



VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

Ieva Dirdaitė

**SUBALANSUOTO ĖMIMO IR KALIBRUOTO
ĮVERTINIO TAIKYMAS
APPLICATION OF BALANCED SAMPLING AND
CALIBRATED ESTIMATOR**

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 621G31001

Statistinių metodų finansuose ir ekonomikoje specializacija

Statistikos studijų kryptis

Vilnius, 2016

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

TVIRTINU
Katedros vedėjas

(Parašas)

Doc. dr. Kazimieras Padvelskis

(Data)

Ieva Dirdaitė

SUBALANSUOTO ĖMIMO IR KALIBRUOTO ĮVERTINIO
TAIKYMAS
APPLICATION OF BALANCED SAMPLING AND CALIBRATED
ESTIMATOR

Baigiamasis magistro darbas

Taikomosios statistikos studijų programa, valstybinis kodas 621G31001

Statistinių metodų finansuose ir ekonomikoje specializacija

Statistikos studijų kryptis

Vadovas: doc. dr. Danutė Krapavickaitė

(Parašas)

(Data)

Lietuvių kalbos konsultantas: dr. Vaida Buivydienė

(Parašas)

(Data)

Vilnius, 2016

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

Statistikos studijų kryptis

Taikomosios statistikos studijų programa, valstybinis kodas 621G31001

Statistinių metodų finansuose ir ekonomikoje specializacija

TVIRTINU

Katedros vedėjas

(Parašas)

Doc. dr. Kazimieras Padvelskis

(Data)

**BAIGIAMOJO MAGISTRO DARBO
UŽDUOTIS**

.....Nr.

Vilnius

Studentei Ievai Dirdaitei

Baigiamojo darbo tema: Subalansuoto ėmimo ir kalibruoto įvertinio taikymas

patvirtinta 2015 m. spalio 28 d. dekanų potvarkiu Nr. 379 fm.

Baigiamojo darbo užbaigimo terminas 2016 m. birželio 3 d.

BAIGIAMOJO DARBO UŽDUOTIS:

Papildomos informacijos naudojimas renkant imtį ar vertinant parametrus gali pagerinti gautamų įvertinių tikslumą, jei papildomi kintamieji yra koreliuoti su tyrimo kintamuoju. Modeliavimo būdu ištirti subalansuoto ėmimo ir kalibruoto įvertinio įtaką sumos įvertinimo tikslumui, kai duomenis pateikė visi imties elementai ir kai apklausoje yra neatsakymų. Modeliavimui naudoti Lietuvos statistikos departamento gyventojų užimtumo statistinio tyrimo duomenis.

Baigiamojo darbo rengimo konsultantai:

dr. Vaida Buivydienė.

Vadovas:
(Parašas)

doc. dr. Danutė Krapavickaitė

Užduotį gavau

.....
(Parašas)

Ieva Dirdaitė

2014 m. spalio 10

Vilniaus Gedimino technikos universitetas
Fundamentinių mokslų fakultetas
Matematinės statistikos katedra

ISBN ISSN
Egz. sk.
Data-.....-.....

Antrosios pakopos studijų **Taikomosios statistikos** programos magistro baigiamasis darbas 4

Pavadinimas **Subalansuoto ėmimo ir kalibruotojo įvertinio taikymas**
Autorius **Ieva Dirdaitė**
Vadovas **Danutė Krapavickaitė**

Kalba: lietuvių

Anotacija

Baigiamajame magistro darbe nagrinėjamas subalansuotas ėmimas taikant kubo metodą bei kalibruotasis sumos įvertinys. Lyginami septynių strategijų, kurios yra subalansuoto ėmimo (naudojant kubo metodą) bei paprastojo atsitiktinio lizdinio ėmimo ir Horvitco ir Tompsono bei kalibruotojo įvertinio kombinacijos, rezultatai. Visais atvejais bent viename iš etapų (imties ėmimo, parametrų vertinimo) yra naudojama papildoma informacija, kurią galima žinoti dar neatlikus apklausos (asmens lytis, amžius, gyvenamoji vieta). Modeliavimui naudojami realūs duomenys (Lietuvos statistikos departamento gyventojų užimtumo statistinio tyrimo (2011) duomenys). Sumos įverčių tikslumui palyginti skaičiuojami santykinės vidutinės kvadratinės paklaidos, poslinkio, dispersijos, variacijos koeficiento įverčiai.

Darbą sudaro 7 dalys: lentelių sąrašas, paveikslų sąrašas, įvadas, teorinė dalis, praktinė dalis, išvados, literatūros sąrašas.

Darbo apimtis - 39 p. teksto be priedu, 3 paveikslai, 13 lentelių, 31 bibliografiniai šaltiniai.

Atskirai pridedami darbo priedai.

Prasminiai žodžiai: subalansuotas ėmimas, kubo metodas, kalibruotas įvertinys.

Vilnius Gediminas Technical University
Faculty of Fundamental Sciences
Department of Mathematical Statistics

ISBN _____ ISSN _____
Copies No.
Date-.....-.....

Master Degree Studies **Applied Statistics** study programme Master Graduation Thesis 4

Title **Application of balanced sampling and calibrated estimator**

Author **Ieva Dirdaitė**

Academic supervisor **Danutė Krapavickaitė**

Thesis language: Lithuanian

Annotation

Application of balanced sampling together with the cube method and calibrated estimator for the total of a study variable defined on a finite population is studied in the final master thesis. Seven different strategies which are combinations of balanced sampling, simple random cluster sampling, Horwitz-Thompson estimator and calibrated estimator are proposed in a paper. Auxiliary information is used in at least one stage (sampling, estimation) of all strategies. Auxiliary information used is known before the survey (gender, age, living place). Real data (Labor force survey data (2011) of Statistics Lithuania) is used for simulation. The estimates of the relative root mean squared error, bias, variance, the coefficient of variation are used to compare accuracy of the estimates obtained.

Structure: list of tables, list of figures, introduction, theoretical part, experimental part, conclusions, references.

Thesis consists of: 39 p. text without appendixes, 3 figures, 13 tables, 31 bibliographical entries.

Appendixes are included.

Keywords: balanced sampling, cube method, calibrated estimator.

Turinys

Įvadas	10
1. Pagrindinės imčių teorijos sąvokos	11
2. Subalansuotas ėmimas	13
2.1. Balansavimo kintamieji esant fiksuotam imties dydžiui	15
2.2. Subalansuotas sluoksninės imties planas	16
2.3. Balansavimo algoritmai	17
2.3.1. Deville'io, Grosbras'o ir Roth'o balansavimo algoritmas lygioms įtrauk-	
ties tikimybėms	17
2.3.2. Subalansuotas ėmimas naudojantis tiesiniu programavimu	18
3. Kubo metodo etapai	18
3.1. Pakilimo fazė	19
3.2. Nusileidimo fazė	21
4. Imties plano svorių kalibravimas	22
5. Logistinės regresijos modelio taikymas praleistoms kintamojo reikšmėms įrašyti	24
6. Statistinis modeliavimas	26
6.1. Tyrimo populiacija	26
6.2. Strategijos	27
6.2.1. Subalansuotas lizdinis ėmimas ir Horvico ir Tompsono įvertinys	28
6.2.2. Paprastas atsitiktinis lizdinis ėmimas ir kalibruotasis įvertinys	30
6.2.3. Subalansuotas lizdinis ėmimas ir kalibruotasis įvertinys	33
6.2.4. Subalansuotas lizdinis ėmimas, kai yra neatsakymų, ir kalibruotasis įver-	
tinys	33
6.2.5. Paprastoji lizdinė atsitiktinė imtis, kai yra neatsakymų, ir kalibruotasis	
įvertinys	35
6.2.6. Subalansuotas lizdinis ėmimas kai yra neatsakymų, ir Horvico ir Tomp-	
sono įvertinys	35
6.2.7. Subalansuotas lizdinis ėmimas su Horvico ir Tompsono įvertiniu, kai yra	
neatsakymų, ir logistinė regresija	36
Išvados	39
Literatūra	40
A. Priedas: <i>balancedcluster</i> funkcijos komentaras	43

B. Priedas: Naudoti R programos kodai	43
B.1. Subalansuotos lizdinės imties ėmimas su Horvico ir Tompsono įvertiniu	43
B.2. Paprastasis atsitiktinis lizdinis ėmimas ir kalibruotasis įvertinys	47
B.3. Tikimybių perskaičiavimas, kai yra neatsakymų	51
B.4. Neatsakymų į apklausą modeliavimas, taikant logistinę regresiją ir Bernulio atsitiktinius dydžius	53

Lentelių sąrašas

1.	10 subalansuotų imčių balansuojančių kintamųjų sumų įverčiai, kai $m = 5\,000$	29
2.	10 subalansuotų imčių tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 5\,000$	29
3.	10 subalansuotų imčių balansuojančių kintamųjų sumų įverčiai, kai $m = 100$, $m = 1\,000$	29
4.	10 subalansuotų imčių tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$	30
5.	10 paprastųjų atsitiktinių imčių kalibruotojo tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$	33
6.	10 subalansuotų lizdinių imčių kalibruoto tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$	34
7.	10 subalansuotų lizdinių imčių, kai yra neatsakymų, kalibruoto tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$	34
8.	10 paprastųjų atsitiktinių lizdinių imčių, kai yra neatsakymų, kalibruotojo tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$	35
9.	10 subalansuotų lizdinių imčių, kai yra neatsakymų, tyrimo kintamojo Horvico ir Tompsono sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$	36
10.	Daugiareikšmio įrašymo metodo rezultatai	37
11.	Visų strategijų tyrimo kintamojo sumos ir jo tikslumo matų įverčių palyginimas, kai $m = 100$	38
12.	Visų strategijų tyrimo kintamojo sumos ir jo tikslumo matų įverčių palyginimas, kai $m = 1\,000$	38
13.	Visų strategijų tyrimo kintamojo sumos ir jo tikslumo matų įverčių palyginimas, kai $m = 5\,000$	38

Paveikslų sąrašas

1.	Visos galimos negražintinės imtys iš $N = 3$ didumo populiacijos	15
2.	Visos galimos negražintinės fiksuoto $n = 2$ dydžio imtys iš $N = 3$ dydžio populiacijos	15
3.	Komentaras, gaunamas išrinkus subalansuotą imtį	43

Įvadas

Darbo tikslas – išnagrinėti subalansuotos imties išrinkimo algoritmą bei jo atvaizdavimą kube, padaryti šio metodo apžvalgą, bei pritaikyti jį socialinio tyrimo duomenims. Kadangi nagrinėjamas metodas yra naujas bei Lietuvoje oficialiems tyrimams dar netaikytas, šis darbas turėtų būti naudingas sprendžiant, ar metodas yra pakankamai efektyvus ir naudingas atliekant imčių tyrimus mūsų šalyje. Darbe bus siekiama parodyti imties rinkimo metodo pranašumus bei trūkumus, pateikti taikymo rekomendacijas. Naudojant subalansuotą imties planą gauti įverčiai bus lyginami su įverčiais, gautais naudojant paprastąjį atsitiktinį lizdinį ėmimą bei regresinius įvertinius, siekiant parodyti subalansuoto ėmimo efektyvumą.

Uždaviniai:

1. Apžvelgti subalansuotos imties išrinkimo algoritmą bei paprastąją atsitiktinę lizdinę imtį su kalibruotuoju įvertiniu.
2. Išnagrinėti kubo metodą.
3. Išnagrinėti R programos paketo *sampling* taikymo galimybes subalansuotai imčiai rinkti.
4. Pritaikyti nagrinėtus metodus realiems duomenims.
5. Palyginti įverčius, gautus naudojant pasirinktas imties išrinkimo strategijas.
6. Pateikti taikymo rekomendacijas.

Tyrimo rezultatai buvo pristatyti 4-ojoje Baltijos – Šiaurės šalių imčių statistikos konferencijoje Helsinkyje (BaNoCoSS 2015) [10].

Apie imties rinkimą naudojant tam tikrą kintamųjų balansavimą buvo mąstoma nuo pat imčių statistikos pradžios, tačiau tam trukdė pernelyg sudėtingi skaičiavimai bei labai ilgas laikas jiems atlikti. Tokio rinkimo užuomazgų galima aptikti ir XIX amžiuje publikuotuose Kiaerio [15] darbuose. Pirmasis apie subalansuotą rinkimą parašė Gini 1921 metais [11]. 1929 metais Gini ir Galvani [12], naudodami subalansuotą ėmimą, iš 214 Italijos apskričių išrinko 29 apskričių imtį. Tačiau Neymanas (1952) [19] bei Yatesas (1949) [14] sukritikavo Gini pasiūlytą metodą, kadangi jame nebuvo naudojamas atsitiktinis ėmimas.

Apie subalansuotą ėmimą taip pat rašė ir Nekrašaitė-Liegė [20]. Jos darbe minima viena iš šio darbo strategijų – subalansuoto imties plano naudojimas kartu su kalibruotuoju įvertiniu taip siekiant pagerinti tyrimo kintamojo populiacijos sumos įvertinį.

Pats kubo metodas pirmą kartą buvo pristatytas 1998 metais. Jo autoriai – Deville'is ir Tille'as [8], tačiau pirmieji publikaciją šia tema paskelbė trys studentai Bousabbas, Lieberis ir Sirollis 1999 metais [1]. Visas metodas pirmą kartą paskelbtas 2001 metais Tille'o [29] bei 2004 metais Deville'io ir Tille'o [9]. Nuo to laiko buvo publikuoti keli šio metodo patobulinimai bei pats metodas praktiškai buvo pritaikytas Prancūzijoje.

Plačiau ir dažniau naudojamas lizdinis ėmimas. Yra minimi keli šio metodo privalumai [24]: lizdinės imtys yra ekonomiškesnės, turint tą patį biudžetą galima gauti didesnę imtį,

nei naudojantis kitais imčių planais, apklausos atlikimas užima mažiau laiko. Tačiau, kaip ir subalansuoto ėmimo atveju, yra ir trūkumų – to pačio lizdo elementų panašumas ir dėl to atsirandantis skirtingos informacijos apribojimas, didelis paklaidų standartinis nuokrypis.

Paprastojo atsitiktinio lizdinio ėmimo rezultatai gali būti tikslesni, jei yra naudojamas kalibruotasis įvertinys. Pagalbinis kintamuosius parametrų vertinimams galima gauti dviem būdais: naudoti papildomus kintamuosius, tokius kaip lytis, išsilavinimas, kuriems nereikalingi atsakymai į apklausos klausimus, arba, kaip rašė Chadyšas [2], pritaikius kelių fazių ėmimą, pirmosios fazės duomenis galima naudoti kaip papildomą informaciją sekančioje imties ėmimo fazėje. Šiame darbe nagrinėsime pirmąjį atvejį – kai turime vienos fazės ėmimą ir naudojama papildoma informacija nesutampa su tyrimo kintamuoju. Kalibruotasis įvertinys buvo pristatytas Deville'io ir Sarndalo 1992 metais [7] kaip procedūra, minimizuojanti atstumo funkciją tarp pradinių ir galutinių svorių kalibravimo lygtyse. Prieš tai Sarndalas, Swensonas ir Wretmanas [26] apibrėžė svorio parametą (*g-weight*), reikalingą kalibruotam įvertiniui.

1. Pagrindinės imčių teorijos sąvokos

Baigtinė populiacija – tai N dydžio elementų aibė $U = \{u_1, \dots, u_k, \dots, u_N\}$.

Imtis – tai populiacijos elementų poaibis. Ją žymėsime $\mathbf{i} = \{I_1 \dots I_k \dots I_N\}$, čia I_k - elemento priklausymo imčiai indikatorius.

Negražintinė imtis. Tokią imtį iš N dydžio populiacijos galime žymėti vektoriumi

$$\mathbf{i} = (I_1 \dots I_k \dots I_N)' \in \{0, 1\}^N, \quad (1)$$

čia, visiems $k \in U$, turime

$$I_k = \begin{cases} 1, & \text{jei elementas su numeriu } k \text{ priklauso } \mathbf{i}, \\ 0, & \text{jei elementas su numeriu } k \text{ nepriklauso } \mathbf{i}, \quad k = 1, \dots, N. \end{cases}$$

I_k - atsitiktinis dydis. Tokiu atveju imties $\mathbf{i}$ dydis $n(\mathbf{i})$ bus

$$n(\mathbf{i}) = \sum_{k \in U} I_k. \quad (2)$$

Gražintinė imtis. Analogiškai negražintinei imčiai, gražintinė imtis taip pat gali būti žymima vektoriumi

$$\mathbf{i} = (I_1 \dots I_k \dots I_N)' \in \mathbb{N}^N,$$

čia N žymi populiacijos dydį, I_k yra k -ojo elemento išrinkimo į imtį kartų skaičius, $\mathbb{N} = \{0, 1, 2, \dots\}$. Imties dydis $n(\mathbf{i})$ skaičiuojamas taip pat pagal (2) formulę, tik gali būti, kad $n(\mathbf{i}) > N$.

