Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kätrin Suvi

Vastanute hulga tasakaalustamine hinnangute täpsustamiseks

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: Imbi Traat, *Ph.D*

Tartu 2017

Vastanute hulga tasakaalustamine hinnangute täpsustamiseks

Tänapäeval on valikuuringutes väga levinud probleemiks mittevastamine, mis toob endaga kaasa kallutatud vastanute hulga ja nihkega hinnangud üldkogumi uuritavatele tunnustele. Käesoleva bakalaureusetöö eesmärgiks on uurida, kas valikuuringute andmekogumisprotsessis valitud abitunnuste abil vastanute hulga tasakaalustatuse jälgimine koguvalimi suhtes hoiab hinnangute nihked madalamad. Esmalt tuuakse ülevaade vastanute hulga tasakaalustamisega seotud mõistetest. Seejärel selgitatakse kahte vastanute hulga tasakaalustamise meetodit - lävendimeetod ning fikseeritud osakaalu meetod. Töö teises osas viiakse läbi simulatsiooniülesanne selgitamaks, kas vastanute hulga tasakaalustamiseks tehtud jõupingutused ka reaalselt täpsemaid hinnanguid annavad. Vastanute hulga genereerimiseks kasutatakse antud töös fikseeritud osakaalu meetodit. Tulemused kinnitavad, et selline andmekogumismeetod tasakaalustab vastanute hulka ja vähendab hinnangute nihkeid.

Märksõnad: valikuuringud, statistiline andmetöötlus, matemaatiline statistika

CERCS teaduseriala: Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika (P160)

Balancing the responding sample to improve the estimates

Nowadays, nonresponse is a common problem in survey sampling. This causes disproportionate response set with respect to the full sample and leads to biased estimates. The purpose of this thesis is to analyze whether the efforts durning data collection to balance the survey responce with respect to selected auxiliary variables will improve the chances for low nonresponse bias in the estimates. Firstly, an overview of the imbalance notion will be given. Then we introduce and describe two balancing methods - the Treshold method and the Fixed proportion method. In the second part of this thesis, a simulated data collection process will be carried out on real data, using fixed proportion method. Results confirm that this data collection method reduces bias in the estimates.

Keywords: sample surveys, statistical data processing, mathematical statistics

CERCS research specialisation: Statistics, operation research, programming, actuarial mathematics (P160)

Sisukord

Sissejuhatus **4**

1 Vastanute hulga tasakaalustamine **5**

1.1 Tähistused 5

1.2 Abivektorid 6

1.3 Tasakaalu mõõtmine 7

1.4 Tasakaalumõõdu jälgimise meetodid 9

1.4.1 Lävendimeetod 10

1.4.2 Fikseeritud osakaalu meetod 11

2 Praktiline ülesanne **12**

2.1 Andmestiku kirjeldus 12

2.2 Ülesande püstitus 12

2.3 Ülesande käik 13

2.3.1 Vastamistõenäosuste leidmine 13

2.3.2 Vastanute hulkade genereerimine 16

2.4 Tulemused 17

3 Kokkuvõte **21**

Viited **22**

Lisa – R-i kood **23**

Sissejuhatus

Käesoleva bakalaureusetöö eesmärgiks on uurida, kas valikuuringute andmekogumisprotsessis tasakaalustatuse järgimine hoiab hinnangute nihked madalamad.

Tänapäeval on valikuuringute suurimaks probleemiks suur kao protsent ja sellest tulenevalt ka kallutatud vastanute hulk. Uuringu andmetes esineb puuduvaid väärtusi, kuna valimisse sattunud objekt jätab mõnele küsimusele vastamata, ei soovi üldse küsimustele vastata või ei õnnestu teda küsitluse vältel kätte saada. Sellest tulenebki kallutatud vastanute hulk, mis mitmete näitajate osas on ebaproportsionaalne algselt võetud valimi suhtes, ning andmetöötluse tulemusena saadakse nihkega hinnangud.

Carl-Erik Särndal [3]-[4] on välja töötanud mitmeid indikaatoreid, mis mõõdavad vastanute hulga tasakaalu abitunnuste suhtes. Indikaatorite abil võrreldakse abitunnuste keskmisi vastanute hulgas ning kogu valimis. Kui keskmised on lähedased, on vastanute hulk tasakaalus. Kasutusele on võetud ka uued andmekogumismeetodid (*adaptive sampling*), mis protsessi käigus muutuvad eesmärgiga saavutada tähtaja lõpuks suurem tasakaalustatus. Samas ei ole üheselt tõestatud, et jõupingutused vastanute hulga tasakaalustamiseks tagavad täpsemad hinnangud võrreldes olukorraga, kus mingeid jõupingutusi ei tehta, aga kasutatakse kalibreerimist abiinformatsiooniga hindamise etapil. Teoreetiliselt on antud probleemi raske uurida. Tasakaalustamise positiivset mõju on näidatud vaid erijuhtudel [2]. Antud töös uurime praktilise näite varal, kas tasakaalu jälgimine hoiab ka hinnangute nihked väiksemad. Kasutame andmete kogumise tasakaalustavat meetodit, mida Carl-Erik Särndal ja Peter Lundquist artiklis [5] nimetavad fikseeritud osakaalu meetodiks.

Töö esimeses peatükis on toodud ülevaade teemaga seotud mõistetest ning tasakaalustamise protsessidest. Teises osas viiakse läbi praktiline ülesanne ning tuuakse ülevaade analüüsi tulemustest.

Bakalaureusetöö kirjutamiseks on kasutatud tekstitöötlus programmi *LaTeX*. Analüüsid on läbi viidud statistikapaketiga *RStudio*. Praktilises ülesandes kasutatavad andmed on saadud *European Social Survey* [7] kodulehelt.

Käesolevaga tänab autor bakalaureusetöö juhendajat Imbi Traati rohkete nõuannete, suunamiste ja paranduste, eriliselt aga entusiastliku koostöö ja pühendatud aja eest. Samuti Natalja Lepikut antud töös kasutatavate andmete jagamise eest.

1 Vastanute hulga tasakaalustamine

Teadagi ei ole ükski uuring kunagi läbi viidud ideaalsetes tingimustes. Tihti esineb mittevastamist, mille tulemusena saadakse vastanute hulk, mis ei ole vaadeldavate tunnuste suhtes tasakaalus algse valimiga. Kallutatud vastanute hulk omakorda suurendab andmetöötlusel saadavate hinnangute nihkeid. Üheks võimaluseks vähendada mittevastamisest tulenevaid nihkeid on hoida küsitluse vältel vastanute hulk tasakaalus valimiga. Järgnevalt uurime uusi väljatöötatud meetodeid saavutamaks tähtaja lõpuks suurem tasakaalustatus. Praktilises ülesandes analüüsime, kas tasakaalustamine toob kaasa ka väiksemad hinnangute nihked.

1.1 Tähistused

Tähistame üldkogumi sümboliga U ja valimi sümboliga s. Olgu üldkogumi maht N ja valimi maht n. Uuringu eesmärgile vastavalt valitakse sobiv valikudisain. Igal üldkogumi U objektil k on valikudisainiga määratud kaasamistõenäosus $\pi_k = \Pr(k \in s)$. On teada valikukaal, mis võrdub kaasamistõenäosuse pöördväärtusega, tähistame $d_k = \frac{1}{\pi_k}$. Küsitluse läbiviimisel esineb mittevastamist, siinkohal tähistame vastanute hulga sümboliga r ja vastanute arvu sümboliga m. Ei ole teada, kuidas vastanute hulk r on genereeritud valimist s, seega on vastamistõenäosused tundmatud. Vastanute hulk r rahuldab tingimust $r \subseteq s \subset U$ ehk et r on valimiga s võrdne või selle osahulk, s omakorda on üldkogumi U osahulk ja r ei ole tühi hulk. Vastamismäär üldistatud kujul on defineeritud valemiga

$$P = \frac{\sum_r d_k}{\sum_s d_k} \ . \tag{1}$$

Paneme tähele, et summa $\sum_{k \in A}$ kirjutatakse kujul $\sum_A$. Kui kaasamistõenäosused on võrdsed, $\pi_k = \frac{n}{N}$, saadakse siit tuntud vastamismäär $P = \frac{m}{n}$. Vastamismäär P kasvab andmete kogumise protsessi edenedes. Teoreetiliselt on võimalik, et $r = s$ ning vastanute hulk on valimisse sattunutega võrdne. Enamasti on aga andmete kogumise aeg piiratud ning lõpetatakse enne, kui kõik valimisse sattunud objektid vastavad. P väärtus 0.5 või väiksem on üsna tavapärane andmete kogumise lõppedes.

Uuringus võib esineda mitmeid tunnuseid, mis pakuvad huvi. Tähistame tüüpilise uuritava tunnuse sümboliga y ja tema väärtuse k-nda objekti korral sümboliga y_k.

