

SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS
and DEMOGRAPHY

3/2024

ročník/volume 34

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: 5

Typ článku/Type of article: informatívny článok/informative article

Strany/Pages: 82 – 100

Dátum vydania/Publication date: 15. júl 2024/July 15, 2024



Informatívny článok/Informative article

Milan TEREK
Vysoká škola manažmentu

VÝBEROVÉ SKÚMANIE NA BÁZE NENÁHODNÝCH VÝBEROV

SAMPLE SURVEY BASED ON NONPROBABILITY SAMPLES

ABSTRAKT

S rozmachom internetu sa vo výberovom skúmaní opäť stalo populárnym nenáhodné vyberanie ako rýchla a lacná metóda tvorby výberov, ktoré možno rýchlo realizovať online. Cieľom príspevku je podať súhrnné hodnotenie štatistických prieskumov založených na nenáhodných výberoch. Základným kameňom hodnotenia kvality náhodných výberov je celková chyba prieskumu. Kvalitu nenáhodného výberu možno ohodnotiť iba posúdením toho, do akej miery sa konečný modelovaný výber zhoduje so základným súborom z hľadiska rôznych charakteristík. Uvedieme závery, ktoré vyplývajú zo štúdií o empirických dôkazoch o presnosti prieskumov založených na náhodných a nenáhodných výberoch. Všeobecne možno odporúčať, tak ako doteraz preferovať prieskumy na báze náhodných výberov.

ABSTRACT

With the spread of the Internet, nonprobability sampling has become popular again in sample surveys as a quick and inexpensive method of creating samples that can be quickly implemented online. The paper aims to provide a comprehensive assessment of statistical surveys based on nonprobability samples. The basis of the quality assessment of probability samples is the total survey error. The quality of a nonprobability sample can only be assessed by considering the extent of matching the final modeled sample with the population in terms of various characteristics. We will present the conclusions emerging from studies on empirical evidence of the accuracy of surveys based on probability and nonprobability samples. In general, it is advisable to prefer, as is the case now, surveys based on probability samples.

KLÚČOVÉ SLOVÁ

náhodný výber, nenáhodný výber, kvalita nenáhodného výberu, presnosť v štatistických prieskumoch

KEY WORDS

probability sample, nonprobability sample, quality of nonprobability sample, accuracy in statistical surveys

1. ÚVOD

Pri riešení nejakej výskumnej otázky pomocou výberového skúmania všeobecne existujú dva rozličné prístupy. Výberové skúmanie možno realizovať na báze náhodného (pravdepodobnostného) alebo nenáhodného (zámerného, nepravdepodobnostného) výberu. V článku si všimneme výsledky výskumu v oblasti hodnotenia kvality nenáhodných výberov a možností jej zlepšovania prostredníctvom ich úprav. Hlboká znalosť možností, ktoré poskytuje tento prístup, je mimoriadne užitočná pri riešení praktických problémov odhadovania veličín základného súboru.

Už v 20. a 30. rokoch 20. storočia vedecká komunita diskutovala o tom, či je v štatistických prieskumoch založených na výberovom skúmaní lepšie realizovať náhodné alebo nenáhodné vyberanie. Skúsenosti ukázali, že hlavnou príčinou nepresných záverov o základnom súbore bolo práve nenáhodné vyberanie. Následne bolo nenáhodné vyberanie vo väčšine najkvalitnejších sociálnych prieskumov realizovaných výberovým skúmaním nahradené náhodným vyberaním. S rozmachom internetu na konci 20. storočia sa však nenáhodné vyberanie opäť stalo populárnym ako rýchla a lacná metóda tvorby výberov [3].

Ďalej budeme uvažovať len o konečnom základnom súbore. Bez ohľadu na to, či sú štatistické prieskumy určené na predpovedanie výsledkov volieb, meranie verejnej mienky alebo na niečo iné a bez ohľadu na to, či sa vykonávajú online alebo off-line, štatistické prieskumy na báze náhodných aj nenáhodných výberov majú často spoločný cieľ: efektívne odhadnúť veličiny základného súboru na základe výsledkov analýzy výberového súboru. Preto prieskumy na báze náhodných aj nenáhodných výberov vyžadujú, aby [3]:

- a) jednotky vo výbere boli zameniteľné s jednotkami, ktoré nie sú vo výbere a ktoré majú rovnaké merané charakteristiky,
- b) žiadne časti základného súboru neboli systematicky úplne vylúčené z výberu,
- c) štruktúra jednotiek vo výbere s ohľadom na pozorované charakteristiky sa má buď zhodovať, alebo môže byť upravená tak, aby sa zhodovala so štruktúrou jednotiek v základnom súbore.

Napriek spoločnému cieľu poskytnúť presné informácie o základnom súbore, prieskumy na báze náhodných aj nenáhodných výberov sa líšia v kritickom aspekte. Kľúčový rozdiel spočíva v type a sile odôvodnenia, prečo by mal prístup viesť k dosiahnutiu presných výsledkov.

Pri riešení konkrétnej výskumnej otázky s využitím štatistického prieskumu na báze výberového skúmania je rozhodnutie o tom, či sa využije náhodný alebo nenáhodný výber, jednou z kľúčových otázok. Preto je pri praktických aplikáciách nevyhnutná hlboká znalosť možností, ktoré poskytuje každá z týchto metód.

Najprv uvedieme možné metódy nenáhodného vyberania, potom si všimneme možnosti hodnotenia kvality náhodných a nenáhodných výberov. Pri hodnotení kvality náhodných výberov sa možno oprieť o celkovú chybu prieskumu. V hodnotení kvality nenáhodných výberov sa odôvodnenie očakávania presných záverov môže opierať len o netestované predpoklady modelovania a možno ju ohodnotiť len posúdením toho, do akej miery sa konečný modelovaný výber zhoduje so základným súborm z hľadiska rôznych charakteristík. Uvedieme niektoré úpravy nenáhodných výberov, ktoré môžu slúžiť na odôvodnenie ich použitia na induktívne úsudky o veličinách základného súboru. Nakoniec uvedieme výsledky niektorých empirických štúdií, ktoré hodnotili presnosť prieskumov na základe náhodných a nenáhodných výberov, pomocou porovnania získaných výsledkov s výsledkami z vyčerpávajúceho skúmania alebo z vysoko kvalitných prieskumov na základe náhodných výberov z príslušných základných súborov.

2. NENÁHODNÉ VYBERANIE

Nenáhodné vyberanie možno rozdeliť do troch skupín [5]:

- a) vyberanie založené na prístupnosti jednotiek (*convenience sampling*),
- b) vyberanie založené na zhode (*sample matching*),
- c) sieťové vyberanie (*network sampling*).

Vyberanie založené na prístupnosti jednotiek je forma nenáhodného vyberania, pri ktorej je prvoradá ľahká lokalizácia a výber jednotiek. Nepoužíva sa žiadna formálna výberová schéma. Medzi niektoré typy takýchto výberov patrí zastavovanie v nákupných centrách (*mall intercepts*), výbery dobrovoľníkov (*volunteer samples*), tokové výbery (*river samples*) a RDE výbery.

Pri zastavovaní v nákupných centrách anketár oslovuje kupujúcich s prosbou o účasť v nejakom prieskume, pričom postupuje voľne, bez nejakej výberovej schémy. Výbery dobrovoľníkov sú bežné v spoločenských vedách, v prieskumoch trhu a pod. Dobrovoľníci sa môžu zúčastniť na jednej štúdii alebo sa stať súčasťou panelu, ktorého členovia môžu participovať v rôznych štúdiách počas dlhšieho časového obdobia. V poslednom čase sú to prihlasovacie webové panely (*opt-in web panels, access panels*), do ktorých sa prijímajú dobrovoľníci (samovýber) pri ich návšteve konkrétnych webových stránok [20]. V prihlasovacom paneli sa demografický profil respondentov dôsledne zhromažďuje, udržiava a aktualizuje v hlavnej databáze. Členovia panelu sa môžu zúčastniť na rozličných prieskumoch, často za odmenu. Je známy profil členov panelu a možno ich cielene osloviť v konkrétnych typoch výskumných štúdií. Panely majú tiež definovaný proces pravidelného čistenia a obnovovania, v ktorom sa panel naplňa novými respondentmi a zároveň sa odstraňujú respondenti, ktorí vykazujú vzor správania nízkej kvality. Online prieskumy najčastejšie využívajú výbery z prihlasovacích webových panelov osôb alebo domácností. Vo väčšine prípadov je nábor účastníkov prieskumu nenáhodný online výber. Panel tu slúži ako výberová báza¹. Pretože sa takéto výberové bázy tvoria nenáhodným vyberaním, výbery z nich sú tiež nenáhodnými výbermi. Nenáhodné online panely však čelia množstvu výziev, ako je nepokrytie ľudí bez prístupu na internet a výberové vychýlenie v dôsledku spoliehania sa na dobre dostupné vzorky dobrovoľníkov, ktorí sa môžu zúčastniť na viacerých online prieskumoch[2]. Napriek tomu sa dnes veľké množstvo prieskumov verejnej mienky vykonáva prostredníctvom nenáhodných online panelov [3, s. 6]. Výsledkom je, že väčšina dát z prieskumov zozbieraných online po celom svete sa dnes spolieha na nenáhodné výbery. Niekedy sa použijú zmiešané výbery, v ktorých časť výberu pochádza z panelu a ďalšia časť pochádza z náboru v reálnom čase mimo internetu. Bez úpravy budú odhady založené na takýchto výberoch vychýlené [16].

