Liet. matem. rink., T. 40, spec. nr., 392–398

Kartotinių imčių panaudojimas baigtinių populiacijų tyrimuose

Danutė KRAPAVICKAITĖ (MII), Genovaitė ŠALUČKIENĖ (VU)
el.-paštas: krapav@ktl.mii.lt, genovaites@mail.std.lt

1. Įvadas

Didelę imčių metodų teorijos dalį sudaro populiacijos parametrų įvertinių dispersijų vertinimas. Dispersijų įvertiniai naudojami parametrų įvertinių tikslumo nustatymui, lyginant imčių planų efektyvumą, nustatant imties didumą populiacijos sluoksniuose.

Tegul $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ – baigtinė populiacija su tyrimo kintamaisiais $y^{(1)}, y^{(2)}, \ldots, y^{(q)}$. Vienas iš dažniausiai praktikoje naudojamų populiacijos parametrų yra suma: $t_{y^{(k)}} = \sum_{i=1}^{N} y_i^{(k)}$, $k = 1, 2, \ldots, q$. Daug netiesinių baigtinės populiacijos parametrų, tokių kaip santykis, regresijos ar koreliacijos koeficientai išsireiškia kaip glodžios sumų funkcijos: $\theta = g(t_{y^{(1)}}, \ldots, t_{y^{(q)}})$, ir gali būti vertinami kaip $\widehat{\theta} = g(\widehat{t}_{y^{(1)}}, \ldots, \widehat{t}_{y^{(q)}})$, naudojant sumų įvertinius $\widehat{t}_{y^{(1)}}, \ldots, \widehat{t}_{y^{(q)}}$. Netiesinių įvertinių $g(\widehat{t}_{y^{(1)}}, \ldots, \widehat{t}_{y^{(q)}})$ dispersijų vertinimui gali būti naudojami skleidimo Teiloro eilute ir kartotinių imčių metodai: *jackknife*, subalansuotų pakartojimų (*balanced repeated replication*) ir *bootstrap* metodai ([2], [3], [4]). Teiloro eilutę naudojantis metodas tinka visiems imčių planams ir parametrų įvertiniams, bet jį naudojant reikia skaičiuoti statistikų dalines išvestines vertinamų populiacijos sumų atžvilgiu, kas kartais būna nelengva sudėtingų imties planų ir sudėtingų įvertinių atveju. Čia gali padėti kartotinių imčių metodai, visoms vertinamoms statistikoms naudojantys tą pačią formulę. Tačiau jackknife ir subalansuotų pakartojimų metodai taikytini grąžintinėms imtims arba negrąžintinėms imtims su labai mažu ėmimo dažniu pirmajame imties rinkimo etape. Bootstrap metodas atrodo plačiausiai taikytinas, bet skaičiavimams reikalauja labai daug kompiuterinio laiko. Aukščiau minėtuose teoriniuose darbuose pateikti gana paprastų įvertinių dispersijų vertinimo pavyzdžiai. Lietuvoje imtims iš baigtinių populiacijų kartotinių imčių metodai parametrų įverčių paklaidų vertinimui praktikoje iki šiol buvo nedaug naudoti. Šio darbo tikslas – išbandyti keletą kartotinių imčių metodų sudėtingų įvertinių tikslumo vertinimui konkrečioje situacijoje ir padaryti išvadas apie jų tinkamumą.

2. Darbo jėgos tyrimas

Šiame darbe kartotinių imčių pritaikymo pavyzdžiu pasirinktas Lietuvos Statistikos Departamente vykdomas Darbo jėgos tyrimas. Tyrimo tikslas – įvertinti užimtųjų ir bedarbių skaičių Lietuvoje ir jos dalyse bei gautųjų įverčių paklaidas.

Tyrimo populiacija – 14 metų ir vyresni Lietuvos gyventojai. Ėmimo sąrašas – gyventojų registras. Iš gyventojų registro išrenkama paprastoji atsitiktinė gyventojų imtis. Po to imtis papildoma visais kartu gyvenančiais tiriamojo amžiaus asmenimis. Tokiu būdu namų ūkis, turintis daugiau tiriamojo amžiaus asmenų, turi ir didesnę tikimybę patekti į imtį. Gaunama lizdinė imtis.

