TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Nora Roosileht

Andmete kogumise juhtimine tasakaaluindikaatori abil

Bakalaureusetöö

Juhendaja:

dotsent Imbi Traat

Tartu

2013

Sisukord

Sissejuhatus

Käesolev bakalaureusetöö käsitleb andmete kogumise juhtimist tasakaaluindikaatorite abil valikuuringutes.

Kadu on sage probleem uuringutes. Valimisse sattunud objektidelt ei saada andmeid, sest inimesed jätavad küsimustele vastamata, neid ei saada kätte või ajapuuduse tõttu ei ole võimalik kõiki objekte küsitleda. Tagajärjeks on kallutatud vastanute hulk, mis on paljude näitajate osas ebaproportsionaalne valimi suhtes, ja mida kasutades tekivad nihkega hinnangud.

Särndal (2011a) on välja töötanud indikaatorid, mis võimaldavad vastanute hulga tasakaalu mõõta. Indikaatorid võrdlevad abitunnuste keskmisi vastanute hulgas ja kogu valimis. Kui keskmised on lähedased, on vastanute hulk tasakaalus. Antud töö eesmärk on leida tasakaaluindikaatori abil mittevastanute hulgast üles need objektid, kelle andmed viiksid kogutud tulemused kõige enam tasakaalu valimi suhtes. Küsitlejate jõupingutused tuleb suunata siis just nende objektide kättesaamisele.

Töö esimeses pooles tutvustatakse tasakaaluindikaatoreid, mille abil on võimalik mõõta vastanute hulga tasakaalu kogu valimi suhtes. Varasemas bakalaureusetöös, Mätik (2012), on juba käsitletud tasakaaluindikaatoreid, nende tuletamist ning omadusi, seetõttu tuuakse sellest materjalist välja vaid vajalik. Antud bakalaureusetöös vaadeldakse kõne all olevate indikaatorite käitumist praktikas. Töö teises pooles rakendataksegi tasakaaluindikaatorite leidmist ning vastanute hulga valimi suhtes tasakaalu viimist reaalsete andmete peal. Autori panuseks on tasakaaluindikaatori töö põhimõtte selgitamine, reaalsete andmete saamine ja kasutatavaks teisendamine, ülesande püstitamine, mis imiteerib andmekogumisprotsessi, programmi kirjutamine, mis võimaldab indikaatori katsetamist andmekogumisprotsessis ja selle protsessi suunamist, ning tulemuste interpreteerimine.

Bakalaureusetöö on kirjutatud tekstitöötlusprogrammiga Texmaker. Programm on koostatud ja arvutused teostatud statistikapaketiga R.

Autor tänab turundusuuringute firmat TNS Emor oma andmete töös kasutamise loa eest.

1 Tasakaaluindikaatorid

1.1 Tasakaalutus valikuuringutes

Ükski uuring pole kunagi läbi viidud ideaalsetes tingimustes. Ikka esineb olukordi, kus valimisse sattunud inimesed jätavad mõnedele küsimustele vastamata, ei soovi uuringus osaleda või pole küsitluse käigus kättesaadavad. Näiteks telefoniküsitluse käigus võib vastanute hulka sattuda rohkem pensionieas inimesi kui nooremaid töölkäijaid, sest nemad viibivad rohkem kodus ja saavad elukohta registreeritud telefonile vastata või pole töötavatel inimestel lihtsalt võimalik töö ajal, kui küsitleja helistab, aega uuringus osaleda. Selline olukord tekitab vastanute hulga tasakaalutuse valimi suhtes ning andmetöötluse tulemusena saadakse nihkega hinnangud.

Tähistame üldkogumit sümboliga U. Vastavalt uuringu eesmärgile valitakse valikudisain, mida üldkogumile rakendada. Valikudisainiga on määratud kaasamistõenäosused ehk valimisse sattumise tõenäosused. Objekti k valimisse sattumise tõenäosust tähistame sümboliga π_k. Valimit, mis saadakse valikudisaini rakendamisel, tähistame sümboliga s. Iga valimisse sattunud objekti jaoks on olemas valikukaal, mida tähistame sümboliga d_k, ning mis võrdub kaasamistõenäosuse pöördväärtusega $d_k = \frac{1}{\pi_k}$. Vastanute hulka tähistame sümboliga r ning vastanute arvu sümboliga m. Me eeldame, et $r \subset s \subset U$ ja r ei ole tühi hulk.

1.2 Abivektorid

Selleks, et andmete kogumise protsessi jälgida, on meil vaja kasutada abiinformatsiooni. Enamasti saadakse abiinformatsiooni erinevatest registritest. Leitakse tunnused, mis on teada nii mittevastanute kui ka vastanute kohta.

Abiinformatsioonina kasutatavaid tunnuseid võib olla palju. Moodustub abivektor (Särndal, 2011b) $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$, 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 vastab omaduse puudumisele ja 1 selle olemasolule.

Ka mittearvulist ehk kvalitatiivset tunnust saab kasutada abivektoris, seda binaarsete tunnuste abil. Näiteks J tasemega mittearvulise tunnuse korral ($J \geq 2$) moodustatakse objektile k (0,1) väärtustega J-dimensionaalne vektor $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk})' = (0, \ldots, 1, \ldots, 0)'$, kus $\gamma_{jk} = 1$, kui objektil k esineb omadus j. 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 $J = 1 + \sum_{i=1}^{I}(J_i - 1)$. Maatriksi singulaarsuse vältimiseks eemaldatakse vektorist $\mathbf{x}_k$ iga mittearvulise tunnuse üks väärtustest. Näiteks olgu tunnuse haridus jaoks kolm erinevat väärtust – põhiharidus, keskharidus, kõrgharidus. Saame objekti k jaoks abivektori $\mathbf{x}_k = (\gamma_{1k}, \gamma_{2k})$, kus $\gamma_{1k} = 1$, kui objektil k on põhiharidus, muidu 0, $\gamma_{2k} = 1$, kui objektil k on keskharidus, muidu 0. Kui $\gamma_{1k} = \gamma_{2k} = 0$, siis on objektil k kõrgharidus.

1.3 Tasakaaluindikaatorid

Seda, kui hästi vastanute hulk iseloomustab valimit, on vaja mõõta. Selleks saab kasutatada ühikskaalal määratud tasakaaluindikaatorit.

Me ütleme, et vastanute hulk r on tasakaalus, kui mõõdetavate abitunnuste keskmised on võrdsed vastanute hulgas r ja kogu valimis s. Defineerime J-dimensionaalsed keskmiste vektorid nii vastanute hulgale kui valimile:

$$\bar{\mathbf{x}}_{r;d} = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k}, \tag{1}$$

$$\bar{\mathbf{x}}_{s;d} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k}, \tag{2}$$

kus vektori indeksitest esimene näitab hulka, kuhu kaasatud objektid kuuluvad, teine tähistab objekti kaalu. Tegemist on kaalutud keskmistega. Disainikaalud d_k tagavad, et $\bar{\mathbf{x}}_{s;d}$ on ligikaudu nihketa hinnang üldkogumi keskmiste vektorile. Kui vastanute hulk on tasakaalust väljas, siis $\bar{\mathbf{x}}_{r;d}$ on nihkega.

