SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2017
ročník/volume 27

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

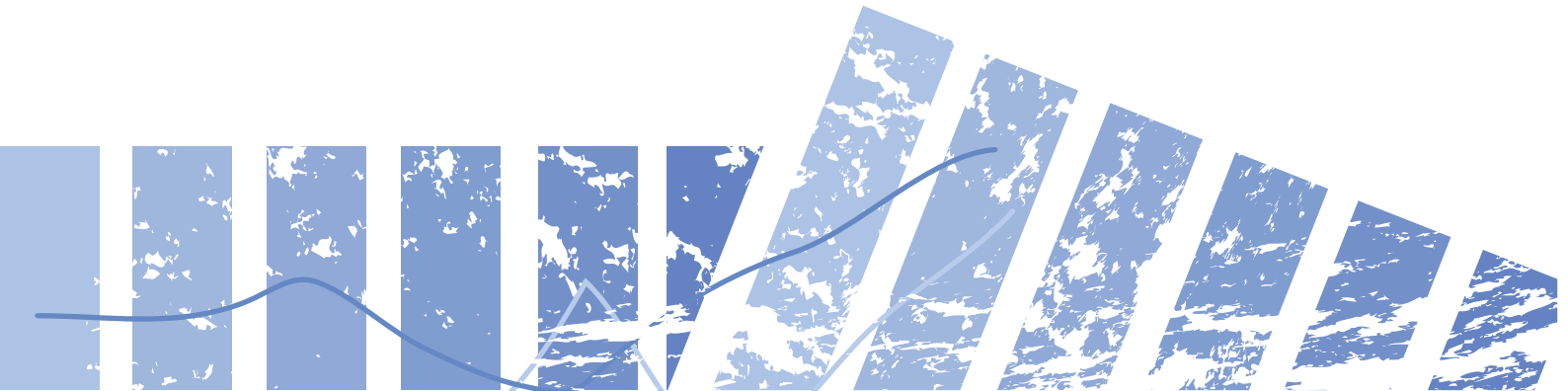
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 2

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

Strany/Pages: 7 – 19

Dátum vydania/Publication date: 15. júl 2017/July 15, 2017



Milan TEREK
Katedra štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave

NAVRHOVANIE KOMPLEXNÝCH ŠTATISTICKÝCH PRIESKUMOV A NIEKTORÉ MOŽNOSTI ANALÝZY Z NICH ZÍSKANÝCH DÁT

DESIGNING COMPLEX STATISTICAL SURVEYS AND SOME POSSIBILITIES OF DATA ANALYSIS OBTAINED THEREFROM

ABSTRAKT
Článok sa zaoberá možnosťami navrhovania komplexných štatistických prieskumov a vybranými analýzami dát získaných z týchto prieskumov. Uvádza sa v ňom všeobecný postup návrhu výberovej schémy komplexného štatistického prieskumu. Analýzu dát z komplexného prieskumu možno realizovať postupnosťou krokov odhadovania charakteristík z najnižšej po najvyššiu úroveň výberovej schémy. Využitie výberových váh môže proces analýzy dát z komplexného prieskumu často značne zjednodušiť. Článok sa venuje opisu formulácie empirickej pravdepodobnostnej funkcie, empirickej distribučnej funkcie a odhadovania mediánu s využitím výberových váh. Na základe dát z komplexného prieskumu EU SILC 2014 v Slovenskej republike sa odhadujú a porovnávajú mediány celkových hrubých príjmov domácností v ôsmich slovenských regiónoch.

ABSTRACT
The paper deals with the possibilities of designing complex statistical surveys and selected analysis of data obtained from these surveys. A description is provided for the general approach to the sampling scheme of the complex statistical survey. Data analysis from the complex survey can be realized by sequence of actions estimating the characteristics from lowest to the highest level of the sampling scheme. The use of sampling weights can often greatly simplify the data analysis process from complex surveys. The paper describes the construction of empirical probability mass function, empirical cumulative distribution function and estimation of median with the use of sampling weights. The medians of the total gross household incomes in eight Slovak regions were estimated and compared on the basis of the data from the Slovak Republic: EU-SILC, 2014.

KĽÚČOVÉ SLOVÁ
komplexný štatistický prieskum, výberové váhy, empirická pravdepodobnostná funkcia, empirická distribučná funkcia, odhadovanie mediánu

KEYWORDS
complex statistical survey, sampling weights, empirical probability mass function, empirical cumulative distribution function, median estimation

1. ÚVOD

Štatistický prieskum[1], ktorý obsahuje viacero takých komponentov, ako je napríklad náhodné vyberanie[2], stratifikácia, skupinové vyberanie, vyberanie s nerovnakými pravdepodobnosťami, pomerové odhadovanie a podobne, sa zvyčajne nazýva komplexný štatistický prieskum. Uvedieme moduly na konštrukciu komplexných štatistických prieskumov a základné možnosti analýzy dát, ktoré sme z nich získali.

Všetky prezentované procedúry vyžadujú využitie pomocných informácií. Všeobecne, pomocné informácie sú ľubovoľné informácie, ktoré nepochádzajú z výberu a ktoré môžu zlepšiť presnosť hodnôt odhadov[3]. Pomocné informácie možno využiť v etape tvorby plánu výberového skúmania aj v etape odhadovania pri konštrukcii bodových odhadov[4]. Slúžia na vytvorenie vhodnej výberovej schémy (*sampling design*) a/alebo na výpočet hodnôt odhadov. V oboch prípadoch sa nazývajú pomocné premenné[5]. Hodnoty pomocných premenných možno často získať z rozličných registrov, napríklad z obchodného registra alebo z registra obyvateľov. V návrhoch komplexných štatistických prieskumov je veľmi dôležitá voľba vhodných pomocných premenných. Metódy na identifikáciu najvhodnejších pomocných premenných v súvislosti s vychýlením bodových odhadov sú podrobne opísané v [10].