Kadangi apklaustųjų žmonių imtis į amžiaus, lyties ir miesto/kaimo grupes pasiskirsto ne visiškai proporcingai tiriamajai populiacijai, tai siekiant šias proporcija suderinti su turimomis demografinėmis konstantomis ir patikslinti įverčius, naudojamas imties sluoksniavimas (*poststratification*), t.y. imties persvėrimas 12-je amžiaus, 2-se lyties ir 10-je apskričių grupių sankirtose, viso $K = 12 \cdot 2 \cdot 10 = 240$ grupėse, kad čia ji atitiktų demografinius duomenis. Tokiu būdu gauto įverčio dispersija yra mažesnė negu klasikinio įverčio, naudojančio vien tik imties planą.

Tyrime yra naudojami dveji imties sluoksniavimo būdu gaunami papildomi svoriai: vieni, skirti užimtiesiems, gauti, naudojant demografinius duomenis, kiti, skirti bedarbiams, – naudojant Darbo biržos duomenis (16 grupių). Būtų naudinga juos kokiu nors būdu apjungti, tačiau sudaryti labai daug grupių negalima, nes jose liktų mažai imties. Todėl, norint imtį suderinti su žinomomis populiacijos konstantomis, bandoma daryti kartotinį persvėrimą pagal skirtingus kintamuosius, naudojant daugiau turimų demografinių duomenų grupių bei prijungiant Darbo biržos duomenis. Tačiau su šiais papildomais svoriais gautų įverčių dispersijų jau negalėtume lengvai įvertinti, nes Teiloro eilutės pagalba užrašyti įvertinių dispersijų įvertiniai tampa labai griozdiški. Todėl yra bandymų įverčių dispersijas vertinti jackknife ir negrąžintinio bootstrap metodais ([1]). Tai daroma ir šiame darbe praktikoje naudojamiems sudėtingiems įvertiniams, naudojant dar ir veidrodinio sutapimo bootstrap metodą.

Vertinant populiacijos sumą darbo jėgos tyrime iš n dydžio gyventojų lizdinės imties, naudojantis vien tik imties plano svoriais w_i, $\widehat{t} = \widehat{t}_y = \sum_{i=1}^{N} w_i y_i$ yra nepaslinktasis tyrimo kintamojo y sumos įvertinys.

Paprasčiausia panaudoti kartotinį persvėrimą būtų galima persveriant du kartus pagal du kintamuosius. Tarkime, kad iš N dydžio populiacijos $\mathcal{U}$ atsitiktinai išrenkame n dydžio imtį s. Pirmą kartą darant imties sluoksniavimą, imtį skaidome į du sluoksnius s_1 ir s_2: $s = s_1 \bigcup s_2$, $s_1 \bigcap s_2 = \oslash$ pagal pirmą grupuojantį kintamąjį, įgyjantį dvi reikšmes. Iš tikrųjų šis kintamasis ir populiaciją suskaido į dvi dalis $\mathcal{U}_1$ ir $\mathcal{U}_2$: $\mathcal{U} = \mathcal{U}_1 \bigcup \mathcal{U}_2$, $\mathcal{U}_1 \bigcap \mathcal{U}_2 = \oslash$. Laikysime, kad yra žinomas populiacijos dalis $\mathcal{U}_1$ ir $\mathcal{U}_2$ sudarančių elementų skaičius M_1 ir M_2. Kita vertus, šias populiacijos konstantas galime įvertinti atitinkamose imties dalyse s_1 ir s_2. Tada gali būti naudojamas tyrimo kintamojo populiacijos sumos įvertinys

$$\begin{aligned}\widehat{t}^{pos} &= \sum_{s_1} w_i \frac{M_1}{\sum_{s_1} w_j} y_i + \sum_{s_2} w_i \frac{M_2}{\sum_{s_2} w_j} y_i \\ &= \sum_{k=1}^{2} \sum_{i \in s} w_i g_k \delta_i(k) y_i = \sum_{i \in s} w_i a_i y_i,\end{aligned}$$

$$a_i = \sum_{k=1}^{2} g_k \delta_i(k), \quad \delta_i(k) = \begin{cases} 1, & i \in s_k, \\ 0 & \text{kitur;} \end{cases} \quad g_k = \frac{M_k}{\sum_{j \in s_k} w_j} = \frac{M_k}{\widehat{M}_k},$$