Okrem online panelov je na vzostupe nepanelový nábor online respondentov, napríklad prostredníctvom tokového vyberania (*river sampling*). Tokové výbery prijímajú členov na viacerých webových stránkach, alebo (menej často) na sociálnych sieťach. Webové stránky sa vyberajú tak, aby sa dosiahol prierez demografickými skupinami. V tejto metóde sa respondenti získavajú tak, že návštevník web stránky je požiadaný o účasť v prieskume kliknutím na odkaz, ktorý ho presmeruje na prieskum [15]. Bežnými poskytovateľmi „prúdového toku“ sú napríklad marketingové agentúry zapojené do sociálnych médií. Potenciálny respondent prieskumu klikne na reklamu

¹ Výberová báza (*sampling frame, základ výberu, opora výberu*) je zoznam zostavený s cieľom tvorby výberu, ktorý označuje jednotky základného súboru, z ktorého sa vyberá tak, aby sa mohli brať do úvahy pri ich skúmaní [24, s. 108].

alebo inú ponuku, môže zvoliť nasmerovanie na niekoľko predbežných kontrolných otázok a na základe svojich odpovedí môže byť nasmerovaný na dotazník prieskumu. Analytik nemá dopredu žiadne informácie o osobách, ktoré sa zapoja do prieskumu. RDE (*Random Device Engagement*) vyberanie je vyberanie potenciálnych respondentov prostredníctvom zapojenia online používateľov na všetkých zariadeniach, ktoré už používajú, či už v rámci reklamných sietí, mobilných aplikácií alebo iných portálov na rôznych zariadeniach. Vyberanie RDE sa môže šíriť prostredníctvom digitálnych prvkov podobných tým, ktoré sa používajú pri tokovom vyberaní, ako sú banery², reklamy a iné miesta na webových stránkach, ako sú tlačidlá. Tieto výzvy na prieskum musia byť umiestnené strategicky, aby ich respondenti ľahko rozpoznali. Musia byť tiež vytvorené tak, aby vyvolali zvedavosť alebo záujem návštevníkov webovej stránky. Keďže tokové a RDE vyberanie je rýchly a ľahko použiteľný spôsob vykonávania výskumu, je to užitočný spôsob vyberania hlavne pre výskumníkov, ktorí nemajú veľa času ani peňazí na investovanie do svojej štúdie. Môžu byť užitočné aj pri oslovovaní širokého spektra demografických skupín. Najväčšou nevýhodou tokového a RDE vyberania je problematickosť nájdania reprezentatívneho výberu³ pre skúmanú podskupinu.

Pri vyberaní založenom na zhode, sa jednotky do nenáhodného výberu vyberajú tak, aby štruktúra výberu zodpovedala štruktúre základného súboru podľa dôležitých charakteristík. Napríklad výber osôb môže byť vytvorený tak, aby jeho rozdelenie podľa veku a pohlavia bolo blízke rozdeleniu základného súboru podľa týchto charakteristík⁴. Kvótové vyberanie⁵ je príkladom vyberania založeného na zhode. Takéto vyberanie je určené na zníženie výberových vychýlení⁶ do tej miery, do akej je možné pri štruktúrovaní použiť kovariáty, ktoré predikujú výstupy prieskumu. Možno zvoliť aj individuálne porovnávanie jednotiek. Pri ňom sa využije výberová báza, ktorá obsahuje veľké množstvo pomocných informácií na vybratie cieľového výberu s využitím známych pravdepodobností vybratia⁷. Pre každú jednotku z cieľového výberu sa vyberie jednotka z panelu, s najväčšou zhodou. Pretože zhody nemusia byť úplné, výsledný výber vyžaduje váženie. V [16] sa uvádza, že pri webových prieskumoch, najmä na báze prihlasovacích webových panelov sa zdá, že výbery založené na zhode sú pri znižovaní vychýlenia efektívnejšie ako tradičné metódy.

² Baner je reklamný formát na internete, ktorý pozostáva zo zahrnutia grafického reklamného prvku do webovej stránky.

³ Reprezentatívnym výberom myslíme výber, ktorý umožňuje vypočítať spoľahlivé odhady veličín základného súboru. Podrobnejšie o rozličných chápaniach termínu reprezentatívny výber pozri v [1], s. 23 – 24.

⁴ Ak vek a pohlavie nie sú študované premenné, ale ovplyvňujú výstupy štúdie, ide o kovariáty (covariates). Všeobecne, kovariáty sú premenné, ktoré ovplyvňujú výstupné premenné, ale nie sú súčasťou štúdie.

⁵ Viac o kvótovom vyberaní pozri v [22].

⁶ Výberové vychýlenie (selection bias) sa týka systematických rozdielov medzi hodnotou odhadu a hodnotou odhadovanej veličiny v základnom súbore a je zapríčinené problémami so zložením výberu [13].

⁷ Majme konečný základný súbor rozsahu N . Konkrétny výber n jednotiek z tohto základného súboru označme S . V náhodnom výbere má každý možný výber S zo základného súboru známu pravdepodobnosť že bude vybraný – pravdepodobnosť vybratia (selection probability) $p(S)$. Suma pravdepodobností výberu všetkých možných výberov sa rovná jednej.

Pri sieťovom vyberaní jednotky z nejakého cieľového základného súboru⁸ (zvyčajne zriedkavejšie skupiny obyvateľov, napríklad geovia) sú požiadaní, aby identifikovali iných členov základného súboru, s ktorými sú nejakým spôsobom spojení. Takto identifikovaní členovia základného súboru sa požiadajú, aby sa pripojili k výberu. Tento spôsob náboru sa môže realizovať v niekoľkých kolách. Vyberanie metódou snehovej gule (*snowball sampling*, *chain sampling*, *chain-referral sampling* alebo *referral sampling*) je príkladom sieťového vyberania. Sieťové vyberanie je vhodný spôsob zhromaždenia veľkého výberu jednotiek zo zriedkavejšieho základného súboru.

Rastúce využívanie prieskumov, ktoré nie sú založené na náhodných výberoch vyvolalo záujem, ale aj obavy zo strany štatistikov. Rastúce náklady na zber dát spolu s klesajúcou mierou odpovedania zdôraznili atraktivnosť lacných prieskumov na báze nenáhodných výberov, ktoré možno rýchlo realizovať online [13]. Samo vyberanie respondentov do týchto prieskumov však spôsobuje, že indukčné metódy použiteľné v prístupe založenom na pláne výberového skúmania⁹ sa dajú použiť len pri splnení istých predpokladov a procedúry váženia dokážu redukovať vychýlenia odhadov len čiastočne.

3. HODNOTENIE KVALITY VÝBEROV A POUŽITEĽNOSŤ NENÁHODNÝCH VÝBEROV NA INDUKTÍVNE ÚSUDKY

3.1. Hodnotenie kvality náhodných výberov

V prípade prieskumov na báze náhodných výberov sa možno oprieť o teóriu pravdepodobnosti, ktorá je založená na súbore zavedených matematických princípov. Základným kameňom hodnotenia kvality náhodných výberov je celková chyba prieskumu (*total survey error*). Ide o rozdiel medzi veličinou základného súboru a hodnotou odhadu tejto veličiny na báze analýzy výberu. Má dve zložky: výberovú chybu (*sampling error*) a nevýberovú chybu (*nonsampling error*). Výberové chyby vznikajú vtedy, keď sa veličiny základného súboru odhadujú z podmnožiny alebo výberu z tohto základného súboru. Keďže výber nezahŕňa všetky jednotky základného súboru, hodnoty výberových charakteristík ktoré slúžia ako hodnoty odhadov veličín v základnom súbore, sa od nich môžu líšiť. Rozdiel medzi výberovou charakteristikou a odhadovanou veličinou základného súboru sa považuje za výberovú chybu [18]. Nevýberová chyba (*nonsampling error*) je súčtom všetkých ostatných chýb. Možno ich rozdeliť na 5 kategórií:

1. Chyba špecifikácie (*specification error*) nastane, keď sa koncept daný výskumnou otázkou líši od konceptu, ktorý sa má v prieskume merať. Chyba špecifikácie je často spôsobená zlou komunikáciou medzi výskumníkom, analytikom dát alebo sponzorom prieskumu a tvorcom dotazníka.
2. Chyba výberovej bázy (*frame error*) zvyčajne vyplýva z procesu jej konštrukcie. Rozlišujú sa tri typy chýb pokrytia: nepokrytie (*noncoverage*) alebo neúplné pokrytie (*undercoverage*), napríklad niektoré jednotky sú vo výberovej báze vynechané, presahujúce pokrytie (*overcoverage*), napríklad zaradenie

⁸ Cieľový základný súbor (*target population*) je základný súbor, o ktorom chceme robiť indukčné úsudky.

