FORUM STATISTICUM SLOVACUM Vol XVI (2020) No 1 pp 57-70

Určenie veľkosti vzorky pre aplikovaný výskum
Sample size determination in applied statistical research

Iveta Stankovičová[a], Boris Frankovič[b]

[a] Univerzita Komenského v Bratislave, Fakulta managementu, Katedra informačných systémov, Odbojárov 10, 820 05 Bratislava, Slovenská republika

[a] Comenius University in Bratislava, Faculty of Management, Department of Information Systems, Odbojárov 10, 820 05 Bratislava, Slovak Republic

[b] Štatistický úrad Slovenskej republiky, Odbor metód štatistických zisťovaní, Miletičova 3, 824 67 Bratislava, Slovenská republika

[b] Statistical Office of the Slovak Republic, Statistical Surveys and Methodology Department, Miletičova 3, 824 67 Bratislava, Slovak Republic

iveta.stankovičová@fm.uniba.sk; boris.frankovic@statistics.sk

Abstrakt: *Aplikovaný štatistický výskum zohráva kľúčovú úlohu pri riešení problémov v rôznych oblastiach ekonomických, spoločenských, technických, poľnohospodárskych vied ale aj zdravotníctva. Mnohé zisťovania sa vykonávajú výberovým spôsobom. Technika výberu vzorky a určenie veľkosti vzorky majú rozhodujúcu úlohu pri záveroch o populácii na základe vzorky vo výskumných problémoch. Pre špecifické výskumné problémy sa používajú osobitné techniky výberu vzoriek, pretože jedna technika nemusí byť vhodná pre všetky problémy. Podobne, ak je nevhodná veľkosť vzorky, môže to viesť k nesprávnym záverom. Tento článok predstavuje zhrnutie spôsobu výpočtu veľkosti vzorky pre empirické štúdie v oblasti sociálnych a ekonomických výskumov.*

Abstract: *Applied statistical research plays a key role in solving problems in various fields of economic, social, technical, agricultural sciences and health care. Many surveys are carried out on a sampling basis. The sampling technique and the determination of the sample size have a decisive role in the conclusions about the population based on the sample in research problems. Specific sampling techniques are used for specific research problems because one technique may not be appropriate for all problems. Similarly, if the sample size is inappropriate, this may lead to incorrect conclusions. This article summarizes how sample size is calculated for empirical studies in social and economic research.*

Kľúčové slová*: populácia, výber, stratifikovaný výber, rozsah vzorky, interval spoľahlivosti, hladina významnosti*

Keywords: *population, sampling, stratification sampling, sample size, confidence interval, level of significance*

1 Úvod

Koľko odpovedí od respondentov na dotazník skutočne potrebujeme? Táto jednoduchá otázka je pre vedcov vždy veľkou dilemou. Väčšia vzorka môže priniesť presnejšie výsledky pre odhady a testovanie rôznych hypotéz, ale na druhej strane uskutočniť rozsiahly výskum môže byť veľmi drahé.

Na začiatku výskumu si teda veľmi často kladieme nasledovné otázky: „Akú veľkú výberovú vzorku potrebujeme dosiahnuť?“ „Aká veľká výberová vzorka je

postačujúca?“ „Koľko štatistických jednotiek potrebujeme vybrať, aby naše odhady boli spoľahlivé?“ Odpovedať na tieto otázky je cieľom tohto článku.

Výber vzorky je starý koncept. Alexander Ivanovič Čuprov začal s výberovými zisťovaniami v cárskom Rusku už v roku 1870. Pierre Simon Laplace odhadol počet obyvateľov Francúzska pomocou tejto techniky v roku 1786.

Výberové zisťovania sú založené na výbere podskupiny jednotlivcov z populácie s cieľom odhadnúť charakteristiky (parametre) pre celú populáciu. Výberové skúmanie má dve hlavné výhody: 1. rýchlejšie získavanie údajov a 2. nižšie náklady (Kish, 1965). Dnes je výberové skúmanie veľmi rozšírené a používa sa na skúmanie javov v rôznych oblastiach: vo výrobných a iných podnikoch, v spoločnosti, v medicíne, vo farmácii, v poľnohospodárstve a podobne. Hlavnou nevýhodou je, že ak vzorku nevyberieme reprezentatívne, tak dostaneme skreslené, resp. chybné výsledné odhady parametrov populácie.

Cieľom predkladaného článku je popísať kritériá a postup pre výpočet veľkosti minimálnej efektívnej vzorky. Predstavíme metódy výberu vzorky a pre dve najčastejšie používané metódy zadefinujeme požadované vstupy, aby sme dokázali stanoviť, resp. vypočítať veľkosť vzorky. Prínosom článku je zhrnutie poznatkov a postupov pri tvorbe reprezentatívnej vzorky z populácie pravdepodobnostným výberom. V empirickej časti uvádzame praktickú ukážku výpočtu veľkosti vzorky z populácie podnikov SR na základe stratifikovaného náhodného výberu pomocou proporcionálnej alokácie s použitím dvoch stratifikačných premenných.

2 Druhy a metódy výberu vzoriek

Pre tvorbu výberových súborov (vzoriek) z populácie môžeme použiť pravdepodobnostný (náhodný) alebo nepravdepodobnostný (nenáhodný) výber. Výbery bez pravdepodobnosti sú také, ktoré nie sú založené na náhode. Pri tomto type výberu neexistuje spôsob, ako presne odhadnúť pravdepodobnosť, že bude vybraný nejaký prvok. Kvalita takejto vzorky závisí od znalostí, úsudku a odbornosti výskumného pracovníka. Vytvoriť takéto vzorky je síce pomerne pohodlné a hospodárne, ale nie je pri nich vopred známe riziko nesprávneho zovšeobecnenia na populáciu.

