Liet. matem. rink., T. 42, spec. nr., 541–547

Kvantilio vertinimas baigtinėje populiacijoje

Danutė KRAPAVICKAITĖ (MII, VGTU)
el. paštas: krapav@ktl.mii.lt

1. Įvadas

Nagrinėjama baigtinė populiacija $\mathcal{U} = \{1, 2, \ldots, N\}$ su tyrimo kintamuoju y, kurio reikšmės $\{y_1, y_2, \ldots, y_N\}$ nežinomos ir gali būti sužinotos imties elementams. Populiacijos parametras $\theta = \theta(y_1, y_2, \ldots, y_N)$ taip pat nežinomas. Jį reikia įvertinti. Tuo tikslu renkama tikimybinė imtis $\mathrm{i} \subset \mathcal{U}$, kurios elementus paprastumo dėlei žymėsime tiesiog $\mathrm{i} = \{1, 2, \ldots, n\}$. Dažniausiai vertinamas populiacijos parametras yra suma

$$t = t_y = y_1 + y_2 + \ldots + y_N.$$

Norint ją įvertinti, turint tikimybinę imtį i, galima pasinaudoti Horvitz–Thompson sumos įvertiniu ([2]):

$$\widehat{t}_y = \sum_{k \in \mathrm{i}} \frac{y_k}{\pi_k}.$$

Jis yra nepaslinktasis. Jo dispersija

$$\mathrm{D}\widehat{t}_y = \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

gali būti vertinama nepaslinktuoju įvertiniu

$$\widehat{\mathrm{D}}\widehat{t}_y = \sum_{k \in \mathrm{i}} \sum_{l \in \mathrm{i}} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

pažymėjus elementų priklausymo imčiai tikimybes $\pi_k = P\{k \in \mathrm{i}\}$, $\pi_{kl} = P\{k \in \mathrm{i},\ l \in \mathrm{i}\}$.

Literatūroje retai nagrinėjamas, bet praktikoje naudojamas populiacijos parametras yra tyrimo kintamojo y pasiskirstymo F bet kurio q lygio, $0 < q < 1$, kvantilis $Q(q)$. Atskiras kvantilio – medianos ($q = 1/2$) ir jos pasikliautinojo intervalo vertinimo atvejis išnagrinėtas [2]. Kai kurios kvantilių funkcijų vertinimo galimybės nagrinėtos [1].

Šiame darbe siūlomas kvantilio – dviejų sumų santykio įvertinys ir kvantilio pasikliautinojo intervalo įvertinys. Šie įvertiniai gali būti taikomi Lietuvos oficialiosios statistikos vykdomame namų ūkių biudžetų tyrime.

2. Kvantilis

Kvantiliu vadinamas toks skaičius $Q(q)$, $0 < q < 1$, kuris tenkina nelygybes

$$F(Q(q)) \leqslant q \leqslant F(Q(q)+0).$$

Kvantilis gali būti lygties

$$F(Q(q)) = q \tag{1}$$

sprendiniu. Tačiau tyrimo kintamojo reikšmių pasiskirstymo funkcija baigtinėje populiacijoje yra laiptuota, ir (1) lygtis turi arba be galo daug sprendinių, arba jų visai neturi. Tyrimo kintamojo reikšmės populiacijoje surūšiuojamos didėjimo tvarka: $y_1^\star \leqslant y_2^\star \leqslant \ldots \leqslant y_N^\star$. Jei egzistuoja toks $k_0 \in \mathcal{U}$, kad $F(y_{k_0}^\star) = q$, tai kvantilis $Q(q) = (y_{k_0+1}^\star + y_{k_0}^\star)/2$. Jei $F(y_k^\star) \neq q$ su visais $k \in \mathcal{U}$, tai atsiras toks k_0, kuriam $F(y_{k_0-1}^\star) < q$, tačiau $F(y_{k_0}^\star) > q$. Tada kvantilis $Q(q) = y_{k_0}^\star$.