Kui $k \in r$, siis on y_k teada, kui $k \in s - r$, siis y_k väärtus puudub, see tähendab ei ole uurijale teada. Eesmärk on hinnata üldkogumi y-kogusummat, $Y = \sum_U y_k$. Olgu vastamisindikaatoriks I, millel on väärtus $I_k = 1$, kui küsitluses osaleja on vastanud ($k \in r$) või $I_k = 0$, kui vastus puudub ($k \in s - r$). Eesmärk praktikas on saada vastajate hulk r, mis teatud tunnuste suhtes on hästi tasakaalustatud. Tasakaalu mõistet täpsustame hiljem. Märgime, et tänu lünkadele võib valimist s erinevate tunnuste puhul saada erinevaid vastajate hulkasid r.

1.2 Abivektorid

Abivektorite kirjeldus on refereeritud Nora Roosilehe bakalaureusetööst [1].

Andmete kogumise protsessi jälgimiseks on tarvis kasutada abiinformatsiooni. Enamasti pärineb abiinformatsioon erinevatest registritest. Leitakse tunnused, mis on teada nii mittevastanute kui ka vastanute kohta. Abiinformatsioonina kasutatavaid tunnuseid võib olla palju. Moodustub abivektor $\mathbf{x}_k = \Big(x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk}\Big)'$, kus x_{jk} on k-nda objektiga seotud j-nda abitunnuse väärtus. Abivektori elemendid võivad olla pidevad või diskreetsed arvulised tunnused, sealhulgas binaarsed $(0, 1)$ tunnused, kus 0 märgib omaduse puudumist ja 1 selle esinemist.

Ka mittearvulist ehk kvalitatiivset tunnust on võimalik kasutada abivektoris binaarsete tunnuste abil. Näiteks $J \geq 2$ tasemega kvalitatiivse tunnuse korral moodustatakse objektile k (0,1) väärtustega J-dimensionaalne vektor

$$\mathbf{x}_k = \Big(\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk}\Big)' = \Big(0, \ldots, 1, \ldots, 0\Big)',$$

kus $\gamma_{jk} = 1$, kui objektil k esineb omadus j, muidu $\gamma_{jk} = 0$. Kui i-ndal kvalitatiivsel tunnusel on J_i võimalikku väärtust, kus $i = 1, \ldots, I$, paigutatakse need vektoris $\mathbf{x}_k$ üksteise kõrvale ja abivektori mõõtmeks on $J = 1 + \sum_{i=1}^{I}(J_i - 1)$. Abivektori kovariatsiooni maatriksi singulaarsuse vältimiseks eemaldatakse vektorist $\mathbf{x}_k$ mittearvulise tunnuse üks väärtustest. Näiteks olgu tunnuse kodakondsus jaoks neli erinevat väärtust - eesti, vene, ukraina, muu. Saame objekti k jaoks abivektori $\mathbf{x}_k = (\gamma_{1k}; \gamma_{2k}; \gamma_{3k})$, kus $\gamma_{1k} = 1$, kui objektil k on eesti kodakondusus, $\gamma_{2k} = 1$, kui objektil k on vene kodakondus. Kui objetkil k on ukraina kodakondsus, siis $\gamma_{3k} = 1$. Kui $\gamma_{1k} = \gamma_{2k} = \gamma_{3k} = 0$, siis on objektil k muu kodakondsus.

1.3 Tasakaalu mõõtmine

Mõistet *tasakaal* kasutatakse viidates kahe kogumi elementide teatud tunnuse keskmiste võrdsusele, kusjuures üks kogum võib olla teise alamhulk. Vastanute hulk r on igal ajahetkel andmete kogumise vältel valimi s rohkem või vähem tasakaalustatud esindushulk. Ütleme, et vastanute hulk r on tasakaalus, kui valitud abitunnuste keskmised on võrdsed vastanute hulgas r ja kogu valimis s [1]. Seoses mittevastamisega soovime mõõta, kui hästi tasakaalus on vastanute hulk r võrreldes valimiga s, mis oleks andnud meile nihketa hinnangud. Antud abivektoril $\mathbf{x}$ on võimalik arvutada keskmised vastanute hulgale r ja valimile s. Defineerime J-dimensionaalsed keskmiste vektorid vastanute hulgas ning valimis järgmiselt:

$$\overline{\mathbf{x}}_r = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k}, \tag{2}$$

$$\overline{\mathbf{x}}_s = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k}. \tag{3}$$

Antud vektorite puhul on tegemist kaalutud keskmistega. Kui keskmised on võrdsed, siis vastanute hulk r on abivektori $\mathbf{x}_k$ suhtes valimiga s perfektselt tasakaalus. See on väga ebatõenäoline tulemus, kuid andmete kogumise protsessis võime jõuda sellele ligilähedale.

Keskmiste vahet näitab vektor $\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s$, mis koosneb elementidest $\overline{x}_{jr} - \overline{x}_{js}$, kus $j = 1, \ldots, J$. Elemendid näitavad erinevust j-nda x-tunnuse vastanute hulga keskmise, $\overline{x}_{jr} = \sum_r d_k x_j k \Big/ \sum_r d_k$, ja valimi keskmise, $\overline{x}_{js} = \sum_s d_k x_j k \Big/ \sum_s d_k$, vahel [5].

Suur erinevus $\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s$ näitab, et vastanute hulk ei ole tasakaalus valimiga. See on mingil määral mõjutatud ka suurest mittevastamise määrast $1 - P$, kuna $\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s = (1 - P)(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_{s-r})$, kus $\overline{\mathbf{x}}_{s-r} = \sum_{s-r} d_k \mathbf{x_k} \Big/ \sum_{s-r} d_k$ näitab mittevastanute hulga $s - r$ keskmist.

Erinevust vastanute hulga r ja valimi s vahel on võimalik mõõta skalaarsete näitajate abil [2] :

$$Q_s = \Big(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\Big)' \Sigma_s^{-1} \Big(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\Big) \ ; \ Q_r = \Big(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\Big)' \Sigma_r^{-1} \Big(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\Big) \ . \tag{4}$$

Ainus erinevus Q_s ja Q_r vahel seisneb $J \times J$ kaalumaatriksites

$$\Sigma_s = \frac{\sum_s d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_s d_k}, \tag{5}$$

$$\Sigma_r = \frac{\sum_r d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_r d_k}. \tag{6}$$

Mõlemad maatriksid eeldatakse olevat mitte-singulaarsed. Eriti tähtis on Q_s, mida kasutame leidmaks tasakaalumõõtu vastanute hulgale r kindlaksmääratud abivektori $\mathbf{x}$ suhtes. Defineerime tasakaalumõõdu järgmiselt [5]:

$$IMB(r, \mathbf{x} \mid s) = P^2\Big(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\Big)' \Sigma_s^{-1} \Big(\overline{\mathbf{x}}_r - \overline{\mathbf{x}}_s\Big) = P^2 Q_s \tag{7}$$

kus P on vastamise määr (1). Enamasti kasutame lühitähistust $IMB(r, \mathbf{x} \mid s) = IMB$, mis tuleneb inglise keelsest sõnast *imbalance*, mis tähendab tasakaalustamatust. Siinkohal tasub märkida, et antud töös räägime tasakaalust keelelise lihtsustamise eesmärgil. Väärtus $IMB = 0$ väljendab tasakaalu hulkade r ja s vahel. IMB kasvades suureneb hulga r tasakaalutus. Tasakaalumõõt (7) sõltub objetide karakteristikutest (väärtused $\mathbf{x}_k$) nii vastanute kui ka mittevastanute hulgas. Siinkohal tuleb silmas pidada, et tasakaalumõõt on arvutatud kindlal ajahetkel andmete kogumise protsessi käigus realiseerunud vastanute hulga, $\mathbf{x}$-vektori ja valimi põhjal. Ühe ja sama vastanute hulga r puhul võib tasakaalumõõt olla erinev, olenevalt kui mitu ning millised abitunnused on $\mathbf{x}$-vektorisse valitud [5]. Tasakaalumõõt (7) annab märgatavalt parema ülevaate uuringu käigus kogutud andmetest, kui lihtsalt vastamise määr (1), mis üksi on ebapiisav ja vaid kaudne indikaator kirjeldamaks vastanute hulga kvaliteeti.

Isegi juhtudel, kui r on vaid väike osa valmist s võib erandkorras juhtuda, et $IMB = 0$, kui ideaalne tasakaal $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$ kehtib. Tavaliselt andmete kogumisel suurenev vastamismäär P toob vastanute hulga keskmised $\overline{\mathbf{x}}_r$ (2) lähemale fikseeritud valimi keskmisele $\overline{\mathbf{x}}_s$ (3). Suure vastamismäära puhul on r ligilähedane valimile s, seega $\overline{\mathbf{x}}_r \approx \overline{\mathbf{x}}_s$.