Pravdepodobnostný (náhodný) výber je taký, v ktorom každý prvok v populácii má známu nenulovú pravdepodobnosť výberu do vzorky (Sudman, 1976, s. 49). Pretože pravdepodobnosť je známa, výberové štatistiky (napr. priemer, podiel, regresné parametre) môžu byť zovšeobecnené pre celú populáciu (aspoň v rámci danej úrovne presnosti). Pravdepodobnostné výbery sa vo všeobecnosti uprednostňujú pred inými druhmi výberov, pretože pri nich je vopred známe riziko nesprávneho zovšeobecnenia na populáciu.

Opora výberu (sampling frame) je zoznam prvkov (štatistických jednotiek), z ktorých bude vzorka vyberaná. Ideálna opora výberu uvádza každý prvok samostatne a iba raz a nič iné sa neobjaví v zozname štatistických jednotiek. Ak opora výberu nie je korektná, tak musíme vykonať opravy, napr. odstrániť duplicity prvkov, vylúčiť jednotky (prvky), ktoré už v danom čase nie sú súčasťou sledovanej populácie (nadmerné pokrytie) a naopak zaradiť jednotky, ktoré sa medzičasom stali súčasťou sledovanej populácie (nedostatočné pokrytie). Ideálna opora musí byť aktuálna, dosiahnuteľná a nesmie obsahovať spomínané chyby.

Po získaní korektnej opory výberu je potrebné sa rozhodnúť pre vhodný postup výberu vzorky z populácie. Ak sme sa rozhodli pre pravdepodobnostný (náhodný) výber, tak existuje niekoľko metód, ktorých použiteľnosť je determinovaná konkrétnymi podmienkami, v ktorých sa výberové skúmanie realizuje. Poznáme nasledovné metódy náhodného výberu:

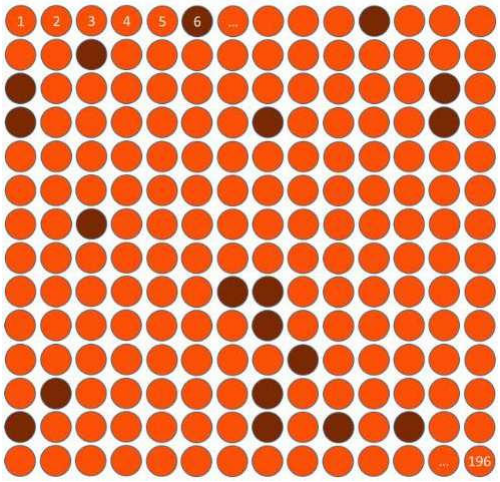
- jednoduchý náhodný výber (simple random sample),
- stratifikovaný náhodný výber (stratified random sample),
- systematický výber (systematic sample),
- oblastný výber (cluster or area sample).

Všetky vyššie uvedené druhy náhodného, resp. pravdepodobnostného výberu sú názorne zachytené na obrázkoch 1 až 4. Obrázky ilustrujú jednotlivé metódy náhodného výberu z populácie 196 loptičiek (*N*), z ktorej vyberáme vzorku 20 loptičiek (*n*).

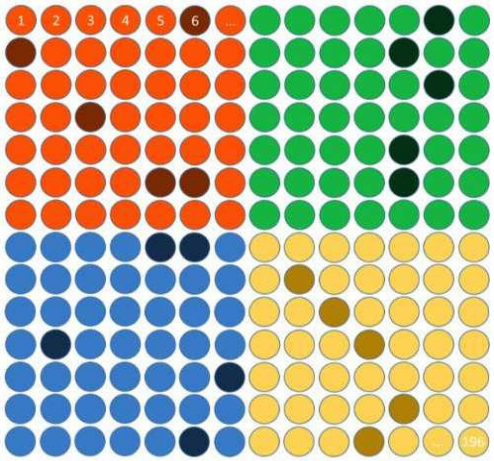
Jednoduchý náhodný výber je najjednoduchšia a najmenej zložitá metóda tvorby vzorky. Pri tejto metóde, t.j. jednoduchý náhodný výber s opakovaním, má každý prvok opory výberu rovnakú pravdepodobnosť výberu do vzorky. Predpokladajme, že vzorka o rozsahu 20 loptičiek (*n*) sa vyberá z populácie o rozsahu 196 loptičiek (*N*) jednoduchým náhodným výberom. Pomocou generátora náhodných čísiel zistíme 20 čísiel, ktoré označujú poradie loptičiek, ktoré majú byť vybraté do vzorky z populácie. Na obr. 1 sú znázornené náhodne vybrané loptičky do vzorky tmavou farbou.

Stratifikovaný náhodný výber sa používa na zlepšenie odhadov parametrov z heterogénnej populácie. Celá populácia sa rozdelí do skupín alebo vrstiev (strát) na základe vybraných charakteristík, tzv. stratifikačných premenných. Voľba stratifikačných premenných vyplýva z účelu a kontextu výberového skúmania. Vek, pohlavie a bydlisko sú niektoré demografické charakteristiky, ktoré sú bežne používané na stratifikáciu vzoriek. V každej vrstve (strate) sa potom vyberajú prvky náhodným výberom. Pri tejto metóde je potrebné, aby sme získali informácie o populácii pred procesom odberu vzoriek. Stratifikované výbery poskytujú presnejšie odhady ako jednoduché náhodné výbery, pretože