Norint gauti kvantilio įvertinį $\widehat{Q}(q)$, reikia nagrinėti populiacijos pasiskirstymo funkcijos $F(x)$ įvertinį $\widehat{F}(x)$, tenkinantį nelygybes

$$\widehat{F}(\widehat{Q}(q)) \leqslant q \leqslant \widehat{F}(\widehat{Q}(q)+0).$$

Imties reikšmės rūšiuojamos didėjimo tvarka: $y_1^\star \leqslant y_2^\star \leqslant \ldots \leqslant y_n^\star$. Kvantilio įvertinys bus

$$\widehat{Q}(q) = \begin{cases} (y_{k_0+1}^\star + y_{k_0}^\star)/2, & \text{jei yra } k_0 \in \mathbf{i}\text{: } \widehat{F}(y_{k_0}^\star) = q, \\ y_{k_0}^\star, & \text{jei } \widehat{F}(y_{k_0-1}^\star) < q,\ \widehat{F}(y_{k_0}^\star) > q \end{cases}$$

kokiam nors $k_0 \in \mathbf{i}$.

Pasiskirstymo funkciją galima užrašyti

$$F(x) = \#A_x/N, \tag{2}$$

čia $A_x = \{k \in \mathcal{U}\text{: } y_k \leqslant x\}$, ir $\#A_x$ reiškia aibės A_x elementų skaičių. Norint įvertinti kvantilį, reikia įvertinti pasiskirstymo funkciją, o ji gali būti užrašyta ir kitaip. Įveskime papildomą kintamąjį $z(x) = \{z_1(x), \ldots, z_N(x)\}$, čia

$$z_k(x) = \begin{cases} 1, & \text{jei } y_k \leqslant x, \\ 0, & \text{jei } y_k > x, \end{cases}$$

$k = 1, 2, \ldots, N$. Tada

$$F(x) = t_{z(x)}/N, \tag{3}$$

ir pasiskirstymo funkcijos (2) vertinimo uždavinys susiveda į dviejų sumų santykio vertinimo uždavinį:

$$\widehat{F}(x) = \frac{\widehat{t}_{z(x)}}{\widehat{N}}, \quad \widehat{t}_{z(x)} = \sum_{k \in \mathbf{i}} \frac{z_k(x)}{\pi_k}, \quad \widehat{N} = \sum_{k \in \mathbf{i}} \frac{1}{\pi_k}.$$

3. Kvantilio pasikliautinasis intervalas

Radus konstantas $c_1 > 0$ ir $c_2 > 0$ tokias, kad

$$P\{c_1 \leqslant \widehat{F}(Q(q)) \leqslant c_2\} \geqslant 1 - \alpha,$$

intervalas $\left(\widehat{F}^{-1}(c_1), \widehat{F}^{-1}(c_2)\right)$ bus α lygio $(0 < \alpha < 1)$ pasikliautinasis intervalas kvantiliui $Q(q)$. Jei santykio įvertinys $\widehat{F}(Q(q))$ turi normalųjį skirstinį su vidurkiu q, tai

$$\begin{aligned} c_1 &= q - Q_{N(0,1)}(\alpha/2)\sqrt{\mathbf{D}\widehat{F}(Q(q))}, \\ c_2 &= q + Q_{N(0,1)}(\alpha/2)\sqrt{\mathbf{D}\widehat{F}(Q(q))}, \end{aligned} \tag{4}$$

čia $Q_{N(0,1)}(\alpha/2)$ yra standartinio normaliojo skirstinio $1 - \alpha/2$ lygio kvantilis.

