

Štatistický úrad Slovenskej republiky
The Statistical Office of the Slovak Republic

SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS
and DEMOGRAPHY

vedecký časopis/scientific journal

1/2023
ročník 33



ŠTATISTICKÝ
ÚRAD
SLOVENSKEJ
REPUBLIKY

ISSN 1339-6854 (online)
ISSN 1210-1095 (tlačené vydanie)

SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

Recenzovaný vedecký časopis založený v roku 1991. Jednotlivé čísla časopisu zverejňujeme aj v elektronickej podobe na ssad.statistics.sk. Názory autorov článkov sa nemusia zhodovať s názormi vydavateľa.

Zahraniční poradcovia/Foreign Consultants

Gabriela Czanner

University of Liverpool
Veľká Británia/United Kingdom

Jitka Langhamrová

Vysoká škola ekonomická v Praze
University of Economics in Prague
Česká republika/Czech Republic

Estefanía Mourelle Espasandín

Universidade da Coruña
Španielsko/Spain

Michaela Potančoková

Joint Research Centre,
European Commission
Taliansko/Italy

Hana Řezanková

Vysoká škola ekonomická v Praze
University of Economics in Prague
Česká republika/Czech Republic

Milan Stehlík

Institute of Statistics, University of Valparaíso
Čile/Chile
Johannes Kepler University Linz
Rakúsko/Austria

Výkonná redaktorka/Executive Editor

Silvia Hudecová

Jazykové redaktorky/Language Editors

Slovenský jazyk/Slovak Language

Silvia Duchková

Anglický jazyk/English Language

Andrea Okenková

SLOVAK STATISTICS AND DEMOGRAPHY

The scientific peer-reviewed journal founded in 1991. Individual copies of the journal are available to readers in electronic form at the website ssad.statistics.sk. The opinions of the authors do not necessarily correlate with the opinions of the publisher.

Redakčná rada/Editorial Board

Ľudmila Ivančíková

(predsedníčka/chairwoman)
Štatistický úrad SR
Statistical Office of the SR

Mikuláš Cár

Slovenská štatistická a demografická spoločnosť
Slovak Statistical and Demographic Society

Helena Glaser-Opitzová

Štatistický úrad SR
Statistical Office of the SR

Ján Haluška

INFOSTAT Bratislava

Iveta Stankovičová

Univerzita Komenského v Bratislave
Comenius University in Bratislava

Erik Šoltés

Ekonomická univerzita v Bratislave
University of Economics in Bratislava

Pavol Tišliar

Univerzita Cyrila a Metoda v Trnave
University of Ss. Cyril and Methodius in Trnava
Masarykova univerzita
Masaryk University

Boris Vaňo

INFOSTAT - Výskumné demografické centrum
INFOSTAT - Demographic Research Centre

Adresa redakcie/Address of Editorial Office

Slovenská štatistika a demografia
Štatistický úrad SR
Lamačská cesta 3/C, 840 05 Bratislava 45
Slovenská republika

E-mailová adresa/E-mail address

SSaD@statistics.sk

ssad.statistics.sk
www.statistics.sk

OBSAH/CONTENTS

I. VEDECKÉ ČLÁNKY/SCIENTIFIC ARTICLES

Milan TEREK 3
METODOLÓGIA URČOVANIA ROZSAHU VÝBEROVÉHO SÚBORU
SAMPLE SIZE DETERMINATION METHODOLOGY

Branislav ŠPROCHA 22
PLODNOSŤ ŽIEN NA SLOVENSKU VO VÝSLEDKOCHE SČÍTANIA OBYVATEĽOV,
DOMOV A BYTOV 2021
WOMEN'S FERTILITY IN SLOVAKIA IN THE RESULTS OF THE 2021 POPULATION
AND HOUSING CENSUS

Peter KNÍŽAT, Helena GLASER-OPITZOVÁ 37
INDEX SPOTREBITEĽSKÝCH CIEN Z WEBS CRAPOVANÝCH ÚDAJOV: ANALÝZA
VYBRANEJ PRODUKTOVEJ SKUPINY
CONSUMER PRICE INDEX FROM WEB-SCRAPED DATA: ANALYSIS OF
SPECIFIC PRODUCT CATEGORY

II. INFORMATÍVNE ČLÁNKY, NÁZORY, RECENZIE, ROZHOVORY, INFORMÁCIE/ INFORMATIVE ARTICLES, OPINIONS, REVIEWS, INTERVIEWS, INFORMATION

Ľudmila IVANČÍKOVÁ 50
MANAŽMENT INOVÁCIÍ V EURÓPSKOM ŠTATISTICKOM SYSTÉME:
OD EXPERIMENTOVANIA K PRODUKCII (KONFERENCIA DGINS)
MANAGEMENT OF INNOVATION IN THE EUROPEAN STATISTICAL SYSTEM:
FROM EXPERIMENTATION TO PRODUCTION (DGINS CONFERENCE)
Informácia/Information

Mária RYBANSKÁ 52
NITRIANSKE ŠTATISTICKÉ DNI 2022
STATISTICAL DAYS IN NITRA 2022
Informácia/Information

III. PRIPRAVUJEME/COMING SOON 54

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.

KLÚČOVÉ 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 štúdiu používať indukívne 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 veľký², 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 uvažlivo. 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 odlišenie 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 indukčných 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 indukčných 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:

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

kde $z_{1-\frac{\alpha}{2}}$ je $(1 - \frac{\alpha}{2}) \cdot 100\%$ 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 :

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

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 chybu³.

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 :

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

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:

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \pi(1-\pi)}{\varepsilon^2 \cdot \pi^2} = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \frac{1-\pi}{\pi}}{\varepsilon^2 \cdot \pi} \quad (5)$$

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:

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

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 rovná⁴ 50.

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

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

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

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

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

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \pi_K(1-\pi_K)}{d^2} \quad (9)$$

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]:

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

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

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \pi_K(1-\pi_K)}{\varepsilon^2 \cdot \pi_K^2} = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot (1-\pi_K)}{\varepsilon^2 \cdot \pi_K} \quad (11)$$

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:

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

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 rozdelenie⁵. V tomto prípade môžeme aj σ vo vzťahu (12), nahradiť hodnotou s výberovej smerodajnej odchýlky⁶.

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:

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

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 výberu⁷ n :

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

⁵ V dôsledku platnosti centrálnej limitnej teóremy.

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:

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot \sigma^2}{\varepsilon^2 \cdot \mu^2} = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot VK^2 \cdot \mu^2}{\varepsilon^2 \cdot \mu^2} = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot VK^2}{\varepsilon^2} \quad (15)$$

kde $VK^2 = \frac{\sigma^2}{\mu^2}$ 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:

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

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

Možno dokázať⁸, 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:

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

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:

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

Možno použiť aj jednoduchší približný vzťah:

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

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:

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

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

$$n = \frac{N \cdot z_{1-\frac{\alpha}{2}}^2 \cdot VK^2}{\varepsilon^2(N-1) + z_{1-\frac{\alpha}{2}}^2 \cdot VK^2} \quad (20)$$

Na výpočet n možno použiť aj jednoduchší, približný vzťah:

$$n = \frac{z_{1-\frac{\alpha}{2}}^2 \cdot VK^2}{\varepsilon^2} \quad (21)$$

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:

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

kde

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

$\hat{\sigma}_K^2$ je nevychýleným bodovým odhadom⁹ rozptylu σ_K^2 a $\hat{\mu}_K^2 = \bar{X}^2$.

2. Pre odhadnuté \widehat{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 $(1 - \alpha) \cdot 100$ % 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 $(1 - \alpha) \cdot 100$ % interval spoľahlivosti pre úhrn τ .

Keď nepoznáme smerodajnú odchýlku σ_K , približný $(1 - \alpha) \cdot 100$ % 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 pravdepodobnosťami¹¹.

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 jablák 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 nerespektuje 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, hviezdikami 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 indukčné 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 ekonomie“.

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=CjwKCAjww0-WBhAMEiwAV4dybZBpza_kMZTd9LDFv5aROlcZDSAqYmO-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 indukzívne 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 Vyskej škole manažmentu v Bratislave. Vede 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

Branislav ŠPROCHA
INFOSTAT – Výskumné demografické centrum
Centrum spoločenských a psychologických vied SAV

**PLODNOSŤ ŽIEN NA SLOVENSKU VO VÝSLEDKOKH
SČÍTANIA OBYVATEĽOV, DOMOV A BYTOV 2021¹**

**WOMEN'S FERTILITY IN SLOVAKIA IN THE RESULTS OF
THE 2021 POPULATION AND HOUSING CENSUS**

ABSTRAKT

Plodnosť na Slovensku má dlhodobý klesajúci trend. Potvrdzujú to nielen prierezové údaje, ale znižovanie intenzity rodenia detí jasne identifikujeme aj v generačnom pohľade. Nezastupiteľné miesto pre hlbšie pochopenie týchto zmien zohrávajú údaje získavané zo sčítania obyvateľov, domov a bytov. V nich je už od roku 1930 integrálnou súčasťou na Slovensku otázka o počte (živo-)narodených detí. Okrem generačnej plodnosti prezentujúcej priemernú realizovanú plodnosť žien narodených v rovnakom roku umožňujú aj hlbší pohľad na štruktúru žien podľa počtu narodených detí, úroveň a vývoj pravdepodobností narodenia prvého, či ďalšieho dieťaťa, či koncentrácie reprodukcie a tým samotnej variability reprodukcie.

Hlavným cieľom tohto článku bola hlbšia analýza generačnej plodnosti na Slovensku v skupinách žien narodených v rokoch 1940 – 1980. Na tieto účely využívame výsledky najnovšieho sčítania obyvateľov, domov a bytov 2021.

Získané údaje potvrdili pokračujúci pokles plodnosti, keď generácie žien z druhej polovice 60. rokov boli poslednými s priemerne viac ako 2 deťmi. U žien narodených v 70. rokoch generačná plodnosť klesla signifikantne pod hranicu 2 detí. To je podmienené znižovaním pravdepodobností narodenia prvého i druhého dieťaťa, čo sa následne prejavuje aj na náraste bezdetnosti, jednodetnosti pri súčasnom poklese podielu dvojdetného modelu rodiny. V najmladších generáciách tiež dochádza k zvyšovaniu variability reprodukcie a znižovaniu rodín a počtu súrodencov v nich.

ABSTRACT

Fertility in Slovakia has a long-term downward trend. This is confirmed not only by cross-sectional data, but we also clearly identify the fertility reduction of in a cohort view. An irreplaceable place for a deeper understanding of these changes is played by data obtained from Population and Housing Census. The question of the number of (live-) births has been their an integral part of them since 1930. In addition to completed cohort fertility presenting the average realized fertility of women born in the same year, they also allow a deeper insight into the parity structure, the level and development of the parity progression ratios, or the concentration of reproduction and thus the variability of the reproduction.

The main goal of this article was a deeper analysis of cohort fertility in Slovakia in groups of women born in the years 1940 – 1980, using the results of the most recent 2021 Population and Housing Census.

The obtained data confirmed the continued decline in fertility, when cohorts of women from the second half of the 1960s were the last with an average of more than 2 children per woman. For women born in the 1970s, completed cohort fertility dropped significantly below the threshold of 2 children. This is conditioned by the reduction of

¹ *Príspevok je výsledkom projektu VEGA č. 2/0064/23 „Časové a priestorové zmeny rodinných domácností na Slovensku a ich možné faktory.“*

the probabilities of the birth of the first and the second child, which is subsequently reflected in the increase in childlessness and one-child families, with a simultaneous decrease in the share of the two-child family model. In the youngest generations, there is also an increase in the variability of reproduction and the decrease in family size and the number of siblings in them.

KLÚČOVÉ SLOVÁ

generačná plodnosť, SODB 2021, plodnosť, konečná plodnosť, štruktúra podľa parity, Slovensko

KEY WORDS

cohort fertility, the 2021 Population and Housing Census, fertility, completed cohort fertility, parity structure, Slovakia

1. ÚVOD

Sčítanie obyvateľov, domov a bytov predstavuje vzhľadom na svoj obsah, podobu vyčerpávajúceho zisťovania a možnosť viacnásobného kombinačného prepojenia premenných v rôznych administratívnych úrovniach v mnohých smeroch na Slovensku nenahraditeľný a jedinečný zdroj údajov. Najčastejšie sa získané výsledky využívajú na hodnotenie počtu a jednotlivých štruktúrnych charakteristík obyvateľstva v priestorovom a pri komparácii viacerých cenzov aj časovom aspekte. Sčítanie obyvateľov je tak vnímané ako kľúčový vstup hodnotenia populačnej statiky. Okrem toho je však nutné podotknúť, že dlhodobou integrálnou súčasťou jednotlivých sčítaní obyvateľov na Slovensku sú aj otázky umožňujúce analýzu demografickej dynamiky. Výnimkou nebolo ani posledné sčítanie obyvateľov, domov a bytov z roku 2021. Z hľadiska demografickej reprodukcie je nenahraditeľnou predovšetkým otázka o počte živonarodených detí ženám. V kombinácii s ďalšími znakmi, ako je napríklad vek, rok narodenia a prípadne aj ďalšie štruktúrne premenné, umožňuje podrobnú generačnú (kohortovú) analýzu realizovanej plodnosti. Nemenej dôležitou je tiež možnosť tieto údaje konštruovať pre rôzne administratívne úrovne, alebo inak vymedzené populačné skupiny. Práve aplikácia rôznych viacnásobných kombinačných triedení a prepojení výraznou mierou rozširuje analytické prístupy zameriavajúce sa na identifikáciu diferencných rozdielov v realizovanej plodnosti a štruktúre žien podľa počtu živonarodených detí (parity). Tým sa význam a nenahraditeľnosť daných výsledkov zo sčítania v dátových podmienkach Slovenska len zvyšuje. Okrem primárneho využitia na analytické účely sú však výsledky o počte živonarodených detí ženám v kombinácii s vekom resp. rokom narodenia kľúčovým vstupom na konštrukciu niektorých demografických modelov (napr. inkrementno-dekrementných tabuliek plodnosti).

Aj napriek tomu, že význam údajov zo sčítaní o počte živonarodených detí ženám na Slovensku sa zdá byť nesporný, ich analytické využitie je skôr výnimkou. To pravdepodobne súvisí jednak s dostupnosťou týchto údajov, ako aj celkovo nižším využívaním longitudinálneho prístupu pri analýze demografickej reprodukcie na Slovensku.

Cieľom predloženého textu je jednak základná analýza vývoja konečnej plodnosti a štruktúry žien podľa počtu narodených detí ženám na Slovensku prostredníctvom údajov získaných v poslednom sčítaní obyvateľov, domov a bytov 2021. Okrem toho sa zameriavame aj na otázky spojené s úrovňou a vývojom pravdepodobností

zväčšenia rodiny (zvýšenia parity) a koncentrácie reprodukcie. V poslednej časti príspevku sa pokúsime prostredníctvom špeciálnej dekompozičnej techniky identifikovať príčiny znižovania realizovanej plodnosti na Slovensku medzi vybranými generáciami žien.

2. ZDROJE ÚDAJOV A METODIKA PRÁCE

Analýzu procesu plodnosti žien na Slovensku opierame o údaje poskytnuté Štatistickým úradom Slovenskej republiky (ďalej ŠÚ SR) o počte živonarodených detí ženám podľa jednotiek veku z definitívnych výsledkov sčítania obyvateľov, domov a bytov 2021. Keďže rozhodujúci okamih sčítania bol stanovený na polnoc z 31.12. 2020 na 1.1. 2021, je možné priamo odvodiť aj rok narodenia (generáciu) týchto žien. Pri samotnom spracovaní údajov sme z analýzy vylúčili osoby, ktoré v sčítaní na otázku o počte živonarodených detí neodpovedali.² Na účely našej analýzy pritom pracujeme s generáciami z rokov 1940 – 1980, pričom ženy narodené v rokoch 1971 – 1980 vzhľadom na svoj vek v čase sčítania (41 – 49 rokov) ešte nemôžeme považovať za osoby s ukončenou reprodukciou. Na druhej strane však pretrvávajúca veľmi nízka intenzita plodnosti žien vo veku 40 a viac rokov (a najmä po dovŕšení 45. roku života) na Slovensku indikuje, že získané výsledky sa už s najväčšou pravdepodobnosťou nebudú výraznejšie líšiť od reálneho stavu.

Základný metodický nástroj na analýzu intenzity generačnej plodnosti predstavuje indikátor označovaný ako konečná plodnosť. Ten v prípade, že je konštruovaný z údajov zo sčítania obyvateľov, vyjadruje priemerný počet detí, ktoré sa narodili jednej žene určitej generácie (ročníka narodenia) do rozhodujúceho okamihu sčítania. Ako sme uviedli vyššie, na naše účely pracujeme len so ženami s ukončenou reprodukciou, teda s tými, ktoré k rozhodujúcemu okamihu mali 50 a viac rokov, respektíve so ženami, ktoré boli vo veku 40 – 49 rokov.