Särndal ja Lundquist [5] on kasutanud kaalumaatriksit Σ_s valemis (7), kuna see annab tasakaalumõõdule ülemise tõkke. Antud valimi s juures saame, et $0 \leq IMB \leq P(1-P)$ suvalise vastanute hulga r ja väärtuste $\mathbf{x}_k$ korral ($k \in s$). Näiteks, olgu mittevastamine $1 - P = 0.1$, siis $0 \leq IMB \leq 0.09$. Kui aga $1 - P = 0.5$, siis $0 \leq IMB \leq 0.25$. Veel tuuakse välja, et andmete kogumise praktikas on siiani tõdetud, et enamasti on IMB palju väiksem, kui antud ülemine piir.

1.4 Tasakaalumõõdu jälgimise meetodid

Järgnev peatükk on refereeritud Särndali ja Lundquisti artiklist [5].

Olgu $\mathbf{x}$ abitunnuste vektor, mille väärtuste abil jälgitakse andmete kogumise vältel vastajate hulga kallutatust. Iga objekti $k \in s$ kohta on teada väärtus $\mathbf{x}_k$. Seega vektoril $\mathbf{x}$ on spetsiaalne ülesanne juhtida andmete kogumist. Me eeldame, et $\mathbf{x}_k$ kohta kehtib ka omadus $\lambda'\mathbf{x}_k = 1$ iga k ja konstantse vektori λ korral. See ei ole eriti kitsendav tingimus, samas aga lihtsustab tuletuskäike.

Tasakaalustatud vastanute hulga saamiseks uuritakse tasakaalumõõtu andmekogumisprotsessi teatud hetkedel, nimetame neid vahelesegamispunktideks. Neis punktides muudetakse andmekogumisprotsessi, mida kirjeldame järgnevates alapunktides. Andmekogumisprotsessi muutmiseks kasutatakse hinnatud vastamistõenäosusi. Tähistame k-nda objekti hinnatud vastamistõenäosuse $\hat{\theta}_k$ ja toome välja tuletuskäigu nende suuruste leidmiseks. Kasutame vastamisindikaatori lineaarset modelleerimist [6]. Kõigepealt hindame mudeli parameetreid $\beta = (\beta_0, \beta_1, \ldots, \beta_J)$ kasutades vähimruutude meetodit, see tähendab minimiseerime β suhtes summa $\sum_s (I_k - \beta'\mathbf{x}_k)^2$. Saame lahendiks

$$\hat{\beta}' = \Big(\sum_s d_k \mathbf{x}_k' I_k\Big)\Big(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1},$$

kus I_k tähistab vastamise indikaatorit. Edasi leiame prognoosi $\hat{\beta}'\mathbf{x}_k$, mis ongi hinnang vastamistõenäosusele:

$$\begin{aligned}\hat{\theta}_k = \hat{\beta}'\mathbf{x}_k &= \Big(\sum_s d_k \mathbf{x}_k' R_k\Big)\Big(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\Big)^{-1}\mathbf{x}_k = \Big(\sum_r d_k \mathbf{x}_k'\Big)\Big(\Sigma_s \cdot \sum_s d_k\Big)^{-1}\mathbf{x}_k = \\ &= \Big(\overline{\mathbf{x}}_r' \cdot \sum_r d_k\Big)\Big(\Sigma_s \cdot \frac{1}{\sum_s d_k}\Big)\mathbf{x}_k = P\overline{\mathbf{x}}_r'\Sigma_s^{-1}\mathbf{x}_k,\end{aligned} \tag{8}$$

kus $\overline{\mathbf{x}}_r$ (2) tähistab keskmiste vektorit vastanute hulgas, Σ_s (5) on $J \times J$ kaalumaatriks.

1.4.1 Lävendimeetod

Andmekogumisprotsessis toimub teatud punktides vahelesegamine. Esimeses vahelesegamispunktis arvutatakse vastamistõenäosuste hinnangud $\hat{\theta}_k$ (8) kõikide objektide $k \in s$ jaoks. Vastanute ja mittevastanute hulgas tuvastatakse need, kel $\hat{\theta}_k$ on suurem, kui varasemalt fikseeritud lävend, näiteks 0.6 (60%). Nende objektidega kontakti luua ei üritata, niiöelda "jäetakse kõrvale" kuni järgmise vahelesegamispunktini.

Teises vahelesegamispunktis arvutatakse $\hat{\theta}_k$ uuesti iga $k \in s$ korral. Eelnevas punktis kõrvale jäetud objektide $\hat{\theta}_k$ võivad olla nüüd mõnevõrra muutunud. Kui ka sel korral on nende vastamistõenäosuse hinnang kõrgem kui lävend, jäetakse need objektid jälle kõrvale. Lisaks tuvastatakse ka uued objektid, kelle $\hat{\theta}_k$ on kõrgem, kui fikseeritud lävend ning jäetakse samuti kõrvale. Sel viisil jätkatakse sama tegevust igas vahelesegamispunktis. Suurused $\hat{\theta}_k$ arvutatakse uuesti igas punktis iga $k \in s$ objekti jaoks, kui $\hat{\theta}_k$ on kõrgem, kui määratud lävend, siis k-s objekt jäetakse kõrvale. Viimases vahelesegamispunktis alles jäänud objektidega üritatakse kontakti luua ning vastuseid kätte saada kuni andmete kogumise aja lõpuni.

Näiteks olgu meil valim s mahuga $n = 10$, objektid nummerdatud $k = 1$ kuni 10. Olgu määratud kaks vahelesegamispunkti. Esimeses vahelesegamispunktis oletame, et objektidel $k = 1, 2, 3, 4$ on $\hat{\theta}_k$ kõrgem, kui lävend. Neist objektid $k = 1, 2$ on vastanud, $k = 3, 4$ on mittevastajad. Ülejäänud objektidega $k = 5, 6, 7, 8, 9, 10$ üritatakse kontakti saada. Teises vahelesegamispunktis arvutatakse $\hat{\theta}_k$ igale objektile uuesti. Olgu nüüd objektide $k = 5, 6, 7$ uued hinnangud $\hat{\theta}_k$ ka kõrgemad, kui lävend. Neist olgu $k = 6, 7$ vastanud, $k = 5$ mittevastaja. Objektidega $k = 8, 9, 10$ üritatakse kontakti saada kuni andmete kogumise aja lõpuni. Oletame, et objektid $k = 8, 9$ vastasid, kuid $k = 10$-ga ei õnnestunud kontakti saada. Lõplik vastanute hulk on seega $r = \{1, 2, 6, 7, 8, 9\}$ ning mittevastanute hulk $s - r = \{3, 4, 5, 10\}$.

Lävendi määramisel tuleb toetuda varasemalt läbi viidud kas sama uuringu või sarnase uuringu tulemustele. Kui uuringu vastamise määr (1) on reaalselt kuskil 65% juures, siis tuleks kasutada lävendit 60% või 55%.

1.4.2 Fikseeritud osakaalu meetod

Selle meetodi puhul jäetakse igas vahelesegamispunktis kõrvale kindlaksmääratud osa valimist s. Vastamistõenäosuste hinnangud $\hat{\theta}_k$ arvutatakse igale objektile $k \in s$ igas vahelesegamispunktis. Suuruste $\hat{\theta}_k$ keskmine tõuseb iga vahelesegamisega. Samuti on suuruste $\hat{\theta}_k$ keskmine võrdne vastamismääraga [5]:

$$P = \frac{\sum_s d_k \hat{\theta}_k}{\sum_s d_k} = \frac{\sum_s d_k P \overline{\mathbf{x}}_r' \Sigma_s^{-1} \mathbf{x}_k}{\sum_s d_k} = \frac{P \overline{\mathbf{x}}_r' \Sigma_s^{-1} \sum_s d_k \mathbf{x}_k}{\sum_s d_k} = P \overline{\mathbf{x}}_r' \Sigma_s^{-1} \Big(\frac{\sum_s d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_s d_k} \Big) \cdot \lambda =$$
$$= P \overline{\mathbf{x}}_r' \Sigma_s^{-1} \Sigma_s \lambda = P \overline{\mathbf{x}}_r' I \lambda = P \frac{\sum_r d_k \mathbf{x}_k^{prime\lambda}}{\sum_r d_k} = P \frac{\sum_r d_k}{\sum_r d_k} = P = \frac{\sum_r d_k}{\sum_s d_k},$$

kus kasutame eelnevalt välja toodud vektori $\mathbf{x}_k$ omadust $\lambda' \mathbf{x}_k = \mathbf{x}_k' \lambda = 1$.

Vahelesegamispunktis järjestatakse objektid $k \in s$ hinnangu $\hat{\theta}_k$ järgi kahanevalt. Valimi mahust $100/(L+1)$ protsenti objekte, suurimate $\hat{\theta}_k$-väärtustega, jäetakse iga vahelesegamise korral välja, neid hiljem edasistes vahelesegamispunktides ei arvestata.