⁹ Vo výberovom skúmaní (*sample survey*, *survey sampling*) sú známe tri rozličné prístupy – prístup založený na pláne výberového skúmania (*design based approach*), prístup založený na modeli (*model based approach*) a prístup s asistenciou modelu (*model-assisted approach*). Podrobnejšie pozri v [10], [12], [22], [24], [25].

jednotiek ktoré nepatria do základného súboru, a duplicitné zaradenie niektorých jednotiek (*duplicate listings*) [20].

3. Chyba neodpovedania (*nonresponse error*) zahŕňa neodpovedanie jednotky (*unit nonresponse*), pri ktorom chýba celý dotazník, a *častočné neodpovedanie jednotky* (*item nonresponse*) pri ktorom jednotka odpovedá na jednu alebo viaceré, ale nie na všetky otázky v dotazníku (dotazník je čiastočne vyplnený) [23].
4. Chyba merania (*measurement error*) nastáva vtedy, keď spôsob získania merania ovplyvňuje zaznamenanú hodnotu, pričom často sa na nej podieľa súčasne respondent, anketár aj dotazník prieskumu.
5. Chyba spracovania (*processing error*) sa vzťahuje na chyby, ktoré sa vyskytnú počas fázy spracovania dát, vrátane chýb pri ich úprave, kódovaní, priradovaní váh a zaraďovaní dát do tabuliek. Pri náhodnom vyberaní možno výberovú chybu kontrolovať pomocou výberovej schémy (*sampling design*). Môže ísť o jednoduché, stratifikované, skupinové, viacstupňové náhodné vyberanie, prípadne náhodné vyberanie s nerovnakými pravdepodobnosťami alebo ich rozličné kombinácie. Tento spoľahlivý teoretický základ umožňuje vypočítať presnosť odhadov (napr. vo forme intervalov spoľahlivosti alebo hraníc chýb) a dáva metóde odhadu univerzálnu platnosť. Okrem toho, keďže poskytovatelia prieskumov na báze náhodných výberov bežne opisujú detaily procesu generovania dát (napr. v [6], [7]), výskumníci sú schopní vykonať úpravy, ktoré zohľadňujú prípadné chyby pokrytia a výberové a nevýberové vychýlenie.

Náhodné vyberanie je bez akejkoľvek diskusie najvhodnejší prístup k tvorbe induktívnych úsudkov o základnom súbore. Náhodný výber vytvára prierez základného súboru a teória výberových skúmaní poskytuje nástroje na opis a ohodnotenie vychýlenia a rozptylu na základe plánu výberového skúmania, ktorý determinuje pravdepodobnosti zahrnutia¹⁰. Za ideálnych okolností výberová база pokrýva základný súbor takmer na 100 %. Výsledkom je, že možno očakávať nevychýlenosť alebo len malé vychýlenia výberových charakteristík [17, s. 3658].

3.2. Hodnotenie nenáhodných výberov a ich použiteľnosť na induktívne úsudky

Tradične sa výberové vychýlenie považuje za výsledok nepokrytia alebo neodpovedania. Tieto koncepty sú spojené s náhodným výberom z konečného základného súboru. Tieto kategórie sa môžu ukázať ako limitujúce, ak sa aplikujú v kontexte nenáhodných výberov. Mnohé prieskumy na báze nenáhodných výberov nemožno spájať s výbermi z nejakej výberovej bázy. Dokonca aj predstava výberu ako konečného súboru jednotiek, z ktorých niektoré nemusia odpovedať, sa nevzťahuje na mnohé prieskumy na báze nenáhodných výberov. Pri takýchto prieskumoch môžu byť procesy zaradenia respondenta do výberu rôznorodé, niekedy úplne nepodobné procesu zaradenia jednotky do výberu v prieskumoch na báze náhodných výberov [13].

¹⁰ Pravdepodobnosť zahrnutia π_i (*inclusion probability*) je pravdepodobnosť, že *i*-tá jednotka bude vo vybranom výbere. V náhodnom vyberaní sú pravdepodobnosti π_i známe predtým ako sa začne vyberať. Predpokladáme, že pre každú jednotku v základnom súbore je pravdepodobnosť zahrnutia väčšia ako nula [12, s. 28]. Výberové váhy w_i pre ľubovoľnú výberovú schému sú definované ako $w_i = 1/\pi_i$. Výberovú váhu *i*-tej jednotky možno interpretovať ako počet jednotiek v základnom súbore, ktoré reprezentuje.

Nenáhodné výbery sa často získavajú spôsobom, ktorý je ťažko kontrolovateľný. Ak majú slúžiť na odhadovanie veličín konkrétneho konečného základného súboru, môžu byť zdrojom množstva potenciálnych vychýlení. Všimnime si možné problémy na príklade prihlasovacích webových panelov, môžu sa však týkať aj iných nenáhodných výberov. Výberová chyba vzniká vtedy, keď sa vybraná časť základného súboru (jednotky vo výbere) líši od jeho nevybranej časti (jednotky ktoré nie sú vo výbere) takým spôsobom, že výber nemožno premietnuť do celého cieľového základného súboru. Veľkým problémom je, či nenáhodný výber pokrýva cieľový základný súbor. Napríklad v prihlasovacom webovom paneli sa môžu k panelu pripojiť iba osoby s prístupom na internet. V [5, s. 253] sú uvedené tri rozličné základné súbory, ktoré umožňujú opísať tri komponenty vychýlenia pokrytia prieskumu: U – cieľový základný súbor, F_{pc} – potenciálne pokrytá časť cieľového základného súboru daná spôsobom zhromažďovania dát a F_c – skutočne pokrytá časť cieľového základného súboru, získaná pre štúdiu prostredníctvom základných podmienok prieskumu. Uvažujme napríklad o prieskume, ktorý sa týka odvykania od fajčenia. Účastníkov prieskumu budeme získavať z prihlasovacieho webového panelu. Cieľový základný súbor U definujme ako fajčiarov vo veku 18 – 65 rokov. Základný súbor F_{pc} by tvorili fajčiari z danej vekovej skupiny, ktorí majú prístup na internet a navštívili stránku s výzvou na účasť v prieskume. Základný súbor F_c by tvorila podmnožina súboru F_{pc} , ktorá prejavila ochotu zúčastniť sa na štúdiu. Je zrejmé, že realizácia výberu len zo súboru F_c je zdrojom výberovej chyby. Výber S tvoria osoby, ktoré sú pozvané na účasť v prieskume a ktoré sa na ňom aj zúčastnia. Rozdiel medzi U a F_{pc} predstavuje osoby, ktoré majú prístup na internet, ale nenavštívili náborové webové stránky, prípadne nemajú prístup na internet. Často môže byť tento rozdiel väčší ako F_c alebo F_{pc} . Výberová chyba môže vzniknúť aj tým, že niektoré skupiny osôb sa prihlásia do panelu s väčšou pravdepodobnosťou ako iné osoby. Prihlasovacie webové panely ovplyvňujú aj niekoľko druhov neodpovedania. Mnohí predajcovia panelov aplikujú na pripojenie k panelu postup dvojitého súhlasu. Osoba najprv zaregistruje svoje meno, e-mail a niektoré demografické údaje. Potom predajca pošle danej osobe e-mail, na ktorý je potrebné odpovedať, aby sa mohla oficiálne pripojiť k panelu. Vyzvaná osoba nemusí na prijatý e-mail odpovedať, ak odpovie, nemusí dokončiť celú registráciu do panelu. Ak je člen panelu vyzvaný na účasť v nejakom prieskume, nemusí odpovedať. Ďalším problémom je možná strata záujmu byť členom panelu. Aj v nenáhodných výberoch je nevyhnutné počítať s chybou merania, ktorá môže byť spôsobená návrhom dotazníka a osobitosťami respondentov. Motiváciou na účasť v štúdiu môže byť pre niektorých pocit altruizmu, pre iných len získanie odmeny (účastníci prieskumu z panelu získavajú peňažné odmeny). Z uvedeného je zrejmé, že celkovú chybu prieskumu nemožno jednoducho aplikovať aj na nenáhodné výbery. Nenáhodné výbery neumožňujú vytvoriť jasné závery na základe známej výberovej schémy. Nemožno ich ohodnotiť kvantifikáciou a sčítaním chýb, ktoré sa vyskytujú v každej fáze procesu vyberania. Odôvodnenie očakávania presných záverov sa opiera o netestované predpoklady modelovania, ktoré sú založené na presvedčení výskumníka o charakteristikách, ktoré odlišujú výber od zvyšku základného súboru, a ako tieto charakteristiky súvisia s témou výskumu.