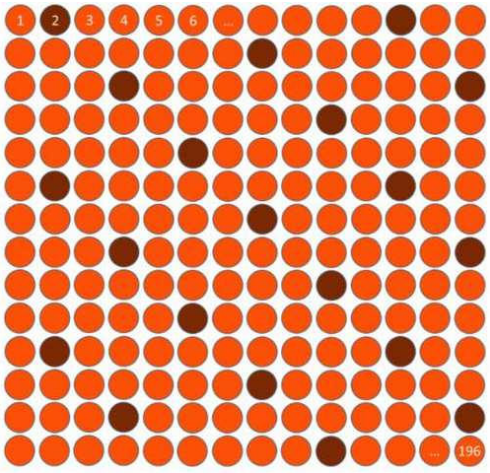
každá skupina (strata) populácie je vo vzorke dobre zastúpená. Výberová chyba závisí od rozptylu populácie vo vrstve, ale nie medzi vrstvami. Na obr. 2 je populácia loptičiek rozdelená do štyroch strát podľa farby a v každej strate bolo vybrané náhodne 5 loptičiek, aby rozsah vzorky bol n = 20.



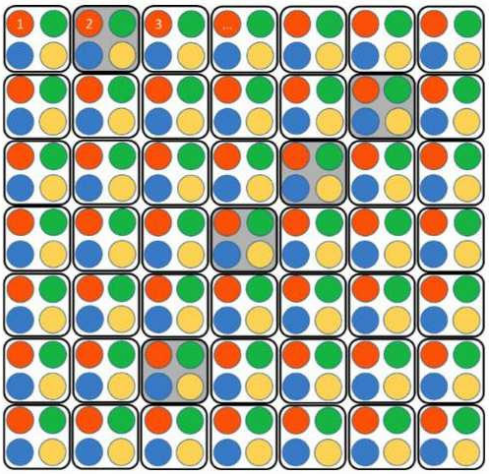
Obr. 1 Jednoduchý náhodný výber (n = 20) *(Zdroj: Israel, 1992)*



Obr. 2 Stratifikovaný výber (4 straty, n = 20) *(Zdroj: Israel, 1992)*



Obr. 3 Systematický výber prvkov (n = 20) *(Zdroj: Israel, 1992)*



Obr. 4 Oblastný výber 5 oblastí (n = 20) *(Zdroj: Israel, 1992)*

Systematický náhodný výber sa bežne používa a dá sa ľahko implementovať. Tak ako jednoduchý náhodný výber, každý prvok populácie má rovnakú pravdepodobnosť sa dostať do vzorky (ale nie nezávislú). Prvý prvok sa vyberie náhodne a potom každý ďalší prvok z opory výberu sa vyberá systematicky. Predpokladajme, že vzorka 20 loptičiek sa vyberá z populácie o rozsahu 196 loptičiek (20/196 = 0,10204 ≈ 10 %). Interval medzi prvou náhodne vybranou loptičkou a druhou bude 10, atď. Na obr. 3 je z populácie loptičiek náhodne

vybraná druhá loptička a potom systematicky s krokom 10 sú vyberané ďalšie loptičky (postupne v riadkoch označené tmavšou farbou). V prípade výskytu nejakého vzorca, resp. periodicity v usporiadaní populácie tento výber však prináša skreslenie odhadov parametrov populácie.

Oblastný náhodný výber je metóda výberu, pri ktorej je celá populácia rozdelená do skupín, resp. oblastí, zhlukov. Náhodne sa vyberajú celé tieto skupiny, čiže do vzorky sa dostanú všetky pozorovania (prvky) z náhodne vybraných skupín (zhlukov). Oblastný výber je vhodný vtedy, keď sa v populácii prejavujú „prirodzené“, ale relatívne homogénne skupiny. Všeobecne sa používa vtedy, keď výskumný pracovník nemá a nemôže získať úplný zoznam prvkov populácie, ktorú chce skúmať, ale môže získať úplný zoznam skupín alebo „zoskupení“ populácie. Táto metóda je často viac praktická a ekonomická ako jednoduchý náhodný alebo stratifikovaný výber. Je vhodná hlavne vtedy, ak neexistuje zoznam celej populácie, je ťažko dostupný alebo drahý. Tvorba vzorky pomocou skupín, resp. oblastí (zhlukov) uľahčuje celý proces výberu. Na obr. 4 je populácia loptičiek rozdelená na 49 oblastí, ktoré obsahujú zhodne po 4 prvky (7 x 7 x 4 = 196). Z nich je náhodne vybraných 5 oblastí (skupín), čiže bola vybraná vzorka o rozsahu 20 loptičiek (5 x 4 = 20).

V porovnaní s jednoduchým náhodným a stratifikovaným výberom má oblastný (skupinový) výber výhody a nevýhody. Napríklad pri rovnakých veľkostiach vzoriek oblastný výber obvykle poskytuje menšiu presnosť odhadov ako jednoduchý náhodný alebo stratifikovaný výber. Na druhej strane, ak sú kontaktné náklady medzi skupinami (zhlukmi) veľmi vysoké, oblastný výber môže byť nákladovo efektívnejší ako iné metódy.