Teeme siinkohal läbi ühe näite fikseeritud osakaalu meetodist. Olgu meil vahelesegamispunkte 4 ja tähistagu L ees ootavaid vahelesegamispunkte. Siis igas vahelesegamispunktis välja jäetavate objektide osakaal on $1/5$ valimist s. Seega esimesel vahelesegamisel jäetakse välja $\Big(100/(L+1)\Big)\% = (100/5)\% = 20\%$ suurima $\hat{\theta}_k$-väärtusega objekti. Neile objektidele enam ei läheneta. Teise vahelesegamise korral arvutatakse alles jäänud $5/6$ objektidele valimist s uuesti $\hat{\theta}_k$-väärtusted. Nüüd on ees ootavaid vahelesegamise punkte veel 3, seega $L = 3$. $\Big(100/(L+1)\Big)\% = (100/4)\% = 25\%$ kõrgeima vastamistõenäosuse hinnanguga $\hat{\theta}_k$ objekti jäetakse selles vahelesegamispunktis jällegi välja. Ja nii edasi, kuni neljanda vahelesegamise korral on alles jäänud 20% objekti valimist s, kellega üritatakse kontakti luua kuni andmete kogumise aja lõpuni.

2 Praktiline ülesanne

2.1 Andmestiku kirjeldus

Antud bakalaureusetöö praktilises ülesandes on kasutatud *European Social Survey* 2014. aasta andmeid Eesti kohta [7]. Küsitlus on läbi viidud silmast silma intervjuu vormis. Uuringu käigus küsiti vastajatelt väga mitmeid küsimusi erinevatest valdkondadest. Andmestikus on kokku 2051 Eesti elanikku vanuses 15-99, kelle kohta mõõdeti 253 tunnust. Käesolevas töös vähendati objektide hulka 1809 küsitletuni puuduvate andmete tõttu. Oma ülesandes kasutame järgnevaid tunnuseid:

1) vastaja sugu (1 - mees, 0 - naine);

2) vastaja vanus;

3) vastaja hariduse kõrgeim aste (ES-ISCED süsteemi alusel). Andmestikus oli algselt välja toodud 7 eirnevat taset, praktilises ülesandes kodeeriti tunnused ümber järgnevalt:

   0 - kuni põhiharidus (k.a.) (ES-ISCED I, ES-ISCED II);

   1 - keskeriharidus, keskharidus või kutseharidus (ES-ISCED IIIb, ES-ISCED IIIa, ES-ISCED IV);

   2 - kõrgharidus, bakalaureuse kraad või kõrgem (ES-ISCED V1, ES-ISCED V2);

4) leibkonna liikmete arv;

5) leibkonna sissetulek detsiilides, vahemikus 1-10.

2.2 Ülesande püstitus

Oletame, et viiakse läbi uuring ning parasjagu on käimas andmete kogumine valimisse s sattunud objektidelt. Vastamistõenäosused p_k olgu meil teada iga objekti kohta (arvutatud logistilise regressiooni valemi järgi). Vastavalt nendele tõenäosustele toimub objekti valimine vastanute hulka. Seda tehakse järjestusvaliku meetodil. Alguses saavad vastanuteks suurimate vastamistõenäosustega m_1 objekti valimisse sattunuist ($m_1 \ll n$). Need objektid moodustavad vastanute hulga r_1. Lisaks olgu

registrite põhjal varasemalt teada abiinformatsioon iga valimisse sattunu kohta. Kasutades valemit (7) leiame vastanute hulgale r_1 tasakaalumõõdu abitunnuste suhtes, tähistame selle IMB_1. Leiame ka hinnangu uuritava tunnuse keskväärtusele, tähistame selle $\hat{\overline{Y}}_1 \coloneqq \frac{\sum_{r_1} d_k y_k}{\sum_{r_1} d_k}$, kus d_k on disainikaal.

Selleks, et hinnata, kas vastanute hulga tasakaalustamine ka hinnangute nihkeid vähendab, läheme sellest hetkest edasi kahte teed pidi.

Esimesel juhul oletame, et jõupingutusi vastanute hulga tasakaalustamiseks ei tehtud ning m_2 objekti vastasid tavapäraste vastamistõenäosustega p_k, moodustades hulga r_2. Kogu vastanute hulgaks on nüüd $r_{tava} = r_1 \cup r_2$ mahuga $m_1 + m_2$. Leiame uuele vastanute hulgale r_{tava} tasakaalumõõdu, tähistame IMB_{tava}, ning hinnangu uuritava tunnuse keskväärtusele $\hat{\overline{Y}}_{tava} \coloneqq \frac{\sum_{r_{tava}} d_k y_k}{\sum_{r_{tava}} d_k}$.

Teisel juhul oletame, et pärast m_1 vastanu andmete saamist tehti vahelesegamine fikseeritud osakaalu meetodil. Leiame hinnatud vastamistõenäosused $\hat{\theta}_k$ (8). Mittevastanute $s - r_1$ hulgast pandi kõrvale pooled objektid, kel olid suurimad vastamistõenäosuste hinnangud $\hat{\theta}_k$. Järelejäänud objektide hulgast võeti vastanuteks m_2 objekti suurimate vastamistõenäosustega p_k. Koos esialgsete vastanutega saame nüüd vastanute hulga r_{fix}, mahuga $m_1 + m_2$. Leiame saadud vastanute hulgale r_{fix} tasakaalumõõdu, tähistame IMB_{fix} ning hinnangu uuritava tunnuse keskväärtusele $\hat{\overline{Y}}_{fix} \coloneqq \frac{\sum_{r_{fix}} d_k y_k}{\sum_{r_{fix}} d_k}$.

Selleks, et teha otsuseid hinnangute $\hat{\overline{Y}}_1, \hat{\overline{Y}}_{tava}$ ja $\hat{\overline{Y}}_{fix}$ ning nende nihete kohta, teeme antud tsükli läbi 1000 korda. Analüüsime tulemusi ning vaatleme, kas fikseeritud osakaalu meetod on efektiivsem ning annab väiksemad nihked, kui tavapärane vahelesegamiseta vastamine.

2.3 Ülesande käik

2.3.1 Vastamistõenäosuste leidmine

Ülesande lahendamiseks on vaja esmalt üldkogumist moodustada valim, sellest omakorda esialgne vastanute hulk ja seejärel kahe erineva meetodi põhjal lõplikud vastanute hulgad. Olgu genereeritud üldkogumist lihtsa juhusliku valiku abil valim s mahuga $n = 1809$. Olgu vastamismäär 40%, $P_1 = 0.4$. Saame esialgse vastanute hulga r_1, mahuga $m_1 = 724$. Kallutatud vastanute hulga saamiseks, mis

oleks ebaproportsionaalne valimi ja üldkogumi suhtes, peavad vastamistõenäosused sõltuma mõõdetavatest tunnustest. Valime tunnusteks, millest vastamistõenäosus sõltub, soo, vanuse ja leibkonna sissetuleku. Vastamine toimub nii, et mida vanem on valimisse sattunu, seda suurema tõenäosusega võtab ta uuringust osa. Naised vastavad suurema tõenäosusega kui mehed ning kõrgema sissetulekuga inimesed vastavad väiksema tõenäosusega kui madalama sissetulekuga inimesed. Vastamistõenäosused p_k genereerime iga objekti k jaoks logistilise regressiooni mudeli järgi:

$$logit(p_k) = l(p_k) = b_0 + b_1 \cdot sugu + b_2 \cdot ts\_vanus + b_3 \cdot lbk\_sissetulek,$$

kus ts_vanus tähendab keskmistatud vanust (objekti vanus miinus tunnuse $vanus$ keskmine valimis) ning $l(p_k)$ on vastamistõenäosuse log-šansid. Keskmise vanusega objekti puhul $ts\_vanus = 0$. Meie andmetes $sugu(naine) = 0$, siis $b_0 + b_3 \cdot 1$ on keskmise vanusega ja madalaima sissetulekuga naise vastamise log-šansid.