Kvalitu nenáhodného výberu možno ohodnotiť iba posúdením toho, do akej miery sa konečný modelovaný výber zhoduje so základným súborom z hľadiska rôznych charakteristík. Formálna štatistická teória ponúka na zdôvodnenie použiteľnosti nenáhodných výberov na induktívne úsudky o základných súboroch len málo. Neexistuje žiadna všeobecná štatistická teória nenáhodného vyberania, ktorá by

odôvodňovala, kedy a prečo možno očakávať na základe nenáhodného výberu presné závery o základnom súbore. Platnosť záverov závisí od témy a konkrétneho prieskumu. Okrem toho poskytovatelia nenáhodných online výberov často považujú svoje postupy zberu dát za dôverné, čo sťažuje alebo robí nemožným zistiť, ktoré faktory treba zahrnúť do ktorého modelu, aby sa redukovala výberová chyba kľúčových odhadov.

Podľa [5, s. 251 – 252], všeobecným rámcom na tvorbu induktívnych záverov na báze nenáhodných výberov je úvaha o združenom rozdelení pravdepodobnosti vektora študovanej premennej v základnom súbore $\mathbf{Y} = (Y_1, Y_2, \dots, Y_N)$ a o vektore 0-1 indikačných premenných v základnom súbore $\delta_S = (\delta_1, \delta_2, \dots, \delta_N)$ pre výber S . Predpokladajme, že \mathbf{X} je matica kovariátov typu $N \times p$, ktorú možno použiť pri navrhovaní výberu alebo pri konštrukcii odhadov. Podmiernená hustota \mathbf{Y} pre dané \mathbf{X} a vektor parametrov $\boldsymbol{\theta}$ je $f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta})$. Vektor parametrov závisí od použitého prístupu k tvorbe modelu. Napríklad pri kvázi-randomizácii ide o parametre použitej výberovej schémy, pri prístupe založenom na superpopulácii ide o parametre rozdelenia superpopulácie (podrobnejšie informácie o týchto dvoch prístupoch uvádzame v časti 3.2.3). Hustota δ_S pre dané \mathbf{Y}, \mathbf{X} a iný neznámy parameter $\boldsymbol{\Phi}$ je $f(\delta_S|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Phi})$. Združená hustota \mathbf{Y} a δ_S je:

$$f(\mathbf{Y}, \delta_S|\mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\Phi}) = f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta})f(\delta_S|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Phi}).$$

Všimnime si, že skutočnosť či jednotka je vo výbere závisí od \mathbf{Y} , to znamená že dáta sú typu NMAR¹¹. V náhodnom výbere bez neodpovedania alebo iných strát ktoré sú mimo kontroly analytika, $f(\delta_S|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Phi}) = f(\delta_S|\mathbf{X})$. Hustota $f(\delta_S|\mathbf{X})$ je randomizačné rozdelenie a je základom indukcie na báze náhodného výberu pomocou prístupu založeného na pláne výberového skúmania. V nenáhodnom výbere môže rozdelenie δ_S závisieť zároveň od \mathbf{Y} a od neznámeho parametra $\boldsymbol{\Phi}$. V závislosti od aplikácie možno induktívne úsudky založiť na $f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta})$, na $f(\delta_S|\mathbf{Y}, \mathbf{X}; \boldsymbol{\Phi})$, prípadne na ich kombinácii.

Vo všeobecnosti sa prijímajú na odôvodnenie použitia nenáhodných výberov na induktívne úsudky takéto tvrdenia [3, s. 8]:

- a) každý výber na skúmanie konkrétnej otázky poskytne rovnaké induktívne závery,
- b) špecifická výberová schéma, súvisiaca s danými otázkami prinesie závery, ktoré odzrkadľujú skutočnosť v skúmanom základnom súbore,
- c) séria analytických krokov bude brať do úvahy všetky rozdiely medzi výberom a základným súborom,
- d) konkrétna kombinácia výberových a/alebo analytických prístupov bude viesť k presným hodnotám odhadov parametrov základného súboru.

¹¹ Podrobnejšie o dátach MCAR (missing completely at random), MAR (missing at random) a NMAR (not missing at random) pozri v [1, s. 185].

Predpoklad, že použitie nejakej konkrétnej metódy pri riešení konkrétnej výskumnej otázky je vhodné, závisí od základných tvrdení o tejto otázke, o výbere a od použitej procedúry úpravy. Všimneme si každý z uvedených štyroch bodov podrobnejšie.

3.2.1. Nedôležitosť výberovej schémy daná charakterom skúmanej otázky

V niektorých prípadoch nemusí byť skúmaná otázka spojená s výberovou schémou. Napríklad sa študuje nejaký fyziologický alebo psychologický proces, pri ktorom možno predpokladať, že je podobný u všetkých jednotlivcov. Vtedy môže byť rozumné predpokladať, že ktorákoľvek skupina jednotlivcov sa bude správať podobne ako ktorákoľvek iná skupina. Jednoducho možno oprávnene predpokladať, že skúmaný proces je univerzálny a teda rozličné skupiny jednotlivcov sa budú správať podobne. Potom je irelevantné či sa realizuje náhodné alebo nenáhodné vyberanie. Za zmienku stoja aj niektoré teoretické a empirické dôvody veriť, že niektoré triedy induktívnych záverov môžu byť viac alebo menej spojené so zložením výberu. Niektorí autori napríklad tvrdia, že trendy v čase čo sa týka postojov a správania sú menej spojené so zložením výberu ako hodnoty odhadov rozdelenia postojov a správania. Podobne, niektoré empirické štúdie ukázali, že odhady vzťahov medzi premennými menej záviseli od toho, či je výber náhodný alebo nenáhodný, ako iné typy odhadov. Tvrdenie, že niektoré druhy induktívnych záverov možno urobiť rovnako dobre bez ohľadu na stratégiu vyberania, môže byť niekedy správne. Otázkou je, kedy [3, s. 8 – 9].

3.2.2. Vhodnosť výberovej schémy

Niekedy analytik jednoducho prijme tvrdenie, že použitá výberová schéma znižuje všetky vychýlenia, ktoré môžu v procese vyberania vzniknúť. Klasický príklad predstavuje kvótové vyberanie. Pri kvótovom vyberaní je základný súbor rozdelený do podskupín rovnako ako pri stratifikovanom náhodnom vyberaní, ale s jedným dôležitým rozdielom. Pri tvorbe výberu sa nepoužíva náhodné vyberanie. V extrémnych verziách kvótového výberu je výber jednotiek vo výbere výlučne na uvážení anketára. Pri kvótovom výbere sa v konečnom výbere vyžadujú špecifikované počty (kvóty) jednotlivých typov jednotiek základného súboru [12, s. 96 – 97]. Kvótové výbery sú zvyčajne navrhnuté tak aby sa súbor respondentov zhodoval so základným súborom v určitých kľúčových demografických parametroch. Predpokladá sa, že demografické parametre, ktoré tvoria bázu na tvorbu kvót pokrývajú zdroje vychýlenia induktívnych záverov. Induktívne závery na báze kvótového výberu možno považovať za presné do tej miery, do akej je to tak, pretože všetky potenciálne vychýlenia sú neutralizované výberovou schémou.

Okrem demografických kvót sú známe aj iné postupy. Možno napríklad porovnávať osoby v nenáhodných výberoch s osobami v náhodných výberoch a na základe toho vybrať respondentov [16]. Je známy aj postup na zmiešanie výberov získaných zo zdrojov, ktoré majú známe opačné vychýlenia [19]. Ak je analytik presvedčený, že zvolená stratégia vyberania eliminuje všetky potenciálne vychýlenia, čo sa týka skúmanej otázky, potom táto stratégia môže viesť k presným induktívnym záverom.

Problém s týmito druhmi prístupov spočíva v tom, že presnosť kritických predpokladov možno skutočne stanoviť len empiricky. Nie je tiež jasné, čo by mohlo dokázať, že konkrétny záver je robustný voči konkrétnemu rozhodnutiu o tvorbe výberu. Môže sa stať, že povaha otázky a/alebo typ induktívneho záveru ho robí presným pre akúkoľvek podobnú otázku na základe podobne získaného výberu, alebo

sa môže stať, že zvláštnosti jednej analýzy priniesli rovnaké závery len náhodou. Neexistuje žiadny zrejmy spôsob, ako určiť, o ktorú z týchto možností ide.

