SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

1/2023
ročník/volume 33

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

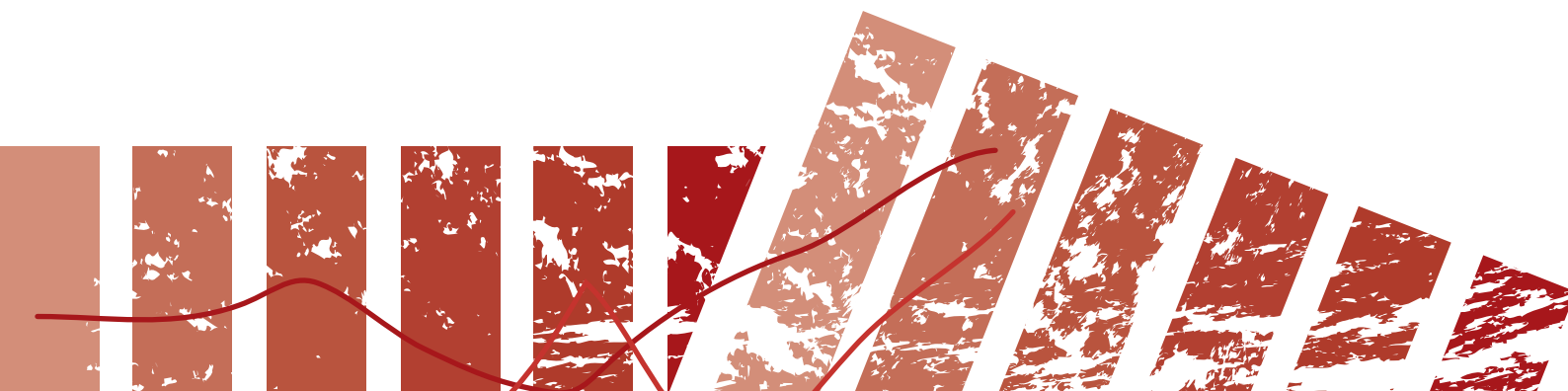
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 1

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

Strany/Pages: 3 – 21

Dátum vydania/Publication date: 15. január 2023/January 15, 2023



Milan TEREK
Vysoká škola manažmentu

METODOLÓGIA URČOVANIA ROZSAHU VÝBEROVÉHO SÚBORU

SAMPLE SIZE DETERMINATION METHODOLOGY

ABSTRAKT
V štatistickom prieskume treba určiť taký rozsah výberového súboru, ktorý s vysokou spoľahlivosťou zabezpečí dostatočnú významnosť získaných výsledkov v praxi. Tiež je dôležité, aby výberový súbor nemal väčší rozsah ako je nevyhnutné. Cieľom príspevku je prezentovať možnosti určovania potrebného rozsahu výberu pri odhadovaní podielu, strednej hodnoty a úhrnu pomocou intervalu spoľahlivosti v štatistických prieskumoch na základe náhodného výberu z nekonečne veľkého aj z konečného základného súboru a poukázať na časté chyby, ktoré sa v tejto súvislosti vyskytujú. To môže uľahčiť výber vhodného postupu pri určovaní rozsahu výberu a správnu interpretáciu výsledkov odhadovania.

ABSTRACT
In a statistical survey, it is necessary to determine such sample size, which ensures high significance of the obtained results in practice with sufficient reliability. It is also important that the sample size is not larger than necessary. The aim of the paper is to present the possibilities of determining the necessary sample size when estimating the proportion, mean and total using a confidence interval in statistical surveys, based on a random sample from both an infinite and a finite population, and to point out frequent errors occurring in this context. This can facilitate the selection of an appropriate sample size procedure and the correct interpretation of the estimation results.

KĽÚČOVÉ SLOVÁ
rozsah výberu, najväčšia chyba odhadu, relatívna presnosť, chyby pri využívaní vzťahov na výpočet potrebného rozsahu výberu

KEY WORDS
sample size, margin of error, relative precision, errors in using the relations to calculate the needed sample size

1. ÚVOD
Ak chceme v rámci kvantitatívneho výskumu v nejakej vedeckej štúdii používať induktívne štatistické metódy, pracujeme s náhodným (pravdepodobnostným) výberom zo základného súboru. Výhodou náhodného vyberania je to, že všeobecne možno identifikovať výberové rozdelenie príslušnej výberovej charakteristiky. Výberové rozdelenie možno potom použiť na tvorbu pravdepodobnostných záverov o chybe spojenej s použitím výsledkov analýzy výberu na tvorbu úsudkov o základnom súbore [2, s. 353]. Treba určiť taký rozsah výberového súboru, ktorý s vysokou spoľahlivosťou zabezpečí dostatočnú praktickú významnosť získaných výsledkov. Tiež je dôležité, aby výberový súbor nemal väčší rozsah, ako je nevyhnutné, teda aby náklady na štúdiu neboli väčšie ako je nevyhnutné [9, s. XV].

Náhodný výber je výber *n* jednotiek vybraných zo základného súboru tak, že každá z možných kombinácií *n* jednotiek má konkrétnu (*particular*) pravdepodobnosť, že

bude vybraná. Keď má každá z možných kombinácií *n* jednotiek rovnakú pravdepodobnosť, že bude vybraná, ide o jednoduchý náhodný výber [5][1]. V tomto príspevku budeme termín jednoduchý náhodný výber chápať tak, ako je definovaný v [5].

Pri náhodnom vyberaní z konečného základného súboru treba vykonať postupnosť štyroch krokov: vytvoriť výberovú bázu (základ výberu, oporu výberu), ktorá obsahuje úplný zoznam *N* jednotiek základného súboru z ktorého sa vyberá, priradiť jednotkám výberovej bázy čísla od 1 po *N*, určiť rozsah *n* náhodného výberu, vybrať náhodným vyberaním s opakovaním alebo bez opakovania *n* čísel z množiny čísel 1 až *N*, pričom vybratie každého čísla má rovnakú pravdepodobnosť. Výsledkom náhodného vyberania s opakovaním alebo bez opakovania z konečného základného súboru je jednoduchý náhodný výber. Na náhodné vyberanie jednotiek, ktoré budú v jednoduchom náhodnom výbere sa použije napríklad tabuľka náhodných čísel, generátor pseudonáhodných čísel, prípadne nejaký iný randomizačný nástroj.

Keď je základný súbor z ktorého sa vyberá nekonečne veľký[2], nemožno pri náhodnom vyberaní vytvoriť výberovú bázu a postupovať pri vyberaní ako v prípade konečného základného súboru. Za náhodný výber z nekonečne veľkého základného súboru sa považuje výber *n* jednotiek zo základného súboru, ktorý sa získa tak, že sa rešpektujú dve podmienky: každá vybraná jednotka je z toho istého základného súboru a každá jednotka je vybraná nezávisle [2, s. 324]. Potom sú pozorovania štatisticky nezávislé a rovnako rozdelené náhodné premenné. Pri implementácii náhodného vyberania z nekonečne veľkého základného súboru treba postupovať veľmi uvážlivo. Každý konkrétny prípad môže vyžadovať rozličnú procedúru vyberania [2, s. 324].

Sú známe aj iné výberové schémy s prvkami randomizácie, napríklad stratifikované náhodné vyberanie, skupinové vyberanie, viacstupňové vyberanie, vyberanie s nerovnakými pravdepodobnosťami a rozličné ich kombinácie.

Predpokladajme, že máme náhodný výber získaný náhodným vyberaním s opakovaním z konečného alebo nekonečne veľkého základného súboru alebo náhodným vyberaním bez opakovania z nekonečne veľkého základného súboru. Vtedy sú pozorovania štatisticky nezávislé a rovnako rozdelené náhodné premenné. Poznamenajme, že pri náhodnom vyberaní s opakovaním z konečného základného súboru môžeme nakoniec realizovať nekonečne veľký výber z toho istého základného súboru, teda aj v tomto prípade môžeme uvažovať o výbere z nekonečne veľkého základného súboru. Preto sa niekedy pri náhodnom vyberaní s opakovaním z konečného alebo nekonečne veľkého základného súboru a náhodnom vyberaní bez opakovania z nekonečne veľkého základného súboru hovorí súhrnne o náhodnom vyberaní z nekonečne veľkého základného súboru a pri náhodnom vyberaní bez opakovania z konečného základného súboru sa hovorí len o náhodnom vyberaní z konečného základného súboru. Takto to budeme používať aj v tomto článku.