3 Kritériá pre veľkosť vzorky

Stanovenie veľkosti vzorky závisí od niekoľkých faktorov. Na prvom mieste si musíme určiť cieľ a účel výskumu (štúdie). Keď je nám to jasné, tak na výpočet veľkosti minimálnej efektívnej výberovej vzorky si potrebujeme ďalej ujasniť tieto vstupy, resp. kritériá:

1. *Veľkosť populácie (population size)*. Koľko respondentov celkovo vyhovuje požadovanej demografickej skupine/skupinám? Napríklad, ak robíme výskum o priemyselných podnikoch v SR, veľkosť populácie bude celkový počet priemyselných podnikov v SR evidovaných v registroch, tzv. opora výberu. V praxi je však časté, že veľkosť populácie nie je známa alebo je len aproximovaná.
2. *Prípustná chyba odhadu (sampling error, margin of sampling error) pre intervaly spoľahlivosti (confidence interval)*. Musíme sa rozhodnúť, aká veľkosť chyby je pre nás prípustná ešte pred samotným výskumom. V správach o výsledkoch politických prieskumov, sa často vyskytuje výrok typu: „*Na danú*

otázku odpovedalo „áno“ 68 % voličov s prípustnou chybou +/- 5 %.“ Čiže na základe prieskumu sa podiel voličov v populácii, ktorí odpovedali na položenú otázku „áno“ bude pohybovať v intervale (63 %, 73 %). Prípustná chyba vlastne vyjadruje maximálnu šírku intervalov spoľahlivosti pre odhady parametrov populácie, ktorú chceme dosiahnuť na základe skúmania vzorky, čiže je to požadovaná presnosť odhadov. Presnosť pre kľúčové ukazovatele si stanoví vopred výskumník sám alebo pre celoštátne pravidelné prieskumy Eurostat. A takisto je potrebné určiť aj prípadné subdomény (podmnožiny populácie, napr. kraje), na ktorých chceme dodržať túto stanovenú presnosť.

3. *Spoľahlivosť odhadu (confidence level, označenie (1 – α))* - Do akej miery chceme byť presvedčení, resp. istí, že skutočná stredná hodnota v populácii patrí do intervalu spoľahlivosti? Najčastejšie sa v praxi používajú intervaly spoľahlivosti pre odhady parametrov populácie 90 %, 95 % a 99 %. Zaužívaná je spoľahlivosť 95 %, ktorá sa používa ja v softvéroch.
4. *Štandardná odchýlka (standard of deviation)* – akú variabilitu skúmaného javu očakávame v prieskume? Vysoká variabilita skúmaného javu implikuje potrebu väčšej výberovej vzorky. Naopak, pri nízkej variabilite stačí aj menšia vzorka. Pre zistenie variability skúmaného javu sa v praxi používajú tri scenáre:
    a) Ak poznáme hodnoty skúmaného javu z predchádzajúcich zisťovaní alebo z iného zdroja údajov, tak vypočítame variabilitu na ich základe.
    b) Ak nepoznáme hodnoty skúmaného javu a tento má kvantitatívny charakter, tak uskutočníme pilotnú štúdiu, kde na malej vzorke odhadneme hodnotu rozptylu skúmaného javu.
    c) Ak nepoznáme hodnoty skúmaného znaku a tento má kvalitatívny charakter, tak buď uskutočníme pilotnú štúdiu, ale použijeme hodnotu rozptylu rovnú 0,25.

Keď máme vyššie uvedené kritériá stanovené, tak si treba ujasniť stratégiu pre výpočet veľkosti vzorky. Stratégií je niekoľko a závisia od veľkosti populácie. Ak je populácia malá, tak nie je potrebné robiť výberové zisťovanie, ale treba uskutočniť cenzus. Cenzus znamená uskutočnenie zberu údajov na všetkých prvkoch populácie (tzv. vyčerpávajúce zisťovanie). Cenzus eliminuje výberové chyby pre odhady. Pre veľké populácie je však cenzus neefektívny, lebo znamená vysoké náklady na zber. V dnešnej praxi sa preto zvyčajne robia výberové zisťovania.

Rozsah vzorky n pri výberových zisťovaniach je možné stanoviť na základe veľkosti vzorky, ktorá bola použitá v minulosti pre podobné štúdie. Je však potrebná opatrnosť, lebo táto stratégia môže viesť k opakovaniu chýb z minulosti.

Môžeme tiež použiť publikované tabuľky, ktoré uvádzajú veľkosť vzorky pre zvolenú kombináciu vstupných kritérií (Bartlett et al., 2001, Singh a Masuku, 2014). Treba si však uvedomiť, že v tomto prípade *n* znamená počet získaných odpovedí a nie počet oslovených respondentov, resp. zaslaných dotazníkov. Ďalej aj to, že veľkosť vzorky v týchto tabuľkách je stanovená za predpokladu normálneho rozdelenia skúmaného javu a jednoduchého náhodného výberu.

Inou stratégiou na výpočet rozsahu vzorky je použiť vzorce na to odvodené v odbornej literatúre (napr. Cochran, 1963 a 1977, Kish, 1965). Výber vhodného vzorca závisí od druhu výberu a od parametra populácie, ktorý chceme odhadnúť na základe vzorky.

Spôsob výberu vzorky bude do značnej miery závisieť od množstva informácií, ktoré sú dostupné pre populáciu. Ak sú známe charakteristiky populácie, na získanie presnejších údajov sa môže použiť stratifikovaná vzorka. Ak je o populácii málo informácií, môže sa použiť menej zložitá konštrukcia, ako napríklad jednoduché náhodné alebo systematické vzorky. Ak zoznam prvkov populácie nie je k dispozícii alebo je neúplný, najlepšou voľbou môže byť oblastný výber. V prípade veľkých národných alebo celoštátnych prieskumov sa tieto metódy môžu kombinovať, napríklad stratifikovaný viacstupňový oblastný výber, kedy získame veľmi užitočné a nákladovo efektívne vzorky.