Tyrimo kintamasis – tai kintamasis, kurio reikšmės domina tyrimo metu. Šio kintamojo y k -ojo elemento reikšmę žymėsime y_k , $k \in N$. Tyrimo kintamojo sumą populiacijoje žymėsime

$$t_y = \sum_{i=1}^N y_i.$$

Pagalbiniai kintamieji – tai kintamieji, kurių reikšmės kiekvienam populiacijos elementui žinome prieš pradėdami tyrimą ir kurie yra naudojami vertinant pagrindinio tyrimo kintamojo reikšmių funkcijas. Šie kintamieji, kurių reikšmės žinomos visiems populiacijos elementams, taip pat gali padėti subalansuoti imtį. Tarkime, kintamajam y turime J pagalbinių kintamųjų, kurių reikšmės užrašome vektoriumi

$$\mathbf{x}_k = (x_{k1} \dots x_{kj} \dots x_{kJ}), \quad k = 1, 2, \dots, N,$$

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)'.$$

Tada žinome ir šių papildomų kintamųjų populiacijos reikšmių sumą

$$t_{\mathbf{x}} = \sum_{j=1}^J \mathbf{x}_j = (t_{\mathbf{x}_1}, t_{\mathbf{x}_2}, \dots, t_{\mathbf{x}_J})',$$

arba

$$t_{\mathbf{x}_j} = \sum_{i=1}^N x_{ji}, \quad j = 1, \dots, J.$$

Ji gali būti įvertinta, taikant Horvico ir Tompsono¹ sumos įvertinį

$$\hat{t}_{\mathbf{x}HT} = \sum_{k \in U} \frac{\mathbf{x}_k I_k}{\pi_k}, \quad (3)$$

čia π_k – k -ojo elemento įtraukties į imtį tikimybė. Sumos įvertinio dispersijos įvertinys skaičiuojamas pagal formulę

$$\hat{\mathbf{D}}\hat{t}_{\mathbf{x}HT} = \sum_{k \in U} \left(\frac{1 - \pi_k}{\pi_k^2} \right) y_k^2 I_k + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \frac{y_k I_k y_l I_l}{\pi_{kl}}.$$

Imties plano nešėju $\mathbf{Q}$ vadinsime imčių aibę, kurių imties plano tikimybės yra teigiamos. Imties plano nešėjas $\mathbf{Q}$ (aibė, kurios elementai yra imtys) yra vadinamas **simetriniu**, jei, $\forall \mathbf{i} \in \mathbf{Q}$, visi įmanomi $\mathbf{i} = (I_1, I_2, \dots, I_N)$ koordinačių perstatiniai taip pat priklauso $\mathbf{Q}$. Negrąžintinės imties plano simetrinis nešėjas (*symmetric support*): yra aibė

$$\mathbf{S} = \{0, 1\}^N.$$

Kaip tyrimo kintamojo tikslumo matai buvo pasirinkti šie įverčiai:

- variacijos koeficiento įvertis

$$\hat{c}v(\hat{t}) = \frac{\hat{\mathbf{D}}(\hat{t})}{\hat{t}}; \quad (4)$$

¹Horwitz–Thompson

- vidutinės santykinės kvadratinės paklaidos įvertis

$$\widehat{VSKP}(\hat{t}) = \frac{\sqrt{\widehat{Posl}^2(\hat{t}) + \hat{\mathbf{D}}(\hat{t})}}{\hat{t}}; \quad (5)$$

- poslinkio įvertis

$$\widehat{Posl}(\hat{t}) = \hat{t} - t. \quad (6)$$

2. Subalansuotas ėmimas

Subalansuotas ėmimas – tai toks imčių išrinkimo būdas, kai, naudojantis papildomais kintamaisiais, tų kintamųjų Horvico ir Tompsono (3) sumos įvertinių reikšmės yra tokios pat arba beveik tokios pat kaip ir tikrosios tų kintamųjų populiacijos sumos. Kitaip sakant, šis metodas leidžia fiksuoti ne vieną, o daugiau kintamųjų (pavyzdžiui: vyrų, moterų, vaikų skaičius populiacijoje, dirbančių ir bedarbių skaičius populiacijoje) ir naudojant juos kaip papildomą informaciją, užtikrinti, kad atsitiktinai išrinktos imtys neiškreiptų viso tyrimo rezultatų [27]. Yra ir daugiau subalansuoto ėmimo apibrėžimų, pabrėžiančių tokio ėmimo naudą. Taip pat iš apibrėžimo seka, jog subalansuota imtis yra tokia imtis, kurios papildomo kintamojo vidurkio įvertis yra lygus visos populiacijos papildomo kintamojo vidurkiui [18].

Tarkime, turime N dydžio baigtinę populiaciją U bei iš jos elementų sudarytą imtį $\mathbf{i}$. Imties planas $p(\mathbf{i})$ bus visų galimų imčių $S = \{\mathbf{i}\}$ išrinkimo tikimybių rinkinys, tenkinantis sąlygą

$$p(\mathbf{i}) > 0, \quad \sum_{\mathbf{i} \in S} p(\mathbf{i}) = 1.$$

Tegu k -ojo elemento tyrimo kintamojo y reikšmė yra y_k . Pažymime t_y visos populiacijos tyrimo kintamojo y reikšmių sumą

$$t_y = \sum_{k \in U} y_k.$$

Jei visi $\pi_k > 0$, kai $k \in U$, tai Horvico ir Tompsono (1952) [13] sumos įvertis apibrėžiamas formule

$$\hat{t}_y = \sum_{k \in U} \frac{y_k I_k}{\pi_k}$$

ir yra nepaslinktasis sumos t_y įvertis.

Tarkime, kad turime pagalbinių kintamųjų rinkinį $\mathbf{x}_1, \dots, \mathbf{x}_J$, kurių reikšmės $x_{k1}, \dots, x_{kJ}$ žinome kiekvienam k , $k \in U$. Tegu šie kintamieji yra koreliuoti su tyrimo kintamuoju y . Imties planas $p(\mathbf{i})$, $\mathbf{i} \in S$ bus vadinamas **subalansuotu** pagalbinio kintamojo x atžvilgiu tada ir tik tada [30], kai bus tenkinama balansavimo lygtis

$$\hat{t}_{\mathbf{x}HT} = t_{\mathbf{x}}, \quad (7)$$

arba, kai bus teisinga

$$\sum_{k \in U} \frac{\mathbf{x}_k I_k}{\pi_k} = \left(\sum_{k \in U} x_{k1}, \sum_{k \in U} x_{k2}, \dots, \sum_{k \in U} x_{kJ} \right)',$$

arba

$$\mathbf{D}(\hat{t}_{\mathbf{x}HT}) = 0.$$

Imtys, kurios tenkina balansavimo lygtį (7), turi griežtai teigiamas išrinkimo tikimybes. Tokiu atveju **subalansuoto imties plano nešėjas** bus aibė imčių

$$\mathbf{Q} = \left\{ \mathbf{i} \in \mathbf{S} \mid \sum_{k \in U} \frac{\mathbf{x}_k I_k}{\pi_k} = t_{\mathbf{x}} \right\}.$$

Tais atvejais, kai $J = 2$, $x_{k1} = \pi_k$ ir $x_{k2} = 1$, imtis visada bus subalansuota [28]. Tai parodysime 2.1 skyriuje.

Imties plano simetrinis nešėjas ypatingas tuom, kad ne tik pasirinkta imtis yra subalansuota, bet ir bet kaip perstačius imties (1) elementus, ji vis tiek liks subalansuota. Tarkime, turime imtį $\mathbf{i}_1 = \{0, 0, 1, 1\}$ ir žinome, kad ji gauta pagal subalansuotą imties planą. Jei perstatyta imtis $\mathbf{i}_2 = \{1, 0, 1, 0\}$, bei visos kitaip perstatytos imtys, taip pat yra subalansuotos – turėsime simetrinį imties plano nešėją. Kitaip sakant, turėsime C_4^2 skirtingų imčių, sudarančių simetrinį imties plano nešėją, o iš 4 elementų populiacijos iš viso galima sudaryti 2^4 įvairaus dydžio imčių. Geometriškai simetrinis nešėjas gali būti vaizduojamas kaip visų N -mačio hiperkubo viršūnių aibė [30]. Jei imties dydis yra fiksuotas ir lygus n , tai negražintinės imties plano simetrinis nešėjas apibrėžiamas

$$\mathbf{S}_n = \left\{ \mathbf{i} \in \mathbf{S} \mid \sum_{k \in U} I_k = n \right\}.$$

Norint sudaryti subalansuotos imties planą, reikia apibrėžti kainos funkciją $Kaina(\mathbf{i})$ visoms galimoms imtims. Ši funkcija turi tenkinti dvi sąlygas:

1. $Kaina(\mathbf{i}) > 0$, visoms galimoms imtims;
2. $Kaina(\mathbf{i}) = 0$, jei $\mathbf{i} \in \mathbf{Q}$.

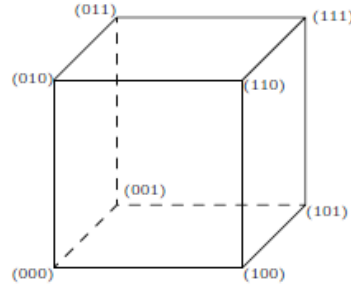
Kainos funkcija gali būti apibrėžta taip:

$$Kaina(\mathbf{i}) = \sum_{j=1}^J \frac{(\hat{t}_{\mathbf{x}_j HT}(\mathbf{i}) - t_{\mathbf{x}_j})^2}{t_{\mathbf{x}_j}^2}. \quad (8)$$

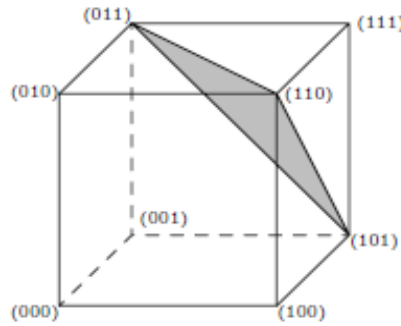
Kitaip sakant, jei skirtumas tarp išrinktos imties Horvico ir Tompsono papildomų kintamųjų sumos įvertinių bei tikrųjų reikšmių yra lygus nuliui ((8) išraiškoje skaitiklis lygus 0), tai imtis yra subalansuota.

Kaip atrodo visos galimos imtys iš $N = 3$ populiacijos, galima pamatyti 1 paveiksle ([30]). Kiekviena galima imtis nusakoma trijų komponentių vektoriumi, kurio vienetui lygi komponentė reiškia, kad atitinkamas elementas priklauso imčiai ir 0 lygi komponentė reiškia, kad

atitinkamas elementas imčiai nepriklauso. 2 paveiksle matomos pasirinktų pagalbinių kintamųjų atžvilgiu subalansuotos $N = 3$ dydžio populiacijos $n = 2$ dydžio imtys, sudarančios trikampio viršūnes [30]. Viršūnių yra gerokai mažiau, nei imčių yra iš viso.



1 pav. Visos galimos negrąžintinės imtys iš $N = 3$ didumo populiacijos



2 pav. Visos galimos negrąžintinės fiksuoto $n = 2$ dydžio imtys iš $N = 3$ dydžio populiacijos

Apžvelgsime kelis subalansuoto ėmimo atvejus:

1. Kai imties dydis yra fiksuotas,
2. Kai yra naudojamas sluoksninės imties planas,
3. Kai imčiai išrinkti yra naudojamas kubo metodas.

2.1. Balansavimo kintamieji esant fiksuotam imties dydžiui

Fiksuoto n dydžio imties planas yra subalansuotas pagalbinio kintamojo $x = (x_1, \dots, x_N)'$ su reikšmėmis $x_k = \pi_k$, $k \in U$, atžvilgiu. Iš tikrųjų,

$$\sum_{k \in U} \frac{x_k I_k}{\pi_k} = \sum_{k \in U} I_k = n. \quad (9)$$

Kai įtraukties į imtį tikimybės π_k yra skirtingos, Horvico ir Tompsono populiacijos dydžio N įvertinys yra atsitiktinis dydis [30], išreiškiamas

$$\hat{N} = \hat{t}_{N,HT} = \sum_{k \in U} \frac{I_k}{\pi_k}.$$

Kai populiacijos dydis N yra žinomas dar prieš renkant imtį, galima išrinkti atsisveriančią imtį, tai yra tokią imtį, kuri tenkintų sąlygą:

$$\sum_{k \in U} \frac{I_k}{\pi_k} = N.$$

Ši sąlyga yra balansavimo lygtis, kurioje balansavimo kintamasis $x_k \equiv 1, k \in U$. Šios sąlygos tenkinimas gali būti užtikrinamas taikant kubo metodą.

2.2. Subalansuotas sluoksninės imties planas

Tarkime, kad populiaciją galime suskirstyti į H nesikertančių dalių. Jas žymėsime $U_h, h = 1, \dots, H$, o jų dydžius – raidėmis $N_h, h = 1, \dots, H$. Šiose dalyse esantys elementai panašūs vieni į kitus (pavyzdžiui, sluoksniais populiacija gali būti skirstoma pagal gyvenvietės dydį). Tai nėra reikalaujama, bet ši sąlyga turi įtakos populiacijos sumos įvertinio dispersijos dydžiui. Iš kiekvieno sluoksnio nepriklausomai yra renkamos fiksuotų dydžių $n_1, \dots, n_H$ paprastosios atsitiktinės imtys.

Apibrėšime elemento priklausymo sluoksniui indikatorius $\delta_{k1}, \dots, \delta_{kH}$:

$$\delta_{kh} = \begin{cases} 1, & \text{jei } k \in U_h, \quad k = 1, \dots, N, \\ 0, & \text{jei } k \notin U_h, \quad k = 1, \dots, N. \end{cases}$$

Sluoksninės imties planas bus subalansuotas šių indikatorių atžvilgiu, jei

$$\hat{N}_h = \sum_{k \in U} \frac{I_k \delta_{kh}}{\pi_k} = \sum_{k \in U} \delta_{kh} = N_h, \quad h = 1, \dots, H.$$

Kitaip sakant, sluoksnio dydžio N_h Horvico ir Tompsono įvertinys $\hat{N}_j$ yra lygus tikrajai N_h reikšmei.

Y. Tille'as [30] aprašė sluoksniniame ėmime (bei visuose imties planuose) esančią apvalinimo problemą. Jei įtraukties į imtį tikimybių $\pi_k, k = 1, \dots, N$, suma $\sum_{k=1}^N \pi_k$ nėra lygi sveikam skaičiui, tai išrinkti paprastąją atsitiktinę imtį su fiksuotu imties dydžiu yra neįmanoma. Kai naudojame sluoksniavimą, kuriame sluoksnių imties dydžiai proporcingi sluoksnių dydžiams, visos įtraukties į imtį tikimybės atskiruose sluoksniuose yra lygios, todėl šiuo atveju yra neįmanoma nurodyti imties dydžio, kuris būtų sveikasis skaičius, sluoksnyje, ir kad įtraukties į imtį tikimybės būtų lygios visuose sluoksniuose. Ši apvalinimo problema išnyksta renkant didelio dydžio imtį.

2.3. Balansavimo algoritmai

2.3.1. Deville'io, Grosbras'o ir Roth'o balansavimo algoritmas lygioms įtraukties tikimybėms

1988 metais Deville'is, Grosbras'as ir Roth'as [6] aprašė kelis imties rinkimo algoritmus populiacijos elementams, kurie turi vienodas arba nevienodas įtraukties į imtį tikimybes. Jie taip pat apibūdino metodą, skirtą subalansuoti imčiai gauti naudojantis tam tikrais algoritmais. Apžvelgsime vieną iš jų siūlytų būdų.