[1] Jednoduchý náhodný výber sa často nazýva len náhodný výber. Prívlastok „jednoduchý“ sa potom používa na odlíšenie tohto typu vyberania od zložitejších výberových schém, ktoré majú tiež prvky randomizácie [1, s. 15].

[2] Za nekonečne veľké sa považujú aj konečné základné súbory, v ktorých je zaznamenanie každej jednotky nemožné alebo nerealizovateľné v reálnom čase [15, s.109].

Uvažovali sme o reálnych základných súboroch, z ktorých sa náhodne vyberá. V tradičnom prístupe sa pri tvorbe induktívnych úsudkov o základnom súbore pracuje s pravdepodobnostným modelom konečného alebo nekonečne veľkého reálneho základného súboru. Prijíma sa predpoklad, že hodnoty v základnom súbore tvoria realizácie náhodnej premennej s nejakým rozdelením pravdepodobnosti. Potom možno hovoriť, že máme náhodný výber z nejakého, napríklad z normálneho rozdelenia. V prístupe známom ako výberové skúmanie (*sample survey, survey sampling*) sa pracuje priamo len s konečnými základnými súbormi a pri tvorbe induktívnych úsudkov o nich sa neuvažuje o ich pravdepodobnostných modeloch.

Často sa v ekonomických a sociálnych štúdiách odhaduje podiel, stredná hodnota a úhrn. Pri odhadovaní týchto parametrov sa rozsah výberu určuje na základe voľby želanej spoľahlivosti pri intervalovom odhadovaní a najväčšej chyby (*margin of error*) v intervale spoľahlivosti alebo relatívnej presnosti (*relative precision*) bodového odhadu odhadovaného parametra [4, s. XII].

2. POTREBNÝ ROZSAH VÝBERU PRI INTERVALOVOM ODHADOVANÍ PODIELU V ZÁKLADNOM SÚBORE

Budeme rozlišovať dva prípady – náhodný výber z nekonečne veľkého základného súboru a náhodný výber z konečného základného súboru.

2.1 NÁHODNÝ VÝBER Z NEKONEČNE VEĽKÉHO ZÁKLADNÉHO SÚBORU

Ide o náhodný výber s opakovaním z konečného základného súboru alebo o náhodný výber s opakovaním alebo bez opakovania z nekonečne veľkého základného súboru. Najprv definujeme $(1-\alpha)\cdot 100\ \%$ interval spoľahlivosti pre podiel π v základnom súbore.

Keď *K* má binomické rozdelenie pravdepodobnosti s parametrami n a π, n je veľké číslo a $p=\frac{k}{n}$ je hodnota výberového podielu, potom interval:

$$p - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \qquad (1)$$

kde $z_{1-\frac{\alpha}{2}}$ je $\left(1-\frac{\alpha}{2}\right)\cdot 100\ \%$ kvantil normovaného normálneho rozdelenia, je približný $(1-\alpha)\cdot 100\ \%$ interval spoľahlivosti pre podiel π v základnom súbore. Vzťah (1) sa odporúča používať vtedy, keď: $np>5$ a súčasne $n(1-p)>5$. Vtedy možno aproximáciu binomického rozdelenia normálnym považovať za dobrú.

Pri určovaní rozsahu výberu pri odhadovaní podielu π v základnom súbore podľa vzťahu (1) postupujeme tak, že pre danú spoľahlivosť $(1-\alpha)$ stanovíme najväčšiu chybu d, teda takú hodnotu, pre ktorú:

$$P(|P-\pi|<d)=1-\alpha$$

kde P je výberový podiel.

Najväčšia chyba je:

$$d = z_{1-\frac{\alpha}{2}} \cdot \sigma_P = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi(1-\pi)}{n}} \quad (2)$$

kde $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$ je smerodajná odchýlka výberového podielu P.

Po jednoduchých úpravách vzťahu (2) dostaneme vzťah na výpočet potrebného rozsahu výberu n:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \pi(1-\pi)}{d^2} \quad (3)$$

Hodnotu π nepoznáme. Pri určovaní potrebného rozsahu výberu podľa uvedeného vzťahu možno namiesto neznámej hodnoty π, použiť jej plánovaciu (*planning*) hodnotu, ktorú možno získať realizáciou niektorej z týchto procedúr [2, s. 395]:

1. použiť ako plánovaciu hodnotu π, hodnotu výberového podielu p z predchádzajúceho výberu rovnakých alebo podobných jednotiek,
2. použiť pilotnú štúdiu na realizáciu predbežného výberu. Hodnotu výberového podielu tohto výberu možno použiť ako plánovaciu hodnotu π,
3. použiť subjektívne ohodnotenie π,
4. keď nie je vhodná ani jedna z predošlých procedúr, možno použiť ako plánovaciu hodnotu $\pi = 0{,}5$. Rozptyl Bernoulliho (alternatívneho) rozdelenia $\pi(1-\pi)$, ktorý vystupuje v čitateli vzťahu (2) je totiž najväčší pre $\pi = 0{,}5$. Rozsah výberu n je úmerný tomuto rozptylu a teda voľba $\pi = 0{,}5$ garantuje, že rozsah výberu bude dostatočný na splnenie požiadavky na najväčšiu chybu[3].

Niektorí autori vychádzajú pri odvodení vzťahu na výpočet potrebného rozsahu výberu n zo vzťahu na výpočet najväčšej chyby, v ktorom je π odhadnuté hodnotou výberového podielu p:

$$d = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$$

Rovnakým postupom dospejeme ku vzťahu na výpočet *n*:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot p(1-p)}{d^2} \quad (4)$$

kde $p = \frac{k}{n}$ je hodnota výberového podielu P a k je počet úspechov vo výbere rozsahu n.

Hodnotu výberového podielu p budeme poznať až po získaní náhodného výberu. Namiesto neznámej hodnoty p možno vo vzťahu (4) použiť jej plánovaciu hodnotu,

[3] *Zdôvodnenie tohto postupu pozri aj v [15, s. 143 – 144].*

ktorú možno získať pomocou jednej zo štyroch predtým uvedených procedúr, v ktorých nahradíme π, symbolom p [2, s. 395].

Výpočet rozsahu vzorky podľa (3) alebo (4) má zmysel len vtedy, keď chceme získať náhodný výber z nekonečne veľkého základného súboru a potom počítať a interpretovať intervaly spoľahlivosti pre podiel π. Rozsah výberu n vypočítaný podľa (3) alebo (4) zabezpečí, že so spoľahlivosťou $(1-\alpha)\cdot 100\,\%$ sa bude p vypočítané z výberu v absolútnej hodnote líšiť od neznámej skutočnej hodnoty π v základnom súbore, o menej ako o d.

Pri výpočte n sme vychádzali z určenia najväčšej chyby. Alternatívne možno vyjsť z určenia relatívnej presnosti ε. Relatívna presnosť ε je podiel z neznámej hodnoty odhadovaného parametra, v tomto prípade z hodnoty podielu π. Keď vo vzťahu (3) dosadíme za $d = \varepsilon \cdot \pi$, dostaneme:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \pi(1-\pi)}{\varepsilon^2 \cdot \pi^2} = \frac{z^2_{1-\frac{\alpha}{2}} \cdot (1-\pi)}{\varepsilon^2 \cdot \pi} \qquad (5)$$

Pri odhadovaní π možno opäť použiť niektorú zo štyroch procedúr, ktoré sme uviedli. Výsledok možno interpretovať takto. Rozsah výberu n vypočítaný podľa vzťahu (5) zabezpečí, že so spoľahlivosťou $(1-\alpha)\cdot 100\,\%$ sa hodnota p vypočítaná z výberu bude v absolútnej hodnote líšiť od neznámej hodnoty π v základnom súbore o menej ako o $\varepsilon \cdot 100\,\%$ hodnoty π.

