Milan Terek

MOŽNOSTI RIEŠENIA PROBLÉMU NEODPOVEDANIA V ŠTATISTICKÝCH PRIESKUMOCH[1]

Abstract: *The paper deals with the possibilities of solution to the nonresponse in statistical surveys. Effects of nonresponse on the estimators are described and the causes of nonresponse are analysed. Then the methods of increasing the response rate in statistical surveys are presented. The weighting to reduce nonresponse bias, sampling variance and noncoverage error is studied in the main part of the paper. The weighting class adjustment and poststratification adjustments are described in detail. The application of described methods is illustrated on the example. Finally, the importance of auxiliary information is emphasized.*

Keywords: *statistical survey, nonresponse, coverage errors, weighting class adjustment, poststratification adjustment*

JEL: C 83

Úvod

V posledných rokoch sa výrazne rozšírilo používanie výberových skúmaní na báze štatistických prieskumov na získavanie informácií potrebných pri rozhodovaní. S tým je spojený aj rast množstva komplikácií spojených s dosahovaním vysokej miery odpovedania. To spôsobilo, že sa tejto problematike začala venovať značná pozornosť.

V minulosti nebol problém neodpovedania až taký zásadný. Vplyvom zmien v spoločnosti je v súčasnosti odlišná spoločenská klíma, ktorá veľmi často spôsobuje menšiu ochotu poskytovať dáta. Je nevyhnutné vyrovnať sa v analýzach získaných dát s vyššou mierou neodpovedania (*nonresponse*), ktorá je v súčasných prieskumoch bežná. Vysoká miera neodpovedania môže totiž výrazne znehodnotiť kvalitu a výpovednú schopnosť výsledkov štatistických zisťovaní.

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

[1] Tento článok vznikol s prispením grantovej agentúry VEGA v rámci projektu č. 1/0761/12: *Alternatívne prístupy k meraniu sociálno-ekonomického rozvoja (v kontexte Stratégie 2020 a poučení z globálnej finančnej krízy).*

chýba hodnota aspoň jednej, ale nie všetkých premenných v dotazníku (Särndal, Lundström 2005, s. 15), [8]. Napríklad nevrátenie vyplneného dotazníka znamená neodpovedanie jednotky, vrátenie čiastočne vyplneného dotazníka znamená čiastočné neodpovedanie jednotky. Oba typy neodpovedania zmenšujú presnosť odhadov, spravidla sa im však dá len veľmi ťažko vyhnúť. Pri mnohých prieskumoch si získanie aspoň 50 % miery odpovedania vyžaduje značné množstvo námahy a finančných zdrojov.

Imputovanie (*imputation*) znamená nahradenie chýbajúcich hodnôt premenných blízkymi hodnotami. Imputovanie sa používa v prípade čiastočného neodpovedania jednotky. Najčastejšie sa v prieskumoch postupuje tak, že najprv sa realizuje imputovanie pri jednotkách, ktoré čiastočne neodpovedali, a potom sa už uvažuje len o neodpovedaní jednotiek a uskutoční sa váženie. Takýto prístup sa nazýva kombinovaný (Särndal, Lundström 2005, s. 17), [8].

V článku si podrobnejšie všimneme problém neodpovedania jednotiek a problém nepokrytia.

1 Účinok neodpovedania a presnosť odhadov

Cieľom väčšiny výberových skúmaní je odhadnúť s čo najväčšou presnosťou parametre základného súboru, napríklad strednú hodnotu, úhrn alebo podiel. Sú známe nevychýlené bodové odhady týchto parametrov pre rozličné výberové schémy. Hlavný problém spôsobený neodpovedaním je potenciálne vychýlenie odhadov. Čím je miera neodpovedania väčšia, tým je potenciálne vychýlenie odhadov väčšie.

Všimnime si najprv chápanie pojmu základný súbor. Budeme definovať tri rozličné základné súbory.

- Cieľový základný súbor (*target population*) je základný súbor, o ktorom chceme robiť induktívne úsudky. Napríklad by išlo o súbor všetkých domácností v SR.
- Výberová báza (opora výberu, *frame population*) je[2] zoznam zostavený s cieľom tvorby výberu, ktorý označuje jednotky základného súboru tak, aby sa mohli brať do úvahy pri ich skúmaní. V ideálnom prípade výberová báza reprezentuje presne množinu fyzicky existujúcich jednotiek, ktoré tvoria cieľový základný súbor. V reálnej praxi sa cieľový základný súbor a výberová báza viac alebo menej líšia. Keď sa napríklad realizuje prieskum pomocou elektronickej pošty, často nie je možné poznať všetky jednotky cieľového základného súboru, pretože často nie sú k dispozícii adresy všetkých respondentov z cieľového základného súboru a pod.
- Základný súbor odpovedajúcich (*respondent population*) je podmnožina výberovej bázy, ktorá je reprezentovaná jednotkami, ktoré by v prieskume odpovedali, keby boli vybraté do výberu. Ide o hypotetický koncept, pretože je nemožné identifikovať všetky jednotky tohto základného súboru. Doplnkom základného súboru odpovedajúcich vo výberovej báze je súbor neodpovedajúcich.[3]

[2] Podľa [15].

[3] Sú známe aj iné koncepty, podrobnejšie v [7] a [8].

Predpokladajme, že sa odhaduje stredná hodnota μ skúmanej premennej x v konečnom cieľovom základnom súbore rozsahu N. Predpokladajme, že výberová báza presne pokrýva cieľový základný súbor.

