

Valér Taragel

PROBLÉM NEODPOVEDANIA VO VÝBEROVÝCH SKÚMANIACH¹

Úvod

Nonresponse problem alebo problém neodpovedania je situácia, ktorá vzniká najčastejšie pri výberovom skúmaní na báze štatistický prieskumov. Jedná sa o problém, kedy štatistická jednotka neposkytne žiadne alebo iba časť informácií. Tento problém sa stal vypuklejším za posledných cca 20 rokov. Klesá ochota respondentov zúčastňovať sa rôznych prieskumov, a odpovedať v nich. Počas tohto obdobia s príchodom nových komunikačných médií a celkového zrýchlenia toku informácií, významne narástol počet vykonávaných prieskumov.

1 PROBLÉM NEODPOVEDANIA

Neodpovedanie je bežný jav, ale ktorý je nežiaduci v štatistickom prieskume. Neodpovedanie dokáže vážne znížiť kvalitu výsledkov prieskumu. Dôvodom je skreslenosť/vychýlenosť (*bias*), ktorú so sebou prináša nekompletný výberový súbor.

Vo všeobecnosti môžeme rozlišovať 2 typy neodpovedania:

- *Neodpoveď jednotky (unit nonresponse)* jedná sa o prípad, kedy nemáme žiadne informácie od štatistickej jednotky (jednotka neodpovedala/odovzdala prázdny dotazník/dotazník neodpovedala)
- *Čiastočné neodpovedanie jednotky (item nonresponse)* jedná sa o prípad, keď štatistická jednotka poskytla iba určité informácie, nie všetky (časť dotazníku/niektoré otázky ostali nevyplnené)

Neodpovedanie môže mať rôzne príčiny:

- *Nenadviazanie kontaktu (No-contact)* požiadavka na zodpovedanie nezastihne, alebo nie je doručená štatistickej jednotke. V tomto prípade je problém na strane výskumníka. Je potrebné preveriť, či kontakty na štatistickú jednotku sú aktuálne, či je štatistická jednotka aktívna (napr. či je zapísaná v OR SR, či podniká, či študent ešte študuje atď.)
- *Odmietnutie (Refusal)* v tomto prípade štatistická jednotka odmietla poskytnúť o sebe informácie úplne, alebo čiastočne. Je vhodné zistiť o štatistickej jednotke pomocné a doplnujúce informácie ktoré by eventuálne mohli pomôcť objasniť dôvod neodpovedania, a tak znížiť vychýlenosť skúmania

¹ Tento príspevok vznikol s príspevom grantovej agentúry VEGA v rámci projektu číslo 1/0092/15: Moderné prístupy k navrhovaniu komplexných štatistických prieskumov.

- *Neschopnosť odpovedať (Not-able)* štatistická jednotka nie je schopná odpovedať, buď úplne alebo čiastočne. Problém môže byť vo formulácii otázky, nesprávne zvolenej otázke alebo iné technické príčiny prečo štatistická jednotka nemohla odpovedať (výpadok siete pri telefonickom interview, nesprávne odoslaný dotazník atď.)

1.1 Popisné charakteristiky neodpovedania

Pri analýze neodpovedania a skúmaní príčin sa využíva niekoľko popisných charakteristík pre tento jav:

- *Miera odpovedania (Response rate)* vyjadruje podiel odpovedí, resp. podiel štatistických jednotiek ktoré poskytli všetky informácie z počtu všetkých dopytovaných štatistických jednotiek.
- *Miera spolupráce (Cooperation rate)* vyjadruje podiel odpovedí zo všetkých dopytovaných štatistických jednotiek. Do tejto miery spadajú aj štatistické jednotky ktoré neposkytli všetky informácie.
- *Miera odmietnutia (Refusal rate)* vyjadruje podiel štatistických jednotiek ktoré odmietli odpovedať, alebo prerušili odpovedanie.
- *Miera dostupnosti (Contact rate)* vyjadruje podiel štatistických jednotiek ktoré sa podarilo skontaktovať.