V prípade odhadovania podielu π sa bežne preferuje prvý spôsob výpočtu, v ktorom sa zadá priamo d. Interpretácia výsledku odhadovania je totiž jasnejšia. Ak napríklad vyjde $p = 0{,}34$ a stanovili sme spoľahlivosť $(1-\alpha) = 0{,}95$ a najväčšiu chybu $d = 0{,}03$, potom výpočet n podľa vzťahu (3) alebo (4) zabezpečí, že so spoľahlivosťou 0,95 sa skutočná hodnota π v základnom súbore bude líšiť od 0,34 o menej ako o 0,03. Ešte prístupnejšia je interpretácia v percentách. So spoľahlivosťou 0,95 sa skutočná hodnota π v základnom súbore bude líšiť od 34 % o menej ako o 3 percentuálne body.

2.2 NÁHODNÝ VÝBER Z KONEČNÉHO ZÁKLADNÉHO SÚBORU

Ide o náhodný výber bez opakovania z konečného základného súboru. V tomto prípade nie sú pozorovania štatisticky nezávislé, ani rovnako rozdelené náhodné premenné. K príslušnému symbolu, ktorý označuje parameter základného súboru, budeme kvôli odlíšeniu od predošlého prípadu, pridávať vpravo dolu index K.

Uvedieme definíciu $(1-\alpha)\cdot 100\,\%$ intervalu spoľahlivosti pre podiel π_K v konečnom základnom súbore rozsahu N.

Keď p je hodnota výberového podielu náhodného výberu bez opakovania z konečného základného súboru a n, N a $(N-n)$ sú všetky „dostatočne veľké“ [7, s. 42], potom interval:

$$p - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N} \cdot \frac{p(1-p)}{n-1}} < \pi_K < p + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N} \cdot \frac{p(1-p)}{n-1}} \qquad (6)$$

je približný $(1-\alpha)\cdot 100\ \%$ interval spoľahlivosti pre podiel π_K [6, s. 63], [14, s. 46]. Za „dostatočne veľkú“ hodnotu sa väčšinou považuje hodnota väčšia alebo rovná[4] 50.

Najväčšia chyba je [6, s. 59], [14, s. 40]:

$$d = z_{1-\frac{\alpha}{2}} \cdot \sigma_P = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi_K(1-\pi_K)}{n} \cdot \frac{N-n}{N-1}} \tag{7}$$

kde N je rozsah konečného základného súboru. Po jednoduchých úpravách vzťahu (7) dostaneme vzťah na výpočet potrebného rozsahu výberu *n*:

$$n = \frac{N \cdot z^2_{1-\frac{\alpha}{2}} \cdot \pi_K(1-\pi_K)}{d^2 \cdot (N-1) + z^2_{1-\frac{\alpha}{2}} \cdot \pi_K(1-\pi_K)} \tag{8}$$

Na výpočet n možno použiť aj jednoduchší, približný vzťah (získa sa zo vzťahu (8) vydelením čitateľa aj menovateľa číslom N):

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \pi_K(1-\pi_K)}{d^2} \tag{9}$$

Vzťah (9) sa líši od vzťahu (3) len iným použitým symbolom pre podiel v základnom súbore. Namiesto neznámej hodnoty π_K možno vo vzťahoch (8) a (9) použiť jej plánovaciu hodnotu pomocou jednej zo štyroch predtým uvedených procedúr. Veľmi často sa používa posledná z uvedených procedúr, v ktorej sa položí $\pi_K = 0{,}5$.

Rozsah výberu n vypočítaný podľa vzťahu (8) alebo (9) zabezpečí, že so spoľahlivosťou $(1-\alpha)\cdot 100\ \%$ sa hodnota p vypočítaná z výberu bude v absolútnej hodnote líšiť od neznámej skutočnej hodnoty π_K v základnom súbore, o menej ako o d.

Alternatívne možno vyjsť z určenia relatívnej presnosti ε. Keď vo vzťahu (8) dosadíme za $d = \varepsilon \cdot \pi_K$, dostaneme [6, s. 74], [17, s. 137]:

$$n = \frac{N \cdot z^2_{1-\frac{\alpha}{2}} \cdot \pi_K(1-\pi_K)}{\varepsilon^2 \cdot \pi_K^2 \cdot (N-1) + z^2_{1-\frac{\alpha}{2}} \cdot \pi_K(1-\pi_K)} = \frac{N \cdot z^2_{1-\frac{\alpha}{2}} \cdot (1-\pi_K)}{\varepsilon^2 \cdot \pi_K \cdot (N-1) + z^2_{1-\frac{\alpha}{2}} \cdot (1-\pi_K)} \tag{10}$$

Keď v zjednodušenom vzťahu (8) dosadíme za $d = \varepsilon \cdot \pi_K$, dostaneme:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \pi_K(1-\pi_K)}{\varepsilon^2 \cdot \pi_K^2} = \frac{z^2_{1-\frac{\alpha}{2}} \cdot (1-\pi_K)}{\varepsilon^2 \cdot \pi_K} \tag{11}$$

Rozsah výberu n vypočítaný podľa vzťahu (10) alebo (11) zabezpečí, že so spoľahlivosťou $(1-\alpha)\cdot 100\ \%$ sa hodnota p vypočítaná z výberu bude v absolútnej hodnote líšiť od neznámej skutočnej hodnoty π_K v základnom súbore o menej ako o $\varepsilon \cdot 100$ % hodnoty π_K.

[4] Ak si myslíte, že generujúce rozdelenie sa veľmi nelíši od normálneho, je pravdepodobne bezpečné použiť centrálnu limitnú teorému (Hájek, 1960), keď má výber rozsah aspoň 50 [7, s. 43].

Aj v tomto prípade odhadovania podielu sa bežne preferuje prvý spôsob výpočtu, v ktorom sa priamo určí veľkosť *d*, kvôli jasnejšej interpretácii výsledku odhadovania. Pripomeňme, že výpočet rozsahu vzorky podľa (8), (9), (10) alebo (11) má zmysel len vtedy, keď chceme získať náhodný výber bez opakovania z konečného základného súboru a potom počítať a interpretovať intervaly spoľahlivosti pre podiel π_K.

3. POTREBNÝ ROZSAH VÝBERU PRI INTERVALOVOM ODHADOVANÍ STREDNEJ HODNOTY V ZÁKLADNOM SÚBORE

3.1 NÁHODNÝ VÝBER Z NEKONEČNE VEĽKÉHO ZÁKLADNÉHO SÚBORU

Uvedieme definíciu $(1-\alpha)\cdot 100$ % intervalu spoľahlivosti pre strednú hodnotu μ. Keď $\bar{x}$ je hodnota výberového priemeru náhodného výberu rozsahu *n* zo základného súboru s normálnym rozdelením, so známym rozptylom σ^2, potom interval:

$$\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{12}$$

je $(1 - \alpha)\cdot 100$ % interval spoľahlivosti pre strednú hodnotu μ. Keď je rozsah náhodného výberu dostatočne veľký ($n \geq 30$), možno vzťah (12) použiť na výpočet približného $(1 - \alpha)\cdot 100$ % intervalu spoľahlivosti pre strednú hodnotu μ, aj keď základný súbor, z ktorého je výber, nemá normálne rozdelenie[5]. V tomto prípade môžeme aj σ vo vzťahu (12), nahradiť hodnotou *s* výberovej smerodajnej odchýlky[6].

Pri určovaní rozsahu výberu pri odhadovaní strednej hodnoty μ v základnom súbore postupujeme rovnako ako v prípade odhadovania podielu tak, že pre danú spoľahlivosť $(1-\alpha)$ stanovíme najväčšiu chybu *d*, teda takú hodnotu, pre ktorú:

$$P(|\bar{X} - \mu| < d) = 1 - \alpha$$

kde $\bar{X}$ je výberový priemer.

