TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Maret Muusikus

Valikuuringutes vastanute hulga kvaliteeti mõõtvad indikaatorid

Bakalaureusetöö (9 EAP)

Juhendaja:

MSc Kaur Lumiste

TARTU

2015

Valikuuringutes vastanute hulga kvaliteeti mõõtvad indikaatorid

Bakalaureusetöö

Kokkuvõte

Valikuuringutes on sageliesinev probleem mittevastamine ja sellest tingitud mõju. Seni on mittevastamisest tingitud nihke hindamiseks kõige sagedamini kasutatud vastamismäära, kuid tänapäeval otsitakse alternatiivseid mooduseid. Käesolevas bakalaureusetöös tutvustatakse kahte indikaatorit, mis on abiks, kui uuringus esineb mittevastamine: (i) R-indikaator, mis põhineb esinduslikkuse mõistel, hindab abiinformatsioonist hinnatud vastamistõenäosuste abil sarnasust vastanute hulga ja valimi vahel; (ii) BI-indikaator põhineb tasakaalu mõistel ehk vastanute hulk ja valim on tasakaalus, kui abitunnuste aritmeetiliste keskmiste erinevus on 0 või väga väike.

Toodud indikaatoreid võrreldakse omavahel simuleerimisülesandes.

Märksõnad: valikuuringud, juhuvalik, R (programmeerimiskeel)

Indicators for measuring response set quality in sample surveys

Bachelor thesis

Summary

Nowadays nonresponse and its impact is very common in surveys and usually practitioners focus on response rate but it is a well-developed finding that response rates alone are not the best for detecting nonresponse bias because they are only indirectly connected. This Bachelor thesis presents two quality indicators which are useful when nonresponse occurs. We look at R-indicator to assess the similarity between the gathered set of respondents and the sample of a survey through the concept of representativeness. Secondly we introduce BI-indicator which is based on the concept of balance. Response set is well-balanced when the difference between sample and response set auxiliary variable means is 0 or very small.

The two indicators are compared in a simulation study.

Key Words: sample surveys, random sampling, R (programming language)

Sisukord

Sissejuhatus

Valikuuringutes on sageliesinev probleem mittevastamine ja sellest tingitud mõju. Enamasti kasutatakse mittevastamisest tingitud mõju hindamiseks vastamismäära. Üldtuntud võteteks, mis parandavad vastamismäära, on näiteks oskuslikemate intervjueerijate kasutamine, mittevastanute hulgast alamvalimi võtmine või tugevamate ajendite leidmine uuringus osalemiseks. Artiklis Heerwegh et al. (2007) jõutakse aga järeldustele, et suurem vastanute hulk ei pruugi tagada paremat andmestikku, millelt hinnatud parameetrite nihked oleks väiksemad. Seega otsitakse alternatiivseid indikaatoreid valimi headuse mõõtmiseks. Uuemate võtetena suunatakse intervjueerijate tähelepanu teatud tunnustega valimi objektide poole, mis pole seni kogutud vastanute hulgas piisavalt hästi esindatud (Särndal, 2011). Käesolevas töös tutvustataksegi R-indikaatorit (R - *representativeness*) ja BI-indikaatorit (BI – *balance indicator*), mis hindavad vastavalt kogutud valimi „esinduslikkust“ ja „tasakaalu“. R-indikaatorit pole eestikeelses kirjanduses varem käsitletud. BI-indikaatorit on eelnevalt käsitletud bakalaureusetöös Mätik (2012), seega tutvustatakse antud töös vaid indikaatori konstrueerimise põhimõtet. Töö eesmärgiks on nende indikaatorite võrdlemine ja rakendamine simuleerimisülesandes.

Bakalaureusetöö esimeses osas antakse ülevaade töös vajaminevatest valikuuringutega seotud mõistetest. Teises peatükis tutvustatakse esinduslikkuse mõistel põhinevat R-indikaatorit ja kolmandas peatükis tasakaalu mõistel põhinevat BI-indikaatorit. Neljandas peatükis võrreldakse kahte indikaatorit ning viiendas peatükis rakendatakse indikaatoreid simuleerimisülesandes, mille koodi loomisel on kasutatud mõningaid võtteid bakalaureusetööst Roosileht (2013).

Töö kirjutamiseks on kasutatud tekstitöötlusprogrammi MS Word. Simulatsiooniülesanne viidi läbi tarkvarapaketiga R.

1. Valikuuringutega seotud mõisted

Valikuuringu eesmärgiks on püstitatud probleemülesandele vastava informatsiooni saamine teatud objektide hulga kohta, mida nimetame üldkogumiks ja tähistame sümboliga U. Uuringu maksumuse ja ajakulu vähendamiseks kogutakse andmed vaid juhuslikult valitud üldkogumi alamhulga kohta, mida nimetame valimiks ja tähistame sümboliga s. Valim määratakse statistilise valikumeetodiga ning selle põhjal tehakse järeldused üldkogumi kohta.

Olgu $U = \{1, 2, \dots, i, \dots, N\}$ üldkogum mahuga N, millest valitakse vastavalt valikudisainile juhuslik valim s mahuga n. Valikudisain fikseerib iga objekti $i \in U$ jaoks kaasamistõenäosuse $\pi_i = P(i \in s)$ ehk tõenäosuse, et objekt i kaasatakse valimisse. Objekti i valimisse kuulumist tähistame indikaatoriga I_i, mis omandab väärtuse 1, kui $i \in s$ ning väärtuse 0, kui $i \notin s$. Uuringu käigus aga paljud uuritavad ei vasta (ei saada kätte või keeldutakse vastamast), valimist saadakse lõpuks kätte vaid alamhulk $r \subset s$ mahuga m ehk vastanute hulk. Sarnaselt valikuindikaatoriga tähistame vastamisindikaatori r_i, mis on väärtusega 1, kui objekt $i \in r$ ehk vastas uuringule, vastasel juhul on r_i väärtus 0. Vastamistõenäosuseks nimetatakse tõenäosust, et objekt i satub valimisse ja vastab ehk $P(r_i = 1, I_i = 1) =: \rho_i$.

Uuritav tunnus Y on uuringu raames huvipakkuv tunnus, mida soovitakse eraldi uurida. Väärtus y_i on teada iga objekti kohta, mis kuulub vastanute hulka ehk $i \in r$, aga mittevastamise tõttu ei ole need teada objektide $i \in s - r$ kohta.

Mittevastamisest põhjustatud kao mõju hindamiseks ja vähendamiseks kasutatakse abitunnuseid. Olgu J kasutavate abitunnuste arv. Eeldame, et meil on J-dimensionaalne abitunnuste veeruvektor $\boldsymbol{x}_i: J \times 1$, milles sisalduv informatsioon on teada nii vastanute kui ka mittevastanute kohta. Abivektori element $\boldsymbol{x}_i = \left(x_{1i}, \dots, x_{ji}, \dots, x_{Ji}\right)'$ on saadaval kõigi valimi objektide jaoks, kus x_{ji} on i-nda objekti j-nda abitunnuse väärtus. Seda informatsiooni saab kasutada nii andmete kogumise kui hindamise etapil. Vektor $\boldsymbol{x}_i$ võib sisaldada tunnuseid nagu näiteks sugu ja vanus ning muid tunnuseid, mis võivad olla olulised uuritava tunnuse analüüsimisel.

2. R-indikaator

Järgnev peatükk on koostatud artikli Schouten et al. (2009) põhjal.

R-indikaatori areng sai alguse olukorrast, kus praktikutele pakkus huvi erineval ajal ja eri teemadel läbi viidud uuringute võrdlemine, selgitamaks välja andmekogumisstrateegiate efektiivsust. Seetõttu ei ole mõistlik defineerida esinduslikku vastanute hulka sõltuvalt uuringu teemast või hinnatud üldkogumi parameetrist. Keskendutakse hoopis andmekogumise kvaliteedile, mis viib meid vastanute hulga struktuuri ja valimi võrdlemiseni. Vastanute hulga struktuuri hindamiseks kasutatakse abitunnuseid, millele on ligipääs väljaspool uuringut. Soovime, et vastanute hulga moodustumine oleks sarnane lihtsale juhuvalikule (LJV) ehk vastamine ei sõltuks objekte eristavatest karakteristikutest.

Lisaks on tegemist kogutud vastanute hulga „headust“ mõõtva indikaatoriga, mis hindab sarnasust vastanute hulga ja valimi või uuritava üldkogumi vahel. Sellele sarnasusele viidatakse kui „esinduslik vastanute hulk“ ehk, kui hästi kogutud objektid esindavad valimit (üldkogumit).

2.1. Esinduslikkuse mõiste

Järgnevalt anname esinduslikkusele esmalt tugeva definitsiooni ning seejärel erijuhuna sellest ka nõrga definitsiooni.