$\mathbf{D}\widehat{F}(Q(q))$ apytikslę dispersiją užrašome, naudojantis santykio įvertinio apytikslės dispersijos formule, gaunama, skleidžiant $\mathbf{D}\widehat{F}(Q(q))$ Teiloro eilute:

$$A\mathbf{D}\widehat{F}(Q(q)) = \frac{1}{N^2}\mathbf{D}\left(\widehat{t}_{z(Q(q))} - q\widehat{N}\right). \tag{5}$$

Praktiškai šios apytikslės dispersijos panaudoti negalime, nes ji priklauso nuo visų populiacijos reikšmių, todėl ją dar reikia įvertinti. Įverčiu naudojama

$$\widehat{\mathbf{D}}\widehat{F}(Q(q)) = \frac{1}{N^2}\widehat{\mathbf{D}}(\widehat{t}_{z(\widehat{Q}(q))} - q\widehat{N}).$$

Tokiu būdu iš (4) gauname q lygio kvantilio $Q(q)$ pasikliautinąjį intervalą (α lygio)

$$\left(\widehat{F}^{-1}\left(q - Q_{N(0,1)}(\alpha/2)\sqrt{\widehat{\mathbf{D}}\widehat{F}(Q(q))}\right),\right.$$
$$\left.\widehat{F}^{-1}\left(q + Q_{N(0,1)}(\alpha/2)\sqrt{\widehat{\mathbf{D}}\widehat{F}(Q(q))}\right)\right). \tag{6}$$

4. Pavyzdys

Pajamų vienam namų ūkio nariui pasiskirstymo kvantilių ir jų pasikliautinųjų intervalų vertinimas namų ūkių biudžetų tyrime

Šiame tyrime naudojama sluoksninė imtis su dviejų etapų lizdine imtimi sluoksniuose. Pirmąjį sluoksnį sudaro trijų didžiųjų miestų: Vilniaus, Kauno, Klaipėdos, Šiaulių ir Panevėžio namų ūkiai. Antrajam priklauso vidutinių ir mažųjų miestų gyventojų namų ūkiai, ir trečiajam – kaimo gyventojų namų ūkiai. Tarkime, kad visi gyventojai yra įtraukti į gyventojų registrą. 10 680 namų ūkių imtis paskirstoma į sluoksnius proporcingai jų dydžiui. Imčių planai sluoksniuose yra tokie:

1-*asis sluoksnis*. Iš gyventojų registro renkama paprastoji atsitiktinė 4 476 asmenų imtis. Tiriami išrinktųjų asmenų namų ūkiai.

2-*asis sluoksnis*. Vidutinių ir mažųjų miestelių gyventojai apjungiami į lizdus, turinčius nuo 1 000 iki 10 000 gyventojų. 1-ajame etape su tikimybėmis, proporcingomis lizdo dydžiui, renkama grąžintinė 20 lizdų imtis iš 140 populiacijoje esančių lizdų. 2-ajame etape kiekviename iš išrinktųjų lizdų iš gyventojų registro renkama paprastoji atsitiktinė 132 asmenų imtis. Jų namų ūkiai laikomi išrinktais į imtį.

3-*asis sluoksnis*. Suformuojami 463 lizdai, kuriuose yra nuo 300 iki 2 000 namų ūkių. 1-ajame etape su tikimybėmis, proporcingomis lizdo dydžiui, išrenkama grąžintinė 33 lizdų imtis. 2-ajame etape išrinktuosiuose lizduose iš gyventojų registro renkama paprastoji atsitiktinė 108 asmenų imtis. Jų namų ūkiai laikomi išrinktais į imtį.

Tyrimo kintamasis y – asmens pajamos. Jų pasiskirstymo funkcija $F(x)$, pasinaudojus (3), gali būti užrašyta tokiu būdu:

$$F(x) = \frac{t_{z(x)}}{N} = \sum_{h=1}^{3} \frac{N_h}{N} \frac{t_{z(x)}^{(h)}}{N_h} = \sum_{h=1}^{3} \frac{N_h}{N} F_h(x),$$

čia N – populiacijos dydis ėmimo sąraše (gyventojų registre), N_h – sluoksnio dydis ėmimo sąraše, $F_h(x)$ – pajamų pasiskirstymo funkcija h-ajame sluoksnyje, $t_{z(x)}^{(h)}$ – kintamojo $z(x)$ reikšmių suma h-ajame sluoksnyje. Kvantilis vertinamas, įprastiniu būdu nagrinėjant funkciją $F(x)$.

