Liet. matem. rink., T. 41, spec. nr., 444–450

Sumos vertinimas asimetriškoje populiacijoje

Danutė KRAPAVICKAITĖ (MII), Jurgita TURKUVIENĖ (VU)
el. paštas: krapav@ktl.mii.lt, jurgutet@takas.lt

1. Įvadas

Nagrinėjama baigtinė populiacija $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ su tyrimo kintamuoju y, kurio reikšmės $\{y_1, y_2, \ldots, y_N\}$ nežinomos ir gali būti sužinotos imties elementams. Nagrinėjamas toks kintamasis, kurio skirstinys populiacijoje yra asimetriškas, kartais atsitiktinai įgyja nulines reikšmes ir turi didelę populiacijos dispersiją. Vertinamas parametras yra populiacijos suma $t_y = \sum_{i=1}^{N} y_i$. Yra ir papildomas kintamasis x, charakterizuojantis populiacijos elementų dydį, koreliuotas su tyrimo kintamuoju, kurio reikšmės $\{x_1, x_2, \ldots, x_N\}$ yra žinomos visiems populiacijos elementams. Atliekamas tyrimas tikslu kuo tiksliau įvertinti tokio tyrimo kintamojo populiacijos sumą.

Yra įvairių metodų, kaip išrinkti imtį, įvertinti populiacijos sumą ir šio įvertinio dispersiją, atsižvelgiant į papildomo kintamojo reikšmes, kai populiacijos elementų dydžiai skirtingi. Kai kurie iš metodų: paprastoji atsitiktinė populiacijos elementų imtis ir santykinis sumos įvertinys; sistemingoji populiacijos elementų imtis su tikimybėmis, proporcingomis papildomo kintamojo dydžiui ([3]), Pareto imtis ([4]). Tai imties planu pagrįsti metodai. Gali būti naudojami ir populiacijos modeliu pagrįsti metodai ([2]).

Minėtomis savybėmis pasižyminti populiacija yra Lietuvos ūkininkų ūkiai su tyrimo kintamaisiais – įvairių žemės ūkio kultūrų pasėlių plotais. Lietuvos Statistikos Deparatmente atliekamo tyrimo tikslas – išrinkti atsitiktinę imtį ir, surinkus iš jos duomenis, įvertinti įvairių kultūrų pasėlių plotus Lietuvoje – populiacijos sumą. Ėmimo sąrašas – ūkininkų registras. Jame esanti informacija apie ūkinko turimą žemės ūkio naudmenų dydį – tai papildomas ūkio dydžio kintamasis, koreliuotas su kai kurių kultūrų pasėlių plotais.

Šis darbas skirtas žemės ūkio kultūrų pasėlių sumos įverčių modeliavimui, siekiant Lietuvos duomenims parinkti tinkamiausią imties išrinkimo ir vertinimo metodą. Modeliavimui naudojami Lietuvos Statistikos Departamente 1999 m. atlikto tyrimo imties duomenys. Tyrimo duomenys šiame darbe laikomi tyrimo populiacija, kurioje žinomos ir tikrosios parametrų reikšmės. Iš šios populiacijos renkamos atsitiktinės imtys ir vertinami parametrai bei gautų įverčių tikslumas.

2. Paprastoji atsitiktinė imtis

Paprastoji atsitiktinė imtis – tai tokia n skirtingų elementų imtis iš N dydžio baigtinės populiacijos, kai bet kuri n elementų kombinacija turi vienodą tikimybę būti išrinkta.