Definitsioon (tugev). *Vastanute hulk on valimi suhtes esinduslik, kui vastamistõenäosused* ρ_i *on konstantsed kõigi üldkogumi objektide jaoks ehk* $\rho_i = \rho,\ \forall i \in U$ *ning kui objekti* i *vastamine ei sõltu teiste objektide vastamisest.*

Vastamistõenäosuste konstantsus tähendab seda, et iga objekti jaoks on vastanute hulka r kuulumise tõenäosus võrdne. See tähendab, et ükski abitunnus ei määra objektide kuulumist vastanute hulka ning vastanute hulga moodustumisel on tegemist LJV-ga. Esinduslikkuse mõiste seotakse LJV-ga, kuna üldiselt annab LJV esindusliku vastanute hulga.

Definitsioon (nõrk). *Olgu meil H kihiga tunnus X. Vastanute hulk esindab tunnust X, kui keskmine vastamistõenäosus alamklassides on konstantne*

$$\bar{\rho}_h = \frac{1}{N_h}\sum_{i\in h}\rho_{hi} = \rho, \qquad h = 1, \dots, H\,, \tag{2.1}$$

kus N_h on klassi h maht, ρ_{hi} on klassi h objekti i vastamistõenäosus ja summeerimine toimub üle kõigi selle klassi objektide.

Nõrga definitsiooni järgi on vastamine tunnuse X suhtes juhuslik. See tähendab, et tunnuse X põhjal pole võimalik vastajaid mittevastajatest eristada.

2.2. R-indikaatori definitsioon

Defineerime esmalt uuringu disainikaaludega kaalutud vastamismäära:

$$P = \frac{\sum_{i\in s} r_i d_i}{\sum_{i\in s} d_i}, \tag{2.2}$$

kus $d_i = 1/\pi_i$ on disainikaal, mis näitab, mitut üldkogumi objekti esindab objekt i.

Eeldame hüpoteetiliselt, et individuaalsed vastamistõenäosused ρ_i on teada. Sellisel juhul on võimalik testida tugevat definitsiooni, mõõtes vastamistõenäosuste ρ_i varieeruvust. Mida väiksem on varieeruvus, seda esinduslikum on vastanute hulk tugeva definitsiooni mõttes. Olgu $\rho = (\rho_1, \rho_2, \dots, \rho_N)'$ vastamistõenäosuste vektor, $\mathbf{1} = (1, 1, \dots, 1)'$ ühtede N-vektor ja $\rho_0 = \mathbf{1} \times \bar{\rho}$ vektor, mille elementideks on üldkogumi keskmine vastamistõenäosus. Artiklis Schouten et al. (2009) on kasutatud hälbe leidmiseks tugevast esinduslikkusest Eukleidilist kaugusfunktsiooni, kuid võib ka teisi kaugusfunktsioone kasutada. Rakendades Eukleidilist kaugust vektorite ρ ja ρ_0 vahel, saame mõõtme, mis on proportsionaalne vastamistõenäosuste standardhälbe valemiga

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{i\in U}(\rho_i - \bar{\rho})^2}\,. \tag{2.3}$$

Saab näidata, et

$$S(\rho) \le \sqrt{\bar{\rho}(1-\bar{\rho})} \le \frac{1}{2}. \tag{2.4}$$

Defineerime R-indikaatori järgmiselt:

$$R(\rho) = 1 - 2S(\rho). \tag{2.5}$$

Eelnevalt välja toodud seose (2.4) põhjal on selge, et R-indikaatori väärtused jäävad intervalli [0,1]. R-indikaatori väärtus 1 tähendab tugevat esinduslikkust, kuna sellisel juhul on vastamistõenäosuste varieeruvus 0. Indikaatori väärtuse 0 korral on vastamistõenäosuste varieeruvus suurim ning hälve tugevast esinduslikkusest maksimaalne.

Praktikas pole vastamistõenäosused ρ_i teada ning meil on informatsioon vaid valimi objektide, mitte terve üldkogumi kohta. Seega peaksime vastamistõenäosused hindama, näiteks logistilise regressiooniga, mistõttu on R-indikaatori hinnangu $\hat{R}(\rho)$ korral tegemist juhusliku suurusega. Vastamistõenäosuse hinnangu tähistame $\hat{\rho}_i$ ning selle leidmiseks kasutatakse abitunnuseid. Hindame vastamistõenäosuste kaalutud keskmist järgmiselt:

$$\hat{\bar{\rho}} = \frac{1}{N}\sum_{i\in U} \hat{\rho}_i I_i d_i \,. \tag{2.6}$$

R-indikaatori hinnangu saame seega kujul

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i\in U} I_i d_i (\hat{\rho}_i - \hat{\bar{\rho}})^2}\,. \tag{2.7}$$

Valemites (2.6) ja (2.7) on summa liikmed läbi kaalutud disainikaaludega d_i ehk kaalutakse valimilt üldkogumi tasemele.

R-indikaator on tugevalt seotud abitunnuste kättesaadavuse ning kasutamisega. Indikaatorit võib andmete kogumise faasil pidevalt arvutada, et hinnata esinduslikkust tol hetkel ning vajadusel suunata andmete kogumist nii, et esinduslikkus paraneks fikseeritud vektori $\boldsymbol{x}_i$ suhtes. Vastamistõenäosusi pole võimalik hinnata ilma abitunnuseid kasutamata, seega nõrga esinduslikkuse kohta saame järeldusi teha vaid kasutatud abitunnuste komplekti raames. Juhul kui indikaatorit kasutatakse uuringute võrdlemiseks, peavad vastamistõenäosuste hindamiseks kasutatud abitunnused uuringutes olema samad.

3. BI-indikaator

Järgnev peatükk on koostatud artikli Särndal (2011) ja bakalaureusetöö Mätik (2012) põhjal.

Mittevastamise mõju vähendamisega on võimalik tegeleda andmete kogumise faasil ja/või hindamise etapil. Hindamise etapil kasutatakse näiteks järelkaalumise või kalibreerimise meetodeid. Andmete kogumise perioodil kasutatakse sekkuvat disaini (*responsive design*), modifitseeritakse algset valikudisaini, eriti andmekogumise hilisematel etappidel. Selle tulemuseks on paremini „tasakaalus“ või „esinduslikum“ lõplik vastanute hulk. Seega esmalt konstrueeritakse valim, kogutakse mingi hulk valimi elemente ja hiljem tasakaalustatakse ehk hangitakse lisaks objekte, mis muudaksid vastanute hulga karakteristikuid lähedasemaks valimi karakteristikutega. Tasakaalustamine hõlmab aga taas abitunnuste kasutamist. Uuemate võtetena suunatakse intervjueerijate tähelepanu teatud tunnustega valimi objektide poole, mis pole piisavalt hästi esindatud seni kogutud vastanute hulgas. Selline lähenemisviis võib viia paremini tasakaalustatud lõpliku vastajate hulgani. Järelikult on vaja indikaatorit, mis andmete kogumise jooksul mõõdaks, kuidas vastanute hulga tasakaalustatus on muutunud.

3.1. Tasakaalus vastanute hulk ja tasakaalutuse indeks

Abivektori element $\boldsymbol{x}_i = \left(x_{1i}, \dots, x_{ji}, \dots, x_{Ji}\right)'$ sisaldab mitmeid abitunnuseid, mille koguarvu tähistus on J. Erijuhul on kaasatud vaid üks kategooriline abitunnus, millel on $J \geq 2$ taset. Siis avaldub abivektor kujul $\boldsymbol{x}_i = \left(\gamma_{1i}, \dots, \gamma_{ji}, \dots, \gamma_{Ji}\right)' = (0, \dots, 1, \dots, 0)'$, kus ainuke 1 asub selle taseme kohal, mida objekt i omab. Praktikas on abitunnuseid tavaliselt mitu. Kasutame abivektoreid $\boldsymbol{x}_i$, mille puhul kehtib võrdus $\boldsymbol{\mu}'\boldsymbol{x}_i = 1,\ \forall i \in U$, kus $\boldsymbol{\mu} \neq \mathbf{0}$ on mingi konstantne vektor. Tegemist ei ole väga range piirava tingimusega. Näiteks, kui $\boldsymbol{x}_i = (0, \dots, 1, \dots, 0)'$, kus ainuke 1 asub selle taseme kohal, mida objekt i omab, siis sobib vektoriks $\boldsymbol{\mu} = (1, \dots, 1, \dots, 1)'$. Kui aga abitunnus x_i on pidev tunnus, siis tuleks võtta $\boldsymbol{x}_i = (1, x_i)'$ ja konstantseks vektoriks sobib $\boldsymbol{\mu} = (1,0)'$.