3.2.3. Prístupy k úprave výberu

Opíšeme modelové prístupy, ktoré možno použiť na zlepšenie presnosti dát z nenáhodných výberov. Sú známe procedúry na úpravu systematických vychýlení v náhodných výberoch. Niektoré možno použiť aj na úpravu výberových vychýlení v nenáhodných výberoch. Možno ich rozdeliť na dve skupiny: globálne úpravy (*global adjustments*) a úpravy vzhľadom na špecifický výstup (*outcome-specific adjustments*). Globálne úpravy spočívajú v použití modelu na vytvorenie úpravy, ktorá je použiteľná v ľubovoľnej nasledujúcej analýze, bez ohľadu na študovaný výstup. Úpravy vzhľadom na špecifický výstup spočívajú v tvorbe modelu úpravy „šitého“ na špecifický študovaný výstup [3, s. 10].

GLOBALNE ÚPRAVY

Všeobecne používaný prístup je kalibračné váženie (*calibration weighting*). Kalibrácia je postup, pri ktorom sa výberové váhy upravujú tak, aby sa odhadované úhrny pomocných premenných v základnom súbore zhodovali so skutočnými úhrnmi týchto premenných v základnom súbore [12, s. 154]. Kalibračné váženie spočíva vo vážení skupiny respondentov tak, aby vážené výberové úhrny hodnôt nejakej charakteristiky zodpovedali známym úhrnom hodnôt tejto charakteristiky v základnom súbore. Známe úhrny v základnom súbore môžu pochádzať napríklad z makrocenzu alebo z iných zdrojov, o ktorých sa predpokladá, že sú veľmi kvalitné. Procedúra generuje globálne váhy, ktoré možno použiť v analýze ľubovoľnej výstupnej premennej. Táto stratégia predpokladá, že vybrané jednotky zo špecifikovaných podskupín základného súboru budú zhruba rovnaké ako nevybrané jednotky z týchto podskupín vzhľadom na indukívne závery, ktoré sa budú robiť na základe výberových dát. Kalibračné váženie teda vyžaduje len prístup k referenčným dátam na úrovni základného súboru. V [17] sa napríklad skúmajú dva online prieskumy na báze nenáhodných výberov s cieľom preskúmať stratégie na úpravu a kalibráciu nenáhodných výberov. Pre obe štúdie bolo spoločné, že úprava podľa demografických premenných sama o sebe neznížila žiadne vychýlenie.

V kvázi-randomizácii (*quasi-randomization*) sa odhadujú pravdepodobnosti pseudo-zahrnutia (*pseudo-inclusion probabilities*), ktoré sa použijú na korekciu výberového vychýlenia [5, s. 255]. Na základe týchto odhadnutých pravdepodobností sa vypočítajú výberové pseudo-váhy (*pseudo-weights*) ako ich obrátené hodnoty. Tieto pseudo-váhy sa potom použijú v nasledujúcich analýzach na báze nenáhodných výberov na výpočet hodnôt odhadov veličín základného súboru podľa vzťahov z prístupu založeného na pláne výberového skúmania. Keď použijeme rovnakú symboliku ako v časti 3.2, cieľom je odhadnúť $f(\delta_S | \mathbf{Y}, \mathbf{X}; \Phi)$ alebo $f(\delta_S | \mathbf{X}; \Phi)$. Situácia, v ktorej pravdepodobnosti zahrnutia nezávisia od Y_i je ideálna, pretože Y_i pre jednotky, ktoré neboli vybrané, nepoznáme. Overenie platnosti tejto situácie je však vo väčšine aplikácií nemožné. V praxi sa teda väčšinou odhaduje $f(\delta_S | \mathbf{X}; \Phi)$ bez overenia platnosti tejto podmienky [5, s. 255].

Jedným z prístupov v kvázi-randomizácii je referenčný prieskum (*reference survey*). V tomto prístupe sa spolu s prieskumom na báze nenáhodného výberu využíva referenčný prieskum. Môže to byť prieskum na báze náhodného výberu zo základného súboru osôb, ktoré majú prístup na internet alebo zo všetkých osôb bez ohľadu na to

či majú prístup na internet. Referenčný výber môže byť aj cenzus, ktorý pokrýva celú populáciu [5, s. 255]. Referenčný výber sa kombinuje s výberom dobrovoľníkov z panelu tak, že dáta z výberu dobrovoľníkov z panelu sa vertikálne zreťazia (*vertically concatenate*)¹² s dátami z referenčného výberu. Potom sa sformuluje logit alebo probit model¹³ a odhadnú sa jeho parametre na báze dát z oboch výberov. Odhadnutý model sa použije na predikcie pravdepodobností pseudo-zahrnutia jednotiek v nenáhodnom výbere. Na základe nich sa vypočítajú pseudo-výberové váhy ako ich obrátené hodnoty. Kľúčovou požiadavkou referenčného prieskumu je, aby zahŕňal tie isté kovariáty ako prieskum dobrovoľníkov v paneli, aby bolo možné použiť lineárny model pre binárnu premennú na odhadnutie pravdepodobností pseudozahrnutia jednotiek v nenáhodnom výbere.

Zhoda výberov (*sample matching*) je ďalší prístup v rámci kvázi-randomizácie. Vyvážený nenáhodný výber možno zostaviť napríklad výberom jednotiek z veľmi veľkej výberovej bázy, napríklad zo zoznamu registrovaných členov prihlasovacieho webového panelu, na základe pomocných charakteristík (často demografických), ktoré sa tesne zhodujú¹⁴ s charakteristikami jednotiek z referenčného náhodného výberu. Zhoda na agregátnej úrovni spočíva v zhode rozdelenia početností nenáhodného výberu a rozdelenia konečného základného súboru. Príkladom je kvóťový výber. Napríklad rozdelenie výberu podľa veku a pohlavia má byť rovnaké ako v základnom súbore. Keď máme rozsiahly panel dobrovoľníkov, vyberie sa malý výber tak, aby sa dodržalo toto rozdelenie. Každéj osobe sa priradí rovnaká váha, čo je rovnaké ako pri náhodnom vyberaní s konštantným výberovým pomerom¹⁵. Kvóťový výber možno takto považovať za kvázi-randomizáciu [5, s. 256].

Uvedené prístupy využívajú referenčné dáta len na tvorbu váh, párovanie prípadne na imputácie¹⁶, v analýze nenáhodného výberu sa však nevyužívajú. Alternatívne prístupy kombinujú oba zdroje dát a analyzujú ich spoločne. Jedným z takých prístupov je odhadovanie založené na pseudo-dizajne (*pseudo design-based estimation*) [5]. V ňom sa odhadnú pravdepodobnosti pseudo-zahrnutia pre jednotky nenáhodného výberu na základe množiny premenných spoločných pre náhodný aj nenáhodný výber. Keď sa pravdepodobnosti pseudo-zahrnutia priradia všetkým jednotkám z nenáhodného výberu, potom sa s týmito jednotkami môže pracovať, ako keby boli vybrané pomocou rovnakého mechanizmu ako jednotky v náhodnom výbere. Súbor dát sa potom môžu spojiť a analyzovať spoločne s použitím skutočných váh a pseudo-váh.

ÚPRAVY VZHLADOM NA ŠPECIFICKÝ VÝSTUP

Tieto prístupy používajú modely úpravy „presne šité“ na výstupnú premennú. To znamená že upravujú mechanizmus vyberania jednotiek do nenáhodného výberu vzhľadom na výstupnú študovanú premennú. Takéto prístupy sa pokúšajú kontrolovať premenné, ktoré riadia proces výberu jednotiek a sú korelované s cieľovou výstupnou premennou. Jeden príklad takého prístupu je teoretická koncepcia, v ktorej sa predpokladá, že náhodný aj nenáhodný výber sa vyberajú z hypotetickej nekonečne

¹² Vertikálne zreťazenie matice je operácia, ktorá spojí dve submatice vertikálne do jednej matice. Počet stĺpcov oboch submatic musí byť rovnaký.

¹³ Základné informácie o logit a probit modeloch pozri v [11].

¹⁴ Mierou zhody môže byť napríklad euklidovská vzdialenosť.

¹⁵ Výberový pomer je pomer medzi rozsahom výberu a rozsahom konečného základného súboru.