Kvantilis, būdamas (1) lygties sprendiniu, tuo pačiu yra lygties

$$\sum_{h=1}^{3} \frac{N_h}{N} \widehat{F}_h(Q(q)) = q$$

sprendiniu. Norint rasti kvantilio pasikliautinąjį intervalą (6), reikia skaičiuoti funkcijos $\widehat{F}(Q(q)) = \sum_{h=1}^{3} \frac{N_h}{N} \widehat{F}_h(Q(q))$ apytikslę dispersiją

$$A\mathrm{D}\widehat{F}(Q(q)) = A\mathrm{D}\left(\sum_{h=1}^{3} \frac{N_h}{N} \widehat{F}_h(Q(q)) \right) = \sum_{h=1}^{3} \frac{N_h^2}{N^2} A\mathrm{D}\widehat{F}_h(Q(q)). \qquad (7)$$

Apytikslės dispersijos $AD\widehat{F}(Q(q))$ skaičiavimas

1-*ajame sluoksnyje* elemento tikimybė priklausyti imčiai $\pi_{1k} = n_1 m_{1k}/N_1$, čia n_1 – 1-ojo sluoksnio imties dydis, m_k – k-ojo namų ūkio registruotų asmenų skaičius, $k \in \mathcal{U}_1$, $\mathcal{U}_1$ – 1-ojo sluoksnio gyventojų populiacija. Tikimybė, kad k-sis ir l-sis namų ūkiai kartu priklausys imčiai, yra

$$\pi_{1kl} = \frac{n_1(n_1-1)m_{1k}m_{1l}}{N_1(N_1-1)}.$$

Tada, pasinaudojus Horvitz–Thompson sumos įvertiniu ir (5), gauname

$$\begin{aligned} AD\widehat{F}_1(Q(q)) &= \frac{1}{N_1^2}\mathbf{D}\big(\widehat{t}^{(1)}_{z(Q(q))} - F_1(Q(q))\widehat{N}_1\big) \\ &= \frac{1}{N_1^2}\sum_{k\in\mathcal{U}_1}\sum_{l\in\mathcal{U}_1}(\pi_{1kl}-\pi_{1k}\pi_{1l})\frac{z_{1k}(Q(q)) - F_1(Q(q))}{\pi_{1k}}\cdot\frac{z_{1l}(Q(q)) - F_1(Q(q))}{\pi_{1l}} \\ &= -\frac{1}{N_1^2}\Big(1+\frac{1}{N_1-1}\Big)\Big(1-\frac{n_1}{N_1}\Big)\frac{1}{n_1}\sum_{k\in\mathcal{U}_1}\sum_{l\in\mathcal{U}_1}u_{1k}u_{1l}, \end{aligned} \tag{8}$$

čia $u_{1k} = z_k(Q(q)) - F_1(Q(q))$ yra kintamojo $u_1 = z(Q(q)) - F_1(Q(q))$ reikšmė.

2-*jame ir* 3-*jame sluoksniuose* turima dviejų etapų lizdinė imtis. I etape – grąžtinė lizdinė imtis su lizdų išrinkimo tikimybėmis $p_{hi} = L_{hi}/\sum_{j\in\mathcal{U}_h} L_{hj}$, imties dydis n_h, todėl lizdo priklausymo imčiai tikimybė

$$\pi_{hi} = \frac{n_h L_{hi}}{\sum_{j\in\mathcal{U}_h} L_{hj}},$$

$\mathcal{U}_h$ – lizdų populiacija h-ajame sluoksnyje, $h = 2, 3$, L_{hi} – registruotų asmenų skaičius lizde.

