SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

4/2020
ročník/volume 30

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

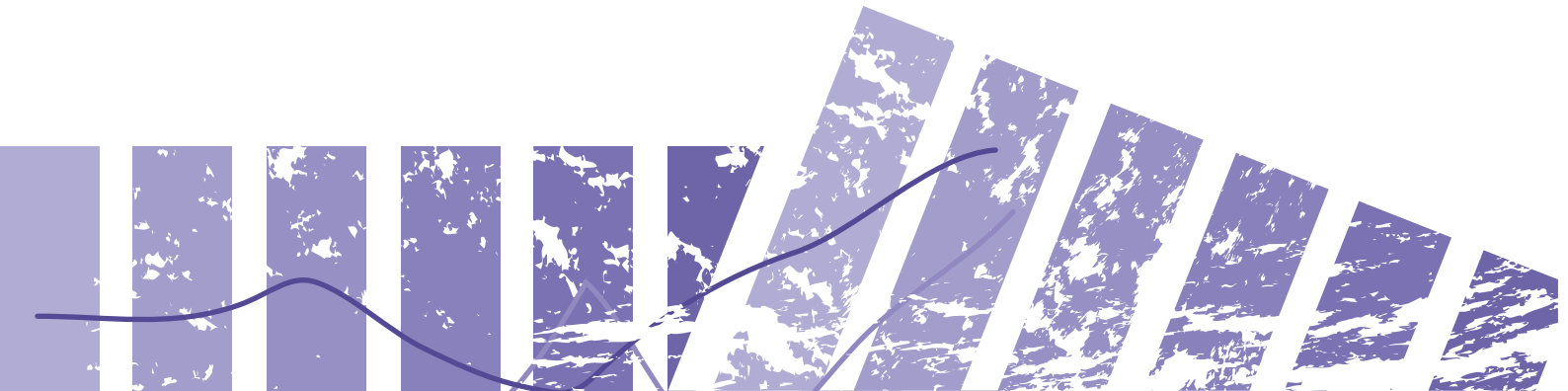
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 28 – 41

Dátum vydania/Publication date: 15. október 2020/October 15, 2020



Milan TEREK
Vysoká škola manažmentu

MOŽNOSTI RIEŠENIA PROBLÉMU NEODPOVEDANIA V ANALÝZACH DÁT PRI VYČERPÁVAJÚCOM SKÚMANÍ PROSTREDNÍCTVOM DOTAZNÍKOVÝCH ZISŤOVANÍ

POSSIBILITIES FOR SOLVING THE PROBLEM OF NONRESPONSE IN ANALYSES OF DATA IN CENSUSES REALIZED THROUGH QUESTIONNAIRE SURVEYS

ABSTRAKT
Cieľom článku je návrh spôsobu zohľadnenia neodpovedania pri odhadovaní charakteristík základného súboru pri vyčerpávajúcom skúmaní. Opisujeme účinky neodpovedania na bodové odhady, príčiny a prevencia neodpovedania a využívanie váh v poststratifikácii vo výberovom skúmaní. Navrhujeme metódu výpočtu konečných váh zohľadňujúcich neodpovedanie v poststratifikácii. Nakoniec navrhujeme spôsob výberu najvhodnejšej skupiny poststratifikačných premenných.

ABSTRACT
The aim of the paper is to propose how nonresponse has been taken into account in estimating the characteristics of the population in the censuses. The effects of nonresponse on the point estimates, the causes and the prevention of nonresponse and the use of weights in poststratification in sample surveys are described. The calculation method of final weights taking into account nonresponse in poststratification is suggested. Finally, a method of selecting the most appropriate group of poststratification variables is proposed.

KĽÚČOVÉ SLOVÁ
neodpovedanie pri vyčerpávajúcom skúmaní, váhy v poststratifikácii, poststratifikačné premenné

KEY WORDS
nonresponse in censuses, poststratification using weights, poststratification variables

1. ÚVOD
V minulosti nebol problém neodpovedania v štatistických prieskumoch (zisťovaniach) taký významný. Vplyvom zmien v spoločnosti je v súčasnosti odlišná spoločenská klíma, ktorá veľmi často spôsobuje menšiu ochotu poskytovať dáta. Je nevyhnutné vyrovnať sa v analýzach dát získaných v štatistických prieskumoch s vyššou mierou neodpovedania (*nonresponse*). Vysoká miera neodpovedania môže významne znehodnotiť kvalitu a vypovedaciu schopnosť výsledkov štatistických prieskumov. V štatistickom prieskume sa väčšinou vytvorí zoznam otázok, ktoré sa zhromaždia v dotazníku, ktorý sa potom poskytne na vyplnenie jednotkám. Vtedy možno hovoriť o dotazníkovom prieskume (podrobnejšie pozri v [9]). V príspevku uvažujeme len o dotazníkových prieskumoch.

Všeobecne možno uvažovať o dvoch typoch neodpovedania: neodpovedanie jednotky (*unit nonresponse*), pri ktorom chýbajú hodnoty všetkých premenných v dotazníku. Čiastočné neodpovedanie jednotky (*item nonresponse*) znamená, že

chýba hodnota aspoň jednej, ale nie všetkých premenných v dotazníku [5]. Nevrátenie vyplneného dotazníka znamená neodpovedanie jednotky, vrátenie čiastočne vyplneného dotazníka znamená čiastočné neodpovedanie jednotky. Oba typy neodpovedania zmenšujú presnosť odhadovania, spravidla sa im však dá len veľmi ťažko vyhnúť.

Imputovanie (*imputation*) znamená nahradenie chýbajúcich hodnôt premenných blízkymi hodnotami. Používa sa v prípade čiastočného neodpovedania jednotky. Najčastejšie sa v prieskumoch postupuje tak, že sa najprv realizuje imputovanie v rámci jednotiek, ktoré čiastočne neodpovedali, a potom sa už uvažuje len o neodpovedaní jednotiek a realizuje sa váženie (kombinovaný prístup [5]).