Defineerime kaks J-dimensionaalset keskmiste vektorit ja kaks $J \times J$ pööratavat kaalumaatriksit, mis on igal andmekogumise hetkel arvutatavad.

$$\bar{\boldsymbol{x}}_r = \frac{\sum_{i \in s} r_i d_i \boldsymbol{x}_i}{\sum_{i \in s} r_i d_i} \tag{3.1}$$

$$\bar{\boldsymbol{x}}_s = \frac{\sum_{i \in s} d_i \boldsymbol{x}_i}{\sum_{i \in s} d_i} \tag{3.2}$$

$$\sum_r = \frac{\sum_{i \in s} r_i d_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sum_{i \in s} r_i d_i} \tag{3.3}$$

$$\sum_s = \frac{\sum_{i \in s} d_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sum_{i \in s} d_i} \tag{3.4}$$

Defineerime tähtsa J-dimensionaalse vektori $\boldsymbol{D} \coloneqq \bar{\boldsymbol{x}}_r - \bar{\boldsymbol{x}}_s = \left(D_1, \dots, D_j, \dots, D_J\right)'$ hindamaks sarnasust ehk tasakaalu vastanute hulga ja valimi vahel. Vektori elementideks on

$$D_j = \bar{x}_{j|r} - \bar{x}_{j|s} = \frac{\sum_{i \in s} r_i d_i x_{ji}}{\sum_{i \in s} r_i d_i} - \frac{\sum_{i \in s} d_i x_{ji}}{\sum_{i \in s} d_i}, j = 1, \dots, J; \tag{3.5}$$

kus $\bar{x}_{j|r}$ on abitunnuse x_j vastanute hulga keskväärtus ning $\bar{x}_{j|s}$ on abitunnuse x_j valimi keskväärtus.

Definitsioon. *Vastanute hulk r on täiuslikus tasakaalus valimiga s (fikseeritud vektori $\boldsymbol{x}_i$ korral), kui keskmised $\bar{\boldsymbol{x}}_r$ ja $\bar{\boldsymbol{x}}_s$ on võrdsed ehk $\boldsymbol{D} = \boldsymbol{0}$.*

Juhul kui $\boldsymbol{D} = \boldsymbol{0}$ on mõõdetud tunnuste väärtused vastanute hulga ja terve valimi korral keskmiselt samad. Tavaliselt ei ole $\boldsymbol{D} \neq \boldsymbol{0}$, mis tähendab, et esineb suuremal või väiksemal määral kõrvalekalle tasakaalust. Teisendades mitmemõõtmelise vektori $\boldsymbol{D}$ ühemõõtmeliseks saaksime mõõta tasakaalu puudumist või tasakaalutust realiseerunud (s, r) jaoks. Sellest lähtuvalt defineerime tasakaalu puudumise ehk tasakaalutuse (*lack of balance*) indeksi

$$\boldsymbol{D}' \sum_s^{-1} \boldsymbol{D} = (\bar{\boldsymbol{x}}_r - \bar{\boldsymbol{x}}_s)' \sum_s^{-1} (\bar{\boldsymbol{x}}_r - \bar{\boldsymbol{x}}_s), \tag{3.6}$$

kus $\sum_s$ on antud valemiga (3.4). Täiusliku tasakaalu korral on ka ruutvorm $\boldsymbol{D}' \sum_s^{-1} \boldsymbol{D} = 0$. Kasvavad erinevused vastanute hulga ja valimi keskmise vahel kasvatavad tavaliselt ka tasakaalutuse indeksit $\boldsymbol{D}' \sum_s^{-1} \boldsymbol{D}$.

3.2. Tasakaaluindikaator

BI-indikaator peaks saavutama oma maksimaalse väärtuse 1, kui $\boldsymbol{D} = \boldsymbol{0}$ (täiuslik tasakaal) ning seda iga vektori $\boldsymbol{x}_i$ ülesehituse korral. Järgnevalt tutvustame ühte Särndali (2011) tuletatud kolmest indikaatorist, mis on $\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D}$ funktsioon ning sisaldab kaalutud vastamismäära (2.2).

$$BI = 1 - 2P(\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D})^{1/2} \tag{3.7}$$

Täiusliku tasakaalu korral saavutab toodud indikaator oma maksimaalse väärtuse 1. Indikaatori ühelähedane väärtus tähendab seda, et $\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D}$ on nullilähedane.

Tasakaaluindikaatori väärtuse tõlgendamine on otseselt seotud välja valitud vektoriga $\boldsymbol{x}_i$, mida kasutatakse $\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D}$ leidmisel. Vektor $\boldsymbol{x}_i$ fikseeritakse kohe alguses ning selle suhtes hinnatakse tasakaalu. Tasakaaluindikaatorit võib andmete kogumise faasil pidevalt arvutada, et hinnata tasakaalu tol hetkel ning vajadusel suunata andmete kogumist nii, et tasakaal paraneks fikseeritud vektori $\boldsymbol{x}_i$ suhtes. Selleks, et võrrelda mõne riigi erinevates uuringutes saavutatud tasakaale või sama uuringu tasakaale erinevate riikide korral, peab abitunnuste vektor $\boldsymbol{x}_i$ olema fikseeritud.

4. R-indikaatori ja BI-indikaatori võrdlus

R-indikaatori tõlgendus põhineb esinduslikkuse mõistel. Kogutud valim on seda esinduslikum, mida väiksem on hinnatud vastamistõenäosuste varieeruvus. BI-indikaator mõõdab aga tasakaalu. Kogutud valim on tasakaalustatud, kui erinevused mõõdetud abitunnuste väärtuste keskmistes on vastanute hulgas ja terves valimis võimalikult väikesed.

Osutub, et BI-indikaator on R-indikaatori erijuht, sest BI-indikaatori saab esitada vastamistõenäosuste hinnangute dispersiooni kaudu, kui vastamistõenäosuste hinnangud on leitud vähimruutude meetodiga (Särndal, 2011). Näitame järgneva seose kehtimist:

$$BI = 1 - 2P(\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D})^{\frac{1}{2}} = 1 - 2\hat{S}_{lin}(\rho) = \hat{R}_{lin}(\rho), \tag{4.1}$$

kus $\hat{S}_{lin}(\rho)$ on lineaarselt hinnatud vastamistõenäosuste ρ_i standardhälve. Seos (4.1) kehtib tingimusel

$$\hat{S}^2_{lin}(\rho) = P^2\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D}\,. \tag{4.2}$$

Kasutame BI-indikaatorile vastamistõenäosuste hindamiseks lineaarset regressiooni ehk $\hat{\rho}_i = \widehat{\boldsymbol{\beta}}'\boldsymbol{x}_i$. Vastamistõenäosuste lineaarsed hinnangud on toodud artiklis Särrndal (2011), eeldusel, et s on fikseeritud. Vastamistõenäosuste hinnangud avaldatakse valemiga

$$\hat{\rho}_i = (\sum_{i\in s} r_i d_i \boldsymbol{x}_i)'(\sum_{i\in s} d_i \boldsymbol{x}_i \boldsymbol{x}_i{}')^{-1}\,\boldsymbol{x}_i\,. \tag{4.3}$$

Vastamistõenäosuste dispersiooni hinnang kogutud valimis avaldub valemiga

$$\hat{S}^2_{lin}(\rho) = \frac{\sum_{i\in s} d_i(\hat{\rho}_i - \hat{\bar{\rho}}_s)^2}{\sum_{i\in s} d_i},$$

kus $\hat{\bar{\rho}}_s = \frac{\sum_{i\in s} d_i\hat{\rho}_i}{\sum_{i\in s} d_i}$ on hinnatud vastamistõenäosuste keskmine valimis. Vastamistõenäosuste keskmine vastanute hulgas avaldub valemiga $\hat{\bar{\rho}}_r = \frac{\sum_{i\in s} r_i d_i\hat{\rho}_i}{\sum_{i\in s} r_i d_i}$. Tingimuse (4.2) tõestamiseks võtame arvesse, et kehtivad alljärgnevates lausetes 4.1–4.3 toodud seosed, mille tõestustega saab huvi korral tutvuda bakalaureusetöös Mätik (2012).