Soovime, et madalaima sissetulekuga naise šansid vastamiseks ja mittevastamiseks oleksid võrdsed, see tähendab $e^{b_0+b_3 \cdot 1} = 1$. Seega nõuame, et $b_0 = -b_3$. Suurus e^{b_1} näitab, mitu korda on meeste šansid vastamiseks suuremad, kui naistel. Meie soovime, et meeste šansid oleksid väiksemad, kuna tavapäraselt kipuvad mehed naistest harvemini küsitlustest osa võtma. Seega soovime, et parameeter $b_1 < 0$. Valime $b_1 = -0.5$, siis $e^{b_1} \approx 0.61$. See tähendab, et mehe šansid küsitlusele vastata on $\frac{1}{0.61} \approx 1.64$ korda väiksemad kui naisel. Vanuse puhul soovime, et vanemate inimeste šansid küsitluses osaleda oleksid suuremad nooremate inimeste omast. Nõudes $b_2 > 0$, valime parameetriks $b_2 = 0.05$. Sel juhul saame, et 10 aastat vanema inimese vastamise šanss suureneb $e^{b_2 * 10} \approx 1.65$ korda. Sissetuleku puhul soovime, et kõrgema sissetulekuga inimese vastamistõenäosus oleks väiksem. Seega soovime, et $b_3 < 0$ ja valime parameetriks $b_3 = -0.05$. Olgu meil kaks samast soost ja ühevanust valimisse sattunut. Oletame, et esimese objekti leibkonna sissetulek on ühe detsiili võrra kõrgem, siis $e^{b_3 * 1} \approx 0.95$ ning tema šansid vastamiseks on $\frac{1}{0.95} \approx 1.05$ korda väiksemad. Lõplik logistilise regressiooni mudel vastamistõenäosuste p_k arvutamiseks on seega järgmine:

$$l(p_k) = 0.05 - 0.5 \cdot sugu + 0.05 \cdot ts\_vanus - 0.05 \cdot lbk\_sissetulek.$$

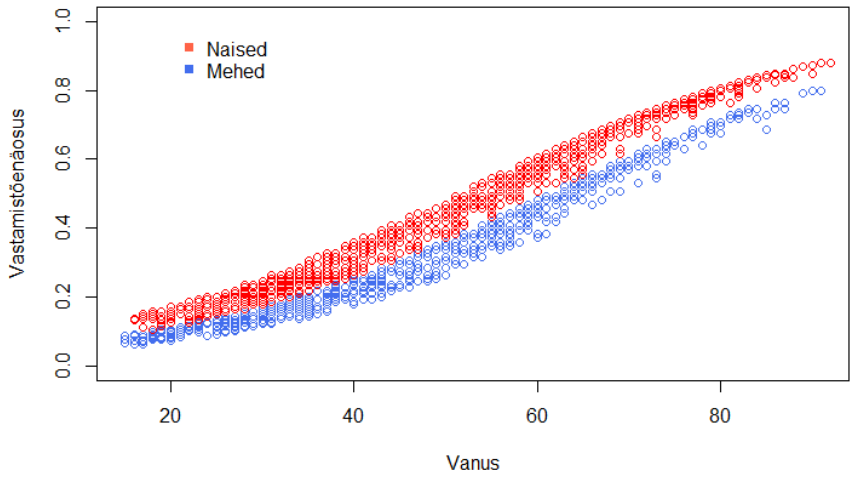
Valemi

$$p_k = \frac{e^{l(p_k)}}{1 + e^{l(p_k)}}$$

järgi saame kätte vastamistõenäosused p_k iga objekti k jaoks. Olgu meil näiteks keskmise vanusega naine, kellel on madalaim sissetulek. Saame võrrandi $l(p_k) = 0.05 - 0.5 \cdot (sugu = 0) + 0.05 \cdot (ts\_vanus = 0) - 0.05 \cdot (lbk\_sissetulek = 1) = 0.05 - 0 + 0 - 0.05 = 0$. Siit edasi $l(p_k) = ln\left(\frac{p_k}{1-p_k}\right)$ ehk $\frac{p_k}{1-p_k} = e^0 = 1$, ja $p_k = 0.5$. Seega selliste tunnustega naise vastamise tõenäosus on 50%. Madalaima sissetulekuga keskmise vanusega mehe vastamise tõenäosus tuleb valemi järgi 38%. Suurima sissetulekuga (väärtusega 10) ja keskmise vanusega naise vastamise tõenäosus on 38%. Keskmise vanuse ja suurima sissetulekuga mehe vastamise tõenäosus on 27%.

Meil on tarvis, et vastamistõenäosused oleksid normeeritud, nii et $\sum_s p_k^* = m$. Nõue tuleneb sellest, et vaatame fikseeritud mahuga m vastanute hulkasid. Siis kehtib $\sum_s I_k = m$, kus I_k on vastamisindikaator. Järelikult kehtib ka $m = E(\sum_s I_k) = \sum_s E(I_k) = \sum_s p_k^*$. Leiame lõpliku vastamise tõenäosuse igale valimi objektile k, kasutades juba leitud väärtust p_k , järgneva eeskirja järgi:

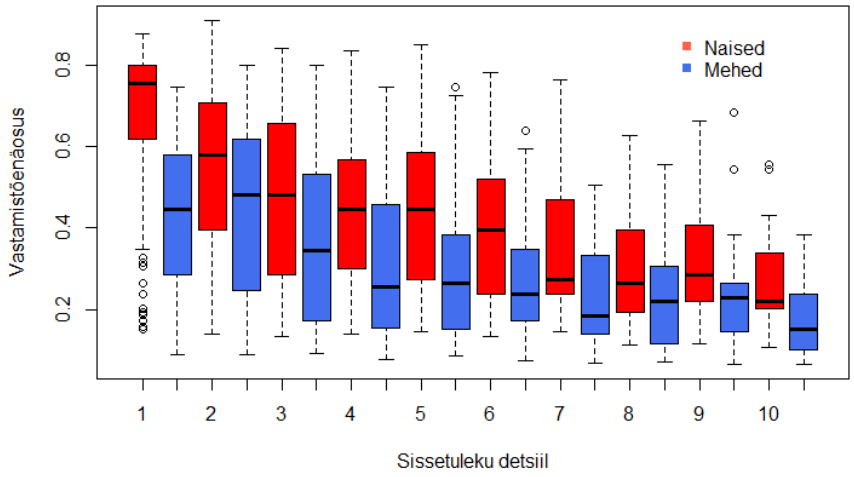
$$p_k^* = \frac{m \cdot p_k}{\sum_s p_k}.$$



Joonis 1: Vastamistõenäosused soo ja vanuse kaupa

Jooniselt 1 on näha, et oleme leidnud lõplikud vastamistõenäosused soost ja vanusest sõltuvalt just eelpool kirjeldatud eeskirja järgi. Vanuse kasvades tõuseb ka

vastamistõenäosuse väärtus ehk vanemaid inimesi kaasatakse vastanute hulka rohkem. Punasega on joonisele märgitud naised, sinisega mehed. Jooniselt on selgelt eristatav, et naiste vastamistäenäosus on suurem, mida me ka soovisime. Seega antud juhul kaasatakse küsitlusse kõige suurema tõenäosusega vanemaid naisterahvaid.



Joonis 2: Vastamistõenäosused soo ja sissetulekute kaupa

Joonisel 2 on karpdiagrammil välja toodud vastamistõenäoste jagunemine olenevalt vastaja soost ning tema leibkonna sissetuleku detsiilist. Väga selgesti on eristatav, et kõrgema sissetulekuga objektide vastamistõenäosus on keskmiselt väiksem kui madalama sissetulekuga objektidel. Samuti näeme jällegi, et naiste vastamistõenäosused on sama suure sissetuleku detsiili korral keskmiselt kõrgemad kui meestel.

2.3.2 Vastanute hulkade genereerimine

Vastanute hulk on juhuslik ja selle tekitame järjestusvaliku alusel [6]. Genereerime iga objekti $k \in s$ jaoks väärtuse ühtlasest jaotusest ja jagame selle vastamistõenäosusega p_k^* :

$$u_k \sim \frac{U(0,1)}{p_k^*}.$$

Mida suurem on vastamistõenäosus p_k^*, seda kitsamas vahemikus $[0, \frac{1}{p_k^*}]$ asub u_k.

Järjestame nüüd valimi tunnuse u_k järgi kasvavalt. Esialgse vastamise korral on meil valitud vastamismääraks $P = 0.4$. Seega vastanute hulga r_1 genereerimiseks võtame valimist s vastanute hulka r_1 esimesed 40% ehk 724 objekti.

Vastanute hulga r_{tava} puhul on vastamismäär $P = 0.6$. Selle vastanute hulga genereerimiseks võtame tunnuse u_k järgi kasvavalt järjestatud valimist s vastanute hulka r_{tava} esimesed 60% ehk 1085 objekti.

Vastanute hulga r_{fix} genereerimiseks kasutame fikseeritud osakaalu meetodit. Selleks arvutame kõigepealt vastamistõenäosuste hinnangud $\hat{\theta}_k$ kõikidele objektidele $k \in s$. Nüüd järjestame esialgselt mittevastanute hulga $s - r_1$ hinnangute $\hat{\theta}_k$ järgi kahanevalt. Esimesed $50\% = 543$ jäävad kõrvale, neid mittevastanuid me tabada ei ürita. Ülejäänud 542 objekti üritatakse uuringule vastamiseks kätte saada. Selleks, et selle ülesande korral vastanute hulkade r_{tava} ja r_{fix} mahud oleksid võrdsed ning neid analüüsivas osas võrrelda saaksime, oletame, et lisandus 20% vastanuid. Järjestame hulga, mis sisaldab objekte, keda vastamiseks kätte saada üritati, jällegi algselt genereeritud vastamistõenäosuste u_k järgi kasvavalt ning valime vastanuteks 20% koguvalimi mahust n ehk 361 esimest objekti. Saame vastanute hulga r_{fix}, mille maht on 1085.