Najväčšia chyba je:

$$d = z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{X}} = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{13}$$

kde $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ je smerodajná odchýlka výberového priemeru $\bar{X}$.

Po jednoduchých úpravách vzťahu (13) dostaneme vzťah na výpočet potrebného rozsahu výberu[7] *n*:

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{d}\right)^2 \tag{14}$$

[5] *V dôsledku platnosti centrálnej limitnej teorémy.*
[6] *Postup formulácie intervalu (12) pozri napr. v [15, s. 132 – 137].*
[7] *Podrobnejšie pozri v [15, s. 137 – 138], [16, s. 60 – 61].*

Použitie vzťahu (14) predpokladá znalosť smerodajnej odchýlky σ. Keď ju nepoznáme, možno do vzťahu dosadiť za σ, jej plánovaciu hodnotu. Pri jej určovaní možno v praxi použiť niektorú z týchto procedúr [2, s. 391].

1. hodnotu odhadu smerodajnej odchýlky, vypočítanú z dát predošlých štúdií ako plánovaciu hodnotu σ.
2. pilotnú štúdiu na realizáciu predbežného výberu. Hodnotu výberovej smerodajnej odchýlky predbežného výberu možno použiť ako plánovaciu hodnotu σ.
3. možno odhadnúť najväčšiu a najmenšiu hodnotu v základnom súbore. Ich rozdiel je hodnota odhadu rozpätia v základnom súbore. Štvrtina rozpätia sa často považuje za dobrú aproximáciu hodnoty smerodajnej odchýlky a teda prijateľnú plánovaciu hodnotu σ.

Aj v tomto prípade má výpočet rozsahu vzorky podľa (14) zmysel len vtedy, keď chceme získať náhodný výber z nekonečne veľkého základného súboru a chceme počítať a interpretovať intervaly spoľahlivosti pre strednú hodnotu μ.

Rozsah výberu *n* vypočítaný podľa vzťahu (14) zabezpečí, že so spoľahlivosťou $(1 - \alpha) \cdot 100$ % sa bude $\bar{x}$ vypočítané z výberu v absolútnej hodnote líšiť od neznámej strednej hodnoty μ o menej ako o *d*.

Alternatívne možno vyjsť z určenia relatívnej presnosti *ε*. Najväčšiu chybu vyjadríme ako podiel *ε* zo strednej hodnoty: $d = \varepsilon \cdot \mu$. Po dosadení do vzťahu (14) dostaneme:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \sigma^2}{\varepsilon^2 \cdot \mu^2} = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \mathrm{VK}^2 \cdot \mu^2}{\varepsilon^2 \cdot \mu^2} = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \mathrm{VK}^2}{\varepsilon^2} \tag{15}$$

kde $\mathrm{VK}^2 = \frac{\sigma^2}{\mu^2}$ je štvorec variačného koeficienta.

Keď nemáme žiadnu apriórnu informáciu o VK, možno postupovať tak, že na obmedzenom výbere sa odhadne VK (rozptyl σ^2 sa odhadne pomocou výberového rozptylu $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ a stredná hodnota *μ* sa odhadne pomocou výberového priemeru $\bar{X} = \sum_{j=1}^{n} X_j$). Pre odhadnutý variačný koeficient sa vypočíta *n*. Potom sa vykoná dodatočný výber tak, aby rozsah predbežného a dodatočného výberu spolu bol *n*.

Rozsah výberu *n* vypočítaný podľa vzťahu (15) zabezpečí, že so spoľahlivosťou $(1 - \alpha) \cdot 100$ % sa hodnota $\bar{x}$ vypočítaná z výberu bude v absolútnej hodnote líšiť od neznámej strednej hodnoty *μ* o menej ako o $\varepsilon \cdot 100$ % hodnoty *μ*.

3.2 NÁHODNÝ VÝBER Z KONEČNÉHO ZÁKLADNÉHO SÚBORU

Všimneme si náhodný výber bez opakovania z konečného základného súboru rozsahu *N*. Pozorovania nie sú štatisticky nezávislé, ani rovnako rozdelené náhodné premenné.

Uvedieme definíciu $(1 - \alpha) \cdot 100$ % intervalu spoľahlivosti pre strednú hodnotu μ_K.

Keď $\bar{x}$ je hodnota výberového priemeru náhodného výberu bez opakovania z konečného základného súboru a n, N a $(N - n)$ sú všetky „dostatočne veľké“ [7, s. 43], potom interval:

$$\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma_K}{\sqrt{n}} < \mu_K < \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma_K}{\sqrt{n}} \tag{16}$$

je približný $(1 - \alpha) \cdot 100$ % interval spoľahlivosti pre strednú hodnotu μ_K [6, s. 61 – 62], [14, s. 42].

Možno dokázať[8], že nevychýleným bodovým odhadom rozptylu výberového priemeru $\sigma_{\bar{X}}^2$ je:

$$\hat{\sigma}_{\bar{X}}^2 = \frac{N-n}{N} \cdot \frac{S^2}{n}$$

Keď nepoznáme rozptyl $\sigma_{\bar{X}}^2$ možno ho odhadnúť pomocou $\hat{\sigma}_{\bar{X}}^2$ a približný $(1 - \alpha) \cdot 100$ % interval spoľahlivosti pre strednú hodnotu μ_K vypočítať podľa vzťahu:

$$\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}} < \mu_K < \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}} \tag{17}$$

Určíme rozsah náhodného výberu bez opakovania potrebný na odhadnutie strednej hodnoty μ_K so zvolenou spoľahlivosťou $(1 - \alpha)$ a požadovanou najväčšou chybou d. Je známe, že:

$$d = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma_K}{\sqrt{n}}$$

Z toho po jednoduchých úpravách dostaneme:

$$n = \frac{N \cdot z_{1-\frac{\alpha}{2}}^2 \cdot \sigma_K^2}{d^2(N-1) + z_{1-\frac{\alpha}{2}}^2 \cdot \sigma_K^2} \tag{18}$$

Možno použiť aj jednoduchší približný vzťah:

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \sigma_K^2}{d^2} \tag{19}$$

Rozsah výberu n vypočítaný podľa vzťahu (18) alebo (19) zabezpečí, že so spoľahlivosťou $(1 - \alpha) \cdot 100$ % sa bude hodnota výberového priemeru $\bar{x}$ vypočítaná z výberu v absolútnej hodnote líšiť od neznámej strednej hodnoty μ_K o menej ako o d.

[8] *Pozri napr. v [14, s. 38].*

Alternatívne možno postupovať tak, že vo vzťahu na výpočet potrebného rozsahu výberu použijeme relatívnu presnosť ε a variačný koeficient VK v konečnom základnom súbore:

$$\mathrm{VK} = \frac{\sigma_K}{\mu_K}$$

Keď vo vzťahu (18) vydelíme čitateľa aj menovateľa μ_K^2 a za *d* dosadíme $\varepsilon \cdot \mu_K$, po jednoduchých úpravách dostaneme:

$$n = \frac{N \cdot z^2_{1-\frac{\alpha}{2}} \cdot \mathrm{VK}^2}{\varepsilon^2 (N-1) + z^2_{1-\frac{\alpha}{2}} \cdot \mathrm{VK}^2} \tag{20}$$

Na výpočet *n* možno použiť aj jednoduchší, približný vzťah:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \mathrm{VK}^2}{\varepsilon^2} \tag{21}$$

Problematické môže byť zadanie VK v uvedených vzťahoch na výpočet *n*. Keď nemáme žiadnu apriórnu informáciu o VK, odporúča sa postupovať pri jeho odhadovaní a pri výpočte *n* v dvoch krokoch takto:

1. Na obmedzenom predbežnom výbere sa odhadne VK^2 takto:

$$\widehat{\mathrm{VK}}^2 = \frac{\hat{\sigma}_K^2}{\hat{\mu}_K^2}$$

kde

$$\hat{\sigma}_K^2 = \frac{N-1}{N} S^2$$

$\hat{\sigma}_K^2$ je nevychýleným bodovým odhadom[9] rozptylu σ_K^2 a $\hat{\mu}_K^2 = \bar{X}^2$.