Daugelyje vadovėlių, pavyzdžiui, [1], rašoma apie tai, kad paprastojoje atsitiktinėje imtyje sumos įvertinys

$$\widehat{t}_{papr} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

yra nepaslinktasis. Jo dispersija yra

$$D\widehat{t} = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}, \quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \mu\right)^2, \quad \mu = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Šios dispersijos įvertinys $\widehat{D}\widehat{T} = N^2(1 - nN)\widehat{s}^2/n$ yra nepaslinktasis,

$$\widehat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Turint papildomą kintamąjį x, pakankamai koreliuotą su tyrimo kintamuoju, kurio reikšmės žinomos visiems populiacijos elementams, paprastojoje atsitiktinėje imtyje galima gauti ir tikslesnį asimptotiškai nepaslinktąjį santykinį sumos įvertinį $\widehat{t}_{sant} = t_x\bar{y}/\bar{x}$, $t_x = \sum_{i=1}^{N} x_i$. Jo apytikslė dispersija lygi

$$D\widehat{t}_{sant} = N^2\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}, \quad s_r^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - rx_i)^2, \quad r = \frac{t_y}{t_x}.$$

Ši dispersija gali būti vertinama nepaslinktuoju įvertiniu

$$\widehat{D}\widehat{t}_{sant} = N^2\left(1 - \frac{n}{N}\right)\frac{\widehat{s}_r^2}{n}, \quad \widehat{s}_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{r}x_i)^2, \quad \widehat{r} = \frac{\widehat{t}_y}{\widehat{t}_x}.$$

3. Imtis, išrinkta su tikimybėmis, proporcingomis papildomo kintamojo dydžiui

Vertinant populiacijos sumą, geresnį tikslumą gauti gali padėti ir tokios imtys, kuriose elementų priklausymo imčiai tikimybės yra proporcingos tų elementų dydžiui. Nagrinėsime dvi tokių fiksuoto dydžio imčių rūšis: sistemingąją ir Pareto. Šiose n dydžio imtyse elementų tikimybės priklausyti imčiai yra lygios

$$\pi_i = n\frac{x_i}{t_x}, \quad i = 1, 2, \ldots, N, \quad t_x = \sum_{j=1}^{N} x_j.$$

3.1. *Sistemingoji imtis, išrinkta su tikimybėmis, proporcingomis papildomo kintamojo dydžiui*

Surūšiavę populiacijos elementus atsitiktine tvarka, pažymėkime papildomo kintamojo x sukauptąsias sumas $v_0 = 0$, $v_i = \sum_{j=1}^{i} x_j$, $i = 1, \dots, N$. Pastebėsime, kad $v_N = t_x$. Susiekime intervalus $I_i = (v_{i-1}, v_i]$ su atitinkamais populiacijos elementais u_i. Norint išrinkti n dydžio atsitiktinę imtį, randamas skaičius $h = t_x/n$, kuris toliau naudojamas kaip ėmimo žingsnis. Sumodeliuojama tolygiai interavle $(0, h]$ pasiskirsčiusio atsitiktinio dydžio q_1 reikšmė ir apskaičiuojami dydžiai $q_i = q_1 + (i-1)h$, $i = 2, \dots, n$. Tarus, kad $x_i < t_x/n$, $i = 1, 2, \dots, N$, į imtį imami populiacijos elementai, atitinkantys tuos intervalus I_i, į kuriuos patenka dydžiai $q_1, q_2, \dots, q_n$. Tokiu būdu gautoji imtis vadinama imtimi, išrinkta su tikimybėmis $p_i = x_i/t_x$, porporcingomis papildomo kintamojo dydžiui. Tokioje imtyje

$$\widehat{t}_{sispp} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

yra nepaslinktasis populiacijos sumos įvertinys. Apskaičiavus

$$C_1 = \sum_{i=1}^{n} \frac{y_i \ln(1 - np_i)}{p_i^2} (1 - np_i),$$

$$C_2 = \sum_{i=1}^{n} \frac{\ln(1 - np_i)}{p_i} (1 - np_i), \quad A = \frac{C_1}{C_2},$$

gaumanas asimptotiškai nepaslinktasis įvertinio $\widehat{t}_{sispp}$ apytikslės dispersijos įvertinys ([3]):

$$\widehat{D\widehat{t}}_{sispp} = \frac{n}{n-1} \sum_{i=1}^{n} \left(\frac{y_i}{np_i} - A \right)^2 (1 - np_i).$$

3.2. *Pareto imtis*

Norint išrinkti n dydžio Pareto imtį, pirmiausiai apskaičiuojamos populiacijos elementų priklausymo imčiai tikimybės $\pi_1, \pi_2, \dots, \pi_N$. Po to modeliuojamos tolygiai intervale [0,1] pasiskirsčiusių atsitiktinių dydžių $\xi_1, \xi_2, \dots, \xi_N$ reikšmės, atitinkančios kiekvienam iš populiacijos elementų, ir apskaičiuojami dydžiai

$$q_i = \frac{\xi_i (1 - \pi_1)}{\pi_i (1 - \xi_i)}, \quad i = 1, 2, \dots, N.$$

Į imtį imami tie n populiacijos elementų, kuriems q_i reikšmės yra mažiausios.