Nech:

N_R – počet jednotiek v základnom súbore odpovedajúcich,
N_{NR} – počet jednotiek v súbore neodpovedajúcich[4] ($N_{NR} = N - N_R$),
μ_R – stredná hodnota základného súboru odpovedajúcich,
μ_{NR} – stredná hodnota súboru neodpovedajúcich,

$\mu = \frac{N_R\mu_R + N_{NR}\mu_{NR}}{N}$ je stredná hodnota premennej x v cieľovom základnom súbore.

Uvažujme teraz o jednoduchom náhodnom vyberaní. Keď náhodný výber n jednotiek obsahuje n_R jednotiek, ktoré odpovedali, a $\overline{X}$ je výberový priemer týchto n_R jednotiek, potom

$$E(\overline{X}) = \mu_R$$

a vychýlenie bodového odhadu $\overline{X}$ je

$$B(\overline{X}) = \mu_R - \mu = \mu_R - \frac{N_R\mu_R + N_{NR}\mu_{NR}}{N} = \frac{N_{NR}}{N}(\mu_R - \mu_{NR})$$

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.

2 Príčiny neodpovedania

Príčiny neodpovedania možno rozdeliť napríklad takto (Lohr 1999, s. 260), [6]:

- obsah výberového skúmania,
- metódy zhromažďovania dát,
- charakteristiky respondentov.[5]

[4] Súbor neodpovedajúcich je reprezentovaný neodpovedajúcimi jednotkami vo výbere.
[5] Podrobnejšie o príčinách neodpovedania pozri ([6], s. 259 – 262).

Často sa pri príprave plánu výberového skúmania venuje málo času analýze problému možného neodpovedania. Mnoho menej skúsených, ale niekedy aj skúsenejších osôb jednoducho začne zhromažďovať dáta bez toho, aby dôkladne premysleli riziká neodpovedania. Výskumník, ktorý dobre pozná základný súbor, by mal byť schopný predvídať príčiny neodpovedania a urobiť účinnú prevenciu. Na poznávanie príčin neodpovedania možno využiť navrhovanie experimentov a aplikáciu metód zlepšovania kvality v procese zhromažďovania a spracúvania dát.

Všeobecne treba vyvinúť maximálne úsilie na získanie odpovedí všetkých respondentov, aj keby sa musel redukovať rozsah pôvodného výberového súboru, aby sa neprekročil finančný limit určený na prieskum.

Z uvedených troch základných príčin neodpovedania možno účinne ovplyvniť hlavne metódy zhromažďovania dát. Podrobnejšie si všimneme niektoré možnosti rastu podielu odpovedania, ktoré súvisia s metódami zhromažďovania dát.

3 Metódy rastu podielu odpovedania v štatistických prieskumoch

Všimneme si najmä prieskumy v domácnostiach. Uvedieme niekoľko základných odporúčaní.

3.1 Zvyšovanie počtu úspešne kontaktovaných domácností

V prieskumoch prostredníctvom priamych interview sa môže objaviť problém, že pri návšteve domácnosti nie je nikto doma. Cez deň je väčšina rodičov v práci, deti v škole, preto sa odporúča navštevovať domácnosti osobne, alebo telefonovať večer.

V prieskumoch prostredníctvom pošty sa zasa môže stať, že rodina už nebýva na zaznamenanej adrese. Ak je vo výberovej báze len adresa, nie meno rodiny, je potrebná návšteva na získanie mena novej rodiny, ktorá býva na zaznamenanej adrese. Prípadne možno na list uviesť adresu: „Pán XY alebo súčasný obyvateľ".

3.2 Zvyšovanie podielu odpovedí v dotazníkoch zasielaných poštou

Materiál zasielaný do domácností by mal obsahovať aj starostlivo pripravený list, v ktorom sa vysvetľuje cieľ prieskumu, uvádza sa identifikácia organizácie, ktorá prieskum realizuje, ďalej ubezpečenie, že poskytnuté informácie sú prísne dôverné a použijú sa len v agregátnej forme na účely štatistiky. Hlavne ubezpečenie o dôvernosti poskytnutých dát je mimoriadne dôležité.

Podstatné sú aj zdanlivé maličkosti. Materiály by mali byť veľmi kvalitné (kvalitný papier, kvalitná tlač a pod.) a zaslané prvou triedou. Treba priložiť obálku na zaslanie vyplneného dotazníka, takisto prvou triedou. Ľudia totiž radšej pracujú s dotazníkom, ktorý má atraktívny vzhľad a naznačuje profesionálnu zdatnosť organizátora prieskumu.

Podiel odpovedí možno zvýšiť, ak prieskum odporúča nejaká oficiálna, známa organizácia. V prieskume poštou môže byť odporúčanie v liste, ktorý je priložený

k zasielaným materiálom. Odporúčací list by mal obsahovať aj logo odporúčajúcej organizácie a podpis jej vysokého predstaviteľa. Odporúčanie je mimoriadne dôležité hlavne v prieskume organizácií. Napríklad pri prieskume nemocníc je dôležité priložiť k zaslaným materiálom odporúčanie napríklad Ministerstva zdravotníctva alebo od Slovenskej lekárskej komory a pod.

Pri prieskume domácností môže byť užitočné priložiť k zaslaným materiálom odmenu pre respondentov, tzv. stimul (*incentive*). Možno napríklad zaslať úplne novú 20 centovú mincu a pod. Možno použiť aj iné stimuly, napr. magnetku, vreckový kalendár a pod., podobné stimuly sú ale obyčajne atraktívne len pre určité skupiny respondentov.

Skúsenosti ukazujú, že treba formulovať dotazníky, ktorých vyplnenie nepresiahne 30 minút, inak sa riziko neodpovedania výrazne zvyšuje.