$k = 1, 2$, $i = 1, \ldots, n$. Antrą kartą imtį skaidome į kitas dvi dalis – sluoksnius s_3 ir s_4: $s = s_3 \bigcup s_4$, $s_3 \bigcap s_4 = \oslash$ pagal kitą kintamąjį, įgyjantį dvi reikšmes. Šio kintamojo populiacijos konstantos M_3 ir M_4 taip pat yra žinomos, bet jas galima galima ir įvertinti imtyse s_3 ir s_4. Taip gaunami nauji papildomi svoriai ir naujas populiacijos sumos įvertinys

$$\hat{\hat{t}}^{pos} = \sum_{i=1}^{n} w_i a_i b_i y_i, \tag{2.1}$$

$$b_i = \sum_{l=3}^{4} g_l \delta_i(l), \quad g_l = \frac{M_l}{\sum_{j \in s_l} w_j a_j} = \frac{M_l}{\widehat{M}_l}, \quad l = 3, 4, \quad i = 1, \ldots, n.$$

Šio įvertinio dispersijos vertinimui bus naudojami kartotinių imčių metodai. Be to, šiais metodais vertinama ir įvertinio $\hat{\hat{t}}^{pos}$ dispersija, kai minėtas kartotinis persvėrimas pagal 2 kintamuosius atliekamas 5 kartus:

$$\hat{\hat{\hat{t}}}^{pos} = \sum_{i=1}^{n} w_i a_i^{(1)} b_i^{(1)} a_i^{(2)} b_i^{(2)} \cdot \ldots \cdot a_i^{(5)} b_i^{(5)} y_i, \tag{2.2}$$

$$a_i^{(1)} = \sum_{k=1}^{2} g_k^{(1)} \delta_i(k), \quad g_k^{(1)} = \frac{M_k}{\sum_{j \in s_k} w_j}, \quad k = 1, 2, \quad i = 1, \ldots, n,$$

$$b_i^{(1)} = \sum_{l=3}^{4} g_l^{(1)} \delta_i(l), \quad g_l^{(1)} = \frac{M_l}{\sum_{j \in s_l} w_j a_j^{(1)}}, \quad l = 3, 4, \quad i = 1, \ldots, n,$$

$$a_i^{(2)} = \sum_{k=1}^{2} g_k^{(2)} \delta_i(k), \quad g_k^{(2)} = \frac{M_k}{\sum_{j \in s_k} w_j a_j^{(1)} b_j^{(1)}}, \quad k = 1, 2, \quad i = 1, \ldots, n,$$

$$b_i^{(2)} = \sum_{l=3}^{4} g_l^{(2)} \delta_i(l), \quad g_l^{(2)} = \frac{M_l}{\sum_{j \in s_l} w_j a_j^{(1)} b_j^{(1)} a_j^{(2)}}, \quad l = 3, 4, \quad i = 1, \ldots, n,$$

ir t.t.,

$$\delta_i(k) = \begin{cases} 1, & i \in s_k, \\ 0 & \text{kitur;} \end{cases} \quad k = 1, 2, \quad i = 1, \ldots, n.$$

3. Kartotinių imčių metodai

3.1. *Jackknife metodas*

Jį naudojant, imtį reikia padalinti į dalis ir perskaičiuoti dominančias statistikas, išmetant po vieną dalį, vėl sugrįžti ir kartoti šį procesą. Iš gautų rezultatų vertinamos originalios statistikos savybės. Turint lizdinę imtį, lizdai (namų ūkiai) ir bus tos dalys. Tegul

L – sluoksnių skaičius,
N_h – lizdų skaičius h-jame populiacijos sluoksnyje, $h = 1, \ldots, L$,
n_h – imtyje esančių lizdų skaičius iš h-ojo sluoksnio, $h = 1, \ldots, L$,
θ – vertinamas parametras,
$\widehat{\theta}$ – jo įvertinys,
$\widehat{\theta}_h$ – vertinamo parametro įvertinys h-jame sluoksnyje, $h = 1, \ldots, L$,
$\widehat{\theta}_{hj}$ – įvertinys, gaunamas, pašalinus j-jį lizdą iš h-ojo sluoksnio, $h = 1, \ldots, L$, $j = 1, \ldots, L$.