2. Pre odhadnuté $\widehat{\mathrm{VK}}^2$ sa podľa vzťahu (20) alebo (21) vypočíta *n*. Realizuje sa doplnkový výber tak, aby rozsah predbežného a doplnkového výberu bol spolu aspoň *n*.

3. Rozsah výberu *n* vypočítaný podľa vzťahu (20) alebo (21) zabezpečí, že so spoľahlivosťou (1 – *α*) ·100 % sa hodnota $\bar{x}$ vypočítaná z výberu bude v absolútnej hodnote líšiť od neznámej strednej hodnoty μ_K o menej ako o $\varepsilon \cdot 100$ % hodnoty μ_K.

[9] *Pozri v [14, s. 38].*

4. POTREBNÝ ROZSAH VÝBERU PRI INTERVALOVOM ODHADOVANÍ ÚHRNU V KONEČNOM ZÁKLADNOM SÚBORE

Keď vzťahy (16) a (17) vynásobíme rozsahom konečného základného súboru *N*, dostaneme vzťahy na výpočet intervalov spoľahlivosti pre úhrn τ.

Keď $\bar{x}$ je hodnota výberového priemeru náhodného výberu bez opakovania z konečného základného súboru a *n*, *N* a (*N* – *n*) sú všetky „dostatočne veľké“, potom interval:

$$N \cdot \bar{x} - N \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma_K}{\sqrt{n}} < \tau < N \cdot \bar{x} + N \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma_K}{\sqrt{n}} \quad (22)$$

je približný (1 – *α*) ·100 % interval spoľahlivosti pre úhrn τ.

Keď nepoznáme smerodajnú odchýlku σ_K, približný (1 – *α*) ·100 % interval spoľahlivosti pre úhrn τ možno vypočítať podľa vzťahu:

$$N \cdot \bar{x} - N \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}} < \tau < N \cdot \bar{x} + N \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}} \quad (23)$$

Výpočet rozsahu výberu je rovnaký ako pri intervalovom odhadovaní strednej hodnoty konečného základného súboru, teda podľa vzťahov (18), (19) alebo (20), (21).

Aj v tomto prípade má výpočet rozsahu vzorky podľa uvedených vzťahov zmysel len vtedy, keď chceme získať náhodný výber bez opakovania z konečného základného súboru a chceme počítať a interpretovať intervaly spoľahlivosti pre strednú hodnotu μ_K alebo intervaly spoľahlivosti pre úhrn τ.

Keď je rozsah výberu oveľa menší ako rozsah konečného základného súboru ($\frac{n}{N} \leq 0{,}05$), možno prijať predpoklad, že základný súbor je nekonečne veľký, aj keď je v skutočnosti konečný a uplatniť postupy ako pri náhodnom vyberaní z nekonečne veľkého základného súboru.

Všetky intervaly spoľahlivosti ktoré sme uviedli, sú formulované ako otvorené sprava aj zľava. Alternatívne môžu byť formulované ako sprava aj zľava uzavreté. Potom sa v ich interpretácii zmení výraz „bude sa v absolútnej hodnote líšiť o menej ako o“ na výraz „nebude sa v absolútnej hodnote líšiť o viac ako o“.

Príklad: Miestne zastupiteľstvo jednej mestskej časti zisťuje pomocou dotazníkového prieskumu niektoré skutočnosti o životných podmienkach domácností, ktoré žijú v bytových domoch. V tejto mestskej časti žije v bytových domoch *N* = 2 000 domácností. Jedna z otázok v dotazníku sa týka nákladov spojených s užívaním bytu v roku 2021. Cieľom je odhadnúť ich strednú hodnotu μ_K a úhrn τ. Chceme, aby sa so spoľahlivosťou 0,95 hodnota odhadu odchyľovala od neznámej skutočnej strednej hodnoty resp. úhrnu nákladov o menej ako o 5 % týchto parametrov.

Jedna z otázok v prieskume je „Máte doma internet?“ Chceme, aby sa so spoľahlivosťou 0,95 hodnota odhadu podielu domácností, ktoré majú internet *p* odchyľovala od neznámeho skutočného podielu π_K o menej ako o 5 percentuálnych bodov.

Realizoval sa obmedzený predbežný náhodný výber bez opakovania rozsahu *n* = 20. Na báze dát z tohto obmedzeného výberu vyšli priemerné náklady jednej domácnosti $\bar{x}$ = 2 500, – eur, smerodajná odchýlka nákladov *s* = 700, – eur a podiel domácností s internetom *p* = 0,75. Vypočítame najmenší potrebný rozsah náhodného výberu bez opakovania.

Najmenší potrebný rozsah výberu treba vypočítať pre:

1. Intervalové odhadovanie strednej hodnoty a úhrnu nákladov spojených s užívaním bytu podľa vzťahu (20), kde VK^2 odhadneme hodnotou:

$$\widehat{VK}^2 = \frac{\frac{N-1}{N} \cdot s^2}{\bar{x}^2} = \frac{\frac{2000-1}{2000} \cdot 700^2}{2500^2} \approx 0{,}078$$

Po dosadení do (20) dostaneme:

$$n = \frac{2000 \cdot 1{,}96^2 \cdot 0{,}078}{0{,}05^2(2000-1) + 1{,}96^2 \cdot 0{,}078} \approx 114$$

Rozsah výberu *n* = 114 zabezpečí, že so spoľahlivosťou 0,95 sa priemerné náklady spojené s užívaním bytu vypočítané z výberu $\bar{x}$ budú odchyľovať od skutočnej neznámej strednej hodnoty nákladov v základnom súbore 2000 domácností μ_K, o menej ako o 5 % strednej hodnoty nákladov μ_K. Podobne pri úhrne nákladov.

2. Pre intervalové odhadovanie podielu bytov, ktoré majú internet, vypočítame potrebný rozsah výberu *n*, podľa vzťahu (8). Po dosadení do (8) dostaneme:

$$n = \frac{2000 \cdot 1{,}96^2 \cdot 0{,}75 \cdot 0{,}25}{0{,}05^2(2000-1) + 1{,}96^2 \cdot 0{,}75 \cdot 0{,}25} \approx 252$$

Rozsah výberu *n* = 252 zabezpečí, že so spoľahlivosťou 0,95 sa hodnota výberového podielu domácností ktoré majú internet *p* bude odchyľovať od neznámeho skutočného percentuálneho podielu π_K v základnom súbore 2000 domácností, o menej ako o 5 percentuálnych bodov.

Rozsah náhodného výberu bez opakovania musí byť väčšie z čísiel, teda 252. V predbežnom výbere sme už realizovali 20 pozorovaní, bude treba ešte získať ďalších 232, aby celkový počet pozorovaní bol 252.

Pri odhadovaní strednej hodnoty a úhrnu sme požadovali spoľahlivosť 0,95 a relatívnu presnosť 0,05. Pre rozsah výberu *n* = 252 môžeme napríklad fixovať spoľahlivosť a vypočítať novú relatívnu presnosť podľa vzťahu (20):

$$\frac{2000 \cdot 1{,}96^2 \cdot 0{,}078}{\varepsilon^2(2000-1) + 1{,}96^2 \cdot 0{,}078} = 252$$

z toho

$$\varepsilon \approx 0{,}032$$

So spoľahlivosťou 0,95 sa hodnota odhadu bude odchyľovať od neznámej skutočnej strednej hodnoty resp. úhrnu nákladov o menej ako o 3,2 % týchto parametrov. Alternatívne možno fixovať relatívnu presnosť a vypočítať novú hodnotu spoľahlivosti.

Všeobecne sa v štatistickom prieskume môžu odhadovať rozličné veličiny, rozličných premenných, s rozličnou požadovanou spoľahlivosťou a najväčšou chybou alebo relatívnou presnosťou, pre ktoré môže vyjsť rozličný požadovaný rozsah výberu. Bude sa realizovať najväčší z nich. V ostatných sa podobne, ako sme to ukázali na príklade, vypočíta nová hodnota relatívnej presnosti, najväčšej chyby alebo spoľahlivosti pre rozsah výberu, ktorý sa bude realizovať.