Pareto imtyje sumos įvertinys

$$\widehat{t}_{Par} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

yra nepaslinktasis.

$$\widehat{D}\widehat{t}_{Par} = \frac{n}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i} - \frac{\sum_{j=1}^{n} y_j(1-\pi_j)/\pi_j}{\sum_{j=1}^{n}(1-\pi_j)}\right)^2 (1-\pi_i)$$

yra suderintasis jo apytikslės dispersijos įvertinys ([4]).

4. Tyrimo kintamojo reikšmių modeliavimas

Tyrimo kintamojo reikšmių sumą galima užrašyti kaip n esančių imtyje ir $N-n$ nesančių imtyje elementų sumą:

$$t = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} y_i.$$

Įvertinę imtyje nesančių elementų tyrimo kintamojo reikšmes, gausime sumos įvertinį

$$\widehat{t} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \widehat{y}_i.$$

4.1. *Lognormalusis modelis*

Tarkime, kad esant asimetriškam teigiamo tyrimo kintamojo y skirstiniui, jo reikšmių logaritmų $z_i = \ln x_i$, $i = 1, 2, \ldots, N$ sąlyginis skirstinys yra normalusis: $z_i | \boldsymbol{x}_i \sim N(\mu_i, \sigma^2)$, su vidurkiais $\mu_i = \mu(\boldsymbol{x}_i) = \alpha \boldsymbol{x}_i = \alpha_0 + \alpha_1 x_i$ ir vienodomis dispersijomis $\sigma^2 > 0$. Čia $\boldsymbol{x}_i = (1, x_i)^T$. Vektorius $\alpha = (\alpha_0, \alpha_1)$ ir dydis σ yra nežinomi. Pažymėkime vektorių $\boldsymbol{z} = (z_1, \ldots, z_n)$, matricą

$$\boldsymbol{x} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}.$$

Laikant, kad matrica $\boldsymbol{x}\boldsymbol{x}^T$ yra teigiamai apibrėžta, galima apibrėžti dydžius

$$a_{ij} = \boldsymbol{x}_i^T (\boldsymbol{x}\boldsymbol{x}^T)^{-1} \boldsymbol{x}_j, \quad i, j, = 1, 2, \ldots, n.$$

Tada

$$\widehat{t}_{logn} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} e^{\widehat{z}_i} \exp(\widehat{z}_i) \exp\left(\frac{\widehat{\sigma}^2}{2}(1 - a_{ii}) - \frac{\widehat{\sigma}^4}{4n}\right),$$

$$\widehat{z}_i = \widehat{\alpha} \boldsymbol{x}_i, \quad \widehat{\alpha} = (\boldsymbol{z}\boldsymbol{x}^T)(\boldsymbol{x}\boldsymbol{x}^T)^{-1}, \quad \widehat{\sigma} = \frac{\boldsymbol{z}\boldsymbol{z}^T - \widehat{\alpha}(\boldsymbol{x}\boldsymbol{x}^T)\widehat{\alpha}^T}{n-2}$$

yra apytiksliai nepaslinktasis sumos įvertinys ([2]). Šio įvertinio dispersijos įvertinys

$$\widehat{D}\widehat{T} = \sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}\omega_{ij}(\widehat{\alpha},\widehat{\sigma}),$$

$$D_{ij} = \exp\left\{ -\frac{\widehat{\sigma}^2}{2}(a_{ii}+a_{ij}+a_{ji}+a_{jj}) - \frac{\widehat{\sigma}^4}{n}\right\},$$