Keskmiste vahet tähistame sümboliga $\mathbf{D}$ ning

$$\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d}. \tag{3}$$

Tasakaaluindikaatorid kasutavad teatud viisil normeeritud vahet $\mathbf{D}$. Särndal (2011a) pakkus välja kolm indikaatorit:

$$BI_1 = 1 - \frac{\mathbf{D'\Sigma_s^{-1}D}}{Q-1}, \tag{4}$$

$$BI_2 = 1 - 4P^2\mathbf{D'\Sigma_s^{-1}D}, \tag{5}$$

$$BI_3 = 1 - 2P(\mathbf{D'\Sigma_s^{-1}D})^{\frac{1}{2}}. \tag{6}$$

Suurust $\mathbf{D'\Sigma}_s^{-1}\mathbf{D}$ nimetatakse tasakaalutuse indeksiks, kus

$$\mathbf{\Sigma}_s = \frac{\sum_s d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_s d_k}. \tag{7}$$

Kuna tasakaaluindikaator on määratud ühikskaalal (Särndal, 2011), siis

$$0 \leq BI_i \leq 1, i = 1, 2, 3$$

ning lisaks kehtib järjestus

$$0 \leq BI_1 \leq BI_2 \leq 1 \text{ ja } 0 \leq BI_3 \leq BI_2 \leq 1$$

iga realisatsiooni (s, r) ja abivektori $\mathbf{x}_k$ korral.

Suurust P nimetakse uuringu vastamismääraks ja suurus Q on selle pöördväärtus:

$$P = \frac{\sum_r d_k}{\sum_s d_k}, \tag{8}$$

$$Q = \frac{1}{P} = \frac{\sum_s d_k}{\sum_r d_k}. \tag{9}$$

Valemid (1)-(9) lihtsustuvad fikseeritud mahuga isekaaluva disaini korral, mille puhul $\pi_k = \frac{n}{N}$ ja $d_k = \frac{N}{n}$ $\forall k \in s$, kus N on üldkogumi maht ja n valimi maht.

Näeme, et siis

$$\sum_s d_k = \sum_s \frac{N}{n} = n \cdot \frac{N}{n} = N, \tag{10}$$

$$\sum_r d_k = \sum_r \frac{N}{n} = m \cdot \frac{N}{n}, \tag{11}$$

kus m on vastanute hulga r maht.

Valemi (7) jaoks saame alljärgneva lihtsustatud kuju:

$$\mathbf{\Sigma}_s = \frac{\sum_s d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_s d_k} = \frac{1}{n} \sum_s \mathbf{x}_k \mathbf{x}_k'.$$

Samuti lihtsustuvad valemid (1) ja (2) kujule

$$\bar{\mathbf{x}}_{r;d} = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k} = \frac{1}{m} \sum_r \mathbf{x}_k$$

ja

$$\bar{\mathbf{x}}_{s;d} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k} = \frac{1}{n} \sum_s \mathbf{x}_k.$$

Seega isekaaluva disaini korral sisaldab vektor $\mathbf{D}$ abitunnuste tavaliste valimikeskmiste vahesid nii vastanute hulgas kui koguvalimis.

Vastamismäär P ja selle pöördväärtus Q saavad harjumuspärase tähenduse valemeid (10) ja (11) kasutades:

$$P = \frac{m \cdot \frac{N}{n}}{n \cdot \frac{N}{n}} = \frac{m}{n}$$

ning

$$Q = \frac{1}{P} = \frac{n}{m}.$$

2 Andmete kogumise juhtimine

2.1 Ülesande püstitus

Oletame, et toimumas on andmete kogumine valimisse s sattunud objektidelt. Töö lõputähtaeg hakkab saabuma, kuid andmed on saadud vaid m ($m << n$) objektidelt. Need objektid moodustavad vastanute hulga r. Ajapuudus ei võimalda kõiki mittevastanuid enam kätte saada neid korduvalt taga ajades. Seetõttu tuleks jõupingutused suunata just nende objektide kättesaamisele, kes vastanute hulga r tasakaalu kõige enam valimi s suhtes suurendaksid.

Objektide valikul kasutame tasakaaluindikaatoreid BI_1, BI_2 ja BI_3, leidmaks optimaalset objekti vastanute hulgast. Otsime sellist objekti, kelle lisamine suurendab indikaatoreid maksimaalselt. Seejärel lülitame ta vastanute hulka ning asume uue optimaalse objekti otsingule. Kordame protsessi nii kaua, kuni oleme indikaatoritega rahul. Seejuures püüdleme võimalikult suure vastanute arvu poole.

2.2 Andmestiku kirjeldus

Antud bakalaureusetöös on kasutatud sotsio-demograafilisi taustaandmeid, mis on kogutud TNS Emori ühe regulaarse omnibuss uuringu käigus 1999. aastal, mis viidi silmast silma intervjuu vormis. TNS Emor teostab sellist uuringut kaks korda kuus ning kogutakse andmeid vastavalt klientide soovile. Klientide küsitud andmed on andmestikust eemaldatud. Andmestikus on kokku 3524 Eesti elanikku vanuses 16-74, kelle koh-ta mõõdeti 54 tunnust. Objektide arv üldkogumis vähendati 1538 küsitletu peale puuduvate andmete tõttu. Käesolevas töös kasutati järgmiseid tunnuseid:

1) vanus;

2) sugu (1 – mees, 0 – naine);

3) keel (1 – eesti, 0 – vene);

4) haridus, mis on mittearvuline nelja kategooriaga tunnus:

   1 – põhiharidus ja alla selle;

   2 – keskharidus (üldkeskharidus);

   3 – keskeriharidus (kutse- või keskeriharidus pärast põhiharidust, keskeriharidus pärast keskharidus);

   4 – kõrgharidus (diplom, bakalaureus, magister, doktor (PhD), teaduste kandidaat, teaduste doktor);

5) elukoht, mis on mittearvuline nelja kategooriaga tunnus:

   1 – pealinn;

   2 – suur linn (Tartu, Pärnu, Narva, Kohtla-Järve);

   3 – muu linn;

   4 – maa-asula (maakonnakeskus, alevik, küla);

6) sissetulek (küsitlusele vastaja isiklik sissetulek ühes kuus);

2.3 Ülesande teostus

Meie ülesande jaoks on vaja üldkogumist moodustada valim ja sellest omakorda vastanute hulk. Genereerime üldkogumist lihtsa juhusliku valiku abil valimi mahuga $n = 770$. Olgu vastamismäär 50%. Seega saame oma vastanute hulga mahuks $m = 385$. Kui vastamistõenäosused sõltuvad mõõdetavatest tunnustest, siis on vastanute hulk kallutatud ehk ebaproportsionaalne valimi ja üldkogumi suhtes. Tunnusteks, millest küsitlusele vastamine sõltub, valime tunnused vanus ja sugu. Vastamine toimub nii, et mida vanem on inimene, seda suurema tõenäosusega võtab ta uuringust osa ning naised

vastavad parema meelega küsitlusele kui mehed. Vastamistõenäosused p_k genereerime iga objekti k jaoks logistilise regressiooni mudeli järgi:

$$l(p_k) = a + b_1 * sugu + b_2 * tsvanus,$$

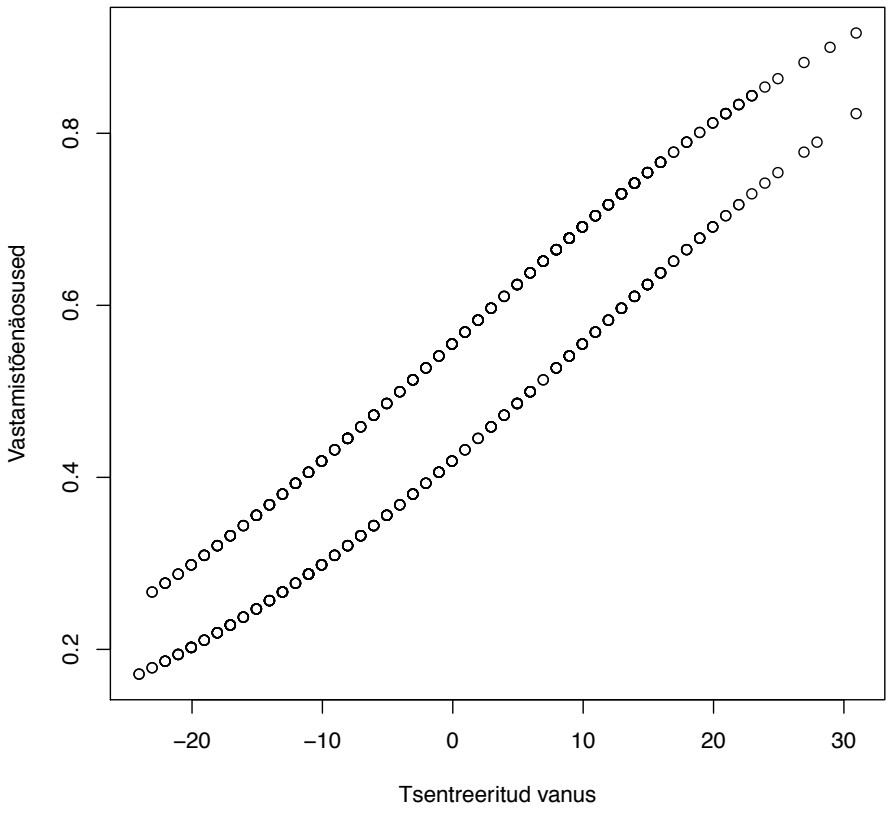
kus $l(p_k)$ on vastamistõenäosuse log-šansid ja $tsvanus$ tähendab tsentreeritud vanust (vanus miinus tema valimikeskmine). Keskmise vanusega objekti korral $tsvanus = 0$. Kuna $sugu(naine) = 0$, siis a on keskmise vanusega naise vastamise log-šansid. Soovime, et naise šansid vastamiseks ja mittevastamiseks oleks võrdsed, st $e^a = 1$. Saame, et $a = 0$. Suurus e^{b_1} näitab, mitu korda on meeste šansid suuremad naiste šanssidest küsimustele vastamiseks. Meie aga soovime, et meeste šansid oleks väiksemad, sest mehed kipuvadki naistest vähem küsitlustest osa võtma. Seetõttu, on vaja, et parameeter $b_1 < 0$. Kui valime näiteks $b_1 = -0.5$, 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 korral on meie jaoks oluline, et vanemate inimeste šansid uuringus osaleda oleksid suuremad nooremate inimeste omast. Seega soovime, et parameeter $b_2 > 0$. Valime näiteks $b_2 = 0.05$. Siit saame, et vanuse kasvades 10 aasta võrra vastamise šanss suureneb $e^{b_2 \cdot 10} \approx 1.65$ korda.

Valemi

$$p_k = \frac{e^{l(p_k)}}{1 + e^{l(p_k)}} \tag{12}$$

järgi saame kätte p_k väärtused iga objekti k jaoks. Kuna vastamistõenäosused peavad olema normeeritud nii, et $\sum_s p_k = m$, siis leiame lõplikud vastamistõenäosused igale valimi objektile, kasutades juba leitud väärtust p_k, järgmise eeskirja järgi:

$$\theta_k = \frac{m \cdot p_k}{\sum_s p_k}. \tag{13}$$



Joonis 1: Vastamistõenäosused valimis

Jooniselt 1 näeme, et oleme genereerinud vastamistõenäosused just eelkirjeldatud eeskirja järgi. Vanuse kasvades kasvab ka vastamistõenäosuse väärtus ehk vastanute hulka kaasatakse rohkem vanemaid inimesi. Kaks paralleelselt asetsevat joont kirjeldavad vastamistõenäosuste erinevust naiste ja meeste vahel. Ülemine joon iseloomustab naiste vastamistõenäosust, sest soovime antud ülesande jaoks suuremat uuringust osavõttu just naiste hulgas. See tähendab, et kõige suurema tõenäosusega kaasatakse küsitlusse vanemaid naisterahvaid.

Vastanute hulk on juhuslik ja selle tekitame järjestusvaliku alusel (Rosén, 1997). Esmalt genereerime iga objekti k jaoks väärtuse ühtlasest jaotusest

ning jagame selle vastamistõenäousega θ_k:

$$u_k \sim \frac{U(0,1)}{\theta_k}. \tag{14}$$

Mida suurem on vastamistõenäosus, seda kitsamal lõigul $[0, \frac{1}{\theta_k}]$ asub u_k. Seejärel järjestame valimi objektid tunnuse u_k järgi kasvavalt ning võtame vastanute hulka neist esimesed 385 objekti.

Lisame tunnuste vanus ja sugu jaotuskarakteristikute tabeli illustreerimaks üldkogumit, valimit ning vastanute hulka.

Tabel 1: Vanuse jaotuskarakteristikud

	Min	Mediaan	Keskmine	Max
Üldkogum	16.00	44.50	43.77	74.00
Valim	19.00	43.00	43.06	74.00
Vastanud	19.00	49.00	47.23	74.00

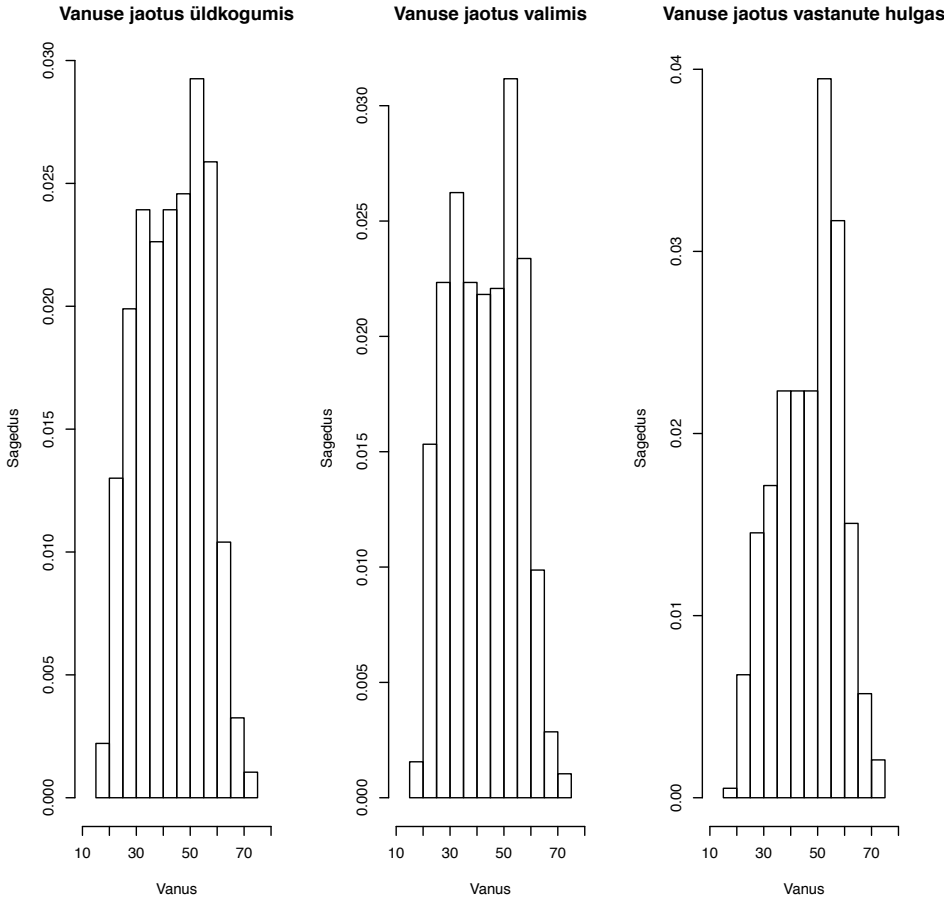
Tabel 2: Soo jaotuskarakteristikud

	Mediaan	Keskmine
Üldkogum	0	0.439
Valim	0	0.453
Vastanud	0	0.371

Kuna valim on võetud lihtsa juhusliku valikuga, siis on see tasakaalus üldkogumi suhtes. Tabelist 1 näeme, et valimi mediaan ja keskmine on ligikaudu võrdsed üldkogumi omadega. Samas näeme Tabelist 1, et vastanute hulgas on keskmine vanus ning mediaan mitme aasta võrra suuremad kui valimis. On näha, et vastanute hulk on vanuse poolest kallutatud valimi suhtes, mis oli ka eespool kirjeldatud protseduuri eesmärk.