Je veľa štúdií, ktoré sa venujú riešeniam problému neodpovedania pri výberovom skúmaní. Väčšinou sú zamerané na minimalizáciu negatívneho dosahu neodpovedania na presnosť hodnôt bodových odhadov. Niekedy je k dispozícii zoznam všetkých jednotiek v základnom súbore. To je bežné napríklad vtedy, keď sa robí prieskum v nejakej organizácii, v ktorej je k dispozícii kontakt na každého pracovníka organizácie, prípadne máme kontakty na všetky jednotky základného súboru z nejakého registra alebo databázy. Keď sa finančné, časové a iné náklady na vyčerpávajúce skúmanie prakticky nelíšia od nákladov na výberové skúmanie, nemá realizácia náhodného vyberania jednotiek zmysel. Jednoducho pošleme dotazník všetkým jednotkám v základnom súbore s prosbou o jeho vyplnenie. Nie sú však známe štúdie o riešení problému neodpovedania pri vyčerpávajúcom skúmaní.

Príspevok je venovaný jednej možnosti riešenia problému neodpovedania pri vyčerpávajúcom skúmaní. Navrhujeme modifikáciu jednej metódy minimalizácie negatívneho vplyvu neodpovedania na presnosť hodnôt odhadov, známej vo výberovom skúmaní. Metóda spočíva vo výpočte konečných váh v poststratifikácii, v ktorých sa kompenzuje neodpovedanie.

Často môže byť k dispozícii viac potenciálnych poststratifikačných premenných. To umožňuje uvažovať v poststratifikácii o viacerých skupinách poststratifikačných premenných. V príspevku navrhujeme spôsob výberu najvhodnejšej skupiny.

2. PREHĽAD LITERATÚRY A METÓDY

2.1 ÚČINOK NEODPOVEDANIA NA PRESNOSŤ ODHADOV

Predpokladajme, že sa odhaduje (podrobnejšie o odhadoch pozri napr. v [7], [8]) stredná hodnota μ skúmanej premennej *y* v konečnom základnom súbore rozsahu *N*.

Ak
N_R – počet jednotiek v základnom súbore, ktoré by odpovedali, keby boli vybrané[1],
N_{NR} – počet jednotiek v základnom súbore, ktoré by neodpovedali, keby boli vybrané[2] $(N_{NR} = N - N_R)$,
μ_R – stredná hodnota súboru N_R jednotiek, ktoré by odpovedali, keby boli vybrané,
μ_{NR} – stredná hodnota súboru N_{NR} jednotiek, ktoré by neodpovedali, keby boli vybrané.

[1] *Základný súbor odpovedajúcich (respondent population).*
[2] *Základný súbor neodpovedajúcich (non-respondent population).*

potom

$$\mu = \frac{N_R\mu_R + N_{NR}\mu_{NR}}{N} \tag{1}$$

je stredná hodnota premennej *y* v celom základnom súbore rozsahu *N*.

Keď máme náhodný výber *n* jednotiek, potom v skutočnosti na základe tohto náhodného výberu odhadujeme μ_R, nie μ. Uvažujme teraz o jednoduchom náhodnom vyberaní. Keď náhodný výber *n* jednotiek obsahuje len n_R jednotiek, ktoré odpovedali, a $\bar{Y}$ je výberový priemer týchto n_R jednotiek, potom:

$$E(\bar{Y}) = \mu_R \tag{2}$$

a vychýlenie bodového odhadu $\bar{Y}$ je

$$B(\bar{Y}) = \mu_R - \mu = \mu_R - \frac{N_R\mu_R + N_{NR}\mu_{NR}}{N} = \frac{N_{NR}}{N}(\mu_R - \mu_{NR}) \tag{3}$$

Všeobecne účinok neodpovedania závisí od podielu jednotiek, ktoré by neodpovedali a od rozdielu medzi strednými hodnotami jednotiek, ktoré by odpovedali, a jednotiek ktoré by neodpovedali. Žiaľ, hodnoty N_{NR}, μ_R a μ_{NR} spravidla nepoznáme.

Posledný vzťah ukazuje, že vychýlenie dané neodpovedaním je nezávislé od *n*, a nemožno ho redukovať zväčšením rozsahu výberu. Možno ho však redukovať napríklad zmenšením podielu $\frac{N_{NR}}{N}$ jednotiek ktoré by neodpovedali. To naznačuje veľký význam preventívnych opatrení na zmenšenie podielu jednotiek, ktoré by neodpovedali (podrobnejšie pozri v [3, s. 397 – 399], [4, s. 330 – 332], a v [6]).

2.2 PRÍČINY A PREVENCIA NEODPOVEDANIA

Často sa pri príprave plánu výberového skúmania venuje málo času analýze problému možného neodpovedania. V [4] sa na s. 333 – 336 uvádza, že príčiny neodpovedania najčastejšie súvisia s obsahom prieskumu, metódami zhromažďovania dát alebo s charakteristikami respondentov, prípadne aj s časom prieskumu, majstrovstvom anketárov, návrhom dotazníka (podrobnejšie pozri aj v [1] a v [9, s. 9 – 18]), prezentáciou dôležitosti prieskumu a s pripomenutím prieskumu v prípade neodpovedania.

Analytik, ktorý dobre pozná základný súbor, by mal byť schopný predvídať príčiny neodpovedania a realizovať účinnú prevenciu. Na poznávanie príčin neodpovedania možno využiť navrhovanie experimentov a aplikáciu metód zlepšovania kvality v procese zhromažďovania a spracovania dát. Všeobecne treba vyvinúť maximálne úsilie na získanie odpovedí všetkých respondentov. Aj po dôkladnej analýze možných príčin neodpovedania a realizácii účinných preventívnych opatrení však treba vždy počítať s istou mierou neodpovedania.

Niekedy sa pri odhadovaní neodpovedanie úplne alebo čiastočne ignoruje. Jednoducho sa zhromaždia úplne a čiastočne vyplnené dotazníky, potom sa prípadne aplikuje niektorá z metód imputovania (podrobnejšie pozri napr. v [3, s. 408 – 419] a v [4, s. 346 – 351]) na doplnenie chýbajúcich odpovedí a ďalej sa aplikujú bežné

metódy odhadovania bez zohľadnenia neodpovedania jednotiek. Ignorovanie účinkov neodpovedania však môže spôsobiť vážne vychýlenia odhadov.