¹⁶ V štatistike je imputácia proces nahradenia chýbajúcich dát náhradnými hodnotami.

veľkej superpopulácie (*superpopulation*). V tomto prístupe sa štatistický model pre Y odhadne z výberu a použije sa na premietnutie výberu do celého základného súboru. To znamená, že indukčné úsudky sú založené na $f(Y|X; \Theta)$. Tento prístup môže byť samozrejme použitý aj pri náhodnom výbere. Na rozdiel od indukcie na báze prístupu založeného na pláne výberového skúmania v ktorom je randomizačné rozdelenie pod kontrolou analytika, pri nenáhodnom výbere pod kontrolou nie je. Mechanizmus tvorby výberu možno ignorovať pri aplikácii prístupu založeného na modeli k indukčným úsudkom o rozdelení Y , keď $f(\delta_S|Y, X; \Phi) = f(\delta_S|X; \Phi)$, čo je formálne odôvodnenie použitia len $f(\delta_S|X; \Phi)$. Môžu existovať nenáhodné výbery, ktoré túto podmienku spĺňajú. V nenáhodných výberoch, v ktorých vyberanie jednotiek nie je dobre kontrolované, nemusí byť táto podmienka splnená. Vtedy sa odporúča kombinovať prístup založený na kvázi-randomizácii s prístupom založeným na superpopulácii [5, s. 257 – 258].

Cieľom analytika je modelovať proces generovania dát príslušnej superpopulácie tak aby boli v modeli analýzy zohľadnené všetky relevantné premenné. V praxi to znamená, že výskumník zostaví predikčný model pre nejakú premennú Y na základe výberu, ktorý je k dispozícii. Ten sa použije na predikciu hodnôt Y pre jednotky, ktoré nie sú vo výbere. Vybrané a nevybrané jednotky sa potom kombinujú a na základe nich sa odhadne skúmaná veličina (napr. stredná hodnota, úhrn alebo regresný koeficient) v celom základnom súbore. Kľúčovými predpokladmi tohto prístupu je, že premenná Y je vysvetlená prostredníctvom modelu, ktorý je spoločný pre vybrané aj nevybrané jednotky, a že všetky parametre, ktoré riadia model superpopulácie, sa v modeli kontrolujú. Tento prístup tiež vyžaduje dostupnosť pomocných dát o základnom súbore na tvorbu predikcií pre jednotky ktoré nie sú vo výbere [3, s. 12 – 13].

Ďalším prístupom založeným na modeli, ktorý možno použiť v teórii superpopulácie, je kalibrácia s asistenciou modelu (*model-assisted calibration*). Tento prístup zahŕňa konštrukciu kalibrovaných váh pomocou modelu na predikcie hodnôt študovanej premennej. Kalibrované váhy sa generujú na základe obmedzení na rozsah a úhrn predikovaných hodnôt v základnom súbore. Na modelovanie študovanej premennej boli navrhnuté rozličné prístupy k výberu modelov (pozri napríklad v [8]). Kľúčovým predpokladom metódy je, že model je korektné špecifikovaný a schopný robiť spoľahlivé predikcie na báze rôznych výberov zo základného súboru. Tiež sa predpokladá, že v modeli sú zahrnuté všetky relevantné parametre superpopulácie.

Ďalším prístupom používaným na odhadovanie špecifického výstupu na báze nenáhodného výberu je viacúrovňová regresia (*multilevel regression*) a poststratifikácia [4]. Základnou myšlienkou je odhadnúť viacúrovňový regresný model¹⁷ na predikciu výstupu pre danú množinu kovariát. Použitie viacúrovňového modelu umožňuje začlenenie veľkého počtu kovariát alebo interakcií vyššieho rádu do predikčného modelu. Model sa potom použije na odhad strednej hodnoty pre veľký počet poststratifikačných buniek definovaných krížovou klasifikáciou všetkých premenných použitých v regresnom modeli. Je potrebné, aby bola známa relatívna veľkosť každej bunky v základnom súbore alebo aby sa dala spoľahlivo odhadnúť

¹⁷ Viacúrovňové modely sú štatistické modely parametrov, ktoré sa menia na viac ako jednej úrovni. Napríklad model výkonnosti študenta obsahuje miery pre individuálneho študenta aj miery pre triedy do ktorých sú študenti zaradení. Tieto modely možno považovať za zovšeobecnenie lineárnych modelov (podrobnejšie pozri v [14]).

z externých zdrojov dát, ako je cenzus alebo register obyvateľstva. Veličiny základného súboru sa odhadujú agregovaním týchto predikovaných stredných hodnôt buniek, pričom každá bunka sa váži úmerne k jej podielu na základnom súbore. Viacúrovňový regresný model umožňuje generovanie odhadov na úrovni buniek, aj keď bunky vo výbere obsahujú len niekoľko jednotiek. Rovnako ako všetky vyššie uvedené prístupy založené na modeli sa predpokladá, že model riadi všetky premenné, ktoré ovplyvňujú pravdepodobnosť zahrnutia v nenáhodnom výbere. Metóda vyžaduje dobrý odhad modelu, preferujú sa veľké rozsahy buniek, aby sa na úrovni buniek generovali robustné hodnoty odhadov [3, s. 13].

Celkovo tieto prístupy k používaniu nenáhodných výberov na získanie induktívnych záverov o základnom súbore závisia od kombinácie predpokladov o ignorovateľných chybách a dostupnosti informácií o zdrojoch nenáhodnosti pri výbere respondentov, ktoré možno použiť na takú úpravu, ktorá redukuje relevantné chyby. Hoci sa tieto závislosti zásadne nelíšia od predpokladov, ktoré sú základom použitia náhodných výberov, je ťažšie sa na ne spoľahnúť, pretože vieme málo o faktoroch, ktoré vedú jednotlivcov k tomu, aby sa zaradili do nenáhodných výberov.

3.2.4. Presnosť v štatistických prieskumoch na základe náhodných a nenáhodných výberov

Množstvo štúdií empiricky hodnotilo presnosť prieskumov na základe náhodných a nenáhodných výberov, pomocou porovnania získaných výsledkov s výsledkami z vyčerpávajúceho skúmania alebo z vysoko kvalitných prieskumov na základe náhodných výberov z príslušných základných súborov. Porovnania sa týkali rozličných krajín, rozličných tém a časového rámca od roku 2007 do roku 2018. Všetky tieto štúdie dospeli k záveru, že prieskumy založené na náhodných výberoch vedú k presnejším hodnotám odhadov ako v prípade nenáhodných výberov ([3], s. 18). Viaceré štúdie dospeli k záveru, že váženie nedostatočne redukuje vychýlenie v prieskumoch na báze nenáhodných výberov. Prieskumy na báze náhodných výberov sú všeobecne presnejšie ako prieskumy na báze nenáhodných výberov, aj keď sa v nich použilo váženie [3, s. 21]. Celkovo, väčšina štúdií o vážení a presnosti prišla k záveru, že nepresnosť prieskumov založených na nenáhodných výberoch nemožno spoľahlivo odstrániť využitím váhových procedúr. Okrem toho niektorí autori ukazujú, že na získanie presných odhadov z prieskumov založených na nenáhodných výberoch je dôležitejšia dostupnosť premenných, ktoré predikujú výstupy, ako to, ktorá štatistická metóda sa použila [3, s. 22].

Treba poznamenať, že jedným z problémov podobných štúdií je oddeliť účinky režimu (*mode effects*)¹⁸ od účinku vyberania (výberové vychýlenie). Je to dané tým, že prieskumy založené na náhodných výberoch sa väčšinou realizujú off-line (osobne alebo telefonicky), zatiaľ čo prieskumy založené na nenáhodnom výbere sa väčšinou realizujú online prostredníctvom online nenáhodných panelov. V mnohých štúdiách sa však uvádza, že je možné rozlíšiť tieto dva účinky porovnaním off-line a online prieskumov založených na náhodných výberoch (účinnok režimu) a porovnaním online prieskumov založených na náhodných a nenáhodných výberoch (účinnok vyberania). Väčšina týchto štúdií dospela k záveru, že off-line aj online prieskumy založené na náhodných výberoch sú presnejšie ako online prieskumy založené na nenáhodných výberoch [3, s. 19].

¹⁸ Ide o vychýlenie spôsobené meraním.

Ďalšou otázkou je, ako merať presnosť prieskumu tak, aby sa brala do úvahy súčasne variabilita vyberania aj systematické vychýlenie. Väčšinou je totiž k dispozícii len jeden prieskum na základe náhodného a jeden na základe nenáhodného výberu. Sú však známe aj štúdie v ktorých sa porovnáva väčší počet prieskumov na báze náhodného výberu s väčším počtom prieskumov na báze nenáhodných výberov. V nich autori dospeli k záveru, že prieskumy na báze náhodných výberov sú konzistentne presnejšie ako na báze nenáhodných výberov. To naznačuje, že aj keď sú prieskumy variabilné v mnohých faktoroch (napr. rozličné motivačné schémy, rozličná frekvencia kontaktov a pod.) a niekedy môžu byť viac alebo menej presné náhodou, výbery majú všeobecne väčšiu vierohodnosť viesť k presnejším výsledkom, keď sa získali procedúrami náhodného vyberania ako keď boli získané procedúrami nenáhodného vyberania [3, s. 19 – 20].