Lause 4.1. *Iga realiseerunud* (s,r) *ja abivektori* $\boldsymbol{x}_i$ *korral, mis rahuldab nõuet* $\boldsymbol{\mu}'\boldsymbol{x}_i = 1$ *kehtivad võrdused*

$$\overline{\boldsymbol{x}}_r{}'\sum_r^{-1}\overline{\boldsymbol{x}}_r = \overline{\boldsymbol{x}}_r{}'\sum_r^{-1}\overline{\boldsymbol{x}}_s = \overline{\boldsymbol{x}}_r{}'\sum_s^{-1}\overline{\boldsymbol{x}}_s = \overline{\boldsymbol{x}}_s{}'\sum_s^{-1}\overline{\boldsymbol{x}}_s = 1\ . \tag{4.4}$$

Lause 4.2. *Suurus* $\hat{\rho}_i$*, mis leitakse valemiga (4.3), rahuldab tingimust*

$$\sum_{i\in s} d_i\hat{\rho}_i^2 = \sum_{i\in s} r_i d_i\hat{\rho}_i\ . \tag{4.5}$$

Lause 4.3. *Suuruste* $\hat{\bar{\rho}}_s$*,* $\hat{\bar{\rho}}_r$*, vastamismäära P ning tasakaalutuse indeksi* $\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D}$ *vahel kehtivad järgmised seosed*

$$\hat{\bar{\rho}}_s = P\ , \tag{4.6}$$

$$\hat{\bar{\rho}}_r = P\overline{\boldsymbol{x}}_r{}'\sum_s^{-1}\overline{\boldsymbol{x}}_r\ . \tag{4.7}$$

Toetudes seostele (4.4) – (4.7), näitame nüüd, et $\hat{S}_{lin}^2(\rho) = P^2\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D}$. Alljärgnev tõestus on samuti olemas lõputöös Mätik (2012), kuid kuna see on kahe indikaatori võrdlemise koha pealt kõige tähtsam tulemus, siis on see eraldi välja toodud.

$$\begin{aligned}\hat{S}_{lin}^2(\rho) &= \frac{\sum_{i\in s} d_i(\hat{\rho}_i - \hat{\bar{\rho}}_s)^2}{\sum_{i\in s} d_i} = \\ &= \frac{\sum_{i\in s} d_i(\hat{\rho}_i^2 - 2\hat{\rho}_i\hat{\bar{\rho}}_s + \hat{\bar{\rho}}_s^2)}{\sum_{i\in s} d_i} = \\ &= \frac{\sum_{i\in s} d_i\hat{\rho}_i^2}{\sum_{i\in s} d_i} - \frac{2\hat{\bar{\rho}}_s\sum_{i\in s} d_i\hat{\rho}_i}{\sum_{i\in s} d_i} + \hat{\bar{\rho}}_s^2\end{aligned}$$

Kasutades seost (4.5) ja vastamistõenäosuste hinnatud keskmise valemit valimis, saame

$$\begin{aligned}\hat{S}_{lin}^2(\rho) &= \frac{\sum_{i\in s} r_i d_i\hat{\rho}_i}{\sum_{i\in s} d_i} - 2\hat{\bar{\rho}}_s^2 + \hat{\bar{\rho}}_s^2 = \\ &= \frac{\sum_{i\in s} r_i d_i\hat{\rho}_i}{\sum_{i\in s} d_i}\frac{\sum_{i\in s} r_i d_i}{\sum_{i\in s} r_i d_i} - \hat{\bar{\rho}}_s^2\end{aligned}$$

Kasutades vastamismäära valemit (2.2) ja seejärel seost (4.6) ning vastamistõenäosuste hinnatud keskmise valemit vastanute hulgas, saame

$$\begin{aligned}\hat{S}_{lin}^2(\rho) &= \frac{\sum_{i\in s} r_i d_i\hat{\rho}_i}{\sum_{i\in s} r_i d_i}P - \hat{\bar{\rho}}_s^2 = \\ &= \hat{\bar{\rho}}_r\hat{\bar{\rho}}_s - \hat{\bar{\rho}}_s^2 = \\ &= \hat{\bar{\rho}}_s(\hat{\bar{\rho}}_r - \hat{\bar{\rho}}_s)\end{aligned}$$

Edasi, kasutades seoseid (4.6), (4.7) ja (4.4), saame lõpuks

$$\hat{S}_{lin}^2(\rho) = P(P\overline{\boldsymbol{x}}_r{}'\sum_s^{-1}\overline{\boldsymbol{x}}_r - P) =$$

$$= P^2(\overline{\boldsymbol{x}}_r{}'\sum_s^{-1}\overline{\boldsymbol{x}}_r - \overline{\boldsymbol{x}}_s{}'\sum_s^{-1}\overline{\boldsymbol{x}}_s) =$$

$$= P^2(\overline{\boldsymbol{x}}_r - \overline{\boldsymbol{x}}_s)'\sum_s^{-1}(\overline{\boldsymbol{x}}_r - \overline{\boldsymbol{x}}_s) =$$

$$= P^2\boldsymbol{D}'\sum_s^{-1}\boldsymbol{D}$$

Oleme saanud, et vastamistõenäosuseid lineaarselt hinnates on BI-indikaator R-indikaatori erijuht.

5. Simuleerimisülesanne

5.1. Üldkogum

Aastatel 2004–2007 läbi viidud Eesti Leibkonna Uuringust kogutud andmete põhjal koostati tehislik üldkogum, mille mahuks tuli kokku 13 914 leibkonda. Iga leibkonna kohta leiduvad andmestikus väärtused järgmistele tunnustele:

- Leibkonnapea majanduslik aktiivsus (binaarne tunnus) – väärtustega 1 – töötab, 0 – ei tööta;
- Leibkonnapea sugu (binaarne tunnus) – väärtustega 1 – mees, 2 – naine;
- Leibkonnapea haridustase (nominaalne tunnus) – väärtustega 1 – algtase, 2 – kesktase, 3 – kõrgtase;
- Netosissetulek ehk netotulu (pidev tunnus) – leibkonna sissetulek koos maksudega uuringukuul (sisaldab sissetulekut palgatööst, tulu põllumajanduslikust tegevusest, omanditulu, siirdetulu ja finantsvahendite müügitulu);
- Siirdetulu (pidev tunnus) – riigilt või omavalitsuselt saadud rahaline toetus (nt pensionid, töötu abiraha, lastetoetus jne);
- Tarbimiskulu (pidev tunnus) – leibkonna tarbimiskulutused uuringukuul;
- Leibkonna suurus (diskreetne tunnus) – inimeste arv leibkonnas;
- Alla 16-aastaste laste arv leibkonnas (diskreetne tunnus).

Simuleerimisülesandes on üldkogumi mahuks 13 914 leibkonda. Olgu uuritavaks tunnuseks leibkonna netotulu. Abitunnusteks, mida kasutatakse R-indikaatori ja BI-indikaatori arvutamisel, valiti leibkonnapea sugu, haridustase, majanduslik aktiivsus ja leibkonna suurus.

5.2. Kallutatud vastanute hulk

Üldkogumist võeti juhusliku valikuga valim mahuga $n = 500$ objekti ja sellest omakorda tekitati kallutatud vastanute hulk mahuga $m = 250$ objekti, seega vastamismäär oli $P = 0.5$.

Vastamistõenäosused tekitati logistilist regressiooni kasutades, mille mudelis valiti vastamist mõjutavateks tunnusteks leibkonnapea sugu ning haridustase. Olgu vastamise eeskiri selline, et meessoost leibkonnapea vastab küsitlusele väiksema tõenäosusega kui naissoost leibkonnapea. Haridustase mõjutagu vastamist selliselt, et mida kõrgem on leibkonnapea haridustase, seda suurem vastamistõenäosus. *Logit* seosefunktsiooni parameetrid valiti järgmiselt:

$$logit(p) = \\ -0.4 \cdot sugu(mees) + 0.3 \cdot haridus(kesk) + 0.45 \cdot haridus(kõrg)\ . \tag{5.1}$$

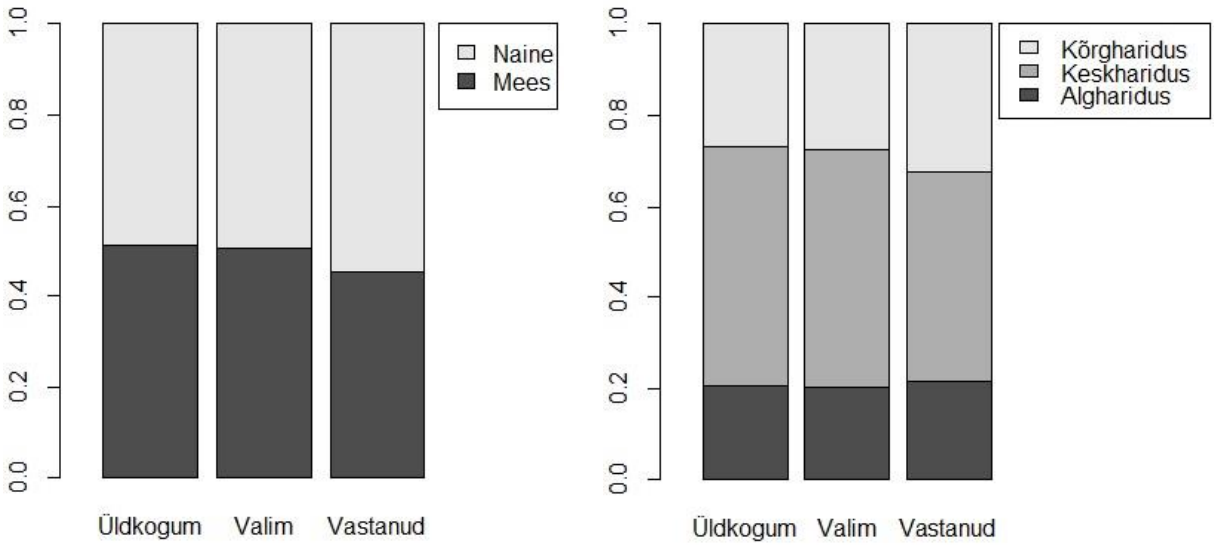
Šansside suhe $e^{-0.4} \approx 0.67$ tähendab, et meessoost leibkonnapea küsitlusele vastamise šansid on $1/e^{-0.4} = 1.49$ korda väiksemad kui naissoost leibkonnapeal. Šansside suhted $e^{0.3} \approx 1.35$ ja $e^{0.45} \approx 1.57$ näitavad, mitu korda on küsitlusele vastamise šansid suuremad vastavalt keskharidusega ning kõrgharidusega leibkonnapea jaoks võrreldes alghariduse omava leibkonnapeaga.