1.2 Vychýlenosť odhadov v dôsledku neodpovedania

Úlohou štatistického zisťovania je, získať informácie o neznámych charakteristikách základného súboru. Tieto charakteristiky sa nazývajú parametre a na základe informácií získaných z výberového súboru sa snažíme úsudky o ňom indukovať na základný súbor. Jedná sa najčastejšie o parametre ako napr. stredná hodnota, podiel, alebo úhrn. Najväčším problémom neodpovedania je vychýlenie resp. skreslenie odhadov.

Uvažujme príklad, v ktorom by sme zisťovali priemernú hodnotu telesnej váhy na určitej populácii. Pomocou náhodného vyberania by sme získali výberový súbor o počte 100 respondentov. Muži aj ženy by boli v tomto výberovom súbore zastúpené vďaka náhodnému výberu približne v rovnako ako v základnom súbore. Výberový súbor by obsahoval 50 mužov a 50 žien. Po zodpovedaní otázok by sme zistili že odpovedalo iba 80 účastníkov výberového skúmania. Odpovede od týchto účastníkov by sme spracovali a vyhodnotili. Získali by sme tak odhad priemernej telesnej váhy. Avšak ak by z určitého dôvodu bolo daných 20 účastníkov výberového skúmania ktorí neodpovedali iba mužského alebo iba ženského pohlavia, získali by sme tak značne nadhodnotený alebo podhodnotený odhad priemernej telesnej váhy.

Čím je miera neodpovedania vyššia, tým je väčšie vychýlenie odhadov.

Uvažujme že odhadujeme strednú hodnotu μ premennej x v konečnom základnom súbore rozsahu N .

N_1 - počet jednotiek v základnom súbore, ktoré by odpovedali

N_2 - počet jednotiek v základnom súbore, ktoré by neodpovedali ($N_2 = N - N_1$)

μ_1 - stredná hodnota súboru N_1 jednotiek, ktoré by odpovedali

μ_2 - stredná hodnota súboru N_2 jednotiek, ktoré by neodpovedali

$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N}$ je stredná hodnota premennej x v celom základnom súbore rozsahu N .

Pri náhodnom výbere n jednotiek v tomto prípade odhadujeme μ_1 a nie μ . Ak náhodný výber obsahuje iba n_1 jednotiek, ktoré odpovedali a $\bar{X}$ je výberový priemer týchto n_1 jednotiek, potom:

$$E(\bar{X}) = \mu_1$$

a vychýlenie $\bar{X}$ je

$$B(\bar{X}) = \mu_1 - \mu = \frac{N_2}{N}(\mu_1 - \mu_2)$$

Z posledného vzťahu možno konštatovať že, vychýlenie spôsobené neodpovedaním je nezávislé od n_1 a nemožno ho zredukovať zväčšením rozsahu výberu n . Je ale možné ho zredukovať zmenšením podielu $\frac{N_2}{N}$ jednotiek ktoré neodpovedali.

Možno teda tvrdiť že, efekt ktorý spôsobí neodpovedanie závisí od podielu jednotiek, ktoré neodpovedali a od rozdielu medzi strednými hodnotami skupiny jednotiek ktoré odpovedali, a skupiny jednotiek ktoré neodpovedali. Avšak hodnoty N_2 , n_1 a n_2 nepoznáme.

Miera neodpovedania sa však veľmi líši v závislosti od prieskumu. Každá štatistická jednotka má za určitých podmienok sklon resp. existuje nenulová pravdepodobnosť že neodpovie. Túto pravdepodobnosť môžeme označiť ako ρ_i . Vychýlenosť môžeme potom vyjadriť aj ako:

$$B(\bar{X}) \approx \frac{\sigma_{yp}}{\bar{\rho}}$$

Pričom σ_{yp} je kovariancia medzi skúmanou premennou pre y a pravdepodobnosťou ρ_i a $\bar{\rho}$ je priemerná pravdepodobnosť neodpovedania.

Výraz ρ_i je náhodná premenná a vychýlenosť $B(\bar{X}) \approx \frac{\sigma_{yp}}{\bar{\rho}}$ je funkciou toho ako vzájomne koreluje pravdepodobnosť neodpovedania a hodnoty skúmanej premennej.