Základom na výpočet konečnej plodnosti z výsledkov sčítaní obyvateľov je štruktúra žien podľa roku narodenia k rozhodujúcemu okamihu sčítania a počtu živonarodených detí [bližšie 4, 19]. Pre podiely žien podľa počtu živonarodených detí z generácie (g) môžeme odvodiť nasledujúce vzťahy:

podiel žien bez detí (bezdetné ženy):

$$pF_0^g = \frac{F_0^g}{F^g} \quad (1)$$

F_0^g je počet žien z generácie (g), ktoré v sčítaní obyvateľov 2021 uviedli, že sú bezdetné,

F^g je počet všetkých žien generácie (g) sčítaných k rozhodujúcemu okamihu sčítania obyvateľov 2021;

podiel žien s jedným dieťaťom:

$$pF_1^g = \frac{F_1^g}{F^g} \quad (2)$$

² Podiel žien vo veku 40 – 80 rokov, ktoré v sčítaní neodpovedali na otázku o počte živonarodených detí sa pohyboval v rozmedzí 0,4 – 2,0 %.

F_1^g je počet žien z generácie (g), ktoré v sčítaní obyvateľov 2021 uviedli, že sa im narodilo jedno dieťa;

a analogicky pre podiel žien s (n) deťmi:

$$pF_n^g = \frac{F_n^g}{Fg} \quad (3)$$

F_n^g je počet žien z generácie (g), ktoré v sčítaní obyvateľov 2021 uviedli, že sa im narodilo n - detí.

Hodnotu konečnej plodnosti môžeme následne odvodiť z čiastkových konečných plodností jednotlivých poradí, keď platí:

$$KP^g = \sum_{i=0}^n KP_i^g \quad (4)$$

pričom na výpočet jednotlivých konečných plodností podľa poradia narodeného dieťaťa použijeme nasledujúce vzťahy:

pre prvé poradie:

$$KP_1^g = 1 - pF_0^g \quad (5)$$

pre druhé a n -té poradie:

$$KP_2^g = KP_1^g - pF_1^g \quad (6)$$

$$KP_n^g = KP_{n-1}^g - pF_{n-1}^g \quad (7)$$

Samotná konečná plodnosť a štruktúra žien podľa počtu živonarodených detí je podmienená pravdepodobnosťami narodenia dieťaťa danej parity. Tie môžeme odvodiť prostredníctvom ukazovateľa, ktorý je v česko-slovenskom prostredí známy ako pravdepodobnosť zväčšenia rodiny (v anglickej literatúre ide o pravdepodobnosť zvýšenia parity, *parity progression ratio*). Predmetný indikátor empiricky určuje pravdepodobnosť, že sa žene s (i) deťmi narodí ďalšie dieťa ($i + 1$) ($PZR_{i,i+1}$), resp., že sa bezdetnej žene narodí prvé dieťa ($PZR_{0,1}$) [pozri 19].

Pravdepodobnosť narodenia prvého dieťaťa bezdetnej žene ($i = 0$) je daná vzťahom:

$$PZR_{0,1}^g = KP_1^g \quad (8)$$

Pre deti ďalšieho poradia potom vo všeobecnosti platí:

$$PZR_{i,i+1}^g = \frac{KP_{i+1}^g}{KP_i^g} \quad (9)$$

a pre posledný otvorený paritný interval aplikujeme:

$$PZR_{(n-1)+;(n+)}^g = \frac{KP_{n+}^g}{KP_{(n+1)+}^g} \quad (10)$$

Konečná plodnosť ako priemerný počet narodených detí na jednu ženu prináša len jednu časť informácie o charaktere reprodukčného modelu a reprodukčnej histórii. Je výsledkom rôznej štruktúry žien podľa počtu narodených detí, ktorá sa následne odzrkadľuje na samotnej variabilite procesu plodnosti. Vo všeobecnosti pritom platí, že ak v nejakej generácii identifikujeme výraznejšie diferencie v štruktúre podľa parity, môže vzniknúť situácia, keď ženy s nízkym počtom detí (napr. jedno, dve) aj pri svojom výrazne vyššom zastúpení v danej kohorte prispievajú k celkovej plodnosti výrazne nižšou váhou ako ženy s vyšším počtom detí, ale výrazne menej početne zastúpené a naopak. Snahou tak je určiť, aká časť žien sa podieľala na akej časti reprodukcie v danej generácii, teda číselne vyjadriť mieru prerozdelenia medzi počtom detí a počtom žien.

Úroveň tejto variability je možné analyzovať prostredníctvom koeficientu koncentrácie. Ten sa pohybuje v rozmedzí 0 – 1, pričom s rastúcou úrovňou dochádza k zvyšovaniu vnútornej variability procesu plodnosti vedúcej k nerovnomernému rozloženiu reprodukcie medzi ženami [7, 8]. Jeho hodnoty možno interpretovať ako priemernú vnútornú diferenciáciu počtu detí narodených ženám vo vzťahu k priemernému počtu detí [11]. Samotný výpočet koeficientu koncentrácie sa potom opiera o nasledujúci vzťah:

$$Kkr = 1 - \sum_{i=1}^m f_i \cdot (Z_{i-1} + Z_i) \quad (11)$$

kde:

f_i je relatívna početnosť detí narodených v sledovanej generácii príslušnej parity (i),
 Z_i je kumulatívna relatívna početnosť žien v sledovanej generácii, ktorým sa narodilo maximálne i -detí.

Okrem koeficientu koncentrácie je možné stupeň variability reprodukcie vyjadriť aj prostredníctvom tzv. *half-statistics* [bližšie 2, 3, 18]. Konkrétne ide o *havehalf*, ktorý identifikuje podiel žien, ktorým sa v analyzovanej generácii narodila polovica v nej živonarodených detí. Analogicky potom *halfhave* určuje podiel detí, ktoré sa narodili polovici všetkých žien v príslušnom populačnom ročníku.

Samotná konštrukcia týchto štatistických indikátorov sa opiera o relatívne početnosti žien (z_i) a detí (f_i) v sledovanej generácii triedené podľa parity (i). Z nich sú odvodené kumulatívne relatívne početnosti (Z_i , F_i), a to od najvyššej parity (m) po najnižšiu (0, bezdetné ženy). V prípade *have-half* je následne identifikovaný mediánový interval, v ktorom kumulatívna váha detí dosahuje hranicu 50 %. Potom platí, že:

$$havehalf = Z_a + h_z \cdot \frac{50\% - F_a}{f_a} \quad (12)$$

kde

Z_a je kumulatívna relatívna početnosť žien v sledovanej generácii rozdelených podľa parity po mediánový interval,

h_z je rozpätie mediánového intervalu,

F_a je kumulatívna relatívna početnosť detí rozdelených podľa poradia pôrodu v sledovanej kohorte po mediánový interval,

f_a je rozpätie mediánového intervalu.

Pre *half-have* analogicky najprv hľadáme interval, v ktorom podiel žien prekračuje hranicu 50 % a následne získané informácie dosadíme do rovnice:

$$\text{halfhave} = F_a + h_f \cdot \frac{50\% - Z_a}{z_a} \quad (13)$$

F_a je kumulatívna relatívna početnosť detí v sledovanej generácii rozdelených podľa poradia pôrodu po mediánový interval,

h_f je rozpätie mediánového intervalu,

Z_a je kumulatívna relatívna početnosť žien rozdelených podľa parity v sledovanej generácii po mediánový interval,

z_a je rozpätie mediánového intervalu.

3. KONEČNÁ PLODNOSŤ ŽIEN PODĽA VÝSLEDKOV SODB 2021

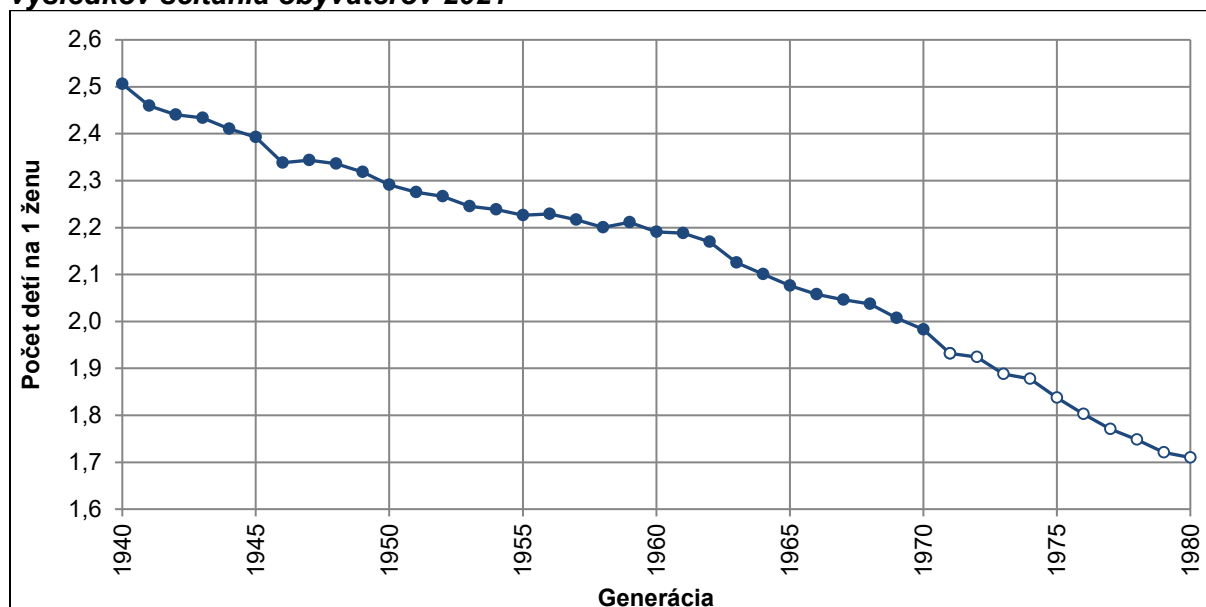
Hlavným vývojovým znakom konečnej plodnosti žien na Slovensku je jej dlhodobý medzigeneračný postupný pokles [4, 16, 17]. Potvrdzujú to aj výsledky sčítania obyvateľov 2021. Kým u žien narodených na začiatku 40. rokov minulého storočia sa priemerný počet živonarodených detí pohyboval nad hranicou 2,5 dieťaťa, v generáciách zo začiatku 50. rokov to bolo už len približne 2,3 dieťaťa a u osôb narodených o desaťročie neskôr približne 2,2 dieťaťa. Z uvedeného je zrejmé, že o niečo dynamickejší pokles realizovanej plodnosti nachádzame v najstarších analyzovaných generáciách zo 40. rokov (–0,19 dieťaťa medzi generáciami 1940 a 1949), aby následne v 50. rokoch nastalo určité zmiernenie dynamiky znižovania konečnej plodnosti (–0,08 dieťaťa). K ďalšej dynamizácii tohto procesu dochádza v generáciách zo 60. rokov (–0,18 dieťaťa), no kľúčovým je predovšetkým vývoj u žien narodených v 70. rokoch. Medzi generáciou 1970 a 1979 sa totižto priemerný počet detí na jednu ženu znížil o približne –0,26 dieťaťa. Údaje o väčšine žien narodených v 70. rokoch je však ešte potrebné vzhľadom na vek v čase sčítania (41 – 49 rokov) brať ako predbežné. Keďže však úroveň plodnosti v tomto vekovom spektre je vo všeobecnosti na Slovensku stále veľmi nízka, je potrebné tiež doplniť, že výsledná hodnota sa už s najväčšou pravdepodobnosťou významnejšie nezmení.

Podľa získaných údajov zo sčítania obyvateľov 2021 sa konečná plodnosť z približne 2 detí na ženu znížila do generácie 1975 na necelých 1,84 dieťaťa a v generácii 1980³ by dokonca len mierne prekračovala hranicu 1,71 dieťaťa. Potvrdzujú sa tak niektoré predpoklady [napr. 5, 12, 14], že ženy najmä z druhej polovice 70. rokov už na Slovensku budú mať výrazne menej ako dve deti.

Výsledky posledného sčítania obyvateľov 2021 tak poukazujú na ďalšie zrýchlenie medzigeneračného poklesu realizovanej plodnosti, ale aj na skutočnosť, že práve ženy narodené na konci 60. rokov sú poslednými, ktoré mali na Slovensku v priemere viac ako 2 deti (graf č. 1).

³ Tieto ženy v čase sčítania obyvateľov 2021 mali 40 rokov.

Graf. č. 1: Vývoj priemerného počtu živonarodených detí ženám na Slovensku s ukončenou reprodukciou alebo na jej konci v generáciách 1940 – 1980 podľa výsledkov sčítania obyvateľov 2021



Pozn.: Prázdne značky prezentujú údaje v generáciách, ktorých ženy v čase sčítania boli vo veku 40 – 49 rokov.

Zdroj: ŠÚ SR, SODB 2021, výpočty autora

4. ŠTRUKTÚRA ŽIEN PODĽA POČTU NARODENÝCH DETÍ

Konečná plodnosť predstavuje priemernú hodnotu, ktorá sa odvíja od štruktúry žien podľa počtu živonarodených detí. V podstate rovnaká konečná plodnosť môže tak znamenať úplne odlišný obraz paritného zloženia. Práve štruktúra žien po skončení reprodukčného obdobia prináša ďalší dôležitý rozmer pre komplexné zhodnotenie plodnosti a jej prípadných zmien. Predovšetkým poskytuje detailnú analýzu nastavenia charakteru reprodukčných modelov, ktoré aplikovali jednotlivé generácie žien.

Celkovo vyššia konečná plodnosť na Slovensku v rámci európskeho priestoru [1] bola dlhodobo podmienená predovšetkým častejším rodením tretieho a ďalších detí, pri súčasne pomerne nízkej bezdetnosti a jednodetnosti [16, 17]. V kontexte špecifických reprodukčných podmienok vytvorených počas minulého politického režimu [16] sa však postupne do popredia dostával dvojdetný model rodiny [4, 16, 17]. Ten podľa výsledkov sčítania obyvateľov, domov a bytov 2021 svoju maximálnu úroveň dosahoval v generáciách žien z druhej polovice 50. a prvej polovice 60. rokov, keď sa ich podiel pohyboval na úrovni 47 – 48 % (graf č. 2). V najstarších analyzovaných skupinách žien zo začiatku 40. rokov pritom tento model dosahoval ešte len okolo 41 – 42 % a miernu prevahu mali ženy s tromi a viac deťmi (42 – 43 %). Ich podiel však postupne medzigeneračne klesal, až sa v generáciách z druhej polovice 60. rokov sa dostal pod hranicu 25 % a u žien narodených na začiatku 70. rokov (posledné generácie s ukončenou reprodukciou v čase sčítania 2021) dosahoval takmer rovnaký podiel ako model rodiny s jedným dieťaťom (23 %). V ešte mladších generáciách žien, ktoré v čase sčítania boli v poslednej dekáde svojho reprodukčného veku, sa podiel modelu rodiny s tromi a viac deťmi dostal dokonca výrazne pod hranicu 20 % a v podstate sa vyrovnal so zastúpením bezdetných (graf č. 2). Aj v tomto prípade však ešte môžeme očakávať určitú miernu zmenu, ktorá nemôže nastavený trend významnejšie zmeniť.