Tada jackknife įvertinio $\widehat{\theta}$ dispersijos įvertiniu laikoma

$$\widehat{\mathbf{D}}_{jack}\widehat{\theta} = \sum_{h=1}^{L}\left(1 - \frac{n_h}{N_h}\right)\frac{n_h - 1}{n_h}\sum_{j=1}^{n_h}(\widehat{\theta}_{hj} - \widehat{\theta}_h)^2.$$

3.2. *Negrąžintinis bootstrap metodas (bootstrap without replacement)*

Metodo panaudojimas susideda iš tokių žingsnių:

1) Sudaroma $N_h = k_h n_h$ dydžio pseudo-populiacija h-ajame sluoksnyje, t.y. imtis nepriklausomai pakartojama kiekviename sluoksnyje k_h kartų.
2) Iš pseudo-populiacijos išrenkama n_h dydžio paprastoji atsitiktinė negrąžintinė imtis $y^*_{h1}, y^*_{h2}, \ldots, y^*_{hn_h}$ ir įvertinama statistika $\widehat{\theta}^* = \widehat{\theta}(y^*_{h1}, y^*_{h2}, \ldots, y^*_{hn_h}, h = 1, \ldots, L)$.
3) Žingsnis 2) nepriklausomai kartojamas B kartų, ir gaunami statistikos įvertiniai $\widehat{\theta}^*_1, \widehat{\theta}^*_2, \ldots, \widehat{\theta}^*_B$.
4) Įvertinio $\widehat{\theta}$ dispersijos bootstrap metodu gautas įvertinys yra $\widehat{\mathbf{D}}_{BWO}\widehat{\theta} = \frac{1}{B-1} \cdot \sum_{i=1}^{B}(\widehat{\theta}^*_i - \widehat{\theta})^2$. Jei n_h nėra sveikasis skaičius, tai galima imti $N_h = k_h n_h + r_h$, $0 \leqslant r_h \leqslant n_h - 1$ kiekvienam h su sveikaisiais skaičiais k_h ir r_h. Tam, kad šis įvertinys būtų suderintasis, R.Sitter ([4]) siūlo rinkti paprastąją atsitiktinę $m_h = n_h - \left(1 - \frac{n_h}{N_h}\right)$ didumo imtį ir kartoti ją $k_h = \frac{N_h}{n_h}\left(1 - \frac{1-\frac{n_h}{N_h}}{n_h}\right) = \frac{N_h}{n_h}\frac{m_h}{n_h}$ kartų, o po to atlikti 3), 4) žingsnius.

3.3. *Veidrodinio sutapimo bootstrap metodas (mirror match bootstrap)*

Jis susideda iš tokių etapų:

1) Iš pradinės imties nepriklausomai kiekviename sluoksnyje išrenkama n^*_h dydžio paprastoji atsitiktinė imtis. Ji atspindi pradinę ėmimo schemą.
2) Pirmasis žingsnis kartojamas nepriklausomai k_h kartų kiekviename sluoksnyje, $h = 1, \ldots, L$. Tokiu būdu, kiekviename sluoksnyje gaunama $m_h = k_h n^*_h$ dydžio imtis $y^*_{h1}, \ldots, y^*_{hm_h}$ ir statistika θ įvertinama $\widehat{\theta}^* = \widehat{\theta}(y^*_{h1}, \ldots, y^*_{hn_h}, h = 1, \ldots, L)$.
3) Pakartojus negrąžintinio bootstrap metodo 3) ir 4) žingsnius, gaunamas dispersijos įvertinys $\widehat{\mathbf{D}}_{MM}\widehat{\theta}$ vietoje $\widehat{\mathbf{D}}_{BWO}\widehat{\theta}$.

Specialus n_h^* ir k_h parinkimas:

$$n_h^* = n_h\Big(1 - \frac{n_h^*}{n_h}\Big)/\Big(1 - \frac{n_h}{N_h}\Big), \quad k_h = n_h\Big(1 - \frac{n_h^*}{n_h}\Big)/n_h^*\Big(1 - \frac{n_h}{N_h}\Big)$$

leidžia gauti suderintuosius įvertinio dispersijos įvertinius.

4. Skaičiavimo rezultatai

Skaičiavimams naudojama realaus 1999 m. lapkričio mėnesio tyrimo imtis iš 7542 asmenų, kuri šiame darbe yra laikoma populiacija $\mathcal{U}$. Ji sudaryta iš trijų sluoksnių: apie 1000 namų ūkių paimta iš prieš tai buvusio tyrimo, apie 1000 – iš dar anksčiau daryto tyrimo ir 1000 naujų namų ūkių. Šioje populiacijoje žinomos tikrosios bedarbių ir užimtųjų skaičių reikšmės, todėl jos gali būti lyginamos su įverčių, apskaičiuotų skirtingoms imtims, vidurkiais.