1.3 Možnosti riešenia problému neodpovedania

Ak po ukončení fázy zberu údajov zistíme, že sa v prieskume vyskytli prípady neodpovedania máme niekoľko možností ako tento problém riešiť. Veľmi účinné je aj

zamerať sa na prevenciu tohto problému, a navrhnúť plán výberového skúmania s ohľadom na tento problém, ale ak je prieskum už vo fáze ukončenia zberu dát je možné aplikovať nasledovné metódy:

- *Imputovanie údajov* – Metóda spočíva v nahradení chýbajúcich hodnôt premenných blízkymi hodnotami. Imputovanie sa odporúča použiť iba prípade čiastočného neodpovedania štatistickej jednotky a následne použiť metódu výberových váh pre prípady úplného neodpovedania štatistickej jednotky.
- *Modifikácia výberového súboru pomocou poststratifikácie* – Táto metóda umožňuje korigovať štruktúru výberového súboru porušenú neodpovedaním. Ako riadiaci znak stratifikácie sa odporúča použiť premennú ktorá rozdelenie jasne deformované problémom neodpovedania a zároveň je silno korelovaná so skúmanými premennými.
- *Použitie pomocného výberového súboru – chýbajúcich odpovedí* – Metóda spočíva v dodatočnom získaní odpovedí od štatistických jednotiek ktoré boli náhodne vybrané z množiny štatistických jednotiek ktoré neodpovedali. Takto možno získať nevychýlený odhad.
- *Využitie výberových váh* – Táto metóda spočíva v úprave základných váh odpovedí štatistických jednotiek vzhľadom na ich neodpovedanie alebo vyskytujúci sa efekt nedostatočného pokrytia.

2 VÝBEROVÉ SKÚMANIE A JEHO KLÚČOVÉ BODY

Cieľom štatistického prieskumu je získať informácie o vlastnostiach určitých parametrov základného súboru. V rámci jedného prieskumu sa zisťujú informácie o viacerých parametroch. Množstvo prieskumov sa vykonáva pravidelne, a je tak možné skúmať dynamiku týchto parametrov v čase.

Prieskumy sa vykonávajú z toho dôvodu, že zistené informácie nie je možné získať iným spôsobom (neexistuje databáza, ani iný zdroj).

Prvým krokom pri tvorbe prieskumu je stanovenie jasných a dosiahnuteľných cieľov. Následne je potrebné vykonať množstvo príprav a rozhodnutí ako napr. zvoliť vhodnú metódu zberu dát, formu otázok/dotazníku, metóda výberu vhodných štatistických jednotiek, anketári musia byť riadne zaškolení, atď.

Všeobecne možno tento postup prípravy návrhu výberového skúmania zhrnúť do 4 krokov:

- Štúdium relevantných informácií o téme
- Definovanie predmetu skúmania
- Definovanie základného súboru výberovej bázy
- Formulácia plánu výberového skúmania

Štatistické jednoty vyberáme z tzv. cieľového základného súboru (*target population*). Jedná sa teda o súbor o ktorom chceme robiť indukčné úsudky. Môžu to byť napr. študenti určitej školy. Konkrétny výber potom robíme z tzv. výberovej bázy (*frame*

population). Jedná sa o zoznam, v ideálnom prípade všetkých štatistických jednotiek z cieľového základného súboru.

2.1 Zdroje chýb vo výberovom skúmaní

Pri vyhodnocovaní výberových skúmaní/prieskumov sa môžeme stretnúť s tromi typmi chýb:

- *Výberová chyba (sampling error)* jedná sa o chybu, kde bol nevhodne zvolený výberový súbor.
- *Chyba neodpovedania (nonresponse error)* jedná sa o chybu, kde nie všetky štatistické jednotky poskytli potrebné informácie.
- *Chyba pokrytia (coverage error)* jedná sa o chybu, kde neboli vo výberovom súbore zahrnuté všetky štatistické jednotky ktoré v ňom mali byť.

Chyba pokrytia (*coverage error*)

Chyba pokrytia sa môže v prieskume vyskytnúť v troch formách:

- Nedostatočné pokrytie (*undercoverage*)
- Prekrytie (*overcoverage*)
- Duplicita (*duplicate listings*)

V prípade duplicit sa jedná o chybu kedy jedna štatistická jednotka odpovedala dva krát alebo jej odpoveď bola do prieskumu zahrnutá viac krát.