2.3 VÝBEROVÉ VÁHY A ICH MODIFIKÁCIA

V pravdepodobnostnom výbere má každá jednotka v základnom súbore známu pravdepodobnosť, že bude vo výberovom súbore. Pravdepodobnosti

$$\pi_i = P \text{ (jednotka } i \text{ bude vo výberovom súbore)}$$

sa obyčajne nazývajú pravdepodobnosti zahrnutia (*inclusion probabilities*) a sú známe pre ľubovoľnú výberovú schému, už pred začiatkom realizácie výberového skúmania [4, s. 28]. Výberové váhy (základné, *base weights, design weights*) w_{Bi} sú pre ľubovoľnú výberovú schému definované ako:

$$w_{Bi} = \frac{1}{\pi_i} \tag{4}$$

Základnú váhu jednotky *i* možno interpretovať ako počet jednotiek v základnom súbore, reprezentovaných jednotkou *i*.

Výberová báza (opora výberu, *frame population*) je zoznam zostavený s cieľom tvorby výberu, ktorý označuje jednotky základného súboru tak, aby sa mohli brať do úvahy pri ich skúmaní [11]. V ideálnom prípade výberová báza reprezentuje presne množinu fyzicky existujúcich jednotiek, ktoré tvoria cieľový základný súbor. Cieľový základný súbor (*target population*) je základný súbor, o ktorom chceme robiť induktívne úsudky. V praxi sa cieľový základný súbor a výberová báza viac alebo menej líšia.

Predpokladajme, že sme náhodne vybrali *n* jednotiek so známymi pravdepodobnosťami zahrnutia π_i, i = 1, ...,*n*. Keď niektoré jednotky vo výbere neodpovedali a niektoré jednotky z cieľového základného súboru nie sú vo výberovej báze, možno konečnú váhu w_i pozorovania *i* vyjadriť ako súčin troch komponentov

$$w_i = w_{Bi} \cdot w_{NRi} \cdot w_{NCi} \tag{5}$$

kde

w_{NRi} je faktor úpravy vzhľadom na neodpovedanie,

w_{NCi} je faktor kompenzácie nepokrytia. Nepokrytie (*non-coverage*) alebo neúplné pokrytie je spôsobené tým, že niektoré jednotky z cieľového základného súboru nie sú vo výberovej báze. Výberová báza nepokrýva celý cieľový základný súbor.

Váhy, o ktorých uvažujeme, majú tieto vlastnosti (sú kalibrované[3]):

1. $\sum_{i=1}^{n_R} w_{Bi} = N_R$
2. $\sum_{i=1}^{n_R} w_{Bi} \cdot w_{NRi} = N_F$
3. $\sum_{i=1}^{n_R} w_i = \sum_{i=1}^{n_R} w_{Bi} \cdot w_{NRi} \cdot w_{NCi} = N$

[3] Kalibrácia je procedúra, v ktorej sú váhy upravené tak, že odhadnuté úhrny pomocných premenných sú v súlade s aktuálnymi úhrnmi týchto premenných v základnom súbore [4, s. 154].

kde n_R je rozsah výberu zo základného súboru odpovedajúcich rozsahu N_R , N_F je rozsah výberovej bázy a *N* je rozsah cieľového základného súboru. Základný súbor odpovedajúcich respondentov je podmnožina výberovej bázy, ktorá je reprezentovaná jednotkami, ktoré by v prieskume odpovedali, keby boli vybraté do výberu. Ide len o teoretický koncept, pretože je nemožné identifikovať jednotky tohto základného súboru.

2.4 SKLON K ODPOVEDANIU

Ak Z_i je premenná, ktorá indikuje prítomnosť jednotky vo výberovom súbore, pravdepodobnosť *P* (Z_i = 1) = π_i, náhodnú premennú R_i definujme:

$$R_i = \begin{cases} 1, \text{ keď jednotka } i \text{ odpovedá}, \\ 0, \text{ keď jednotka } i \text{ neodpovedá}. \end{cases}$$

Po realizácii náhodného vyberania sú realizácie náhodnej premennej R_i známe pre všetky jednotky v náhodnom výbere. Nech y_i je hodnota študovanej premennej. Hodnota y_i sa zaznamená, keď r_i, realizácia náhodnej premennej R_i, sa rovná 1. Pravdepodobnosť, že jednotka *i* vybratá do výberového súboru bude zodpovedať

$$\varphi_i = P\ (R_i = 1),$$

je neznáma a predpokladáme, že je väčšia ako nula. Pravdepodobnosť φ_i sa nazýva sklon k odpovedaniu (*response propensity*) *i*-tej jednotky. Keď R_i je nezávislá od Z_i, potom pravdepodobnosť, že jednotka *i* bude meraná je:

$$P \text{ (jednotka } i \text{ je vo výbere a bude odpovedať)} = \pi_i \varphi_i.$$

Sklon k odpovedaniu φ_i sa odhaduje pre každú jednotku vo výbere na základe doplnkových informácií, ktoré sú známe pre všetky jednotky vo výberovom súbore. Konečná váha pre respondenta je potom $1/(\pi_i \hat{\varphi}_i)$, kde $\hat{\varphi}_i$ je odhadnutý sklon k odpovedaniu. Nech $\boldsymbol{x}_i$ je vektor informácií, známy o jednotke *i* vo výbere. Ak φ_i závisí od $\boldsymbol{x}_i$ ale nie od y_i, dáta sú typu "náhodne chýbajúce" (*missing at random* - MAR *data*). Podrobnejšie o *Missing Completely at Random* (MCAR *data*), *Missing at Random* (MAR *data*) a *Not Missing at Random* (NMAR *data*), pozri napr. v [4, s. 338 – 340].