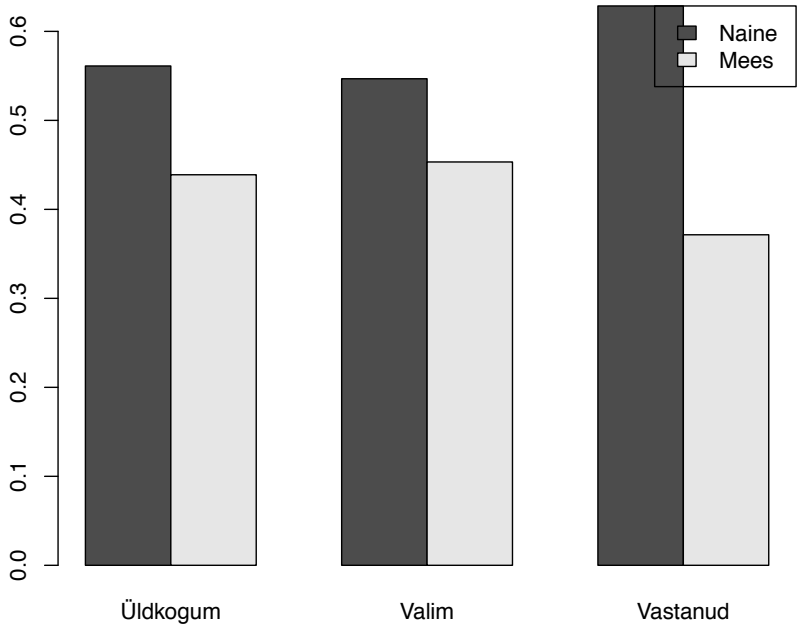
Tabelist 2 näeme, et ka tunnus sugu ei peegelda vastanute hulgas valimit. On näha, et meeste osakaal on vastanute hulgas (37%) väiksem kui valimis (45%).

Tekkinud olukordi kirjeldab ka Joonis 2 ja Joonis 3.



Joonis 2: Vanuse jaotused

Jooniselt 2 on näha, et üldkogumis ja valimis on nii noorema- kui ka vanemapoolsed elanikud jaotunud võrdlemisi ühtlaselt. Vastanute hulka kirjeldavalt graafikult aga nähtub, et vanemate vastajate osakaal on suurem kui nooremate vastajate osakaal. Seega on saadud valimi suhtes ebaproportsionaalne vastanute hulk, mis tekitab hindamisel nihkega hinnangud.



Joonis 3: Soo jaotused

Jooniselt 3 näeme,et meeste osakaal nii üldkogumi kui valimi tasemel on juba naiste osakaalust väiksem. Vastanute hulgas oleme tekitanud eel kirjeldatud mudeli abil veel suurema soolise lõhe, kui see juba oli. Seega on naised antud ülesandes küsitlusele vastanud igatahes suuremal määral kui mehed.

Leiame tasakaaluindikaatorid ilma sugu ja vanust vastanute hulgas arvestamata ning seejärel indikaatorid, kui need samad vastamist mõjutavad tunnused on lisatud abivektorisse. Abivektor **b** sisaldab üht binaarset tunnust soo (mees) ja keele (eesti) kohta, kolme binaarset tunnust hariduse ja elukoha jaoks ning tunnust vanus:

$\mathbf{b}$ = (sugu(mees), keel(eesti), haridus(põhiharidus), haridus(keskharidus), haridus(keskeriharidus), elukoht(pealinn), elukoht(suur linn), elukoht(muu linn), vanus).

Vektorist $\mathbf{a}$ puuduvad eel loetletust vanus ja sugu. Alljärgnevast tabelist näeme, kuidas indikaatorid sellisel juhul teineteisest erinevad.

Tabel 3: Tasakaaluindikaatorid

Abivektor	BI_1	BI_2	BI_3
$\mathbf{a}$	0.995	0.995	0.932
$\mathbf{b}$	0.912	0.912	0.703

Paneme tähele Tabelist 3, et indikaator BI_3 on kasvab aeglasemalt kui BI_1 ja BI_2, mis on põhjustatud ruutjuurest valemis (6). Indikaatorid BI_1 ja BI_2 kasvavad kolmandast indikaatorist kiiremini ning $BI_1 = BI_2$. Kuna valitud ülesande tingimuste kohaselt $P = \frac{m}{n} = \frac{385}{770} = 0.5$ ja $Q = \frac{n}{m} = \frac{770}{385} = 2$, siis BI_1 ja BI_2 võrdsustuvad. Näeme valemitest (4) ja (5), et

$$BI_1 = 1 - \frac{\mathbf{D'\Sigma_s^{-1}D}}{Q-1} = 1 - \mathbf{D'\Sigma_s^{-1}D}$$

ja

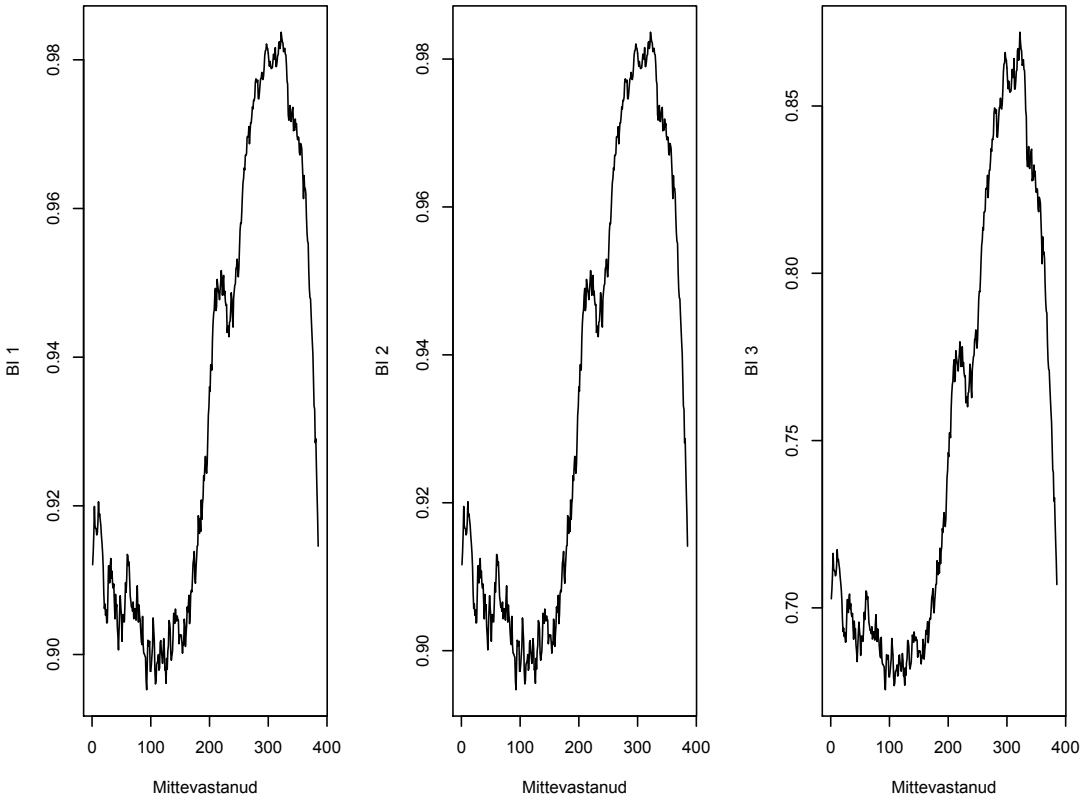
$$BI_2 = 1 - 4P^2\mathbf{D'\Sigma_s^{-1}D} = 1 - \mathbf{D'\Sigma_s^{-1}D}.$$

Oluline tähelepanek on see, et kui abivektor ei sisalda tunnuseid, millest sõltub vastamine (abivektor $\mathbf{a}$), siis tasakaaluindikaatorid kallutatust ei avasta.