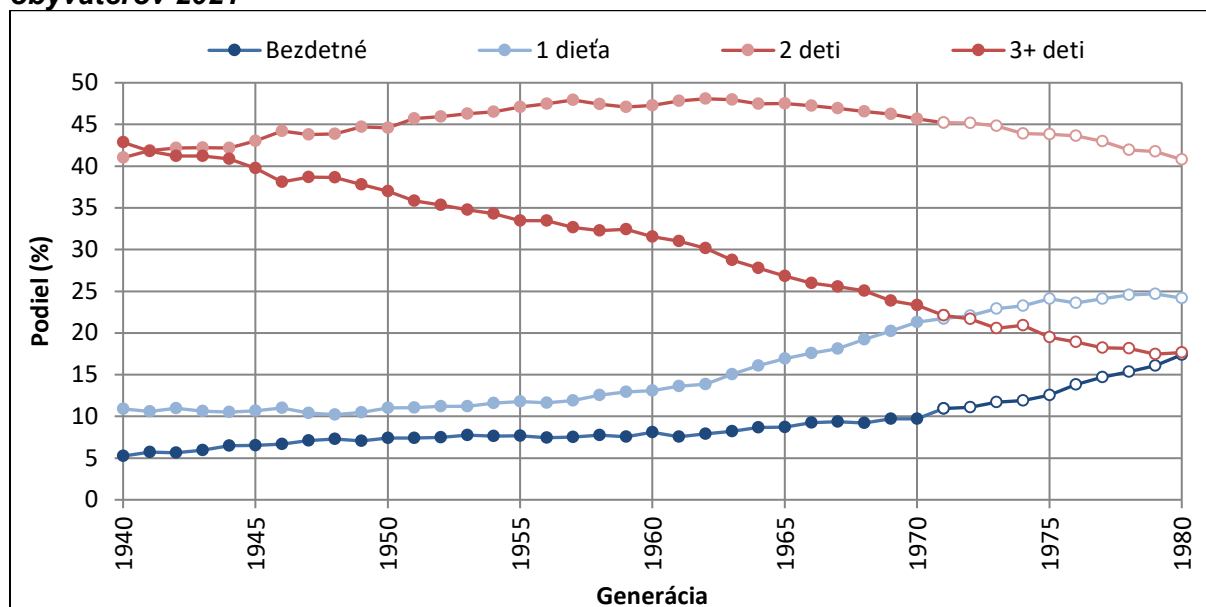
Náznak poklesu identifikujeme aj u žien s dvomi deťmi, a to už od generácií z polovice 60. rokov. Ich zastúpenie z viac ako 48 % kleslo na približne 45 % v skupinách žien zo začiatku 70. rokov. Smerom k ešte mladším skupinám žien v čase sčítania vo veku 40 – 49 rokov sa podiel osôb s dvomi deťmi dostal výraznejšie pod hranicu 45 % a u najmladších generácií zo začiatku 80. rokov dokonca ich váha len mierne prekračovala 40 % (graf č. 2).

Získané výsledky zo sčítania 2021 potvrdzujú zistenia z viacerých čiastkových predchádzajúcich analýz [5, 12, 16], ktoré od generácií žien z druhej polovice 60. rokov ukazujú na prehlbujúce sa odkladanie rodenia detí a v spojitosti s najmä druhými a ďalšími paritami aj pomerne nízku úroveň dobiehania tejto odloženej plodnosti vo vyššom veku. To by sa malo následne odzrkadľovať na rastúcej bezdetnosti a najmä jednodetnosti v neprospech klesajúceho zastúpenia žien s dvomi a viac deťmi.

Bezdetnosť bola na Slovensku s výnimkou žien narodených na konci 19. a začiatku 20. storočia dlhodobo relatívne nízka [13, 15, 16]. Navyše v európskom priestore pritom platilo, že smerom ku generáciám z 30. rokov dochádzalo k jej všeobecnej redukcii [6, 10]. Platilo to aj v prípade Slovenska, čo potvrdzujú údaje o podiele bezdetných žien na začiatku 40. rokov (graf č. 2). Smerom k mladším generáciám však pozorujeme postupný medzigeneračne pretrvávajúci rast bezdetnosti. Podľa výsledkov sčítania obyvateľov 2021 sa podiel bezdetných žien narodených medzi rokom 1940 a 1970 v podstate zdvojnásobil z približne 5 na takmer 10 %. U žien zo 70. rokov však vidíme ďalšie pokračovanie tohto trendu. Môžeme bezpečne predpokladať, že v ich prípade zostane viac ako desatina bez skúsenosti s biologickým materstvom, pričom u najmladších kohort sa dá očakávať aj bezdetnosť nad úrovňou 15 %. Pre doplnenie ešte dodávame, že údaj o skupine žien narodených v roku 1980, ktoré v čase sčítania 2021 boli na začiatku poslednej dekády reprodukčného obdobia, pritom hovorí, že bezdetnou v tomto veku bolo ešte viac ako 17 % žien. Výsledky sčítania obyvateľov z roku 2021 tak na jednej strane potvrdzujú predpoklad o pokračujúcom medzigeneračnom raste bezdetnosti na Slovensku [9, 13], no na druhej strane poukazujú na jej o niečo nižšiu reálnu úroveň (pod hranicou 20 %), ako tieto prognózy očakávali.

Dlhodobo marginálnym reprodukčným modelom na Slovensku bola okrem bezdetnosti aj jednodetnosť. Podiel žien v generáciách zo 40. rokov, ktorým sa počas ich reprodukčného obdobia narodilo len jedno dieťa, sa podľa výsledkov sčítania obyvateľov 2021 v podstate stabilne pohyboval na úrovni 10 – 11 % (graf č. 2). Až u žien narodených v 50. rokoch pozorujeme nástup rastúceho trendu, ktorý pretrval v pozvoľnej podobe až do generácií zo začiatku 60. rokov, keď podiel žien s jedným dieťaťom dosiahol približne 14 %. Smerom k mladším skupinám však dochádza k zrýchleniu nárastu jednodetnosti, čím v generáciách 1969 a 1970 tento model rodiny prekročil hranicu 20 %. Ako je však zjavné z grafu č. 2, u žien zo 70. rokov nastáva určité spomaľovanie tohto trendu. V najmladších analyzovaných generáciách z konca 70. rokov podiel žien s jedným dieťaťom pravdepodobne výraznejšie neprekročí hranicu 25 %. Vzhľadom na vyššie identifikovaný vývoj v prípade podielu žien s dvomi deťmi uvedené zistenie o váhe jednodetnosti medzi najmladšími populačnými ročníkmi žien nie je ani tak dôsledkom častejšieho rodenia druhého a ďalšieho dieťaťa, ale najmä zistenej rastúcej bezdetnosti.

Graf č. 2: Vývoj štruktúry žien podľa počtu narodených detí na Slovensku s ukončenou reprodukciou alebo na jej konci v generáciách 1940 – 1980 podľa výsledkov sčítania obyvateľov 2021



Pozn.: Prázdné značky prezentujú údaje v generáciách, ktorých ženy v čase sčítania boli vo veku 40 – 49 rokov.

Zdroj: ŠÚ SR, SODB 2021, výpočty autora

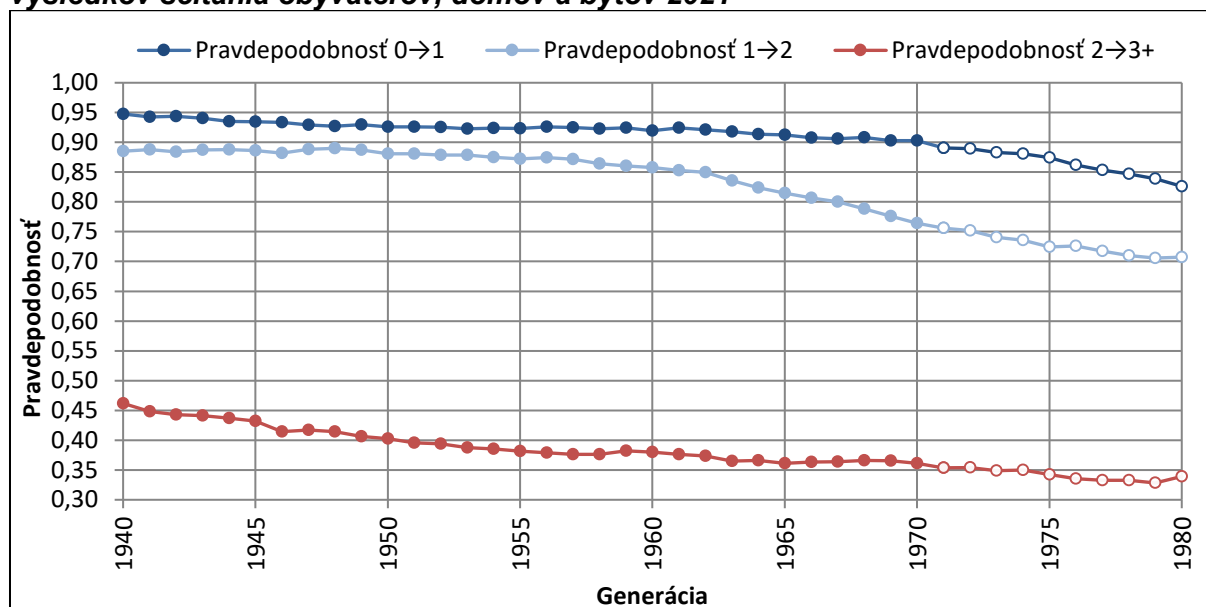
5. PRAVDEPODOBNOŠŤ ZVÄČŠENIA RODINY

Štruktúra žien podľa počtu živonarodených detí (paritná štruktúra) a s tým spojená konečná plodnosť sú úzko previazané s úrovňou pravdepodobnosti narodenia detí jednotlivých parít [19]. Pre obe je potom dôležité nielen to, akú úroveň dosahujú, ale aj ako sa táto úroveň medzigeneračne menila.

Zistená dlhodobou prevládajúca pomerne nízka bezdetnosť na Slovensku bola odrazom existujúcej vysokej pravdepodobnosti, že sa bezdetnej žene narodí aspoň jedno dieťa. Jej hodnota sa podľa výsledkov sčítania obyvateľov 2021 stabilne medzi generáciami 1940 – 1969 stabilne pohybovala nad hranicou 0,9 (90 %). Na druhej strane je však z grafu č. 3 zrejme, že postupne nastal určitý mierny medzigeneračný pokles, a preto v posledných skupinách žien s ukončenou reprodukciou v čase sčítania došlo k prelomeniu vyššie spomenutej hranice. Súčasne je zrejme, že najvyššie šance stať sa aspoň raz matkou dosahovali ženy narodené na začiatku 40. rokov. Ide o osoby, ktoré v podstate celé svoje reprodukčné obdobie prežili v minulom politickom režime. To nepriamo svedčí o vysokej normativite materstva, ktorá v slovenskej spoločnosti v týchto časoch prevládala.

Údaje za generácie zo 70. rokov, teda skupiny žien s neukončenou reprodukciou, však v čase sčítania s vekom v poslednej dekáde reprodukčného obdobia, hovoria o ďalšom poklese pravdepodobnosti narodenia prvého dieťaťa bezdetným. Tá sa tak dostala významne pod 90 % a u najmladších z nich dokonca prelomila aj úroveň 85 %. Posledná zaznamenaná hodnota pre kohortu žien z roku 1980 predstavovala menej ako 83 % (graf č. 3).

Graf č. 3: Vývoj pravdepodobnosti zväčšenia rodiny vybraných parít žien na Slovensku s ukončenou reprodukciou alebo na jej konci v generáciách 1940 – 1980 podľa výsledkov sčítania obyvateľov, domov a bytov 2021



Pozn.: Pravdepodobnosť 0→1 je pravdepodobnosť narodenia prvého živého dieťaťa bezdetným ženám; pravdepodobnosť 1→2 je pravdepodobnosť narodenia druhého živého dieťaťa ženám s jedným dieťaťom; pravdepodobnosť 2→3+ je pravdepodobnosť narodenie tretieho a ďalšieho dieťaťa ženám s dvomi deťmi.

Prázdne značky prezentujú údaje v generáciách, ktorých ženy v čase sčítania boli vo veku 40 – 49 rokov.

Zdroj údajov: ŠÚ SR, SODB 2021, výpočty autora

Pravdepodobnosť narodenia druhého dieťaťa žene s jedným dieťaťom dosahovala podľa výsledkov sčítania 2021 v najstarších analyzovaných generáciách približne stabilnú úroveň 88 – 89 %. Približne od žien narodených v prvej polovici 50. rokov však nastáva postupný pokles. Preto v generáciách zo začiatku 60. rokov táto pravdepodobnosť dosiahla 85 % a u žien narodených na konci tejto dekády dokonca klesla pod hranicu 80 % (graf č. 3). Zistený odklon od dvojdetnej rodiny potvrdzuje aj pokračujúci pokles pravdepodobností v generáciách zo 70. rokov. U žien narodených v roku 1970 tak pravdepodobnosť narodenia druhého dieťaťa klesla na približne 76 %. V generáciách z polovice 70. rokov to bolo niečo viac ako 72 % a u najmladšej skupiny z rokov 1979 a 1980 sa dokonca dostala na hranicu 70 %.

Znamená to, že kým v generáciách realizujúcich prevažnú časť svojej reprodukcie počas minulého politického režimu až takmer 9 z desiatich žien majúcich jedno dieťa malo napokon aj druhé, v skupinách, ktorých reprodukčné správanie bolo najviac zasiahnuté celospoločenskou transformáciou po roku 1989, to je (zatiaľ) len približne v 70 – 75 % prípadov.

Znižujúci sa podiel žien s tromi a viac deťmi bol medzigeneračne podmienený predovšetkým prehlbujúcou sa snahou o kompletizáciu veľkosti rodiny na dve deti. Potvrdzuje to aj vývoj pravdepodobností narodenia tretieho a ďalšieho dieťaťa ženám s dvomi deťmi (graf č. 3). Kým v najstarších analyzovaných generáciách platilo, že až v 45 % prípadov sa ženám s dvomi deťmi narodilo do konca ich reprodukčného obdobia ďalšie dieťa, u žien zo začiatku 50. rokov to bolo už len približne v 40 % prípadov. Pokračovanie tohto vývojového trendu potvrdili aj výsledky sčítania

obyvateľov 2021. Na druhej strane je pritom potrebné tiež poukázať na určité spomalenie až stagnáciu u žien narodených v druhej polovici 60. rokov a naopak, na ďalšie zrýchlenie u osôb zo 70. rokov (graf č. 3). Preto o najmladších z analyzovaných generácií platí, že ženy s dvomi deťmi mali len približne tretinovú pravdepodobnosť stať sa matkou aj tretieho a ďalšieho dieťaťa (graf č. 3).

6. KONCENTRÁCIA REPRODUKČIE

Pokles bezdetnosti a jej stabilizácia na pomerne nízkej úrovni, odklon od viacdetej rodiny a naopak, výrazná preferencia dvojdetného modelu priniesli na Slovensku z hľadiska procesu plodnosti značnú homogenizáciu reprodukčného správania najmä u žien realizujúcich svoje reprodukčné zámery v období minulého politického režimu. Tento vývoj bol zaznamenaný aj mimo bývalého východného bloku [pozri napr. 3, 7, 8, 11]. Došlo tak k poklesu vnútornej diverzity žien podľa paritnej štruktúry. Keďže sme tiež boli svedkami redukcie bezdetnosti, dá sa povedať, že na výslednej realizovanej plodnosti sa podieľala drvivá väčšina všetkých žien. Smerom k mladším generáciám však títo autori upozorňovali na postupnú heterogenizáciu a tým aj vnútornú diverzifikáciu reprodukcie. Tento trend môžeme na základe výsledkov sčítania obyvateľov, domov a bytov 2021 identifikovať aj na Slovensku.

Ako je zrejmé z grafu č. 4, v najstarších analyzovaných skupinách žien narodených v 40. a 50. rokoch nachádzame pomerne stabilné hodnoty koeficientu koncentrácie. To potvrdzuje značnú unifikáciu reprodukčného správania, ktorá vrcholila práve v generáciách žien z konca 40. a začiatku 50. rokov (graf č. 4). V mladších skupinách žien narodených od približne polovice 50. rokov však dochádza k pomerne významnému rastu variability reprodukcie. Ten je podmienený jednak zvyšujúcou sa bezdetnosťou a jednodetnosťou, ako aj poklesom dominantného modelu dvojdetnej rodiny. Podľa údajov zo sčítania 2021 tak koeficient koncentrácie dosahuje v najmladších analyzovaných generáciách žien z konca 70. rokov už takmer o polovicu vyššiu úroveň, ako tomu bolo v prípade spomínaných kohort z konca 40. a prvej polovice 50. rokov (graf č. 4). Aj to poukazuje, ako výrazne sa medzigeneračne zmenil model reprodukčného správania u žien najviac zasiahnutých celospoločenskou transformáciou po roku 1989.