2.5 VYUŽITIE VÁH V POSTSTRATIFIKÁCII

Váhové metódy (*weighting methods*) predpokladajú MAR dáta. Potom možno sklony k odpovedaniu odhadnúť na základe premenných, ktorých hodnoty sú známe pre všetky jednotky vo výbere. Budeme uvažovať o poststratifikácii, ktorá patrí medzi váhové metódy. Zároveň ide o špeciálny prípad kalibračnej metódy [4, s. 346]. Predpokladá sa, že jednotky, ktoré odpovedali aj ktoré neodpovedali v tom istom poststrate, sú podobné. Váhy jednotiek z toho istého poststrata, ktoré odpovedali sa zvýšia tak, že reprezentujú okrem seba aj jednotky, ktoré neodpovedali.

Po realizácii jednoduchého náhodného vyberania sa jednotky zaradia do *H* rozličných poststrát. Základný súbor má N_h jednotiek v *h*-tom poststrate. Z nich n_h

jednotiek bolo vybratých a z nich n_{hR} odpovedalo. Sklon k odpovedaniu pre každú jednotku v poststrate *h* odhadneme pomocou váženého pomeru odpovedania (*weighted response rate*) RR_{w_h}. Pre každého respondenta *i* v poststrate *h* sa sklon k odpovedaniu odhadne pomocou:

$$RR_{w_h} = \frac{\sum_{i=1}^{n_{hR}} w_{Bi}}{N_h}, \tag{6}$$

a faktor úpravy vzhľadom na neodpovedanie w_{NRi} je:

$$w_{NRi} = \frac{1}{RR_{w_h}} \tag{7}$$

Alternatívne možno pri odhadovaní sklonu k odpovedaniu použiť aj logistickú regresiu (podrobnejšie napr. v [3, s. 504 – 505]). Pri použití poststratifikácie na kompenzáciu neodpovedania modifikujeme základné váhy tak, že výber je v poststratách kalibrovaný na úhrny v základnom súbore [4, s. 342]. Poststratifikačný bodový odhad strednej hodnoty alebo úhrnu je približne nevychýlený, keď v každom poststrate *h*:

a) výstup y_i je nekorelovaný so sklonom k odpovedaniu φ_i ,
b) sklon k odpovedaniu φ_i je rovnaký pre každú jednotku, alebo
c) hodnota študovanej premennej y_i je rovnaká.

Odporúča sa použiť čo najviac poststrát, aby bolo splnenie uvedených predpokladov hodnovernejšie [4, s. 343 – 344].

Keď w_{NRi} > 2, poststratum obsahuje viac jednotiek ktoré neodpovedali, ako tých, čo odpovedali. V takých prípadoch sa rozptyl hodnôt odhadov zvyšuje, váhy nie sú stabilné. Odporúča sa zlúčiť susedné poststratá (*collapsing*) tak, aby w_{NRi} bolo menšie alebo rovné 2 [4, s. 342]. To isté sa odporúča, keď je počet jednotiek v poststrate menší ako 20. V [2] sa odporúča zlučovať poststratá, ktoré majú podobné stredné hodnoty kľúčových premenných.

2.6 ODHADOVANIE ZALOŽENÉ NA VÝBEROVÝCH VÁHACH

Ak y_i je meranie na jednotke *i*, a w_i je výberová váha jednotky *i*, realizovaný výber označíme *S*. Ide o podmnožinu *n* jednotiek zo základného súboru *U*. Bodový odhad úhrnu je [4, s. 286]:

$$\hat{\tau} = \sum_{i \in S} w_i y_i \tag{8}$$

Bodovým odhadom strednej hodnoty je

$$\hat{\mu}_K = \frac{1}{\sum_{i \in S} w_i} \sum_{i \epsilon S} w_i y_i, \tag{9}$$

kde $\sum_{i \in S} w_i$ odhaduje počet jednotiek v základnom súbore.

Ak y_i sa rovná 1, keď jednotka má nejaký znak, a rovná sa 0, keď ho nemá, potom podiel π je

$$\pi = \frac{\sum_{i=1}^{N} y_i}{N} \tag{10}$$

a π sa odhaduje $\hat{\pi} = \hat{\mu}_K$.

3. NEODPOVEDANIE PRI VYČERPÁVAJÚCOM SKÚMANÍ

Aj pri vyčerpávajúcom skúmaní možno mieru neodpovedania redukovať vhodnými preventívnymi opatreniami, ktoré sú rovnaké ako v prípade výberového skúmania. Vyčerpávajúce skúmanie budeme charakterizovať trochu inak ako obvykle. Doteraz sme uvažovali o výberovej schéme, ktorá obsahuje aj jednoduché náhodné vyberanie. Pri vyčerpávajúcom skúmaní základného súboru rozsahu *N* sa vyberú všetky jednotky základného súboru. To je vlastne prípad náhodného vyberania bez opakovania rozsahu *N* s jediným malým rozdielom, že posledná jednotka zo základného súboru je vybraná nenáhodne. Pravdepodobnosť zaradenia jednotky *i* v takomto výbere sa rovná jednej, rovnako ako jej základná výberová váha.

Pri vyčerpávajúcom skúmaní možno váhy použiť aj na zohľadnenie neodpovedania. Majme náhodnú premennú R_i:

$$R_i = \begin{cases} \text{1, keď jednotka } i \text{ odpovedá,} \\ \text{0, keď jednotka } i \text{ neodpovedá.} \end{cases}$$

Hodnota y_i sa zaznamená, keď r_i, realizácia náhodnej premennej R_i, sa rovná 1. Pravdepodobnosť, že jednotka bude odpovedať

$$\varphi_i = P(R_i = 1),$$

je neznáma, ale predpokladáme, že je väčšia ako nula. Pravdepodobnosť φ_i je sklon k odpovedaniu pre jednotku *i* aj pri vyčerpávajúcom skúmaní. Tu sa však pravdepodobnosť že jednotka *i* je vybraná a bude odpovedať, redukuje na pravdepodobnosť, že jednotka *i* bude odpovedať, pretože sú vybrané všetky jednotky základného súboru. Potom

$$P \text{ (jednotka } i \text{ je vo výbere a bude odpovedať)} = \varphi_i$$

Pravdepodobnosť odpovedania φ_i sa odhaduje pomocou $\hat{\varphi}_i$ pre každú jednotku v základnom súbore, s využitím pomocných informácií, ktoré sú známe pre všetky jednotky v základnom súbore. Konečná váha pre respondenta je potom $1/\hat{\varphi}_i$. Aj v tomto prípade predpokladáme MAR dáta. Potom možno využiť váhové metódy.