2.4 Tulemused

Olgu antud uuringus uuritavaks tunnuseks leibkonna sissetuleku detsiil. Vastamistõenäosused u_k on genereeritud eespool väljatoodud viisil, arvestades vastaja sugu, vanust ja uuritavaks tunnuseks olevat leibkonna sissetuleku detsiili. Abitunnused tasakaalumõõdu arvutamiseks valime järgneva loogika põhjal. Valida saame neid tunnuseid, mis on saadaval küsitluse väliselt, see tähendab on teada kõigi objektide kohta valimis. Mõistlik on valida selliseid abitunnuseid, mis on korreleeritud uuritava tunnusega. Olukorras, kus vastanute hulka satub pensioniealisi või alaealisi objekte rohkem, kui töötavaid inimesi, võib uuritava tunnuse väärtus vastanute hulgas erineda suuresti valimikeskmisest. Samuti sõltub sissetuleku suurus haridustasemest - mida kõrgem on vastaja haridus, seda kõrgem on eeldatavasti ka tema sissetulek. Uuritava tunnuse väärtus käib vastaja leibkonna kohta, kus võib olla ka mitu sissetulekuga liiget. Seetõttu on vastanute hulga ja koguvalimi vahelise tasakaalumõõdu leidmisel vaja arvestada ka leibkonna suurusega. Kokkuvõttes võtame tasakaalumõõdu IMB arvutamiseks vajalikeks abitunnusteks vastaja soo, vanuse ja tema haridustaseme ning leibkonna suuruse.

Esmalt vaatleme ühekordsel läbitegemisel saadud tulemusi. Tabelis 1 on välja toodud abitunnuste keskmiste vektorid $\overline{\mathbf{x}}_s, \overline{\mathbf{x}}_{r1}, \overline{\mathbf{x}}_{tava}, \overline{\mathbf{x}}_{fix}$ vastavalt koguvalimi, suurimate vastamistõenäosuste abil genereeritud esialgse vastanute hulga r_1, suurendatud vastamismääraga hulga r_{tava} ning fikseeritud osakaalu meetodi abil genereeritud vastanutehulga r_{fix} kohta.

Tabel 1: Abitunnuste keskmised võrdlused koguvalimi ja erinevate vastanute hulkade korral

	vanus	lbk liikmeid	sugu=naine	sugu=mees	haridus=0	haridus=1
Valim s	51.5	2.44	0.610	0.390	0.153	0.558
Hulk r_1	61.1	2.17	0.704	0.296	0.166	0.572
Hulk r_{tava}	60.4	2.19	0.699	0.301	0.167	0.570
Hulk r_{fix}	51.7	2.46	0.637	0.363	0.153	0.537

Tulemuste interpreteerimisel peame meeles, keskmised valimis s on nihketa hinnangud üldkogumi U keskmistele. Suure n korral, nagu see on meil, on valimi hinnangud väga lähedal üldkogumi tegelikule keskväärtusele. Hea hinnang vastanute hulgal on järelikult selline, mis on lähedal valimi s hinnangule.

Vaatleme kõigepealt tunnust *vanus*. Koguvalimis s oli objektide keskmiseks vanuseks 51.5 aastat. Vastanute hulga r_1 korral oli vastajate keskmine vanus 61.1 aastat ja erinevus valimikeskmisest on selle hulga puhul kõige suurem. Vastanute hulga r_{tava} puhul oli keskmine vanus 60.4 aastat. Siinkohal saame tõdeda, et suurema arvu vastanute kuid sama genereerimisviisi korral ei liikunud vaadeldava tunnuse keskmine koguvalimi keskmisele eriti palju lähemale. Fikseeritud osakaalu meetodil leitud hulga r_{fix} keskmine vanus 51.7 aastat erines valimikeskmisest kõige vähem, vaid 0.02 aasta võrra.

Ka leibkonna liikmete arvu korral on fikseeritud osakaalu meetodil leitud valimi keskmine 2.46 inimest koguvalimi keskmisele 2.44 inimest kõige lähemal. Tavapäraselt vastamistõenäosuste järgi genereeritud vastajate hulkades on jällegi näha, et suurema arvu vastajate puhul hulgas r_{tava} on leibkonna keskmine liikmete arv 2.19 peaaegu sama nagu esialgselt vastanute hulga r_1 keskmine 2.17 inimest. Tulemused viitavad fikseeritud osakaalude meetodi kasulikkusele. Samasugune hinnangute muster on ka ülejäänud tunnuste hulgas.

Valimisse sattunud objektide hulgas oli naisi 61% ja mehi 39%. Vastanute hulgas r_1 oli naiste osakaal tunduvalt suurem − 70.4% ning mehi vaid 29.6%. Suurendatud vastamismääraga hulgas r_{tava} on need protsendid jäänud peaaegu samaks, naisi

69.9% ja mehi 30.2%. Hulga r_{fix} korral on sugude osakaal kõige rohkem sarnane valimile. Naiste osakaal on selles hulgas 63.7% ning mehi 36.3%.

Kuni põhiharidusega inimeste osakaal valimis on 15.4%, hulgas r_1 16.6%, hulgas r_{tava} 16.7% ning vastanute hulgas r_{fix} 15.3%. Keskharidus on valimisse sattunuist 55.8%-l, hulgas r_1 57.2% vastanutel, 57%-l hulga r_{tava} objektidel ning 53.7% vastanutel hulgas r_{fix}. Kõrgharidusega objekte on seega valimis s 28.9%, vastanute hulgas r_1 26.2%, vastanute hulgas r_{tava} 26% ning fikseeritud osakaalu meetodil saadud vastanute hulgas 31%. Valimiga kõige rohkem sarnane selle tunnuse korral on jällegi hulk r_{fix}.

Oleme iga abivektorisse valitud tunnuse korral tõdenud, et valimile s kõige sarnasem vastanute hulk on r_{fix}. Tabelis 2 on välja toodud tasakaalumõõdud IMB ning uuritava tunnuse keskväärtuse hinnangud iga vastanute hulga korral.

Tabel 2: Tasakaalumõõt ja uuritava tunnuse hinnang

Vastanute hulk	IMB	$\widehat{\overline{Y}}$
Valim s	0	3.98
Hulk r_1	0.045	3.36
Hulk r_{tava}	0.086	3.40
Hulk r_{fix}	0.002	3.89

Tabelist 2 on näha, et vastanute hulga r_1 tasakaalumõõt $IMB_1 = 0.045$, hulga r_{tava} korral välja arvutatud tasakaalumõõt $IMB_{tava} = 0.086$ ning hulga r_{fix} välja arvutatud tasakaalumõõt $IMB_{fix} = 0.002$. Tuletame meelde, et mida väiksem on tasakaalumõõt IMB, seda rohkem on vastanute hulk tasakaalus algselt võetud valimiga. Siinkohal saame kinnitust, et hulk r_{fix} on algse valimiga enim tasakaalus. Samas näeme ka, et mõnikord võik väiksem valim olla rohkem tasakaalus, kui suurem, seda muidugi kasutatud abitunnuste suhtes.

Uuritava tunnuse ehk leibkonna sissetuleku detsiili keskmine väärtus koguvalimis on 3.89, hulgas r_1 3.36, hulgas r_{tava} 3.40 ning hulgas r_{fix} 3.89. Näeme, et fikseeritud osakaalu meetodil saadud vastanute hulga puhul on uuritava tunnuse keskväärtuse hinnang kõige lähemal valimi keskväärtusele. Märgime ka, et uuritav tunnus ei olnud otseselt sees tasakaalustamise eeskirjas. Tasakaalustamise positiivne mõju uuritavale tunnusele tuleneb korreleeritusest abitunnustega.

Järgnevalt genereerime 1000 korda samast valimist s samade eeskirjade järgi vastanute hulgad r_1, r_{tava} ja r_{fix}. Leiame igal korral genereeritud vastanute hulkade keskväärtuste vektorid ning uurime nende keskväärtusi. Teeme seda selleks, et veenduda, et Tabelist 2 järelduv fikseeritud osakaalu meetodi kasulikkus ei ole mitte ühekordse juhusliku kokkusattumise tulemus.

Tabel 3: Abitunnuste keskmised keskmistatuna üle 1000 korduse

	vanus	lbk liikmeid	sugu=naine	sugu=mees	haridus=0	haridus=1
Valim s	51.5	2.44	0.610	0.390	0.153	0.558
Hulk r_1	60.7	2.14	0.698	0.302	0.172	0.568
Hulk r_{tava}	60.1	2.16	0.691	0.309	0.166	0.571
Hulk r_{fix}	51.6	2.48	0.640	0.360	0.153	0.544

Tabelist 3 näeme, et koguvalimis s on arvud samad nagu Tabelis 1. Ülejäänud arvud on lähedased Tabeli 1 arvudele. Keskmiste muutumise muster on sarnane Tabeli 1 omaga. Seega annab fikseeritud osakaalude meetod vastanute hulga r_{fix}, millelt arvutatud hinnangud on lähedased koguvalimilt arvutatud nihketa hinnangutele.