Šis metodas yra paremtas grupėmis, kurios yra sudaromos naudojantis visų pagalbinių kintamųjų vidurkiais. Jei kintamųjų skaičius J nėra labai didelis, galime sukonstruoti 2^J grupių pagal galimų vektorių komponentų ženklą:

$$\text{sign}(x_{k1} - \bar{\mu}_1, \dots, x_{kj} - \bar{\mu}_j, \dots, x_{kJ} - \bar{\mu}_J),$$

čia $\bar{\mu}_j$ yra j -ojo pagalbinio kintamojo reikšmių vidurkis

$$\bar{\mu}_j = \frac{1}{N} \sum_{k \in U} x_{kj}, \quad j = 1, \dots, J.$$

Šiame poskyryje aprašomo algoritmo tikslas – išrinkti imtį, kuri būtų subalansuota vektorių $(\bar{\mu}_1, \dots, \bar{\mu}_j, \dots, \bar{\mu}_J)$ atžvilgiu. Naudojantis šiuo metodu gaunamos įtraukties į imtį tikimybės yra apytiksliai lygios duotosioms.

Šis algoritmas susideda iš 8 žingsnių:

1. Tarkime, kad populiacijos vidurkis $\bar{\mu} = (\bar{\mu}_1, \dots, \bar{\mu}_J)' \in \mathbb{R}^J$ yra vektorius, sudarytas iš 0. Suskaidykime populiaciją į G grupių, kur $G > J$, o J – balansavimo kintamųjų skaičius. Grupės turi būti sudarytos taip, kad jas sudarančių elementų skaičiai N_g būtų apytiksliai lygūs. Pažymėkime $\bar{x}_g \in \mathbb{R}^J$, $g = 1, \dots, G$, g -osios grupės vidurkį. Toliau paskirstykime imties dydį į grupes proporcingu būdu: $n_g = n \frac{N_g}{N}$. Skaičiai n_g yra nebūtinai sveikieji.
2. Imties grupių dydžių apvalinimas. Visi skaičiai n_g yra apvalinami [6]. Iš populiacijos $\{1, \dots, g, \dots, G\}$ su nelygiomis imties išrinkimo tikimybėmis $(n_1 - [n_1], \dots, n_g - [n_g], \dots, n_G - [n_G])'$, čia $[n_i]$ žymi sveikąją n_i dalį, išrenkame imtį, kurią pažymime indikatoriu vektoriumi $(I_1, \dots, I_g, \dots, I_G)$, čia I_g – grupės indikatorius. Suapvalintas grupės dydis bus $m_g = [n_g] + I_g$, $g = 1, \dots, G$.
3. Apskaičiuojame išrinktos imties vidurkį. Pažymime $\bar{x}(t) = \sum_{g=1}^G \pi_g \bar{x}_g$, čia π_g – g -osios grupės tikimybė būti išrinktai.
4. Skaičiuojame $\nu = (\nu_1, \dots, \nu_G)'$, minimizuojantį funkciją

$$\sum_{g=1}^G \frac{\nu_g^2}{N_g},$$

ir tenkinantį sąlygas

$$\sum_{g=1}^G \nu_g = 0, \quad \sum_{g=1}^G \nu_g \bar{x}_g = -n\bar{x}(t).$$

5. Apvaliname $\boldsymbol{\nu}$. Kiekviena ν_g reikšmė apvalinama taip pat, kaip ir n_g . Iš populiacijos $\{1, \dots, g, \dots, G\}$ su nelygiomis įtraukties į imtį tikimybėmis $(\nu_1 - [\nu_1], \dots, \nu_g - [\nu_g], \dots, \nu_G - [\nu_G])'$ išrenkame imtį $(I_1, \dots, I_g, \dots, I_G)'$. Suapvalintas ν_g reikšmės pažymime $\nu'_g = [\nu_g] - I_g$, $g = 1, \dots, G$.
6. Stabdymo taisyklė. Jei $\sum_g |\nu'_g| \geq A$, imtis yra tinkama. Priešingu atveju imame $t = t + 1$ ir skaičiuojame $A = \sum_g |\nu'_g|$.
7. Imties atnaujinimas. Jei $\nu'_g \geq 0$, pasinaudodami paprastuoju atsitiktiniu ėmimu išrenkame $|\nu'_g|$ reikšmių ir pašaliname jas iš imties. Jei $\nu'_g < 0$ - išrenkame $|\nu'_g|$ naujų elementų iš atitinkamų grupių pasinaudodami paprastuoju atsitiktiniu ėmimu ir įtraukiame juos į imtį.
8. Apskaičiuojame naują imties vidurkį $\bar{x}(t)$ ir tęsiame algoritmą nuo 4 žingsnio.

2.3.2. Subalansuotas ėmimas naudojantis tiesiniu programavimu

Šio algoritmo esmė [6] – apskaičiuoti

$$\min_{p(\cdot)} \sum_{\mathbf{i} \in S} p(\mathbf{i}) K_{aina}(\mathbf{i}),$$

kai

$$\sum_{\mathbf{i} \in S} p(\mathbf{i}) = 1, \quad \sum_{\mathbf{i} \in S} \mathbf{i} p(\mathbf{i}) = \boldsymbol{\pi}, \quad p(\mathbf{i}) \geq 0, \quad \forall \mathbf{i} \in S,$$

čia S – visų įmanomų imčių $\mathbf{i}$ aibė.

Toliau sudarome sąrašą visų galimų imčių $\{\mathbf{i}_1, \dots, \mathbf{i}_j, \dots, \mathbf{i}_L\}$ su jų išrinkimo tikimybėmis. Sugeneruojame tolygiai intervale $[0, 1]$ pasiskirsčiusį atsitiktinį dydį ξ ir išrenkame imtį tokią, kad

$$\sum_{i=1}^{l-1} p(\mathbf{i}_i) \leq \xi \leq \sum_{i=1}^l p(\mathbf{i}_i).$$

Naudojant tiesinį programavimą yra gaunamas bent vienas sprendinys, paremtas mažiausiu imties plano nešėją sudarančių aibių skaičiumi.

3. Kubo metodo etapai

Kubo metodą pirmieji pristatė Deville'is ir Tille'as 2004 metais [9]. Šis metodas – tai tam tikras algoritmas, kuris padeda atsitiktinai išrinkti subalansuotas imtis, kurių elementai turi iš anksto duotas įtraukties į imtį tikimybes.

Šis metodas yra taikomas prancūzų statistikų. Jų darbe tos savivaldybės, kurių gyventojų skaičius nesiekia 10 000, naudojantis subalansuotu imties planu su lygiomis įtraukties tikimybėmis $\pi_k = \frac{1}{5}$, $k = 1, \dots, N$, buvo suskirstytos į penkias nesikertančias rotacijos grupes. Kiekvienais metais penktadalis savivaldybių yra traukiamos į imtį. Didesnėse savivaldybėse penkios nesikertančių adresų imtys kasmet yra išrenkamos su įtraukties tikimybėmis $\pi_k = \frac{1}{8}$, $k = 1, \dots, N$. Būtent šioje stadijoje yra taikomas kubo metodas. Po 5 metų jose turėtų būti apklausta 40% gyventojų. Šiuo atveju pagalbiniai balansuojantys kintamieji yra gaunami iš naujausio gyventojų surašymo socialinių bei demografinių duomenų [27].

Negražintiniame ėmime kiekvienas vektorius $\mathbf{i}$ gali būti sutapatinamas su N -mačio kubo viršūne. Taigi toks kubas turi tiek viršūnių, kiek yra skirtingų imčių. Ėmimo planas su įtraukties į imtį tikimybėmis π_k turi priskirti tikimybę $p(\mathbf{i})$ kiekvienai N -mačio kubo viršūnei bei tos tikimybės turi tenkinti lygtį

$$E(S) = \sum_{\mathbf{i} \in S} p(\mathbf{i}) \mathbf{i} = \boldsymbol{\pi},$$

čia $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)'$ yra įtraukties į imtį tikimybių vektorius.

Balansavimo lygtis (7) gali būti užrašyta ir taip:

$$\sum_{k \in U} \check{\mathbf{x}}_k I_k = \sum_{k \in U} \check{\mathbf{x}}_k \pi_k, \quad k \in U, \quad (10)$$

čia $\check{\mathbf{x}}_k = \frac{\mathbf{x}_k}{\pi_k}$. Lygtis (10) apibrėžia tiesinį $\mathbb{R}^N$ poerdvį Q , kurio matavimas yra $N - J$, čia J - papildomų kintamųjų skaičius. Šį poerdvį galime apibrėžti ir kaip $Q = \boldsymbol{\pi} + \text{Ker} \mathbf{A}$, kur

$$\text{Ker} \mathbf{A} = \{\mathbf{v} \in \mathbb{R}^N | \mathbf{A}\mathbf{v} = 0\}$$

yra $J \times N$ dydžio matricos $\mathbf{A}$ branduolys ($\mathbf{A} = (\check{\mathbf{x}}_1 \dots \check{\mathbf{x}}_k \dots \check{\mathbf{x}}_N)$). Norint gauti subalansuotą imtį reikia pasirinkti tokią N -mačio kubo viršūnę, kuri arba priklausytų poerdviui Q , arba būtų nedaug nuo jo nutolusi, jei nepavyksta rasti viršūnės iš poerdvio.

Tarkime, $C = [0, 1]^N$ žymi N -matį kubą, kurio viršūnės atitinka visoms įmanomoms imtims iš populiacijos U . Tada aibių C ir Q sankirta bus netuščia aibė, kadangi $\boldsymbol{\pi}$ yra aibės C išorinis taškas ir kartu priklauso aibei Q .

Kubo metodą sudaro dvi pagrindinės fazės – pakilimo (*flight*) bei nusileidimo (*landing*) fazės. Pakilimo fazės metu reikia atsitiktinai suapvalinti beveik visas įtraukties tikimybes iki 0 arba 1 arba kitaip – atsitiktinai išrinkti viršūnę iš aibės $Q \cap C$. Nusileidimo fazėje yra siekiama kiek galima tiksliau patenkinti subalansavimo lygtį (9).

3.1. Pakilimo fazė

Pakilimo fazės tikslas – atsitiktinai išrinkti N -mačio kubo viršūnę taip, kad įtraukties į imtį tikimybės būtų lygios π_k visiems imties elementams bei pagrindinė balansavimo lygtis (7) būtų patenkinta. Nusileidimo fazė reikalinga tik tada, kai pasirinkta kubo viršūnė priklauso aibei Q , bet nepriklauso aibei C , t. y. tik apytiksliai tenkina balansavimo lygtį (7). Bendru atveju,

norint praeiti šią fazę, galima naudotis subalansuoto martingalo algoritmu.

Atsitiktinis procesas yra atsitiktinių dydžių šeima $\{Z_t, t \in \mathbb{T}\}$, kur indeksas t priklauso indeksų aibei $\mathbb{T}$. Jei $\mathbb{T}$ yra natūraliųjų skaičių aibės $\mathbb{N}$ poaibis, Z vadinsime **diskretaus laiko atsitiktiniu procesu** [16].

Diskretaus laiko atsitiktinis procesas $\boldsymbol{\pi}(t) = (\pi_1(t), \dots, \pi_N(t))'$, $t = 0, 1, \dots$, $k = 1, \dots, N$, yra vadinamas įtraukties į imtį tikimybių vektoriaus $\boldsymbol{\pi}$ bei pagalbinių kintamųjų $\mathbf{x}_1, \dots, \mathbf{x}_J$ **subalansuotu martingalu**, jei:

1. $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$;
2. $E\{\boldsymbol{\pi}(t) | \boldsymbol{\pi}(t-1), \dots, \boldsymbol{\pi}(0)\} = \boldsymbol{\pi}(t-1), t = 1, 2, \dots$;
3. $\boldsymbol{\pi}(t) \in K = \{[0, 1]^N \cap (\boldsymbol{\pi} + \text{Ker} \mathbf{A})\}$, kur $\mathbf{A}$ yra $J \times N$ dydžio matrica $\mathbf{A} = (\frac{\mathbf{x}_1}{\pi_1}, \dots, \frac{\mathbf{x}_N}{\pi_N})$.

Teiginys [1] Jei $\boldsymbol{\pi}(t)$ – subalansuotas martingalas, tai:

1. $E[\boldsymbol{\pi}(t)] = E[\boldsymbol{\pi}(t-1)] = \dots = E[\boldsymbol{\pi}(0)] = \boldsymbol{\pi}$;
2. $\sum_{k \in U} \check{\mathbf{x}}_k \pi_k(t) = \check{\mathbf{x}}_k \pi_k = t_{\mathbf{x}}, t = 0, 1, \dots$;
3. Kai subalansuotas martingalas pasiekia aibės C tašką, jis jo nebepalieka.

Algoritmas I Pradėkime nuo $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Tille'as [27] siūlo tokią bendrą subalansavimo algoritmą. Pakilimo fazė:

1. Sugeneruojame bet kokią (nebūtinai atsitiktinį) vektorių $\mathbf{u}(t) = (u_1(t), \dots, u_N(t)) \neq (0, \dots, 0)$ tokį, kad $\mathbf{u}(t)$ priklausytų matricos $\mathbf{A}$ branduoliui ir $u_k(t) = 0$, jei $\pi_k(t)$ yra sveikas skaičius.
2. Randame $\lambda_1^*(t)$ ir $\lambda_2^*(t)$ – didžiausias iš visų tokių $\lambda_1(t)$ ir $\lambda_2(t)$, kurioms

$$\boldsymbol{\pi}(t) + \lambda_1(t) \mathbf{u}(t) \leq 1,$$

$$0 \leq \boldsymbol{\pi}(t) - \lambda_2(t) \mathbf{u}(t) \leq 1.$$

3. Renkame $\boldsymbol{\pi}(t+1)$ tokį, kad

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t) \mathbf{u}(t) & \text{su tikimybe } q_1(t), \\ \boldsymbol{\pi}(t) - \lambda_2^*(t) \mathbf{u}(t) & \text{su tikimybe } q_2(t). \end{cases}$$

4. Kartojame 1-3 žingsnius tol, kol įmanoma atlikti pirmąjį algoritmo žingsnį, kai $t = 1, \dots, T$.

Chauvetas ir Tille'as [3] pasiūlė greitą pakilimo fazės algoritmą. Jie siūlė naudoti ne visą matricą $\mathbf{A}$, o tik jos dalinę matricą $\mathbf{B}$, sudarytą iš $J+1$, $J < N$ matricos $\mathbf{A}$ stulpelių. Šios matricos rangas q yra ne didesnis už J . Tokiu atveju nereikėtų naudoti visų $\mathbf{A}$ matricos

stulpelių. Nepaisant visų tokio algoritmo privalumų, apvalinimo problema išlieka, kadangi išrikiavus baigtinės populiacijos elementus pagal dydžio kintamąjį mažėjimo tvarka, pirmiausiai bus balansuojami didžiausią dydžio kintamojo reikšmę turintys elementai, o patys mažiausi gali susidurti su apvalinimo sunkumais.

3.2. Nusileidimo fazė

Pakilimo fazės pabaigoje subalansuotas martingalas yra pasiekęs aibės Q viršūnę, kuri nebūtinai yra C viršūnė. Ši viršūnė yra pažymima $\boldsymbol{\pi}^* = \{\pi_k^*\} = \boldsymbol{\pi}(T)$. Tarkime, šią viršūnę sudaro q komponentų, kurios yra ne sveikieji skaičiai. Jei $q = 0$ tai algoritmas baigtas, priešingu atveju – visi apribojimai negali būti tiksliai įvykdyti. Šiuo atveju galima naudotis išrikiavimo algoritmu. Visų pirma apibrėšime suderintą imtį.