Meid huvitab, mis juhtub tasakaaluindikaatoritega, kui kaasata ükshaaval objekte mittevastanute hulgast Vastanute hulka. Seejuures ei jäta me vaadeldavaid objekte vastanute sekka, st vastanute hulga maht on iga kaasamise korral $m+1$. Nii näeme, millised objektid tasakaalu suurendavad või vähendavad.

Jooniselt 4 näeme, et mõni mittevastanute hulgast suvaliselt kaasatud objekt

võib vastanute hulga hoopis rohkem tasakaalust välja viia, kuigi eesmärk on muuta vastanute hulk just sarnasemaks valimile.



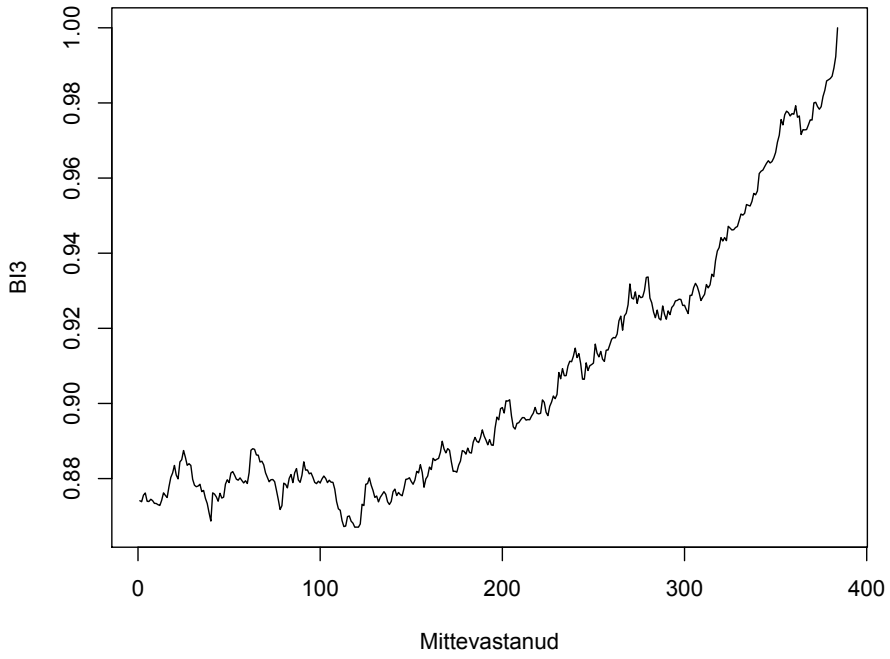
Joonis 4: Tasakaaluindikaatorite käitumine

Jooniselt 4 nähtub ka, et juba ühe objekti lisamisel, saaksime indikaatorite BI_1 ja BI_2 järgi vastanute hulga tasakaalu. Meie ülesandes on kõigi kolme indikaatori puhul maksimiseerivaks objektiks mittevastanute hulgas objekt 322, kes on 30-aastane mees. Vastavateks indikaatorite väärtusteks on $BI_1 = 0.984$, $BI_2 = 0.984$ ja $BI_3 = 0.872$

Kuigi indikaatori väärtus on kõrge, ei ole mõttekas esimese sellise juhtumi leidmisel tööd lõpetada. Teoreetiliselt on olukordi, kus juba ühe objekti lisamisel näitab indikaator tasakaalu. Näiteks, kui valime abivektorisse üksnes

tunnuse vanus, võib indikaatori abil leida sellise mittevastanu, kelle lisamine vastanute hulka võrdsustab keskmised vanused vastanute hulgas ja koguvalimis. Samal ajal m kasvab väga vähe. Oluline on ikkagi ka vastanute hulga maht. Mida lähemale valimi mahule saab vastanute hulga maht, seda täpsemad hinnangud on võimalik leida, seda kõikide uuritavate tunnuste osas. Seega pingutused tuleks suunata vastanute hulga taskaalu saamiseks võimalikult paljude mittevastanute kaasamisel, samal ajal indikaatoreid jälgides.

Järgmiseks vaatlemegi, kuidas muutub indikaator BI_3 objektide järjestikusel lisamisel vastanute hulka. Igal sammul lisatakse objekt, mille korral tasakaaluindikaator saavutab maksimumi antud hetke mittevastanute hulgal. Võtame vaatluse alla just selle indikaatori, sest see läheneb 1-le aeglasemalt kui indikaatorid BI_1 ja BI_2 ning ei omandanud juba esimese objekti lisamisel mittevastanute hulgast maksimaalset väärtust üle 0.9. Igal ringil viskame valitud objekti mittevastanute hulgast välja ehk küsitlejal õnnestus antud isikuga küsitlus läbi viia ning edaspidi kuulub ta vastanute hulka.



Joonis 5: Indikaatori BI_3 käitumine objektide vastanute hulka kaasamise korral

Jooniselt 5 näeme, kuidas muutub indikaator BI_3 vastanute hulga mahu suurenedes indikaatorit maksimaalseks muutvaid objekte juurde lisades. Mahu suurenemisel läheneb indikaator järjest arvule 1, sest vastanute hulk muutub üha rohkem proportsionaalsemaks valimi suhtes. Graafikult nähtuvad aeg-ajalt tekkinud langused indikaatori väärtuse osas. Olukorda selgitab see, et vastanute hulka lisatakse kas vanuse või soo järgi mitu liiga ühesugust objekti, mis tekitavad nö üleküllluse ühes suunas ja esineb taas mingisugune tasakaalutus. Meie ülesande korral on vastanute hulgas esindatud suuremal hulgal vanemapoolsed naised. Eelkirjeldatud meetodi järgi hakatakse lisama sellisel juhul esmajoones nooremaid mehi. Mingil hetkel sellise objekti lisades on neid juba liiga palju, mis viib indikaatori selle objekti kaasamisel madalamaks. See objekt jääb sinna hulka ning hakatakse otsima uute omadustega

tasakaalustavaid objekte, näiteks võivad need sellises olukorras olla vanemapoolsed mehed. Kuna aga vastanute hulga m suurus järjest läheneb valimi s suurusele, on indikaator ikkagi üldjoontes oma väärtuse poolest kasvav.

Samalt jooniselt näeme, kui palju peaksime mittevastanute hulgast objekte kaasama, et indikaatori väärtusega rahul olla. Kui seame eesmärgiks, et vastanute hulk kirjeldaks 90% valimist, siis piisaks 200-st sellises järjekorras vastanute hulka lisatud objekti andmete kättesaamisest. Kui aga tahame, et selle indikaatori väärtus oleks ligikaudu sama, kui siis, kui uuringus osalemine ei sõltu isiku vanusest ja soost (93%, Tabel 3), peaksime kätte saama andmed rohkem kui 300-lt selles järjekorras lisatud objektilt. See tähendaks rohkem kui poole mittevastanute hulga läbi küsitlemist.