Logit seosest leitakse vastamistõenäosused järgmise valemiga:

$$\rho_i = \frac{e^{logit(p_i)}}{1 + e^{logit(p_i)}}. \tag{5.2}$$

Vastanute hulk r genereeritakse, kasutades tekitatud vastamistõenäosusi. See tähendab, et valimist võetakse juhuslik valik ning kaasamistõenäosuste asemel kasutatakse vastamistõenäosusi. Selliselt on meessoost ja alghariduse leibkonnapeadel väiksem võimalus „valituks“ saada ehk vastata. Tulemusena tekib kallutatud vastanute hulk vastavalt eelpool toodud eeskirjale.

Jooniselt 1 on näha, et üldkogumis on mehi rohkem kui naisi (52%). Juhuslikult valitud valimis langes meeste osakaal 51%-le. Vastamise eeskirja kohaselt oli vastamine kallutatud selliselt, et naised vastavad suurema tõenäosusega, seetõttu on vastanute hulgas meeste osakaal langenud 46%-le. Keskhariduse osakaal on kõigis hulkades ülekaalukalt suurim (ligi pooled leibkonnapeadest). LJV-ga valitud valimis on kõrgharidusega leibkonnapeade osakaal 27% ja alghariduse leibkonnapeade osakaal 21%. Kallutatud vastanute hulgas on kõrghariduse osakaal aga 32%, sest eeskirja järgi vastavad kõrgharidusega leibkonnapead suurema tõenäosusega kui teiste haridusgruppide esindajad. Alghariduse osakaal kasvas 1% võrra, mis vastamismudeli järgi oleks võinud kahaneda, aga oma osa on ka juhuslikkusel.



Joonis 1. Leibkonnapea soo ja haridustaseme jaotused.

Vastanute hulgas r on R-indikaatori ja BI-indikaatori ligikaudsed väärtused vastavalt 0.82294 ja 0.82297, seega mõlemad indikaatorid tagastavad vastanute hulga kvaliteedi hinnanguna sarnase väärtuse.

5.3. Parima uue objekti lisamine vastanute hulka

Vastanute hulga suurendamisel on eesmärgiks lisada selliseid objekte, mis tagavad parima võimaliku esinduslikkuse ja tasakaalu. Esmalt lisatakse üks uus leibkond mittevastanute hulgast. Selleks läbitakse terve mittevastanute hulk ning iga leibkonna korral uuritakse, kuidas ta mõjutab indikaatorite väärtuseid. Kokkuvõtvad tulemused on toodud tabelis 1.

Tabel 1. R-indikaatori ja BI-indikaatori jaotuskarakteristikud.

	Min	Mediaan	Keskmine	Max
R-indikaator	0.8153	0.8250	0.8235	0.8305
BI-indikaator	0.8153	0.8250	0.8235	0.8306

Tulemused kahe indikaatori jaoks on väga sarnased. Parim objekt, mida mittevastanute hulgast lisada on kummagi indikaatori jaoks erinev, millest tulenevalt saavutavad indikaatorid nende objektide lisamisel veidi erinevad maksimaalsed väärtused. Kõige madalama väärtuse tagastab mõlema indikaatori jaoks üks ja sama objekt.

Järgnevalt võeti võrdluseks kolm vastanute hulka r ja

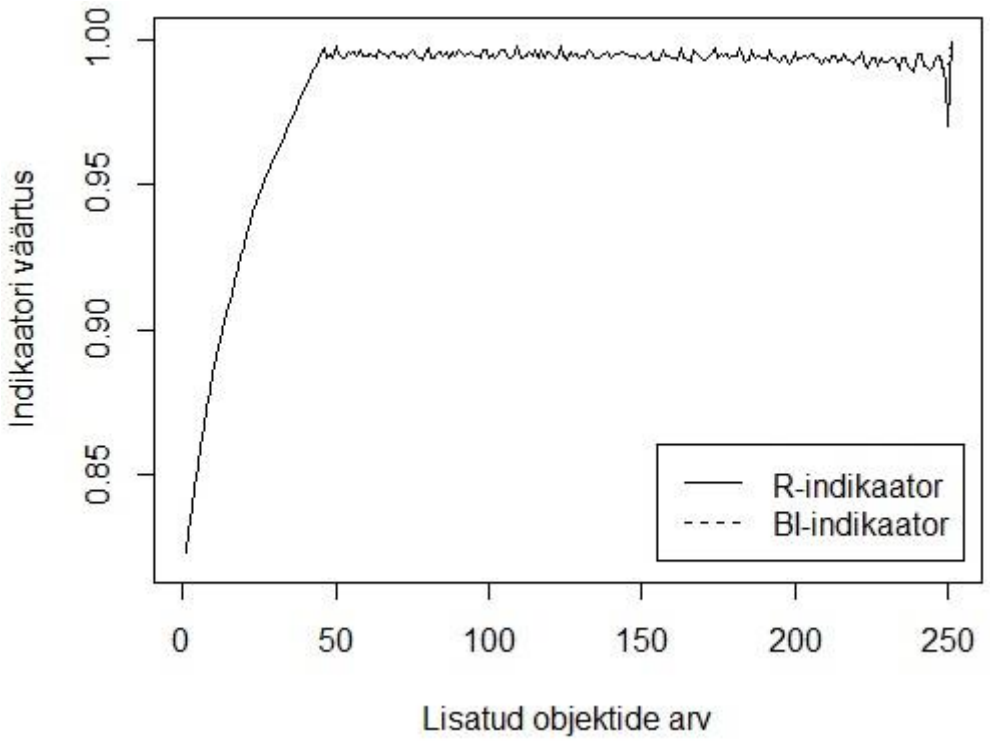
1. esimesse hakati lisama objekte, mis andsid suurima R-indikaatori väärtuse tõusu (tähistus V_R);
2. teise hakati lisama objekte, mis andsid suurima BI-indikaatori väärtuse tõusu (tähistus V_{BI});
3. kolmandasse lisati objekte juhuslikult, et imiteerida olukorda praktikas, kus vastanute hulka tasakaalustavaid indikaatoreid ei kasutataks ja mis oleks võrdluse baasiks (tähistus V_{juh}).

Vastanute hulka tehakse leibkondade lisamisi mittevastanute hulgast 250 korral ehk tsükli lõppedes on kätte saadud kogu valim ning indikaatorid saavutavad oma maksimaalse väärtuse 1.

Iga uue leibkonna lisamisel eelmainitud hulkadesse mõõdeti ka keskmise netopalga erinevust vastanute hulga ja valimi vahel, mis annab aimu nihke suurusest.

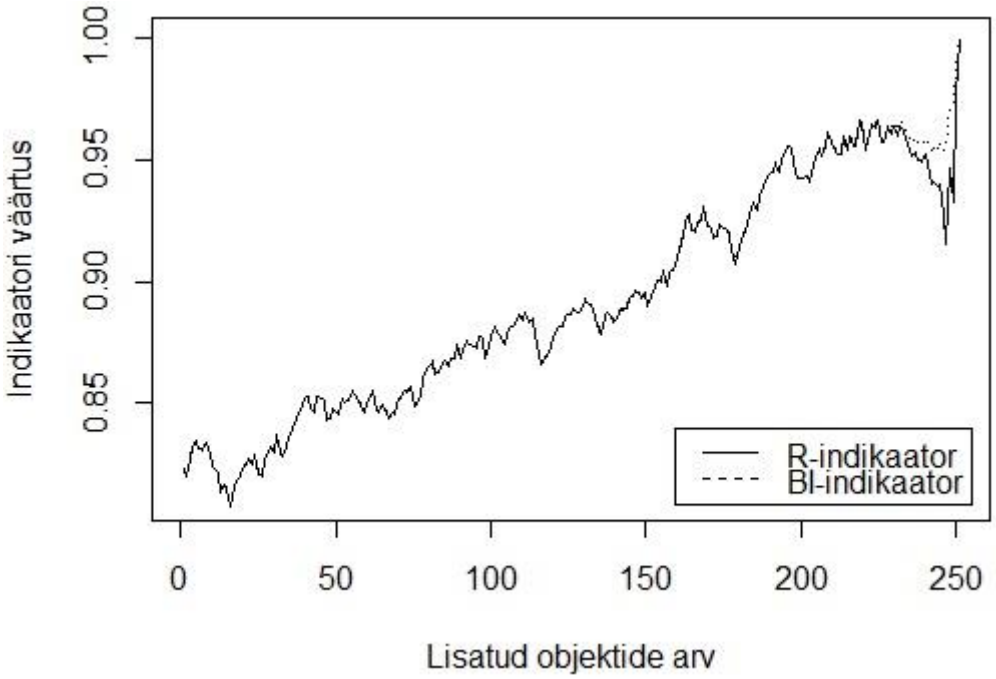
5.4. Tulemuste analüüs pärast kõigi mittevastanute lisamist vastanute hulka