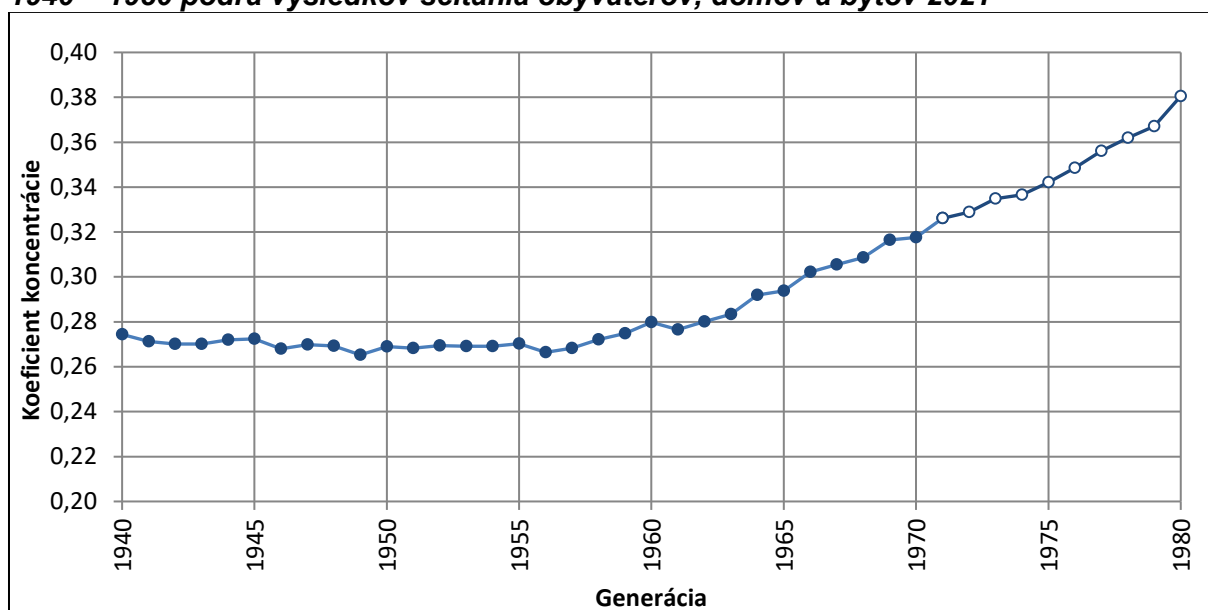
V dôsledku nízkej bezdetnosti, jednodetnosti a preferencii dvojdetného modelu rodiny v najstarších analyzovaných skupinách žien zo 40. rokov platilo, že viac ako tretina z nich sa podieľala na polovici celkového počtu narodených detí v týchto generáciách.

V mladších generáciách však identifikujeme postupný pokles hodnoty indikátora *havehalf* (graf č. 5). Znamená to, že čoraz menšia časť žien sa podieľala na narodení polovice všetkých detí, ktoré sa narodili v príslušnom populačnom ročníku. Podľa výsledkov sčítania 2021 v generácii 1965 sa na polovici narodených detí nepodieľalo už ani 30 % všetkých žien a v generáciách z druhej polovice 70. rokov to už nie je ani štvrtina žien (graf č. 5). V konečnom dôsledku to tiež znamená, že dochádza k určitej heterogenizácii, a teda nerovnomernejšiemu rozloženiu reprodukcie mladších generácií.

Potvrdzuje to aj ukazovateľ *halfhave* (graf č. 6). Vzhľadom na klesajúcu variabilitu a tiež zvyšujúci sa podiel žien spolupodieľajúcich sa na realizovanej plodnosti nastal najprv určitý pokles jeho hodnôt. Najnižšiu úroveň podľa výsledkov sčítania obyvateľov

2021 dosahoval v generáciách zo začiatku 50. rokov, keď polovica žien z nich mala niečo viac ako dve tretiny detí. V dôsledku spomínanej zvyšujúcej sa variability reprodukcie však u mladších generácií začína *halfhave* opätovne rásť. Podľa údajov zo sčítania 2021 polovica žien narodených na začiatku 70. rokov mala približne 70 % všetkých detí. U žien z polovice 70. rokov to bolo už viac ako 72 % a v najmladších generáciách z konca 70. a začiatku 80. rokov sa hodnota *halfhave* dostala už takmer k hranici 76 %.

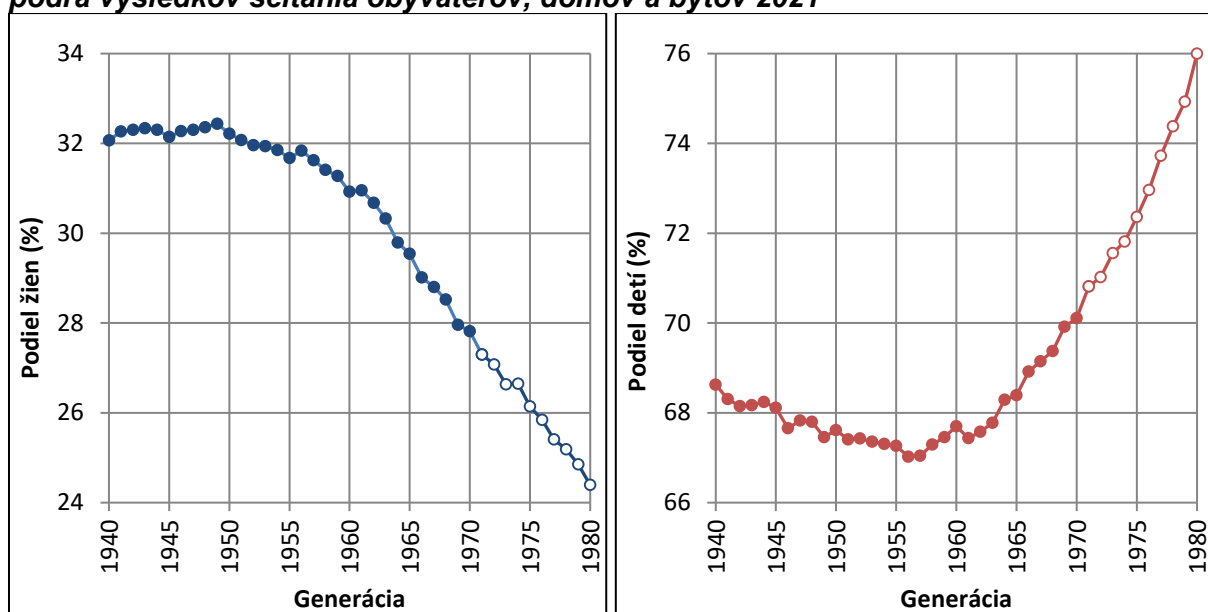
Graf č. 4: Vývoj koeficientu koncentrácie plodnosti žien Slovenska v generáciách 1940 – 1980 podľa výsledkov sčítania obyvateľov, domov a bytov 2021



Pozn.: Prázdné značky prezentujú údaje v generáciách, ktorých ženy v čase sčítania boli vo veku 40 – 49 rokov.

Zdroj údajov: ŠÚ SR, SODB 2021, výpočty autora

Graf č. 5 a 6: Podiel žien majúcich polovicu detí (*havehalf*) a podiel všetkých detí pripadajúcich na polovicu žien (*halfhave*) v generáciách 1940 – 1980 na Slovensku podľa výsledkov sčítania obyvateľov, domov a bytov 2021



Pozn.: Prázdné značky prezentujú údaje v generáciách, ktorých ženy v čase sčítania boli vo veku 40 – 49 rokov. Zdroj údajov: ŠÚ SR, SODB 2021, výpočty autora

7. ZÁVER

Výsledky najnovšieho sčítania obyvateľov domov a bytov 2021 potvrdili dlhodobý medzigeneračný pokles priemerného počtu živonarodených detí ženám na Slovensku. Tento trend sa navyše v generáciách zo 70. rokov pravdepodobne ešte zrýchlil. V dôsledku toho je zrejmé, že ženy narodené v 60. rokoch sú poslednými, ktoré v priemere mali viac ako dve deti, pričom hodnota konečnej plodnosti u žien narodených v druhej polovici 70. rokov klesne pod úroveň 1,8 dieťaťa.

V pozadí týchto zmien sa ukazujú byť významné posuny v štruktúre žien podľa počtu živonarodených detí podmienené meniacimi sa pravdepodobnosťami zväčšenia rodiny. Predovšetkým sme v najmladších analyzovaných populačných ročníkoch svedkami znižovania pravdepodobností, že sa bezdetná žena stane do konca reprodukčného obdobia matkou ako aj toho, že sa žene s jedným dieťaťom narodí druhé.

To má za následok v najmladších generáciách postupný pokles podielu žien s dvomi a viac deťmi a naopak, zvyšujúce sa zastúpenie bezdetných žien a žien, ktorým sa do dovŕšenia 50. roku života narodí len jedno dieťa. Podľa údajov u žien vo veku v čase sčítania v poslednej dekáde reprodukčného obdobia (pozri graf č. 1) by jednodetnosť mohla dosahovať takmer štvrtinu a bezdetnosť niečo viac ako 15 % v príslušnej generácii.

Meniaca sa štruktúra žien podľa počtu narodených detí sa odzrkadľuje aj na rastúcej variabilite reprodukcie. Kým ženy realizujúce svoje reprodukčné zámery počas minulého politického režimu vďaka nízkej bezdetnosti a jednodetnosti a značnému príklonu k dvojdetnému modelu rodiny vytvárali značne unimorfny model reprodukcie, v najmladších generáciách sledujeme postupný rast hodnoty koeficientu koncentrácie plodnosti a tým aj samotnú heterogenizáciu reprodukcie. To sa napokon prejavuje aj na miere zapojenia žien do realizovanej plodnosti. Kým podiel žien, ktorým sa narodila polovica všetkých detí v danej generácii, postupne medzigeneračne klesá, opačný trend má vývoj podielu detí, ktoré sa narodili polovici všetkých žien. Aj to nepriamo naznačuje, že čoraz väčšia časť žien sa nezapája do reprodukčného procesu, alebo sú ich príspevky vzhľadom na svoju početnosť neadekvátne nízke.

LITERATÚRA

- [1] FREJKA, T. – SARDON, J-P.: Childbearing Trends and Prospects in Low-Fertility Countries. A Cohort Analysis. European Studies of Population, roč. 13, Dordrecht, Boston a Londýn : Kluwer Academic Publishers, 2004.
- [2] GOODWIN, D. – VAUPEL, J.: Concentration Curves and Have-Statistics for Ecological Analysis of Diversity: Part III: Comparison of Measures of Diversity. Working Paper 85-91. Laxenburg, Austria: IIASA. 28 pp. 1985.
- [3] LUTZ, W.: On The Concentration of Childbearing and Rearing in Austria. Working Paper IIASA The Division of Labor for Societies Reproduction, 1986.
- [4] POTANČOKOVÁ, M.: Plodnosť žien na Slovensku v období rokov 1950 – 2007 v generačnom pohľade. Bratislava: INFOSTAT, 2008.
- [5] POTANČOKOVÁ, M.: Zmena reprodukčného správania populácie Slovenska poroku 1989: trendy, príčiny a dôsledky. In: Piscová, M. (ed.) Desaťročia premien slovenskej spoločnosti. Bratislava: VEDA, SAV, 2011, s. 142 – 159.
- [6] ROWLAND, D. T.: Historical Trends in Childlessness. In: Journal of Family Issues. 2007, roč. 28, č. 10, s. 1314 – 1315.

- [7] SHKOLNIKOV, V. – ANDREEV, E. – HOULE, R. – VAUPEL, J.: To concentration of reproduction in cohorts of US and European women. MPIDR Working Paper WP 2004-027, 2004.
- [8] SHKOLNIKOV, V. – ANDREEV, E. – HOULE, R. – VAUPEL, J.: The Concentration of Reproduction in Cohorts of Women in Europe and the United States. *Population and Development Review*, 2007, roč. 33, č. 1, s. 67 – 99.
- [9] SOBOTKA, T.: Childlessness societies? Trends and projections of childlessness in Europe and the United States. In: *Population Association of America 2005 Annual Meeting Program*. Philadelphia, Pennsylvania March 31- April 2, 2005
- [10] SOBOTKA, T.: Childlessness in Europe: Reconstructing Long-Term Trends Among Women Born in 1900–1972. In: Kreyenfeld, M. – Konietzka, D. (eds.): *Childlessness in Europe: Contexts, Causes, and Consequences*. Demographic Research Monographs, 2017, s. 17 – 53.
- [11] SPIELAUER, M.: Concentration of Reproduction in Austria: General Trends and Differentials by Educational Attainment and Urban-Rural Setting. *Vienna Yearbook of Population Research*, 2005, s. 171 – 195.
- [12] ŠPROCHA, B.: Odkladanie a rekuperácia plodnosti v kohortnej perspektíve v Českej republike a na Slovensku. In: *Demografie*, 2014, roč. 56, č. 3, s. 219 – 233.
- [13] ŠPROCHA, B.: Bezdetnosť a odkladanie rodenia prvých detí v krajinách Vyšehradskej skupiny. In: *Demografie*, 2022, roč. 64, č. 2, s. 139 – 157.
- [14] ŠPROCHA, B. – ĎURČEK, P.: Odkladanie materstva na Slovensku v generačnej perspektíve. In: *Sociológia*, 2018, roč. 50, č. 5, s. 550 – 575.
- [15] ŠPROCHA, B. – ŠTASTNÁ, A. – ŠÍDLO, L.: Bezdetnosť – nový fenomén v životných dráhach žien na Slovensku. In: *Sociológia*, 2017, roč. 49, č. 5, s. 561 – 587.
- [16] ŠPROCHA, B. – TIŠLIAR, P.: Transformácia plodnosti žien Slovenska v 20. a na začiatku 21. storočia. Bratislava: Muzeológia a kultúrne dedičstvo, 2016. 376 s.
- [17] ŠPROCHA, B. – TIŠLIAR, P.: 100 rokov obyvateľstva Slovenska: od vzniku Československa po súčasnosť. Bratislava: Muzeológia a kultúrne dedičstvo, 2018. 445 s.
- [18] VAUPEL, J. – GOODWIN, D.: The Concentration of Reproduction among US Women, 1917–80. In: *Population and Development Review*, 1987, roč. 13, č. 4, s. 723-730.
- [19] ZEMAN, K. – BEAUJOUAN, E. – BRZOSOWSKA, Z. – SOBOTKA, T.: Cohort fertility decline in low fertility countries: Decomposition using parity progression ratios. In: *Demographic Research*, 2018, roč. 38, č. 25, s. 651 – 690.

RESUMÉ

Sčítanie obyvateľov, domov a bytov na Slovensku predstavuje jeden z najdôležitejších zdrojov údajov na generačnú analýzu procesu plodnosti. Už od roku 1930 je integrálnou súčasťou obsahu sčítaní otázka o počte (živo)narodených detí ženám. V kombinácii s vekom, respektíve rokom narodenia tak umožňuje analyzovať nielen úroveň realizovanej plodnosti, ale aj štruktúru žien podľa počtu narodených detí, či vnútorné nastavenie procesu reprodukcie. Hlavným cieľom príspevku bola podrobná analýza generačnej plodnosti žien na Slovensku prostredníctvom výsledkov najnovšieho sčítania obyvateľov, domov a bytov 2021. Zamerali sme sa pritom na medzigeneračný vývoj konečnej plodnosti, štruktúry žien podľa počtu živonarodených detí, pravdepodobnosti zväčšenia rodiny a koncentrácie reprodukcie. Získané výsledky potvrdili pokračujúci postupný pokles priemerného počtu detí narodených jednej žene na Slovensku. Generácie žien zo 60. rokov sú poslednými,

ktoré v priemere mali viac ako dve deti. U žien zo 70. rokov môže konečná plodnosť poklesnúť až k hranici 1,7 dieťaťa na ženu. Tento vývoj je podmienený najmä znižujúcou sa pravdepodobnosťou narodenia prvého dieťaťa bezdetným ženám a druhého ženám s jedným dieťaťom. To sa následne odzrkadľuje aj na rastúcom podiele bezdetných žien a žien len s jedným dieťaťom pri súčasnom znižovaní zastúpenia dvojdetného modelu rodiny. Rovnako výsledky sčítania obyvateľov, domov a bytov 2021 poukazujú na rastúcu variabilitu reprodukcie pri znižovaní aktívnej účasti čoraz väčšej časti žien na jej realizácii.

RESUME

The Population and Housing Census in Slovakia represents one of the most important sources of data for cohort analysis of fertility. Since 1930, the question of the number of (live-) births to women has been an integral part of the census content. In combination with age or the year of birth, thus it enables to analyze allows us analyzing not only the level of realized fertility, but also the parity structure of women, or the internal setting of the reproduction process. The main goal of the article was in a detailed analysis of the cohort fertility of women in Slovakia by means of through the results of the latest 2021 Population and Housing Census. We focused on the inter-cohort development of completed fertility, parity structure, parity progression ratios and concentration of reproduction. The obtained results confirmed the continuing gradual decrease in the average number of children born to one woman in Slovakia. The cohorts of women from the 1960s are the last to have more than two children on average. For women from the 1970s, cohort completed fertility may drop to the limit of 1.7 children per woman. This development is mainly conditioned by the decreasing probability of the birth of the first child for childless women and the second for women with one child. This is subsequently reflected in the growing shares of childless women and women with only one child, while simultaneously reducing the representation of the two-child family model. Likewise, the results of the 2021 Population and Housing Census point to the growing variability of reproduction while reducing the active participation of an increasing number of women in its implementation.