5. INÉ VÝBEROVÉ SCHÉMY

Vzťahy na výpočet potrebného rozsahu výberu, ktoré sme uviedli, platia len pri realizácii náhodného vyberania z konečného alebo nekonečne veľkého základného súboru, pri ktorom sa jednotky vyberajú priamo z celého základného súboru. V prípade iných výberových schém je všetko inak.

Všimnime si napríklad výberovú schému – stratifikované náhodné vyberanie. Pripomeňme, že stratifikácia je rozdelenie konečného základného súboru na vzájomne sa vylučujúce a základný súbor celkom pokrývajúce podsúbory (stratá alebo vrstvy), ktoré sa vzhľadom na skúmanú premennú považujú za viac homogénne ako celý základný súbor [5]. Pri stratifikovanom náhodnom vyberaní sa často uvažuje o pomernej alokácii pozorovaní medzi stratá (*proportional allocation*), v ktorej sa uplatní konštantný výberový pomer. Z každého strata sa napríklad náhodným vyberaním bez opakovania vyberie 5 % jednotiek. Vybrané jednotky zo všetkých strát spolu tvoria stratifikovaný náhodný výber. Keď je vo všetkých stratách rozptyl približne rovnaký, je pomerná alokácia asi najlepší spôsob zvyšovania presnosti odhadovania. Keď rozptyly v stratách významne kolíšu, môže použitie optimálnej alokácie (*optimal allocation*) viesť k nižším nákladom [7, s. 87][10].

Pri poststratifikácii sa stratá definujú až po vytvorení výberu jednoduchým náhodným vyberaním. Pri odhadovaní strednej hodnoty, úhrnu a podielu sú tiež známe vzťahy na výpočet najmenšieho rozsahu výberu [6, s. 175 – 176]. Skupinové vyberanie (*cluster sampling*) je náhodné vyberanie skupín. Do výberového súboru sú zaradené všetky jednotky z náhodne vybraných skupín. Skupina (*cluster*) je časť základného súboru, ktorý je rozdelený na navzájom sa vylučujúce skupiny jednotiek, ktoré určitým spôsobom súvisia [5]. Pri dvojstupňovom skupinovom vyberaní sa najprv jednoduchým náhodným vyberaním vyberú skupiny – primárne jednotky a z každej vybranej skupiny sa jednoduchým náhodným vyberaním s rovnakým výberovým pomerom vyberú sekundárne jednotky. Pri odhadovaní strednej hodnoty, úhrnu a podielu sú pri skupinovom aj dvojstupňovom skupinovom vyberaní známe vzťahy na výpočet najmenšieho rozsahu výberu. Podobne pri náhodnom vyberaní s nerovnakými pravdepodobnosťami[11].

Vzťahy na výpočet najmenšieho potrebného rozsahu výberu má zmysel počítať vtedy, keď chceme realizovať intervalové odhadovanie a závisia od odhadovaného parametra základného súboru a od použitej výberovej schémy. Podrobnejšie sme opísali len vzťahy použiteľné pri intervalovom odhadovaní podielu, strednej hodnoty

[10] Podrobnejšie o optimálnom rozdelení výberového súboru pozri v [7, s. 87 – 91].
[11] Podrobnejšie pozri v [6] a [7].

a úhrnu na základe náhodného výberu získaného z nekonečne veľkého základného súboru alebo náhodného výberu z konečného základného súboru.

6. DOSTUPNÉ KALKULÁTORY POTREBNÉHO ROZSAHU VÝBERU A ČASTÉ CHYBY PRI JEHO VÝPOČTE A INTERPRETÁCII

Na internete sú bežne dostupné mnohé kalkulátory na výpočet potrebného rozsahu výberu. Väčšina z nich umožňuje len výpočet potrebného rozsahu výberu n pri odhadovaní podielu π na základe náhodného výberu z nekonečne veľkého alebo z konečného základného súboru, napríklad [8], [10], [12], [18]. Niektoré ponúkajú aj viac alebo menej stručný a jasný opis spôsobu výpočtu. Možno odporúčať napríklad Sample Size Calculator [10], ktorý síce neponúka vysvetlenie postupu, ale ponúka stručný a jasný návod na použitie, a čo treba mimoriadne oceniť, hneď v prvej vete upozorňuje, že kalkulátor sa má používať len na náhodné výbery. Niektoré kalkulátory umožňujú aj iné výpočty, napríklad [11].

6.1 ČASTÉ CHYBY PRI VÝPOČTE A INTERPRETÁCII POTREBNÉHO ROZSAHU VÝBERU

Pri výpočte a interpretácii potrebného rozsahu výberu sa často vyskytujú chyby, ktoré možno rozdeliť na chyby, týkajúce sa charakteru výberu a chyby týkajúce sa dosahu výsledkov.

6.1.1 CHYBY, KTORÉ SA TÝKAJÚ CHARAKTERU VÝBERU

Niekedy sa nejaká výberová schéma zámerného (nenáhodného) vyberania považuje za náhodné vyberanie. Pripomeňme, že pri zámernom vyberaní výber jednotiek závisí od znalostí a úsudku osoby, ktorá vyberanie realizuje. Niekedy sa v praxi možno stretnúť s nenáhodným vyberaním založeným na prístupnosti jednotiek (*convenience sampling*). Takéto vyberanie môže byť lacné a jednoducho realizovateľné. Napríklad odberateľ dodávky jabĺk vyberie z niekoľkých debničiek jablko na kontrolu bez toho, že by uplatnil nejaký mechanizmus náhodného vyberania. Veľa spoločností dáva návštevníkom svojich webových stránok možnosť vyplniť dotazník a elektronicky ho poslať. Získaný výber môže byť veľký, ale je založený na samo výbere návštevníkov webovej stránky (*self-selection survey*). Sama jednotka rozhodne, či sa zaradí do výberového súboru. Niekedy analytik zaradí do výberu jednotky o ktorých sa nazdáva, že najlepšie reprezentujú jednotky v základnom súbore (*judgement sampling*). Jednotky vo výbere závisia od subjektívneho úsudku analytika. Niekedy sa štatistické metódy určené na náhodné výbery použijú v praxi na analýzu dát z nenáhodných výberov, pričom sa argumentuje, že nenáhodný výber bol vytvorený tak, že má vlastnosti ako náhodný výber. To nemožno akceptovať.

Väčšinou pri zámene zámerného vyberania za náhodné nejde o zámer. Totiž, v hovorovom jazyku sa napríklad realizácia vyberania založeného na prístupnosti jednotiek bežne označí za náhodné vyberanie. Podobne samo výber sa často chápe ako náhodné vyberanie. Ak si nie sme istí, či zvolená výberová schéma je alebo nie je výberovou schémou jednoduchého náhodného vyberania z konečného základného súboru, možno sa pri rozhodovaní oprieť o jednoduché pravidlo. Každá schéma vyberania z konečného základného súboru s prvkami randomizácie totiž umožňuje určiť pre každú jednotku v základnom súbore pravdepodobnosť, že bude zaradená do výberového súboru (pravdepodobnosť zaradenia – *inclusion probability*). Ak sa výber získa napríklad tak, že na sociálnu sieť sa dá dotazník s prosbou o jeho vyplnenie, ani v prípade, že základný súbor tvoria všetky osoby prihlásené na sociálnej

sieti, nejde o náhodný výber, pretože nemožno pre každú jednotku v základnom súbore určiť pravdepodobnosť že bude zaradená do výberového súboru.

V prípade vyberania z nekonečne veľkého základného súboru treba overiť, či boli pri vyberaní dodržané pravidlá: každá vybraná jednotka je z toho istého základného súboru a každá jednotka je vybraná nezávisle. Ak nejaká výberová schéma, o ktorej uvažujeme nerešpektuje aspoň jednu z týchto dvoch podmienok, nejde o výberovú schému náhodného vyberania z nekonečne veľkého základného súboru.