Tabel 4: Keskmine tasakaalumõõt ja uuritava tunnuse hinnangud üle 1000 korduse

Vastanute hulk	IMB	$\hat{\overline{Y}}$	nihe B	$\hat{\overline{Y}}$ dispersioon	$\hat{\overline{Y}}$ standardhälve
Valim s	0	3.98			
Hulk r_1	0.042	3.35	-0.63	0.0044	0.066
Hulk r_{tava}	0.082	3.39	-0.59	0.0016	0.039
Hulk r_{fix}	0.002	3.90	-0.02	0.0021	0.046

Tabelis 4 toodud arvud on sarnased Tabeli 2 omadele. Ka siinkohal saime tulemuseks, et fikseeritud osakaalu meetodil saadud vastanute hulga puhul on uuritava tunnuse keskväärtuse hinnang keskmiselt parim, see tähendab kõige lähemal valimi keskväärtusele.

Hulgas r_1 on hinnangu $\hat{\overline{Y}}$ nihe B_{r_1} kõige suurem. Nihkest rääkides loeme $\hat{\overline{Y}} = 3.98$ õigeks keskväärtuseks, mistõttu $B_{r_1} = 3.35 - 3.98 = -0.63$. Nihet ei parandatud ka vastanute arvu kasv 20% võrra, $B_{r_{tava}} = 3.39 - 3.98 = -0.59$. Küll vähenes aga nihe praktiliselt olematuks, kui vastanute arv kasvas 20% võrra fikseeritud osakaalude meetodit kasutades. Tabelis 4 on toodud ka hinnangute standardhälbed, mis näitavad hinnangute väikest varieeruvust kordustes.

3 Kokkuvõte

Valikuuringutes on sagedaseks probleemiks mittevastamine. Seega tekib uurijal töötlemiseks ja analüüsimiseks üldkogumist võetud valimist veel väiksema mahuga hulk küsitlusele vastanud objektidest. Lisaks väiksemale mahule, on saadud vastanute hulk ka ebaproportsionaalne valimiga teatud abitunnuste suhtes. On teada, kuidas leida nihketa ja hea täpsusega hinnanguid valimilt. See teooria ei kehti aga kallutatud vastanute hulga korral. Selleks, et vastanute hulgalt uuritavatele tunnustele võimalikult täpseid hinnanguid saada, peame saama vastanute hulga, mis esindaks valimit võimalikult hästi.

Käesoleva bakalaureusetöö eesmärgiks oli uurida, kas vastanute hulga tasakaalustamine abitunnuste suhtes aitab meil uuritava tunnuse nihet vähendada. Tasakaalu teatud abitunnuste suhtes vastanute hulga ning valimi vahel on võimalik mõõta tasakaalumõõduga (7). Inglise keelses kirjanduses kasutatakse siinkohal mõistet *imbalance* ehk tasakaalutus, seega mida madalam on arvutatud tasakaalumõõt, seda enam on vastanute hulk tasakaalus valimiga.

Praktilises osas võtsime vaatluse alla fikseeritud osakaalu meetodi vastanute hulga tasakaalustamiseks. Omistasime objektidele vastamistõenäosused, mis sõltusid ka uuritavast tunnusest. Genereerisime ühe vastanute hulga nende vastamistõenäosuste abil ilma vahelesegamiseta. Teise hulga genereerisime samade tõenäosustega, kuid ühe vahelesegamisega. Vahelesegamispunktis jäeti teatud osa objekte vaatlusest kõrvale (fikseeritud osakaalu meetod). Nii saadi tasakaalustatud vastanute hulk. Mõlemal juhul leidsime nii abitunnustele kui ka uuritavale tunnusele keskväärtuse hinnangud. Veendusime, et tasakaalustatud vastanute hulga korral olid nihked väiksemad kui tavapärase vastamise puhul. Tavapäraselt vahelesegamiseta vastanute hulkade puhul tõdesime, et olenemata vastanute hulga suurenevast mahust, ei paranenud tasakaal abitunnuste suhtes ega ei vähenenud ka hinnangute nihe valimi keskväärtuse suhtes.

Seega võime antud bakalaureusetöös läbi viidud simulatsiooniülesande põhjal väita, et tasakaalustatud vastanute hulga saamiseks tehtud jõupingutused kannavad vilja ning nende abil on võimalik hinnangute nihkeid vähendada.

Viited

[1] Roosileht, N. (2013) *Andmete kogumise juhtimine tasakaaluindikaatori abil*, Tartu Ülikool;

[2] Särndal, C.-E., Lumiste, K., Traat, I. (2016) *Reducing the response imbalance: Is the accuracy of the survey estimates improved?*, Survey Methodology, 42(2): 219-238;

[3] Särndal, C.-E. (2011a) *The 2010 Morris Hansen Lecture. Dealing with Survey Nonresponse in Data Collection, in Estimation*, Journal of Offcial Statistics. 27(1): 1-21;

[4] Särndal, C.-E. (2011b) *Three Factors to Signal Nonresponse Bias - With Applications to Categorical Variables*, Statistics Sweden, Research and Development Department - Methodology Reports from Statistics Sweden. (1): 1-49;

[5] Särndal, C.-E., Lundquist, P. (2014) *Balancing the response and adjusting estimates for nonresponse bias: Complementary activities*, Journal de la Société Frencaise de Statistique, 155(4): 28-50.

[6] Lepik, N., Traat, I. (2017) *Downward calibration property of estimated response propensities*, Acta Et Commentationes Universitatis Tartuensis De Mathematica, ilmumas.