PROFESIJNÝ ŽIVOTOPIS

RNDr. Branislav Šprocha, PhD., absolvoval magisterské štúdium na Prírodovedeckej fakulte Univerzity Karlovej v Prahe v odbore demografia a geodemografia (2006). V roku 2011 ukončil doktorandské štúdium v programe demografia a v roku 2021 sa na Prírodovedeckej fakulte Univerzity Komenského v Bratislave habilitoval v odbore humánna geografia. Od roku 2007 je vedeckovýskumným pracovníkom Výskumného demografického centra pri INFOSTATE a od roku 2009 vedeckým pracovníkom Prognostického ústavu Centra spoločenských a psychologických vied SAV. V roku 2015 sa stal vedúcim Výskumného demografického centra. V oblasti demografie sa špecializuje na problematiku rodinného a reprodukčného správania a ich vplyvu na spoločnosť. Okrem toho sa zameriava na analýzu vybraných populačných štruktúr, reprodukčného správania rómskeho obyvateľstva na Slovensku a otázky konštrukcie populačných prognóz.

KONTAKT

branislav.sprocha@gmail.com

Peter KNÍŽAT, Helena GLASER-OPITZOVÁ
Štatistický úrad Slovenskej republiky, Ekonomická univerzita v Bratislave

**INDEX SPOTREBITEĽSKÝCH CIEN Z WEBS CRAPOVANÝCH ÚDAJOV:
ANALÝZA VYBRANEJ PRODUKTOVEJ SKUPINY**

**CONSUMER PRICE INDEX FROM WEB-SCRAPED DATA:
ANALYSIS OF SPECIFIC PRODUCT CATEGORY**

ABSTRAKT

V dôsledku zmien spotrebiteľského správania sa spotrebiteľ orientuje viac na nákup cez internet. Štatistické inštitúcie zodpovedné za zber cien tovarov a služieb sú nútené prehodnotiť tradičný zber cien pre oblasť cenovej štatistiky a v niektorých prípadoch ho potenciálne nahradiť automatickým zberom cien cez internet, tzv. webscrapingom. Implementácia takéhoto zdroja údajov prináša so sebou rôzne výzvy, od otázok v metodologickej oblasti až po významnú zmenu procesov spracovania údajov. Ide o spracovanie veľkého množstvo údajov, vrátane hodnotenia ich kvality, výberu reprezentantov a určenie cien jednotlivých tovarov, ktoré sú obvykle scrapované na dennej báze. Ďalšou výzvou je výber metódy na výpočet indexu spotrebiteľských cien, ktorá sa môže zásadne odlišovať od metódy výpočtu indexu spotrebiteľských cien použitej pri tradičnom zbere údajov. Cieľom tohto článku je predstaviť teoretický rámec na implementáciu webscrapovaných údajov do produkcie cenových štatistík. V prípadovej štúdii sme použili údaje o cenách pre produktovú skupinu chladničky, ktoré boli scrapované z webového porovnávacieho portálu <https://www.heureka.sk/>.

ABSTRACT

As a consequence of changes in the consumer behaviour, a consumer prefers to shop online. Statistical institutions responsible for the collection of prices of goods and services for the area of price statistics are obligated to reconsider the traditional collection of prices and in some cases potentially replace it with automated collection of prices through internet, also called web-scraping. The implementation of this type of data sources entails various challenges, from questions in the methodological field to a significant changes of data processing. This involves the processing of big data including the evaluation of their quality, the selection of representatives and the determination of prices of individual goods, which are usually scraped on a daily basis. Another challenge is the selection of methodology for estimating the consumer price index (CPI) that can be fundamentally different from CPI estimation methodology used in the traditional data collection. The aim of this study is to present a theoretical framework for the implementation of web-scraped data in the production for price statistics. In the case study, we used data for the product category of refrigerators, scraped from the comparison website heureka.sk.

KLÚČOVÉ SLOVÁ

webscrapované údaje, index spotrebiteľských cien, Jevonsov index, multilaterálne indexy

KEY WORDS

web-scraped data, consumer price index, Jevons index, multilateral indexes

1. ÚVOD

Ceny tovarov, ktoré vstupujú do výpočtu indexu spotrebiteľských cien sa tradične zberajú v kamenných predajniach, vyhľadávaním na internetových stránkach alebo štatistickým zisťovaním. Tovary v spotrebnom koši sa každoročne aktualizujú na základe spotrebiteľského správania spoločnosti a vývoj ich cien sa sleduje a zaznamenáva mesačne. Index spotrebiteľských cien sa počíta na mesačnej báze.

V posledných rokoch dochádza k zmene spotrebiteľského správania. Spotrebiteľ sa viac orientuje na nákup cez internet, ktorý ponúka väčší sortiment niektorých tovarov, rýchle porovnanie cien a parametrov kvality jednotlivých tovarov a v niektorých prípadoch nižšiu cenu v porovnaní s cenami v kamenných predajniach.

V dôsledku tejto zmeny spotrebiteľského správania sú štatistické inštitúcie zodpovedné za zisťovanie cien tovarov a služieb a výpočet indexov spotrebiteľských cien nútené k prehodnoteniu a revízii tradičného zberu cien a jeho potenciálnemu nahradeniu automatickým zberom cien cez internet, tzv. webscrapovaním údajov.

Webscrapovanie údajov je sťahovanie informácií z internetu pomocou automatických robotov vytvorených v programovacích jazykoch, ktoré po naprogramovaní nepotrebujú žiaden ľudský zásah – intervencia je potrebná len pri nutnom preprogramovaní, napríklad pri zmene štruktúry webovej stránky, blokovaní robota majiteľom webovej stránky atď.

Štatistický úrad SR inicioval projekt zberu údajov webscrapovaním v spolupráci s Infostatom v roku 2020. Produkty, ktoré sa scrapujú, patria najmä do skupiny bielej a čiernej techniky a stiahnuté údaje obsahujú identifikátor tovaru, cenu a charakteristické vlastnosti jednotlivých tovarov. Sťahovanie prebieha každý tretí deň v nočných hodinách, aby sme neprímerane nezaťažovali webovú stránku <https://www.heureka.sk/>.

Dôsledkom takéhoto spôsobu zberu údajov je potreba omnoho náročnejšieho spôsobu spracovania, čo zahŕňa analýzu kvality údajov, ich očistenie, výber tovarov do spotrebného koša a ich agregáciu na mesačnú úroveň predtým, než vôbec vstúpia do výpočtu indexu spotrebiteľských cien. Aj napriek z výšeným nárokom na IT infraštruktúru a tiež IT zručnosti zamestnancov môže mať podľa nášho názoru tento zdroj údajov potenciál na jeho využitie v praxi.

Autori vo vedeckom článku [11] testovali implementáciu webscrapingu údajov o cenách spotrebnej elektroniky a leteniek do výpočtu talianskeho harmonizovaného indexu spotrebiteľských cien. Ich štúdia sa zaoberá samotným scrapingovým nástrojom a IT infraštruktúrou potrebnou pre webscraping. Autori konštatujú, že scrapovanie údajov z webu má veľký potenciál nahradiť tradičné postupy zberu údajov, ale pre štatistické inštitúcie predstavuje veľkú výzvu z hľadiska spracovania veľkého množstva údajov (big data) pre oficiálnu štatistiku.

Ďalšou otázkou je výpočet samotného indexu spotrebiteľských cien. Vzorec na výpočet indexu spotrebiteľských cien sa môže zásadne odlišovať od vzorca, ktorý sa používa pri tradičnom zbere údajov, keďže pri webscrapingu je zozbierané veľké množstvo údajov pre všetky dostupné tovary v online predaji.

V online predaji pri niektorých druhoch tovarov dochádza k ich častej obmene. Napríklad, na trh vstupujú nové tovary a vystupujú z neho staré, v niektorých prípadoch ide o opätovný vstup toho istého tovaru na trh s viac alebo menej vylepšenými vlastnosťami (parametrami). Autori v článku [5] nazývajú tento jav zmenou kvality tovaru, ktorú je potrebné pri výpočte cenovej zmeny zohľadniť. Na meranie zmeny kvality medzi novým a starým tovarom sú vhodné a v akademickej literatúre odporúčané hedonické a tzv. time-dummy regresné modely. Pre podrobný popis a použitie pozri vedecká štúdia [11], alebo obsiahla príručka [16] pre hedonické regresie a úpravu kvality v cenových indexoch.

Ďalej môže dochádzať k dočasnej nedostupnosti cien niektorých tovarov v niektorých mesiacoch, ktoré by pri bilaterálnych indexoch spotrebiteľských cien založených na fixnom spotrebnom koši neboli zohľadnené vo výpočte. Fixný spotrebný koš sa môže pri využití scrapovaných údajov nahradiť dynamickým spotrebným košom, keď výpočet indexu spotrebiteľských cien je medzimesačný s tzv. reťazením na bázi obdobia, čo však môže viesť k vychýleniu indexu, tzv. „driftu“ [5].

Autori vo vedeckej štúdi [10] vyriešili problém vychýlenia reťazeného indexu spotrebiteľských cien, ktoré nastáva pri mesačnom reťazení, rozšírením metódy GEKS ([8], [7] a [14]), ktorá sa toho času používala na porovnávanie priemernej zmeny cien medzi krajinami, na porovnávanie priemernej zmeny cien v čase. GEKS patrí do skupiny multilaterálnych indexov.

Pri použití multilaterálnych indexov, keď sa porovná priemerná zmena cien tovarov medzi viac ako dvomi časovými obdobiami súčasne, je možné zobrať do úvahy aj tovary, ktorých ceny v niektorých mesiacoch chýbajú. V uvedenej metóde GEKS je potrebná zhoda tovarov len pre každé dva porovnávané mesiace. Pri použití regresných typov indexov nie je potrebná žiadna zhoda tovarov medzi porovnávanými mesiacmi.

Vo vedeckej štúdi [5] autori vypočítali indexy spotrebiteľských cien pomocou metódy GEKS s hedonickými imputáciami a bez imputácií, t. j. tovary, ktorých ceny sa nenachádzali v oboch porovnávaných mesiacoch boli vylúčené. Taktiež vypočítali index spotrebiteľských cien pomocou regresnej metódy s časovou umelou premennou, tzv. time-product dummy (TPD) index spotrebiteľských cien. V prípadovej štúdi použili transakčné, tzv. scanner, údaje cien elektroniky predávanej na Novom Zélande, ktoré sa získavajú priamo od obchodných reťazcov, Tieto údaje obsahujú informácie o predaji, tržbách a predanom množstve jednotlivých tovarov, takže autori mohli vypočítať vážené indexy spotrebiteľských cien. V článku konštatovali, že najvhodnejšia metóda na výpočet indexu spotrebiteľských cien je GEKS s hedonickými imputáciami.

Autori v článku [6] použili na výpočet indexov spotrebiteľských cien údaje scrapované z internetu holandským štatistickým úradom. Článok konštatuje, že ceny tovarov získané online by mohli čiastočne nahradiť ceny zberané tradičným spôsobom, keďže online zber cien je omnoho efektívnejším spôsobom zberu. Autori vypočítali indexy spotrebiteľských cien v dennej a týždennej periodicite pre 3 produktové kategórie: dámske tričká, mužské hodinky a kuchynské potreby. V článku sa konštatuje, že TPD index spotrebiteľských cien má praktickú výhodu, najmä ak je naším cieľom vypočítať vysokofrekvenčný index spotrebiteľských cien

vyžitím online údajov o cenách tovarov. Keďže štatistické inštitúcie počítajú najčastejšie indexy spotrebiteľských cien v mesačnej alebo štvrťročnej periodicite, táto prípadová štúdia je skôr vhodná na akademické účely.

Ďalej je nutné doplniť, že jeden z indexov spotrebiteľských cien, harmonizovaný index spotrebiteľských cien, sa zostavuje na základe európskeho nariadenia¹. Eurostat postupne vydáva oficiálne usmernenia týkajúce sa implementácie nových zdrojov údajov do cenovej štatistiky². Usmernenie týkajúce sa implementácie údajov získaných webscrapingom však v súčasnosti neexistuje.

V prípadovej štúdii sme analyzovali produktovú skupinu chladničky³, ktorá zahŕňa kompletnú 5-miestnu kategóriu Európskej klasifikácie individuálnej spotreby podľa účelu (ďalej ako „ECOICOP“) [12], na ktorej úrovni sa vypočítava základný index spotrebiteľských cien. Spôsob výpočtu základného indexu spotrebiteľských cien nie je stanovený európskym nariadením a je zvolený na základe analýzy danej produktovej skupiny. Agregácia základných indexov spotrebiteľských cien do vyšších úrovní kategórií ECOICOP je už stanovená európskym nariadením, pozri poznámka č. 1 pod čiarou. Za produktovú skupinu chladničky máme k dispozícii scrapingom zozbieraný časový rad cien jednotlivých typov chladničiek dostupných v online predaji na webovom portáli <https://www.heureka.sk/> od januára 2021 do januára 2022. Táto dĺžka časového radu cien už umožňuje aj výpočet multilaterálnych typov indexov spotrebiteľských cien.

Zámerom tohto článku je predstaviť teoretický rámec na agregovanie denných scrapovaných cien na mesačnú úroveň pre jednotlivé tovary, keďže Štatistický úrad SR počíta a publikuje index spotrebiteľských cien s mesačnou periodicitou. Ďalej predostrieme možnosti pre viacero spôsobov výpočtu indexu spotrebiteľských cien. Cieľom prípadovej štúdie je aplikovať tento teoretický rámec na pozorované webscrapované údaje a vyhodnotiť správanie rôznych typov indexov spotrebiteľských cien. Konečným cieľom nie je návrh najvhodnejšej metódy na výpočet indexu spotrebiteľských cien, ale prezentovanie súčasného stavu poznania v tejto oblasti v podmienkach Štatistického úradu SR.

Zlepšenie kvality výpočtu indexov spotrebiteľských cien implementáciou rôznych zdrojov údajov je kľúčovým cieľom viacerých európskych projektov [11]. Aj štúdia prezentovaná v tomto článku vznikla ako súčasť projektu dynamický cenový model (311071AA56), ktorý je spolufinancovaný Európskou úniou cez Európsky fond regionálneho rozvoja. Účelom realizácie projektu je otestovanie možností využitia iných, ako v súčasnosti využívaných údajov pre cenové štatistiky, ktoré realizuje Štatistický úrad SR. Ide hlavne o využitie moderných metód alternatívneho zberu údajov a porovnanie výsledkov týchto zisťovaní so štandardizovanými formami sledovania cien.

¹ Nariadenie Európskeho parlamentu a Rady (EÚ) 2016/792 z 11. mája 2016 o harmonizovaných indexoch spotrebiteľských cien a indexe cien nehnuteľností na bývanie a o zrušení nariadenia Rady (ES) č. 2494/95.

² Nedávno publikovaná príručka Eurostatu [2] na výpočet multilaterálnych indexov spotrebiteľských cien sa zaoberá spracovaním a použitím transakčných, tzv. scanner, údajov, ktoré sa získavajú priamo od maloobchodných predajcov a obsahujú informácie o predajoch jednotlivých tovarov. Transakčné údaje majú úplne inú štruktúru ako webscrapované údaje.

³ Do produktovej skupiny chladničky sú tiež zahrnuté chladničky s mrazničkou a mrazničky.

2. TEORETICKÉ VÝCHODISKÁ

V tejto kapitole predstavíme teoretický rámec na počiatočnú analýzu webscrapovaných údajov a pre výpočet indexu spotrebiteľských cien.

V prvej časti predstavíme teoretický koncept, ktorý by mohol byť najvhodnejší na agregovanie denných cien na mesačné. V druhej časti predstavíme teoretický rámec na výpočet nevážených bilaterálnych a multilaterálnych indexov spotrebiteľských cien.

2.1 ANALÝZA WEBS CRAPOVANÝCH ÚDAJOV

Pri súčasnom tradičnom zbere údajov pre cenovú štatistiku sa ceny tovarov zisťujú jedenkrát mesačne. Tovary, ktoré boli vybrané do spotrebného koša sú fixne definované, revidované sú raz ročne a k zmene dochádza len pri výpadku tovaru z predaja u maloobchodného predajcu.

V online prostredí sa zber cien tovarov aplikuje na všetky dostupné tovary v predaji v daný deň. Webscrapovanie normálne prebieha na dennej báze, niekedy sa môže zúžiť na určitý počet dní v týždni. Jednotlivé tovary ponúkajú rôzni online predajcovia na tej istej webovej platforme za rozdielne ceny, takže spotrebiteľ má možnosť výberu na základe preferencie ceny alebo predajcu. Týmto sa akumuluje veľké množstvo údajov („big data“).

Z dôvodu volatility cien určíme dennú cenu tovaru ako geometrický priemer jednotlivých cien, za ktoré sa tovary ponúkajú.