Efekt nedostatočného pokrytia spôsobujú štatistické jednotky, ktoré sú v cieľovom základnom súbore ale chýbajú vo výberovej báze. Ako príklad môžeme uviesť prieskum ktorého cieľovým základným súborom sú všetky právnické jednotky zapísané v ORSR v Bratislavskom kraji, ktoré majú určitý predmet podnikania. V procese navrhovania výberového skúmania je určitý časový interval medzi definovaním základného súboru, vytvorením výberovej bázy a realizáciu získavania informácií. Tento časový interval by mal byť čo najmenší z toho dôvodu že, sa v čase môže významne zmeniť cieľový základný súbor. Napr. niektoré právnické osoby môžu zaniknúť, iné vzniknúť, a iné môžu zmeniť predmet podnikania.

Efekt prekrytia vzniká ak štatistické jednotky ktoré sú obsiahnuté vo výberovej báze, ale nenachádzajú sa v cieľovom základnom súbore.

Efekt prekrytia aj nedostatočné prekrytie je možné znížiť tým že, sa zníži čas medzi tvorbou výberovej bázy a štatistickým zisťovaním. Efekt nedostatočného pokrytia je taktiež možné znížiť využitím výberových váh.

Výberová chyba a chyba neodpovedania

Výberová chyba v ponímaní výberového skúmania je chyba, ktorá je spôsobená že hodnoty premenných boli zisťované iba pre výberový súbor a nie pre celý základný súbor. Ak by boli zisťované pre celý základný súbor, výberová chyba by bola samozrejme nula. Táto chyba je meraná ako odchýlka skutočnosti od odhadu premennej. Odchýlka odhadu je teda priemer štvorcov odchýlok hodnôt od ich priemeru.

Výberová chyba je neznáma veličina. Závisí od všetkých hodnôt základného súboru. Výberovú chybu odhadujeme pomocou pozorovaných hodnôt výberového súboru.

Chyba neodpovedania vzniká v prípade ak nemáme k dispozícii odpoveď od štatistickej jednoty. Či už z dôvodu toho že štatistická jednotka neodpovedala vedome, nebola zastihnutá alebo už neexistuje. V medzinárodnom porovnaní úspešnosti výberových skúmaní sa veľmi líši miera neodpovedania ako medzi krajinami tak medzi zameraním výberových skúmaní. Tejto problematike sa venovali vo svojom výskume de Leeuw a de Heer (2002). Zozbierali dáta z 10 rozličných opakujúcich sa výberových skúmaní (ročných, štvrťročných atď.) v 16 rozvinutých krajinách. Každá krajina disponovala dátami zo skúmania pracovných síl (v našom prípade ekvivalent k evidencii nezamestnanosti a výberovému skúmaniu pracovných síl), niektoré mali k dispozícii výsledky skúmaní spojené so zdravotníctvom, výdavkami domácností, cestovaním a príjmom. Autori dáta analyzovali pomocou modelov logistickej regresie kde premenné miera odpovedania, miera nenadviazania kontaktu, a miera odmietnutia boli závislé premenné. Vysvetľujúce premenné boli rok konania skúmania, krajina a druh prieskumu. Nasledujúce závery prieskumu boli vyhodnotené ako štatisticky významné:

- Miera odpovedania klesala s pribúdajúcimi rokmi. Krajiny ako aj typy prieskumu sa v miere neodpovedania líšili.
- Miera nenadviazania kontaktu rástla taktiež s pribúdajúcimi rokmi, a líšila sa v rámci krajín.
- Miera odmietnutia rástla s pribúdajúcimi rokmi, a líšila sa v rámci prieskumov.

Skúsenosti získané analýzou neodpovedania v prieskumoch ukázali že, v prípade jednotlivcom majú vyšší sklon k neodpovedaniu osoby ktoré sú obyvateľmi miest, slobodné, obyvatelia domácnosti bez detí, staršie osoby, rozvedené prípadne ovdovené osoby, osoby s nižším vzdelaním, a samostatne zárobkovo činné osoby.

3 POMOCNÉ INFORMÁCIE

Pomocné informácie sú všetky informácie, ktoré nepochádzajú z výberového skúmania a umožňujú zvýšiť presnosť výsledkov prieskumu. Týmto informáciami bývajú najčastejšie informácie o základnom súbore. Môže sa jednať o znalosť hodnôt určitých premenných alebo nejakú funkciu týchto hodnôt.