Joonisel 2 on toodud R-indikaatori optimeerimine ja kõrval võrdluseks BI-indikaator hulgas V_R. Joonise alguspunkt on ligikaudu 0.82 (indikaatorite väärtused vastanute hulgas r). Mõlemat indikaatorit iseloomustab kiire väärtuse tõus esimese 40-ne objekti lisamisel ning pärast seda jäävad väärtused vahemikku $(0.99; 1)$ kõikuma. Kõikumine toimub seetõttu, et teatud punktist alates saavutavad indikaatorid päris kõrgeid väärtuseid, tsükkel aga sunnib järgmist objekti võtma, mis võib vastanute hulga esinduslikkuse/tasakaalu kehvemaks muuta. Järgnev objekt võib seda taas parandada, mistõttu tekib kõikumine. Objekte lisatakse, kuni kätte on saadud terve valim ning indikaatorid saavutavad väärtuse 1. Kahte indikaatorit on joonisel raske eristada, kuna väärtused erinevad vaid tuhandeliste poolest. Silma torkab R-indikaatori väärtuse langus 0.97 lähedale viimaste objektide lisamisel, lõpuks saavutatakse siiski väärtus 1. BI-indikaatorit optimeerides oli tulemuseks samasugune joonis, mistõttu pole seda eraldi välja toodud. Lisa 1 joonisel 5 on suurendatud joonist 2 kõikuva piirkonna lähemalt nägemiseks.



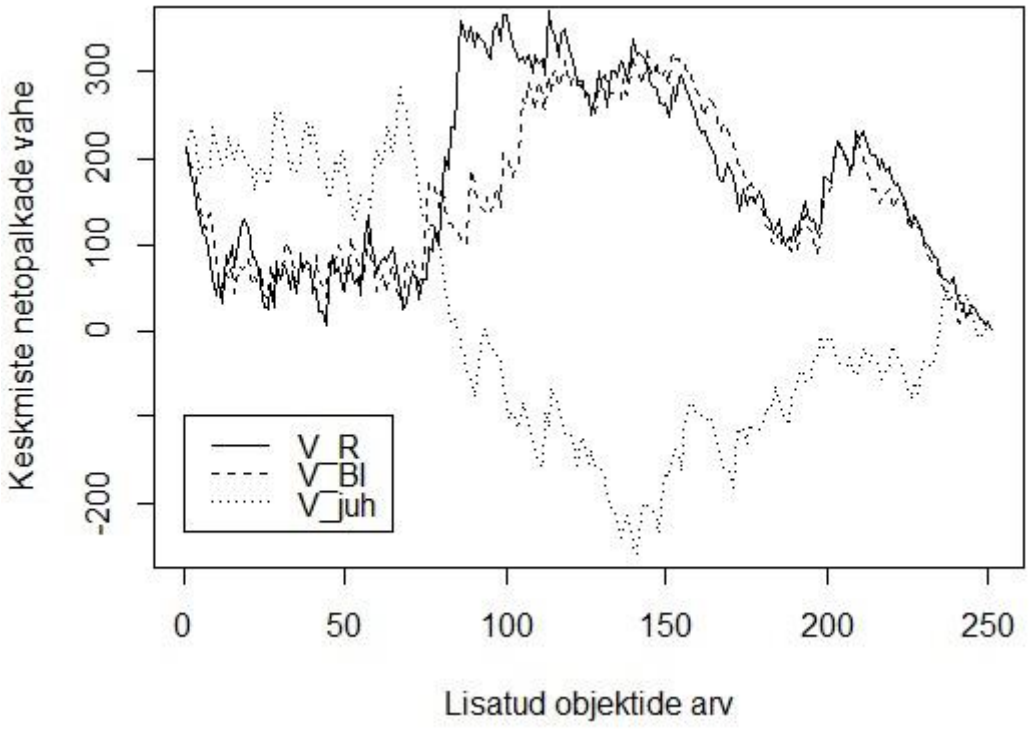
Joonis 2. R-indikaatori optimeerimine ja BI-indikaator hulgas V_R.

Joonisel 3 on näha, kuidas käituvad R-indikaator ja BI-indikaator, kui vastanute hulka objekte juhuslikult lisada. Võrreldes optimeerimisega on joonisel 3 näha, et puudub indikaatorite kiire kasv esimese paarikümne objekti lisamisel. Indikaatorite väärtused kasvavad lisamisprotsessi jooksul. Suurema osa leibkondade lisamisel on indikaatorite erinevused väikesed ning joonisel ei ole võimalik neid eristada. Erinevus paistab silma viimaste objektide lisamisel, kus BI-indikaatori väärtused on suuremad. R-indikaatori väärtuse langus viimaste objektide lisamisel oli näha ka joonisel 2.



Joonis 3. R-indikaator ja BI-indikaator juhusliku lisamisega hulgas V_{juh}.

Joonisel 4 kujutatakse keskmise netopalga erinevust vastanute hulga ja valimi vahel, tehes seda eraldi hulkade V_R, V_{BI} ja V_{juh} jaoks. Arvutades erinevused iga leibkonna lisamise järel, saame jälgida, kuidas muutub vastanute hulkade ja valimi tasakaal uuritava tunnuse netopalga suhtes. Joonise alguspunkt on kõigis hulkades $214\,kr$, tegemist on vastanute hulga r keskmiste erinevusega. On selge, et kui lisada kõik 250 mittevastanut, siis on erinevus 0 ehk tegemist on täiusliku tasakaaluga. Joonisel on näha, et kui indikaatorite väärtuse muutused olid uute objektide lisamisel peaaegu eristamatud, siis keskmiste netopalkade erinevused paistavad rohkem silma. Kõikide hulkade korral on tähelepandav, et parema tasakaalu saavutamiseks piisab alla saja objekti lisamisest. Objektide vahemikus $(90; 170)$ lisamisel nihkub tasakaal kõvasti paigast ära. Tsükli lõppedes hakkab tasakaal taas paranema, sest jõutakse valimi mahule lähemale. Tähelepandav on ka hulkade V_R ja V_{BI} justkui peegeldamine hulga V_{juh} poolt. Pärast 90-nda objekti lisamist on netopalga keskmine hulgas V_{juh} pidevalt väiksem valimi keskmisest netopalgast. Hulkades V_R ja V_{BI} on olukord vastupidine.



Joonis 4. Keskmise netopalga erinevus vastanute hulkade ja valimi vahel.

Uuritavaks tunnuseks oli leibkonna netotulu, seega uuringu eesmärgiks oleks hinnata elanike keskmist netopalka. Olgu uuringu tellijate soov tõsta vastanute hulga maht 300 leibkonnani, see tähendab, et mittevastanute hulgast tuleb lisada 50 objekti. Objekte lisatakse optimeerides R-indikaatorit, BI-indikaatorit ja juhuslikult. Järgnevasse tabelisse on koondatud tunnuse netopalk arvkarakteristikud üldkogumi, valimi, vastanute hulga r ja lõplike vastanute hulkade (V_R, V_{BI} ja V_{juh}) jaoks.

Tabel 2. Leibkonna netopalga arvkarakteristikud üldkogumis, valimis ja simuleeritud vastanute hulkades.

	Maht	Min	Mediaan	Keskmine	Max
Üldkogum	13914	3	7300	9619	137500
Valim	500	124	7401	9601	66690
Vastanute hulk r	250	227	7380	9815	66690
V_R	300	227	7380	9684	66690
V_{BI}	300	227	7448	9669	66690
V_{juh}	300	227	7532	9767	66690

Tabelist 2 on näha, et 50 uue objekti lisamisel vastanute hulka on tulemused kooskõlas joonisega 4. Juhusliku objektide lisamise korral on keskmiste vahe üle $100\ kr$. Indikaatoreid optimeerides on vastanute hulgad valimiga paremas tasakaalus (keskmiste netopalkade vahe on alla $100\ kr$). Siiski on joonise 4 tulemustest näha, et meetodi headus sõltub lisatud objektide arvust.