Imtis $\mathbf{i}$ yra **suderinta** su vektoriumi $\boldsymbol{\pi}^*$, jei $\pi_k^* = I_k$ visiems k tokiems, kuriems π_k^* yra sveikasis skaičius. Tarkime, $\mathbf{C}(\boldsymbol{\pi}^*)$ yra aibė suderintų imčių su 2^q skaičiumi elementų su $\boldsymbol{\pi}^*$. Nusileidimo fazėje skaičiavimus užtenta atlikti populiacijos poaibyje $\mathbf{C}(\boldsymbol{\pi}^*)$. Reikia apskaičiuoti

$$\min_{p^*(\cdot)} \sum_{\mathbf{i} \in \mathbf{C}(\boldsymbol{\pi}^*)} K a_{\mathbf{i}}(\mathbf{i}) p^*(\mathbf{i}),$$

čia

$$\sum_{\mathbf{i} \in \mathbf{C}(\boldsymbol{\pi}^*)} p^*(\mathbf{i}) = 1, \quad \sum_{\mathbf{i} \in \mathbf{C}(\boldsymbol{\pi}^*)} \mathbf{i} p^*(\mathbf{i}) = \boldsymbol{\pi}^*, \quad 0 \leq p^*(\mathbf{i}) \leq 1, \quad \forall \mathbf{i} \in \mathbf{C}(\boldsymbol{\pi}^*).$$

Atlikus šiuos skaičiavimus, pagal imties planą $p^*(\cdot)$ reikia išrinkti imtį. Kadangi $q \leq J$, tai skaičiavimai nepriklauso nuo populiacijos dydžio, bet priklauso nuo balansuojančių kintamųjų. Jei įtraukties į imtį tikimybės yra pagalbiniai kintamieji ir jų suma yra sveikasis skaičius, tai toks tiesinio programavimo algoritmas gali būti taikomas tik kai

$$C_n(\boldsymbol{\pi}^*) = \left\{ \mathbf{i} \in \mathbf{C}(\boldsymbol{\pi}^*) \mid \sum_{k \in U} I_k = \sum_{k \in U} \pi_k = n \right\}.$$

Kubo metodas taip pat naudoja atsitiktinį apvalinimą. Apvalinimo problema šiuo atveju priklauso tik nuo $q \leq J$ reikšmių. Tille'as [30] pateikė kelias išvadas, svarbias naudojant kubo metodą. Visų pirma, tai naudojant kubo metodą yra teisinga

$$|\hat{t}_{\mathbf{x}_j HT} - t_{\mathbf{x}_j}| \leq J \times \max_{k \in U} \left| \frac{x_{kj}}{\pi_k} \right|.$$

Taip pat jei įtraukties tikimybių suma yra sveikasis skaičius ir jei imties dydis yra fiksuotas, tai kubo metodui yra teisinga:

$$|\hat{t}_{\mathbf{x}_j HT} - t_{\mathbf{x}_j}| \leq (J - 1) \times \max_{k \in U} \left| \frac{x_{kj}}{\pi_k} - \frac{N \bar{t}_{\mathbf{x}_j}}{n} \right|,$$

čia

$$\bar{t}_{\mathbf{x}_j} = \frac{1}{N} \sum_{k \in U} x_{kj}.$$

Praktikoje apvalinimo problema aiškiai matoma tik mažų imčių atveju bei sluoksninio ėmimo atveju, kai iš sluoksnių renkamos mažo dydžio imtys. Nusileidimo fazėje kaip įmanoma labiau yra stengiamasi išvengti šios problemos.

4. Imties plano svorių kalibravimas

Gauti sumos įvertinius su savybe, kad papildomų kintamųjų sumos vertinamos tiksliai, galima naudojantis ne tik subalansuotu ėmimu, tačiau ir kalibruotaisiais įvertiniais, kuriuose taip pat naudojama žinoma informacija apie populiacijos elementų papildomus kintamuosius [21]. Tarkime, turime populiaciją U bei tyrimo kintamąjį y . Norime įvertinti sumą:

$$t_y = \sum_{k=1}^N y_k.$$

Turėdami n dydžio tikimybinę imtį, sumą t_y galima vertinti Horvico ir Tompsono įvertiniu

$$\hat{t}_{y\pi} = \sum_{k \in U} \frac{y_k I_k}{\pi_k} = \sum_{k \in U} d_k y_k I_k,$$

čia π_k yra k -ojo populiacijos elemento priklausymo imčiai tikimybė, $d_k = \frac{1}{\pi_k}$ imties plano svoriai, $k \in U$.

Jei turime papildomos informacijos apie populiaciją, galima sukonstruoti ir tikslesnius įvertinius, turinčius mažesnes dispersijas, negu Horvico ir Tompsono įvertinys. Kalibruotieji įvertiniai būtent tokie ir yra. Tarkime, turime J toje pačioje populiacijoje apibrėžtų pagalbinių kintamųjų $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_J$. Pažymėkime pagalbinių kintamųjų sumas

$$t_{\mathbf{x}_1} = \sum_{k=1}^N x_{1k}, \quad t_{\mathbf{x}_2} = \sum_{k=1}^N x_{2k}, \quad \dots, \quad t_{\mathbf{x}_J} = \sum_{k=1}^N x_{Jk},$$

o kintamųjų vektoriaus $\mathbf{x}$ populiacijos reikšmių sumą

$$\mathbf{t}_{\mathbf{x}} = \sum_{k=1}^N \mathbf{x}_k = (t_{\mathbf{x}_1}, t_{\mathbf{x}_2}, \dots, t_{\mathbf{x}_J})'.$$

Laikysime, kad sumą $\mathbf{t}_{\mathbf{x}}$ žinome. Visų populiacijos elementų kintamųjų vektoriaus $\mathbf{x}$ sumos gali būti žinomos iš ankstesnių tos pačios populiacijos tyrimų arba kitų duomenų bazių.

Kintamojo y sumos t_y kalibruotuoju įvertiniu vadinamas įvertinys

$$\hat{t}_{yw} = \sum_{k \in U} w_k y_k I_k,$$

kurio svoriai w_k fiksuotoje imtyje tenkina sąlygas:

1. w_k kaip galima mažiau skiriasi nuo imties plano svorių d_k , minimizuodami funkciją

$$L(w_k, d_k, k \in U) = \sum_{k \in U} \frac{(w_k - d_k)^2 I_k}{d_k q_k} \longrightarrow \min,$$

čia q_k , $k \in U$, laisvai pasirenkamos konstantos. Specialiai parenkant papildomus svorius q_k galime gauti skirtingus įvertinius. Praktikoje dažniausiai $q_k = 1$. Šios prielaidos laikysimės ir savo darbe.

2. w_k tenkina kalibravimo lygtį

$$\hat{\mathbf{t}}_{xw} = \sum_{k \in U} w_k \mathbf{x}_k I_k = \mathbf{t}_x,$$

čia $\hat{\mathbf{t}}_{xw} = (\hat{t}_{\mathbf{x}_1 w}, \dots, \hat{t}_{\mathbf{x}_J w})'$, $\hat{t}_{\mathbf{x}_j w} = \sum_{k \in U} w_k x_{jk} I_k$, $j = 1, \dots, J$. Ši kalibravimo lygtis reiškia, kad papildomų kintamųjų $\mathbf{x}_j$ sumų įvertiniai $\hat{t}_{\mathbf{x}_j w}$ sutampa su šių kintamųjų populiacijos sumomis $t_{\mathbf{x}_j}$, $j = 1, \dots, J$.

Įvertinio $\hat{t}_{yw}$ svoriai w_k , $k \in U$, vadinami kalibruotaisiais. Jie priklauso nuo imties, todėl yra atsitiktiniai.

Tarkime, kad tarp pagrindinio kintamojo y ir papildomų kintamųjų yra ganėtinai didelė koreliacija. Kadangi svorius w_k naudojančios papildomų kintamųjų sumų įvertiniai sutampa su tikrosiomis šių kintamųjų sumų reikšmėmis, yra pagrindo tikėtis, kad ir tiriamo kintamojo y sumos t_y kalibruotasis įvertinys $\hat{t}_{yw}$ bus tikslesnis (turės mažesnę dispersiją) negu Horvico ir Tompsono įvertinys $\hat{t}_{y\pi}$.

Šio kalibruotojo sumos įvertinio $\hat{t}_{yw}$

1. Svoriai yra

$$w_k = d_k (1 + q_k (t'_k - \sum_{k \in U} I_k d_k \mathbf{x}'_k) (\sum_{i \in U} I_k d_i q_i \mathbf{x}_i \mathbf{x}'_i)^{-1} \mathbf{x}_k);$$

2. Apytikslė dispersija yra

$$AD \hat{t}_{y w} = \sum_{k, l=1}^N (\pi_{kl} - \pi_k \pi_l) (y_k - \mathbf{x}'_k \mathbf{b}_w) (y_l - \mathbf{x}'_l \mathbf{b}_w),$$

čia vektorius $\hat{\mathbf{b}}_w$ yra lygties

$$(\sum_{k \in U} I_k w_k q_k \mathbf{x}_k \mathbf{x}'_k) \hat{\mathbf{b}}_w = \sum_{k \in U} I_k w_k q_k \mathbf{x}_k y_k$$

sprendinys.

Jei tarp tyrimo kintamojo ir papildomų kintamųjų vektoriaus yra tiesinė priklausomybė $y_k = \boldsymbol{\beta}' \mathbf{x}_k$, $k = 1, \dots, N$, su konstantų vektoriumi $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)'$, tai, vertindami kintamojo y

sumą kalibruotuoju įvertiniu, gauname tikrąją sumos t_y reikšmę

$$\hat{t}_{yw} = \sum_{k \in U} w_k \beta' \mathbf{x}_k I_k = \beta' \sum_{k \in U} w_k \mathbf{x}_k I_k = \beta' \sum_{k=1}^N \mathbf{x}_k = \sum_{k=1}^N y_k = t_y.$$

Kaip jau buvo minėta šio skyriaus pradžioje, kalibruotieji svoriai yra atsitiktiniai. Gali būti, kad kai kurie iš jų gali būti netgi neigiami, tai reikštų, kad imties elementas pats sau neatstovauja. Siekiant išvengti neigiamų svorių, jiems daromas papildomas apribojimas $a \leq \frac{w_k}{d_k} \leq b$, $k \in U$, su pasirinktais skaičiais a , b .

5. Logistinės regresijos modelio taikymas praleistoms kintamojo reikšmėms įrašyti

Kaip žinome iš 2 skyriaus, subalansuotas ėmimas yra skirtas išrinkti imtį iš populiacijos taip, kad ji kuo geriau atspindėtų pačią populiaciją pasirinktų kintamųjų atžvilgiu. Tačiau esant išrinktų į imtį elementų neatsakymams imtis yra išbalansuojama ir, kaip vėliau matysime iš praktinių skaičiavimų, gautas įvertis įgyja didelį poslinkį. Vienas iš būdų išvengti poslinkio – naudoti kalibruotą įvertinį, tačiau tokiu atveju du kartus tenka naudoti papildomą informaciją. Ši strategija bus išbandyta. Kitas būdas – užpildyti praleistas reikšmes. Įrašymui naudosime logistinę regresiją. Logistinės regresijos modelis buvo pristatyta 1958 metais statistiko Davido Coxo [4].

Dvireikšmė logistinė regresija [5] – tai toks apibendrintas tiesinis modelis, kuriame vienam priklausomam dvireikšmiui kintamajam įtaką daro vienas ar keli nepriklausomi aiškinamieji kintamieji. Mūsų atveju būtent tokia situacija ir yra – mūsų tyrimo kintamasis gali įgyti dvi reikšmes: 1, jei asmuo bedarbis ir 0 – kitu atveju. Aiškinamieji kintamieji šiuo atveju būtų lytis, amžiaus grupė, gyvenamoji vieta. Modelio išraiška:

$$P(y = 1) = \frac{e^{Z(\mathbf{x})}}{1 + e^{Z(\mathbf{x})}}, \quad (11)$$

čia y – tyrimo kintamojo reikšmių vektorius, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_J)'$ – pagalbinių kintamųjų reikšmių vektorius, o

$$Z(\mathbf{x}) = \alpha + \beta_1 \mathbf{x}_1 + \dots + \beta_J \mathbf{x}_J.$$

Kitaip pagrindinė modelio lygtis (11) gali būti užrašyta taip:

$$\ln \frac{P(y_k = 1)}{P(y_k = 0)} = Z(\mathbf{x}_k), \quad (12)$$

čia $Z(\mathbf{x}_k) = \alpha + \beta_1 x_{k1} + \dots + \beta_J x_{kJ}$, $k = 1, \dots, A$, kai norime įrašyti A praleistų reikšmių.

Pačiam duomenų įrašymui galima naudoti bet kokius pasirinktus kintamuosius. Kadangi subalansuotam ėmimui buvo pasirinkti 3 balansavimo kintamieji, tai, laikantis prielaidos, jog turime tuos pačius papildomus duomenis, paprastajam atsitiktiniam lizdiniam ėmimui bei logis-

tiniam praleistų reikšmių užpildymui naudosime tuos pačius pagalbinius kintamuosius. Papildomai galima naudoti darbo biržos kintamąjį, kuris parodo, ar žmogus yra joje užsiregistravęs, tačiau laikantis prielaidos, jog turimi tie patys duomenys, kurie buvo naudoti ir subalansuotame ėmime, šio kintamojo neįtrauksime.

Logistiniam regresijos modeliui įvertiname asmens tikimybę būti bedarbiu $p_k = P(y_k = 1)$ iš (12):

$$\hat{p}_k = \frac{e^{\hat{Z}(\mathbf{x}_k)}}{1 + e^{\hat{Z}(\mathbf{x}_k)}},$$

čia $\hat{Z}(\mathbf{x}_k) = \hat{\alpha} + \hat{\beta}_1 x_{1k} + \dots + \hat{\beta}_J x_{Jk}$. Tada modeliuojame Bernulio atsitiktinį dydį y_k , įgyjantį 1 su įvertintąja tikimybe $\hat{p}_k$. Gautas reikšmes įrašome į praleistas y reikšmių vietas ir sumą vertiname taip, tarsi turėtume visas y reikšmes.

Norėdami išsiaiškinti galimą Horvico ir Tompsono sumos įverčio dispersijos padidėjimą, naudosime daugiareikšmio įrašymo metodą [21]. Šį metodą pasiūlė Rubinas [23], Litlas ir Rubinas [17]. Šio metodo esmė – kiekvienai praleistai reikšmei pasirinktu būdų (šiuo atveju – naudojantis logistinės regresijos modeliu) modeliuojama keletas reikšmių. Taip sudarius pasirinktą skaičių pilnų imčių tyrimo kintamojo suma vertinama taip, lyg nebūtų buvę jokių praleistų reikšmių. Patys modelio autoriai manė, jog šis metodas tinkamiausias naudoti, kai kintamojo reikšmės yra praleistos atsitiktinai (nepriklauso nuo tyrimo kintamojo reikšmių). Darysime prielaidą, kad mūsų atveju būtent taip ir yra.

Trumpai apžvelgsime daugiareikšmio įrašymo metodą. Tarkime, turime imtį su praleistomis reikšmėmis. Jas sumodeliavus ir įrašius vieną kartą gaunamas pilnas duomenų rinkinys. Šiam rinkiniui skaičiuojamas tyrimo kintamojo sumos įvertis bei jo dispersijos įvertis. Tai pačiai pradinei imčiai su tomis pačiomis praleistomis reikšmėmis vėl modeliuojame trūkstamas reikšmes ir atliekame sumos vertinimą. Tai kartojame pasirinktą skaičių C kartų. Modelio autorių nuomone, C neturi būti labai didelis, dažnai užtenka rinktis ir $C = 10$. Būtent tokį kiekį kartų ir naudosime. Tada gausime sumų įvertinius

$$\bar{t}_{yC} = \frac{1}{C} \sum_{c=1}^C \hat{t}_{yc}$$

čia $\hat{t}_{yc}$ – sumos įvertis, kai praleistos reikšmės yra įrašomos c -ąjį kartą, $c = 1, \dots, C$, ir dispersijos įvertinius

$$\hat{\mathbf{D}}_{\bar{t}_{yC}} = \bar{W}_C + \frac{C+1}{C} \hat{B}_C.$$

Čia $\bar{W}_C$ – papildytų duomenų rinkinių vidinės dispersijos įvertinys

$$\bar{W}_C = \frac{1}{C} \sum_{c=1}^C \hat{\mathbf{D}}_{\hat{t}_{yc}},$$

o $\hat{B}_C$ – dispersijos tarp papildytų duomenų rinkinių įvertinys

$$\hat{B}_C = \frac{1}{C-1} \sum_{c=1}^C (\hat{t}_{yC} + \bar{\hat{t}}_{yC})^2.$$

Jei iš skirtingų papildytų duomenų rinkinių gaunami sumų įvertiniai labai skiriasi, tai dispersijos įvertinio dėmens $\hat{B}_C$ indėlis į bendrą dispersijos įvertinį $\hat{D}\bar{\hat{t}}_{yC}$ yra didelis.

Šis metodas turi tiek privalumų, tiek trūkumų [21]. Įrašius visas praleistas reikšmes yra gaunamas pilnas duomenų rinkinys ir yra išsaugoma imties plano svorių sistema, bet tokiu atveju sumos t_y įverčiai gali turėti poslinkį.

6. Statistinis modeliavimas

6.1. Tyrimo populiacija

Modeliavimui naudojami Lietuvos statistikos departamento gyventojų užimtumo statistinio tyrimo (2011) duomenys. Klausimyną sudaro 98 klausimai, duomenis pateikė 28 058 imčiai priklausantys asmenys. Laikome, kad šio baigiamojo darbo tyrimo populiacija yra visi asmenys, dalyvavę gyventojų užimtumo tyrime. Fiktyvią populiaciją sudaro 21 318 asmenys nuo 16 iki 69 metų amžiaus.