3.1 MODIFIKÁCIA POSTSTRATIFIKÁCIE, KTORÁ VYUŽÍVA VÁHY

Nech základný súbor má N_h jednotiek v poststrate *h*, a nech z nich N_{hR} jednotiek odpovedalo. Sklon k odpovedaniu pre každú jednotku v poststrate *h* odhadneme

pomocou váženého pomeru odpovedania. Potom pre každého respondenta *i* v poststrate *h* je sklon k odpovedaniu odhadnutý pomocou

$$\hat{\varphi}_i = RR_{w_h} = \frac{\sum_{i=1}^{N_{hR}} w_{Bi}}{N_h} = \frac{N_{hR}}{N_h} \tag{11}$$

a konečná váha pre jednotku *i* je

$$w_i = \frac{1}{RR_{w_h}} = \frac{N_h}{N_{hR}} \tag{12}$$

Pri vyčerpávajúcom skúmaní sa konečná váha rovná faktoru úpravy vzhľadom na neodpovedanie w_{NRi}.

Príklad. Na univerzite pracuje 892 učiteľov. Manažment univerzity má informácie o veku a pracovnej pozícii každého učiteľa. Rozdelenie učiteľov podľa veku a pracovnej pozície je v tabuľke č. 1. Manažment univerzity robil dotazníkový prieskum o názore učiteľov na niektoré pracovné podmienky. Všetci učitelia boli oslovení prostredníctvom e-mailu s prosbou o vyplnenie priloženého dotazníka. Počty učiteľov, ktorí vrátili vyplnený dotazník v určenom termíne, sú v tabuľke č. 1 v zátvorkách.

Tabuľka č. 1: Rozdelenie učiteľov podľa veku a pracovnej pozície

Pracovná pozícia **Veková kategória**	**Asistent**	**Odborný asistent**	**Docent**	**Profesor**	**Spolu**
24 – 34	22 (19)	41 (27)	10 (8)		73 (54)
35 – 44		81 (19)	131 (37)		212 (56)
45 – 54		21 (3)	326 (64)	76 (26)	423 (93)
55 – 64			26 (5)	96 (15)	122 (20)
65 –			16 (3)	46 (11)	62 (14)
Spolu	22 (19)	143 (49)	509 (117)	218 (52)	892 (237)

Zdroj: vlastné výpočty

Jedna otázka v dotazníku bola „Preferujete v zabezpečení stravovania na univerzite stravné lístky alebo finančný príspevok na stravovanie?”, s možnými odpoveďami “stravné lístky” alebo “finančný príspevok”. Chceme odhadnúť podiel všetkých učiteľov univerzity, ktorí preferujú stravné lístky.

Pretože sa realizovalo vyčerpávajúce skúmanie, základná váha každej jednotky vo výbere sa rovná jednej. Všetky kombinácie kategórií premenných „pracovná pozícia“ a „veková kategória“, budú tvoriť poststratá. Celkovo máme 4 *x* 5 = 20 poststrát. Ďalej budeme uvažovať len o 12 poststratách, v ktorých je nenulový počet jednotiek. V každom z týchto poststrát poznáme počet jednotiek v základnom súbore. Tieto pomocné informácie možno využiť pri modifikácii základných váh vzhľadom na neodpovedanie. Sklon k odpovedaniu pre každú jednotku v poststrate *h* bude odhadnutý podľa (4) a pre každého učiteľa *i* v poststrate *h*, ktorý odpovedal, sa konečná váha vypočíta podľa (5). Konečné váhy sú v tabuľke č. 2. Počty učiteľov, v

poststratách, ktorí vo vrátenom dotazníku odpovedali “stravné lístky”, sú v tabuľke č. 2 v zátvorkách.

Tabuľka č. 2: Konečné váhy pre poststratá

Pracovná pozícia **Veková kategória**	**Asistent**	**Odborný asistent**	**Docent**	**Profesor**
24 – 34	1,158 (18)	1,519 (22)	1,250 (3)	
35 – 44		4,263 (11)	3,541 (12)	
45 – 54		7,000 (2)	5,094 (25)	2,923 (10)
55 – 64			5,200 (1)	6,400 (3)
65 –			5,333 (1)	4,182 (2)

Zdroj: vlastné výpočty

Nech y_i sa rovná 1 keď učiteľ odpovedal „stravné lístky“ a rovná sa 0 keď odpovedal „finančný príspevok“. Potom, podľa (2)[4]

$$\hat{\pi} = \hat{\mu}_K = \frac{356{,}074}{892} = 0.3992 \qquad (13)$$

Odhadujeme, že približne 39,92 % učiteľov preferuje stravné lístky.

Na porovnanie, keby sme odhadli podiel π pomocou výberového podielu P, čo je pri riešení podobných problémov bežná prax, výsledok by bol:

$$P = \frac{110}{237} = 0{,}4641 \qquad (14)$$

Pomocou výberového podielu P by sme odhadli, že približne 46,41 % učiteľov preferuje stravné lístky. Rozdiel medzi výsledkami odhadovania je značný.