Pomocné informácie možno využiť:

- Pri tvorbe návrhu výberového skúmania
- Pri odhadovaní parametrov

V prípade že pomocné informácie sa využívajú pri návrhu výberového skúmania, hľadá sa taký návrh ktorý, poskytuje presné odhady pri danej cene prieskumu, alebo ktorý je menej nákladný pri danej presnosti odhadov. V prípade ak sa pomocné informácie využívajú pri odhadovaní parametrov slúžia ako verifikačný prostriedok ktorým, si overíme skutočnosti zistené výberovým skúmaním.

Záver

Problém neodpovedania je často sa vyskytujúci problém vo výberových skúmaniach ktorý, dokáže významne skresliť výsledky takéhoto skúmania. Je prítomný takmer pri každom výberovom skúmaní avšak najčastejšie ním trpia výberové skúmania, v ktorých figurujú ako štatistické jednotky fyzické osoby. Tento problém je možno riešiť dvoma spôsobmi, a to preventívne, alebo dodatočnou úpravou výberového súboru. Medzi metódy na úpravu výberového súboru patria: imputácia údajov, modifikácia výberového súboru pomocou poststratifikácie, použitie pomocného výberového súboru alebo použitie výberových váh.

Kľúčové slová

chyba neodpovedania, miera neodpovedania, pomocné informácie, problém neodpovedania, výberové skúmanie, štatistika, vychýlenosť odhadov

Klasifikácia JEL

C83

LITERATÚRA

- [1] CORNISH, J.: *Response problems in surveys Improving response & minimising the load*. Seminar „Good Practices in the Organisation and Management of Statistical Systems“. Dostupné na: „http://unstats.un.org/unsd/methods/statorg/Workshops/Yangon/Session4_SurSurv_Cornish_Paper.pdf“ 2002.
- [2] CHAJDIAK, J.: *Štatistika jednoducho*. Bratislava : STATIS, 2003. ISBN 80-85659-28-X.
- [3] GROVES, M. R.: *Nonresponse Rates and Nonresponse Bias in Household Surveys*. Online dostupné na: „<http://poq.oxfordjournals.org/content/70/5/646.full>“.
- [4] SÄRNDAL, C.-E. - LUNDSTROM, S.: *Estimation in Surveys with Nonresponse*. John Wiley and Sons, Ltd, 2005. ISBN 0-470-01133-5.
- [5] TEREK, M. – HRNČIAROVÁ, Ľ.: *Výberové skúmanie*. Bratislava : Ekonóm, 2008. ISBN 978-80-225-2440-7.
- [6] TEREK, M.: *Možnosti riešenia problému neodpovedania v štatistických prieskumoch*. Ekonomické rozľady/Economic Review 2/2014, s. 150 - 165.
- [7] TEREK, M.: *Problém neodpovedania v štatistických prieskumoch prostredníctvom pošty*. FORUM STATISTICUM SLOVACUM 6/2013, s. 156 - 161. ISSN 1336 – 7420.

RESUMÉ

Cieľom práce je popísať problém neodpovedania, objasniť na príklade jeho vplyv na výsledky výberového skúmania a navrhnúť možné riešenie tohto problému. Ako možné riešenia boli navrhnuté 2 metódy a to: prevencia a úprava výberového súboru. V ďalších kapitolách sa práca venuje kľúčovým bodom výberového skúmania a pomocným informácia, ktorých znalosť a zvládnutie sú nevyhnuté pre vytvorenie kvalitného výberového súboru ktorý je minimálne vychýlený a poskytuje čo najmenej skreslený obraz o základom súbore.

SUMMARY

The aim of this work is to describe the non-response problem, clarify its impact on the results of a sample survey and suggesting possible solutions to this problem. As a possible solution have been proposed two methods, namely: prevention and modification of the sample. In the following chapter, the work addresses the key point of the sample survey and review of auxiliary information, which knowledge and management are essential for the creation of a quality sample that has minimal bias and provides the least disorted picture of target population.

Kontakt

Ing. Valér Taragel, Janšákova 14, 841 07 Bratislava, e-mail: valer.taragel@gmail.com