3.1 Výpočet veľkosti vzorky pri jednoduchom náhodnom výbere

Ak máme veľmi veľkú populáciu a skúmame výskyt kvalitatívneho, resp. kategoriálneho javu, čiže podiel, tak Cochran (1977, s. 75) odvodil vzorec pre výpočet rozsahu vzorky vyberanej jednoduchým náhodným výberom. Vychádzal z intervalu spoľahlivosti pre odhad podielu (proporcie) javu v populácii, ktorý má tvar:

$$P\left(p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} < \pi < p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha\ . \qquad (1)$$

Časť vzorca (1), ktorý obsahuje rozsah vzorky (n) nazývame prípustná chyba odhadu a označujeme ako:

$$\Delta = z_{1-\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} \quad \text{resp. jednoduchšie} \quad \Delta = z\sqrt{\frac{pq}{n}}\ , \qquad (2)$$

kde p označuje odhadovaný podiel javu vo vzorke, $q = 1 - p$ je jeho doplnok. Symbolom $z_{1-\alpha/2}$ označujeme príslušný kvantil normovaného normálneho rozdelenia. Pretože podiel p spravidla nepoznáme, tak sa dosadzuje $p = 0{,}5$ a potom aj $q = 0{,}5$. Súčin $pq = 0{,}25$ znamená disperziu javu a takto dosahuje maximálne možnú hodnotu.

Po úpravách získame vzorec pre výpočet minimálnej veľkosti reprezentatívnej vzorky pre podiel:

$$n_0 = \frac{z^2 pq}{\Delta^2} \ . \tag{3}$$

Ak je populácia menšieho konečného rozsahu N, tak je možné veľkosť vzorky n_0 zredukovať podľa nasledovného vzťahu (Cochran, 1977, str. 76):

$$n = \frac{n_0}{1+\frac{(n_0-1)}{N}} \ . \tag{4}$$

Ak skúmame kvantitatívny spojitý jav X, tak chceme zvyčajne odhadnúť strednú hodnotu μ tohto javu. V tomto prípade vychádzame z intervalu spoľahlivosti pre strednú hodnotu, ktorý má tvar:

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \ , \tag{5}$$

kde $\bar{x}$ označuje bodový odhad strednej hodnoty μ, čiže $est\ \mu = \bar{x}$. Časť vzorca (5), ktorý obsahuje rozsah vzorky *n* nazývame prípustná chyba odhadu a označujeme:

$$\Delta = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{resp. zjednodušene.} \quad \Delta = z\, SE \ , \tag{6}$$

kde *σ* označuje štandardnú odchýlku premennej X v populácii, SE je štandardná chyba a $z_{1-\alpha/2}$ označuje príslušný kvantil normovaného normálneho rozdelenia.

Zvyčajne disperziu premennej v populácii (σ^2) nepoznáme, chceme ho práve výberovým zisťovaním zistiť. Vo vzorcoch (5) a (6) preto musíme použiť odhad pomocou výberového rozptylu s^2:

$$est\ \sigma^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ . \tag{7}$$

Vo vzorci (6) musíme potom použiť namiesto $z_{1-\alpha/2}$ príslušný kvantil Studentovho t-rozdelenia, ktorého hodnota závisí aj od veľkosti vzorky, čiže od počtu stupňov voľnosti $k = n - 1$. Ak spoľahlivosť odhadu nastavíme na zvyčajnú hodnotu 95 %, tak kvantil t-rozdelenia $t_{1-\alpha/2;n-1}$ môžeme aproximovať kvantilom normovaného normálneho rozdelenia, ktorý nie je závislý od rozsahu vzorky a platí $t_{1-\alpha/2;n-1} \approx z_{1-\alpha/2}$. Ak použijeme spoľahlivosť odhadu 95 %, tak $z_{1-0,025} = z_{0,975} = 1{,}965$ čo je približne rovné číslu 2.

Po úpravách získame vzorec pre výpočet minimálnej veľkosti reprezentatívnej vzorky z veľmi veľkej populácie:

$$n_0 = \frac{t^2 s^2}{\Delta^2} = \frac{s^2}{\left(\frac{\Delta}{t}\right)^2} \approx \frac{z^2 s^2}{\Delta^2} \approx \frac{2^2 s^2}{\Delta^2} \approx \frac{s^2}{\left(\frac{\Delta}{2}\right)^2} \ . \tag{8}$$

Zo vzorca (8) je zrejmé, že veľkosť vzorky závisí priamo úmerne od veľkosti rozptylu skúmaného javu a nepriamo do zvolenej prípustnej chyby odhadu. Ak je populácia menšieho konečného rozsahu N, tak vzorec (8) je potrebné rozšíriť nasledovným spôsobom:

$$n = \frac{s^2}{\left(\frac{\Delta}{t}\right)^2 + \frac{1}{N}s^2} \quad . \tag{9}$$

Vo vzorcoch (8) a (9) je viacero veličín, ktorých hodnoty treba zistiť. Informácie o variabilite (s^2) skúmaného kvantitatívneho javu, resp. javov sa často používajú z minulých zisťovaní. Ak nie sú k dispozícii, tak sa pri výpočtoch pre veľkosť vzorky často preferuje vzorec pre podiel javu v populácii, kde sa fixne použije pre variabilitu najhorší možný scenár, čiže p = 0,5.

3.2 Výpočet veľkosti vzorky pri stratifikovanom náhodnom výbere

Ak sú známe charakteristiky populácie, tak na získanie presnejších odhadov parametrov populácie sa môže použiť stratifikovaný náhodný výber. V prieskumoch sa bežne používajú tri typy stratifikovaných výberov vzoriek:

- vzorky proporcionálne,
- vzorky neproporcionálne alebo optimálne alokované a
- vzorky o rovnakej veľkosti.

V proporcionálnych stratifikovaných vzorkách je ich veľkosť v každej vrstve (strate) úmerná veľkosti populácie vrstvy (Kish, 1965). Napríklad, ak 16% populácie je staršia ako 65 rokov, potom 16% vzorky by malo obsahovať respondentov v tejto vekovej skupine.