V článku sa podrobnejšie zmienime o možnostiach využitia výberových váh pri konštrukcii empirickej pravdepodobnostnej funkcie, empirickej distribučnej funkcie a pri odhadovaní niektorých charakteristík základného súboru. Postupy budeme ilustrovať na analýze dát z komplexného štatistického prieskumu EU SILC (European Union Statistics on Income and Living Conditions), ktorý sa na Slovensku realizoval v roku 2014. Budeme analyzovať regionálnu štruktúru príjmov na základe hodnôt odhadov mediánu celkových hrubých príjmov domácností v ôsmich slovenských regiónoch.

2. MATERIÁL A METÓDY

Článok poskytne charakteristiku niektorých základných komponentov komplexných štatistických prieskumov, spôsob ich navrhovania a základné spôsoby analýzy dát z komplexných štatistických prieskumov. Uvedieme možnosti využívania výberových váh pri odhadovaní rozdelenia pravdepodobnosti pre základný súbor a pri odhadovaní niektorých charakteristík základného súboru.

---

[1] *Štatistický prieskum (statistical survey, survey) je proces zhromažďovania dát prostredníctvom zisťovania odpovedí jednotiek (osôb, domácností, firiem a pod.) na otázky. V rovnakom význame sa používajú aj termíny anketa alebo zisťovanie.*

[2] *Výberový súbor alebo výber (sample) je vybratá časť základného súboru. Postup získavania výberu nazveme vyberanie (sampling). Náhodné alebo pravdepodobnostné vyberanie (random sampling, probability sampling) je také vyberanie, že pravdepodobnosť každého výberu z daného základného súboru je známa. Množina hodnôt pozorovaní, ktorá sa získala náhodným vyberaním, sa nazýva náhodný výber (random sample). V induktívnej štatistike sa hodnoty v náhodnom výbere považujú za realizácie náhodných premenných – pozorovaní. Množina týchto pozorovaní sa tiež nazýva náhodný výber.*

[3] *Keď sa použije hodnota v nejakej výberovej charakteristike V na odhadnutie charakteristiky θ základného súboru, ide o bodové odhadovanie (point estimation). Samotná výberová charakteristika V sa nazýva bodovým odhadom (point estimator) charakteristiky θ a jej hodnota v sa nazýva hodnotou bodového odhadu (point estimate) V charakteristiky θ. Pri bodovom odhadovaní sa charakteristika základného súboru odhaduje jediným číslom alebo jediným bodom na osi reálnych čísel.*

[4] *Podrobnejšie pozri v [4].*

[5] *Podrobnejšie pozri napríklad v [9], [10].*

2.1. Navrhovanie komplexných štatistických prieskumov

Budeme charakterizovať viaceré komponenty komplexného štatistického prieskumu: jednoduché náhodné vyberanie, stratifikácia, skupinové vyberanie atď. Ďalej uvedieme, ako z nich možno vytvoriť jedinú výberovú schému.

2.1.1. Komponenty komplexného štatistického prieskumu

Jednoduché náhodné vyberanie je najjednoduchšia forma náhodného vyberania. Ide o náhodné vyberanie jednotiek[6] bez opakovania alebo s opakovaním, ktoré sa realizuje priamo z celého základného súboru. Pri jednoduchom náhodnom vyberaní má každá možná podmnožina *n* jednotiek v základnom súbore rovnakú pravdepodobnosť tvoriť náhodný výber rozsahu *n*. Výsledkom jednoduchého náhodného vyberania je jednoduchý náhodný výber.

Jednoduché náhodné vyberanie je najjednoduchšia výberová schéma. Všimnime si teraz stratifikované náhodné vyberanie. Ide o náhodné vyberanie, v ktorom sa základný súbor delí na vzájomne sa vylučujúce a základný súbor celkom pokrývajúce podsúbory, ktoré sa nazývajú stratá. Stratá sa vzhľadom na skúmanú premennú považujú za viac homogénne ako celý základný súbor. Z každého strata sa získa jednoduchý náhodný výber, pričom jednotlivé výbery sa zo strát vyberajú nezávisle. Súbor vytvorený zo všetkých získaných výberov tvorí stratifikovaný náhodný výber zo základného súboru. Stratá sa najčastejšie definujú na základe záujmových podskupín základného súboru (domén), napríklad môže ísť o regióny krajiny pri prieskumoch, ktoré sa týkajú obyvateľstva krajiny, alebo veľkostné kategórie firiem pri prieskumoch, ktoré sa týkajú firiem. Jednotky v tom istom strate majú obyčajne tendenciu viac sa podobať ako jednotky náhodne vybraté z celého základného súboru, preto stratifikácia často zvyšuje presnosť hodnôt odhadov. Stratifikácia sa často používa na redukciu variability bodových odhadov a na získavanie separovaných hodnôt odhadov pre domény. Stratifikované viacstupňové vyberanie je založené na využití hierarchickej štruktúry jednotiek v každom strate.