[7] European Social Survey (2014) *http://www.europeansocialsurvey.org/*

Lisa - R-i kood

```
##Loeme andmed sisse
rm(list=ls(all=TRUE))
data=read.csv(file="C:/Users/suvi/Desktop/bakatöö/eestilyhendatud.csv",
              head=TRUE, sep=",",dec=",", na.strings="")
dim(data) # andmemaht, 2051 küsitletut, 253 tunnust

##Valime vajalikud tunnused töötamiseks ja loome tööfaili
tunnused=c("X","HINCTNEE", "HHMMB", "GNDR", "AGEA",  "EISCED")
toofail=data[, tunnused, drop=FALSE]

##Mittevastanute eemaldamine
is.na(toofail$HHMMB[toofail$HHMMB %in% c(77,88,99)])=TRUE
is.na(toofail$GNDR[toofail$GNDR %in% c(9)])=TRUE
is.na(toofail$AGEA[toofail$AGEA %in% c(999)])=TRUE
is.na(toofail$EISCED[toofail$EISCED %in% c(77,88,99)])=TRUE
is.na(toofail$HINCTNEE[toofail$HINCTNEE %in% c(77,88,99)])=TRUE
toofail=na.omit(toofail)
dim(toofail) # alles jäi 1809 objekti, 6 tunnust

## Kodeerin ümber tunnuse haridus: kuni põhi (k.a.), kesk, kõrg
toofail$EISCED[toofail$EISCED %in% c(1,2)]=0
toofail$EISCED[toofail$EISCED %in% c(3,4,5)]=1
toofail$EISCED[toofail$EISCED %in% c(6,7)]=2

##tunnuste karakteristikud analüüsiks
summary(toofail) #saame koguvalimi karakteristikud

## VASTAMISTõENÄOSUSTE ARVUTAMINE
## Oletame, et vastamine sõltub soost, vanusest, sissetulekust:
### naised vastavad suurema tõenäosusega, vanemad inimesed suurema tn-ga,
###  suurema sissetulekuga väiksema tõenäosusega

#esmalt asendame soo puhul naised=0, mitte 2 ja tsentreerime vanuse
toofail$GNDR[toofail$GNDR==2]=0
CENT_AGE=toofail$AGEA-mean(toofail$AGEA)
toofail=cbind(toofail,CENT_AGE)


## Loome vastamistõenosused igale objektile etteantud eeskirja järgi
logit=0.05-0.5*toofail$GNDR+0.05*toofail$CENT_AGE-0.05*toofail$HINCTNEE
p_k=exp(logit)/(1+exp(logit))
sum_pk=sum(p_k)
toofail=cbind(toofail, p_k)
```

```
## Joonised genereeritud p_k kontrolliks
plot(toofail$AGEA[toofail$GNDR==1],p_k[toofail$GNDR==1], col="royalblue2",
     xlab="Vanus", ylab="Vastamistõenäosus", ylim = 0:1)
points(toofail$AGEA[toofail$GNDR==0],p_k[toofail$GNDR==0], col="red")
legend("topleft", legend = c("Naised", "Mehed"), col=c("tomato" , "royalblue2"),
       pch = 15,bty = "n",   horiz = F, inset = c(0.1, 0.05))

##Näeme, et naistel + vanematel kõrgem vastamistõenäosus

boxplot(toofail$p_k ~ toofail$GNDR*toofail$HINCTNEE,
        ylab ="Vastamistõenäosus", xlab ="Sissetuleku detsiil",
        col=c("red", "royalblue2" ),names=c("1","","2","","3","",
                                            "4","","5","","6","","7",
                                            "","8","","9","","10",""))
legend("topright", legend = c("Naised", "Mehed"),
       col=c("tomato" , "royalblue2"),
       pch = 15, bty = "n",   horiz = F, inset = c(0.1, 0.05))


#### Praktiline ülesanne

### ABItunnused tasakaalumõõdu leidmiseks
#binaarsed tunnused soo jaoks
s1=1*(toofail$GNDR==0)
s2=1*(toofail$GNDR==1)

#binaarsed tunnused hariduse jaoks
h1=1*(toofail$EISCED==0)
h2=1*(toofail$EISCED==1)
#h3=1*(toofail$HaridusEISCED==2) - jätame haridus=3 välja,
## et maatriks oleks mittesingulaarne
toofail=cbind(toofail, s1, s2, h1, h2)


##Loome tühjad listid andmete kogumiseks
IMB_1=rep(0,1000)
IMB_tava=rep(0,1000)
IMB_fix=rep(0,1000)
Y_keskalg=rep(0,1000)
Y_kesktava=rep(0,1000)
Y_keskfix=rep(0,1000)
kesk_s_hulk=data.frame("AGEA"=numeric(0), "HHMMB"=numeric(0), "s1"=numeric(0),
                       "s2"=numeric(0), "h1"=numeric(0), "h2"=numeric(0))
kesk_r_hulk=data.frame("AGEA"=numeric(0), "HHMMB"=numeric(0), "s1"=numeric(0),
                       s2=numeric(0), h1=numeric(0), h2=numeric(0))
kesk_r2tava_hulk=data.frame("AGEA"=numeric(0), "HHMMB"=numeric(0),
"s1"=numeric(0), s2=numeric(0), h1=numeric(0),
```

```
h2=numeric(0))
kesk_r2fix_hulk=data.frame("AGEA"=numeric(0), "HHMMB"=numeric(0),
"s1"=numeric(0), s2=numeric(0), h1=numeric(0),
 h2=numeric(0))



### Loome vastanute hulkade genereerimiseks tsükli, kordade arv kokku 1000
kordus=1
while(kordus<=1000){

##Vastanute hulk järjestusvalikuga
set.seed(10*kordus)
n=nrow(toofail)
u_k=runif(n)/toofail$p_k
toofail$u_k=u_k

#sorteerime u_k järgi:
toofail=toofail[order(toofail$u_k),]

x_s=data.frame("AGEA"=toofail$AGEA, "HHMMB"=toofail$HHMMB,
               "s1"=toofail$s1, "s2"=toofail$s2,
               "h1"=toofail$h1, "h2"=toofail$h2)
x_s=as.matrix(x_s)
kesk_s=colMeans(x_s)
kesk_s_hulk[nrow(kesk_s_hulk)+1,] <- kesk_s

#Sigma leidmine
Sigma=t(x_s)%*%(x_s/nrow(x_s))
Sigma_inv=solve(Sigma)

## 1. vastanutehulga r_1 (P=0.4) genereerimine,
###      vastanutehulgale keskmiste vektorid
P_1=0.4 ##esialgne vastamise määr
m_1=round(P_1 * nrow(toofail))
m_1 # saame esialgsesse valimisse 724 objekti
RespInd=rep(0,n)
RespInd[1:m_1]=1
toofail=cbind(toofail,RespInd)
x_r=x_s[1:m_1, ]
head(x_r)
kesk_r=colMeans(x_r)
kesk_r_hulk[nrow(kesk_r_hulk)+1,] <- kesk_r
d=kesk_r-kesk_s


##HINNANGUD
```

```
##Meil tuleb arvutada $\hat{\theta}_k$, tasakaalumõõt IMB_1$,
#### sissetuleku keskmine.

##vastamistõenäosuse hinnang
theta_k=P_1*t(kesk_r)%*%solve(Sigma)%*%t(x_s)
#var(t(theta_k))  - kontrolliks

#IMB
IMB_r1=(P_1**2)*t(d)%*%Sigma_inv%*%d
IMB_1[kordus]=IMB_r1
#Keskmised
Y_kesk=mean(toofail$HINCTNEE[toofail$RespInd==1])
Y_keskalg[kordus]=Y_kesk

####### Esimene variant- leiame r_tava,
###         võtame 20% vastanuid u_k järgi juurde
vastanud_A=toofail
P_2=0.6 ##vastamise määr
m2_tava=round(P_2 * nrow(vastanud_A))
RespInd2=rep(0,n)
RespInd2[1:m2_tava]=1
vastanud_A=cbind(vastanud_A,RespInd2)

#leiame vastanutehulga r_tava keskmiste vektori
x_r2tava=x_s[1:m2_tava, ]
kesk_r2tava=colMeans(x_r2tava)
kesk_r2tava_hulk[nrow(kesk_r2tava_hulk)+1,] <- kesk_r2tava

d_tava=kesk_r2tava-kesk_s

#IMB
IMBtava=(P_2**2)*t(d_tava)%*%Sigma_inv%*%d_tava
IMB_tava[kordus]=IMBtava
#Keskmised
Y_tavakesk=mean(vastanud_A$HINCTNEE[vastanud_A$RespInd2==1])
Y_kesktava[kordus]=Y_tavakesk


### b) variant - järjestame s-r_1 hulga theeta_k hinnangute järgi,
##### jätame 50% suuremaid kõrvale ja võtame
#####  järelejäänutest u_k järgi suurimad 20% vastajateks
vastanud_B=cbind(toofail, "theta_k"=t(theta_k))
P_2=0.6 ##vastamise määr
m2_fix=round(P_2 * nrow(vastanud_B))
vastanud_B=vastanud_B[order(vastanud_B$RespInd, vastanud_B$theta_k,
                            decreasing=TRUE), ]
abi_Ind=rep(1,n)
```

```
abi_Ind[(n-round(m2_fix/2)):n]=0
sum(abi_Ind)##1809-543, kontrolliks
vastanud_B=cbind(vastanud_B, abi_Ind)
vastanud_B=vastanud_B[order(vastanud_B$abi_Ind, vastanud_B$u_k), ]
vastanud_B=cbind(vastanud_B,"RespInd3"=vastanud_B$RespInd)
vastanud_B$RespInd3[1:n*0.2]=1
sum(vastanud_B$RespInd3)

x_r2fix=data.frame("AGEA"=vastanud_B$AGEA[vastanud_B$RespInd3==1],
                   "HHMMB"=vastanud_B$HHMMB[vastanud_B$RespInd3==1],
                   "s1"=vastanud_B$s1[vastanud_B$RespInd3==1],
                   "s2"=vastanud_B$s2[vastanud_B$RespInd3==1],
                   "h1"=vastanud_B$h1[vastanud_B$RespInd3==1],
                   "h2"=vastanud_B$h2[vastanud_B$RespInd3==1])
kesk_r2fix=colMeans(x_r2fix)
kesk_r2fix_hulk[nrow(kesk_r2fix_hulk)+1,] <- kesk_r2fix

d_fix=kesk_r2fix-kesk_s

#IMB
IMBfix=(P_2**2)*t(d_fix)%*%Sigma_inv%*%d_fix
IMB_fix[kordus]=IMBfix

#Keskmised
Y_fixkesk=mean(vastanud_B$HINCTNEE[vastanud_B$RespInd3==1])
Y_keskfix[kordus]=Y_fixkesk

toofail[,c("u_k","RespInd")] <- list(NULL)

toofail=toofail[order(toofail$X),]
kordus=kordus+1
}


##Ühekordsel läbilaskmisel tulemused, set.seed(10)
IMB_r1
IMBtava
IMBfix

Y_kesk
Y_tavakesk
Y_fixkesk
mean(toofail$HINCTNEE)

kesk_s
kesk_r
kesk_r2tava
```

```
kesk_r2fix


#1000-kordne tsükkel ja selle tulemused
mean(IMB_1)
mean(IMB_tava)
mean(IMB_fix)

mean(toofail$HINCTNEE)

mean(Y_keskalg)
var(Y_keskalg)
sd(Y_keskalg)
mean(Y_kesktava)
var(Y_kesktava)
sd(Y_kesktava)
mean(Y_keskfix)
var(Y_keskfix)
sd(Y_keskfix)

colMeans(kesk_s_hulk)
colMeans(kesk_r_hulk)
colMeans(kesk_r2tava_hulk)
colMeans(kesk_r2fix_hulk)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Kätrin Suvi** (sünnikuupäev: 13.02.1994)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Vastanute hulga tasakaalustamine hinnangute täpsustamiseks", mille juhendaja on Imbi Traat,

   1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 09.05.2017