Schafer a Olsen, 1998.

Potřebný počet opakování algoritmu závisí na konkrétní úloze, na rozsahu datového souboru, charakteru proměnných v něm a především na podílu chybějících údajů. Pokud není příliš velký, postačuje i jen malý počet opakování algoritmu (pět) k dosažení relativně vydatných odhadů.

4. Chybějící údaje v souborech z dotazníkových šetření

Předpoklad vícerozměrné normality v datových souborech z terénních průzkumů realizovaných dotazováním není většinou příliš realistický. Měřitelné proměnné jsou v takových datech v menšině, převládají proměnné nominální a ordinální (tedy kategoriální). Modelovány jsou četnosti v *p*-rozměrné kontingenční tabulce a uvažovaná rozdělení jsou tedy nespojitá.

Pro jejich rekapitulaci uvažujme $p = 2$, a tedy dvourozměrnou kontingenční tabulku vzniklou tříděním podle R kategorií veličiny $X_1 = X$ a S kategorií veličiny $X_2 = Y$ (tato symbolika bude nyní výhodnější). V případě X je k dispozici všech n údajů, v případě Y však m hodnot chybí − k dispozici je tedy $n - m = r$ pozorování (Tabulka 1).

Tabulka 1: Kontingenční tabulka – symbolika

Údaje	disponibilní				∑ disp.	chybějící	celkem
	y_1	y_2	…	y_S	r_{i+}	m_i	n_{i+}
x_1	r_{11}	r_{12}	…	r_{1S}	r_{1+}	m_1	n_{1+}
x_2	r_{21}	r_{22}	…	r_{2S}	r_{2+}	m_2	n_{2+}
…	…	…	…	…	…	…	…
x_R	r_{R1}	r_{R2}	…	r_{RS}	r_{R+}	m_R	n_{R+}
r_{+j}	r_{+1}	r_{+2}	…	r_{+S}	r	m	n

x_i kategorie veličiny X,
y_j kategorie veličiny Y,
r_{ij} ($i = 1, 2 \ldots, R$; $j = 1, 2 \ldots, S$) sdružené četnosti získané tříděním disponibilních dat,
m_i četnosti údajů chybějících v jednotlivých řádcích tabulky,
r_{i+} četnosti v řádcích tabulky po odečtení počtu chybějících údajů,
n_{i+} marginální četnosti řádkové,
r_{+j} marginální četnosti sloupcové;

$$\sum_j r_{ij} = r_{i+};\ r_{i+} + m_i = n_{i+};\ \sum_i r_{i+} = r;\ \sum_i m_i = m;\ r + m = n\ .$$

Pro vzorek (pořízený prostým náhodným výběrem) o rozsahu n, který neobsahuje žádné chybějící údaje o obou proměnných, kdy $r_{ij} = n_{ij}$, platí:

- rozdělení každé četnosti n_{ij} v tabulce je binomické s parametry n a π_{ij}, střední hodnotou $n\pi_{ij}$ a rozptylem $n\pi_{ij}(1 - \pi_{ij})$;
- četnosti n_{ij} nejsou vzájemně nezávislé (protože $\Sigma\Sigma n_{ij} = n$), jejich kovariance jsou $-n\pi_{ij}\pi_{i'j'}$;
- rozdělení marginálních četností $n_{i+} = r_{i+} + m_i$ jsou multinomická s parametry n,π_{i+};
- pro $\sum_j n_{ij} = r_{i+}$ jsou podmíněná rozdělení $n_{i1}, n_{i2}, \ldots, n_{iS}$ v jednotlivých řádcích kontingenční tabulky nezávislá multinomická s parametry n_{i+}, $\pi_{j\,|\,i} = \pi_{ij} / \pi_{i+}$;
- rozdělení $R \times S$ četností v tabulce je multinomické s parametry $n,\pi_{11}, \pi_{12}, \ldots, \pi_{RS}$ (např. Jobson, 1992; Agresti, 2002).