Skupinové vyberanie je náhodné vyberanie, v ktorom každá jednotka patrí do určitej skupiny a skupiny sa vyberajú podľa konkrétnej výberovej schémy. Pri jednostupňovom skupinovom vyberaní je každá jednotka z náhodne vybratých skupín zaradená do výberu. Pri dvojstupňovom skupinovom vyberaní sa náhodne vyberú len niektoré jednotky z náhodne vybratých skupín. Všeobecne je možné navrhnúť ľubovoľný konečný počet stupňov. Skupinové vyberanie, v ktorom všetky skupiny v základnom súbore nemajú rovnakú pravdepodobnosť byť vybraté do výberu, sa nazýva skupinové vyberanie s nerovnakými pravdepodobnosťami. Často sa navrhuje skupinové vyberanie s pravdepodobnosťami výberu úmernými veľkosti skupiny. Skupinové vyberanie, niekedy s nerovnakými pravdepodobnosťami, sa zvyčajne navrhuje z dôvodu redukcie nákladov spojených s prieskumom. Často sa v komplexných štatistických prieskumoch využíva pomerové a regresné odhadovanie.

2.1.2. Postup pri navrhovaní komplexných štatistických prieskumov

Pri návrhu komplexného štatistického prieskumu sa odporúča postupovať tak, že sa známe koncepty formulujú v modulárnej forme[7]. Moduly sa potom v rozličných možných kombináciách integrujú do jedinej výberovej schémy. Prvý modul je

[6] *Nie skupín jednotiek.*
[7] *Podrobnejšie v [7], s. 281 – 282.*

založený na stratifikácii, druhý na skupinovom vyberaní s opakovaním[8] a tretí na skupinovom vyberaní bez opakovania[9].

Stratifikácia je obyčajne jadrom výberovej schémy. Stratá môžu byť napríklad regióny v krajine, typy sídiel a podobne. Skupiny (niekedy viacero stupňov skupín) sa vyberajú z každého strata v návrhu a vnútri skupín sa môže objaviť dodatočná stratifikácia. Veľa prieskumov používa stratifikované viacstupňové vyberanie, v ktorom sa využíva stratifikovaný výber primárnych jednotiek a podvýbery sekundárnych jednotiek sa vyberajú z každej vybratej primárnej jednotky. Keď sa komplexný prieskum skladá z viacstupňového skupinového vyberania a stratifikácie, je užitočné vytvoriť diagram alebo tabuľku výberovej schémy. Niekedy sa oplatí využiť pri odhadovaní aj pomerové odhadovanie. To sa bežne využíva na ľubovoľnej úrovni výberovej schémy.

V [7] sa na s. 282 – 283 uvádza zaujímavý príklad. V roku 1991 sa v Gambii vo vidieckych sídlach realizoval prieskum zameraný na mieru používania ochranných sietí na posteľ impregnovaných insekticídom proti komárom prenášajúcim maláriu. Výberová báza (opora výberu) pozostávala z vidieckych sídiel v Gambii s 3 000 a menej obyvateľmi. Vidiecke sídla boli stratifikované podľa dvoch stratifikačných premenných − regiónu (východný, centrálny a západný) a existencie verejnej kliniky (áno, nie). Stratifikácia sa realizovala v troch stupňoch. V každom regióne sa vybralo päť okresov (*districts*) s pravdepodobnosťami úmernými počtu obyvateľov okresu. V druhom stupni sa v každom vybratom okrese vybrali štyri vidiecke sídla opäť s pravdepodobnosťami úmernými počtu obyvateľov – dve s verejnými klinikami a dve bez nich. Nakoniec sa v každom vidieckom sídle náhodne vybralo šesť objektov (*compounds*) a v nich sa zistil počet postelí a ochranných sietí spolu s inými informáciami. Výberová schéma je charakterizovaná v tabuľke 1.

Tabuľka č. 1: Výberová schéma

Stupeň	Výberová jednotka	Stratifikácia
1	okres	región
2	vidiecke sídlo	existencia verejnej kliniky
3	objekt	

Zdroj údajov: vlastné spracovanie podľa [7], s. 283

Pri výpočte hodnôt odhadov alebo smerodajných chýb sa začína na 3. stupni a postupuje sa smerom nahor. Odhadovanie úhrnu (celkového počtu) ochranných sietí vo vidieckych sídlach v Gambii pozostávalo zo 6 krokov. Zaznamenal sa celkový počet ochranných sietí pre každé vidiecke sídlo, vypočítala sa hodnota odhadu úhrnu sietí pre každé vidiecke sídlo, hodnota odhadu úhrnu pre vidiecke sídla s verejnou klinikou a rovnako bez verejnej kliniky v každom okrese, hodnota odhadu počtu sietí v každom okrese, hodnota odhadu úhrnu sietí pre každý región a nakoniec hodnota odhadu úhrnu sietí v Gambii. Podobne sa odhadli aj rozptyly. Celý postup je pomerne komplikovaný. Vo všeobecnosti nie je nevyhnutné pri odhadovaní v komplexných štatistických prieskumoch používať pomerne zložité postupy ako v spomínanom príklade. V mnohých prípadoch odhadovanie uľahčuje využitie výberových váh.