Rinkdami imtį, populiaciją skaidysime į lizdus. Lizdu laikysime namų ūkį, lizdo dydžiu – namų ūkio narių skaičių. Į imtį rinksime lizdus. Lizdą išrinkus į imtį, imčiai priklausys visi lizdo nariai.

Nors klausimyną sudaro beveik šimtas klausimų, tačiau baigiamojo darbo uždaviniui spręsti visų jų nereikia. Darbo tikslas – įvertinti bedarbių dalį populiacijoje. Kadangi imtis yra balansuojama ją renkant, reikėjo nuspręsti, kurie kintamieji gali būti pagalbiniai. Buvo pasirinkti trys pagalbiniai kintamieji, kurių duomenis galime gauti ir neatlikę apklausos – tai lytis, amžius bei gyvenamoji vieta (miestas arba kaimas). Lyties kintamajame 1 reiškia moterį, 2 reiškia vyrą. Miesto ar kaimo kintamajame 1 reiškia miestą, 2 reiškia kaimą. Taigi šie du kintamieji gali įgyti tik dvi reikšmes. Amžiaus kintamasis buvo suskirstytas į 6 grupes: pirmą grupę sudaro asmenys, kurių amžius priklauso intervalui [16;22], antrą grupę – asmenys, kurių amžius priklauso intervalui [23;29], trečią grupę – [30;39], ketvirtą grupę – [40;49], penktą grupę – [50;59], šeštą grupę – [60;69].

Modeliavimas atliekamas trimis skirtingo dydžio imtimis – 100, 1 000 ir 5 000 lizdų. Skirtingo dydžio imtys pasirinktos norint įvertinti, kokio dydžio imtys yra tinkamiausios skirtingoms strategijoms bei kuriuo atveju galime gauti mažiausią poslinkį. Kiekvienu atveju yra renkama 10 atitinkamo dydžio imčių, iš kurių vertiname populiacijos sumos vidurkį, bei skaičiuodamas jų vidurkis.

6.2. Strategijos

Naudosime tokias 7 strategijas:

1. Subalansuotas lizdinis ėmimas ir Horvico ir Tompsono sumos įvertinys. Ėmimo stadijoje naudojamas kubo metodas bei papildoma informacija.
2. Paprastasis atsitiktinis lizdinis ėmimas ir kalibruotasis įvertinys. Papildoma informacija naudojama tik vertinimo stadijoje.
3. Subalansuotas lizdinis ėmimas, kai yra neatsakymų ir kalibruotasis įvertinys.
4. Subalansuotas lizdinis ėmimas ir kalibruotasis įvertinys.
5. Paprastasis atsitiktinis lizdinis ėmimas, kai yra neatsakymų ir kalibruotasis įvertinys.
6. Subalansuotas lizdinis ėmimas, kai yra neatsakymų ir Horvico – Tompsono įvertinys.
7. Subalansuotas lizdinis ėmimas, kai yra neatsakymų, bei logistinės regresijos modelio taikymas praleistoms tyrimo kintamojo reikšmėms įrašyti ir Horvico ir Tompsono įvertinys.

Renkant subalansuotą imtį, kai papildoma informacija yra naudojama ėmimo stadijoje, išrinktų asmenų neatsakymai į apklausą didina populiacijos sumos įvertinio poslinkį, dispersiją, taip pat išbalansuoja pačią imtį. Siekiant atsižvelgti į neatsakymus ir sumažinti poslinkį, populiacijos sumos vertinimo etape naudosime kalibruotą įvertinį. Kitaip sakant, naudodami trečią strategiją tirsime, ar papildomos informacijos naudojimas imties ėmimo bei parametrų vertinimo stadijose esant neatsakymams gali sumažinti sumos įvertinio poslinkį lyginant su penktąja bei šeštąja strategijomis, kur papildoma informacija yra naudojama tik vienoje iš stadijų.

Neatsakymus į apklausą reikia sumodeliuoti. Laikysime, kad į imtį išrinktas lizdas atsako su tikimybe, lygia $\pi_{ats} = 0,9$. Šiuo atveju turėsime naudoti į apklausą atsakančių lizdų įtraukties į imtį tikimybių persvėrimą. Tada šių lizdų įtraukties į imtį tikimybės bus lygios

$$\pi_k^* = \pi_k * \pi_{ats},$$

čia π_k – pradinė lizdo įtraukties į imtį tikimybė. Šis įtraukties į imtį tikimybių persvėrimas bus naudojamas ir penktoje bei šeštoje strategijose.

Kaip jau buvo minėta teorinėje dalyje, subalansuotas ėmimas nėra tinkamas mažoms imtims, kadangi gautų imčių sumos įvertinių poslinkiai yra dideli. Norėdami išspręsti šią problemą, naudosime ketvirtą strategiją, kai papildoma informacija yra naudojama du kartus – imties ėmimo ir parametrų vertinimo stadijose. Kartu įvertinsime, ar vidutinio dydžio bei didelės imtims verta naudoti papildomą informaciją du kartus bei ar galime gauti mažesnius poslinkius, nei vertinant sumos populiaciją papildomą informaciją naudojant tik vienoje iš stadijų.

Septintoji strategija skirta ištirti, ar galima sumažinti subalansuoto ėmimo, kai turime neatsakymų, populiacijos sumos įvertinio poslinkį, kai vietoje papildomos informacijos naudojimo vertinimo stadijoje užpildome praleistas tyrimo kintamojo reikšmes naudodami logistinės regresijos modelį.

6.2.1. Subalansuotas lizdinis ėmimas ir Horvico ir Tompsono įvertinys

Rinkdami subalansuotą nelygių tikimybių negrąžintinę lizdinę imtį naudosimės R programos paketu *sampling* bei funkcija *balancedcluster*:

```
balancedcluster(X, m, cluster, selection=1, comment=TRUE, method=1).
```

Šios funkcijos argumentai yra:

1. X – pagalbinių kintamųjų reikšmių matrica;
2. m – skaičius lizdų, renkamų į imtį (skiriasi nuo išrinktų elementų skaičiaus);
3. $cluster$ – natūraliųjų skaičių, identifikuojančių lizdo numerį, vektorius;
4. $selection$ – argumentas, nusakantis įtraukties tikimybės. Jei 1 – tikimybės proporcingos lizdo dydžiui, jei 2 – lygios tikimybės;
5. $comment$ – renkant imtį pateikiamas papildomas komentaras, jei pasirenkama *TRUE* reikšmė;
6. $method$ – *samplecube* funkcijoje naudojamas imties išrinkimo metodas. Funkcija *samplecube* išrenka subalansuotą imtį ir yra naudojama *balancedcluster* subalansuotos lizdų imties rinkimo funkcijoje.

Pasinaudoję funkcija *balancedcluster*, gauname komentarą, pavaizduotą A priede. Komentare aprašomos pakilimo bei nusileidimo fazės. Pakilimo fazės apraše pateikiamas balansavimo kintamųjų skaičius, populiacijos dydis, įtraukties į imtį tikimybių vektoriaus reikšmių suma bei iš karto nurodoma, kiek tikimybių reikės apvalinti. Nusileidimo fazės apraše konkrečiu atveju matyti, kad yra įmanomos 4 skirtingos subalansuotos imtys su mūsų pasirinktu imties dydžiu. Taip pat yra pateikiama $Kaina$ reikšmė, kuri skaičiuojama pagal (8) formulę. Jei $Kaina$ reikšmė beveik lygi 0 – tai reiškia, jog imtis yra beveik tiksliai subalansuota.

Imtis renkama su tikimybėmis, proporcingomis lizdo dydžiui. Tarkime, mūsų i -ojo lizdo dydis yra m_i . Iš viso populiaciją sudaro M lizdų ir $N = \sum_{i=1}^M m_i$ elementų. Tokiu atveju i -ojo lizdo tikimybė priklausyti imčiai bus $\pi_i \cong m \frac{m_i}{N}$, kur m – lizdų imties dydis.

Išrenkame 10 atsitiktinių imčių su $m = 5000$. Apskaičiuojame balansavimo kintamųjų sumos įverčius kiekvienai grupei (vyrams, moterims, visoms 6 amžiaus grupėms, gyventojams mieste bei kaime). Suvidurkinti rezultatai pateikiami 1 lentelėje. Tuo atveju, kai pagalbinis kintamasis gali įgyti tik dvi reikšmes (pvz.: lytis, gyvenamoji vieta), pagalbinių kintamųjų įverčiai yra tokie patys (suapvalinus) arba beveik tokie patys kaip ir tikrosios reikšmės. Kitaip sakant – yra patenkinta balansavimo lygtis 9. Taip pat matome, kad, naudojant pagalbinį kintamąjį, kuris gali įgyti daugiau nei vieną reikšmę (asmenų amžius suskirstytas į 6 grupes), atsiranda didesni poslinkiai tarp įverčių ir tikrųjų reikšmių.

2 lentelėje pateikiame mūsų tyrimo kintamojo įverčius, dispersijas, variacijos koeficientus bei santykinę vidutinę kvadratinę paklaidą, skaičiuotus pagal (4), (5), (6) formules. Parametrų

1 lentelė. 10 subalansuotų imčių balansuojančių kintamųjų sumų įverčiai, kai $m = 5\,000$

j	i	x_{ji}	$\bar{t}_{x_{ji}}$	$t_{x_{ji}}$	Posl
1	1	Vyras	9 674	9 674	0
	2	Moteris	11 644	11 644	0
2	1	1 amžiaus gr.	3 109	3 096	15
	2	2 amžiaus gr.	1 680	1 700	-20
	3	3 amžiaus gr.	2 891	2 899	-8
	4	4 amžiaus gr.	4 964	4 978	-14
	5	5 amžiaus gr.	5 172	5 107	65
	6	6 amžiaus gr.	3 502	3 538	-34
3	1	Gyvenantis mieste	14 634	14 635	-1
	2	Gyvenantis kaime	6 684	6 683	1

vertinimus atlikus su visomis 10 imčių buvo pastebėta, kad, nors ir amžiaus pagalbinio kintamojo įverčių poslinkiai egzistuoja, tačiau bedarbių ir dirbančiųjų skaičius yra vertinamas ganėtinai tiksliai, skirtingų imčių dispensijos skiriasi taip pat nedaug, o variacijos koeficientai skiriasi vos tūkstantosiomis ar dešimttūkstantosiomis dalimis.

Matome, kad nors pagalbinio (amžiaus) kintamojo įverčiai ir nėra lygūs tikrosioms reikšmėms, tačiau tyrimo kintamojo įvertis yra artimas tikrajai reikšmei.

2 lentelė. 10 subalansuotų imčių tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 5\,000$

	$\bar{t}_y$	$\bar{D}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 725	2 186	0,027	0,027	1 732	-7
Užimtieji	19 593	5 459	0,040	0,040	19 586	7

Subalansuotas imtis buvo bandyta rinkti ir mažesnes, pavyzdžiui 100 ar 1 000 lizdų, tačiau tokiais atvejais pablogėja subalansuotumas, taip pat atsiranda gyventojų skaičiaus įverčių poslinkiai. Tikslūs rezultatai vaizduojami 3 lentelėje, o ieškomų koeficientų įverčiai – 4 lentelėje.

3 lentelė. 10 subalansuotų imčių balansuojančių kintamųjų sumų įverčiai, kai $m = 100$, $m = 1\,000$

j	i	x_{ji}	$t_{x_{ji}}$	$\bar{t}_{x_{ji}, m=100}$	$\widehat{Posl}_{m=100}$	$\bar{t}_{x_{ji}, m=1000}$	$\widehat{Posl}_{m=1000}$
1	1	Vyras	9 674	9 664	-10	9 788	114
	2	Moteris	11 644	11 654	10	11 530	-114
2	1	1 amžiaus gr.	3 096	3 613	517	3 393	297
	2	2 amžiaus gr.	1 700	1 261	-439	1 652	-48
	3	3 amžiaus gr.	2 899	2 494	-405	2 381	-518
	4	4 amžiaus gr.	4 978	4 477	-501	5 561	583
	5	5 amžiaus gr.	5 107	5 877	770	4 086	-1 021
	6	6 amžiaus gr.	3 538	3 596	58	4 246	650
3	1	Gyvenantis mieste	14 635	14 496	-139	14 709	74
	2	Gyvenantis kaime	6 683	6 822	139	6 609	-74

4 lentelė. 10 subalansuotų imčių tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$

$m = 100$	$\hat{t}_y$	$\hat{D}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\mathcal{P}osl$
Bedarbiai	1 595	169 191	0,259	0,272	1 732	-137
Užimtieji	19 723	565 531	0,038	0,039	19 586	137
$m = 1\,000$	$\hat{t}_y$	$\hat{D}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\mathcal{P}osl$
Bedarbiai	1 716	17 168	0,076	0,077	1 732	-16
Užimtieji	19 601	50 531	0,012	0,011	19 586	16

Matome, kad dispersijos bei variacijos koeficientų įverčiai yra gerokai didesni, nei tuo atveju, kai rinkome 5 000 lizdų imtį. Kaip jau buvo minėta teorinėje dalyje, subalansuotas ėmimas turi apvalinimo problemą mažoms imtims ir tai labai aiškiai matosi mūsų gautuose rezultatuose.

6.2.2. Paprastasis atsitiktinis lizdinis ėmimas ir kalibruotasis įvertinys

Bedarbių skaičiaus įverčius, gautus naudojant subalansuotą lizdų imties išrinkimą, lyginsime su populiacijos sumų įverčių rezultatais, gautais renkant paprastąją atsitiktinę lizdinę imtį ir vėliau naudojant imties plano svorių kalibravimą, tai yra priskiriant svorius, kurie priklauso nuo tų pačių, subalansuotame ėmime naudotų, pagalbinių kintamųjų.

Paprastąją atsitiktinę lizdinę imtį rinksime naudodami *sampling* paketo funkciją *cluster*:

```
cluster(data, clustertype, size, method=c("srswor", "srswr",
      "poisson", "systematic"), pik, description=FALSE).
```

Šios funkcijos argumentai yra:

1. *data* – duomenų matrica, kuri turi N eilučių;
2. *clustertype* – kintamojo, kuris skaido duomenis į lizdus, stulpelio vardas;
3. *size* – lizdų imties dydis;
4. *method* – lizdų rinkimo metodas. Gali būti: paprastasis atsitiktinis negrąžintinis ėmimas (*srswor*), paprastasis atsitiktinis grąžintinis ėmimas (*srswr*), Puasono ėmimas (*poisson*), sisteminis ėmimas (*systematic*). Jei metodas nėra nurodomas, automatiškai renkama paprastoji atsitiktinė negrąžintinė imtis;
5. *pik* – įtraukties į imtį tikimybių vektorius. Šis argumentas naudojamas tik nelygių tikimybių ėmimui.
6. *description* – jei pasirenkama reikšmė *TRUE*, papildoma informacija (išrinktų lizdų skaičius, populiacijos elementų skaičius bei imties elementų skaičius) yra pateikiama.

Funkcija *cluster* pateikia duomenų reikšmes, kurias sudaro išrinkti lizdai, identifikacinius elementų numerius išrinktuose lizduose, galutines įtraukties į imtį tikimybės.

Išrinkę paprastąją atsitiktinę lizdinę imtį, toliau turime apskaičiuoti svorius, kuriuos naudosime kalibravimui. Tam naudosime funkciją *calib*:

```
calib (Xs, d, total, q=rep(1, length(d)), method=c("linear", "raking",
"truncated", "logit"), bounds=c(low=0, upp=10), description=FALSE,
max_iter=500) .
```

Šios funkcijos argumentai yra:

1. *Xs* – kalibravimo kintamųjų (mūsų atveju - amžius, lytis, gyvenamoji vieta) matrica;
2. *d* – imties plano svorių vektorius. Mūsų atveju $d_k = \frac{1}{\pi_k}$, $k=1, \dots, M$;
3. *total* – papildomų kintamųjų sumų vektorius;
4. *q* – teigiamų reikšmių vektorius. Mūsų atveju tai bus vektorius, sudarytas iš vienetų, kurio ilgis bus lygus imties dydžiui;
5. *method* – kalibravimo metodas;
6. *bounds* – svorių (*g-weights*), naudojamų *truncated* ir *logit* methoduose, apribojimų vektorius. Svoriai $g_k = \frac{w_k}{d_k}$, kur w_k – kalibruotieji svoriai.
7. *description* – pasirinkus reikšmę TRUE, programa pateiks tarpinius ir galutinius svorius, jų histogramas;
8. *max_iter* – maksimalus iteracijų skaičius.