Iš kiekvieno populiacijos $\mathcal{U}$ sluoksnio išrenkama atsitiktinė negrąžintinė 40 lizdų (namų ūkių) imtis su tikimybėmis, proporcingomis lizdo dydžiui, t.y. 120 lizdų iš populiacijos. Iš gautos imties, naudojant imties sluoksniavimą, įvertinamas bedarbių ir užimtųjų gyventojų skaičius. 120 lizdų imtis renkama nepriklausomai 10000 kartų. Skaičiuojami du įverčiai: $\tilde{t}^{pos}$ (su kartotiniu persvėrimu pagal lytį ir miestą/kaimą) ir $\hat{\tilde{t}}^{pos}$ (su 5 kartus pakartotu persvėrimu). Gautų įverčių dispersijos laikomos tikrosiomis įverčių populiacijos dispersijomis. Poslinkiu laikomas gautų įverčių vidurkio ir tikrosios reikšmės skirtumas, VKP – vidutinė kvadratinė paklaida.

1 lentelė.

Bedarbių skaičiaus vertinimas

Metodas	Įverčio poslinkis	Standartinė paklaida	VKP įvertis
Imties sluoksniavimas 1 kartą			
Tikrosios reikšmės	0	113	12737
Jackknife metodas	−44	107	13357
Negrąžintinis bootstrap metodas	44	103	12599
Veidrodinio sutapimo bootstrap metodas	47	74	7745
Imties sluoksniavimas 5 kartus			
Tikrosios reikšmės	1	219	47931
Jackknife metodas	48	187	37176
Negrąžintinis bootstrap metodas	148	246	82548
Veidrodinio sutapimo bootstrap metodas	−141	947	916833

2 lentelė.

Užimtųjų skaičiaus vertinimas

Metodas	Įverčio poslinkis	Standartinė paklaida	VKP įvertis
Imties sluoksniavimas 1 kartą			
Tikrosios reikšmės	1	219	47931
Jackknife metodas	−13	218	47532
Negrąžintinis bootstrap metodas	46	196	40438
Veidrodinio sutapimo bootstrap metodas	20	137	19232
Imties sluoksniavimas 5 kartus			
Tikrosios reikšmės	442	653	621172
Jackknife metodas	644	742	965077
Negrąžintinis bootstrap metodas	1091	1232	2708229
Veidrodinio sutapimo bootstrap metodas	−1545	1501	4640682

5. Išvados

Jackknife metodu gauta VKP tik truputį didesnė už tikrąją, o negrąžintiniu bootstrap gauta VKP netgi truputį mažesnė už tikrąją. Tačiau, pakartojus imties sluoksniavimą 5 kartus, tiek sumos įverčio dispersija, tiek ir kartotinių imčių metodais gauti jos įverčiai išauga. Panagrinėjus tarpinius skaičiavimo rezultatus, pasirodo, kad formulėje (2.1) imties sluoksniavimo svorių $a_i b_i$ vidurkis artimas 1, bet imties sluoksniavimo svorių $a_i^{(1)} b_i^{(1)} \cdot \ldots \cdot a_i^{(5)} b_i^{(5)}$ vidurkis (formulėje (2.2)), atlikus kartotinį imties sluoksniavimą 5 kartus, siekia net 2, renkant imtį 10000 kartų, o negrąžintinio bootstrap atveju – kartais net 319. Todėl įverčiai ir jų dispersijos gaunasi labai dideli. Veidrodinio sutapimo bootstrap metodu gauti dispersijų įverčiai yra patenkinami, atliekant imties sluoksniavima 1 kartą, ir per dideli, atliekant imties sluoksniavimą 5 kartus.

Vadinasi, kartotinių imčių metodus galima naudoti įverčių dispersijų vertinimui baigtinėse populiacijose, darant kartotinį persvėrimą. Tačiau šie metodai netinka, darant kartotinį persvėrimą daugiau kartų. Reikėtų įsitikinti analiziškai, ar, atliekant kartotinį persvėrimą, procesas konverguoja.

Literatūra

[1] A.J. Canty, A.C. Davison, Resampling-based variance estimation for labour force surveys, *The Statistician*, **48**(3), 379–391 (1999).

[2] J.N.K. Rao, C.F.J. Wu, Resampling inference with complex survey data, *Journal of the American Statistical Association*, **83**(401), 231–241 (1988).

[3] J. Shao, D. Tu, *The Jackknife and Bootstrap*, Springer-Verlag, New York (1995).

[4] R.R. Sitter, Comparing three bootstrap methods for survey data, *The Canadian Journal of Statistics*, **20**(2), 135–154 (1992).

Resampling methods in survey sampling

D. Krapavickaitė, G. Šalučkienė

Aim of this paper is to estimate variances of the estimates of the population totals using resampling methods and to compare them with the known variances of the estimates.