[8] *Náhodným vyberaním s opakovaním sa vyberie n skupín – primárnych jednotiek.*
[9] *Náhodným vyberaním bez opakovania sa vyberie n primárnych jednotiek.*

2.2. Výberové váhy a ich využitie pri odhadovaní

Vo výberových skúmaniach sa najčastejšie odhaduje stredná hodnota, úhrn, podiel alebo pomer. Niekedy je potrebné odhadovať medián a iné kvantily[10]. Vhodným prostriedkom nato sú výberové váhy. Výberové váhy umožňujú aj konštrukciu empirického rozdelenia pre základný súbor[11]. Výberové váhy možno vypočítať na základe pomocných informácií.

Predpokladajme že poznáme rozsah *N* konečného základného súboru *U*. Symbolom *x* označíme študovanú premennú aj jej hodnoty, $U = \{1, 2, \ldots N\}$ je množina indexov jednotiek v základnom súbore. Symbol *S* označuje výber zo základného súboru – podmnožinu, ktorá obsahuje *n* jednotiek z *U*. Nech π_i je pravdepodobnosť, že jednotka $i \in U$ bude v náhodnom výbere. Výberové váhy w_i pre ľubovoľnú výberovú schému sú definované takto:

$$w_i = \frac{1}{\pi_i} \ . \tag{1}$$

Výberovú váhu jednotky *i* možno interpretovať ako počet jednotiek v základnom súbore, ktoré sú reprezentované jednotkou *i* vo výbere.

V jednoduchom náhodnom výbere má každá jednotka v základnom súbore pravdepodobnosť $\pi_i = n/N$ byť v náhodnom výbere. V tejto výberovej schéme je výberová váha každej jednotky v základnom súbore $w_i = 1/\pi_i = N/n$. Každá jednotka v jednoduchom náhodnom výbere reprezentuje samu seba a $N/(n - 1)$ ďalších nevybratých jednotiek v základnom súbore. V jednoduchom náhodnom výbere

$$\sum_{i \in S} w_i = \sum_{i \in S} \frac{N}{n} = N \ . \tag{2}$$

Je známe, že stredná hodnota μ_K premennej *x* konečného základného súboru je definovaná takto:

$$\mu_K = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3}$$

a úhrn τ konečného základného súboru je

$$\tau = \sum_{i=1}^{N} x_i = N \mu_K \ . \tag{4}$$

V jednoduchom náhodnom výbere je výberový priemer $\bar{X}$ nevychýleným bodovým odhadom strednej hodnoty μ_K. Jeho hodnota $\bar{x}$ sa vypočíta podľa vzťahu

$$\bar{x} = \frac{1}{n} \sum_{i \in S} x_i \ . \tag{5}$$

[10] *Podrobnejšie o kvantiloch pozri napríklad v [13].*
[11] *V skutočnosti ide o empirické rozdelenie pozorovania zo skúmaného konečného základného súboru.*

Na odhadovanie úhrnu τ možno použiť nevychýlený bodový odhad $N\overline{X}$.

Pomocou výberových váh možno úhrn a strednú hodnotu konečného základného súboru odhadnúť nasledujúcimi hodnotami:

$$N\overline{x} = \sum_{i \in S} \frac{N}{n} x_i = \sum_{i \in S} w_i x_i \tag{6}$$

$$\overline{x} = \frac{N\overline{x}}{N} = \frac{\sum_{i \in S} w_i x_i}{\sum_{i \in S} w_i} \ . \tag{7}$$

Všimnime si teraz stratifikovaný základný súbor. Nech je základný súbor rozsahu *N* rozdelený na *H* strát. Stredná hodnota *h*-teho strata je definovaná takto:

$$\mu_h = \frac{1}{N_h} \sum_{i=1}^{N_h} x_{hi} \ , \tag{8}$$

kde x_{hi} je hodnota premennej *x*, *i*-tej jednotky v *h*-tom strate, N_h – rozsah *h*-teho strata v základnom súbore.

Stredná hodnota μ_K premennej *x* je definovaná takto:

$$\mu_K = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \tag{9}$$

Úhrn τ v stratifikovanom základnom súbore je definovaný takto:

$$\tau = \sum_{h=1}^{H} N_h \mu_h \tag{10}$$

Hodnota $\overline{x}_h$ výberového priemeru $\overline{X}_h$ v stratách sa vypočíta podľa vzťahu

$$\overline{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi} \ , \tag{11}$$

kde n_h je rozsah výberu z *h*-teho strata.

Z každého strata sa náhodným vyberaním bez opakovania získa výberový súbor. Výberový priemer

$$\overline{X}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{X}_h = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi} = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{N_h}{n_h} X_{hi} \tag{12}$$

je nevychýleným bodovým odhadom strednej hodnoty μ_K stratifikovaného základného súboru. Výberový úhrn

$$N\overline{X}_{str} = \sum_{h=1}^{m} N_h \overline{\overline{X}}_h = \sum_{h=1}^{m} \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^{m} \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{N_h}{n_h} X_{hi} \tag{13}$$

je nevychýleným bodovým odhadom úhrnu τ stratifikovaného základného súboru. Výberová váha *i*-tej jednotky z *h*-teho strata je

$$w_{hi} = \frac{N_h}{n_h} . \tag{14}$$

Výberovú váhu w_{hi} možno interpretovať ako počet jednotiek v *h*-tom strate základného súboru reprezentovaných *i*-tou jednotkou z *h*-teho strata vo výbere. Zrejme

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} = N. \tag{15}$$

Pomocou výberových váh možno strednú hodnotu a úhrn stratifikovaného základného súboru odhadnúť hodnotami

$$\overline{x}_{str} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} x_{hi} \tag{16}$$

a

$$N\overline{x}_{str} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} x_{hi} . \tag{17}$$