Na popis použijeme nasledujúcu formuláciu. Máme denné ceny tovarov $[c_i^d]_p$, kde $i = 1, \dots, N$ sú jednotlivé tovary ponúkané v daný deň $d = 0, \dots, D$, jednotlivými online predajcami $p = 1, \dots, P$. Prvý deň zberu definujeme ako bázické obdobie $d = 0$.

Geometrický priemer na výpočet dennej ceny jednotlivých tovarov je definovaný takto:

$$c_i^d = \prod_{p=1}^P ([c_i^d]_p)^{\frac{1}{P}} \quad (1)$$

Keďže index spotrebiteľských cien sa zostavuje na mesačnej báze, v ďalšom kroku musíme ceny jednotlivých tovarov agregovať na mesačnú úroveň. Podobne ako v predchádzajúcom kroku použijeme geometrický priemer denných cien jednotlivých tovarov ponúkaných v daný mesiac.

Geometrický priemer na výpočet mesačnej ceny c_i^m jednotlivých tovarov je definovaný takto:

$$c_i^m = \prod_{d=1}^D (c_i^d)^{\frac{1}{D}} \quad (2)$$

kde $m = 0, \dots, M$ sú jednotlivé mesiace a $m = 0$ je bázické obdobie.

2.2 VÝPOČET INDEXU SPOTREBITEĽSKÝCH CIEN

V tejto časti predstavíme rôzne možnosti výpočtu indexov spotrebiteľských cien, ktoré sú vhodné pre webscrapované údaje. Keďže webscrapované údaje neobsahujú tržby a množstvo predaného tovaru za daný mesiac, môžeme použiť len vzorce na výpočet indexov, ktoré sú definované ako nevážené.

Jedným zo spôsobov ako sa rozhodnúť pre vhodný vzorec indexu je požadovať, aby spĺňal určité špecifické axiómy alebo testy. Tieto testy objasňujú vlastnosti indexov, ktoré nemusia byť na prvý pohľad zrejmé. Medzi štyri základné testy, ktorými môžeme ilustrovať axiomatický prístup patria [3]: test proporcionality, test súmernosti, časovo reverzný test a test tranzitivity. Jevonsov index na rozdiel od iných indexov spotrebiteľských cien spĺňa všetky uvedené testy a je teda z axiomatického hľadiska jednoznačne index s najlepšimi vlastnosťami.

Jevonsov index možno tiež interpretovať ako geometrický priemer zmeny cien a je definovaný takto [3]:

$$I_{Jevons}^{0,m} = \prod_{i=1}^N \left(\frac{c_i^m}{c_i^0} \right)^{\frac{1}{N}} \quad (3)$$

kde 0 je bázické časové obdobie, ku ktorému porovnávame priemernú zmenu cien tovarov aktuálneho časového obdobia m . Spotrebný kôš tovarov je fixný a je definovaný v bázickom období 0.

Ďalej pokračujeme predstavením multilaterálnych indexov spotrebiteľských cien v kontexte webscrapovaných údajov. Multilaterálne indexy spotrebiteľských cien porovnávajú priemernú zmenu cien medzi viac ako dvoma časovými obdobiami súčasne, v tzv. časovom okne M . Výhodou multilaterálnych indexov spotrebiteľských cien je, že dokážu zachytiť dynamiku predaja tovarov v časovom okne, napríklad vstupy a výstupy jednotlivých tovarov, chýbajúce ceny v niektorých mesiacoch. Ich nevýhodou je komplexnosť výpočtu a tým aj nejednoznačná interpretácia, ktorá môže byť podstatne rozdielna v závislosti od použitého typu multilaterálneho indexu.

Multilaterálna metóda GEKS používa bilaterálne indexy spotrebiteľských cien vypočítané medzi dvoma časovými obdobiami v definovanom časovom okne, ako tzv. základné elementy, ktoré vstupujú do výpočtu. Metóda GEKS vyžaduje výber vhodného bilaterálneho indexu spotrebiteľských cien. Pri použití Jevonsovho indexu, multilaterálny GEKS-Jevonsov index je definovaný nasledovne [2]:

$$I_{GEKS-Jevons}^{0,m} = \prod_{\substack{l=0 \\ l \neq m}}^M (I_{Jevons}^{0,l} \times I_{Jevons}^{l,m})^{\frac{1}{M+1}} \quad (4)$$

kde $m = 0, \dots, M$ sú časové obdobia.

Bilaterálne Jevonsove indexy sú vypočítané medzi časovými obdobiami l a m a spojené metódou GEKS, geometrickým priemerom, s $m = 0$ ako bázickým časovým obdobím. Spotrebný kôš tovarov medzi porovnávanými časovými obdobiami sa musí zhodovať. Detaily výpočtu metódy GEKS sú popísané v [2]. Nevýhodou metódy GEKS

je, že chýbajúce ceny tovarov porovnávaných v jednotlivých časových obdobiach, je nutné imputovať, ak chceme, aby ich zmena bola reflektovaná vo výpočte indexu spotrebiteľských cien.

Priamej imputácii pri chýbajúcich cenách tovarov za niektoré časové obdobia je možné predísť použitím multilaterálnej metódy, ktorá je definovaná vo forme štatistickej regresie, tzv. modelu fixných efektov. Log-lineárny štatistický regresný model, tzv. time-product dummy (TPD) model, je definovaný takto [6]:

$$\ln c_i^m = \partial^0 + \sum_{m=1}^M \partial^m F_i^m + \sum_{i=1}^{N-1} \gamma_i F_i + \varepsilon_i^t \quad (5)$$

kde F_i je tzv. binárna premenná, ktorá nadobúda hodnotu 1 ak ide o cenu tovaru i a 0 v opačnom prípade, ε_i^t sú náhodné chyby regresie, ktoré majú normálne rozdelenie so strednou hodnotou 0 a konštantnou štandardnou odchýlkou σ a F_i^m je tzv. binárna časová premenná, ktorá nadobúda hodnotu 1 ak cena tovaru spadá do časového obdobia m a 0 v opačnom prípade. Regresné parametre $\partial^0, \partial^m, \gamma_i$ sa vypočítajú odhadom použitím metódy najmenších štvorcov.

Rovnica (5) patrí do tzv. regresných modelov s fixnými efektami, kde fixné efekty sú definované v parametri γ_i pre každú vzorku i v regresii, v našom prípade tovar. Podrobný popis týchto typov modelov a odhadov ich parametrov je uvedený v [1].

F_i^m sa vynechá pre $m = 0$, takže máme výpočtom odhadnuté parametre $\hat{\partial}^m$ za časové obdobia $m = 1, \dots, M$. TPD index spotrebiteľských cien sa vypočíta takto [5]:

$$I_{TPD}^{0,m} = \exp(\hat{\partial}^m) = \frac{\prod_{i \in S^m} (c_i^m)^{\frac{1}{N^m}}}{\prod_{i \in S^0} (c_i^0)^{\frac{1}{N^0}}} \exp \left[\sum_{i=1}^N (\hat{\gamma}_i^0 - \hat{\gamma}_i^m) \right] \quad (6)$$

kde $m = 1, \dots, M$. sú časové obdobia a S^m a S^0 sú skupiny tovarov v časových obdobiach 0 a m . Ďalej $\hat{\gamma}_i^0 = \sum_{i \in S^0} \hat{\gamma}_i / N^0$ a $\hat{\gamma}_i^m = \sum_{i \in S^m} \hat{\gamma}_i / N^m$ sú priemery odhadovaných parametrov fixných efektov $\hat{\gamma}_i$ pre časové obdobie 0 a m . Ak sa skupiny tovarov medzi S^m a S^0 zhodujú, rovnica (6) sa zjednoduší na bilaterálny Jevonsov index definovaný v rovnici (3). Podrobné odvodenie rovnice (6) je uvedené vo vedeckej štúdii [6].

Ďalšou možnou regresnou metódou na výpočet indexu spotrebiteľských cien je tzv. hedonická regresia, pri ktorej sa použijú charakteristické vlastnosti jednotlivých tovarov ako nezávislé premenné vysvetľujúce logaritmus ceny daného tovaru. Hedonický regresný model definujeme takto [5]:

$$\ln c_i^m = \partial^0 + \sum_{k=1}^K \beta_k Z_{ik} + \varepsilon_i^t \quad (7)$$

kde Z_{ik} je matica charakteristických premenných súvisiacich s tovarom i a $k = 1, \dots, K$ je počet charakteristických premenných s korešpondujúcim parametrami β_k . Ostatné premenné a parametre vrátane ich odhadov sú definované ako v rovnici (5).

Hedonická regresia v rovnici (7) má využitie aj na priame imputácie chýbajúcich cien tovarov. Ak cena daného tovaru chýba v aktuálnom mesiaci, ale bola pozorovaná v predchádzajúcom mesiaci, odhadnutý hedonický model sa môže použiť na predikciu ceny v aktuálnom mesiaci.

Pridaním binárnej časovej premennej F_i^m do rovnice (7) sa hedonický index spotrebiteľských cien odvodí podobne ako v prípade rovníc (5) a (6).

Nevýhodou multilaterálnych indexov spotrebiteľských cien je, že pri výpočte indexu za nadchádzajúce časové obdobie $M + 1$ sa predchádzajúce indexy v časovom okne $M = 1, \dots, M + 1$ znova prepočítajú, čím sa zmenia. Zmena indexov za predchádzajúce časové obdobia, v dôsledku prepočtu, je nežiaduca. Revízií predchádzajúcich indexov spotrebiteľských cien je možné predísť použitím tzv. splicingovej (spájacej) metódy, keď sa index spotrebiteľských cien vypočítaný za nové časové obdobie $M + 1$ naviaže na predchádzajúce indexy spotrebiteľských cien. Keďže téma ohľadom aplikácie rôznych splicingových metód je veľmi rozsiahla, nie je súčasťou tohto článku.

3. VÝSLEDKY

V tejto kapitole aplikujeme uvedený teoretický rámec na webscrapované údaje za produktovú skupinu chladničky⁴, kde typy chladničiek reprezentujú jednotlivé tovary v skupine, ktoré boli scrapované každý tretí deň za časové obdobie od 1. januára 2021 do 31. januára 2022 z porovnávacieho webového portálu <https://www.heureka.sk/>.

V nasledovnej tabuľke uvádzame sumárne informácie webscrapovaných údajov za uvedené časové obdobie.

Tabuľka č. 1: Sumárne informácie webscrapovaných údajov – produktová skupina chladničky

Produktová skupina	Počet cien tovarov	Počet tovarov (jednotlivých typov chladničiek)	Počet online predajcov	Počet charakteristických premenných
chladničky	1 468 279	3 332	121	10

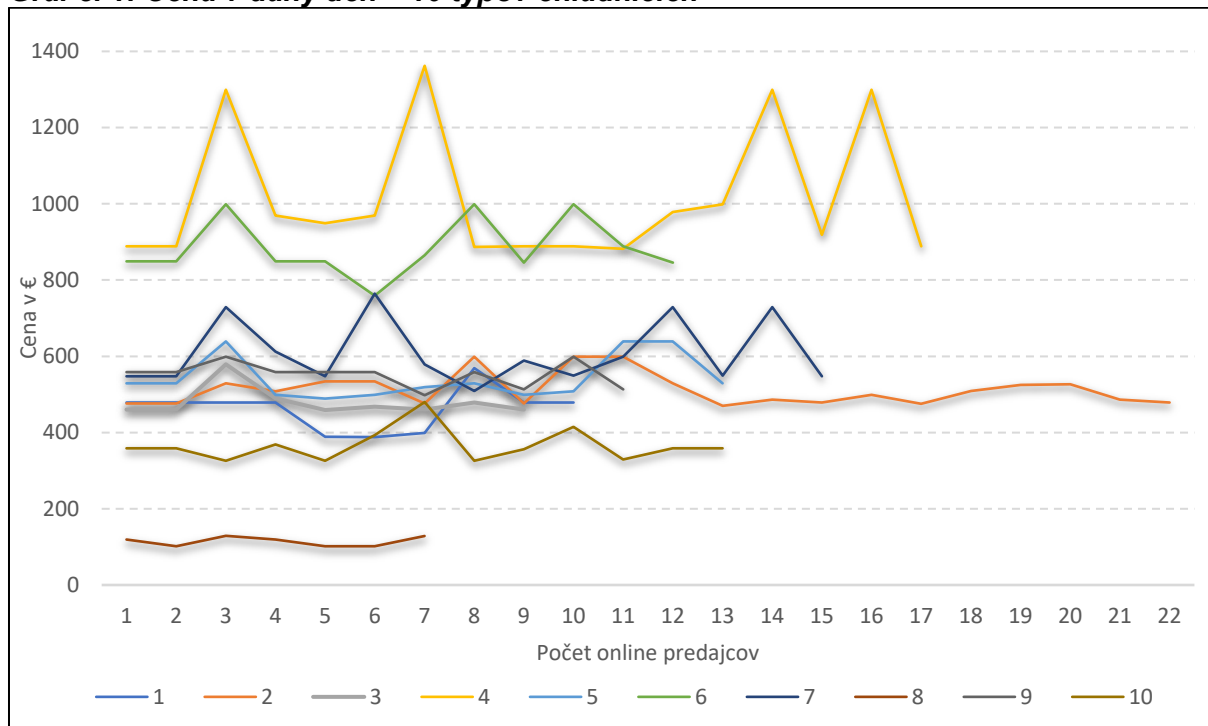
Zdroj údajov: ŠÚ SR, výpočty autorov

V tabuľke č. 1 vidíme, že webscraping umožnil za dané sledované obdobie na dennej báze získať ceny pre 3 332 jednotlivých typov chladničiek, ktoré ponúkalo 121 online predajcov. Týmto bol výrazne rozšírený spotrebný kôš pre túto produktovú skupinu a na rozdiel od tradičného spôsobu zberu, pri ktorom je zaznamenávaná cena tovaru zvyčajne len raz mesačne, v údajoch z webscrapingu pozorujeme ceny tovarov s dennou alebo viacdennou periodicitou.

V nasledujúcom kroku aplikujeme agregovanie cien jednotlivých tovarov na mesačnú úroveň použitím rovníc (1) a (2). Na ukážku, že použitie rôznych priemerov pri agregovaní cien môže viesť k rozdielnym mesačným cenám vyberieme vzorku 10 tovarov, t. j. typov chladničiek, a zobrazíme ich ceny, ktoré v daný deň ponúkali rozdielni online predajcovia.

⁴ Do produktovej skupiny chladničky sú zahrnuté aj chladničky s mrazničkou a mrazničky.

Graf č. 1: Cena v daný deň – 10 typov chladničiek

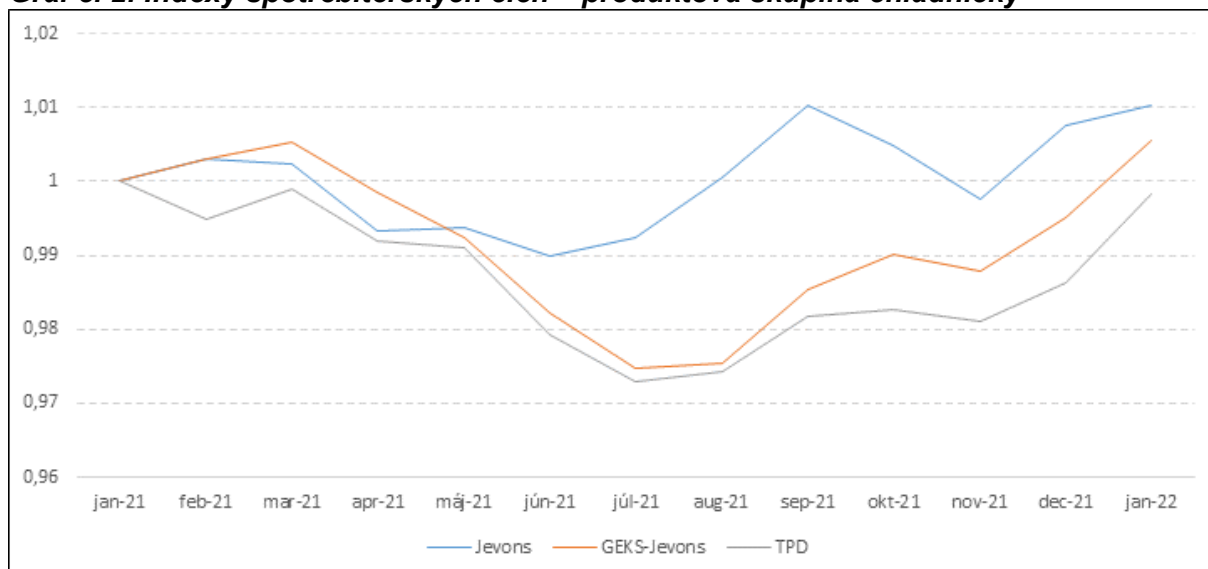


Zdroj údajov: ŠÚ SR, výpočty autorov

Graf č. 1 ukazuje, že cena tovaru od rôznych predajcov, v daný špecifický deň, môže byť podstatne rozdielna. Napríklad, tovar 2 ponúka 22 predajcov a tovar 4 má najväčší rozdiel v cenách v daný deň. V našej štúdii pri agregovaní cien používame geometrické priemery, rovnice (1) a (2), keďže tieto využijú všetky dostupné denné ceny vo výpočte a zredukujú efekt extrémnych hodnôt na výslednú dennú a mesačnú cenu.