Hodláme-li metodou maximální věrohodnosti odhadnout parametry multinomického rozdělení π_{ij}, $i = 1, 2 \ldots, R$ a $j = 1, 2 \ldots, S$ pro nekompletní údaje v Tabulce 1, lze na základě faktorizace věrohodnosti (Little a Rubin, 2002) vyjádřit odhad parametrů poměrně jednoduchým způsobem:

$$\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{j\,|\,i} = \frac{r_{i+} + m_i}{n} \cdot \frac{r_{ij}}{r_{i+}} = \frac{n_{ij} + (r_{ij} / r_{i+})m_i}{n}. \qquad (5)$$

Četnosti jednotek s chybějícími údaji se tak rozdělují do jednotlivých polí v kontingenční tabulce proporčně podle toho, jak se podílí četnost v příslušném poli na rozsahu vzorku bez jednotek s chybějícími údaji.

Popsaný relativně jednoduchý postup lze analogicky rozšířit na větší počet proměnných, a tedy rozměrů kontingenční tabulky, ovšem za podmínky, že chybějící údaje mají monotónní charakter. To znamená, pokud třídíme podle proměnných X_1, X_2, X_3 atd., že chybějící údaj například u proměnné X_2 znamená chybějící údaje také u všech následujících proměnných.

Pro ilustraci uvedeného postupu použijeme data v Tabulce 2. Hodláme zkoumat závislost ochoty zákazníka firmy přejít na novou značku nealkoholického nápoje (alternativní proměnná; ano, ne) na pohlaví a věku (tři věkové skupiny). Tabulka 2 obsahuje simulovaná data (Pecáková, 2011), k nimž byly připojeny další jednotky tak, aby jejich chybějící údaje měly monotónní charakter. Rozsah výběrového souboru je 312 dotázaných osob (předpokládáme prostý náhodný výběr s vracením z rozsáhlé populace).

Pohlaví je u všech dotázaných známo. Chybějící údaje splňují podmínku monotónnosti; u 51 osob chybí pouze odpověď na otázku ohledně značky; 31 osob, které neuvedly věk, neuvedlo ani odpověď ohledně značky.

Žen je celkem 160, z toho 134 uvedlo věk (86 v první, 23 ve druhé a 25 ve třetí věkové skupině). Z 86 žen v první věkové skupině odpovědělo na otázku ohledně nealkoholického nápoje 70 (z toho 45 kladně), ve druhé věkové skupině z 23 odpovědělo

Tabulka 2: Data

	věk I		věk II		věk III		
	žena	muž	žena	muž	žena	muž	
ano	45	28	3	24	3	21	124
ne	25	5	14	25	14	23	106
celkem	70	33	17	49	17	44	230
chybějící odpověď	16	5	6	6	8	10	51
chybějící věk	26 žen			5 mužů			31
celkem							312

Zdroj: Pecáková, 2011

17 (3 kladně) a ve třetí věkové skupině z 25 odpovědělo rovněž 17 (a z toho 3 kladně). Potom odhad parametru pro ženy v první věkové skupině odpovídající ano je:

$$\hat{\pi}_{ž11} = \frac{160}{312} \cdot \frac{86}{134} \cdot \frac{45}{70} = 0,211 .$$

Analogicky pro muže v první věkové skupině parametr odhadneme jako

$$\hat{\pi}_{m11} = \frac{152}{312} \cdot \frac{38}{147} \cdot \frac{28}{33} = 0,107 .$$

Obdobným způsobem pořízené odhady všech parametrů multinomického rozdělení metodou maximální věrohodnosti obsahuje Tabulka 3; výsledné četnosti pak Tabulka 4.