Ši funkcija mums pateikia svorių (*g-weights*) vektorių.

Mūsų imtis bus kalibruojama pagal tris pagalbinius kintamusius, kurie iš viso turi 10 galimų reikšmių (vyras arba moteris, 6 amžiaus grupės bei asmuo, gyvenantis mieste arba kaime). Kalibruojančių kintamųjų matrica bus sudaryta iš 8 stulpelių. Kiekvienas stulpelis reikš atitinkamą požymį, o 1 – jog atitinkamas elementas tą požymį turi, ir 0 – jei neturi. Amžiaus kintamąjį pasirenkame kaip pagrindinį – kiekviena jo reikšmė turės savo stulpelį pagalbinių kintamųjų matricoje. Siekiant išvengti tiesinės kintamųjų priklausomybės, lyties ir gyvenamosios vietos kintamieji turės po vieną stulpelį šioje matricoje – iš lyties kintamojo pasirenkame vyrus, o iš gyvenamosios vietos – gyvenančius mieste.

Gavę kalibravimui reikalingus svorius, toliau galime skaičiuoti įverčius. Tam naudosimės funkcija *calibev*:

```
calibev (Ys, Xs, total, pikl, d, g, q=rep(1, length(d)), with=FALSE, EPS=1e-6) .
```

Šios funkcijos argumentai yra:

1. *Ys* – tyrimo kintamojo reikšmių vektorius, kurio ilgis yra lygus imties dydžiui;
2. *Xs* – kalibravimo kintamųjų (mūsų atveju - amžius, lytis, gyvenamoji vieta) matrica;

3. *total* – papildomų kintamųjų sumų vektorius;
4. *pikl* – elementų porų įtraukties į imtį tikimybių matrica. Bendrų įtraukties į imtį tikimybių matcija bus sudaroma taip: matrica turės n eilučių ir n stulpelių, kur n – imties dydis. Tiek stulpeliai, tiek eilutės reikš imties elementus, išrikiuotus ta pačia tvarka pagal lizdus.
5. d – imties plano svorių vektorius. Mūsų atveju $d_k = \frac{1}{\pi_k}$, $k=1, \dots, M$;
6. g – kalibravimo svorių, gautų naudojantis *calib* funkcija, vektorius;
7. q – teigiamų reikšmių vektorius. Mūsų atveju tai bus vektorius, sudarytas iš vienetų, kurio ilgis bus lygus imties dydžiui;
8. *with* – jei pasirenkama TRUE reikšmė, dispersijos įverčio skaičiavime naudojami pradiniai svoriai d . Priešingu atveju – galutiniai svoriai yra $w=g*d$ (būtent jie ir yra naudojami);
9. *EPS* – paklaidos dėl kalibravimo leidžiama maksimali reikšmė.

Ši funkcija pateikia dvi reikšmes: *cest* – tyrimo kintamojo sumos kalibruotąją įvertį bei *evar* – sumos įverčio dispersijos įvertinį.

Toliau parašysime išraiškas *pikl* reikšmėms. Tarkime, kad turime du lizdus $u_k = \{e_{ki}, i = 1, \dots, m_k\}$ ir $u_l = \{e_{li}, i = 1, \dots, m_l\}$, čia m_k ir m_l žymi elementų skaičių atitinkamame lizde. Atskiriems elementams turime

$$\pi_{ki,lj} = P(e_{ki} \in \mathbf{i}, e_{lj} \in \mathbf{i}) = P(e_{ki} \in \mathbf{i} | e_{lj} \in \mathbf{i}) P(e_{lj} \in \mathbf{i}), P(e_{ki} \in \mathbf{i} | e_{lj} \in \mathbf{i}) =$$

$$= \begin{cases} P(e_{ki} \in \mathbf{i} | e_{kj} \in \mathbf{i}) = 1, & \text{kai } k = l, \quad k, l = 1, \dots, M, \\ P(e_k \in \mathbf{i} | e_l \in \mathbf{i}) = \frac{n-1}{N-1}, & \text{kai } k \neq l, \quad k, l = 1, \dots, M, \\ P(e_{ki} \in \mathbf{i}) = \frac{n}{N}. \end{cases}$$

Iš čia

$$\pi_{ki,lj} = \begin{cases} \frac{n}{N}, & \text{kai } k = l, \quad k, l = 1, \dots, M \\ \frac{n-1}{N-1} \frac{n}{N}, & \text{kai } k \neq l, \quad k, l = 1, \dots, M. \end{cases}$$

Kaip ir subalansuoto ėmimo atveju, buvo renkamos trijų skirtingų dydžių lizdų imtys: 100, 1 000 ir 5 000. Kiekvienu atveju rinkta po 10 imčių ir skaičiuoti reikalingų įverčių vidurkiai. Rezultatai matomi 5 lentelėje.

Matome, kad paprastas lizdinis atsitiktinis ėmimas su kalibruotu įverčiu, priešingai nei subalansuotas ėmimas, yra tinkamas ir mažoms imtims tirti – gauti poslinkiai mažų imčių atveju yra mažesni. Šiuo atveju poslinkiai lygūs vos 2 vienetams. Tiek mažų, tiek didelių imčių atvejais poslinkiai gauti mažesni, nei subalansuoto ėmimo atveju. Taip pat matome, kad

5 lentelė. 10 paprastųjų atsitiktinių imčių kalibruotojo tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$

$m = 100$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 730	137 509	0,211	0,214	1 732	-2
Užimtieji	19 588	137 509	0,018	0,019	19 586	2
$m = 1\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 768	12 863	0,064	0,067	1 732	36
Užimtieji	19 550	12 863	0,006	0,006	19 586	-36
$m = 5\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 744	1 530	0,022	0,023	1 732	12
Užimtieji	19 574	1 530	0,002	0,002	19 586	-12

variacijos bei santykinės vidutinės kvadratinės paklaidos koeficientai gauti maži, o tai reiškia, kad gauti įverčiai yra labai arti tikrųjų reikšmių.

6.2.3. Subalansuotas lizdinis ėmimas ir kalibruotasis įvertinys

Šios strategijos tikslas – išsiaiškinti, ar papildomos informacijos naudojimas du kartus gali duoti tikslesnius įverčius, nei papildomos informacijos naudojimas tik vienoje iš stadijų. Kaip ir ankstesniais atvejais, buvo renkamos 3 skirtingų dydžių imtys – 100, 1 000 ir 5 000 lizdų. Kiekvienu atveju rinkta po 10 imčių ir skaičiuotas įverčių vidurkis.

Vėl tiksliausi rezultatai gauti mažoms ir didelėms imtims, tačiau vidutinio dydžio imtims poslinkis yra didelis. Rezultatai matomi 6 lentelėje. Matome, kad mažoms imtis sumos įvertinio empirinis poslinkis yra mažesnis, nei subalansuoto lizdinio ėmimo su Horvico–Tompsono įvertiniu atveju. Galime daryti išvadą, kad kubo metodo mažų imčių įtraukties į imtį tikimybių apvalinimo problemą kalibruotasis įvertinys gali išspręsti ir suteikti tikslesnius įverčius. Tačiau palyginus paprastojo atsitiktinio ėmimo su kalibruotu įverčiu bei šios strategijos rezultatus matome, kad mažų imčių įverčiai abiem atvejais yra labai panašūs, taigi galima svarstyti galimybę naudoti paprastąjį atsitiktinį ėmimą su kalibruotu įverčių vietoje šios strategijos ir taip sutaupyti laiko bei kompiuterio resursų.

Didelėms imtims nėra tikslo naudoti papildomą informaciją du kartus, kadangi rezultatai yra labai panašūs į tuos, kuriuos gavome papildomą informaciją naudojant tik vienoje stadijoje.

6.2.4. Subalansuotas lizdinis ėmimas, kai yra neatsakymų, ir kalibruotasis įvertinys

Naudojantis šia strategija buvo siekiama ištirti, ar papildomos informacijos naudojimas imties rinkimo stadijoje bei parametrų vertinimo stadijoje yra naudingas, kai ne visi išrinkti asmenys atsako į apklausą. Kitaip sakant – ar kalibruotasis įvertinys gali pataisyti subalansuoto ėmimo rezultatus, kai duomenų gauta ne tiek, kiek buvo tikėtasi. Modeliuojant apklausą neatsakymo lygis buvo parinktas lygus 0,1. Šiuo atveju taip pat reikia naudoti tikimybių pers-

6 lentelė. 10 subalansuotų lizdinių imčių kalibruoto tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$

$m = 100$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 740	178 696	0,242	0,243	1 732	8
Užimtieji	19 578	178 696	0,021	0,022	19 586	-8
$m = 1\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 672	17 166	0,079	0,086	1 732	-60
Užimtieji	19 646	17 166	0,007	0,007	19 586	60
$m = 5\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 727	2 157	0,027	0,027	1 732	-5
Užimtieji	19 591	2 157	0,002	0,002	19 586	5

vėrimą. Išrinktų į imtį ir atsakiusių į klausimus elementų tikimybes persveriname, padauginę iš tikimybės atsakyti:

$$\pi_{ats\ k} = \pi_k * 0,9,$$

čia k - atsakiusio lizdo numeris, π_k – pradinė k -ojo lizdo įtraukties į imtį tikimybė.

Iš pradžių išrenkama subalansuota imtis. Vėliau atsitiktiniu būdu 10% išrinktų lizdų pašalinama iš imties laikant juos neatsakiusiais ir bedarbių skaičius vertinamas naudojant kalibruotą įvertinį. Gauti rezultatai matomi 7 lentelėje.

7 lentelė. 10 subalansuotų lizdinių imčių, kai yra neatsakymų, kalibruoto tyrimo kintamojo sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$

$m = 100$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 729	208 781	0,265	0,264	1 732	-3
Užimtieji	19 589	208 781	0,023	0,023	19 586	3
$m = 1\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 621	17 977	0,083	0,107	1 732	-111
Užimtieji	19 697	17 977	0,007	0,009	19 586	111
$m = 5\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 727	2 417	0,028	0,029	1 732	-5
Užimtieji	19 591	2 417	0,003	0,003	19 586	5

Kaip matome, mažiausi poslinkiai gauti mažoms ir didelėms imtimis. Mažiausi variacijos bei santykinės vidutinės kvadratinės paklaidos koeficientų įverčiai gauti didelėms imtimis. Dispersijos bei variacijos koeficiento įverčiai gražiai mažėja imties dydžiui augant. Taigi galime sakyti, kad mūsų atveju didelėms imtimis papildomos informacijos naudojimas vertinimo stadijoje patikslina norimą gauti įvertinį. Mažoms imtimis, nors ir gavus nedidelį poslinkį, tačiau negalime griežtai teigti, kad ši strategija yra tinkamiausia kadangi gautas variacijos koeficientas yra gerokai per didelis.

6.2.5. Paprastoji lizdinė atsitiktinė imtis, kai yra neatsakymų, ir kalibruotasis įvertinys

Šios strategijos tikslas - išsiaiškinti, ar papildomos informacijos naudojimas tik vertinant pasirinktus parametrus yra tiek pat tikslus ir galintis sutaupyti papildomų resursų bei laiko, kiek ir papildomos informacijos naudojimas renkant imtį bei vertinant parametrus kartu.

Kadangi darome prielaidą, kad yra 0,1 neatsakymo tikimybė, vėlgi turime naudoti tikimybių persvėrimą. Kaip ir ankstesniu atveju, išrinkimo į imtį tikimybes tiems elementams, kurie buvo išrinkti ir į apklausą atsakė, dauginame iš 0,9:

$$\pi_{ats\ k} = \pi_k * 0,9,$$

čia k - atsakiusio lizdo numeris, π_k – pradinė k -ojo lizdo įtraukties į imtį tikimybė.

Kaip matome iš 8 lentelės, geresnių rezultatų, nei naudojant papildomą informaciją imties ėmimo bei parametrų vertinimo metu, gauti nepavyko. Poslinkiai gauti gerokai didesni, nei minėtu atveju. Tačiau vidutinio dydžio imtims poslinkį bei variacijos koeficientą sumažinti pavyko.

8 lentelė. 10 paprastųjų atsitiktinių lizdinių imčių, kai yra neatsakymų, kalibruotojo tyrimo kintamojo sumo bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$

$m = 100$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 692	208 339	0,276	0,271	1 732	-40
Užimtieji	19 626	208 339	0,023	0,023	19 586	40
$m = 1\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 768	18 508	0,077	0,080	1 732	36
Užimtieji	19 550	18 508	0,007	0,007	19 586	-36
$m = 5\,000$	$\hat{t}_y$	$\hat{\mathbf{D}}\hat{t}_y$	$\widehat{cv}(\hat{t}_y)$	$\widehat{SVKP}(\hat{t}_y)$	t_y	$\widehat{Posl}$
Bedarbiai	1 714	2 408	0,029	0,030	1 732	-18
Užimtieji	19 604	2 408	0,003	0,003	19 586	18

Vėliau matysime, jog naudojantis šia strategija įverčiai gaunami tikslesni, nei naudojant subalansuotą ėmimą kai yra neatsakymų.

6.2.6. Subalansuotas lizdinis ėmimas kai yra neatsakymų, ir Horvico ir Tompsono įvertinys

Šios strategijos tikslas – naudojant subalansuotą lizdinį ėmimą kartu su kubo metodu įvertinti bedarbių skaičių, kai yra neatsakymų. Čia taip pat naudojame tikimybių persvėrimą. Kaip matėme iš ankstesnių strategijų rezultatų, tyrimo kintamojo sumos bei jo tikslumo matų įvertinių rezultatai yra susiję, taigi toliau skaičiavimus atliksime tik bedarbių sumai populiacijoje vertinti.

Kaip matome iš 9 lentelės, rezultatai visais atvejais gauti prasčiausi lyginant su kitomis strategijomis. Neatsakymai visiškai išbalansuoja išrinktą imtį ir papildomos informacijos naudojimas imties rinkimo etape nebetenka prasmės. Tai pastebėjo ir Sarndalas bei Lundquistas [25].

9 lentelė. 10 subalansuotų lizdinių imčių, kai yra neatsakymų, tyrimo kintamojo Horvico ir Tompsono sumos bei jo tikslumo matų įverčiai, kai $m = 100$, $m = 1\,000$, $m = 5\,000$

$m = 100$	$\hat{t}_y$	$\bar{\mathbf{D}}\hat{t}_y$	$\overline{cv}(\hat{t}_y)$	$\overline{SVKP}(\hat{t}_y)$	t_y	$\overline{Posl}$
Bedarbiai	1 992	228 248	0,244	0,273	1 732	-260
$m = 1\,000$	$\hat{t}_y$	$\bar{\mathbf{D}}\hat{t}_y$	$\overline{cv}(\hat{t}_y)$	$\overline{SVKP}(\hat{t}_y)$	t_y	$\overline{Posl}$
Bedarbiai	1 899	21 374	0,077	0,117	1 732	-167
$m = 5\,000$	$\hat{t}_y$	$\bar{\mathbf{D}}\hat{t}_y$	$\overline{cv}(\hat{t}_y)$	$\overline{SVKP}(\hat{t}_y)$	t_y	$\overline{Posl}$
Bedarbiai	1 924	2 914	0,028	0,104	1 732	-192

Mažų imčių atveju empirinis poslinkio vidurkis lygus -260, o tai sudaro apie 15% bedarbių populiacijoje. Mažiausias empirinio poslinkio įvertinys gautas vidutinėms imtims, tačiau ir čia jis yra didesnis nei 100.

6.2.7. Subalansuotas lizdinis ėmimas su Horvico ir Tompsono įvertiniu, kai yra neatsakymų, ir logistinė regresija

Kaip matome, neatsakymai visiškai išbalansuoja sumos įverčio rezultatus. Norint to išvengti, buvo nuspręsta pritaikyti logistinę regresiją praleistoms duomenų reikšmėms sumodeliuoti. Jos pagalba tiems asmenims, kurie buvo neatsakę į apklausą, sumodeliavome tikimybes būti bedarbiais. Turint šias tikimybes buvo įrašomos praleistos tyrimo kintamojo reikšmės: modeliuojami Bernulio atsitiktiniai dydžiai, kurie galėjo įgyti reikšmę 1 (bedarbis) su sumodeliuotomis tikimybėmis. Mūsų atveju vienos imties sumos įvertinys lygus

$$\hat{t}_y = \sum_{k \in \mathbf{i}^{(a)}} w_k y_k + \sum_{k \in \mathbf{i}^{(n)}} w_k \hat{y}_k,$$

čia $w_k = \frac{1}{\pi_k}$, $\mathbf{i}^{(n)}$ – neatsakiusių asmenų imtis, $\mathbf{i}^{(a)}$ – atsakiusių asmenų imtis, $\hat{y}(k)$ – logistinės regresijos pagalba sumodeliuotos praleistos reikšmės neatsakiusiems asmenims.