Majme základný súbor, v ktorom je napríklad 2 000 mužov a 500 žien. Nech je pohlavie stratifikačnou premennou. Uvažujeme teda o dvoch stratách – muži a ženy. Jednoduchým náhodným vyberaním sme z každého strata vybrali 100 jednotiek. V stratifikovanom náhodnom výbere je potom 100 mužov a 100 žien. Pravdepodobnosť výberu muža je 100/2 000 = 1/20 a pravdepodobnosť výberu ženy je 100/500 = 1/5. Výberová váha každého muža vo výbere je 20, výberová váha každej ženy vo výbere je 5. Každý muž vo výbere reprezentuje 20 mužov v základnom súbore, každá žena vo výbere reprezentuje 5 žien v základnom súbore. Zrejme suma váh sa rovná rozsahu základného súboru

$$\sum_{h=1}^{2} \sum_{i=1}^{100} w_{hi} = 100 \cdot 20 + 100 \cdot 5 = 2\,500.$$

Bodové odhady uvedených a iných charakteristík konečného základného súboru v skupinovom vyberaní a v iných výberových schémach vrátane ich kombinácií takých, ako napríklad viacstupňové stratifikované skupinové vyberanie, možno vyjadriť pomocou výberových váh.

Výberové váhy sa dajú modifikovať vzhľadom na neodpovedanie a na chyby pokrytia. Doteraz sme uvažovali len o tzv. základných váhach (*base weights*), ktoré sú odvodené z plánu výberového skúmania. Označme ich w_{Bi}. Váhy w_{NRi} sa považujú

za faktory úpravy vzhľadom na neodpovedanie. Váhy w_{NCi} sa považujú za faktory kompenzácie nepokrytia[12]. Posledné dva uvedené typy váh sa považujú za faktory úpravy základných váh[13].

Konečná výberová váha w_i pre *i*-tú jednotku vo výbere je

$$w_i = w_{Bi} \cdot w_{NRi} \cdot w_{NCi} . \tag{18}$$

Výberové váhy pre všetky pozorovania sú rovnaké v samovážiacich výberoch. Takéto výbery možno považovať[14] za reprezentatívne v tom zmysle, že každá pozorovaná jednotka reprezentuje rovnaký počet nepozorovaných jednotiek v základnom súbore[15]. Výberové váhy pre všetky pozorovania nie sú rovnaké v nesamovážiacich výberoch.

2.2.1. Odhadovanie empirickej pravdepodobnostnej funkcie, empirickej distribučnej funkcie a niektorých charakteristík základného súboru pomocou výberových váh

Predpokladajme, že sú známe hodnoty premennej *x* pre všetkých *N* jednotiek v základnom súbore. Hodnota pravdepodobnostnej funkcie v bode *x* je

$$p(x) = \frac{N_{(x)}}{N} , \tag{19}$$

kde $N_{(x)}$ je počet jednotiek, ktoré majú hodnotu premennej *x*. Hodnota distribučnej funkcie v bode *x* je

$$F(x) = \sum_{y \le x} p(y) . \tag{20}$$

Poznamenajme, že ide o pravdepodobnostnú funkciu a distribučnú funkciu pozorovania zo základného súboru[16].

Výberové váhy umožňujú formulovať empirickú pravdepodobnostnú funkciu a empirickú distribučnú funkciu. Empirická pravdepodobnostná funkcia $\hat{p}(x)$ je definovaná ako suma váh všetkých pozorovaní, ktoré majú hodnotu *x,* delená sumou všetkých váh

$$\hat{p}(x) = \frac{\sum_{i \in S ; x_i = x} w_i}{\sum_{i \in S} w_i} . \tag{21}$$

---

[12] Jednotky, ktoré sú v cieľovom základnom súbore, ale nie sú vo výberovej báze, vytvárajú nepokrytie alebo neúplné pokrytie.

[13] Podrobnejšie o výberových váhach pozri napríklad v [6] alebo v [11].

[14] V prípade absencie nevýberových chýb.

[15] Termín „reprezentatívny výber“ môže mať rozličné významy. V [2], s. 23 – 24 je uvedených deväť rozličných významov tohto termínu, ktoré sa bežne používajú. Odporúča sa používať tento termín len s vysvetlením jeho chápania v konkrétnom texte.

[16] Uvažujeme o prístupe k výberovému skúmaniu známemu ako prístup „bez modelu“ alebo „bez rozdelenia“ (podrobnejšie pozri v [3], s. 8 – 9).

Empirická distribučná funkcia $\hat{F}(x)$ je

$$\hat{F}(x) = \sum_{y \le x} \hat{p}(y). \tag{22}$$

2.2.2. Odhadovanie niektorých charakteristík základného súboru pomocou výberových váh

Pomocou empirickej pravdepodobnostnej funkcie $\hat{p}(x)$ a empirickej distribučnej funkcie $\hat{F}(x)$ možno odhadovať charakteristiky základného súboru. Napríklad strednú hodnotu základného súboru možno odhadnúť hodnotou

$$\hat{\mu}_K = \sum_x x\hat{p}(x) = \frac{\sum_{i \in S} w_i x_i}{\sum_{i \in S} w_i}. \tag{23}$$