4 Možnosti eliminácie vplyvu neodpovedania

Napriek vhodným preventívnym opatreniam vždy treba počítať s istou mierou neodpovedania. Je známych viacero možností minimalizácie vplyvu neodpovedania na presnosť odhadovania. Prvou z nich je použitie pomocného výberového súboru chýbajúcich odpovedí. Všimneme si použitie pomocného výberového súboru v prieskumoch prostredníctvom pošty.

Metodológia spočíva v predbežnom zhromaždení dát náhodným vyberaním, nasledovanom rozdelením základného súboru na dve stratá.[6] Jednotky zo základného súboru odpovedajúcich tvoria prvé strátum, jednotky zo súboru neodpovedajúcich tvoria druhé strátum. Jednotky v náhodnom výbere, ktoré odpovedali, sú z prvého strata, tie ktoré neodpovedali, sú z druhého strata. Potom sa realizuje náhodné vyberanie z jednotiek z druhého strata a od vybratých jednotiek sa získajú odpovede prostredníctvom intenzívneho úsilia (*intensive effort*)). Konečný odhad je váženou kombináciou odhadov v jednotlivých stratách. Dvojetapovú výberovú procedúru možno všeobecne využiť aj v situáciách, v ktorých sú vychýlenia spôsobené inými príčinami, nie neodpovedaním.[7]

Ďalšou možnosťou je využitie výberových váh, ktoré okrem eliminácie vplyvu neodpovedania umožňujú eliminovať aj vplyv nepokrytia[8] cieľového základného súboru. Túto možnosť si všimneme podrobnejšie.

---

[6] Stratifikácia je rozdelenie základného súboru na navzájom sa vylučujúce a základný súbor úplne pokrývajúce podsúbory (stratá alebo vrstvy), ktoré sa vzhľadom na skúmanú premennú považujú za viac homogénne ako celý základný súbor. Stratifikované vyberanie (odber vzoriek) je vyberanie zo stratifikovaného základného súboru tak, že určená časť výberu je vybratá z rôznych strát a z každého strata je vybratá aspoň jedna jednotka. Keď stratifikácia predchádza realizácii vyberania, ide o stratifikáciu (stratifikáciu a priori), keď sa stratifikácia používa v etape extrapolácie výsledkov a je založená na pomocných informáciách, ide o poststratifikáciu (stratifikáciu a posteriori).

[7] Podrobnejšie aj s ilustratívnym príkladom pozri [12].

[8] Nepokrytie 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.

4.1 Využitie výberových váh

V mnohých rozsiahlych prieskumoch sa na zohľadnenie vplyvu stratifikácie a skupinového vyberania, neodpovedania a nepokrytia používajú výberové váhy. Všimneme si základné váhy, ich úpravu vzhľadom na neodpovedanie pomocou váhových tried a vzhľadom na nepokrytie pomocou poststratifikačnej úpravy. Uvedené postupy možno považovať za súčasť všeobecného prístupu k váženiu, ktorý sa nazýva kalibračný (*calibration approach to weighting*). Tento prístup je podrobne opísaný v (Särndal, Lundström 2005, s. 19 – 151), [8].

Ďalej budeme predpokladať, že výberová báza nepokrýva celý cieľový základný súbor. Budeme uvažovať o jednoduchom náhodnom vyberaní rozsahu n.

Nech N_F je počet jednotiek (neduplicitných) vo výberovej báze a N je počet jednotiek (neduplicitných) v cieľovom základnom súbore. Keď $N_F < N$, hovoríme o chybe z nepokrytia, pretože $N - N_F$ osôb, ktoré nie sú vo výberovej báze, má nulovú pravdepodobnosť dostať sa do výberu. Neuvažujeme o neodpovedaní.[9] Strednú hodnotu premennej x vo výberovej báze označíme μ_F. Bodový odhad $\overline{X}$ je nevychýleným bodovým odhadom μ_F, ale môže byť vychýleným bodovým odhadom μ, s vychýlením

$$B_F = \gamma_{NF}(\mu_F - \mu_{NF})$$

kde $\gamma_{NF} = 1 - N_F/N$ je podiel cieľového základného súboru, ktorý nie je pokrytý výberovou bázou a μ_{NF} je stredná hodnota jednotiek z cieľového základného súboru, ktoré nie sú vo výberovej báze.

Uvažujme teraz o neodpovedaní. Podiel výberovej bázy, ktorá patrí do základného súboru odpovedajúcich označíme γ_R. Jednotky vo výbere, ktoré odpovedali, možno považovať za výber zo základného súboru odpovedajúcich. Výberový priemer je nevychýleným bodovým odhadom strednej hodnoty základného súboru odpovedajúcich μ_R, ale môže byť vychýleným bodovým odhadom μ_F, s vychýlením

$$B_R = \gamma_{NR}(\mu_R - \mu_{NR})$$

kde $\gamma_{NR} = (1 - \gamma_R)$ je podiel výberovej bázy, ktorá patrí do súboru neodpovedajúcich.

Vychýlenie z nepokrytia B_F a vychýlenie z neodpovedania B_R možno jednoducho spočítať (Levy, Lemeshow, 2008, s. 496), [5]. Úpravy po prieskume sú navrhnuté na redukciu účinkov každého druhu vychýlenia separátne alebo kumulatívne. Napríklad poststratifikácia, ktorá je navrhnutá na redukciu vychýlenia z nepokrytia, môže v určitom rozsahu redukovať aj vychýlenie spôsobené neodpovedaním.

Základné váhy (*base weights*, *design weights*) w_{Bi} sú odvodené z plánu výberového skúmania. Ostatné faktory sa považujú za faktory úpravy základných váh. Váhy

[9] Predpokladáme, že odpovie 100 % osôb z výberu.

w_{NRi} sa považujú za faktory úpravy vzhľadom na neodpovedanie. Váhy w_{NCi} sa považujú za faktory kompenzácie nepokrytia. Niekedy sa nazývajú aj faktory poststratifikačnej úpravy (*poststratification adjustment factors*). Tento faktor má často za úlohu aj redukovať rozptyl odhadov.