V tabuľkách č. 1 a 2 vidno, že nie vo všetkých poststratách majú jednotky konečné váhy menšie alebo rovné 2, alebo aspoň 20 učiteľov, ktorí odpovedali. Procedúra zlučovania poststrát je väčšinou úplne aplikovateľná vo „veľkých prieskumoch“, realizovaných najčastejšie štatistickými úradmi. Môže byť však problematická pri „malých prieskumoch“ v organizáciách. V uvedenom príklade sme realizovali niekoľko spôsobov zlučovania strát, no v žiadnom z nich nebolo možné dosiahnuť úplné splnenie oboch podmienok. Na druhej strane sa ale ukázalo, že hodnoty odhadov sa vo všetkých vyskúšaných štruktúrach so zlúčenými poststratami len minimálne líšili od vypočítanej hodnoty. Zdá sa, že zlučovanie poststrát nemá v tomto príklade veľký vplyv na bodový odhad $\hat{\pi}$. Všeobecne možno odporúčať preverenie citlivosti uvažovaných bodových odhadov na zlučovanie poststrát v každom konkrétnom prieskume a na základe toho prijať rozhodnutie o ich zlučovaní.

[4] V čitateli je súčet súčinov konečných váh a počtov učiteľov v poststratách, ktorí vo vrátenom dotazníku odpovedali “stravné lístky” (z tabuľky č. 2), v menovateli je súčet súčinov konečných váh z tabuľky č. 2 a príslušných počtov učiteľov, ktorí vrátili vyplnený dotazník (čísla z tabuľky č. 1 v zátvorkách).

V príklade je však veľký rozdiel medzi hodnotou odhadu získanou cez poststratifikáciu, ktorá využíva váhy, a medzi hodnotou odhadu získanou bežným postupom, v príklade pomocou výberového podielu *P*, ktorý neberie do úvahy neodpovedanie. Vzhľadom na to, že predpokladáme MAR dáta a váhy jednotiek ktoré odpovedali, sa zvýšia tak, že reprezentujú okrem seba aj jednotky, ktoré neodpovedali, je zrejmé, že uvedený postup, ktorý aj pri vyčerpávajúcom skúmaní berie do úvahy úpravu váh vzhľadom na neodpovedanie, je lepší ako prístup, ktorý berie do úvahy len informácie od jednotiek, ktoré odpovedali, bez ohľadu na neodpovedanie.

V príklade sme uvažovali o dvoch poststratifikačných premenných pracovná pozícia“ a „veková kategória“. Často je organizácii prístupných viacero potenciálnych poststratifikačných premenných. V príklade by napríklad mohli byť k dispozícii aj hodnoty premennej „Pohlavie“, prípadne aj iných premenných. To umožňuje uvažovať o viacerých skupinách poststratifikačných premenných. Z toho vyplýva potreba stratégie ich výberu.

Poststratifikácia môže redukovať variabilitu bodových odhadov. Na dosiahnutie najväčšej redukcie variability treba minimalizovať variabilitu študovaných premenných vnútri poststrát, čo možno dosiahnuť takou voľbou poststratifikačných premenných, ktoré sú silno korelované so študovanými premennými. To je prvá stratégia voľby poststratifikačných premenných [3, s. 507]. Ak sú poststratá určené tak, že rozdiel $(\mu_R - \mu_{NR})$ je veľmi malý, vychýlenie odhadu spôsobené neodpovedaním bude tiež malé [3, s. 503]. To sa dá najlepšie dosiahnuť nájdením takých poststratifikačných premenných, ktoré sú vysoko korelované so sklonom k odpovedaniu. Vtedy bude sklon k odpovedaniu v každom poststrate, pre každú jednotku približne rovnaký. Čím silnejší je vzťah medzi poststratifikačnými premennými a sklonom k odpovedaniu, tým sú poststratifikačné premenné vhodnejšie na následné odhadovanie s ohľadom na neodpovedanie. To je druhá stratégia voľby poststratifikačných premenných. Tieto dve stratégie môžu viesť k rozličným výberom poststratifikačných premenných. Pri ich konečnom výbere by sa malo pamätať na obidva ciele [3, s. 507].

Zdá sa, že vhodným nástrojom na meranie sily vzťahu medzi sklonom k odpovedaniu a poststratifikačnými premennými je korelačný pomer. Všeobecne ide o charakteristiku vzťahu medzi variabilitou vnútri individuálnych kategórií a variabilitou v celom základnom súbore alebo vo výbere [10]. Charakteristika je definovaná ako pomer dvoch smerodajných odchýlok, ktoré reprezentujú tieto dva typy variability.

Každé pozorovanie (hodnotu kvantitatívnej premennej *z*) označíme z_{xi}, kde *x* označuje kategóriu, do ktorej pozorovanie patrí a *i* je index pozorovania. Ak n_x je počet pozorovaní v kategórii *x*., potom

$$\bar{z}_x = \frac{\sum_i z_{xi}}{n_x} \tag{15}$$

a

$$\bar{z} = \frac{\sum_x n_x \bar{z}_x}{\sum_x n_x} \tag{16}$$

kde $\bar{z}_x$ je priemer premennej *z* v kategórii *x* a $\bar{z}$ je priemer premennej *z* v základnom súbore. Korelačný pomer $\eta_{(z|x)}$ je definovaný ako druhá odmocnina z $\eta^2_{z|x}$, kde

$$\eta^2_{z|x} = \frac{\sum_x n_x(\bar{z}_x - \bar{z})^2}{\sum_{x,i}\left(z_{x,i} - \bar{z}\right)^2} \tag{17}$$

Korelačný pomer $\eta_{z|x} \in [0, 1]$. Hodnota $\eta_{z|x} = 0$ reprezentuje prípad, keď neexistuje variabilita medzi strednými hodnotami rozličných kategórií, $\eta_{z|x}$ = 1 svedčí o neexistencii variability vnútri príslušných kategórií.

V kontexte analýzy neodpovedania premenná *z* je kvantitatívna diskrétna premenná, ktorá nadobúda 2 hodnoty, 1 a 0 – počet odpovedí. Keď respondent *i* z kategórie *x* odpovie, tak z_{xi} = 1, keď neodpovie, tak z_{xi} = 0. V uvedenom príklade vyjde $\eta_{z|x} = 0{,}3469$.