II etapo lizdas – namų ūkis. h-ojo sluoksnio i-ojo lizdo k-ojo namų ūkio tikimybė priklausyti imčiai yra

$$\pi_{hik} = \frac{\widetilde{n}_{hi} m_{hik}}{L_{hi}},$$

čia $h = 2, 3$, $i \in \mathcal{U}_h$, $h = 2, 3$, $\widetilde{n}_{hi}$ – II etapo imties dydis h-ojo sluoksnio i-ajame lizde, $k = 1, 2, \ldots, m_{hik}$, m_{hik} – asmenų skaičius i-ojo lizdo k-ajame namų ūkyje.

Pažymėjus $\mathbf{i}_h$ I etapo imtį h-ajame sluoksnyje, parametro θ įvertiniui $\widehat{\theta}$ galioja lygybė

$$\mathbf{D}\widehat{\theta} = \mathbf{D}(E(\widehat{\theta}|\mathbf{i}_h)) + E(\mathbf{D}(\widehat{\theta}|\mathbf{i}_h)).$$

Ja naudojamasi, skaičiuojant dispersiją 2-ajame ir 3-ajame sluoksniuose:

$$\begin{aligned} &\mathbf{D}\Big(\widehat{t}^{(h)}_{z(Q(q))} - F_h(Q(q))\widehat{N}_h\Big) \\ &\quad= \mathbf{D}\Big(E\big(\widehat{t}^{(h)}_{z(Q(q))} - F_h(Q(q))\widehat{N}_h|\mathbf{i}_h\big)\Big) + E\Big(\mathbf{D}\big(\widehat{t}^{(h)}_{z(Q(q))} - F_h(Q(q))\widehat{N}_h|\mathbf{i}_h\big)\Big). \end{aligned}$$

2-ojo ir 3-ojo sluoksnių I etape imtį laikysime grąžintine, išrinkta su tikimybėmis, proporcingomis lizdo dydžiui. Tada sumai vertinti I etape galima naudotis Hansen–Hurwitz sumos įvertiniu ([2]). II etape galima naudotis tokiu pat Horvitz–Thompson sumos įvertiniu, kaip ir 1-ajame sluoksnyje. Sukombinavus šiuos įvertinius, gaunama pasiskirstymo funkcijos įvertinio apytikslės dispersijos sluoksnyje išraiška

$$\begin{aligned} A\mathrm{D}\widehat{F}_h(Q(q)) &= \frac{1}{N_h^2}\mathrm{D}\big(\widehat{t}^{(h)}_{zQ(q)} - F_h(Q(q))\widehat{N}_h\big) \\ &= \frac{1}{N_h^2 n_h}\sum_{i\in\mathcal{U}_h}\Bigg[p_{hi}\Big(\frac{\widehat{t}^{(h)}_{ui}}{p_{hi}} - t^{(h)}_u\Big)^2 \\ &\quad - \frac{1}{\widetilde{n}_{hi}L_{hi}^2 p_{hi}}\Big(1-\frac{\widetilde{n}_{hi}}{L_{hi}}\Big)\Big(1+\frac{1}{L_{hi}-1}\Big)\sum_{k\in\mathcal{U}_{hi}}\sum_{l\in\mathcal{U}_{hi}} u_{hik}u_{hil}\Bigg], \end{aligned} \tag{9}$$

$p_{hi} = L_{hi}/\sum_{j\in U_h} L_{hj}$, $t^{(h)}_u = \sum_{i,k} u_{hik}$, $u_{hik} = z_{hik}(Q(q)) - F_h(Q(q))$, $\widehat{t}^{(h)}_{ui} = \sum_{k\in \mathrm{i}_{hi}} u_{hik}/\pi_{hik}$, $h = 2,3$, i_{hi} – imtis h-ojo sluoksnio i-ajame lizde. $\mathcal{U}_{hi}$ – h-ojo sluoksnio i-ojo lizdo namų ūkių populiacija, z_{hik} – kintamojo $z(x)$ reikšmė h-ojo sluoksnio i-ojo lizdo k-ajam asmeniui.