V ďalšom kroku vypočítame Jevonsove, GEKS-Jevonsove a TPD indexy. Graf č. 2 zobrazuje vývoj týchto indexov spotrebiteľských cien za produktovú skupinu chladničky a za obdobie od januára 2021 do januára 2022, v ktorom január 2021 predstavuje základné obdobie.

Graf č. 2: Indexy spotrebiteľských cien – produktová skupina chladničky



Zdroj údajov: ŠÚ SR, výpočty autorov

Graf č. 2 ilustruje, že multilaterálne indexy spotrebiteľských cien za produktovú skupinu chladničky majú podobný vývoj v čase na rozdiel od bilaterálneho indexu spotrebiteľských cien. Hodnota multilaterálneho indexu spotrebiteľských cien v danom období je ovplyvnená vývojom cien vo všetkých obdobiach definovaného časového okna na rozdiel od bilaterálneho indexu spotrebiteľských cien, keď sa porovnávajú len ceny dvoch časových období.

Výpočet hedonického indexu si vyžaduje dôkladnú analýzu charakteristických parametrov jednotlivých tovarov. Je potrebné tzv. čistenie údajov, t. j. odstránenie nesprávnych alebo doplnenie chýbajúcich údajov, štatistická analýza významnosti jednotlivých charakteristických premenných v súvislosti s cenou tovaru, korelačná analýza a ďalšie štatistické overenia správnosti výpočtu hedonickej regresie. Táto analýza bude súčasťou inej štúdie, ktorá sa bude zaoberať výpočtom hedonického indexu.

V rámci prezentovanej štúdie, autori dôkladne analyzovali webscrapované údaje cien v produktovej skupine chladničky. Štúdia bude slúžiť ako základ analýzy ďalších produktových skupín spotrebného koša. Multilaterálne metódy na výpočet indexov spotrebiteľských cien sú vo všeobecnosti vhodnejšie pre rýchloobrátkové tovary, keďže vo výpočte zahŕňajú viac časových období súčasne. V prípadovej štúdii, pri porovnaní s bilaterálnym indexom spotrebiteľských cien, ktorý sa štandardne používa, ak sa ceny tovarov získavajú tradičným spôsobom, hodnota oboch prezentovaných multilaterálnych indexov spotrebiteľských cien mala tendenciu byť pod jeho úrovňou. Rozhodnúť, ktorá z indexových metód je najvhodnejšia pre webscrapované údaje je náročná úloha, ktorá si vyžaduje množstvo ďalších analýz týkajúcich sa napríklad rozšírenia časového okna⁵, spájania indexov tzv. splicingovými metódami a tiež analýzy ďalších produktových skupín spotrebného koša.

4. ZÁVER

Cieľom tohto článku bolo predstaviť teoretický rámec na implementáciu webscrapovaných údajov do produkcie cenových štatistík.

Prípadová štúdia prezentuje postup spracovania denných webscrapovaných údajov na mesačnú úroveň. Volatilita v denných cenách tovarov je významná, čo môže viesť k rozdielnym mesačným cenám, ak sa použijú rôzne typy priemerov pri agregovaní. V našej štúdii sme použili geometrický priemer, ktorý zabezpečí, že všetky ceny ponúkaných tovarov sú súčasťou výpočtu a vplyv extrémnych hodnôt je zredukovaný.

Porovnanie rôznych metód na výpočet indexov spotrebiteľských cien v prípadovej štúdii na produktovej skupine chladničky ukázalo, že výber metódy má podstatný vplyv na jeho vývoj v čase. Multilaterálne indexy spotrebiteľských cien zachytia dynamiku v zmene cien tovarov, keďže porovnávajú zmenu cien tovarov medzi viacerými časovými obdobiami súčasne a do výpočtu zahrnú ceny tovarov, ktoré sa nevyskytujú v každom časovom období. Ich nevýhodou je komplexnosť výpočtu a nejednoznačná interpretácia pre bežných používateľov.

Z prípadovej štúdie pre ďalšie pokračovanie projektu vyplynulo niekoľko dôležitých záverov. Napríklad, prvotný spôsob spracovania údajov a ich následné agregovanie

⁵ V [2], strana 42, je navrhnuté zobrať do úvahy časové okno s 25 časovými obdobiami.

môže viesť k významne rozdielnym mesačným cenám jednotlivých tovarov. Taktiež nie je jednoznačné, na základe akých kritérií je možné vybrať najvhodnejšiu multilaterálnu metódu na výpočet indexu spotrebiteľských cien.

Projekt bude pokračovať dôkladnou analýzou údajov týkajúcich sa charakteristických parametrov jednotlivých tovarov, ktorá je potrebná na výpočet hedonického indexu. Hedonický regresný model bude v budúcnosti možné využiť aj pri priamej imputácii chýbajúcich cien tovarov v niektorých mesiacoch v prípade multilaterálnej metódy GEKS.

Ďalšou dôležitou problematikou skúmania je analýza spôsobov viazania indexov za nové časové obdobie T+1, tzv. splicingovej metódy.

LITERATÚRA

- [1] ALLISON, P. D.: Fixed Effects Regression Methods for Longitudinal Data Using SAS, 2009, SAS Institute Inc., Cary, NC, USA. ISBN 978-1-59047-568-3.
- [2] Eurostat, Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, Manuals and Guidelines (2022), Luxembourg: Publication Office of the European Union. ISBN 978-92-76-44354-4.
- [3] ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). Consumer Price Index Manual: Theory and Practice. ILO Publications, Geneva.
- [4] DE HAAN J.: Hedonic Prices Indexes: A Comparison of Imputation, Time Dummy and 'Re-Pricing' Methods, Jahrbücher f. Nationalökonomie u. Statistik, Lucius & Lucius, Stuttgart, 2010, Bd. (Vol.) 230/6, 772-791.
- [5] DE HAAN J. – KRSINICH F.: Scanner Data and the Treatment of Quality Change in Rolling Year GEKS Price Indexes, 2012, Paper presented at the eleventh Economic Measurement Group Workshop, 21-23 November 2012, Sydney, Australia
- [6] DE HAAN J. – HENDRIKS R.: Online data, fixed effects the construction of high-frequency price indexes, Paper presented at the Economic Measurement Group Workshop, 2013, 28-29 November 2013, Sydney, Australia.
- [7] ELTETŐ, O. – KÖVES, P.: On a Problem of Index Number Computation Relating to International Comparisons, 1964, Statisztikai Szemle 42, s. 507 – 518 (originál v maďarčine).
- [8] GINI, C.: On the Circular Test of Index Numbers, 1931, Metron 9, s. 3 – 24.
- [9] GLASER-OPITZOVÁ H.: Nové zdroje údajov pre cenovú štatistiku a metódy ich spracovania. In: Slovenská štatistika a demografia, 2019, roč. 29, č.4, str. 49 – 66.
- [10] IVANCIC L. – DIEWERT W. E. – FOX K. J.: Scanner Data, Time Aggregation and the Construction of Price Indexes, 2011, In: Journal of Econometrics 161, s. 24 – 35.
- [11] POLIDORO F. – GIONNINI R. – LO CONTE R. – MOSCA S. – ROSSETTI F.: Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation, 2015, In: Statistical Journal of the IAOS 31 (2015) 165-176, IOS Press.
- [12] RAMON - Reference and Management of Nomenclatures: [Europa - RAMON - Classification Detail List](#) [cit. 2022-12-05].
- [13] SILVER, M. – HERAVI, S.: The Measurement of Quality-Adjusted Price Changes. 2003, Pp. 277 – 317. In: Shapiro, M. – Feenstra, R. (eds.): Scanner Data and Price Indexes, NBER, Studies in Income and Wealth, vol. 61. Chicago: University of Chicago Press.

[14] SZULC, B.: Indices for Multiregional Comparisons, 1964, In: Przeglad Statystyczny 3, s. 239 – 254 (originál v poľštine).

[15] Tovarová skupina chladničky: <https://lednice.heureka.sk/> [cit. 2022-12-05].

[16] TRIPLETT, J.E.: Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes. 2006, Organization for Economic Co-operation and Development, Paris.

RESUMÉ

V súčasnosti sa ceny tovarov zberajú tradičným spôsobom, t. j. v kamenných predajniach, vyhľadávaním na internetových stránkach, alebo štatistickým zisťovaním. Webscraping je zber cien tovarov automatickým spôsobom, t. j. sťahovaním údajov z internetu, pomocou tzv. robota, ktorý je vyvinutý v niektorom z programovacích jazykov. Webscrapované údaje umožnia porovnať vývoj cien v časovom období pre všetky dostupné tovary v online predaji. Príspevok je zameraný na analýzu a spracovanie webscrapovaných údajov, ktoré sa získavajú v dennej periodicite. Štatistické inštitúcie zostavujú indexy spotrebiteľských cien s mesačnou periodicitou takže je potrebná agregácia denných webscrapovaných cien na mesačné. Tento krok je veľmi dôležitý, keďže použitím rôznych typov priemerov môže dôjsť k určeniu rozdielnych mesačných cien tovarov. Ďalej sa článok zameriava na prezentovanie vybraných metód na výpočet indexov spotrebiteľských cien. Webscrapované údaje neobsahujú informácie o predaji jednotlivých tovarov, ktoré by zahŕňali ich predané množstvá a tržby v určitom časovom období, čo umožňuje aplikovať len nevážené typy indexov. Prípadová štúdia prezentuje aplikáciu vybraných bilaterálnych a multilaterálnych metód na výpočet indexov spotrebiteľských cien v produktovej skupine chladničky. Multilaterálne indexy spotrebiteľských cien na rozdiel od bilaterálnych dokážu zachytiť dynamiku predaja jednotlivých tovarov tým, že do výpočtu zahrnú viac ako dve časové obdobia súčasne do výpočtu, čo je dôležité pre rýchloobrátkové tovary.

Prípadová štúdia, ktorá bola vypracovaná autormi tohto článku je súčasťou projektu. Dynamický cenový model. Tento projekt sa zaoberá modernizáciou cenových štatistik z pohľadu zdrojov v Štatistickom úrade SR. Webscrapovanie je technologicky náročný zber informácií priamo z internetu, ktorý si vyžaduje kvalifikované ľudské zdroje a pokročilé informačno-technologické systémy. Štatistický úrad SR momentálne uskutočňuje teoretickú a empirickú štúdiu použitia webscrapovaných údajov v cenových štatistikách.

RESUME

Currently, the prices of goods are collected in a traditional way, i.e. in the brick and mortar stores, by searching on the internet, or through statistical surveys. Web-scraping is a collection of prices of goods in an automated way, i.e. by downloading data from the internet using the so-called robot that is developed in a one of the programming languages. The web-scraped data enables a comparison of price changes over time periods for all available goods in online sales. The paper focuses on the analysis and processing of web-scraped data, which are obtained on a daily basis. Statistical institutes compile consumer price indexes on a monthly basis so the aggregation of daily web-scraped prices is necessary on a monthly level. This step is very important, since using different types of averages can result in a determination of different monthly prices of goods. Furthermore, the paper focuses on a presentation of different methods for estimating the consumer price index. The web-scraped data does not contain information on the sale of individual goods, which would include their

sold quantities and turnover in a given time period, which enables to apply only unweighted consumer price indexes. The case study presents the application of selected bilateral and multilateral methods for the calculation of consumer price indexes for the product category of refrigerators. Multilateral consumer price indexes, unlike bilateral ones are able to capture the sales dynamics of individual products by including more than two time periods simultaneously in its calculation, which is important for fast-moving goods.

This study which was developed by the authors of this paper is part of the dynamic pricing model project. This project deals with the modernization of price statistics from the perspective of different data sources in the Statistical Office of the Slovak Republic. Web-scraping is a technologically demanding collection of information directly from the internet, requiring highly qualified human resources and advanced information technology systems. The Statistical Office of the Slovak Republic is currently conducting a theoretical and empirical study of the use of web-scraped data in price statistics.

PROFESIJNÝ ŽIVOTOPIS

Peter Knížat, MSc, je externým študentom doktorandského štúdia na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave. Pracuje ako dátový analytik v sekcii všeobecnej metodiky, registrov a koordinácie národného štatistického systému Štatistického úradu SR, kde je zodpovedný za návrh štatistickej metodiky v cenových štatistikách s využitím webscrapovaných údajov.

Ing. Helena Glaser-Opitzová je generálna riaditeľka sekcie všeobecnej metodiky, registrov a koordinácie národného štatistického systému Štatistického úradu SR a členka riaditeľskej skupiny Eurostatu pre metodológiu (DIME), ktorá poskytuje poradenstvo Európskemu štatistickému výboru (ESSC) v strategických otázkach. Riadila a podieľala sa na mnohých modernizačných aktivitách úradu. V súčasnosti riadi interný projekt úradu zameraný na modernizáciu cenových štatistík.

KONTAKT

peter.knizat@statistics.sk

helena.glaser-opitzova@statistics.sk

Informácia/Information

MANAŽMENT INOVÁCIÍ V EURÓPSKOM ŠTATISTICKOM SYSTÉME: OD EXPERIMENTOVANIA K PRODUKCII (KONFERENCIA DGINS)

MANAGEMENT OF INNOVATION IN THE EUROPEAN STATISTICAL SYSTEM: FROM EXPERIMENTATION TO PRODUCTION (DGINS CONFERENCE)

V historických priestoroch hotela Bristol sa v hlavnom meste Nórska, v Osle v dňoch 25. – 26. 10. 2022 uskutočnilo **107. zasadnutie konferencie DGINS**. DGINS je konferencia predsedov národných štatistických úradov a ako ročné fórum sa zameriava na strategické problémy a súčasný rozvoj Európskeho štatistického systému. Okrem zástupcov z radov štatistikov sa na konferencii pravidelne zúčastňujú predstavitelia Európskej komisie, Centrálnej európskej banky, OECD a OSN.

Téma s názvom **Manažment inovácií v Európskom štatistickom systéme: od experimentovania k produkcii** odrážala obsah aktuálnych diskusií v rámci Európskeho štatistického systému. Cieľom (ktorý sa maximálne podarilo naplniť) bolo reflektovať na konkrétne výstupy inovácií v štatistike – na úrovni produktov a procesov; diskutovať o tom, ako vytvoriť vhodné prostredie pre inovácie v štatistických úradoch; identifikovať rozhodujúce prvky pre inovácie, ako aj podeliť sa o skúsenosti pri prechode od experimentálnej k oficiálnej štatistike.



Zdroj: Studio Elisenberg, https://www.efta.int/Statistics/news/DGNIS_2022

Za Štatistický úrad SR sa na rokovaní zúčastnili predseda úradu Peter Peťko a generálne riaditeľky sekcií Helena Glaser-Opitzová a Ľudmila Ivančíková. Počas dvojdňového rokovania odznelo celkovo 15 príspevkov. Je vysoko potrebné hodnotiť najmä úroveň panelových diskusií k jednotlivým blokom, ktoré boli transparentné a sumarizujúce.

Konferenciu otvorili svojimi príhovormi predseda Nórskeho štatistického úradu Geir Axelsen a generálna riaditeľka Eurostatu Mariana Kotzeva. Nasledoval hlavný príspevok bloku na tému *Inovácia a digitalizácia – výzvy a skúsenosti v organizáciách verejného sektora*, ktorý predniesol H. CH Holte, generálny riaditeľ Nórskeho úradu práce a sociálneho zabezpečenia.

V rámci prvej sekcie s názvom *Inovácie v oficiálnych štatistikách: skúsenosti a poučenia* zástupcovia štyroch krajín prezentovali *použitie súkromných big data na zlepšenie kvality oficiálnych štatistík* (Nórsko), *použitie umelej inteligencie pri odhade spotrebných výdavkov v domácnosti* (Maďarsko), *použitie machine learningu na automatické odkódovanie ručne písaných numerických informácií v papierovom dotazníku pre populačný cenzus 2021* (Luxembursko) a *predstavenie nových dátových zdrojov a rozvoj novej metodológie pre HICP* (Slovinsko).