Kui aga uuringu tähtaeg hakkab kukkuma ja on aega vaid viimaste objektide leidmiseks, ei pruugi nende inimesteni jõudmine lihtsalt võimalik olla. Alternatiivne võimalus uuringu läbiviijatele oleks keskenduda Joonisel 4 indikaatoreid BI_1, BI_2 ja BI_3 iseloomustavate graafikute tipmisele osale. Seal asuvad objektid annavad ükshaaval vastanute hulka kaasamisel kõige kõrgemad väärtused. Olenevalt küsitlejate arvust tuleb valida paar- kuni mõnikümmend objekti nende seast ning jagada intervjueerjate vahel ära, kelle kõik jõupingutused peavad olema suunatud just nende isikuteni jõudmiseks. Kindlasti vajab see alternatiivne võimalus ja paljud teised võimalikud strateegiad veel põhjalikumat uurimist

2.4 Uuritav tunnus

Olgu uuring seatud üles eesmärgiga hinnata elanike keskmist sissetulekut. Oleme eespool näidanud, kuidas kasutada tasakaaluindikaatorit, ning mis on tema eesmärk. Seejuures pole me pööranud tähelepanu sellele, kuidas muutub samal ajal meie uuritava tunnuse keskmine. Antud andmete puhul on sissetulek esitatud tasemetena, mida on kokku 17. Käsitleme seda tunnust arvulise tunnusena, kus 1 tähendab sissetuleku suurust mingil skaalal.

Olgu meil lõplik vastanute hulk v, mille maht on $q = 600$. Antud hulga moodustame eelmises peatükis kirjeldatud meetodil. Leiame mittevastanute hulgas indikaatorit maksimiseeriva elemendi ning lisame selle vastanute hulka ja eemaldame mittevastanute hulgast. Seejärel kordame protsessi. Jooniselt 5 näeme, et kui lisada veidi üle 200 objekti, on tasakaaluindikaatori väärtus ligikudu 0.9. Võtame sellise meetodiga vastanute hulka 215 objekti ning saame hulga v.

Tabel 4: Sissetulek

Hulk	Miinimum	Mediaan	Keskmine	Maksimum
Valim	1	10	9.514	17
Esialgne vastanute hulk r	1	9	9.205	17
Lõplik vastanute hulk v	1	10	9.468	17

Meie jaoks on oluline küsimus, et kas tasakaalu paranedes läheneb ka sissetuleku väärtus valimi väärtusele. Tabelist 4 näeme, et vastanuta hulgas r on palgataseme mediaan 9 ning keskmine 9.2 väiksemad kui valimis. Kuna vastanute hulgas on rohkem vanemaealisi inimesi kui noori tööl käivaid isikuid, siis on tulemus ootuspärane. Tasakaaluindikaatori BI_3 väärtuse kasvades on ka sissetuleku jaotuskarakteristikud lõpliku vastanute hulga korral sarnasemad valimi karakteristikutele.

Vaatame, millised on abivektorisse valitud tunnuste keskmised samal kolmel hulgal (Tabel 4). Vaatleme abivektorit **b**, kus on ka tunnused, millest sõltub vastamine. Kuigi oleme tunnuste vanus ja sugu karakteristikuid eespool käsitlenud, huvitavad meid ka nende tunnuste keskmised lõplikus vastanute hulgas v.

Tabel 5: Abivektoris **b** esinevate tunnuste keskmised

	s_1	k_1	h_1	h_2	h_3	e_1	e_2	e_3	$Vanus$
Valim	0.453	0.770	0.083	0.209	0.386	0.344	0.165	0.209	43.06
Hulk r	0.371	0.777	0.081	0.192	0.379	0.343	0.168	0.216	47.23
Hulk v	0.447	0.778	0.883	0.202	0.388	0.337	0.162	0.215	44.08

Oleme Tabelis 5 kompaktsuse mõttes abivektoris **b** (lk 16) esinevad tunnused tähistanud vastavalt s_1, k_1, h_1, h_2, h_3, e_1, e_2, e_3 ning $Vanus$. Vastamist mitte mõjutavate tunnuste keskmised on nii valimi, vastanute hulga kui lõpliku vastanute hulga tasemel võrdlemisi ühesugused. See tähendab, et keele, hariduse ja elukoha jagunemise osakaal elanike seas on nendes kolmes hulgas sama ja vastanute hulga kallutamisel ei muutunud. Näeme, et lõplikus vastanute hulgas v meeste osakaal sarnaneb valimi osakaalule ning keskmine vanus on lähedane koguvalimi keskmise vanusega. Mõlema tunnuse keskmised erinesid vastanute hulgas r oluliselt valimi keskmisest. Seega oleme ka nende tunnuste osas saanud vastanute hulga valimi suhtes rohkem tasakaalu indikaatorit BI_3 jälgides.

Antud küsitluse korral oleme saanud sellise optimaalse vastanute hulga, mis peegeldab valimit ning, mille puhul ka uuritav tunnus on karakteristikute poolest sarnane karakteristikutega, mis on leitud valimi pealt.

3 Kokkuvõte

Uuringu läbiviimiseks on lisaks muudele ressurssidele vaja varuda aega. Sageli on just ajapuudus see, mis kõikide valimi objektideni jõudmiseks piirid seab. Sellega seoses tekib meil lisaks valimile veel nii üldkogumist kui valimist väiksema mahuga hulk, küsitlusele vastanute hulk. Kuna uuringu eesmärk on leida nihketa hinnangud üldkogumile, peame veenduma, et saadud hulk oleks tasakaalus valimiga. Siin tulebki appi tasakaaluindikaator.

Käesolevas töös võtsime vaatluse alla kolm tasakaaluindikaatorit (Särndal, 2011a). Leidsime nende väärtused reaalsete adnmete pealt ning jälgisime nende käitumist. Selleks kirja pandud programmikood võimaldab leida järjest isikuid mittevastanute hulgast, kelle lülitamine vastanute hulka parandab antud hulga tasakaalu valimi suhtes optimaalsel viisil. Samuti on töös kasutatud meetodi abil võimalik leida need objektid, kelle kaasamine vastupidi vähendaks indikaatorite väärtust veelgi, kui see juba esialgse vastanute hulga korral oli.

Märgime, et tasakaaluindikaatorite läveväärtusi, millest alates lugeda vastanute hulk tasakaalus olevaks, ei ole fikseeritud. Tähtis on indikaatori suurenemine andmete kogumisel. Võttes läveväärtuseks $BI_3 = 0.9$, saime tasakaalus vastanute hulga mahuga 600, mis teeb vastamismääraks $P = \frac{600}{770} \approx 78\%$. Veendusime ka, et tasakaal abitunnuste osas mõjub hästi ka uuritavatele tunnustele. Meie näites sissetuleku keskmine sai ligikaudu võrdseks valimikeskmisega 9.5. Valimikeskmine on aga lihtsa juhusliku valiku korral nihketa hinnang üldkogumi keskmisele.

Selle bakalaureusetöö põhjal näeme, et tasakaaluindikaatorite arvestamine küsitluse läbiviimisel võib osutada kasulikuks ning abi nihketa hinnangute leidmisel. Kõike juhuse hooleks jättes, ei pruugi hoolimata vastanute hulga suurenevast mahust kaasatud objektid tasakaalu parandada.

Summary

Monitoring Data Collection with Balance Indicator

Bachelor thesis

Nowadays, nonresponse is a very common problem in survey sampling. There are always objects, who do not give out the information asked, are unreachable during the survey, or interviewers do not get a chance to get in touch with some objects in time. This causes responding sample being disproportionate with respect to the full sample and leads to biased estimates.

The purpose of this Bachelor thesis was to present briefly tools that measure balance of the response set and then to focus more thoroughly on putting these tools into practice using real data.

In the first part of the thesis the instruments that measure respondents balance to sample were introduced. In the second part of the thesis these instruments, balance indicators, were applied and tested in a simulated data collection process on real data. Data were collected from these nonrespondents who maximally increased the balance indicators. At certain level, e.g. in our case $BI_3 = 0.9$, we declared the response set balanced. For this level the response rate was approximately 78%, and the auxiliary variable means were close to the respective sample means. Furthermore, also study variable mean beacame close to its sample mean. But sample mean is unbiased estimate for the population mean under our design.

We studied that it is useful to monitor balance indicator in a data collection process. A practical experiment carried out on real data confirmed that balance indicators really can show imbalance under dependent nonresponse.