Konečná váha pre *i*-tu jednotku vo výbere je

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

Váhy, o ktorých uvažujeme, majú tieto vlastnosti:

1. $\sum_{i=1}^{n_R} w_{Bi} = N_R$

kde n_R je rozsah výberu zo základného súboru odpovedajúcich;

2. $\sum_{i=1}^{n_R} w_{Bi} \cdot w_{NRi} = N_F$

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

4.1.1 Základné váhy

Základná váha w_{Bi} jednotky *i* je obrátená hodnota pravdepodobnosti vybratia *i*-tej jednotky do výberu.

Príklad 1 V menšom meste sa mestské zastupiteľstvo potrebuje rozhodnúť, či má udeliť povolenie na výstavbu zábavného parku s množstvom herní. Zaujíma ich názor obyvateľov mesta. Výberovým skúmaním chcú odhadnúť podiel obyvateľov mesta, ktorí súhlasia s udelením povolenia. Realizuje sa stratifikovaný, telefonický prieskum. Uvažuje sa o dvoch stratách – muži a ženy. Predpokladajme, že vo výberovej báze je 2 800 telefónnych čísiel mužov a 3 200 telefónnych čísiel žien a že rozsah výberu z prvého strata je 112 mužov, rozsah výberu z druhého strata je 160 žien.[10] Z vybratých 272 osôb odpovedalo n_R = 116 osôb, z toho 40 mužov a 76 žien.

Je známe, že hodnotu nevychýleného[11] bodového odhadu podielu základného súboru $\hat{\pi}_{str}$ možno vypočítať podľa vzťahu:

$$\hat{\pi}_{str} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{N_h}{N \cdot n_h} \cdot x_{hi} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} \cdot x_{hi}$$

kde *H* je počet strát,
n_h – rozsah výberu zo strata *h*,
N_h – rozsah strata *h*,
x_{hi} – hodnota binárnej premennej *x* (nadobúda dve hodnoty – 1, keď respondent sú-

[10] Podrobnejšie o optimálnej alokácii výberu medzi stratá pozri napr. v [5], [6], [9], [14].
[11] V prípade úplného odpovedania a ideálneho pokrytia.

hlasí, a 0, keď nesúhlasí), *i*-teho respondenta vo výbere z *h*-teho strata,

$w_{hi} = \frac{N_h}{n_h}$ – výberová váha.

Výberová váha w_{hi} môže byť chápaná ako počet jednotiek v základnom súbore, reprezentovaných jednotkou (h, i) vo výbere.

V príklade zrejme pravdepodobnosť výberu muža je

$$p_{B1i} = \frac{112}{2800} = 0{,}04$$

a pravdepodobnosť výberu ženy je

$$p_{B2i} = \frac{160}{3200} = 0{,}05$$

Základné váhy sú

$$w_{B1i} = \frac{1}{p_{B1i}} = 25 \quad \text{a} \quad w_{B2i} = \frac{1}{p_{B2i}} = 20$$

Výsledky možno interpretovať takto. Každý muž vo výbere reprezentuje 25 mužov v základnom súbore a každá žena vo výbere reprezentuje 20 žien v základnom súbore. Suma váh v príklade je

$$\sum_{h=1}^{2} \sum_{i=1}^{n_{Rh}} w_{Bhi} = N_R$$

kde n_{Rh} je počet odpovedajúcich vo výbere z *h*-teho strata.

Keď dosadíme do posledného vzťahu, po jednoduchej úprave dostaneme
$40 \cdot 25 + 76 \cdot 20 = 2520$
čo je odhadovaný rozsah základného súboru odpovedajúcich.

4.1.2 Úprava základných váh vzhľadom na neodpovedanie

Nech p_{Ri} je pravdepodobnosť *i*-tej jednotky vo výbere, že odpovie, ktorá reprezentuje sklon (tendenciu) jednotky odpovedať v danom prieskume (sklon k odpovedaniu). Pravdepodobnosť, že *i*-ta jednotka bude vybratá a odpovie, je
$P(i \text{ bude vybratá a odpovie}) = P(i \text{ odpovie} \mid i \text{ bude vybratá}) \cdot P(i \text{ bude vybratá}) = p_{Ri} \cdot p_{Bi}$

Vychýlenie spôsobené neodpovedaním možno eliminovať vážením každej jednotky, ktorá odpovedala váhou

$$w_{NRi} = \frac{1}{p_{Ri}}$$

Na kompenzáciu neodpovedania a nerovnakých pravdepodobností vybratia jednotiek treba použiť váhy

$w_{Bi} \cdot w_{NRi}$

Pretože skutočný sklon k odpovedaniu pre každú jednotku je neznámy, treba ho odhadnúť. Schopnosť faktora úpravy neodpovedania redukovať vychýlenie spôsobené neodpovedaním je len taká veľká, ako dobre sa podarí odhadnúť p_{Ri}. Všimneme si niektoré možnosti odhadovania p_{Ri}.

Úprava pomocou váhových tried

Úprava pomocou váhových tried (*weighting class adjustment* – WCA) sa považuje za najjednoduchší prístup.