Problémom môže byť aj prístupnosť rozsiahlych, vysokokvalitných štatistických prieskumov, ktoré môžu slúžiť ako benchmark. V [3, s. 20] sa uvádzajú niektoré príklady takýchto prieskumov. Všeobecne ide o výsledky z rozsiahlych vysokokvalitných prieskumov na báze náhodných výberov, prieskumov na báze dát z cenzu a z registra obyvateľov a na báze dát z výsledkov volieb. Čo sa týka vysokokvalitných prieskumov na báze náhodných výberov, ich hlavnou prednosťou je, že množina premenných prístupných na porovnanie obyčajne zahŕňa nielen sociálnodemografické, ale aj premenné o postojoch a správaní. Ich nevýhodou je, že môžu obsahovať také typické chyby, ako chyby pokrytia, výberové chyby a chyby neodpovedania. Dáta z cenzu a z registra obyvateľov majú výhodu, že netrpia chybami prieskumu. Na druhej strane dáta z cenzu často nie sú prístupné za aktuálny rok a môžu byť v čase realizácie štúdie neaktuálne. To platí aj pre dáta z registra obyvateľstva, kde nemusí byť presne zachytená emigrácia, imigrácia, prípadne nie sú včas zaznamenané dáta o narodeniach a úmrtiach. Navyše dáta z cenzu a z registra obyvateľov sú typické aj malým počtom sociálnodemografických charakteristík. Čo sa týka výsledkov volieb, tie obyčajne obsahujú veľa zaujímavých premenných hlavne pre výskum v sociálnej oblasti, ak však prieskum neodhadne výsledky volieb pomerne presne, existuje mimo použitej výberovej schémy aj množstvo iných vysvetlení.

4. ZÁVER

Cieľom príspevku bolo podať súhrnný prehľad o hodnotení štatistických prieskumov založených na nenáhodných výberoch. Teória o náhodnom vyberaní je založená na známych poznatkoch z teórie pravdepodobnosti. Nenáhodné vyberanie takúto teoretickú bázu nemá. Zdôvodnenie mnohých štúdií zostáva nejasné a závery z prieskumov na báze nenáhodných výberov vo všeobecnosti vyžadujú silnejšie predpoklady modelovania, než aké sú potrebné v prieskumoch na báze náhodných výberov. Základným problémom týchto modelovacích predpokladov zostáva, že ich nemožno testovať.

Uviedli sme koncepčný rámec prieskumov na báze nenáhodných výberov, ktorý umožňuje vysvetliť tieto predpoklady modelovania vrátane praktických návrhov, kedy by mohlo byť opodstatnené takéto predpoklady použiť. Okrem toho uvádzame závery ktoré vyplývajú zo štúdií o empirických dôkazoch o presnosti prieskumov založených na náhodných a nenáhodných výberoch. Štúdie ukazujú, že aj v dobe rastúcej miery neodpovedania v prieskumoch na báze náhodných výberov je ich presnosť vyššia ako v prípade nenáhodných výberov. Všeobecne možno odporúčať preferovať tak ako doteraz prieskumy na báze náhodných výberov.

Keď sú k dispozícii len dáta z prieskumu na báze nenáhodného výberu, možno odporúčať pozorne vyberať z rôznych prístupov modelovania, ktorých základný prehľad sme uviedli, na báze ich základných predpokladov. Kvôli zdôvodneniu použitého prístupu modelovania, je užitočné získať čo najviac informácií o procese generovania dát. To vyžaduje čo najväčšiu transparentnosť, čo sa týka informácií o zhromažďovaní a spracovaní dát v paneloch. Bez toho štatistikom chýbajú informácie potrebné na pochopenie obmedzení dát zhromaždených ich poskytovateľom. Preto je veľmi dôležité poznať metodiku, ktorá sa použila na zber dát a na manipuláciu s nimi. Zákazník by mal zasa dať poskytovateľovi dát z prieskumu na báze nenáhodného výberu čo najpresnejšie informácie o svojich potrebách. Ak predajca dát patrí do organizácie ESOMAR¹⁹ alebo je certifikovaný podľa ISO 20252 alebo 26362, príslušný kódex alebo certifikácia vyžaduje zverejňovanie istých informácií.

Často sa v praxi na základe nenáhodného výberu indukujú závery o základnom súbore bez toho, že by sa zobrali do úvahy všetky informácie, ktoré by mohli naznačovať potrebu aplikácie niektorého z prístupov modelovania. Takýto postup je nevyhnutné považovať za subjektívny. Prípadné použitie vhodného prístupu modelovania môže znížiť subjektivitu pohľadu. Ak by sme uvažovali o spojitých stupnici subjektivity vo výberovom skúmaní, potom prijímanie záverov o základnom súbore na báze nenáhodného výberu bez uváženia možností aplikácie vhodného prístupu modelovania je jednoznačne subjektívny prístup. Na opačnej strane stupnice je prijímanie záverov o základnom súbore na báze náhodného výberu, v ktorom neexistuje problém s pokrytím a s neodpovedaním. Vtedy možno použiť niektorú z procedúr prístupu založeného na pláne výberového skúmania. Ide o objektívne odôvodniteľný najlepší postup odhadovania veličín konečného základného súboru.

LITERATÚRA

- [1] BETHLEHEM, J.: Applied Survey Methods. A Statistical Perspective. Hoboken: John Wiley & Sons, Inc., 2009.
- [2] BETHLEHEM, J.: Understanding Public Opinion Polls, Boca Raton, FL: Chapman and Hall/CRC, 2017.
- [3] CORNESSE, C. – BLOM, A. G. – DUTWIN, D. – KROSNICK, J. A. – DE LEEUW, E. D. – LEGLEYET, S. – PASEK, J. – PENNAY, D. – PHILLIPS, B. – SAKSHAUG, J. W. – STRUMINSKAYA, B. – WENZ, A.: A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. In: Journal of Survey Statistics and Methodology, 2020, č. 8, s. 4 – 36.
- [4] DOWNES, M. – GURRIN, L. C. – ENGLISH, D. R. – PIRKIS, J. – CURRIER, D. – SPITTAL, M. J. – CARLIN, J. B.: Multilevel Regression and Poststratification: A Modelling Approach to Estimating Population Quantities from Highly Selected Survey Samples, In: American Journal of Epidemiology, 2018, č. 8, s. 1780 – 1790.
- [5] ELLIOTT, M. R. – VALLIANT, R.: Inference for Nonprobability Samples. In: Statistical Science, 2017, č. 2, s. 249 – 264.

¹⁹ ESOMAR (The European Society for Opinion and Marketing Research) je Európske združenie pre marketingový výskum a výskum verejnej mienky. Všetci členovia ESOMAR-u a vedenie ich marketingových výskumných inštitúcií podpísali vyhlásenie, že budú vo všetkých svojich výskumných aktivitách dodržiavať ustanovenia Medzinárodného kódexu praktík marketingových a sociálnych výskumov, ktorý vydala Medzinárodná obchodná komora spolu s ESOMAR-om [9].