Apytikslė pasiskirstymo funkcijos įvertinio dispersija, įstačius dispersijų išraiškas sluoksniuose (8) ir (9) į (7), tampa

$$\begin{aligned} A\mathrm{D}\widehat{F}(Q(q)) = \frac{1}{N^2}\Bigg\{ &-\frac{1}{N_1^2}\Big(1+\frac{1}{N_1-1}\Big)\Big(1-\frac{n_1}{N_1}\Big)\frac{1}{n_1}\sum_{k\in\mathcal{U}_1}\sum_{l\in\mathcal{U}_1} u_{1k}u_{1l} \\ &+\sum_{h=2}^{3}\frac{1}{n_h}\sum_{i\in\mathcal{U}_h}\Bigg[p_{hi}\Big(\frac{\widehat{t}_{u_{hi}}}{p_{hi}} - t_{u_h}\Big)^2 \\ &-\frac{1}{\widetilde{n}_{hi}L_{hi}^2 p_{hi}}\Big(1-\frac{\widetilde{n}_{hi}}{L_{hi}}\Big)\Big(1+\frac{1}{L_{hi}-1}\Big)\sum_{k\in\mathcal{U}_{hi}}\sum_{l\in\mathcal{U}_{hi}} u_{hik}u_{hil}\Bigg]\Bigg\}. \end{aligned}$$

Pasiskirstymo funkcijos įvertinio dispersijos įvertinys

$$\begin{aligned} \widehat{\mathrm{D}}\widehat{F}(Q(q)) = \frac{1}{N^2}\Bigg\{ &-\frac{1}{N_1^2}\Big(1+\frac{1}{N_1-1}\Big)\Big(1-\frac{n_1}{N_1}\Big)\frac{1}{n_1}\sum_{k\in\mathrm{i}_1}\sum_{l\in\mathrm{i}_1}\frac{1}{\pi_{kl}}\widehat{u}_{1k}\widehat{u}_{1l} \\ &+\sum_{h=2}^{3}\frac{1}{n_h(n_h-1)}\sum_{i\in\mathrm{i}_h}\Bigg[\Big(\frac{\widehat{t}^{(h)}_{u_i}}{p_{hi}} - \widehat{t}^{h}_u\Big)^2 \\ &-\frac{1}{\widetilde{n}_{hi}L_{hi}^2 p_{hi}}\Big(1-\frac{\widetilde{n}_{hi}}{L_{hi}}\Big)\Big(1+\frac{1}{L_{hi}-1}\Big)\sum_{k\in\mathrm{i}_{hi}}\sum_{l\in\mathrm{i}_{hi}}\frac{1}{\pi_{kl}}\widehat{u}_{hik}\widehat{u}_{hil}\Bigg]\Bigg\}, \end{aligned}$$

$\widehat{u}_{hik} = z_{hik}(\widehat{Q}(q)) - \widehat{F}_h(\widehat{Q}(q))$, $h = 2,3$, $\widehat{u}_{1k} = z_k(\widehat{Q}(q)) - \widehat{F}_1(\widehat{Q}(q))$, įstatytas į (5), duoda kvantilio apytikslį pasikliautinąjį intervalą.

Autorė dėkoja VGTU studentei Jurgitai Čimžaitei, dalyvavusiai ruošiant šį darbą.

Literatūra

[1] T. Orusild, Confidence intervals for functions of quantiles under finite population sampling, *Central Statistical Bureau of Latvia, Workshop for the Baltic Countries on Survey Sampling Theory and Methodology*, Jurmala, Latvia (1998), pp. 51–55.

[2] C.–E. Sarndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer–Verlag, New–York (1992).

Estimation of a quantile in finite population

D. Krapavickaitė

An estimator of the quantile of a study variable distribution function in finite population and its confidence interval is proposed. The estimator is adapted to the two-stage cluster sample which is used in a household budget survey carried out in Lithuania.