Kokkuvõte

Käesolevas bakalaureusetöös tutvustati kahte vastanute hulga headust mõõtvat indikaatorit. Mittevastamine on valikuuringutes sageliesinev probleem ning on leitud, et vastamismäär pole parim näitaja mittevastamisest tingitud nihke hindamiseks. Seetõttu otsitakse alternatiivseid mooduseid. R-indikaator, mille väärtus sõltub hinnatud vastamistõenäosuste varieeruvusest, mõõdab vastanute hulga esinduslikkust valimi suhtes. BI-indikaatori (mõõdab vastanute hulga tasakaalu valimi suhtes) ideed on varasemalt tutvustatud näiteks bakalaureusetöös Mätik (2012). Antud töö eesmärgiks oli nende kahe indikaatori omavahel võrdlemine. Indikaatorid on algselt üles ehitatud erinevatel põhimõtetel, kuid selgub, et hinnates BI-indikaatori vastamistõenäosused lineaarse regressiooni mudelist vähimruutude meetodiga, on BI-indikaatori puhul tegemist R-indikaatori erijuhuga.

Antud töös viidi läbi simuleerimisülesanne, kus tekitati kallutatud vastanute hulk. Leibkondade vastamist mõjutavateks tunnusteks valiti leibkonnapea sugu ja haridustase. Indikaatorite arvutamisel võeti abitunnusteks leibkonnapea majanduslik aktiivsus, sugu, haridustase ja leibkonna suurus. Algsetesse vastanute hulkadesse hakati lisama leibkondi kolmel eri viisil. Esimesse hulka lisati mittevastanute hulga objekte, mis optimeerisid R-indikaatori väärtust. Teise hulka objektide lisamisel tehti sama kontroll BI-indikaatoriga. Kolmandasse hulka lisati objekte juhuslikult, et imiteerida praktikas esinevat olukorda, kus indikaatorite väärtusi ei vaadelda. Leibkondade lisamist korrati kuni kätte saadi terve valim ning indikaatorid saavutasid väärtuse 1. Tulemused näitasid, et objektide lisamisel vastanute hulka olid indikaatorite väärtused väga sarnased ning kohe alguses toimus kiire kasv, pärast mida jäid väärtused kõikuma. Juhuslikul lisamisel sama kiiret kasvu ei toimunud. Selgus, et indikaatorite sarnane muutus objektide lisamisel ei taganud vastanute hulkades keskmise netopalga ja valimi keskmise netopalga ühesuguseid erinevusi.

Kasutatud kirjandus

Heerwegh, D., Abts, K., Loosveldt, G. (2007), „Minimizing survey refusal and noncontact rates: do our efforts pay off?“, *Survey Research Methods*, vol. 1, no. 1, pp. 3-10.

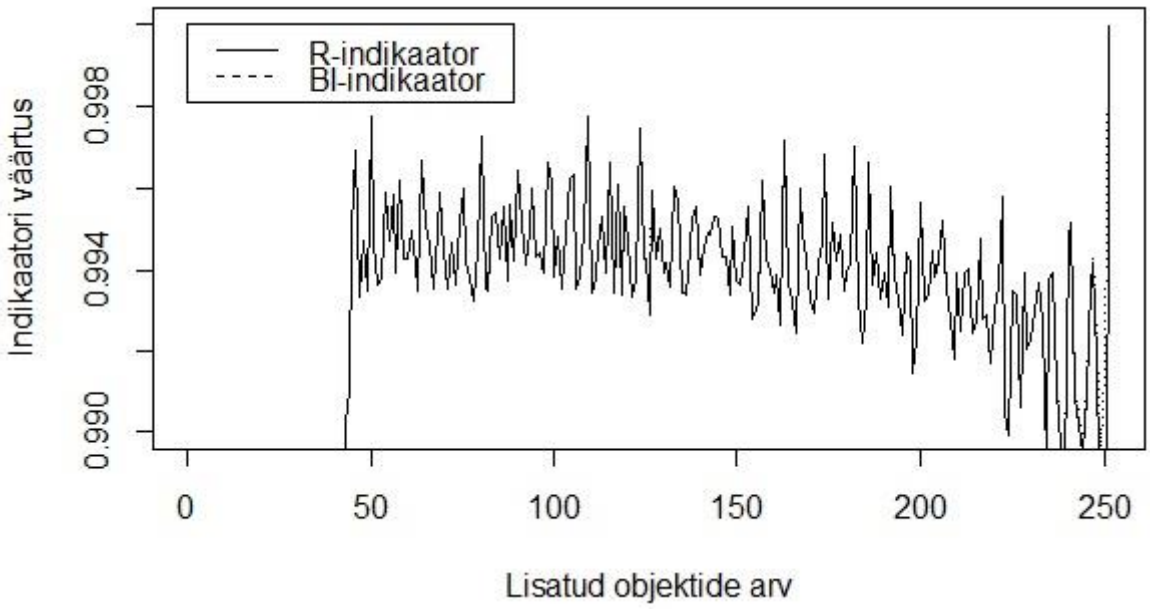
Mätik, M. (2012), „Kao mõju mõõtmine nii valimi võtmise kui ka hindamise etapil“, bakalaureusetöö, Tartu Ülikool.

Roosileht, N. (2013), „Andmete kogumise juhtimine tasakaaluindikaatori abil“, bakalaureusetöö, Tartu Ülikool.

Schouten, B., Cobben, F., Bethlehem, J. (2009), „Indicators for the representativeness of survey response“, *Survey Methodology*, vol. 35, no. 1, pp. 101-113.

Särndal, C.-E. (2011), „The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation“, *Journal of Official Statistics*, vol. 27, no. 1, pp. 1-21.

Lisa 1. Simuleerimisülesande lisajoonis



Joonis 5. R-indikaatori optimeerimine ja BI-indikaator hulgas V_R (lõigus $[0.99; 1]$).