Keby bola distribučná funkcia *F* spojitá, medián základného súboru by bol definovaný ako hodnota $\tilde{\mu}$, pre ktorú $F(\tilde{\mu}) = \frac{1}{2}$. V diskrétnom prípade je medián konečného základného súboru definovaný ako hodnota $\tilde{\mu}$, pre ktorú $F(\tilde{\mu}) = \frac{1}{2}$, ak taká hodnota existuje. Inak je medián konečného základného súboru definovaný ako ľubovoľná hodnota z intervalu $[\tilde{\mu}_1, \tilde{\mu}_2]$, kde $\tilde{\mu}_1$ je najväčšia hodnota *x* v základnom súbore, pre ktorú $F(x) < \frac{1}{2}$ a $\tilde{\mu}_2$ je najmenšia hodnota *x,* pre ktorú $F(x) > \frac{1}{2}$. Všeobecne Q_p je $p \cdot 100$ % kvantil (percentil), keď $F(Q_p) = p$, ak taká hodnota existuje, inak $Q_p \in [a, b]$, kde *a* je najväčšia hodnota *x* v základnom súbore, pre ktorú $F(x) < p$ a *b* je najmenšia hodnota *x,* pre ktorú $F(x) > p$. Keď $p < \frac{1}{N}$, Q_p je najmenšia hodnota *x,* a keď $p > 1 - \frac{1}{N}$, Q_p je najväčšia hodnota *x*.

V ďalšej časti článku ukážeme, ako sa odhadujú kvantily v základnom súbore. Pretože empirická distribučná funkcia $\hat{F}$ je kroková funkcia, na nájdenie jedinej hodnoty kvantilu je obyčajne potrebná interpolácia. Nech y_1 je najväčšia hodnota vo výbere, pre ktorú $\hat{F}(y_1) \le p$ a, nech y_2 je najmenšia hodnota vo výbere, pre ktorú $\hat{F}(y_2) \ge p$, potom

$$\hat{Q}_p = y_1 + \frac{p - \hat{F}(y_1)}{\hat{F}(y_2) - \hat{F}(y_1)}(y_2 - y_1). \tag{24}$$

Poznamenajme, že bodové odhady založené na použití výberových váh nie sú nevyhnutne nevychýlené alebo numericky stabilné. Napriek tomu hodnoty štatistík, ktoré sa počítajú pomocou výberových váh, sú obyčajne oveľa bližšie skutočným hodnotám charakteristík základného súboru ako v prípade použitia bodových odhadov, ktoré neberú do úvahy štruktúru dát ([7], s. 293).

3. ANALÝZA REGIONÁLNEJ ŠTRUKTÚRY PRÍJMOV NA BÁZE DÁT Z EU SILC 2014

Rozdelenia príjmov sú obyčajne výrazne zošikmené a obsahujú odľahlé hodnoty[17]. Výpovedná schopnosť strednej hodnoty v takýchto rozdeleniach je veľmi malá[18] a stredná hodnota sa nepovažuje za vhodnú charakteristiku centra rozdelenia. Stredná hodnota príjmu nie je vhodnou charakteristikou „typického“ príjmu. V takýchto rozdeleniach sa všeobecne považuje za dobrú charakteristiku centra rozdelenia medián. Ide o stabilnú charakteristiku, robustnú voči odľahlým hodnotám. Alternatívne možno odporúčať ako vhodné charakteristiky centra rozdelenia aj niektoré netradičné charakteristiky, napríklad zastrihnutú strednú hodnotu (*trimmed mean*)[19], winsorizovanú strednú hodnotu alebo M-estimátory[20]. Podobné výsledky poskytujú aj tradičné charakteristiky aplikované na redukovaný súbor dát[21].

Prieskumy EU SILC sa vykonávajú každoročne vo všetkých krajinách Európskej únie vrátane Slovenskej republiky. Týkajú sa domácností a osôb. Na úrovni domácností sa zhromažďujú dáta o viacerých kategóriách príjmov. Podobné prieskumy sa vykonávajú aj v mnohých krajinách mimo Európskej únie. V Slovenskej republike sa prieskum EU SILC realizuje ako stratifikovaný, dvojstupňový s dvomi stratifikačnými premennými – región a veľkosť sídla. Zisťovanie EU SILC 2014 sa uskutočnilo vo vybraných 6 010 domácnostiach. Databázu tvoria dáta o 5 490 domácnostiach a 13 433 osobách starších ako 16 rokov. Vypočítané výberové váhy boli modifikované vzhľadom na neodpovedanie. Tieto váhy možno využívať na tvorbu induktívnych úsudkov o domácnostiach v Slovenskej republike. Iné váhy boli vypočítané pre každú z osôb. Výberové dáta z EU SILC sú vo všeobecnosti dátami z komplexného štatistického prieskumu a výberový súbor je nesamovážiaci.

Dáta z EU SILC 2014 sa nachádzajú vo viacerých súboroch. Každej domácnosti je priradené identifikačné číslo[22]. Bola vykonaná analýza celkových hrubých príjmov domácností v ôsmich doménach – slovenských regiónoch. Tieto domény korešpondujú s hodnotami jednej zo stratifikačných premenných. Najskôr sa vykonalo spárovanie dát – výberových váh[23] a celkových hrubých príjmov domácností[24] podľa čísel domácností. Potom sa spárovali dáta rozdelené podľa regiónov. Takýmto spôsobom vzniklo osem súborov dát, jeden pre každý región. Každý región sa analyzoval separovane. Potom sa pre každý región podľa (21) vypočítali hodnoty empirickej pravdepodobnostnej funkcie. Na základe týchto hodnôt sa podľa (22) vypočítali pre každý región hodnoty empirickej distribučnej funkcie. Nakoniec sa pre každý región podľa vzťahu (24) vypočítala hodnota odhadu mediánu celkových hrubých príjmov domácností. Všetky výpočty sa realizovali v programe Excel 2013. Získané výsledky sú v tabuľke č. 2. Hodnotami odhadu mediánu v treťom stĺpci tabuľky odhadujeme mediány celkových hrubých príjmov domácností v slovenských regiónoch v roku 2014.