$$\omega_{ij}(\widehat{\alpha},\widehat{\sigma}) = \exp\{\widehat{\alpha}(\boldsymbol{x}_i+\boldsymbol{x}_j)+\widehat{\sigma}^2\}\left(\exp\left\{\frac{\widehat{\sigma}^2(a_{ij}+a_{ji})}{2}+\frac{\widehat{\sigma}^4}{2n}\right\}-1\right)+\delta_{ij}(\widehat{\alpha},\widehat{\sigma}),$$

$$\delta_{ij}(\widehat{\alpha},\widehat{\sigma}) = \begin{cases} \exp\{2\widehat{\alpha}\boldsymbol{x}_i+\widehat{\sigma}^2\}(\mathrm{e}^{\widehat{\sigma}^2}-1), & i=j,\\ 0, & i\neq j \end{cases}$$

yra nepaslinktasis.

4.2. *Lognormalusis – logistinis modelis*

Nors teigiamų tyrimo kintamojo reikšmių skirstinys populiacijoje ir yra lognormalusis, tačiau kartais tyrimo kintamasis gali įgyti ir reikšmes, lygias nuliui. Todėl nagrinėjamas populiacijos modelis $y_i = \widetilde{y}_i\Delta_i$ su lognormaliąja komponente $\widetilde{y}_i$ ir logistine komponente Δ_i. Δ_i – tai Bernulio skirstinį turintis atsitiktinis dydis, įgyjantis reikšmę 1 su tikimybe p_i ir reikšmę 0 su tikimybe $1-p_i$. Čia tikimybė p_i laikoma turinti pavidalą

$$p_i = P\{\Delta_i = 1|\boldsymbol{x}_i\} = P\{y_i > 0|\boldsymbol{x}_i\} = \frac{\exp(\beta\boldsymbol{x}_i)}{1+\exp(\beta\boldsymbol{x}_i)}.$$

n dydžio imtyje turint n^+ ($n^+ \leqslant n$) teigiamų tyrimo kintamojo reikšmių, įvertinama lognormalioji komponentė $\widehat{\widetilde{y}}_i$, $i = n+1,\dots,N$ (4.1 skyrelis). Pasinaudojus maksimalaus tikėtinumo metodu, įvertinamas logistinis parametras $\beta = (\beta_0,\beta_1)^T$ ir tikimybė

$$\widehat{p}_i = \frac{\exp(\widehat{\beta}\boldsymbol{x}_i)}{1+\exp(\widehat{\beta}\boldsymbol{x}_i)}.$$

Populiacijos sumos įvertinys

$$\widehat{t}_{logn-l} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \widehat{p}_i\widehat{\widetilde{y}}_i$$

yra apytiksliai nepaslinktasis [2].

5. Skaičiavimo rezultatai

Skaičiavimams naudojamos ūkininkų populiacijos dydis – 4066 ūkiai, turintys 88763 ha žemės ūkio naudmenų, iš kurių 17043 ha – žieminių kviečių pasėliai. Koreliacijos koefi-

1 lentelė. Žieminių kviečių pasėlių ploto vertinimas

Imties strategija	Imties dydis	Sumos įvertis	Poslinkio įvertis	Var. koef.	VKP sant.
Paprastoji	50	17738	695	43,3	1
atsitiktinė	100	17131	87	33,5	1
imtis	300	17040	-4	21,2	1
Papr. ats.	50	16957	-86	34,8	0,64
imtis & sant.	100	16839	-204	26,7	0,65
įvertinys	300	16947	-96	17,3	0,69
Sist. imtis	50	17067	24	26,3	0,37
su tik., prop.	100	17115	72	20,1	0,37
dydžiui	300	17063	20	13,6	0,34
Pareto	50	17139	96	26,4	0,37
imtis	100	17014	-29	19,9	0,38
	300	16971	-72	13,1	0,33
Lognormalusis	50	16519	-524	14,2	0,07
modelis	100	16130	-913	9,5	0,10
	300	16313	-730	4,6	0,08
Lognormalusis	50	14344	-2699	28,3	0,24
logistinis	100	14876	-2167	21.2	0,23
modelis	300	16759	-285	9,0	0,16

cientas tarp šių kintamųjų lygus 0,56. Tikslas – išrinkus imtį, įvertinti žieminių kviečių pasėlių plotą.