Výber sa delí na skupiny na základe premenných, ktorých hodnoty sú známe pre jednotky, ktoré by odpovedali, aj pre jednotky, ktoré by neodpovedali, pričom sa predpokladá, že tieto premenné sú spojené so sklonom k odpovedaniu. Napríklad pomer odpovedania (*response rate*) je často v korelácii s vekom a pohlavím respondentov. Keď je známy vek a pohlavie osôb, ktoré neodpovedali, potom všetky jednotky vo výbere možno klasifikovať do tried podľa veku a pohlavia – do váhových tried. Pomer odpovedania pre váhovú triedu sa považuje za sklon k odpovedaniu pre každú jednotku z triedy vo výbere. Úprava neodpovedania pre *i*-tu jednotku je potom jednoducho podiel

w_{Bi} / pomer odpovedania v triede, do ktorej *i*-ta jednotka patrí.

Uvažujme v metóde WCA o K triedach ($j = 1, 2,...K$). Pre j-tu triedu vážený pomer odpovedania (*weighted response rate*) je

$$RR_{w_j} = \frac{\sum_{i=1}^{n_{Rj}} w_{Bi}}{\sum_{i=1}^{n_j} w_{Bi}} \tag{1}$$

kde n_{Rj} je počet jednotiek vo výbere v triede j, ktoré odpovedali,

n_j – počet jednotiek vo výbere v triede j.

Potom pre každú jednotku i v triede j vo výbere, ktorá odpovedala, je faktor úpravy neodpovedania

$$w_{NRi} = \frac{1}{RR_{w_j}} \tag{2}$$

Možno ukázať (Levy, Lemeshow 2008, s. 503), [5], že príspevok k celkovému vychýleniu spôsobenému neodpovedaním pre odhad strednej hodnoty triedy j je

$$B_{R_j} = \pi_j \cdot \gamma_{NRj} \left(\mu_{Rj} - \mu_{NRj} \right)$$

kde π_j je podiel výberovej bázy, ktorá patrí do triedy j,

γ_{NRj} – očakávaný pomer neodpovedania pre jednotky v triede *j*,
μ_{Rj}, μ_{NRj} – v tomto poradí, stredné hodnoty jednotiek, ktoré odpovedia a ktoré neodpovedia v triede *j* [5].

Keď sa triedy nájdu tak, že rozdiely (μ_{Rj} - μ_{NRj}) sú veľmi malé, zvlášť pre veľké π_j, celkové vychýlenie z neodpovedania bude tiež veľmi malé.

Príklad 1 – pokračovanie 1. Vieme, že v prieskume zo 272 oslovených osôb vo výbere odpovedalo n_R = 116 osôb, z toho 40 mužov a 76 žien. To znamená, že celkový pomer odpovedania bol 42,65 %. Na úpravu váh vzhľadom na neodpovedanie budeme aplikovať metódu WCA. Na definovanie tried WCA použijeme rovnakú premennú ako v stratifikácii – pohlavie.

Podľa vzťahu (1), po jednoduchej úprave, dostaneme pre mužov:

$$RR_{w_1} = \frac{40 \cdot 25}{112 \cdot 25} \approx 0{,}35714$$

Podľa vzťahu (2) dostaneme váhu pre všetkých mužov vo výbere, ktorí odpovedali:

$$w_{NR1i} = \frac{1}{RR_{w_1}} = \frac{1}{0{,}35714} = 2{,}8$$

Podobne vypočítame váhu pre všetky ženy vo výbere, ktoré odpovedali:
$w_{NR2i} \approx 2{,}10526$

Po dosadení do vzťahu, ktorý charakterizuje vlastnosť 2, po jednoduchej úprave dostaneme:
$40 \cdot 25 \cdot 2{,}8 + 76 \cdot 20 \cdot 2{,}10526 = 2800 + 3200 = 6000$,
čo je rozsah výberovej bázy. Posledná rovnosť potvrdzuje, že systém váh je kalibrovaný[12] (*calibrated*) alebo konzistentný (*consistent*).

Sú známe aj iné metódy. Metódy úpravy založené na modeli umožňujú odhadovať sklon k odpovedaniu pre rozličné skupiny v základnom súbore. Získané hodnoty odhadov sa použijú na úpravu dát vzhľadom na neodpovedanie. Jedna z najrozšírenejších metód spočíva v aplikácii logistickej regresie na odhadovanie sklonu k odpovedaniu.[13]

4.1.3 Úprava základných váh vzhľadom na nepokrytie

Cieľový základný súbor obsahuje jednotky, ktoré chceme v prieskume obsiahnuť v čase, keď sa dáta zhromažďujú. Tento časový bod sa nazýva referenčný časový bod pre cieľový základný súbor (*reference time point for the target population*). Výberová báza sa obyčajne vytvára predtým, niekedy aj o rok skôr. Tento časový bod sa

[12] Alebo vyvážený.
[13] Podrobnejšie v [5].

nazýva referenčný časový bod pre výberovú bázu (*reference time point for the frame population*). Rozdiel medzi týmito časovými bodmi by mal byť čo najmenší, pretože riziko vzniku chýb pokrytia rastie s rastom tohto rozdielu. Rozlišujú sa tri typy chýb pokrytia: nepokrytie (*noncoverage*) alebo neúplné pokrytie (*undercoverage*), presahujúce pokrytie (*overcoverage*) a duplicitné zaradenie jednotiek (*duplicate listings*) (Särndal, Lundström 2005, s. 8 – 9), [8]. Aj posledné dva uvedené typy chýb pokrytia môžu byť zdrojom vychýlení v procese odhadovania.

Jednotky, ktoré sú v cieľovom základnom súbore, ale nie sú vo výberovej báze, vytvárajú nepokrytie alebo neúplné pokrytie. Napríklad pri prieskumoch firiem, tie, ktoré boli založené až po vytvorení výberovej bázy, patria do cieľového základného súboru, ale nie sú vo výberovej báze. V telefonickom prieskume domácností nemusia mať všetky domácnosti v základnom súbore telefón a pod. Pravdepodobnosť vybratia takýchto jednotiek sa rovná nule.