Optimálna alokácia sa používa vtedy, keď variabilita znaku, resp. znakov je v každej vrstve iná. Základnou myšlienkou optimálneho rozdelenia prvkov je to, že pre vrstvy s vysokým stupňom variability sú potrebné väčšie vzorky, ako pre vrstvy s menšou variabilitou, tak aby sa všade dosiahla rovnaká úroveň presnosti pre odhady parametrov. Táto metóda výberu môže minimalizovať náklady na zber údajov vtedy, keď sa náklady v jednotlivých vrstvách líšia. Pre daný rozpočet a úroveň presnosti sa veľkosť vzorky v každej vrstve určí na základe týchto dvoch kritérií.

Vzorky s rovnakou veľkosťou sú také, v ktorých je veľkosť vzorky v každej vrstve rovnaká. Napríklad, ak vyberáme z populácie 300 prvkov (mužov a žien), tak vyberieme rovnako 150 mužov a 150 žien. Tento typ vzorky uprednostňujeme vtedy, ak sa majú pri hodnotení porovnávať dve alebo viac skupín.

Pri stratifikovanom výbere sa populácia N rozdelí do subpopulácií (označovaných ako straty) na základe stratifikačných premenných, čiže platí

$N_1 + N_2 + \cdots + N_H = N$. Veľkosť populácie v každej strate (N_h) musí byť známa. V každej strate sa vykoná náhodný výber o rozsahu n_h.

V prípade odhadu strednej hodnoty μ je potrebné upraviť vzorec (6) pre výpočet prípustnej chyby nasledovne:

$$\Delta = z\,SE = z\sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}} \quad . \tag{10}$$

Ak použijeme označenie $W_h = \frac{N_h}{N}$, tak dostaneme:

$$\Delta = z\,SE = z\sqrt{\sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}} \quad . \tag{11}$$

Pretože vo vzorci (11) vystupujú veľkosti populácie (N) v každej strate (N_h), tak potrebujeme zaviesť nejaký mechanizmus na rozdelenie celej vzorky na časti v stratách. Spôsob, ktorý zohľadňuje rozptyl skúmaného javu v jednotlivých stratách, sa volá Neymanova alokácia a rozsah vzorky v strate sa vypočíta takto:

$$n_h = n\frac{N_h s_h}{\sum N_h s_h} = n\frac{W_h s_h}{\sum W_h s_h} \quad . \tag{12}.$$

Neymanova alokácia je optimálna metóda na výpočet veľkosti vzorky v stratách, ale je veľmi závislá na kvalite toho istého zisťovania z minulého obdobia, na základe ktorého sa stanovuje veľkosť rozptylu javu v príslušnej strate s_h^2, a tak môže dochádzať k skresleniu výsledkov. Preto je vhodnejšia proporcionálna alokácia vzorky, kedy sa rozsah v strate vypočíta nasledovne:

$$n_h = n\frac{N_h}{N} \quad . \tag{13}$$

Pre výpočet konečného rozsahu výskumnej vzorky (n) dostávame na základe proporcionálnej alokácie nasledovný vzorec:

$$n = \frac{\sum W_h s_h^2}{\left(\frac{\Delta}{t}\right)^2 + \frac{1}{N}\sum W_h s_h^2} \quad , \tag{14}$$

kde podiel $W_h = \frac{N_h}{N}$ je váha danej subpopulácie (strata), s_h^2 označuje rozptyl skúmaného ukazovateľa v strate, Δ prípustná chyba odhadu a t je hodnota kvantilu t-rozdelenia pre zvolenú spoľahlivosť $1 - \alpha$. Hodnota kvantilu t pre zvyčajne používané riziko odhadu α = 0.05 je približne číslo 2, ako sme uviedli už vyššie.

Symbol N nemusí označovať len rozsah celej populácie, ale môže označovať aj veľkosť subdomény. Pomocou vzorca (14) potom môžeme vypočítať aj veľkosť vzorky v subdoménach.

V praxi je potrebné vypočítaný rozsah vzorky navýšiť aj o predpokladanú mieru neodpovede. Ak r bude predpokladaná miera odpovede, tak ak oslovíme n respondentov, získame len $n^* = n/r$ odpovedí.

4 Empirický výskum v oblasti podnikateľskej etiky na Slovensku

Výskumný tím na FM UK v Bratislave pod vedením profesorky Anny Remišovej uskutočnil výskum kvality podnikateľského prostredia na Slovensku z pohľadu etiky[2]. Existujúca literatúra naznačuje, že neetické praktiky sú bežným javom v postsocialistických krajinách, neexistujú však komplexné objektívne poznatky o ich povahe, miere výskytu a príčinách. Tento výskum má vyplniť túto medzeru a poskytnúť obraz o neetických praktikách v slovenskom podnikateľskom prostredí. Pre účely tohto výskumného zámeru bolo potrebné vytvoriť dotazník a získať odpovede od reprezentatívnej vzorky respondentov.

Pre tvorbu reprezentatívnej vzorky bola zvolená metóda stratifikovaného náhodného výberu z populácie podnikov (firiem) SR pomocou proporcionálnej alokácie, pretože podobné zisťovanie ešte nebolo na Slovensku robené. Ako stratifikačné premenné boli použité dve premenné: 1. veľkosť podniku [4 kategórie veľkosti podnikov: 1 – mikro (1 – 9 zamestnancov), 2 – malý (10 – 49 zamestnancov), 3 – stredný (50 – 249 zamestnancov) a 4 – veľký (250 a viac zamestnancov)] a 2. región SR podľa klasifikácie NUTS II [4 regióny SR: Bratislavský kraj (BAK), západné Slovensko (ZSK), stredné Slovensko (SSK), východné Slovensko (VSK)].