Išrinkta po 1000 paprastųjų atsitiktinių, sistemingųjų su tikimybėmis, proporcingomis žemės ūkio naudmenoms, ir Pareto imčių, kurių dydžiai $n = 50, 100$ ir 300. Kiekvienoje iš imčių skaičiuojamas atitinkamas sumos įvertis, paprastojoje atsitiktinėje imtyje – dar ir santykinis sumos įvertis žemės ūkio naudmenų atžvilgiu.

Taikant lognormalųjį modelį žieminių kviečių pasėlių plotams, buvo renkama 1000 to paties dydžio paprastųjų atsitiktinių imčių iš 1507 ūkių, auginusių šią kultūrą, populiacijos.

Taikant lognormalųjį – logistinį modelį, buvo renkama 100 paprastųjų atsitiktinių imčių ir visose jose vertinama suma. Suskaičiavus gautų įverčių vidurkį, dispersiją ir variacijos koeficientą, procesas kartojamas 20 kartų.

1-je lentelėje pateikiami parametrų įverčių, jų poslinkių įverčių, vidutinių kvadratinių paklaidų ir variacijos koeficientų, išreikštų procentais, vidurkiai. Vidutinių kvadratinių paklaidų (VKP) santykis – tai nagrinėjamoje strategijoje gaunamos VKP santykis su VKP paprastojoje atsitiktinėje imtyje.

6. Išvados

1-je lentelėje demonstruojama, kaip mažėja įverčio tikslumą charakterizuojantis variacijos koeficientas, augant imties dydžiui, nulemtas mažėjančios įverčio dispersijos. Imties dydžio įtaka įverčio poslinkiui pastebima paprastojoje atsitiktinėje imtyje ir lognormaliojo – logistinio modelio atveju, kur nesinaudojama įvertinio dispersijos įvertiniu. Sumos įverčiai, gauti, naudojant modelius, yra paslinktieji. To priežastimi yra, gal būt, nepakankamas duomenų atitikimas normaliajam skirstiniui.

VKP santykis parodo papildomos informacijos panaudojimo vertinimo etape naudingumą. Santykinis įvertinys paprastojoje atsitiktinėje imtyje duoda tikslesnius rezultatus negu nepaslinktasis, o imtyse, išrinktose su tikimybėmis, proporcingomis dydžiui, įverčių tikslumas didesnis negu paprastojoje atsitiktinėje imtyje su santykiniu įvertiniu. Sumos įverčio tikslumas sistemingojoje imtyje, išrinktoje su tikimybėmis, proporcingomis dydžiui ir Pareto imtyje pastebimai nesiskiria.

Tiksliausias įvertis gaunamas, naudojant lognormalųjį modelį, tačiau šis įvertis yra ne visiškai palyginamas su kitais įverčiais, kadangi naudojasi papildoma prielaida apie tyrimo kintamojo reikšmių teigiamumą, ir tik tokie duomenys buvo naudojami jo gavimui.

Mažiausią vidutinę kvadratinę paklaidą duoda lognormaliojo – logistinio modelio panaudojimas.

Literatūra

[1] W.G. Cochran, *Sampling Techniques*, 3 rd ed., New York: Wiley & Sons (1977).

[2] F.Karlberg, *Survey Estimation for Highly Skewed Data*, Department of Sattistics, Stockholm University (1999).

[3] B.Rosén, *Variance Estimation for Systematic PPS Sampling*, R&D Report, Statistics Sweden (1991).

[4] B.Rosén, *On Sampling with Probability Proportional to Size*, R&D Report, Statistisc Sweden (1996).

Estimation of total in skewed population

D. Krapavickaitė, J. Turkuvienė

Aim of this paper is to estimate total of the crop of winter wheat in the skewed population of farmers in Lithuania using different sampling strategies, to compare the estimates with the true value, to compare the methods with each other and to find out the most suitable strategy for the estimation of the total.