---

[17] *Odľahlú hodnotu v množine dát možno definovať ako hodnotu (alebo podmnožinu hodnôt), ktorá sa zdá nekonzistentná s ostatnými hodnotami v množine dát ([1], s. 7).*
[18] *Podrobnejšie pozri v [5].*
[19] *Pozri napríklad v [8], s. 55.*
[20] *Podrobnejšie pozri napríklad v [12] alebo v [14].*
[21] *Podrobnejšie pozri v [12].*
[22] *Hodnoty premenných DB030, HB030.*
[23] *Hodnoty premennej DB090.*
[24] *Hodnoty premennej HY010.*

Tabuľka č. 2: Medián celkových hrubých príjmov domácností v slovenských regiónoch v roku 2014

Číslo regiónu	Región	Hodnota odhadu mediánu celkových hrubých príjmov domácností v roku 2014 (v eurách)	Poradie regiónu podľa mediánu celkových hrubých príjmov domácností
1	Bratislava	14 491,37	1.
2	Trnava	13 969,12	4.
3	Trenčín	14 368,47	2.
4	Nitra	12 379,67	7.
5	Žilina	14 054,85	3.
6	Banská Bystrica	11 746,41	8.
7	Prešov	13 595,22	5.
8	Košice	13 118,16	6.

Zdroj údajov: vlastné výpočty

Medián celkového hrubého príjmu domácností v celej Slovenskej republike v roku 2014 odhadujeme hodnotou 13 305,83 eura.

4. ZÁVER

Pri navrhovaní komplexných štatistických prieskumov možno využiť tri základné moduly − stratifikáciu, skupinové vyberanie s opakovaním a skupinové vyberanie bez opakovania. Ich vhodnou kombináciou spolu s prípadným zaradením viacstupňového vyberania, vyberania s nerovnakými pravdepodobnosťami a pomerového alebo regresného odhadovania možno vytvoriť výberovú schému na komplexný štatistický prieskum.

Niekedy sa dajú pri induktívnych úsudkoch o základnom súbore na báze dát z výberového súboru získaného komplexným štatistickým prieskumom použiť štandardné štatistické metódy a bežný softvér, niekedy nie.

Keď bol výber z konečného základného súboru získaný náhodným vyberaním s opakovaním, pozorovania sú štatisticky nezávislé a rovnako rozdelené. Keď sú vo výberovom súbore všetky pozorovania štatisticky nezávislé a rovnako rozdelené, možno na induktívne úsudky o základnom súbore použiť bežné štatistické metódy a bežný štatistický softvér.

Pri náhodných výberoch získaných z komplexných štatistických prieskumov pomocou zložitejších výberových schém nie sú dva spomenuté predpoklady splnené. Keď je výber samovážiaci[25], možno bežné induktívne štatistické metódy a bežný softvér použiť na získanie hodnôt bodových odhadov. Smerodajné chyby, intervaly spoľahlivosti a testy hypotéz, ktoré poskytne bežný softvér, budú už nesprávne.

Keď je výber nesamovážiaci, nemožno bežné induktívne štatistické metódy a bežný softvér použiť ani na bodové odhadovanie. V uvedenej aplikácii bol k dispozícii nesamovážiaci výber, a preto sa nedal odhadovať medián základného

[25] *Napríklad, keď má stratifikovaný náhodný výber rovnaký výberový pomer (z každého strata sa vyberie rovnaké percento jednotiek), majú všetky jednotky vo výbere rovnakú výberovú váhu – výber je samovážiaci.*

súboru pomocou výberového mediánu; bolo nevyhnutné zohľadniť štruktúru dát. Použili sa výberové váhy.

V [7] na s. 287 – 288 sa uvádza: „Keď čítate článok alebo knihu, v ktorej autori analyzujú dáta z komplexného štatistického prieskumu, všimnite si, či zobrali do úvahy štruktúru analyzovaných dát alebo či len realizovali výpočty pomocou bežného softvéru, ktorý nie je určený na analýzy dát z komplexných štatistických prieskumov. Ak nezobrali do úvahy štruktúru dát, mali by ste sa na výsledky, ktoré prezentujú, pozerať s veľkým podozrením."

Analýza regionálnej štruktúry príjmov domácností v roku 2014 podľa mediánu ich celkových hrubých príjmov poskytla zaujímavé výsledky. Obyčajne sa predpokladá, že bratislavský región v príjmoch domácností výrazne prevyšuje ostatné slovenské regióny. Analýza ukázala, že rozdiel medzi prvou Bratislavou a druhým Trenčínom nie je veľký, predstavuje len 122,9 eura. Mediány príjmov tretej Žiliny, štvrtej Trnavy a piateho Prešova sú pomerne blízke a tiež sa priveľmi neodlišujú od bratislavského regiónu. Rozdiel medzi prvou Bratislavou a piatym Prešovom je „len" 896,15 eura. Väčšie rozdiely sú medzi poslednými tromi regiónmi – Košicami, Nitrou a Banskou Bystricou. Medián celkových hrubých príjmov v regióne Banská Bystrica je prekvapivo nízky v porovnaní s Bratislavským krajom – je až o 2 744,96 eura nižší.