4. ZÁVER

Dáta z vyčerpávajúceho skúmania možno chápať ako dáta z náhodného výberu bez opakovania rozsahu *N*, v ktorom je posledná jednotka vybratá nenáhodne. Pravdepodobnosť zahrnutia každej jednotky v takomto výbere sa rovná jednej, rovnako ako jej základná váha. Pri vyčerpávajúcom skúmaní sa pravdepodobnosť, že jednotka *i* je vybraná do výberu a odpovie, redukuje na pravdepodobnosť, že jednotka *i* odpovie, pretože všetky jednotky zo základného súboru sú vybrané.

Predpokladáme MAR dáta. Potom možno použiť metódy váženia. Je navrhnutá modifikácia poststratifikácie, ktorá využíva váhy. Konečná váha pre jednotku *i* sa vypočíta podľa vzťahu (5).

Keď miera odpovedania pre každé poststrátum nie je aspoň 50 % alebo počet pozorovaní v poststrate nie je aspoň 20, odporúča sa zlučovanie susedných poststrát. Procedúra zlučovania poststrát je väčšinou úplne aplikovateľná vo „veľkých prieskumoch“, realizovaných najčastejšie štatistickými úradmi. Môže byť však problematická pri „malých prieskumoch“ v organizáciách. Všeobecne možno odporúčať preverenie citlivosti uvažovaných bodových odhadov na zlučovanie poststrát v každom konkrétnom prieskume a na základe toho prijať rozhodnutie o ich zlučovaní.

Obyčajne je v prieskume k dispozícii viac potenciálnych poststratifikačných premenných. To umožňuje uvažovať pri odhadovaní o rozličných poststratifikačných premenných. Poststratifikačné premenné z hľadiska ich spojenia so sklonom k odpovedaniu možno ohodnotiť pomocou korelačného pomeru. Čím väčšia je hodnota korelačného pomeru, tým je vzťah medzi poststratifikačnými premennými a sklonom k odpovedaniu silnejší. Poststratifikačné premenné s maximálnou hodnotou korelačného pomeru možno odporúčať na použitie v procese úpravy váh vzhľadom na neodpovedanie. Na druhej strane zväčšenie hodnoty korelačného pomeru môže byť sprevádzané horším plnením podmienok na minimálny počet pozorovaní a pomer odpovedania v poststratách. Potom treba hľadať vhodné kompromisné riešenie medzi poststratifikačnými premennými s väčším korelačným pomerom a horším splnením

podmienok a poststratifikačnými premennými s menším korelačným pomerom a lepším splnením podmienok. Nemožno zabúdať ani na požiadavku silnej korelácie medzi poststratifikačnými a študovanými premennými.

LITERATÚRA

[1] BETHLEHEM, J.: Applied Survey Methods. A Statistical Perspective. Hoboken: Wiley and Sons, 2009. 375 s. ISBN 978-0-470-37308-8.
[2] GELMAN, A. – CARLIN, J. B.: Poststratification and weighting adjustments. Groves, R. M. – Dillman, D. – Eltinge, J. and Little, R. (eds.): Survey nonresponse. New York: Wiley and Sons, 2002, s. 289 – 302.
[3] LEVY, P. S. – LEMESHOW, S.: Sampling of Populations. Methods and Applications. Fourth Edition. Hoboken: Wiley and Sons, 2008. 576 s. ISBN 978-0-470-04007-2.
[4] LOHR, S. L.: Sampling: Design and Analysis. 2nd edition. Boston: Brooks/Cole, 2010. 596 s. ISBN-10: 0-495-11084-1.
[5] SÄRNDAL, C.-E. – LUNDSTRÖM, S.: Estimation in Surveys with Nonresponse. Hoboken: Wiley and Sons, 2005. ISBN 0-470-01133-5.
[6] TEREK, M.: Možnosti riešenia problému neodpovedania v štatistických prieskumoch. In: Ekonomické rozhľady 2014, č. 2, s. 150 – 165.
[7] TEREK, M.: Interpretácia štatistiky a dát. Piate, doplnené vydanie. Košice: Equilibria, 2017. 460 s. ISBN 978-80-8143-213-2.
[8] TEREK, M.: Interpretácia štatistiky a dát. Podporný učebný materiál. Piate, doplnené vydanie. Košice: Equilibria, 2017. 244 s. ISBN 978-80-8143-212-5.
[9] TEREK, M.: Dotazníkové prieskumy a analýzy získaných dát. 1. vydanie. Košice: Equilibria, 2019. 202 s. ISBN 978-80-8143-247-7.
[10] Correlation ratio [online]. [cit. 13. 7. 2018] Dostupné na: https://en.wikipedia.org/wiki/Correlation_ratio.
[11] STN ISO 3534-1. Štatistika. Slovník a značky. Časť 1: Všeobecné štatistické termíny a termíny používané v teórii pravdepodobnosti. Bratislava: Slovenský ústav technickej normalizácie, 2008.

Táto práca bola podporená vedeckou grantovou agentúrou VEGA, v rámci projektu číslo 1/0562/18 Vzájomná prepojenosť medzi ľudským kapitálom a informačnými a komunikačnými technológiami.

RESUMÉ

Vysoká miera neodpovedania môže významne znehodnotiť kvalitu a výpovednú schopnosť výsledkov dotazníkových prieskumov. Vychýlenie dané neodpovedaním možno vo výberovom skúmaní redukovať napríklad zmenšením podielu jednotiek ktoré by neodpovedali, keby boli vybrané do výberu. To naznačuje veľký význam preventívnych opatrení na zmenšenie tohto podielu.
Je veľa štúdií, ktoré sa venujú riešeniam problému neodpovedania pri výberovom skúmaní. Väčšinou sú zamerané na minimalizáciu negatívneho dopadu neodpovedania na presnosť hodnôt bodových odhadov. V pravdepodobnostnom výbere v rámci ľubovoľnej výberovej schémy, má každá jednotka v základnom súbore známu pravdepodobnosť zahrnutia. Základné výberové váhy sú definované ako obrátené hodnoty pravdepodobností zahrnutia. Konečné výberové váhy možno získať úpravou základných váh vzhľadom na neodpovedanie a na nepokrytie. Sklon k odpovedaniu sa odhaduje pre každú jednotku vo výbere na základe doplnkových informácií, ktoré sú známe pre všetky jednotky vo výberovom súbore. Konečná váha