Lisa 2. Tarkvarapaketi R kood

Simuleerimisülesande koodi loomisel on kasutatud valimi ja vastanute hulga tekitamisel mõningaid võtteid bakalaureusetööst Roosileht (2013).

```
# Andmestiku sisselugemine
yk=read.csv(file.choose(),header=T,sep=";")
# Andmestikust objektide eemaldamine, kus netopalk=0
yldkogum=yk[yk$Lbk_netopalk>0,]

# Valimi võtmine juhusliku valikuga
N=nrow(yldkogum) # 13914
n=500
kaasamistn=n/N
x=c(1:N)
indeksid=sample(x,size=n,replace=FALSE)
valim=yldkogum[indeksid,]

# 0/1 tunnused soo jaoks
valim$s1=1*(valim$Lbkp_sugu==1) # mees
valim$s2=1*(valim$Lbkp_sugu==2) # naine
# 0/1 tunnused haridustaseme jaoks
valim$h1=1*(valim$Lbkp_haridus==1)
valim$h2=1*(valim$Lbkp_haridus==2)
valim$h3=1*(valim$Lbkp_haridus==3)

# Kallutatud vastanute hulk, mis sõltub soost ja haridustasemest
# Vastamismääraks 0.5, siis vastanute hulga maht on 250
m=250
# Logistiline mudel
logit=-0.4*valim$s1+0.3*valim$h2+0.45*valim$h3
vast_tn=exp(logit)/(1+exp(logit))
# Lisame vastamistn valimisse
valim=cbind(valim,vast_tn)

# Vastanute hulga tekitamine
freim=c(1:n)
indeks=sample(freim,size=m,replace=FALSE,prob=vast_tn)
valim$vastas=0
valim$vastas[indeks]=1 # vastanutele lisatakse ühed
vastanud=valim[indeks,]
mittevastanud=valim[valim$vastas==0,]
#sum(vastanud$vastas) # 250
#sum(mittevastanud$vastas) # 0
valim=valim[order(valim$vastas,decreasing=TRUE),]

# Hariduse ja soo võrdlus üldkogumis, valimis ja vastanute hulgas
sugu=as.matrix(cbind(table(yldkogum$Lbkp_sugu)/N,
```

```
table(valim$Lbkp_sugu)/n,table(vastanud$Lbkp_sugu)/m))
dimnames(sugu)=list(c("Mees","Naine"),c("Üldkogum","Valim","Vastanud
"))
barplot(sugu,xlim=c(0,6),legend.text=TRUE,args.legend =
list(x=5,y=1))

haridus=as.matrix(cbind(table(yldkogum$Lbkp_haridus)/N,
table(valim$Lbkp_haridus)/n,table(vastanud$Lbkp_haridus)/m))
dimnames(haridus)=list(c("Algharidus","Keskharidus","Kõrgharidus"),
c("Üldkogum","Valim","Vastanud"))
barplot(haridus,xlim=c(0,7.5),legend.text=TRUE,args.legend =
list(x=6.2,y=1))

# BI-indikaatori arvutamine
BI.indikaator=function(x){

x_s=as.matrix(x[,c("Lbkp_majakt","Lbkp_sugu","h1","h2","h3","lbk_suu
rus")])
  x_r=x_s[x$vastas==1,]
  D=colMeans(x_r)-colMeans(x_s)
  sigma=(1/n)*(t(x_s) %*% x_s)
  poord=solve(sigma)
  P=(1/n)*sum(x$vastas)
  BI=1-2*P*sqrt(t(D) %*% poord %*% D)
  return(BI)
}
BIindik=BI.indikaator(valim)

# R-indikaatori arvutamine
valim$Lbkp_haridus=as.factor(valim$Lbkp_haridus)
R.indikaator=function(x){

vastamistn.lg=glm(vastas~Lbkp_majakt+Lbkp_sugu+Lbkp_haridus+lbk_suur
us,
  family=binomial(),data=x)
  prob=predict(vastamistn.lg,type="response") # hinnatud vastamistn
  roo_keskm=(1/n)*sum(prob) # kaalutud keskmine
  R=1-2*sqrt(1/(N-1)*1/kaasamistn*sum((prob-roo_keskm)**2))
  return(R)
}
Rindik=R.indikaator(valim)

# Tsüklid
# Hulkade duplikaadid
vastanud1=vastanud
mittevastanud1=mittevastanud
valim1=valim

# BI-indikaatori arvutamine lisades ühe mittevastanu,
# selliselt läbitakse kogu mittevastanute hulk
tasakaal=function(x){

x_s=as.matrix(x[,c("Lbkp_majakt","Lbkp_sugu","h1","h2","h3","lbk_suu
rus")])
  sigma=(1/n)*(t(x_s) %*% x_s)
  poord=solve(sigma)
```

```
  BIindikaatorid=rep(NA,n)
  for (i in (sum(x$vastas)+1):nrow(x)){ # 251-st 500-ni
    vastajad = x
    vastajad$vastas[i] = 1 # lisab ühe uue vastanu
    x_r=x_s[vastajad$vastas==1,]
    D=colMeans(x_r)-colMeans(x_s)
    P=(1/n)*sum(vastajad$vastas)
    BI=1-2*P*sqrt(t(D) %*% poord %*% D)
    BIindikaatorid[i]=BI
  }
  return(BIindikaatorid)
}

yks=tasakaal(valim1)
summary(yks)

# R-indikaatori arvutamine lisades ühe mittevastanu,
# selliselt läbitakse kogu mittevastanute hulk
esinduslikkus=function(x){
  Rindikaatorid=rep(NA,n)
  for (i in (sum(x$vastas)+1):nrow(x)){
    vastajad = x
    vastajad$vastas[i] = 1 # lisab ühe uue vastanu

vastamistn.lg=glm(vastas~Lbkp_majakt+Lbkp_sugu+Lbkp_haridus+lbk_suur
us,
    family=binomial(),data=vastajad)
    prob=predict(vastamistn.lg,type="response")
    roo_keskm=1/N*1/kaasamistn*sum(prob) # kaalutud keskmine
    R=1-2*sqrt(1/(N-1)*1/kaasamistn*sum((prob-roo_keskm)**2))
    Rindikaatorid[i]=R
  }
  return(Rindikaatorid)
}

kaks=esinduslikkus(valim1)
summary(kaks)
which.max(yks) # indeks 287
which.max(kaks) # 428
which.min(yks) # 407
which.min(kaks) # 407
max(yks[251:500]) # 0.8305654
max(kaks[251:500]) # 0.8305459

# Lisame järjest parimaid objekte mittevastanute hulgast
valim.R=valim1
valim.BI=valim1
valim.juhus=valim1
keskmiste_vahe=c(
mean(valim.R$Lbk_netopalk[valim.R$vastas==1])-
mean(valim1$Lbk_netopalk),
mean(valim.BI$Lbk_netopalk[valim.BI$vastas==1])-
mean(valim1$Lbk_netopalk),
mean(valim.juhus$Lbk_netopalk[valim.juhus$vastas==1])-
mean(valim1$Lbk_netopalk))
indikaatorid.R=c(Rindik,BIindik)
```

```
indikaatorid.BI=c(Rindik,BIindik)
juh.indik=c(Rindik,BIindik)
# keskmiste_vahe, indikaatorid.R, indikaatorid.BI ja juh.indik
sisaldavad
# hetkel väärtusi algse kättesaadud vastanute hulga kohta

# juh.indik-sse kogutakse indikaatorite väärtused,
# kui objektide lisamine toimub juhuslikult

for(i in 1:250){
  Rid=esinduslikkus(valim.R)
  BId=tasakaal(valim.BI)
  parim.R=which.max(Rid) # parima indeks
  R_vaartus=max(na.omit(Rid)) # parima väärtus
  parim.BI=which.max(BId)
  BI_vaartus=max(na.omit(BId))
  valim.R$vastas[parim.R]=1 # parim läheb vastanute hulka
  valim.R=valim.R[order(valim.R$vastas, decreasing=TRUE),]
  valim.BI$vastas[parim.BI]=1
  valim.BI=valim.BI[order(valim.BI$vastas, decreasing=TRUE),]
  # Juhusliku olukorra lisamine
  if (i==250){
    valim.juhuslik.obj=500
  } else{
    juhus=c((250+i):500)
    valim.juhuslik.obj=sample(juhus,1)
  }
  valim.juhus$vastas[valim.juhuslik.obj]=1
  valim.juhus=valim.juhus[order(valim.juhus$vastas,
decreasing=TRUE),]
  keskmiste_vahe=rbind(keskmiste_vahe,
  c(mean(valim.R$Lbk_netopalk[valim.R$vastas==1])-
mean(valim1$Lbk_netopalk),
  mean(valim.BI$Lbk_netopalk[valim.BI$vastas==1])-
mean(valim1$Lbk_netopalk),
  mean(valim.juhus$Lbk_netopalk[valim.juhus$vastas==1])-
mean(valim1$Lbk_netopalk)))

indikaatorid.R=rbind(indikaatorid.R,c(R_vaartus,BI.indikaator(valim.
R)))

indikaatorid.BI=rbind(indikaatorid.BI,c(R.indikaator(valim.BI),BI_va
artus))

juh.indik=rbind(juh.indik,c(R.indikaator(valim.juhus),BI.indikaator(
valim.juhus)))
}

# R indikaatorit optimeerime ja võrdleme BI-indikaatorit (R-
indikaatoriga optimeeritud vastanute hulga pealt)
plot(indikaatorid.R[,1],ylim=c(0.82,1),type="l",
xlab="Lisatud objektide arv",ylab="Indikaatori väärtus")
lines(indikaatorid.R[,2],type="l",lty=3)
legend(155,0.86,c("R-indikaator","BI-indikaator"),lty=c(1,2))
# Joonis suurendatult
plot(indikaatorid.R[,1],ylim=c(0.99,1),type="l",
```

```
xlab="Lisatud objektide arv",ylab="Indikaatori väärtus")
lines(indikaatorid.R[,2],type="l",lty=3)
legend(0,1,c("R-indikaator","BI-indikaator"),lty=c(1,2))

# BI indikaatorit optimeerime ja võrdleme R-indikaatorit (BI-
indikaatoriga optimeeritud vastanute hulga pealt)
plot(indikaatorid.BI[,2],ylim=c(0.82,1),type="l",
xlab="Lisatud objektide arv",ylab="Indikaatori väärtus")
lines(indikaatorid.BI[,1],type="l",lty=3)
legend(155,0.86,c("BI-indikaator","R-indikaator"),lty=c(1,2))
# Joonis suurendatult
plot(indikaatorid.BI[,2],ylim=c(0.99,1),type="l",
xlab="Lisatud objektide arv",ylab="Indikaatori väärtus")
lines(indikaatorid.BI[,1],type="l",lty=3)
legend(0,1,c("BI-indikaator","R-indikaator"),lty=c(1,2))

plot(juh.indik[,1],ylim=c(0.81,1),type="l",
xlab="Lisatud objektide arv",ylab="Indikaatori väärtus")
lines(juh.indik[,2],type="l",lty=3)
legend(160,0.84,c("R-indikaator","BI-indikaator"),lty=c(1,2))

plot(keskmiste_vahe[,1],ylim=c(-250,350),type="l",
xlab="Lisatud objektide arv",ylab="Keskmiste netopalkade vahe")
lines(keskmiste_vahe[,2],type="l",lty=2)
lines(keskmiste_vahe[,3],type="l",lty=3)
legend(0,-100,c("V_R","V_BI","V_juh"),lty=c(1,2,3))

summary(yldkogum$Lbk_netopalk)
summary(valim1$Lbk_netopalk)
summary(vastanud1$Lbk_netopalk)
summary(valim.R$Lbk_netopalk[1:300])
summary(valim.BI$Lbk_netopalk[1:300])
summary(valim.juhus$Lbk_netopalk[1:300])
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Maret Muusikus,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Valikuuringutes vastanute hulga kvaliteeti mõõtvad indikaatorid,

mille juhendaja on Kaur Lumiste,

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, **01.12.2015**