- [6] HARTER, R. – BATTAGLIA, M. P. – BUSKIRK, T. D. – DILLMAN, D. A. – ENGLISH, N. – FAHIMI, M. – FRANKEL, M. R. – KENNEL, T. – McMICHAEL, J. P. – McPHEE, C. B. – MONTAQUILA, J. – YANCEY, T. – ZUKERBERG, A. L.: Address-Based Sampling. AAPOR, 2016. [cit. 2024-02-13]. Dostupné na: [Month 200X \(aapor.org\)](#).
- [7] CHA, A. E. – COHEN, R. A.: Demographic Variation in Health Insurance Coverage: United States, 2021, National Health Statistics Reports, Number 177, November 3, 2022. [cit. 2024-02-13]. Dostupné na: [cdc_121999_DS1 \(1\).pdf](#).
- [8] CHEN, J. K. T. – VALLIANT, R. L. – ELLIOTT M. R.: Calibrating Non-Probability Surveys to Estimated Control Totals Using LASSO, with an Application to Political Polling. In: Journal of the Royal Statistical Society: Series C (Applied Statistics), 2019, č. 3, s. 657–681.
- [9] Kódex medzinárodnej obchodnej komory a ESOMAR-u. [cit. 2024-02-13]. Dostupné na: <https://sava.sk/standardy-kvality/esomar-kodex/>.
- [10] LEVY, P. S. – LEMESHOW, S.: Sampling of Populations. Methods and Applications. Fourth Edition. Hoboken: Wiley, 2008.
- [11] Logit. [cit. 2024-03-8]. Dostupné na: [Logit – Wikipedia](#).
- [12] LOHR, S. L.: Sampling: Design and Analysis. Boca Raton: Taylor & Francis Group, 2019.
- [13] MERCER, A. W. – KREUTER, F. – KEETER, S. – STUART, E. A.: Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference. In: Public Opinion Quarterly, 2017, č. S1, s. 250 – 271. [cit. 2024-02-13]. Dostupné na: [Theory and Practice in Nonprobability Surveys | Public Opinion Quarterly | Oxford Academic \(oup.com\)](#).
- [14] Multilevel model. Dostupné na: [Multilevel model – Wikipedia](#).
- [15] River Sampling Versus RDE Sampling: Which is Superior for Market Research. [cit. 2024-02-13]. Dostupné na: <https://resources.pollfish.com/market-research/river-sampling-versus-rde-sampling-which-is-superior-for-market-research/>.
- [16] RIVERS, D.: Sampling for web surveys. [cit. 2024-02-13]. Dostupné na: [Rivers_matching4.pdf \(texastribune.org\)](#).
- [17] ROSHWALB, A. – LEWIS, Z. – PETRIN, R.: The Efficacy of Nonprobability Online Samples, JSM Proceedings, Survey Research Methods Section, 2016, pp. 3657–3666. [cit. 2024-02-13]. Dostupné na: <http://www.asasrms.org/Proceedings/y2016/files/389791.pdf>.
- [18] Sampling error. [cit. 2024-02-13]. Dostupné na: [Sampling error - Wikipedia](#).
- [19] Sampling in the digital age. [cit. 2024-02-13]. Dostupné na: <https://www.cint.com/blog/sampling-in-the-digital-age>.
- [20] SÄRNDAL, C.-E. – LUNDSTRÖM, S.: Estimation in Surveys with Nonresponse. Hoboken: Wiley, 2005.
- [21] SCHONLAU, M. – COUPER, M. P.: Options for Conducting Web Surveys. In: Statistical Science, 2017, č. 2, s. 279 – 292.
- [22] TEREK, M. – HRNČIAROVÁ, Ľ.: Výberové skúmanie. Bratislava: Ekonóm, 2008.
- [23] TEREK, M.: Možnosti riešenia problému neodpovedania v štatistických prieskumoch. In: Ekonomické rozhľady, 2014, č. 2., s. 150 – 165.
- [24] TEREK, M.: Interpretácia štatistiky a dát. 5., doplnené vydanie. Košice: Equilibria, 2017.
- [25] TEREK, M.: Dotazníkové prieskumy a analýzy získaných dát. Košice: Equilibria, 2019.

RESUMÉ

Výberové skúmanie možno realizovať na báze náhodného alebo nenáhodného výberu. V článku si všimame výsledky výskumu v oblasti hodnotenia kvality nenáhodných výberov a možností jej zlepšovania prostredníctvom ich úprav. Hlboká znalosť možností, ktoré poskytuje tento prístup je mimoriadne užitočná pri riešení praktických problémov odhadovania veličín základného súboru. Uvažujeme o konečnom základnom súbore. Štatistické prieskumy na báze náhodných aj nenáhodných výberov majú často spoločný cieľ: efektívne odhadnúť veličiny základného súboru na základe výsledkov analýzy výberového súboru. Nenáhodné vyberanie možno všeobecne rozdeliť do troch skupín: vyberanie založené na prístupnosti jednotiek, vyberanie založené na zhode a sieťové vyberanie. Uvádzame možnosti hodnotenia kvality náhodných a nenáhodných výberov.

V prípade prieskumov na báze náhodných výberov sa možno oprieť o teóriu pravdepodobnosti, ktorá je založená na súbore zavedených matematických princípov. Základným kameňom hodnotenia kvality náhodných výberov je celková chyba prieskumu. Ide o rozdiel medzi veličinou základného súboru a hodnotou odhadu tejto veličiny na báze analýzy výberu. Má dve zložky: výberovú chybu a nevýberovú chybu. Tradične sa výberové vychýlenie považuje za výsledok nepokrytia alebo neodpovedania. Tieto koncepty sú spojené s náhodným výberom z konečného základného súboru. Tieto kategórie sa môžu ukázať ako limitujúce, ak sa aplikujú v kontexte nenáhodných výberov. Mnohé prieskumy na báze nenáhodných výberov nemožno spájať s výbermi z nejakej výberovej bázy. Dokonca aj predstava výberu ako konečného súboru jednotiek, z ktorých niektoré nemusia odpovedať, sa nevzťahuje na mnohé prieskumy na báze nenáhodných výberov. Pri takýchto prieskumoch môžu byť procesy zaradenia respondenta do výberu rôznorodé, niekedy úplne nepodobné procesu zaradenia jednotky do výberu v prieskumoch na báze náhodných výberov. V hodnotení kvality nenáhodných výberov sa odôvodnenie očakávania presných záverov môže opierať len o netestované predpoklady modelovania a možno ju ohodnotiť len posúdením toho, do akej miery sa konečný modelovaný výber zhoduje so základným súborm z hľadiska rôznych charakteristík. Uvádzame niektoré možné úpravy nenáhodných výberov, ktoré môžu slúžiť na odôvodnenie ich použitia na induktívne úsudky o veličinách základného súboru. Uvádzame aj výsledky niektorých empirických štúdií, ktoré hodnotia presnosť prieskumov na základe náhodných a nenáhodných výberov. Ich výsledkom je záver, že náhodné vyberanie je bez akejkoľvek diskusie najvhodnejší prístup k tvorbe induktívnych úsudkov o základnom súbore.

Keď sú k dispozícii len dáta z nenáhodného výberu, možno odporúčať starostlivo uvážiť všetky informácie, ktoré by mohli naznačovať potrebu aplikácie niektorého z prístupov modelovania, ktorých základný prehľad je uvedený. Kvôli zdôvodneniu použitého prístupu modelovania, je užitočné získať čo najviac informácií o procese generovania dát. To vyžaduje čo najväčšiu transparentnosť, čo sa týka informácií o zhromažďovaní a spracovaní dát v paneloch.

Často sa v praxi na základe nenáhodného výberu indukujú závery o základnom súbore bez toho, že by sa zobrali do úvahy všetky informácie, ktoré by mohli naznačovať potrebu aplikácie niektorého z prístupov modelovania. Takýto postup je nevyhnutné považovať za subjektívny. Prípadné použitie vhodného prístupu modelovania môže znížiť subjektivitu pohľadu.

RESUME

Sample surveys can be carried out based on probability or nonprobability samples. In the article, we note the results of research in the field of quality assessment of nonprobability samples and the possibilities of their improvement through modifications. A deep knowledge of the possibilities provided by this approach is extremely useful in solving practical problems of estimating population quantities.

We are considering a finite population. Statistical surveys based on both probability and nonprobability samples often share a common goal: to efficiently estimate population quantities based on the results of the analysis of the sample. Nonprobability sampling can be broadly divided into three groups: convenience sampling, sample matching, and network sampling. We present the possibilities of evaluating the quality of probability and nonprobability samples.

In case of surveys based on probability samples, we can rely on the theory of probability, which is based on a set of established mathematical principles. The total survey error is the basis for assessing the quality of probability samples. It is the difference between the population quantity and the estimate of this quantity based on the sample analysis. It has two components: sampling error and non-sampling error. Selection bias is typically caused by noncoverage or nonresponse. These concepts are associated with probability sampling from a finite population. These categories can prove to be limiting when applied in the context of nonprobability samples. Many surveys based on nonprobability samples cannot be linked to the samples from a sampling frame. Even the idea of a sample as a finite set of units, some of which may not respond, does not apply to many nonprobability surveys. In such surveys, the respondent selection processes can be diverse, sometimes completely dissimilar to those in probability sampling surveys. In assessing the quality of nonprobability samples, the justification for expecting accurate conclusions can only rely on untested modeling assumptions and can only be assessed by considering how well the final modeled sample matches the population in terms of various characteristics. We present some possible modifications of nonprobability samples that may serve to justify their use for inferential judgments about the population quantities. We also present the results of some empirical studies that evaluate the accuracy of surveys based on probability and nonprobability samples. They conclude that probability sampling is, undoubtedly, the most appropriate approach to making inferential judgments about the population.

If only data from a nonprobability sample are available, it may be advisable to carefully consider any information that might indicate the need to apply any of the modeling approaches whose basic overview is given. To justify the modeling approach, obtaining as much information as possible about the data generation process is useful. This requires as much transparency as possible regarding information on the data collection and processing in the panels.

Often, in practice, based on a nonprobability sample, conclusions about the population are inferred without considering all the information that could indicate the need to apply one of the modeling approaches. Such a procedure must be considered subjective. The possible use of an appropriate modeling approach can reduce the subjectivity of the view.

PROFESIJNÝ ŽIVOTOPIS

Prof. Ing. Milan Terek, PhD., od roku 2018 pracuje ako profesor na Vysokej škole manažmentu v Bratislave. Vedie predmety základy 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 predmety 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, 80 článkov vo vedeckých a odborných časopisoch a 115 príspevkov na vedeckých konferenciách publikovaných v zborníkoch.

KONTAKT

mterek@vsm.sk