Na porovnanie sa vykonal aj odhad mediánu celkových hrubých príjmov domácností v slovenských regiónoch pomocou výberového mediánu, teda bez zohľadnenia štruktúry dát prostredníctvom výberových váh[26]. Rozdiel medzi hodnotami odhadov získanými s váhami a bez váh kolíše od (−1,82 %) do 7,50 %, čo nie sú zanedbateľné rozdiely. Pri odhadovaní bez váh vyšlo aj odlišné poradie regiónov: 1. Bratislava, 2. Trenčín, 3. Žilina, 4. Prešov, 5. Košice, 6. Trnava, 7. Nitra, 8. Banská Bystrica.

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

LITERATÚRA

[1] BARNETT, V. − LEWIS, T.: Outliers in Statistical Data. Hoboken: Wiley and Sons, 1994. ISBN 0-471-93094-5.
[2] BETHLEHEM, J.: Applied Survey Methods. A Statistical Perspective. Hoboken: Wiley and Sons, 2009. ISBN 978-0-470-37308-8.
[3] COCHRAN, W. G.: Sampling Techniques. New York: Wiley and Sons, 1977. ISBN 0-471-16240-X.
[4] FULLER, W. A.: Sampling Statistics. USA: Wiley and Sons, 2009. ISBN 978-0-470-45460-2.
[5] HALLEY, R. M.: Measures of Central Tendency, Location, and Dispersion in Wage Survey Research. In: Compensation and Benefits, 2004, No. 36, p. 39-52.
[6] LEVY, P. S. − LEMESHOW, S.: Sampling of Populationas. Methods and Applications. Fourth Edition. Hoboken: Wiley and Sons, 2008. ISBN 978-0-470-04007-2.

[26] *Výpočet bol realizovaný pomocou funkcie Medián v Exceli.*

[7] LOHR, S. L.: Sampling: Design and Analysis. 2nd edition. Boston: Brooks/Cole, 2010. ISBN-10: 0-495-11084-1.
[8] PIEGORSCH, W. W.: Statistical Data Analysis. Foundations for Data Mining, Informatics, and Knowledge Discovery. Chichester: Wiley and Sons, 2015. ISBN 978-1-118-61965-0.
[9] SÄRNDAL, C. E. − SWENSSON, B. − WRETMAN, J.: Model Assisted Survey Sampling. New York: Springer, 2003. ISBN 0-387-40620-4.
[10] SÄRNDAL, C. E. − LUNDSTRÖM, S.: Estimation in Surveys with Nonresponse. Chichester: Wiley and Sons, 2005. ISBN 0-470-01133-5.
[11] TEREK, M.: Možnosti riešenia problému neodpovedania v štatistických prieskumoch. In: Ekonomické rozhľady, 2014, č. 2,, s. 150 – 165.
[12] TEREK, M.: Odľahlé dáta a charakteristiky polohy v analýzach miezd a príjmov. In: Revue sociálno-ekonomického rozvoja, 2016, No. 1.
[13] TEREK, M.: Interpretácia štatistiky a dát. Štvrté doplnené vydanie. Košice: Eqilibria, 2016. ISBN 978-80-8143-177-7.
[14] WILCOX, R. R.: Applying Contemporary Statistical Techniques. Burlington: Academic Press, 2003. ISBN 0-12-751541-0.

RESUME

The statistical survey consisting of more components such as random sampling, stratification, cluster sampling, sampling with unequal probabilities, ratio estimating etc. is commonly referred to as a comprehensive statistical survey. The paper presents the modules for the construction of comprehensive statistical surveys and the main data analysis options from the comprehensive statistical surveys. The above-mentioned procedures require the use of auxiliary information. These information can be used in the planning stage of the sample survey, as well as in the stage of point estimation construction. The article discusses in a detailed manner the use of the construction of sampling weights in empirical probability mass function, empirical cumulative distribution function and in the estimation of some characteristics of the basic file. The procedures are illustrated on the data analysis from the complex EU-SILC survey realized in the Slovak Republic in 2014. The regional structure of incomes was analyzed on the basis of the total gross household incomes median estimates in eight Slovak regions.

PROFESIJNÝ ŽIVOTOPIS

***Prof. Ing. Milan Terek, PhD.**, od roku 1977 do roku 1987 pôsobil na Katedre operačného výskumu a ekonometrie, od roku 1987 pôsobí na Katedre štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. V súčasnosti vyučuje predmety štatistika (v slovenčine aj francúzštine) a aplikácie štatistických metód na prvom stupni štúdia, štatistické riadenie kvality, hĺbková analýza dát a analýza rozhodovania na druhom stupni štúdia a predmety výberové skúmanie a hĺbková analýza dát na treťom stupni štúdia. Na Vysokej škole manažmentu Trenčín/City University of Seattle vedie na doktorandskom štúdiu slovenský aj anglický kurz Kvantitatívne metódy vo výskume v oblasti podnikového manažmentu. Vo výskume sa venuje hlavne aplikáciám výberového skúmania, štatistického riadenia kvality a analýzy rozhodovania v ekonómii a manažmente.*

KONTAKT

milan.terek@euba.sk