Kirjanduse loetelu

1. Mätik, M, 2012. Kao mõju mõõtmine nii valimi võtmise kui ka hindamise etapil. Bakalaureusetöö. Tartu Ülikool;

2. Rosén, B, 1997. Asymptotic Theory for Order Sampling. *Journal of Statistical Planning and Inference.* 62: 135-158.

3. Särndal, C-E, 2011a. The 2010 Morris Hansen Lecture. Dealing with Survey Nonresponse in Data Collection, in Estimation. *Journal of Official 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.

Lisad

R-i programm

```
#ANDMETE SISSELUGEMINE
andmed = read.table("/Users/NoraR/Documents/Bakatöö/andmestik.csv",
header=TRUE,sep=",",dec=".",)
head(andmed)

#Vanuse jaotuskarakteristikud üldkogumis
summary(andmed$VANUS)

#Soo jaotuskarakteristikud üldkogumis
sugu=1*(andmed$SUGU==1)
summary(sugu)

#Sissetuleku jaotuskarakteristikud üldkogumis
summary(andmed$SISSETULEK)

#ÜLDKOGUMIST VALIMI SAAMINE
x=c(1:nrow(andmed)) #set.seed(111) s¡-sample(x,770,replace=FALSE)

#Saan andmestikust kätte valimi andmed
valim=andmed[s,c("KEEL","SUGU","VANUS","HARIDUS","ELUKOHT",
"SISSETULEK")]
nrow(valim)

#Binaarsed tunnused kvalitatiivsetele tunnustele
#Binaarsed tunnused keele jaoks
k1=1*(valim$KEEL==1)
k2=1*(valim$KEEL==2)
#Binaarsed tunnused soo jaoks
s1=1*(valim$SUGU==1)
```

```
s2=1*(valim$SUGU==2)
#Binaarsed tunnused hariduse jaoks
h1=1*(valim$HARIDUS==1)
h2=1*(valim$HARIDUS==2)
h3=1*(valim$HARIDUS==3)
h4=1*(valim$HARIDUS==4)
#Binaarsed tunnused elukoha jaoks
e1=1*(valim$ELUKOHT==1)
e2=1*(valim$ELUKOHT==2)
e3=1*(valim$ELUKOHT==3)
e4=1*(valim$ELUKOHT==4)

#Lisan need uued tunnused valimisse
valim=cbind(valim,k1,k2,s1,s2,h1,h2,h3,h4,e1,e2,e3,e4)

#Vanuse ja sissetuleku jaotuskarakteristikud valimis
summary(valim$SISSETULEK)
summary(valim$VANUS)

#Abitunnuste (k1,s1,h1,h2,h3,e1,e2,e3,vanus) keskmised valimis, vanuse koh-
#ta on leitud eespool
summary(valim$s1)
summary(valim$k1)
summary(valim$h1)
summary(valim$h2)
summary(valim$h3)
summary(valim$e1)
summary(valim$e2)
summary(valim$e3)
```

#VASTANUD JA MITTE VASTANUD

#Mahud

N=1538

n=770

m=385

#Logistiline regressioonimudel

#logit(pii)=a+b1*sugu+b2*tsvanus

tsvanus=valim$VANUS - mean(valim$VANUS)

summary(tsvanus)

logit=-0.5*s1+0.05*tsvanus

#s1 - mees

#plot(tsvanus,logit)

pii=exp(logit)/(1+exp(logit))

summa=sum(pii)

vastamistoen=pii/summa*m

sum(vastamistoen)

#Vastamistõenäosuste joonis

plot(tsvanus,vastamistoen,xlab="Tsentreeritud vanus", ylab="Vastamistõenäosused")

#Lisame saadud vastamistõenäosused valimisee

valim=cbind(valim,vastamistoen)

#Juhuslikkuse sissetoomine: genereerime juhusliku suuruse U(0,1) ning jagame

#vastamistõenäosusega

set.seed(112)