Jednotky, ktoré sú vo výberovej báze, ale nie sú v cieľovom základnom súbore, vytvárajú presahujúce pokrytie. Napríklad firmy, ktoré zanikli medzi dvoma referenčnými časovými bodmi, sú vo výberovej báze, ale nie sú v cieľovom základnom súbore. Pri telefonickom prieskume domácností v nejakej mestskej časti sa do výberovej bázy môžu omylom dostať aj telefónne čísla domácností, ktoré nie sú z tejto mestskej časti a pod. Ideálne je, keď sa zaniknuté firmy, resp. telefónne čísla z iných mestských častí, podarí identifikovať a vyradiť ešte pred začatím vyberania, často to však nie je možné.

Jednotky, ktoré sa vo výberovej báze vyskytujú viackrát, vytvárajú duplicitné zaradenie jednotiek. Napríklad jedna domácnosť má dve telefónne čísla, ktoré môžu byť obe použité v interview. Pri telefonickom prieskume sa odporúča opýtať sa respondenta, na koľkých telefónnych číslach je doma dosiahnuteľný. Získaná informácia sa potom dá využiť na opravu pravdepodobnosti výberu a následne na opravu základnej váhy.

Podrobnejšie si všimneme poststratifikačnú úpravu (PSA – *poststratification adjustment*) na elimináciu vplyvu nepokrytia. Respondenti vo výbere sa klasifikujú do tried, podobne ako v metóde WCA. V porovnaní s WCA sú tu dva rozdiely:

1. na vytvorenie tried stačia len informácie o jednotkách, ktoré odpovedali,
2. pri vytváraní tried je snaha, aby v triedach boli spolu jednotky s podobnými odpoveďami pokiaľ ide o skúmané premenné.

Okrem toho je nevyhnutné mať presnú informáciu o celkovom počte jednotiek v základnom súbore[14] pre každú triedu. Poststratifikačná úprava násobí váhu každej jednotky v triede konštantou tak, aby sa celková váha v každej triede rovnala známemu počtu jednotiek v príslušnej triede základného súboru.

Keď sú známe počty jednotiek v triedach cieľového základného súboru, potom PSA redukuje vychýlenie z nepokrytia. Keď sú známe len počty jednotiek v triedach výberovej bázy, potom môže PSA stabilizovať variabilitu odhadov. Keď časť originálnej výberovej bázy nebola zahrnutá do vyberania, možno známe počty jednotiek

[14] Vo výberovej báze alebo v cieľovom základnom súbore.

v triedach výberovej bázy použiť na redukciu vychýlenia z nepokrytia spojeného s použitím redukovanej výberovej bázy. Napokon, PSA môže byť použitá namiesto, alebo v kombinácii s úpravou neodpovedania na redukciu vychýlenia spôsobeného neodpovedaním.

Pri implementácii metódy treba rozdeliť jednotky vo výbere do *L* tried.[15] Tieto triedy sa môžu líšiť od tried, ktoré sa použijú pri úprave neodpovedania. Napríklad pri prieskumoch domácností sa ale často použijú, rovnako ako pri úprave neodpovedania, premenné vek a pohlavie, pretože počty jednotiek v týchto triedach v cieľovom základnom súbore sú známe z makrocenza.

Nech N_l, $l = 1, 2, ...L$, sú počty jednotiek v triedach v cieľovom základnom súbore. Zrejme:

$$\sum_{l=1}^{L} N_l = N$$

Potom faktor kompenzácie nepokrytia pre *i*-tu jednotku v triede *l* je:

$$w_{NCli} = \frac{N_l}{\sum_{i=1}^{n_{Bl}} w_{Bli} \cdot w_{NRli}} \tag{3}$$

S touto úpravou suma poststratifikovaných, upravených váh pre triedu *l* je:

$$\sum_{i=1}^{n_{Bl}} w_{Bli} \cdot w_{NRli} \cdot w_{NCli} = N_l \tag{4}$$

Keď N_l sú počty jednotiek v triedach cieľového základného súboru, potom, keď sú triedy vhodne určené, PSA môže redukovať vychýlenie z nepokrytia. Možno ukázať, že príspevok k celkovému vychýleniu z nepokrytia triedy *l* je (Levy, Lemeshow, 2008, s. 506 – 507), [5]:

$$B_{NCl} = w_l \cdot \gamma_{NCl} (\mu_{Cl} - \mu_{NCl})$$

kde $w_l = \frac{N_l}{N}$ je podiel základného súboru, ktorý patrí do triedy *l*,

γ_{NCl} – pomer nepokrytia (*noncoverage rate*) pre jednotky v triede *l*,
μ_{Cl} , μ_{NCl} – stredné hodnoty pokrytej a nepokrytej časti základného súboru v triede *l*.

Je zrejmé, že najlepšia stratégia tvorby PSA tried je minimalizovať (μ_C - μ_{NC}) v triedach. V metóde PSA sa v každej triede priraďuje vážená stredná hodnota pokrytého základného súboru všetkým jednotkám nepokrytého základného súboru. Keď sa triedy definujú tak, že rozdiely (μ_{Cl} - μ_{NCl}) sú malé, zvlášť, keď je w_l veľké, bude celkové vychýlenie z nepokrytia tiež malé.

[15] Pri rešpektovaní zásady úplnosti a jednoznačnosti.

Často aspoň niektoré premenné, ktoré sú vhodné pre WCA, sú vhodné aj na redukciu vychýlenia spôsobeného nepokrytím, aj na redukciu rozptylu odhadov. V takom prípade bude použitie týchto premenných redukovať súčasne vychýlenie spôsobené neodpovedaním, nepokrytím, aj rozptyl odhadov.