Pri realizácii štatistického prieskumu treba vždy najprv stanoviť výberovú schému. Keď napríklad zvolíme výberovú schému – náhodné vyberanie bez opakovania n jednotiek z konečného základného súboru rozsahu N, pre každú jednotku v základnom súbore je pravdepodobnosť zaradenia n/N. Potom možno stanoviť potrebný rozsah výberu n. Napokon treba zvoliť vhodný randomizačný nástroj, napríklad generátor pseudonáhodných čísiel.

6.1.2 CHYBY, KTORÉ SA TÝKAJÚ DOSAHU VÝSLEDKOV

Niekedy sa uvádza, že keď sa rozsah výberu vypočíta podľa uvedeného vzorca, získame reprezentatívny výber. V [3, s. 23] sa uvádza 9 rozličných významov výrazu „reprezentatívny výber“. V [3, s. 24] je formulovaný záver: „Vzhľadom na množstvo rôznych významov, ktoré môže mať pojem ‚reprezentatívny výber‘ sa odporúča nepoužívať ho v praxi, pokiaľ nie je jasné, čo sa tým myslí“.

Niekedy sa uvádza, že rozsah výberu vypočítaný podľa uvedeného vzorca je potrebný na dosiahnutie vysokej miery vypovedacej schopnosti výskumu. Niekedy sa uvedie, že vypočítaný rozsah výberu je potrebný na dosiahnutie štatisticky významných výsledkov. Tu treba poznamenať, že výraz „výsledok je štatisticky významný“ sa používa v testovaní štatistických hypotéz, keď p-hodnota je menšia alebo rovná stanovenej hladine významnosti α. Samotný pojem „štatistická významnosť“ je problematický, pretože sa často používa v úplne nepatričných kontextoch. Mnoho štatistikov vyslovilo názor, že tento termín by sa mal prestať používať. Uvedieme citát z Wikipédie: „V roku 2019 viac ako 800 štatistikov a vedcov podpísalo výzvu na upustenie od používania pojmu ‚štatistická významnosť‘, vo vede. Americká štatistická asociácia zverejnila ďalšie oficiálne vyhlásenie, v ktorom sa uvádza (strana 2): „Dospeli sme k záveru, že je čas úplne prestať používať výraz ‚štatisticky významný‘. Nemali by prežiť ani varianty ako ‚výrazne odlišné‘ a ‚nevýznamné‘, či už vyjadrené slovami, hviezdičkami v tabuľke alebo iným spôsobom“ [13]. Tvrdenia, že výpočet potrebného rozsahu výberu podľa jedného z uvedených vzťahov zabezpečí získanie štatisticky významných výsledkov alebo vysokú vypovedaciu schopnosť výsledkov prieskumu, vysoko presahujú význam a dosah uvedených výpočtov rozsahu výberu.

7. ZÁVER

Pri realizácii výberového skúmania treba určiť taký rozsah výberového súboru, ktorý s vysokou spoľahlivosťou zabezpečí dostatočnú praktickú významnosť získaných výsledkov, pričom rozsah výberu by nemal byť väčším ako je nevyhnutné. Keď chceme využívať induktívne štatistické metódy, je nevyhnutné realizovať náhodný výber. Keď uvažujeme o náhodnom vyberaní jednotiek priamo z celého základného súboru, možno rozlíšiť dva rozličné prípady. V prvom prípade ide o náhodné vyberanie z nekonečne veľkého základného súboru, ktoré zahŕňa náhodné vyberanie

s opakovaním z nekonečne veľkého alebo konečného základného súboru a náhodné vyberanie bez opakovania z nekonečne veľkého základného súboru. Tu sú pozorovania štatisticky nezávislé a rovnako rozdelené náhodné premenné. V druhom prípade ide o náhodné vyberanie z konečného základného súboru, čím sa myslí náhodné vyberanie bez opakovania z konečného základného súboru. Tu pozorovania nie sú štatisticky nezávislé ani rovnako rozdelené náhodné premenné.

S uvedenou základnou klasifikáciou súvisia rozličné vzťahy na výpočet najmenšieho potrebného rozsahu výberu pri výpočte intervalov spoľahlivosti. Postup ich odvodenia závisí v oboch prípadoch od toho, či sa okrem želanej spoľahlivosti intervalového odhadu zadá najväčšia chyba odhadu, alebo relatívna presnosť bodového odhadu odhadovaného parametra. V článku sa uvádzajú vzťahy na výpočet najmenšieho potrebného rozsahu výberu pre všetky 4 kombinácie možností (náhodný výber z nekonečne veľkého základného súboru pri zadaní najväčšej chyby odhadu alebo zadaní relatívnej presnosti, náhodný výber z konečného základného súboru pri zadaní najväčšej chyby odhadu alebo relatívnej presnosti) pri výpočte intervalov spoľahlivosti pre podiel, strednú hodnotu a pri náhodnom vyberaní z konečného základného súboru, aj pre úhrn. Všetky uvedené vzťahy sú platné len pre realizáciu najjednoduchšej výberovej schémy, v ktorej sa jednotky vyberajú priamo z celého základného súboru. Keď sa bude realizovať niektorá zo zložitejších schém vyberania s randomizačnými prvkami, z konečného základného súboru, vzťahy a niekedy aj postupy určenia najmenšieho potrebného rozsahu výberu sú iné.

Na internete sú voľne dostupné mnohé kalkulátory najmenšieho potrebného rozsahu náhodného výberu. Väčšinou sú použiteľné len pri odhadovaní podielu pomocou intervalu spoľahlivosti, niektoré aj pri odhadovaní iných parametrov základného súboru. Prístup na niektoré z nich uvádzame v zozname literatúry.

Pri plánovaní štatistického prieskumu s využitím induktívnych štatistických metód sa treba vyvarovať chýb. Niektoré sa vyskytujú pomerne často. V článku sme ich rozdelili na chyby, ktoré sa týkajú charakteru výberu a na chyby, ktoré sa týkajú dosahu výsledkov. Ich znalosť môže byť pri plánovaní štatistického prieskumu určite užitočná.

Táto práca bola podporená vedeckou grantovou agentúrou KEGA v rámci projektu K-20-035-00 „Learn Economics: aplikácia e-vzdelávania ako novej formy výučby ekonómie“.

LITERATÚRA

[1] AGRESTI, A.: Statistical Methods for the Social Sciences. Fifth edition. Boston: Pearson Education, Inc., 2018. 591 s. ISBN 13: 978-0-13-450710-1.
[2] ANDERSON, D. R. – SWEENEY, D. J. – WILLIAMS, T. A. – CAMM, J. D. – COCHRAN, J. J. – FRY, M. J. – OHLMANN, J. W.: Statistics for Business and Economics. 14e Edition. Boston: Cengage Learning, Inc., 2020. 1119 s. ISBN: 978-1-337-90106-2.
[3] BETHLEHEM, J.: Applied Survey Methods. A Statistical Perspective. Hoboken: Wiley and Sons, 2009. 375 s. ISBN 978-0-470-37308-8.
[4] DESU, M. M. – RAGHAVARAO, D.: Sample Size Methodology. San Diego: Academic Press, Inc., 1990. 135 s. ISBN 0-12-212165-1.
[5] ISO 3534–2. Statistics – Vocabulary and symbols – Part 2: Applied Statistics. 2006. 90 s.