U=runif(770)/vastamistoen

```
#Lisame saadud väärtused andmete juurde, sorteerime U kasvamise järjekor-
#ras ning võtame esimesed 385 vastanuteks
valim=cbind(valim,U)
valim=valim[order(U),]
vastanud=valim[1:385,]

#Vanuse ja sissetuleku jaotuskarakteristikud vastanute hulgas
summary(vastanud$VANUS)
summary(vastanud$SISSETULEK)

#Abitunnuste (k1,s1,h1,h2,h3,e1,e2,e3,vanus) keskmised vastanute hulgas,
#vanuse keskmine leitud üleval pool
summary(vastanud$s1)
summary(vastanud$k1)
summary(vastanud$h1)
summary(vastanud$h2)
summary(vastanud$h3)
summary(vastanud$e1)
summary(vastanud$e2)
summary(vastanud$e3)

#Mittevastanud
mittevastanud=valim[386:nrow(valim),]

#Joonis vanuse kohta
op=par(mfrow=c(1,3))
hist(andmed$VANUS,xlim=c(10,80),xlab="Vanus",ylab="Sagedus",freq=F,main="Vanuse
jaotus üldkogumis")
hist(valim$VANUS,xlim=c(10,80),xlab="Vanus",ylab="Sagedus",freq=F,main="Vanuse
jaotus valimis")
```

```
hist(vastanud$VANUS,xlim=c(10,80),xlab="Vanus",ylab="Sagedus",freq=F,main="Vanuse
jaotus vastanute hulgas")
par(op)

#Joonis soo kohta
M=as.matrix(cbind(table(sugu)/N,table(valim$s1)/n,table(vastanud$s1)/m))
dimnames(M)¡-list(c("Naine","Mees"),c("Üldkogum","Valim","Vastanud"))
barplot(M,beside=T,legend=T,args.legend=list(x='topright'))

#TASAKAALUINDIKAATOR ILMA VANUSE JA SOOTA

#Maatriks valimile
xs_vnta=as.matrix(valim[,c("k1","h1","h2","h3","e1","e2","e3")])

#D leidmine
D_vnta=as.matrix(c((mean(vastanud$k1)-mean(valim$k1)),(mean(vastanud$h1)-
mean(valim$h1)),(mean(vastanud$h2)-mean(valim$h2)),(mean(vastanud$h3)-
mean(valim$h3)),(mean(vastanud$e1)-mean(valim$e1)),(mean(vastanud$e2)-
mean(valim$e2)),(mean(vastanud$e3)-mean(valim$e3))))

#Sigma leidmine
M_vnta=1/n*(t(xs_vnta) %*% xs_vnta)
sigma_vnta=solve(M_vnta)

P=m/n
Q=n/m

#BI leidmine, jälgime 3 erinevat tasakaaluindikaatorit

#BI1=1-(t(D1)*Sigma**(-1)*D1/(Q-1))
BI1_vnta=1-((t(D_vnta) %*% sigma_vnta %*% D_vnta)/(Q-1))
BI1_vnta
```

```
#BI2=1-4(P**2)*(t(D1)*Sigma**(-1)*D1)
BI2_vnta=1-4*(P**2)*(t(D_vnta) %*% sigma_vnta %*% D_vnta)
BI2_vnta

#BI3=1-2P(t(D1)*Sigma**(-1)*D1)**(1/2)
BI3_vnta=1-2*P*sqrt(t(D_vnta) %*% sigma_vnta %*% D_vnta)
BI3_vnta

#TASAKAALUINDIKAATOR VANUSEGA JA SOOGA

#Maatriks valimile
xs_vnga=as.matrix(valim[,c("k1","s1","h1","h2","h3","e1","e2","e3","VANUS")])

#D leidmine
D_vnga=as.matrix(c((mean(vastanud$k1)-mean(valim$k1)),(mean(vastanud$s1)-
mean(valim$s1)),(mean(vastanud$h1)-mean(valim$h1)),(mean(vastanud$h2)-
mean(valim$h2)),(mean(vastanud$h3)-mean(valim$h3)),(mean(vastanud$e1)-
mean(valim$e1)),(mean(vastanud$e2)-mean(valim$e2)),(mean(vastanud$e3)-
mean(valim$e3)),(mean(vastanud$VANUS)-mean(valim$VANUS))))

#Sigma leidmine
M_vnga=1/n*(t(xs_vnga) %*% xs_vnga)
sigma_vnga=solve(M_vnga)

P=m/n
Q=n/m

#BI leidmine, jälgime 3 erinevat tasakaaluindikaatorit

#BI1=1-(t(D1)*Sigma**(-1)*D1/(Q-1))
BI1_vnga=1-((t(D_vnga) %*% sigma_vnga %*% D_vnga)/(Q-1))
BI1_vnga
```

```
#BI2=1-4(P**2)*(t(D1)*Sigma**(-1)*D1)
BI2_vnga=1-4*(P**2)*(t(D_vnga) %*% sigma_vnga %*% D_vnga)
BI2_vnga

#BI3=1-2P(t(D1)*Sigma**(-1)*D1)**(1/2)
BI3_vnga=1-2*P*sqrt(t(D_vnga) %*% sigma_vnga %*% D_vnga)
BI3_vnga

#TSÜKKEL
#Leiame ühe kaupa mittevastanuid vastanute hulka lisades maksimiseeriva
objekti

mittevastanud1=mittevastanud[,c("k1","s1","h1","h2","h3","e1",
"e2","e3","VANUS","SISSETULEK")]
vastanud1=vastanud[,c("k1","s1","h1","h2","h3","e1","e2","e3",
"VANUS","SISSETULEK")]
valim1=valim[,c("k1","s1","h1","h2","h3","e1","e2","e3","VANUS")]

#Indikaatorite BI1 arvutamine

ind1=function(x) {
indikaatorid1=rep(NA,nrow(mittevastanud))
for (i in 1:nrow(x)){
vastanud1=rbind(vastanud1,x[i,])
P=nrow(vastanud1)/n
xs=as.matrix(valim1)
D=as.matrix(c((mean(vastanud1$k1)-mean(valim1$k1)),(mean(vastanud1$s1)-
mean(valim1$s1)),(mean(vastanud1$h1)-mean(valim1$h1)),(mean(vastanud1$h2)-
mean(valim1$h2)),(mean(vastanud1$h3)-mean(valim1$h3)),(mean(vastanud1$e1)-
mean(valim1$e1)),(mean(vastanud1$e2)-mean(valim1$e2)),(mean(vastanud1$e3)-
mean(valim1$e3)),(mean(vastanud1$VANUS)-mean(valim1$VANUS))))
```

```
M=1/n*(t(xs) %*% xs)
sigma=solve(M)
BI1=1-((t(D) %*% sigma %*% D)/(Q-1))
indikaatorid1[i]=BI1
vastanud1=vastanud1[-1,]
}
return(indikaatorid1)
}
yks=ind1(mittevastanud1)

#Indikaatorite BI2 arvutamine

ind2=function(x){
indikaatorid2=rep(NA,nrow(mittevastanud))
for (i in 1:nrow(x)){
vastanud1=rbind(vastanud1,x[i,])
P=nrow(vastanud1)/n
xs=as.matrix(valim1)
D=as.matrix(c((mean(vastanud1$k1)-mean(valim1$k1)),(mean(vastanud1$s1)-
mean(valim1$s1)),(mean(vastanud1$h1)-mean(valim1$h1)),(mean(vastanud1$h2)-
mean(valim1$h2)),(mean(vastanud1$h3)-mean(valim1$h3)),(mean(vastanud1$e1)-
mean(valim1$e1)),(mean(vastanud1$e2)-mean(valim1$e2)),(mean(vastanud1$e3)-
mean(valim1$e3)),(mean(vastanud1$VANUS)-mean(valim1$VANUS))))
M=1/n*(t(xs) %*% xs)
sigma=solve(M)
BI2=1-4*(P**2)*(t(D) %*% sigma %*% D)
indikaatorid2[i]=BI2
vastanud1=vastanud1[-1,]
}
return(indikaatorid2)
```

```
}
kaks=ind2(mittevastanud1)

#Indikaatorite BI3 arvutamine

ind3=function(x){
indikaatorid3=rep(NA,nrow(x))
for (i in 1:nrow(x)){
vastanud1=rbind(vastanud1,x[i,])
P=nrow(vastanud1)/n
xs=as.matrix(valim1)
D=as.matrix(c((mean(vastanud1$k1)-mean(valim1$k1)),(mean(vastanud1$s1)-
mean(valim1$s1)),(mean(vastanud1$h1)-mean(valim1$h1)),(mean(vastanud1$h2)-
mean(valim1$h2)),(mean(vastanud1$h3)-mean(valim1$h3)),(mean(vastanud1$e1)-
mean(valim1$e1)),(mean(vastanud1$e2)-mean(valim1$e2)),(mean(vastanud1$e3)-
mean(valim1$e3)),(mean(vastanud1$VANUS)-mean(valim1$VANUS))))
M=1/n*(t(xs) %*% xs)
sigma=solve(M)
BI3=1-2*P*sqrt(t(D) %*% sigma %*% D)
indikaatorid3[i]=BI3
vastanud1=vastanud1[-1,]
}
return(indikaatorid3)
}
kolm=ind3(mittevastanud1)

#Indikaatorid graafikul
op=par(mfrow=c(1,3))
plot(yks,xlab="Mittevastanud",ylab="BI 1",type="l")
plot(kaks,xlab="Mittevastanud",ylab="BI 2",type="l")
```

```
plot(kolm,xlab="Mittevastanud",ylab="BI 3",type="l")
par(op)

#Leiame igast järjendist maksimaalse indikaatori
max(yks)
max(kaks)
max(kolm)
#leiame, mis kohal järjendis maksimaalne element asub
which(yks==max(yks))
which(kaks==max(kaks))
which(kolm==max(kolm))

#Leiame objekti, kelle andmed tasakaalustaksid vastanute hulga.
mittevastanud[which(yks==max(yks)),]
mittevastanud[which(kaks==max(kaks)),]
mittevastanud[which(kolm==max(kolm)),]

#FUNKTSIOONI KORDAV TSÜKKEL
#Leiame ainult hetkel BI3 jaoks

mittevastanud2=mittevastanud1
BI=rep(NA,nrow(mittevastanud2))

for(i in (1:nrow(mittevastanud2))){
a=which(ind3(mittevastanud2)==max(ind3(mittevastanud2)))
vastanud1=rbind(vastanud1,mittevastanud2[a,])
mittevastanud2=mittevastanud2[-a,]
BI[i]=max(ind3(mittevastanud2))
}
#nrow(vastanud1)
#nrow(mittevastanud2)
```

```
BI
plot(BI,ylab="BI3", xlab="Mittevastanud", type="l")

#Moodustame lõpliku vastanute hulga
vastanud_osa=vastanud1[1:600,]
head(vastanud_osa)

#Jaotuskarakteristikud sissetulekule lõplikus vastanute hulgas
summary(vastanud_osa$SISSETULEK)

#Abitunnuste (k1,s1,h1,h2,h3,e1,e2,e3,vanus) keskmised lõplikus vastanute
#hulgas
summary(vastanud_osa$VANUS)
summary(vastanud_osa$s1)
summary(vastanud_osa$k1)
summary(vastanud_osa$h1)
summary(vastanud_osa$h2)
summary(vastanud_osa$h3)
summary(vastanud_osa$e1)
summary(vastanud_osa$e2)
summary(vastanud_osa$e3)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Nora Roosileht, (sünnikuupäev: 10.08.1991),
1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Andmete kogumise juhtimine tasakaaluindikaatori abil,

mille juhendaja on dotsent 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 DSpacei 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, 06.05.2013