Druhej sekcii predchádzala vysoko odborná prezentácia, v rámci ktorej C. Peyrable, viceprezidentka firmy Gartner, vysvetlila podstatu a východiská *inovačného manažmentu*. Sekcia sa zamerala na konkrétne príklady inovačného manažmentu v oficiálnych štatistikách. S. Dufour z kanadského štatistického úradu hovorila o *udržateľnosti úradu prostredníctvom inovácií*, T. Burg z Rakúska prezentoval, *ako použiť strategickú kontrolu na podporu inovácií* a P. Sillard z INSEE nás oboznámil s *modernizáciou zisťovaní v domácnostiach so zameraním na praktickú aplikáciu európskeho stanoviska 2021 k zisťovaniu so zmiešaným spôsobom zberu*. Prezentácia predsedníčky Lotyšského štatistického úradu sa zamerala na *riadenie nepretržitých inovácií, od stratégie k lepšej výkonnosti na zabezpečenie vysokých spoločenských požiadaviek*. Posledná prezentácia sekcie s názvom *Transformácia štatistík: od štatistického úradu k multifunkčnému analytickému dátovému centru* bola zo štatistiky z Litvy.

Tretia sekcia *Od experimentovania k produkcii* zahŕňala príspevok *Premena rozvojových prác na oficiálne štatistiky: praktický prístup* z Poľska, *Poučenia zo skúseností štatistického úradu v rámci ESSnet so smart štatistikou* z Belgicka, *Využitie cloudu a technológie machine learningu na premostenie medzi inováciami a produkciou*, prezentovaný švédskym expertom a nakoniec príspevok Nemeckého federálneho štatistického úradu: *Experimentálne štatistiky: od analýzy k oficiálnemu produktu*.

Panelová diskusia na tému *Inovácia sa len tak neprihodí – stratégia ako dosiahnuť, aby inovácia fungovala* uzatvorila túto na naše podmienky veľmi inšpiratívnu konferenciu.

Hlavné závery a odporúčania konferencie možno zhrnúť do nasledujúcich viet: Inovácia nie je alternatíva. V podmienkach štatistických úradov ju možno chápať ako experiment implementovaný do produkcie. Dôležitým sa ukazuje vzťah s kľúčovými partnermi, najmä vlastníkmí a správcami údajov. Tento musí byť založený na cielenej komunikácii. Okrem partnerstva ďalšími kľúčovými prvkami sú ľudské zdroje, sieťovanie znalostí a technický potenciál. K inováciám je potrebné pristupovať ako k projektu, čiže súčasťou rozhodnutia o zavedení inovácií musia byť aj analýzy nákladov. Celkovo možno povedať, že pre Európsky štatistický systém je len jedna cesta a tou je inovácia. Musíme inovovať rýchlo, interne a externe a adresne toto rozhodnutie odkomunikovať, optimalizovať procesy a vypracovať metodologický manuál na spoločnú stratégiu inovácií. Len tak dokážeme reagovať na zvýšené nároky používateľov na kvalitu štatistických produktov.

PhDr. Ľudmila IVANČÍKOVÁ, PhD.

Autorka je generálna riaditeľka sekcie sociálnych štatistík a demografie Štatistického úradu SR.

Informácia/Information

NITRIANSKE ŠTATISTICKÉ DNI 2022

STATISTICAL DAYS IN NITRA 2022

Konferencia **Nitrianske štatistické dni** si za viac ako desať rokov svojej existencie získala dobré meno medzi slovenskými matematikmi a štatistikmi. Od roku 2015 sa koná raz za dva roky. Rok 2022 je však výnimočný v tom, že pôvodný koncept dvojdňovej konferencie konanej prezenčnou formou sa tím organizátorov rozhodol zmeniť. Konferenciu poňali ako sériu zaujímavých prednášok od odborníkov z rôznych oblastí, ako sú teória vyučovania matematiky, pravdepodobnosti a štatistiky, Sčítanie obyvateľov, domov a bytov 2021, tentokrát nielen prezenčnou, ale aj online formou. Niektoré z príspevkov boli prezentované hybridnou formou (online aj prezenčne) a iné zas iba online. Podujatie bolo naplánované vo vybraných termínoch od 14. 11. 2022 do 15. 12. 2022 a prebiehalo formou prednášok alebo workshopov.



Zdroj: fotografie Univerzity Konštantína Filozofa v Nitre

Záujemcovia o účasť na konferencii si mohli prostredníctvom google formulára vybrať prednášky, na ktorých sa chcú zúčastniť online. V deň konania prednášky im bol zaslaný link, kde sa mohli pripojiť a zúčastniť sa na prednáške, či workshopoch. Práve vďaka tejto možnosti pripojenia sa online a na konkrétny čas (1 – 2 hodiny) sa na konferencii zúčastnilo omnoho viac zástupcov akademickej sféry nielen z radov vyučujúcich, ale aj študentov, ako aj odborníkov z iných oblastí nielen zo Slovenska, ale z celého sveta.

Podujatie, ktoré už tradične organizuje **Katedra matematiky Fakulty prírodných vied a informatiky Univerzity Konštantína Filozofa v Nitre** (FPVal UKF) v spolupráci so **Slovenskou štatistickou a demografickou spoločnosťou** a **Štatistickým úradom Slovenskej republiky** v Nitre, sa začalo v pondelok 14. 11. 2022 na katedre matematiky, kde prezenčne boli prítomní a otvorili konferenciu dekan FPVal UKF v Nitre prof. RNDr. František Petrovič, PhD., vedúca Katedry matematiky FPVal UKF v Nitre doc. PaedDr. Gabriela Pavlovičová, PhD., hlavná organizátorka NŠD 2022 doc. PaedDr. Janka Medová, PhD., ako aj zástupkyňa Pracoviska Štatistického úradu SR v Nitre Mgr. Mária Rybanská. On-line privítala účastníkov

za Slovenskú štatistickú a demografickú spoločnosť jej predsedníčka doc. Ing. Iveta Stankovičová, PhD.

Úvodnú prednášku pod názvom *Statistical Investigation Process with Image-Based Data* predniesla hybridnou formou (prezenčne na Katedre matematiky FPVal UKF v Nitre a súčasne online) viceprezidentka Medzinárodnej asociácie pre vyučovanie štatistiky (International Association of Statistics Education) a podpredsedníčkou pracovnej skupiny o vyučovaní pravdepodobnosti a štatistiky na konferenciách CERME12 a CERME13. doc. Sibel Kazak z Middle East Technical University v Ankare v Turecku. Okrem hostí prítomných na katedre sa na streame prednášky zúčastnilo takmer 120 online účastníkov.

Okrem prvej prednášky boli hybridnou formou realizované aj ďalšie dva príspevky. Z radov akademikov UKF v Nitre ich prezentovali RNDr. Ľubomír Rybanský, PhD., na tému *Teória odpovedí na položku a jej aplikácie (nielen) vo vyučovaní matematiky* a doc. Mgr. Ján Mačutek, PhD., na tému *Aplikácie matematickej štatistiky v lingvistike*.

Zvyšné štyri prednášky a workshopy prebehli on-line prostredníctvom MS Teams alebo Zoomu. UKF v Nitre tu reprezentoval prof. RNDr. Anatolij Dvurečenskij, DrSc., s prednáškou s názvom *Teória pravdepodobnosti: od hazardných hier ku kvantovej mechanike*. Dva workshopy určené prioritne pre študentov matematiky a štatistiky viedla Dr. Edyta Nowinska z Univerzity v Osnabrücker v Nemecku. Oba workshopy prebiehali v angličtine a venovali sa téme z oblasti teórie vyučovania matematiky: *Metacognitive support in class discussion – theoretical considerations on how can teachers support students' metacognition and methodological approaches to analyse metacognitive support in class discussions and Quantitative approach*. Online workshop konaný dňa 1. 12. 2022, zameraný prioritne na prezentáciu a šírenie údajov zo Sčítania obyvateľov, domov a bytov 2021 s názvom *Čo nám hovoria dáta zo SODB 2021*, viedli Mgr. Lucia Vanišová (vedúca oddelenia sčítania obyvateľov, domov a bytov a prierezových štatistík) a Jasmína Stauder (hovorkyňa pre SODB 2021) zo Štatistického úradu SR.

Dvojdňová konferencia sa konala v minulosti vždy za osobnej prítomnosti účastníkov. Prvý deň prebiehal formou prednášok na Katedre matematiky FPVal UKF v Nitre a druhý deň na Pracovisku Štatistického úradu SR v Nitre, kde bol počas workshopu vytvorený priestor na oboznámenie sa s novinkami v oblasti šírenia štatistických informácií, nových digitálnych produktov. Súčasťou boli aj ukážky výberov z databáz Štatistického úradu SR, ako aj diskusia, otázky a podnety zo strany akademikov, študentov a ostatných účastníkov. Keďže posledné roky boli chudobnejšie na osobné stretnutia a organizátori nechceli prísť o možnosť opäť zrealizovať konferenciu, rozhodli sa využiť tento rok celkom nový, hybridný koncept podujatia, aby tak oslovili väčšie fórum, ktoré by nebolo obmedzené časom a priestorom stretnutia. Toto sa ukázalo v konečnom dôsledku ako veľmi úspešné a dobré rozhodnutie: tento ročník Nitrianskych štatistických dní mal doteraz najvyšší počet účastníkov (spolu 657 prihlásených, t. j. priemerne 130 účastníkov na 1 prednášku) a to nielen zo Slovenska a Čiech, ale z viacerých krajín sveta.

Mgr. Mária RYBANSKÁ

Autorka pracuje na odbore informatiky, registrov a informačných služieb Pracoviska Štatistického úradu SR v Nitre.

PRIPRAVUJEME/COMING SOON

**Číslo 2/2023 zamerané na prezentáciu výsledkov
Sčítania obyvateľov, domov a bytov 2021.**

**Issue 2/2023 focused on the presentation of the results
of the 2021 Population and Housing Census.**

* * *

**ONLINE VERZIA ČÍSLA 1/2023 SLOVENSKEJ ŠTATISTIKY A DEMOGRAFIE JE
VEREJNE DOSTUPNÁ na internetovej stránke ssad.statistics.sk od 15. JANUÁRA
2023.**

**THE ONLINE VERSION OF THE JOURNAL SLOVAK STATISTICS AND
DEMOGRAPHY No 1 (2023) IS PUBLICLY BE AVAILABLE at the website
ssad.statistics.sk from JANUARY 15, 2023.**

INFORMÁCIE PRE PRISPIEVATEĽOV

Príspevky prijímame v slovenskom, v českom a v anglickom jazyku. Musia rešpektovať odborné zameranie časopisu a jeho vedecký charakter. Zaslaný príspevok nesmie byť v recenznom konaní v inom časopise, ani uverejnený v odbornej a inej tlači.

Príspevky zasielajte v elektronickej forme vo formáte MS Word alebo Open Office, typ písma Arial, veľkosť 12, riadkovanie 1. Nad titulkom treba uviesť meno autora a jeho pracovisko.

Súčasťou príspevku je abstrakt (základný popis cieľa a spôsobu spracovania faktov v rozsahu do 100 slov), kľúčové slová (maximálne 5), resumé (stručné zhrnutie obsahu článku s dôrazom na jeho prínos a najvýznamnejšie závery v rozsahu do 500 slov), profesijný životopis (v rozsahu do 120 slov) a kontakt (e-mailová adresa autora). Názov článku, abstrakt, kľúčové slová a resumé poskytnite autor aj v anglickom jazyku. Zoznam použitej literatúry v abecednom poradí s úplnými bibliografickými údajmi sa uvádza na konci článku. Odkazy na literatúru sa uvádzajú v texte číslami v hranatých zátvorkách. Poznámky s poradovým číslom sú umiestnené pod čiarou na príslušnej strane textu, ku ktorému sa vzťahujú. Podrobnejšie pokyny nájdete autori na ssad.statistics.sk.

Maximálny rozsah vedeckých článkov je 15 normostrán, informatívnych článkov 6 normostrán, recenzie, rozhovory a informácie publikujeme v rozsahu maximálne 3 normostrany. Tabuľky, mapy, grafy a obrázky musia mať názov a uvedený zdroj údajov; odporúčame, aby kopírovali šírku textu. Skratky sa používajú len minimálne, pri prvom použití je potrebné skratku v zátvorke rozpísať. Redakcia zabezpečuje jazykovú úpravu textu.

Príspevky sú recenzované. Oponentské konanie je obojstranne anonymné. Konečné rozhodnutie o publikovaní článku vydáva redakčná rada.

Redakcia si vyhradzuje právo zverejniť články schválené redakčnou radou v tlačenej a elektronickej podobe na ssad.statistics.sk.

INFORMATION FOR AUTHORS

Articles are accepted in Slovak, Czech and English languages and must comply with the journal's professional specialisation and scientific nature as well. The submitted articles should not be reviewed by another journal and should not have already been published in any specialised or other press.

Please submit your articles in electronic form, in MS Word or Open Office format, Arial font, size 12 and typed in single spacing. The author's name and workplace should be indicated above the title.

Articles should contain an abstract (general description of the objective and the processing methods used up to 100 words), key words (max. 5), resume (brief summary of the article's content emphasizing its contribution and the most important conclusions up to 500 words), curriculum vitae of the author (no more than 120 words) and the author's contact (e-mail address). The author should submit the article's title, abstract, key words and resume in English language. List of the literature used with full bibliographic data should be given in alphabetical order at the end of an article. Bibliographic citations should be given in square brackets. References are indicated by numbers in a text in square brackets. Footnotes should be numbered in the order of the corresponding page of a text. Authors can find more details at the website ssad.statistics.sk.

Maximum scope of a scientific article is up to 15 standard pages, informative articles should be up to 6 standard pages in length, reviews, discussions and information not more than 3 standard pages. Tables, maps, graphs and pictures should have a title and the data source indicated, it is also advised to copy the width of a text. Abbreviations should be used only rarely and should be appropriately explained in parentheses when first used. Language text revisions are provided by the editorial office.

Articles are reviewed. The opponent procedure is mutually anonymous. The final decision on the article's publication is made by the editorial board. The editorial office reserves the right to publish articles approved by the editorial board in printed and electronic form at the website ssad.statistics.sk.

SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

je jediný recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov. Propagujeme miesto a význam slovenskej štatistiky v Európskom štatistickom systéme, spoluprácu Eurostatu a národných štatistických úradov pri harmonizácii zisťovaní a multidimenzionálny rozmer štatistiky. Podporujeme rozvoj štatistickej teórie a jej prepojenie s praxou. Naším cieľom je prispievať k využiteľnosti štatistických výstupov v rôznych oblastiach a k zvyšovaniu ich kvality a efektivity.

Publikujeme analytické články, prognózy, názory, diskusné príspevky, recenzie, rozhovory, informácie a oznamy z rôznych oblastí štatistiky (národné účty, produkčné štatistiky, sociálne štatistiky, štatistika životného prostredia a pod.) a demografie (demografická štatistika, teoreticko-metodologické východiská demografie, historická demografia a pod.), vrátane sčítania obyvateľov, domov a bytov ako neodmysliteľnej súčasti demografickej štatistiky.

Vydáva:

Štatistický úrad SR

Identifikačné číslo vydavateľa:

IČO 00166197

Vychádza:

Štyrikrát ročne

Dátum vydania:

15. január 2023

Tlač:

Reprografické stredisko
Štatistického úradu SR

Predplatné:

20 € (na rok)
5 € (za jeden výtlačok)

Objednávky prijíma:

Informačný servis
Štatistického úradu SR
Tel.: +4212/502 36 339
+4212/502 36 335
E-mail: info@statistics.sk

SLOVAK STATISTICS AND DEMOGRAPHY

is the only scientific reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures. Our aim is to promote the position and importance of Slovak statistics in the European Statistical System, cooperation between the Eurostat and the national statistical offices in the field of survey harmonisation and the multidimensional character of statistics as well. We support the development of statistical theory and its connection with practice. We aim to contribute to the utility of statistical outputs in various fields and to the improvement of quality and efficiency.

We publish analytic articles, prognoses, views, discussion contributions, reviews, discussions, information and announcements from various statistical fields (national accounts, production statistics, social statistics, environmental statistics etc.) and demography (demographic statistics, theoretical and methodological bases of demography, historical demography etc.) including the population and housing census as an essential part of demographic statistics.

Issued by:

Statistical Office of the SR

Company registration number:

00166197

Published:

Four times a year

Date of issue:

15th January 2023

Press:

Reprographic centre of the
Statistical Office of the SR

Subscription:

€20 (per year)
€5 (for one copy)

Orders are to be addressed to:

Information Service of the
Statistical Office of the SR
Tel.: +4212/502 36 339
+4212/502 36 335
E-mail: info@statistics.sk