pre respondenta je potom obrátenou hodnotou súčinu pravdepodobnosti, že jednotka bude vo výbere, a odhadnutého sklonu k odpovedaniu. Predpokladáme MAR dáta.
Pri poststratifikácii je sklon k odpovedaniu pre každú jednotku v poststrate odhadnutý pomocou váženého pomeru odpovedania. Faktor úpravy vzhľadom na neodpovedanie je potom obrátenou hodnotou váženého pomeru odpovedania. Keď poststratum obsahuje viac jednotiek ktoré neodpovedali ako tých, čo odpovedali alebo je počet jednotiek v poststrate menší ako 20, rozptyl hodnôt odhadov sa zvyšuje a váhy nie sú stabilné. Vtedy sa odporúča zlučovanie susedných poststrát. Pomocou konečných výberových váh možno odhadovať charakteristiky konečného základného súboru.
Niekedy je vhodné realizovať vyčerpávajúce skúmanie. Pri vyčerpávajúcom skúmaní základného súboru rozsahu *N* ide vlastne o prípad náhodného vyberania bez opakovania rozsahu *N* s jediným malým rozdielom, že posledná jednotka zo základného súboru je vybraná nenáhodne. Pravdepodobnosť zahrnutia jednotky *i* v takomto výbere sa rovná jednej, rovnako ako jej základná výberová váha. Pravdepodobnosť že jednotka *i* je vybraná a odpovedá sa redukuje na pravdepodobnosť, že jednotka *i* odpovedá, pretože sú vybrané všetky jednotky základného súboru.
V článku je navrhnutá modifikácia poststratifikácie, ktorá využíva váhy upravené vzhľadom na neodpovedanie. V príklade sa odhaduje podiel všetkých učiteľov univerzity, ktorí preferujú v zabezpečení stravovania stravné lístky. Tento podiel sme odhadli pomocou uvedenej modifikovanej metódy a na porovnanie aj „tradične“, pomocou výberového podielu. Rozdiel medzi hodnotami odhadov je značný.
V príklade sa uvažuje o dvoch poststratifikačných premenných. Často môže byť k dispozícii viac potenciálnych poststratifikačných premenných. To umožňuje uvažovať o viacerých skupinách poststratifikačných premenných. Poststratifikačné premenné by mali byť čo najviac korelované so sklonom k odpovedaniu. Zdá sa, že vhodným nástrojom na meranie sily vzťahu medzi sklonom k odpovedaniu a poststratifikačnými premennými je korelačný pomer.

RESUME
A high nonresponse rate can significantly degrade the quality and meaningfulness of questionnaire survey results. The bias given by nonresponse in a sample survey can be reduced, for example, by reducing the proportion of units that would not respond if they were selected into the sample. This indicates the great importance of preventive measures to reduce this proportion.
Many studies have been conducted addressing the problem of nonresponse in a sample survey. They are mostly aimed at minimizing the negative impact of nonresponse on the accuracy of point estimates. In probability sampling within any sampling design, the inclusion probability of each unit in the population is known. Base sample weights are defined as inverse values of inclusion probabilities. Final sampling weights can be obtained by adjusting the base weights to compensate for nonresponse and non-coverage. The response propensity is estimated for each unit in the sample based on additional information that is known for all units in the sample. The final weight for a respondent is then the reciprocal of the product of probability that the unit will be included in the sample and the estimated response propensity. We assume MAR data.
In poststratification, the response propensity for each unit in the poststratum is estimated using a weighted response rate. The nonresponse adjustment factor is then the reciprocal of the weighted response ratio. When the poststratum contains more non-responding than responding units, or the number of units in the poststratum is less

than 20, the variance of the estimates increases and the weights are not stable. Then the collapsing of neighboring poststrata is recommended. Using the final sample weights, the characteristics of the finite population can be estimated.
Sometimes the conduct of the census is advisable. The census of the finite population of size N is actually the random sampling without the replacement of size N with only a small difference, that the last unit is selected non-randomly from the population. The inclusion probability of unit i in such sample is equal to one, equally as its base sampling weight. The probability that unit i is selected and responds is reduced to the probability that unit i responds because all units from the population are selected.
A modification of poststratification using weights, adjusted for nonresponse is proposed in the paper. The example estimates the proportion of all university teachers who prefer meal vouchers in meal provision. This proportion was estimated using the proposed modified method and, for a comparison, also "traditionally", using a sample proportion. There is a considerable difference between the values of the estimates.
The example considers two poststratification variables. Often, more potential poststratification variables may be available. This enables consideration of several groups of poststratification variables. These variables should be correlated as much as possible with the response propensity. A correlation ratio seems to be a suitable tool for measuring the strength of relationship between the response propensity and poststratification variables.

PROFESIJNÝ ŽIVOTOPIS

***Prof. Ing. Milan Terek, PhD.**, od roku 2018 pracuje ako profesor na Vysokej škole manažmentu v Bratislave. Vedie kurzy Úvod do štatistiky, Štatistika, Matematika pre manažérov II, Kvantitatívne metódy pre manažérov a Kvantitatívne metódy vo výskume v oblasti podnikového manažmentu. V rokoch 1977 – 2018 pracoval na Ekonomickej univerzite v Bratislave. Viedol kurzy Štatistika, Štatistické riadenie kvality, Analýza rozhodovania, Hĺbková analýza dát, Výberové skúmanie, Lineárne programovanie, Nelineárne programovanie, Operačný výskum a Systémové modelovanie. Vo výskume sa zameriava na aplikácie štatistických metód v ekonómii a manažmente. Je autorom alebo spoluautorom 6 monografií, 10 vysokoškolských učebníc, 17 skrípt, 72 článkov vo vedeckých a odborných časopisoch a 115 príspevkov na vedeckých konferenciách, publikovaných v zborníkoch.*

KONTAKT

mterek@vsm.sk