Pritaikę logistinės regresijos modelį (žr. 5 skyrių) gauname, kad mūsų lygtis lygi

$$\ln \frac{\hat{P}(y=1)}{\hat{P}(y=0)} = -4,0579 + 0,1625 * x_{vyras} - 0,4259 * x_{miestas} + 1,001 * x_{1 \text{ am. gr.}} + 2,3502 * x_{2 \text{ am. gr.}} \\ + 2,1715 * x_{3 \text{ am. gr.}} + 2,0546 * x_{4 \text{ am. gr.}} + 2,0058 * x_{5 \text{ am. gr.}}$$

Vertinimo rezultatus galime matyti iš 11, 12 ir 13 lentelių.

Mažoms imtims, lyginant su subalansuoto lizdinio ėmimo be neatsakymų strategija, poslinkį pavyko sumažinti beveik 3 kartus. Ir, nors poslinkis ir nėra labai mažas, tačiau vidutinės santykinės kvadratinės paklaidos reikšmė gauta netgi mažesnė, nei kitų 4 strategijų.

Vidutinio dydžio imtims taip pat gauti panašūs rezultatai – poslinkis sumažintas beveik trigubai, vidutinės santykinės kvadratinės paklaidos koeficiento įvertis sumažėjo taip pat apie 30%.

Didelėms imtims poslinkis buvo sumažintas beveik 10 kartų, dispersija gauta viena mažiausių iš visų strategijų. Tačiau vidutinės santykinės kvadratinės paklaidos koeficientas išliko vienas iš didžiausių.

Kadangi taikomas atsitiktinis įrašymas, galima panaudoti daugiareikšmio įrašymo metodą dėl įrašymo atsirandančiam sumos įvertinio dispersijos padidėjimui įvertinti.

10 lentelė. Daugiareikšmio įrašymo metodo rezultatai

	$m = 100$	$m = 1\,000$	$m = 5\,000$
$\hat{t}_{y10}$	2 258	1 758	1 725
$\hat{\mathbf{D}}\hat{t}_{y10}$	241 469	19 779	2 365
$\bar{W}_{10}$	226 576	16 411	2 118
$\hat{B}_{10}$	13 539	3 062	225

Skirtingiems imties dydžiams išrenkame po vieną imtį, kurioje lizdo neatsakymo tikimybė yra 0,1. Naudojantis logistine regresija, $C = 10$ kartų užpildome praleistas reikšmes, kiekvieną kartą išsaugodami tyrimo kintamojo įverčio bei jo dispersijos reikšmes. Rezultatai matomi 10 lentelėje. Mažoms imtims $\hat{B}_{10}$ reikšmė yra ganėtinai didelė ir sudaro apie 6 % sumos įvertinio dispersijos taigi galime teigti, kad skirtingų duomenų rinkinių sumos įvertiniai labai skiriasi. Kaip matome iš 11 lentelės, gaunamas dispersijos $\hat{\mathbf{D}}\hat{t}_{y10}$ įvertinys yra netgi didesnis, nei 6-osios strategijos, bet vidinės įvertinio dispersijos $\bar{W}_{10}$ įvertinys yra netgi mažesnis. Vidutinio dydžio imtims $\hat{B}_{10}$ yra dar didesnė ir sudaro apie 15 % sumos įvertinio dispersijos, tačiau, kaip matome iš 12 lentelės, tiek $\hat{\mathbf{D}}\hat{t}_{y10}$, tiek $\bar{W}_{10}$ įverčiai yra mažesni, nei 6-osios strategijos. Didelėms imtims dėl įrašymo atsirandantis dispersijos įvertinio padidėjimas yra labai nedidelis, pats dispersijos įvertinys yra mažesnis, nei 6-osios strategijos atveju, taigi galime teigti, jog, nors dėl praleistų reikšmių įrašymo ir egzistuoja sumos įvertinio dispersijos padidėjimas, tačiau tai padeda sumažinti sumos įvertinio poslinkį.

Apibendrinti visų 7 strategijų rezultatai matomi 11, 12 ir 13 lentelėse.

11 lentelė. Visų strategijų tyrimo kintamojo sumos ir jo tikslumo matų įverčių palyginimas, kai $m = 100$

	1 str.	2 str.	3 str.	4 str.	5 str.	6 str.	7 str.
$\hat{t}_y$	1 595	1 730	1 740	1 729	1 692	1 992	1 828
$\widehat{Var}(\hat{t}_y)$	169 191	137 509	178 696	208 781	208 339	228 248	201 827
$\widehat{SVKP}(\hat{t}_y)$	0.272	0.214	0.243	0.264	0.271	0.273	0.251
$\widehat{Posl}(\hat{t}_y)$	-137	-2	8	-3	-40	-260	96

12 lentelė. Visų strategijų tyrimo kintamojo sumos ir jo tikslumo matų įverčių palyginimas, kai $m = 1\,000$

	1 str.	2 str.	3 str.	4 str.	5 str.	6 str.	7 str.
$\hat{t}_y$	1 716	1 768	1 672	1 621	1 768	1 899	1 800
$\widehat{Var}(\hat{t}_y)$	17 168	12 863	17 166	17 977	18 508	21 374	18 013
$\widehat{SVKP}(\hat{t}_y)$	0.077	0.067	0.086	0.107	0.080	0.117	0.084
$\widehat{Posl}(\hat{t}_y)$	-16	36	-60	-111	36	-167	68

13 lentelė. Visų strategijų tyrimo kintamojo sumos ir jo tikslumo matų įverčių palyginimas, kai $m = 5\,000$

	1 str.	2 str.	3 str.	4 str.	5 str.	6 str.	7 str.
$\hat{t}_y$	1 725	1 744	1 727	1 727	1 714	1 924	1 750
$\widehat{Var}(\hat{t}_y)$	2 186	1 530	2 157	2 417	2 408	2 914	2 210
$\widehat{SVKP}(\hat{t}_y)$	0.027	0.023	0.027	0.029	0.030	0.104	0.078
$\widehat{Posl}(\hat{t}_y)$	-7	12	-5	-5	-18	-192	18

Išvados

Atlikus tyrimą buvo gauti tokie rezultatai:

1. Įverčiai, turintys mažiausią santikinę vidutinę kvadratinę paklaidą, visoms skirtingo dydžio imtims gauti naudojant paprastąją atsitiktinę imtį su kalibruotu įvertiniu.
2. Mažoms bei didelėms imtims su neatsakymais tinkamiausia strategija – subalansuotas ėmimas kartu su kalibruotu įverčiu. Papildomos informacijos naudojimas du kartus, kai yra neatsakymų, suteikia mums tikslesnius rezultatus.
3. Vidutinio dydžio imtims su neatsakymais geriausia strategija – paprastas atsitiktinis lizdinis ėmimas su kalibruotu įverčiu.
4. Subalansuoto ėmimo atveju neatsakymai į apklausą visiškai išbalansuoja rezultatus. Šios strategijos atveju gauti dideli poslinkiai, dispersijų įverčiai, variacijos koeficientai bei vidutinės santykinės kvadratinės paklaidos įverčiai. Rezultatus galima pagerinti papildomai naudojant kalibruotą įvertį.
5. Mažų imčių atveju įverčiai gali būti patikslinti papildomą informaciją naudojant tiek imties rinkimo, tiek kintamųjų vertinimo stadijose. Kalibruotas įvertinys subalansuoto ėmimo atveju padeda išvengti kubo metodui būdingos apvalinimo problemos ir rezultatų išbalansavimo. Taip pat padeda sumažinti dėl neatsakymų atsiradusį poslinkį. Apie šią strategiją ne tik mažoms imtims rašė ir Tille'as [31].
6. Didelėms imtims nėra jokio tikslo papildomą informaciją naudoti imties ėmimo ir kintamųjų vertinimo stadijose – užtenka papildomą informaciją naudoti vieną kartą. Visų trijų strategijų gauti rezultatai labai panašūs, tačiau naudojant papildomą informaciją du kartus sunaudojama ir daugiau resursų bei laiko.
7. Naudojant subalansuotą ėmimą ir turint neatsakymus į apklausą įverčių tikslumą galima pagerinti naudojant logistinės regresijos pagalba modeliuojamas tyrimo kintamojo reikšmes. Visais atvejais poslinkis sumažėjo 3-10 kartų.
8. Iš atliktų parametrų vertinimų galima daryti išvadą, kad praktikams, atliekantiems realius statistinius tyrimus ir renkant didesnes imtis bei žinant, kad neatsakymų bus mažai, geriau naudoti subalansuotą ėmimą bei Horvico ir Tompsono įvertinį. Naudojantis šia strategija sutaupoma tiek laiko, tiek kompiuterio resursų, o rezultatai gaunami tikslesni, nei paprastosios atsitiktinės lizdinės imties atveju. Net vertinant ir mažą sumą, kaip bedarbių skaičių visoje populiacijoje, įvertinio poslinkis mažesnis, nei kitų strategijų atveju.
9. Tiek vertinant bedarbių, tiek užimtųjų sumą populiacijoje strategijos parodo analogiškus rezultatus – kadangi populiaciją sudaro bedarbių ir užimtųjų bei neaktyvių gyventojų suma, nėra reikalo taikyti skirtingas strategijas šių dviejų parametrų vertinimui.

Literatūra

1. Bousabaa A., Lieber J., Sirolli R., *La Macro Cube*. 1999. Rennes: Tech. rept. INSEE.
2. Chadyšas V., *Statistical Estimators of the Finite Population Parameters in the Case of Sample Rotation*. 2009. Vilnius: Technika.
3. Chauvet G., Tillé Y., *Fast SAS Macros for Balancing Samples: User's Guide*. 2005. University of Neuchatel.
4. Cox D., *Planning of Experiments*. 1992. Oxford: Wiley Classics Library.
5. Čekanavičius V., Murauskas G., *Statistika ir jos taikymai III*. 2000. Vilnius: TEV.
6. Deville J. C., Grosbras J. M., Roth N., *Proceedings in Computational Statistics: Efficient Sampling Algorithms and Balanced Samples*. 1988. Heidelberg: Physica Verlag.
7. Deville J. C., Särndal C. E., Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*. 1992, 87(418): 376–382.
8. Deville J. C., Tillé Y., Unequal Probability Sampling without Replacement through a Splitting Method, *Biometrika*. 1998, 85(1): 89–101.
9. Deville J. C., Tillé Y., Efficient Balanced Sampling: The Cube Method, *Biometrika*. 2004, 91(4): 893–912.
10. Dirdaitė I., *Relationship between Balanced Sampling and Calibrated Estimator* [interaktyvus]. 2015. Helsinki [žiūrėta 2015 m. rugsėjo 24 d.]. Prieiga per internetą: https://wiki.helsinki.fi/display/banocoss2015/Presentations?preview=/149296295/170626600/Dirdaite_Abstract.pdf.
11. Gini C., Une Application de la Methode Representative aux Materiaux du Dernier Recensement de la Population Italienne, *Bulletin of the International Statistical Institute*. 1921, 23(2): 198–215.
12. Gini C., Galvani L., Di una Applicazione del Metodo Rappresentative all'ultimo Censimento Italiano della Popolazione (1 Decembri, 1921), *Annali di Statistica*. 1929, 6(4): 1–107.
13. Horvitz D. G., Thompson D. J., A Generalization of Sampling without Replacement from a Finite Universe, *Journal of the American Statistical Association*. 1952, 47(260): 663–685.
14. Yates F., Sampling Methods for Census and Surveys, *Journal of the American Statistical Association*. 1949, 44(248): 580–582.
15. Kiaer a., Observations et experiences concernant des denombrements representatifs, *Bulletin de l'Institut International de Statistique*. 1896, 9(2): 176–183.

16. Kubilius K., *Stochastiniai matematiniai modeliai* [interaktyvus]. 2012. Vilnius [žiūrėta 2014 m. gegužės 23 d.]. Prieiga per internetą: www.techmat.vgtu.lt/konspektai/stochastiniai.pdf.
17. Little R. J. A., Rubin D. B., *Statistical Analysis with Missing Data*. 2002. Hoboken: John Wiley and Sons.
18. Nedyalkova D. N., *Evaluation and Development of Strategies for Sample Coordination and Statistical Inference in Finite Population Sampling*. 2009. University of Neuchatel.
19. Neyman J., *Lectures and Conferences on Mathematical Statistics and Probability*. 1952. Washington: Graduate School, U.S. Department of Agriculture.
20. Nekrašaitė-Liegė V., *Estimation Strategy for Small Areas, a Case Study*. 2012. Vilnius: Technika.
21. Plikusas A., Krapavickaitė D., *Imčių teorijos pagrindai*. 2005. VGTU leidykla Technika.
22. Pukėnas K., *Binarinė Logistinė Regresija* [interaktyvus]. Kaunas [žiūrėta 2015 m. spalio 15 d.]. Prieiga per internetą: emokymas.lsu.lt/mod/resource/view.php?id=6268.
23. Rubin N. B., *Multiple Imputation for Nonresponse in Surveys*. 1987. New York: John Wiley and Sons.
24. Saifuddin A., *Methods in Sample Surveys*. 2009. Johns Hopkins Bloomberg School of Public Health.
25. Särndal C. E., Lundquist P., Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation, *Journal of Survey Statistics and Methodology*. 2014, 2: 361–387.
26. Särndal C. E., Swenson B., Wretman J. H., *The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total*. 1989. Oxford: Biometrika.
27. Tillé Y., *Balanced Sampling* [interaktyvus]. 2010. StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies [žiūrėta 2014 m. guodžio 10 d.]. Prieiga per internetą: <http://statprob.com/encyclopedia/BalancedSampling.html>.
28. Tillé Y., *Balanced Sampling by Means of the Cube Method* 2010. University of Neuchatel.
29. Tillé Y., *Theorie des Sondages: Echantillonnage et Estimation en Populations Finies*. 2001. Paris: Dunod.
30. Tillé Y., *Sampling Algorithms*. 2006. New York: Springer Science+Business Media.

31. Tillé Y., 10 Years of Balanced Sampling with the Cube Method: An Appraisal, *Statistics Canada*. 2011, 12(37): 215–226.