Tabulka 3: Odhadnuté parametry

	věk I		věk II		věk III		celkem
	žena	muž	žena	muž	žena	muž	
ano	0,211	0,107	0,016	0,089	0,017	0,085	0,525
ne	0,118	0,019	0,072	0,093	0,079	0,094	0,475
celkem	0,329	0,126	0,088	0,182	0,096	0,179	1,000

Zdroj: vlastní výpočet

Tabulka 4: Odhadnuté četnosti

	věk I		věk II		věk III		
	žena	muž	žena	muž	žena	muž	celkem
ano	65,9	33,4	5,0	27,8	5,3	26,5	163,9
ne	36,8	5,9	22,5	29,0	24,6	29,3	148,1
celkem	102,7	39,3	27,5	56,8	29,9	55,8	312

Zdroj: vlastní výpočet

V dotazníkových šetřeních monotónní charakter chybějících dat není příliš reálný. Odhad parametrů multinomického rozdělení lze pak realizovat EM algoritmem v iteračních krocích. Nejprve jsou na základě struktury kompletních dat odhadnuty podmíněné pravděpodobnosti příslušnosti nekompletních jednotek k polím v tabulce a jednotky jsou zatříděny. Zvýší se tak původní četnosti a pozmění struktura výchozí kontingenční tabulky. Znovu je proveden odhad podmíněných pravděpodobností a opraveny četnosti atd. Algoritmus se ukončí, pokud další změny již nejsou podstatné.

Pro názornost nyní ke kompletním jednotkám v Tabulce 2 připojíme chybějící údaje tak, aby neměly monotónní charakter, a provedeme opět odhad parametrů, resp. výpočet příslušných četností.

Tabulka 5: Obecný charakter chybějících dat

	věk I		věk II		věk III		celkem	chyb. věk	
	žena	muž	žena	muž	žena	muž	ž + m		
ano	45	28	3	24	3	21	51 + 73	20	6
ne	25	5	14	25	14	23	53 + 53	3	2
celkem	70	33	17	49	17	44	230	23	8
chybějící odpověď	16	5	6	6	8	10	51		
celkem								312	

Zdroj: vlastní výpočty

Například pro ženy v první věkové skupině odpovídající ano dostáváme

$$\hat{n}_{\check{z}11} = 45 + \frac{45}{51} \cdot 20 + \frac{45}{70} \cdot 16 = 72{,}9 \ ,$$

pro muže v první věkové skupině odpovídající ano dostáváme

$$\hat{n}_{m11} = 28 + \frac{28}{73} \cdot 6 + \frac{28}{33} \cdot 5 = 34{,}6$$

atd. Z Tabulky 6 vyplývá, že stačí dva kroky; změny v četnostech jsou velmi malé a pokračovat dále v jejich přepočítávání není nutné.

Tabulka 6: Opravené četnosti

		věk I		věk II		věk III			
krok		žena	muž	žena	muž	žena	muž	ž + m	celkem
I	ano	72,9	34,6	5,2	28,9	5,6	27,5	83,7 + 91,0	174,7
	ne	32,2	5,9	19,7	29,0	21,4	29,1	73,3 + 64,0	137,3
	celkem	105,1	40,5	24,9	57,9	27,0	56,6		312
II	ano	73,5	34,5	5,5	28,9	6,0	27,7	85,0 + 91,1	176,1
	ne	31,2	5,9	19,6	28,9	21,2	29,1	72,0 + 63,9	135,9
	celkem	104,7	40,4	25,1	57,8	27,2	56,8		312

Zdroj: vlastní výpočty

Pokud jde o mechanismus vzniku chybějících údajů, nejspíš nepůjde o MCAR: zdá se, že tendenci neuvádět věk mají spíše ženy. Zvláště starší ženy mají také tendenci nápoj odmítnout. V SPSS *orientačně* provedený Littleho test (Little, 1988) předpoklad MCAR zamítl; proměnné však podmínky použití testu nesplňují. Otázkou je však i neověřitelný předpoklad MAR. Zdá se totiž, že věk chybí spíše u starších žen, které na otázku ohledně značky odpovídají ano. Pokud jsou však chybějící údaje NMAR, je podmínkou použití metody maximální věrohodnosti při řešení jejich výskytu, jak jsme již konstatovali, nejen stanovení pravděpodobnostního modelu pro všechny proměnné v datové matici, ale také modelu pro mechanismus jejich vzniku.