Príklad 1 – pokračovanie 2. Na úpravu váh vzhľadom na nepokrytie použijeme metódu PSA. Na vytvorenie tried použijeme dve premenné – pohlavie a vek respondentov. Vo výbere sa u všetkých odpovedajúcich zistil okrem ich názoru na povolenie aj ich vek, čo umožnilo zaradiť ich do tried v tabuľke č. 1. Z posledného makrocenza je známe rozdelenie početností v cieľovom základnom súbore podľa týchto premenných (v tabuľke č. 1 dáta v okrúhlych zátvorkách). Vypočítame faktory kompenzácie nepokrytia a celkové váhy pre jednotlivé triedy PSA (indexy tried PSA sú v tabuľke č. 1 v hranatých zátvorkách). Potom vypočítame celkové váhy pre jednotlivé triedy PSA.

Tab. č. 1

Rozdelenie podľa pohlavia a veku

Pohlavie Vek	Muž	Žena	Spolu
18 – 23	24 (800) [1]	14 (860) [4]	38 (1660)
24 – 62	10 (1810) [2]	50 (2210) [5]	60 (4020)
63 a viac	6 (412) [3]	12 (386) [6]	18 (798)
Spolu	40 (3022)	76 (3456)	116 (6478)

Pre mužov vo veku 18 – 23 rokov po dosadení do (3), po jednoduchej úprave dostaneme:

$$w_{NC1i} = \frac{800}{24 \cdot 25 \cdot 2{,}8} \approx 0{,}47619$$

Podobne dostaneme: $w_{NC2i} \approx 2{,}58571$, $w_{NC3i} \approx 0{,}98095$, $w_{NC4i} \approx 1{,}45893$, $w_{NC5i} \approx 1{,}04975$, $w_{NC6i} \approx 0{,}76393$.

Nakoniec vypočítame celkové váhy pre jednotlivé triedy PSA podľa vzťahu:
$w_{li} = w_{Bli} \cdot w_{NRli} \cdot w_{NCli}$

Napríklad pre $l = 1$, dostaneme:
$w_{1i} = w_{B1i} \cdot w_{NR1i} \cdot w_{NC1i} = 25 \cdot 2{,}8 \cdot 0{,}47619 \approx 33{,}33$

Podobne dostaneme: $w_{2i} \approx 181$, $w_{3i} \approx 68{,}6665$, $w_{4i} \approx 61{,}42854$, $w_{5i} \approx 44{,}19993$, $w_{6i} \approx 32{,}16669$.

Podľa vzťahu (4) možno overiť kalibráciu systému váh. Pre $l = 1$ napríklad po jednoduchej úprave dostaneme:
$33{,}33 \cdot 24 = 800$

čo je početnosť triedy $l = 1$ v PSA v cieľovom základnom súbore. Podobne sa potvrdí správnosť výpočtov váh aj v ostatných triedach PSA. Systém váh je kalibrovaný.

Konečné váhy možno použiť na odhadovanie. Postup budeme ilustrovať na uvedenom príklade.

Príklad 1 – pokračovanie 3. Predpokladajme, že respondenti odpovedali tak, ako je to v tab. č. 2. Na základe znalosti celkových váh odhadneme podiel obyvateľov mesta, ktorí sú za udelenie povolenia.

Tab. č. 2

Rozdelenie respondentov, ktorí súhlasia, podľa pohlavia a veku

Pohlavie Vek	Muž	Žena	Spolu
18 – 23	23	4	27
24 – 62	3	20	23
63 a viac	2	11	13
Spolu	28	35	63

Hodnotu odhadu vypočítame podľa vzťahu:

$$\hat{\pi} = \frac{1}{\sum_{l=1}^{L}\sum_{i=1}^{n_{Rl}} w_{li}} \sum_{l=1}^{L}\sum_{i=1}^{n_{Rl}} w_{li} \cdot x_{li} \qquad (5)$$

Suma váh v menovateli vzťahu (5) vyjde 6478, čo je rozsah cieľového základného súboru. Po dosadení do vzťahu (5) a jednoduchých úpravách dostaneme:

$$\hat{\pi} = \frac{1}{6478}(33{,}33 \cdot 23 + 181 \cdot 3 + 68{,}6665 \cdot 2 + 61{,}42854 \cdot 4 + 44{,}19993 \cdot 20 +$$

$$+ 32{,}16669 \cdot 11) \approx 0{,}45237$$

Odhadujeme, že za udelenie povolenia je len približne 45,24 % dospelých obyvateľov mesta. Väčšina dospelých obyvateľov mesta nie je za udelenie povolenia. Pri výpočte bol okrem stratifikácie zohľadnený aj vplyv neodpovedania a nepokrytia. Získal sa podklad pre rozhodnutie o neudelení povolenia.

5 Pomocné informácie

Pri stratifikácii aj pri zohľadnení vplyvu neodpovedania a nepokrytia v príklade sme využili tzv. pomocné informácie (*auxiliary information*). Konkrétne bola použitá pomocná premenná (*auxiliary variable*) pohlavie s dvoma hodnotami – muž a žena. Minimálna požiadavka, aby bolo možné nejakú premennú označiť za pomocnú, je, aby boli jej hodnoty známe pre všetky jednotky vo výbere (pre jednotky, ktoré odpovedali, aj ktoré neodpovedali).

Pomocné informácie možno využiť v etape tvorby plánu výberového skúmania, aj v etape odhadovania, pri konštrukcii bodových odhadov.[16]

Hodnoty pomocných premenných možno často získať z rozličných registrov, napríklad z obchodného registra alebo z registra obyvateľov.