Informácie o rozsahu populácie pre výskum boli získané zo Štatistického úradu SR (Infostat), ktorý spravuje Register organizácií SR (RO – Databáza ekonomických subjektov SR). Na základe vyššie uvedených vzorcov (pozri aj napr. Cochran, 1977, kapitola 5) sme vypočítali minimálny rozsah výskumnej vzorky. Rozsah populácie bol 313 231 podnikov (firiem) SR. Konečný rozsah pre minimálnu efektívnu vzorku sme stanovili na 1 022 podnikov.

Získavanie údajov od respondentov trvalo šesť mesiacov (január až jún 2019). Údaje sa zbierali hlavne prostredníctvom elektronického dotazníka. Členovia výskumného tímu oslovili potenciálnych respondentov prostredníctvom náhodne vybraných e-mailových adries zahrnutých do databázy slovenských spoločností poskytovanej spoločnosťami FinStat a Cribis. Dotazník bol súčasne distribuovaný aj v tlačenej podobe na ôsmich konferenciách pre profesionálov z oblasti podnikania a obchodu. Celkove sa nakoniec podarilo získať 1 295 odpovedí.

[2] Informácie o výskumnom projekte zameraného na oblasť kvality podnikateľského prostredia na Slovensku z pohľadu etiky sú na stránke projektu APVV-16-0091 (https://www.fm.uniba.sk/veda-a-vyskum/vedecke-projekty-granty-a-spolupraca/narodne-projekty/apvv-16-0091).

Štruktúra populácie podnikov SR podľa dvoch zvolených stratifikačných premenných je uvedená v nasledujúcich tabuľkách (tabuľky 1 až 3). Veľkosť a štruktúru vzorky na základe výpočtov uvádzame tiež v ďalších tabuľkách (pozri tabuľky 4 a 5).

Tab. 1 Veľkosť a štruktúra populácie podnikov podľa dvoch stratifikačných premenných *(Zdroj: vlastné spracovanie na základe dát zo Štatistického úradu Slovenskej republiky)*

Veľkosť	Počet	v %
1 - mikro	289 189	92,32%
2 - malý	18 749	5,99%
3 - stredný	4 442	1,42%
4 - veľký	851	0,27%
Spolu	**313 231**	**100,00%**

Kraj4	Počet	v %
1-BAK	58 566	18,70%
2-ZSK	102 111	32,60%
3-SSK	77 261	24,67%
4-VSK	75 293	24,04%
Spolu	**313 231**	**100,00%**

Tab. 2 Veľkosť populácie podnikov *(Zdroj: vlastné spracovanie na základe dát zo Štatistického úradu Slovenskej republiky)*

Kraj4 [NUTS2]	Veľkosť podniku				Populácia Spolu
	1 - mikro	2 - malý	3 - stredný	4 - veľký	
1-BAK	53 255	3 873	1 149	289	**58 566**
2-ZSK	94 365	6 148	1 348	250	**102 111**
3-SSK	71 905	4 241	949	166	**77 261**
4-VSK	69 664	4 487	996	146	**75 293**
Spolu	**289 189**	**18 749**	**4 442**	**851**	**313 231**

Tab. 3 Štruktúra populácie podnikov *(Zdroj: vlastné spracovanie na základe dát zo Štatistického úradu Slovenskej republiky)*

Kraj4 [NUTS2]	Veľkosť podniku				Populácia Spolu
	1 - mikro	2 - malý	3 - stredný	4 - veľký	
1-BAK	17,00%	1,24%	0,37%	0,09%	**18,70%**
2-ZSK	30,13%	1,96%	0,43%	0,08%	**32,60%**
3-SSK	22,96%	1,35%	0,30%	0,05%	**24,67%**
4-VSK	22,24%	1,43%	0,32%	0,05%	**24,04%**
Spolu	**92,32%**	**5,99%**	**1,42%**	**0,27%**	**100,00%**

Pre výpočty rozsahu výberovej vzorky pre výskum v oblasti podnikateľskej etiky na Slovensku boli použité nasledovné vstupné kritériá:

1. Veľkosť populácie: N = 313 231 podnikov (firiem) SR rozdelených podľa dvoch stratifikačných premenných do 16-tich strát (4 x 4 = 16).
2. Prípustná chyba (presnosť) bola zvolená vo výške 0,06 (presn2).
3. Spoľahlivosť odhadu - použili sme bežne používanú hodnotu 0,05.
4. Štandardná odchýlka pre skúmané javy bola stanovená ako najhorší možný scenár, t.j. 0,5 (sh = 0,5 a potom sh2 = $0,5^2$ = 0,25).

Výsledky výpočtov sú v nasledujúcej tabuľke 6. Pre ilustráciu uvádzame aj podrobné výpočty pre riadok č. 1 na základe vzorcov (13) a (14):

- váha subdomény: Wh = 53255/289189 = 0,184;
- vstupy: sh = 0,5 a potom sh2 = 0,25 a uvažujeme Δ = 0,06 (presn2), t = 2;
- rozsah subdomény: velkost2 = $0,25/[(0,06/2)^2 + 1/289189*0,25] = 277,51$;
- veľkosť vzorky v strate: nh2 = Wh * velkost2 = 0,184 * 277,51 = 51,1.