Analýze chybějících údajů se věnují mnohé komerční i volně šiřitelné statistické systémy − z první skupiny uveďme SPSS, SAS, MPLUS či SOLAS, z druhé především R. Systém SPSS, který je v analýze dat z dotazníkových šetření používán patrně nejčastěji, umožňuje v rámci procedury *Multiple Imputation* doplňování chybějících údajů u konkrétních jednotek; pro kategoriální proměnné je přitom použita logistická, nikoliv lineární regrese, a jak z označení procedury vyplývá, lze imputaci dat realizovat jako vícenásobnou.

Tabulka 7: Srovnání výsledků EM algoritmu a vícenásobné imputace

		věk I		věk II		věk III		
metoda		žena	muž	žena	muž	žena	muž	celkem
EM	ano	65,9	33,4	5,0	27,8	5,3	26,5	163,9
	ne	36,8	5,9	22,5	29,0	24,6	29,3	148,1
	celkem	102,7	39,3	27,5	56,8	29,9	55,8	312
MI	ano	66,0	33,6	6,4	27,8	5,4	28	167,2
	ne	37,6	5,8	21,6	28	23	28,8	144,8
	celkem	103,6	39,4	28,0	55,8	28,4	56,8	312

Zdroj: vlastní výpočty

V SPSS (verze 18) jsme realizovali vícenásobnou imputaci (MI; pětinásobné opakování) pro data z Tabulky 2 a provedli třídění jednotek. Tabulka 7 porovnává výsledky aplikace EM algoritmu a vícenásobné imputace. Analogické srovnání pro data z Tabulky 5 obsahuje Tabulka 8.

Tabulka 8: Srovnání výsledků EM algoritmu a vícenásobné imputace

		věk I		věk II		věk III		
metoda		žena	muž	žena	muž	žena	muž	celkem
EM	ano	73,5	34,5	5,5	28,9	6,0	27,7	176,1
	ne	31,2	5,9	19,6	28,9	21,2	29,1	135,9
	celkem	104,7	40,4	25,1	57,8	27,2	56,8	312
MI	ano	71,8	35,2	6,6	30,2	6,6	28,2	178,6
	ne	31,0	6,0	20,2	27,0	20,8	28,4	133,4
	celkem	102,8	41,2	26,8	57,2	27,4	56,6	312

Zdroj: vlastní výpočty

Výsledky se v obou případech od sebe příliš neliší. Rozvržením jednotek s chybějícími údaji do kontingenční tabulky se zvětšil rozsah souboru, který můžeme využít pro analýzu. To je pozitivní důsledek pro realizaci induktivních úsudků, které jsou v analýze kategoriálních dat často založeny na asymptotických rozděleních.

Závěr

Chybějící data jsou natolik běžnou součástí datových souborů, že jejich výskyt bývá podceňován. Obvykle jsou jednotky s chybějícími údaji vypouštěny z analýzy bez ohledu na mechanismus, jaký chybějící data generuje. Jsou-li k dispozici informace o populaci, je nutno do určité míry korigovat nepříznivé důsledky takového postupu konstrukcí vah. Stanovení směrodatných chyb následně realizovaných odhadů je však problematické.

Metody doplňování chybějících údajů jsou pochopitelně závislé na charakteru proměnných v datové matici. Dotazníková šetření generují převážně kategoriální data. Průměr nebo regresní odhady na základě lineární regresní funkce tak nelze pro náhradu chybějících kategoriálních dat použít; jako alternativa se nabízí logistická regrese. Obecně však jednorázová aplikace regrese připadá v úvahu pouze v případě chybějících údajů MCAR a vede k podhodnocování směrodatných chyb.

Parametry modelu pro nekompletní data tříděná do kontingenční tabulky lze odhadnout metodou maximální věrohodnosti, pokud jsou chybějící údaje MCAR či MAR a pokud mají monotónní charakter; druhý předpoklad není příliš realistický, obvykle je nutno odhad realizovat s využitím EM algoritmu. Postup je použitelný i v případě, že chybějící údaje jsou NMAR, je však nutno najít model pro mechanismus jejich vzniku.