V prezentovaných a podobných analýzach je veľmi dôležitá voľba vhodných pomocných premenných. Metódy na identifikáciu najvhodnejších pomocných premenných s ohľadom na vychýlenie bodových odhadov sú podrobne opísané v (Särndal, Lundström 2005, s. 109 – 133), [8].

Záver

Často sa proces váženia zdá byť „čiernou skrinkou". Konečné váhy niektorých pozorovaní môžu byť relatívne malé – napríklad 20, niektoré oveľa väčšie – napríklad 3000. V uvedenom príklade je najmenšia konečná váha 32,16669, najväčšia je 181. V príspevku sme ukázali, ako je to možné.

Keby sme pri odhadovaní podielu obyvateľov, ktorí súhlasia s udelením povolenia, neuvažovali o stratifikácii, ani o vplyve neodpovedania a nepokrytia, mohli by sme dostať výrazne odlišný výsledok. Predpokladajme napríklad, že by sme jednoduchým náhodným vyberaním získali 116 odpovedí, z ktorých by bolo 63 súhlasných s vydaním povolenia.[17] Keby sme odhadli podiel obyvateľov, ktorí súhlasia s udelením povolenia hodnotou výberového podielu, dostali by sme výsledok: $63/116 \approx 0{,}54310$. Podľa tohto výsledku väčšina obyvateľov je za vydanie povolenia. Získal sa podklad pre opačné rozhodnutie. Tu sa však neberú do úvahy váhy, ktoré udávajú, koľko jednotiek v cieľovom základnom súbore reprezentuje každá odpovedajúca jednotka vo výbere.

Možno oprávnene predpokladať, že hodnota odhadu získaná pomocou konečných váh, ktoré okrem použitej výberovej schémy zohľadňujú aj vplyv neodpovedania a nepokrytia, reálnejšie odráža skutočnosť.

Keď je variabilita váh príliš veľká, môže to spôsobovať zväčšenie rozptylu odhadov. Vtedy môže byť užitočné zastrihnutie váh (*weight trimming*).[18]

Poznamenávame, že kroky procesu konštrukcie váh sú pomerne subjektívne. Treba rozhodnúť, či základné váhy treba upraviť vzhľadom na neodpovedanie a nepokrytie[19], ako vytvoriť váhové triedy v metóde WCA a poststratifikačné triedy v metóde PSA, aké dáta použiť v procese úpravy váh, či sa váhy zastrihnú atď. Niekedy môže byť dokonca aj výpočet základných váh subjektívny. Napríklad v telefonickom prieskume nie sú k dispozícii informácie o počte telefónnych čísiel v domácnosti a nie je možné ich využiť pri výpočte pravdepodobností vybratia. Existuje však všeobecná dohoda, že pri odhadovaní parametrov konečného základného súboru sa majú vždy použiť váhy (Levy, Lemeshow, 2008, s. 514), [5].

[16] Podrobnejšie v [2].

[17] Tak ako je to v tabuľke č. 1 a č. 2.

[18] Podrobnejšie o tejto problematike pozri v [5].

[19] Niekedy môže byť vplyv úpravy na odhad veľmi malý. Vtedy nemá zmysel robiť tieto úpravy.

Literatúra

[1] COCHRAN, W. G. (1977): *Sampling Techniques*. New York : J. Wiley and Sons. ISBN 0-471-16240-X.

[2] FULLER, W. A. (2009): *Sampling Statistics*. USA: Wiley. ISBN 978-0-470-45460-2.

[3] GRAIS, B.: *Méthodes statistiques* (2000). Paris : Dunod. ISBN 2-10-001264-9.

[4] GROVES, R. M. – FOWLER, F. J. Jr. – COUPER, M. P. – LEPKOWSKI, J. M. – SINGER, E. – TOURANGEAU, R. (2004): *Survey Methodology*. USA: Wiley. ISBN 0-471-48348-6.

[5] LEVY, P. S. – LEMESHOW, S. (2008): *Sampling of Populationas. Methods and Applications.* Fourth Edition. USA: Wiley. ISBN 978-0-470-04007-2.

[6] LOHR, S. L. (1999): *Sampling: Design and Analysis*. Duxbury Press. ISBN 0-534-35361- 4.

[7] SÄRNDAL, C.-E. – SWENSSON, B. – WRETMAN, J. (2003): *Model Assisted Survey Sampling.* New York: Springer. ISBN 0-387-40620-4.

[8] SÄRNDAL, C.-E. – LUNDSTRÖM, S. (2005): *Estimation in Surveys with Nonresponse.* USA: Wiley. ISBN 0-470-01133-5.

[9] TEREK, M. – HRNČIAROVÁ, Ľ. (2008): *Výberové skúmanie.* Bratislava : Ekonóm. ISBN 978-80-225-2440-7.

[10] TEREK, M. (2013 – 1): *Interpretácia štatistiky a dát. Druhé, doplnené vydanie.* Košice : Equilibria ISBN 978-80-8143-100-5.

[11] TEREK, M. (2013 – 2): *Interpretácia štatistiky a dát. Podporný učebný materiál. Druhé, doplnené vydanie.* Košice: Equilibria. ISBN 978-80-8143-101-2.

[12] TEREK, M. (2013 – 3): Problém neodpovedania v štatistických prieskumoch prostredníctvom pošty. In: *Forum statisticum slovacum* 6/2013**,** ISSN 1336-7420.

[13] TILLÉ, Y.: *Théorie des sondages* (2001). Paris : Dunod. ISBN 2 10 005484 8.

[14] TRYFOS, P. (1996): *Sampling Methods for Applied Research*. USA: Wiley. ISBN 0-471-04727-9.

[15] 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.