A. Priedas: *balancedcluster* funkcijos komentaras

SELECTION OF A SAMPLE OF CLUSTERS

BEGINNING OF THE FLIGHT PHASE

The matrix of balanced variable has 4 variables and 11236 units
The size of the inclusion probability vector is 11236
The sum of the inclusion probability vector is 5178
The inclusion probability vector has 11144 non-integer elements
Step 1

BEGINNING OF THE LANDING PHASE

At the end of the flight phase, there remain 4 non integer probabilities
The sum of these probabilities is 1
This sum is integer
The linear program will consider 4 possible samples
The mean cost is 0.0003915188
The smallest cost is 7.857133e-05
The largest cost is 0.0007608437
The cost of the selected sample is 0.0002619493

QUALITY OF BALANCING

	TOTALS	HorvitzThompson_estimators	Relative_deviation
pik	5178	5178.00	1.076710e-11
2	81868	81876.70	1.063234e-02
3	32962	32962.48	1.467091e-03
4	28001	27999.93	-3.811447e-03

3 pav. Komentaras, gaunamas išrinkus subalansuotą imtį

B. Priedas: Naudoti R programos kodai

B.1. Subalansuotos lizdinės imties ėmimas su Horvico ir Tompsono įvertiniu

```
#Nurodome programai kelią, kur yra nuskaitymui paruoštas failas:  
setwd("C:\\Users\\Ieva\\Magistras")
```

```
#Nuskaitome failą su apdorotais duomenimis:
```

```
library(gdata)
```

```
df = read.xls ("C:\\Users\\Ieva\\Magistras\\duom.xlsx",  
              perl = "C:\\Perl64\\bin\\perl.exe", sheet =1)
```

```
#Renkame subalansuotą lizdinę imtį:
```

```
library(sampling)
```

```
#Nurodome duomenų stulpelį, kuris padės identifikuoti lizdo numerį:  
cluster<-df[2]
```

```

cluster = unlist(cluster, use.names=FALSE)

#Pasirenkame imties dydį m:
m<-1000

#Iš pagalbinių kintamųjų susidarome atskirą kintamąjį:
X<-cbind(df[4],df[5],df[7])

#Renkame imtį. Pasirenkame selection=1, t.y. išrinkimo tikimybės
#bus proporcingos imties dydžiui. Rezultate gausime du stulpelius:
#pirmas bus sudarytas iš 1 (jei elementas išrinktas į imtį) ir
#0 priešingu atveju, antras sudarytas iš įtraukties į imtį tikimybių:
s<-balancedcluster(X,m,cluster,selection=1,comment=TRUE, method=1)

#Susikuriame kintamąjį, sudarytą iš išrinktų lizdų numerių:
lizdai<-unique(cluster[s[,1]==1])
lizdai<-unlist(lizdai, use.names=FALSE)

#Prie duomenų rinkinio df pridedame gautus rezultatus:
for (i in 1:nrow(df)){
  df[i,10]<-s[i,1]
}

for (i in 1:nrow(df)){
  df[i,11]<-s[i,2]
}

#Skaičiuojame balansavimo kintamųjų įverčius:
#Lytis: kadangi moterų kintamojo reikšmė yra 2, pakeičiame ją į 0.
a=df[df[,10]==1,]
l<-a[,5]

for (i in 1:nrow(df[df[,10]==1,])){
  if (l[i]==2){
    l[i]=0
  }
}

```

```

d<-a[,11]
vyrai=HTestimator(l,d)
l<-a[,5]

for (i in 1:nrow(df[df[,10]==1,])){
  if (l[i]==1){
    l[i]=0
  }
  if (l[i]==2){
    l[i]=1
  }
}

d<-a[,11]
moterys=HTestimator(l,d)

#Miesto bei kaimo įverčiai:
k<-a[,7]

for (i in 1:nrow(df[df[,10]==1,])){
  if (k[i]==2){
    k[i]=0
  }
}

d<-a[,11]
miestai=HTestimator(k,d)
k<-a[,7]

for (i in 1:nrow(df[df[,10]==1,])){
  if (k[i]==1){
    k[i]=0
  }
  if (k[i]==2){
    k[i]=1
  }
}

```

```

d<-a[,11]
kaimas=HTestimator(k,d)

#Amžius. Analogiškai įverčius paskaičiuojame visoms amžiaus grupėms.

f<-a[,4]
for (i in 1:nrow(df[df[,10]==1,])){
  if (f[i]==1){
    f[i]=1
  }
  else {
    f[i]=0
  }
}

d<-a[,11]
amzius1=HTestimator(f,d)

#Skaičiuojame bedarbių ir dirbančiųjų sumas įverčius, dispersijas,
#variacijos koeficientus:
g<-a[,9]
bedarbiai=HTestimator(g,d)
bed_disp=varest(Ys=g, pik=d)
bed_koef=sqrt(bed_disp)/bedarbiai

for (i in 1:nrow(df[df[,10]==1,])){
  if (g[i]==1){
    g[i]=0
  }
  else{
    g[i]=1
  }
}

d<-a[,11]
dirbantys=HTestimator(g,d)
dirb_disp=varest(Ys=g, pik=d)
dirb_koef=sqrt(dirb_disp)/dirbantys

```

B.2. Paprastasis atsitiktinis lizdinis ėmimas ir kalibruotasis įvertinys

```
#Rinksime 10 imčių, iš karto išsisaugosime kintamojo įvertį,  
#jo dispersiją bei variacijos koeficientą kiekvienai imčiai.  
#Reikšmes1 - bedarbių įverčiai, reikšmes2 - dirbančiųjų įverčiai.  
reikšmes1=matrix(, nrow = 3, ncol = 10)  
reikšmes2=matrix(, nrow = 3, ncol = 10)  
  
for (k in 1:10){  
  
library(sampling)  
  
#Imtis renkama 1000 lizdų:  
n=1000  
library(plyr)  
  
#Renkame PA lizdinę imtį:  
cl=cluster(df,clustername=c("NU_NR"),size=n,method="srswor",  
description=TRUE)  
lizdai<-unique(cl[,1])  
lizdai<-unlist(lizdai, use.names=FALSE)  
  
#Prie nuskaitytų duomenų prijungiamo įtraukties tikimybės:  
for(i in 1:nrow(df)){  
  df[i,10]=cl[1,3]  
}  
  
#Išrinktiems elementams priskiriame 1, neišrinktiems - 0:  
for (i in 1:nrow(df)){  
  df[i,11]=0  
}  
  
for (i in 1:nrow(df)){  
  for (j in 1:nrow(cl)){  
    if (i==cl[j,2]){  
      df[i,11]=1  
    }  
  }  
}
```

```

    }
}

imtis=df[df[,11]==1,]

#Susidarome pagalbinių kitnamų reikšmių matricą:
for (i in 1:nrow(imtis)){
  if (imtis[i,5]==2){
    imtis[i,5]=0
  }
}

for (i in 1:nrow(imtis)){
  if (imtis[i,7]==2){
    imtis[i,7]=0
  }
}

amzius=matrix(, nrow = nrow(imtis), ncol = 6)

for (i in 1:nrow(imtis)){
  if (imtis[i,4]==1){
    amzius[i,1]=1
  }
  else {
    amzius[i,1]=0
  }
}

for (i in 1:nrow(imtis)){
  if (imtis[i,4]==2){
    amzius[i,2]=1
  }
  else {
    amzius[i,2]=0
  }
}

```

```

for (i in 1:nrow(imtis)){
  if (imtis[i,4]==3){
    amzius[i,3]=1
  }
  else {
    amzius[i,3]=0
  }
}

for (i in 1:nrow(imtis)){
  if (imtis[i,4]==4){
    amzius[i,4]=1
  }
  else {
    amzius[i,4]=0
  }
}

for (i in 1:nrow(imtis)){
  if (imtis[i,4]==5){
    amzius[i,5]=1
  }
  else {
    amzius[i,5]=0
  }
}

for (i in 1:nrow(imtis)){
  if (imtis[i,4]==6){
    amzius[i,6]=1
  }
  else {
    amzius[i,6]=0
  }
}

X<-cbind(imtis[,5],imtis[,7],amzius)

```

```

#Skaičiuojame imties plano svorius:
d<-1/imtis[,10]

#Skaičiuojame pagalbinių kintamųjų sumas:
vyrai=sum(df[,5]==1)
miestai=sum(df[,7]==1)
amzius1=sum(df[,4]==1)
amzius2=sum(df[,4]==2)
amzius3=sum(df[,4]==3)
amzius4=sum(df[,4]==4)
amzius5=sum(df[,4]==5)
amzius6=sum(df[,4]==6)
total=c(vyrai,miestai,amzius1,amzius2,amzius3,amzius4,amzius5,amzius6)

#Skaičiuojame svorius:
gsvoriai=calib(Xs=X, d, total,q=rep(1,length(d)),method="linear",
               description=TRUE, max_iter=10)

#Pažiūrime, ar gauti svoriai galimi naudoti kalibruojant:
checkcalibration(X,d,total, gsvoriai)

#Susidarome porų įtraukties tikimybių matricą:
pikl=matrix(, nrow = nrow(imtis), ncol = nrow(imtis))
tik1=n/11236
tik2=n*(n-1)/(11236*11235)

for (i in 1:nrow(imtis)){
  for (j in 1:nrow(imtis)){
    if (imtis[i,2]==imtis[j,2]){
      pikl[i,j]=tik1
    }
    else{
      pikl[i,j]=tik2
    }
  }
}

#Vertiname bedarbių skaičių:

```

```

ats=calibev(Ys=imtis[,9], Xs=X, total, pikl, d, g=gsvoriai,
            q=rep(1,length(d)), EPS=1e-6)
reiksmes1[1,k]=ats$scalest
reiksmes1[2,k]=ats$evar
reiksmes1[3,k]=sqrt(ats$evar)/ats$scalest

#Vertiname dirbančiųjų skaičių:
for (i in 1:nrow(imtis)){
  if (imtis[i,9]==0){
    imtis[i,9]=1
  }
  else {
    imtis[i,9]=0
  }
}

ats=calibev(Ys=imtis[,9], Xs=X, total, pikl, d, g=gsvoriai,
            q=rep(1,length(d)), EPS=1e-6)
reiksmes2[1,k]=ats$scalest
reiksmes2[2,k]=ats$evar
reiksmes2[3,k]=sqrt(ats$evar)/ats$scalest
}

```

B.3. Tikimybių perskaičiavimas, kai yra neatsakymų

```

#Tarkime 10 procentų lizdų neatsako į apklausą. Atsitiktinai
#juos parinkę,
#išmetame iš išrinktos imties:
lizdats=matrix(,nrow=2, ncol=m)
ats=runif(m,0,1)

for (i in 1:m){
  lizdats[1,i]=lizdai[i]
  lizdats[2,i]=ats[i]
}

a=order(lizdats[2,], decreasing=TRUE)
lizdats2=matrix(,nrow=2, ncol=m)

```

```

for(i in 1:m){
  lizdats2[1,i]=lizdats[1,a[i]]
  lizdats2[2,i]=lizdats[2,a[i]]
}

mm=round(m*0.9,0)
lizdai2=matrix(,nrow=1, ncol=mm)

for (i in 1:mm){
  lizdai2[i]=lizdats2[1,i]
}

ismesti=matrix(,nrow=m-mm, ncol=1)
sk=m-mm

for(i in 1:sk){
  ismesti[i,1]=lizdats[1,a[m-i+1]]
}

for (i in 1:sk){
  for(j in 1:nrow(df)){
    if (ismesti[i,1]==df[j,2]){
      df[j,11]=0
    }
  }
}

#Perskaičiuojame išrinktų į imtį ir atsakiusių lizdų tikimybes:
imtis=df[df[,11]==1,]

for (i in 1:nrow(df)){
  if (df[i,11]==1){
    df[i,10]=df[i,10]*0.9
  }
}

```

B.4. Neatsakymų į apklausą modeliavimas, taikant logistinę regresiją ir Bernulio atsitiktinius dydžius

```
#Atskiriame duomenis tų asmenų, kurie buvo išrinkti į imtį:
a=df[df[,10]==1,]

#Atsitiktiniu būdu išrenkame 10\% lizdų, kurie "neatsakys"
#į apklausą:
lizdats=matrix(nrow=2, ncol=m)
ats=runif(m,0,1)

for (i in 1:m){
  lizdats[1,i]=lizdai[i]
  lizdats[2,i]=ats[i]
}

aa=order(lizdats[2,], decreasing=TRUE)
lizdats2=matrix(,nrow=2, ncol=m)

for(i in 1:m){
  lizdats2[1,i]=lizdats[1,aa[i]]
  lizdats2[2,i]=lizdats[2,aa[i]]
}

mm=round(m*0.9,0)
lizdai2=matrix(,nrow=1, ncol=(m-mm))

for (i in 1:(m-mm)){
  lizdai2[i]=lizdats2[1,i]
}

for(j in 1:(m-mm)){
  for(i in 1:nrow(a)){
    if (a[i,2]==lizdai2[j]){
      a[i,10]<-0
    }
  }
}
}
```

```

neatsake=a[a[,10]==0,]
atsake=a[a[,10]==1,]

#Sukuriame logistinės regresijos pagalbinių kintamųjų matricą:
pagalbiniai=matrix(,nrow=nrow(atsake), ncol=8)

#Vyro kintamasis:
for (i in 1:nrow(atsake)){
  if (atsake[i,5]==2){
    atsake[i,5]=0
  }
}

#Miesto kintamasis:
for (i in 1:nrow(atsake)){
  if (atsake[i,7]==2){
    atsake[i,7]=0
  }
}

amzius=matrix(, nrow = nrow(atsake), ncol = 6)

#Amžiaus kintamasis:
for (i in 1:nrow(atsake)){
  if (atsake[i,4]==1){
    amzius[i,1]=1
  }
  else {
    amzius[i,1]=0
  }
}

for (i in 1:nrow(atsake)){
  if (atsake[i,4]==2){
    amzius[i,2]=1
  }
  else {

```

```

        amzius[i,2]=0
    }
}

for (i in 1:nrow(atsake)){
    if (atsake[i,4]==3){
        amzius[i,3]=1
    }
    else {
        amzius[i,3]=0
    }
}

for (i in 1:nrow(atsake)){
    if (atsake[i,4]==4){
        amzius[i,4]=1
    }
    else {
        amzius[i,4]=0
    }
}

for (i in 1:nrow(atsake)){
    if (atsake[i,4]==5){
        amzius[i,5]=1
    }
    else {
        amzius[i,5]=0
    }
}

for (i in 1:nrow(atsake)){
    if (atsake[i,4]==6){
        amzius[i,6]=1
    }
    else {
        amzius[i,6]=0
    }
}

```

```

}

pagalbiniai<-cbind(atsake[,5],atsake[,7],amzius,atsake[,9])
pagalbiniai=as.data.frame(pagalbiniai)

#Pritaikome logistinę regresiją bedarbio tikimybėms gauti:
mylogit <- glm(pagalbiniai[,9] ~ pagalbiniai[,1]+pagalbiniai[,2] +
  pagalbiniai[,3]+pagalbiniai[,4]+pagalbiniai[,5]+pagalbiniai[,6]+
  pagalbiniai[,7]+pagalbiniai[,8], data = pagalbiniai,
  family = binomial(link="logit"))

for(i in 1:nrow(neatsake)){
  if(neatsake[i,4]==1){
    neatsake[i,12]=mylogit$coefficients[1] +
      neatsake[i,5]*mylogit$coefficients[2]+
      neatsake[i,7]*mylogit$coefficients[3]+
      mylogit$coefficients[4]
  }
  else {
    if (neatsake[i,4]==2){
      neatsake[i,12]=mylogit$coefficients[1] +
        neatsake[i,5]*mylogit$coefficients[2]+
        neatsake[i,7]*mylogit$coefficients[3]+
        mylogit$coefficients[5]
    }
    else {
      if (neatsake[i,4]==3){
        neatsake[i,12]=mylogit$coefficients[1] +
          neatsake[i,5]*mylogit$coefficients[2]+
          neatsake[i,7]*mylogit$coefficients[3]+
          mylogit$coefficients[6]
      }
      else {
        if (neatsake[i,4]==4){
          neatsake[i,12]=mylogit$coefficients[1] +
            neatsake[i,5]*mylogit$coefficients[2]+
            neatsake[i,7]*mylogit$coefficients[3]+
            mylogit$coefficients[7]
        }
      }
    }
  }
}

```

```

    }
    else {
      if (neatsake[i,4]==5){
        neatsake[i,12]=mylogit$coefficients[1] +
          neatsake[i,5]*mylogit$coefficients[2]+
          neatsake[i,7]*mylogit$coefficients[3]+
          mylogit$coefficients[8]
      }
      else {
        neatsake[i,12]=mylogit$coefficients[1] +
          neatsake[i,5]*mylogit$coefficients[2]+
          neatsake[i,7]*mylogit$coefficients[3]
      }
    }
  }
}

```

```

for (i in 1:nrow(neatsake)){
  logar<-neatsake[i,12]
  neatsake[i,13]<-exp(logar)/(1+exp(logar))
}

```

#Modeliuojame Bernulio atsitiktinius dydžius naudojantis gautomis
#naujomis tikimybėmis. Juos naudosime vietoj tikrųjų neatsakiusiųjų
#reikšmių:

```

for (i in 1:nrow(neatsake)){
  reiks=rbinom(1,1,neatsake[i,13])
  neatsake[i,14]=reiks
}

```

```

for (i in 1:nrow(a)){
  a[i,12]=a[i,9]
}

```

```

for (i in 1:nrow(a)){

```

```

for (j in 1:nrow(neatsake)){
  if (as.numeric(rownames(a[i,]))==
    as.numeric(rownames(neatsake[j,]))){
    a[i,12]=neatsake[j,14]
  }
}
}

#Skaičiuojame bedarbių ir dirbančiųjų įverčius:
d<-a[,11]
g<-a[,12]
bedarbiai=HTestimator(g,d)
bed_disp=varest(Ys=g, pik=d)
bed_koef=sqrt(bed_disp)/bedarbiai

for (i in 1:nrow(a)){
  if (g[i]==1){
    g[i]=0
  }
  else{
    g[i]=1
  }
}

d<-a[,11]
dirbantys=HTestimator(g,d)
dirb_disp=varest(Ys=g, pik=d)
dirb_koef=sqrt(dirb_disp)/dirbantys

```