Výsledky srovnatelné s aplikací EM algoritmu (a s podobnými vlastnostmi odhadů) poskytuje vícenásobná imputace. S využitím vhodného softwaru lze dosáhnout relativně vysokého počtu opakování, chybějící údaje doplnit a zvětšit tak rozsah souboru, který můžeme využít pro analýzu. Zejména v případě vícerozměrného třídění je to velmi žádoucí.

Způsob využití věrohodnosti a zejména metod vícenásobné imputace při modelování chybějících dat má blízko k bayesovské statistice. Nutno konstatovat, že prameny, které jsou dnes v této oblasti považovány za klíčové (např. Little a Rubin, 2002), se v mnohém o bayesovský přístup opírají.

Literatura

AGRESTI, A. 2002. *Categorical Data Analysis*. 2. vyd. New Jersey : Wiley, 2002. ISBN 0-471-36093-7.

ALLISON, P. D. 2001. *Missing Data*. SAGE Publications, 2001. ISBN 978-14-1298-507-9.

ALLISON, P. D. 2009. Missing Data. In MILLSAP, R. – MAYDEU-OLIVARES, A. (ed.). *Sage Handbook of Quantitative Methods in Psychology*. SAGE Publications, 2009.

COCHRAN, W. G. 1997. *Sampling Techniques*. 3. vyd. New Jersey : Wiley, 1977. ISBN 978-0-471-16240-7.

FIENBERG, E. S. 1970. An Alternative Procedure for Estimation in Contingency Table. *The Annals of Mathematical Statistics* 1970, roč. 41, 907–917.

JOBSON, J. D. 1992. *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariuate Analysis*. New York : Springer-Verlag, 1992. ISBN 0-387-97804-6.

LITTLE, R. 1988. A test of MCAR for Multivariate Data With Missing Values. *Journal of the American Statistical Association* 1988, roč. 83, 1198–1202.

LITTLE, R.; RUBIN, D. 2002. *Statistical Analysis with Missing Data*. New Jersey : Wiley, 2002. ISBN 0-471-18386-5.

PECÁKOVÁ, I. 2011. *Statistika v terénních průzkumech*. 2. vyd. Praha : Professional Publishing 2011. ISBN 978-80-7431-039-3.

PIGOTT, T. 2001. A Review of Methods for Missing Data. *Educational Research and Evaluation* 2001, roč. 7, 353–383.

RUBIN, D. B. 1976. Inference and missing data. *Biometrika* 1976, roč. 63, 581-590.

SCHAFER, J. L.; OLSEN, M. K. 2014. *Multiple Imputation for Missing-data problems: a Data Analyst's Perspective* [online, cit. 2014-09-10]. http://webdocs.cs.ualberta.ca/~ajit/ impute.pdf

SPSS, 2007. *SPSS – Missing Value Analysis*. Chicago : SPSS Inc., 2007.

PROBLEM OF MISSING DATA IN QUESTIONNAIRE SURVEYS

Abstract: Almost any data set can be encountered to the problem of missing data; it is well known in the phenomena relating to people populations and researched in sample surveys. In recent decades, the issue of missing data received considerable attention, because the simple omission of units, for which data are lacking, from the analysis may lead to erroneous conclusions. The approach that accepts the existence of missing data through the modification of the probabilities of units selection with probabilities of obtaining data on them, leads to the construction and use of the weights. Different solution lies in filling in missing data. Using the arithmetic mean or a regression function, recommended for this purpose before, leads at the relevant variables at least to an underestimation of variability; furthermore, it is applicable only for measurable variables. Alternative approaches to missing data are based on the likelihood of collected data assuming some model. Two directions of their development can be distinguished again, estimating population parameters without imputation of missing data on the one hand (EM algorithm) and multiple imputation methods on the other.

Key words: sample surveys, missing data, categorical data, data imputation

JEL Classification: C10, C18, C83