[6] LEVY, P. S. – LEMESHOW, S.: Sampling of Populations. Methods and Applications. Fourth Edition. Hoboken: Wiley and Sons, 2008. 576 s. ISBN 978-0-470-04007-2.
[7] LOHR, S. L.: Sampling: Design and Analysis. 2nd edition. New York: CRC Press Taylor & Francis Group, 2019. 596 s. ISBN 978-0-3672-7346-0.
[8] *Raosoft* [cit. 2022-02-12]. Dostupné na: http://www.raosoft.com/samplesize.html
[9] RYAN, T. P.: Sample Size Determination and Power. Hoboken: John Wiley & Sons, Inc., 2013. 374 s. ISBN 978-1-118-43760-5.
[10] Sample Size Calculator (Australian Bureau of Statistics) [cit. 2022-02-12]. Dostupné na: https://www.abs.gov.au/websitedbs/d3310114.nsf/home/sample+size+calculator
[11] Sample Size Calculator - Easy to Use with Description [cit. 2022-02-12]. Dostupné na: https://www.benchmarksixsigma.com/calculators/sample-size-calculator-for-1-sample-t-test-finite-population/?gclid=CjwKCAjwwo-WBhAMEiwAV4dybZBpza_kMZTd9LDFv5aROlcZDSAgYmO-6lectTNV1Rn4moZAKemtGRoC0UMQAvD_BwE
[12] Sample Size Methodology [cit. 2022-02-12]. Dostupné na: https://www.macorr.com/sample-size-calculator.htm
[13] Statistical Significance [cit. 2022-02-12]. Dostupné na: https://en.wikipedia.org/wiki/Statistical_significance
[14] TEREK, M. – HRNČIAROVÁ, Ľ.: Výberové skúmanie. Bratislava: Ekonóm, 2008.118 s. ISBN 978-80-225-2440-7.
[15] TEREK, M.: Interpretácia štatistiky a dát. Piate, doplnené vydanie. Košice: Equilibria, 2017. 460 s. ISBN 978-80-8143-213-2.
[16] TEREK, M.: Interpretácia štatistiky a dát. Podporný učebný materiál. Piate, doplnené vydanie. Košice: Equilibria, 2017. 244 s. ISBN 978-80-8143-212-5.
[17] TEREK, M.: Dotazníkové prieskumy a analýzy získaných dát. Košice: Equilibria, 2019. 202 s. ISBN 978-80-8143-247-7.
[18] The Survey System [cit. 2022-02-12]. Dostupné na: https://www.surveysystem.com/sscalc.htm#two

RESUMÉ

V štatistickom prieskume treba určiť taký rozsah výberového súboru, ktorý s vysokou spoľahlivosťou zabezpečí dostatočnú praktickú významnosť získaných výsledkov. Tiež je dôležité, aby výberový súbor nemal väčší rozsah, ako je nevyhnutné. Ak chceme v analýze používať induktívne štatistické metódy, treba pracovať s náhodným (pravdepodobnostným) výberom zo základného súboru. Cieľom článku je prezentovať možnosti určovania potrebného rozsahu náhodného výberu v prípade realizácie najjednoduchšej výberovej schémy – náhodného vyberania jednotiek priamo z celého základného súboru – pri odhadovaní podielu, strednej hodnoty a v prípade konečného základného súboru aj úhrnu, pomocou intervalu spoľahlivosti, prezentovať niektoré bežne dostupné kalkulátory rozsahu výberu a poukázať na časté chyby, ktoré sa v tejto súvislosti vyskytujú.
Keď uvažujeme o náhodnom vyberaní jednotiek priamo z celého základného súboru, možno rozlíšiť dva rozličné prípady. V prvom prípade ide o náhodné vyberanie z nekonečne veľkého základného súboru, ktoré zahŕňa náhodné vyberanie s opakovaním z nekonečne veľkého alebo konečného základného súboru a náhodné vyberanie bez opakovania z nekonečne veľkého základného súboru. Tu sú pozorovania štatisticky nezávislé a rovnako rozdelené náhodné premenné. V druhom prípade ide o náhodné vyberanie z konečného základného súboru, čím sa myslí

náhodné vyberanie bez opakovania z konečného základného súboru. Tu pozorovania nie sú štatisticky nezávislé ani rovnako rozdelené náhodné premenné. S uvedenou základnou klasifikáciou súvisia rozličné vzťahy na výpočet najmenšieho potrebného rozsahu výberu pri výpočte intervalov spoľahlivosti. Postup ich odvodenia závisí v oboch prípadoch od toho, či sa okrem želanej spoľahlivosti intervalového odhadu zadá najväčšia chyba odhadu, alebo relatívna presnosť bodového odhadu odhadovaného parametra. V článku sa uvádzajú vzťahy na výpočet najmenšieho potrebného rozsahu výberu pre všetky 4 kombinácie možností (náhodný výber z nekonečne veľkého základného súboru pri zadaní najväčšej chyby odhadu alebo pri zadaní relatívnej presnosti, náhodný výber z konečného základného súboru pri zadaní najväčšej chyby odhadu alebo zadaní relatívnej presnosti) pri výpočte intervalov spoľahlivosti pre podiel, strednú hodnotu a pri náhodnom vyberaní z konečného základného súboru, aj pre úhrn.
Keď sa bude realizovať niektorá zo zložitejších schém vyberania s randomizačnými prvkami z konečného základného súboru, vzťahy a niekedy aj postupy určenia najmenšieho potrebného rozsahu výberu sú iné. V súvislosti s určovaním najmenšieho potrebného rozsahu výberu sa často vyskytujú niektoré chyby. Ich znalosť môže byť pri plánovaní štatistického prieskumu užitočná. V článku sú rozdelené na chyby, ktoré sa týkajú charakteru výberu a na chyby, ktoré sa týkajú dosahu výsledkov. Sú opísané v časti 6.1.

RESUME
In a statistical survey, it is necessary to determine the sample size, which ensures sufficient practical significance of the obtained results with high confidence. It is also important that the sample size is not larger than necessary. If we want to use inferential statistical methods in the analysis, we must work with a random (probability) sample from the population. The aim of the paper is to present the possibilities of determining the necessary size of random sample in case of the implementation of the simplest sampling scheme - random sampling of units directly from the entire population, when estimating the proportion, mean and, in case of the finite population, also the total, using a confidence interval, to present some commonly available sample size calculators and, point out frequent errors occurring in this context.
When we consider the random sampling of units directly from the entire population, two different cases can be distinguished. In the first case, it is random sampling from an infinite population, including random sampling with replacement from an infinite or finite population and random sampling without replacement from an infinite population. Here, the observations are statistically independent and equally distributed random variables. In the second case, it is a random sampling from a finite population, referring to a random sampling without replacement from a finite population. Here the observations are not statistically independent or equally distributed random variables. Various relations for the calculation of the smallest necessary sample size in the calculation of confidence intervals are related to the mentioned basic classification. The procedure for their derivation depends in both cases on whether, in addition to the desired confidence of the interval estimate, the margin of error of the estimate is entered, or the relative precision of the point estimator of the estimated parameter. The paper provides relations for calculating the smallest required sample size for all 4 combinations of options (random sample from an infinite population and given the margin of error or given relative precision, random sample from a finite population and given the margin of error or specifies the relative precision) when calculating

confidence intervals for the proportion, the mean, and during random sampling from the finite population, also for the total.
When one of the more complex sampling schemes from a finite population with randomization elements will be implemented, the relationships and sometimes the procedures for determining the smallest sample size required are different. In connection with the determination of the smallest necessary sample size, certain errors occur frequently. Their acknowledgement can be useful when planning a statistical survey. In the article, they are divided into errors related to the nature of the sample and errors related to the impact of the results. They are described in the section 6.1.

PROFESIJNÝ ŽIVOTOPIS

***Prof. Ing. Milan Terek, PhD.**, od roku 2018 pracuje ako profesor na Vysokej škole manažmentu v Bratislave. Vedie kurzy Úvod do štatistiky, Štatistika, Matematika pre manažérov II, Kvantitatívne metódy pre manažérov a Kvantitatívne metódy vo výskume v oblasti podnikového manažmentu. V rokoch 1977 – 2018 pracoval na Ekonomickej univerzite v Bratislave. Viedol kurzy Štatistika, Štatistické riadenie kvality, Analýza rozhodovania, Hĺbková analýza dát, Výberové skúmanie, Lineárne programovanie, Nelineárne programovanie, Operačný výskum a Systémové modelovanie. Vo výskume sa zameriava na aplikácie štatistických metód v ekonómii a manažmente. Je autorom alebo spoluautorom 6 monografií, 10 vysokoškolských učebníc, 17 skrípt, 76 článkov vo vedeckých a odborných časopisoch a 115 príspevkov na vedeckých konferenciách publikovaných v zborníkoch.*

KONTAKT
mterek@vsm.sk