Tab. 4 Veľkosť vzorky podnikov pre výskum v oblasti podnikateľskej etiky (počet) *(Zdroj: vlastné spracovanie na základe dát zo Štatistického úradu Slovenskej republiky)*

Kraj4 [NUTS2]	Veľkosť podniku				Vzorka (n) Spolu
	1 - mikro	2 - malý	3 - stredný	4 - veľký	
1-BAK	51	56	68	71	**246**
2-ZSK	91	90	79	61	**321**
3-SSK	69	62	56	41	**228**
4-VSK	67	65	59	36	**227**
Spolu	**278**	**273**	**262**	**209**	**1022**

Tab. 5 Veľkosť vzorky podnikov pre výskum v oblasti podnikateľskej etiky (v %) *(Zdroj: vlastné spracovanie na základe dát zo Štatistického úradu Slovenskej republiky)*

Kraj4 [NUTS2]	Veľkosť podniku				Vzorka (n) Spolu
	1 - mikro	2 - malý	3 - stredný	4 - veľký	
1-BAK	4,99%	5,48%	6,65%	6,95%	**24,07%**
2-ZSK	8,90%	8,81%	7,73%	5,97%	**31,41%**
3-SSK	6,75%	6,07%	5,48%	4,01%	**22,31%**
4-VSK	6,56%	6,36%	5,77%	3,52%	**22,21%**
Spolu	**27,20%**	**26,71%**	**25,64%**	**20,45%**	**100,00%**

Tab. 6 Vstupy a výpočty na určenie veľkosti vzorky pre výskum podnikateľskej etiky *(Zdroj: vlastné spracovanie na základe dát zo Štatistického úradu Slovenskej republiky)*

Veľkosť	Kraj4	Nh	N	Wh	sh2	t	presn1	presn2	velkost1	velkost2	nh1	nh2
1	1	53255	289189	0,184	0,25	2	0,05	**0,06**	399,45	**277,51**	73,56	**51,10**
1	2	94365	289189	0,326	0,25	2	0,05	**0,06**	399,45	**277,51**	130,34	**90,55**
1	3	71905	289189	0,249	0,25	2	0,05	**0,06**	399,45	**277,51**	99,32	**69,00**
1	4	69664	289189	0,241	0,25	2	0,05	**0,06**	399,45	**277,51**	96,22	**66,85**
2	1	3873	18749	0,207	0,25	2	0,05	**0,06**	391,64	**273,72**	80,90	**56,54**
2	2	6148	18749	0,328	0,25	2	0,05	**0,06**	391,64	**273,72**	128,42	**89,76**
2	3	4241	18749	0,226	0,25	2	0,05	**0,06**	391,64	**273,72**	88,59	**61,92**
2	4	4487	18749	0,239	0,25	2	0,05	**0,06**	391,64	**273,72**	93,73	**65,51**
3	1	1149	4442	0,259	0,25	2	0,05	**0,06**	366,96	**261,43**	94,92	**67,62**
3	2	1348	4442	0,303	0,25	2	0,05	**0,06**	366,96	**261,43**	111,36	**79,34**
3	3	949	4442	0,214	0,25	2	0,05	**0,06**	366,96	**261,43**	78,40	**55,85**
3	4	996	4442	0,224	0,25	2	0,05	**0,06**	366,96	**261,43**	82,28	**58,62**
4	1	289	851	0,340	0,25	2	0,05	**0,06**	272,10	**209,42**	92,41	**71,12**
4	2	250	851	0,294	0,25	2	0,05	**0,06**	272,10	**209,42**	79,94	**61,52**
4	3	166	851	0,195	0,25	2	0,05	**0,06**	272,10	**209,42**	53,08	**40,85**
4	4	146	851	0,172	0,25	2	0,05	**0,06**	272,10	**209,42**	46,68	**35,93**
Spolu		**313231**		**4,000**					5720,60	**4088,33**	**1430,15**	**1022,08**

5 Záver

Dobre navrhnutý postup pre tvorbu vzorky z populácie poskytuje reprezentatívne údaje, ktoré sú užitočné na vyhodnotenie výskumných štúdií. Celý proces sa začína úvahami o účele výskumu, charakteristikách a veľkosti populácie, dostupnosti informácií o prvkoch populácie a postupoch tvorby vzorky. Správne riešenie týchto problémov spolu s určením veľkosti vzorky prispeje k dôveryhodnému a dôslednému hodnoteniu skúmaných javov.

Použitie výberového skúmania s primeranou veľkosťou vzorky spolu s kvalitným zberom údajov vedie k spoľahlivým a všeobecné platným výsledkom. Vedie to tiež aj k úspore času a financií pri výskumoch.

6 Literatúra

Bartlett, J. E., Kotrlik, J. W., Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. Information Technology, Learning, and Performance Journal, 19(1): 43-55.

Cochran, W. G. (1977). Sampling techniques. 3. vyd. New York : Wiley.

Israel, G. D. (1992). Determining sample size. [online]. Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD6. November 1992. [cit. 2020-05-06]. Dostupné na internete: <https://www.tarleton.edu/academicassessment/documents/Samplesize.pdf>.

Israel, G. D. (1992). Sampling the evidence of extension program impact. [online]. Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD5. October 1992. [cit. 2020-05-06]. Dostupné na internete: <https://edis.ifas.ufl.edu/pdffiles/PD/PD00500.pdf>.

Kish, L. (1965). Survey sampling. New York : Wiley.

Singh, A. S., Masuku, M. B. (2014). Sampling techniques & determination of sample size in applied statistics research: An overview. International Journal of Economics, Commerce and Management 2(11): 1-22.

Sudman, S. (1976). Applied sampling. New York : Academic Press.

Taherdoost, H. (2017). Determining sample size; how to calculate survey sample size. International Journal of Economics and Management Systems 2: 237-239.

7 Poďakovanie

Tento príspevok bol spracovaný s podporou Agentúry na podporu výskumu a vývoja na základe zmluvy č. APVV-16-0091.