Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Mariliis Kütt

Adaptiivne uuringu disain

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: MSc Kristi Lehto,
MSc Mare Vähi

Tartu 2024

ADAPTIIVNE UURINGU DISAIN

Bakalaureusetöö

Mariliis Kütt

Lühikokkuvõte

Suurenev kadu on valikuuringutes laialt levinud probleem, mis võib kaost tingitud nihke tõttu viia ebatäpsete hinnanguteni uuringu põhinäitajate leidmisel. Vastanute hulga kvaliteeti hinnatakse sageli vastamismäära abil. On aga näidatud, et vastamismäära võime prognoosida kaost tingitud nihet on pigem nõrk. Ühe alternatiivina on välja töötatud R-indikaator, mille abil mõõdetakse vastanute hulga esinduslikkust teatava hulga abitunnuste suhtes. Olemasoleva abiinformatsiooni põhjal andmete kogumise juhtimine on keskne idee adaptiivsetes uuringu disainides.

Bakalaureusetöö eesmärk on tutvustada adaptiivse uuringu disaini põhimõtteid ning kirjeldada metoodikat vastamistõenäosuste ja R-indikaatori hindamiseks. Töö teises pooles rakendatakse teooriat Eesti tööjõu-uuringu andmetel, et analüüsida vastanute hulga kvaliteeti esinduslikkusest lähtuvalt.

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

Märksõnad: Valikuuringud, adaptiivne disain, R-indikaator, statistiline andmetöötlus

ADAPTIVE SURVEY DESIGN

Bachelor thesis

Mariliis Kütt

Abstract

Increasing nonresponse is a common challenge in most sample surveys, leading to a higher risk of nonresponse bias in the survey target variables. Response quality is often assessed by monitoring the response rate. However, it has been shown that response rate is an inadequate predictor of nonresponse bias. As an alternative, the R-indicator has been developed which is used to measure the representativeness of the survey response with respect to a set of auxiliary variables. Targeted data collection based on available auxiliary information is a central concept in adaptive survey designs.

The goal of this bachelor thesis is to provide an overview of adaptive survey design principles and to describe the methodology behind estimated response probabilites and the R-indicator. In the second half of the thesis, the methodology is applied to data from the Estonian Labour Force survey to analyse the response quality from the perspective of response representativeness.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics.

Keywords: Sample surveys, adaptive design, R-indicator, statistical data processing

Sisukord

Sissejuhatus

Eesti töö-jõu uuring (ETU) on Statistikaameti poolt aastaringselt läbi viidav isiku-uuring, millega hinnatakse Eesti tööjõu olukorda ja elanike majanduslikku aktiivsust. Tööjõu-uuringus, nagu ka isiku-uuringutes üldiselt, on aktuaalseks probleemiks mittevastanute osakaalu suurenemine. Kui 2012. aastal oli tööjõu-uuringu riiklikuks vastamismääraks 73,8%, siis 2022. aastaks langes näitaja ligikaudu viie protsendipunkti võrra 69,3%-ni (Statistikaamet, 2024). Suurenev kadu võib aga viia ebatäpsete hinnanguteni uuringu põhinäitajate arvutamisel.

On leitud, et vastamismäära seos kaost tingitud nihkega on pigem nõrk, mistõttu peetakse seda kvaliteediindikaatorina ebapiisavaks näitajaks (Shlomo *et al.* 2012). Ühe alternatiivina on välja töötatud vastanute hulga esinduslikkusega seotud R-indikaator, mis põhineb üldkogumi objektide vastamistõenäosuste hajuvusel. Indikaatorit kasutatakse erinevate uuringute või ühe pidevuuringu eri järkude võrdlemiseks, aga ka andmete kogumise juhtimiseks uuringujärgu vältel. Andmete kogumise juhtimine mittevastanute hulga omadustest lähtuvalt on keskne motiiv adaptiivse uuringu disaini (AUD) rakendustes.

Bakalaureusetöö eesmärk on anda teoreetiline ülevaade AUD põhimõtetest, keskendudes seejuures vastanute hulga esinduslikkusega seotud metoodikale. Töö teises pooles rakendatakse R-indikaatorit ETU andmetel, et läbi viia vastanute hulga analüüs möödunud aasta uuringukvartalite lõikes.

Töö esimeses peatükis antakse ülevaade valikuuringutega seotud mõistetest ja tähistustest, millele töös läbivalt tuginetakse. Teises peatükis tutvustatakse AUD komponente, sh metoodikat valimiobjektide vastamistõenäosuste ja R-indikaatori hindamiseks. Kolmandas peatükis antakse lühidalt ülevaade ETU metoodikast, neljandas peatükis kirjeldatakse kasutatud andmestikku ja andmeanalüüsi tulemusi.

1 Valikuuringute mõisted ja tähistused

Olgu $U = \{1, 2, \ldots, N\}$ lõplik üldkogum mahuga N, millest on tõenäosusliku valikuga võetud valim s mahuga n. Kasutatud valikudisain määrab kaasamistõenäosuse $\pi_i = P(i \in s)$, st tõenäosuse, millega objekt $i \in U$ valimisse kaasatakse. Indikeerigu s_i valimisse kaasatust ($s_i = 1$, kui $i \in s$, vastasel juhul $s_i = 0$). Eeldusel, et tegu on tagasipanekuta valikuga, on disainikaal kaasamistõenäosuse pöördväärtus, $d_i = \frac{1}{\pi_i}$.

Üldiselt esineb valikuuringus objekti kadu, st erinevatel põhjustel on andmestikust puudu terve(d) objekt(id) – valimisse sattunu keeldub uuringus osalemast, isik pole elukoha või kontaktandmete muutuse tõttu kättesaadav jms. Tähistagu r realiseerunud vastanute hulka, mille maht on m ning mis rahuldab tingimusi $r \subseteq s \subset U$, $r \neq \varnothing$. Defineerime tunnuse r_i, mis indikeerib küsitlusele vastamist ($r_i = 1$, kui $i \in r$, vastasel juhul $r_i = 0$).

Uuringus võivad vaatluse all olla mitmed üldkogumi parameetrid, mida soovitakse valimi põhjal hinnata. Tähistagu y uuritavat tunnust (e funktsioontunnust), mille väärtus objekti i jaoks on y_i. Eelnevast nähtub, et kui uuringus esineb mittevastamist, ei ole y väärtused teada kõikide valimi objektide jaoks, mis paraku mõjub negatiivselt leitavate hinnangute kvaliteedile. Kaoviga kirjeldatakse lähemalt järgmises alapeatükis.

Lisaks funktsioontunnustele on uuringus relevantsed ka abitunnused, mille väärtused on teada üldkogumi (või valimi) iga objekti kohta. Abiinformatsioon pärineb üldjuhul kas erinevatest registritest (nt rahvastikuregister, hooneregister), eelnevatest uuringutest või käimasoleva uuringu jooksul kogutud paraandmetest (näiteks hiireklikid ja vastamise ajatemplid veebiküsitluses (Schouten *et al.* 2017)).

Olgu kasutatavaid abitunnuseid J. Sel juhul märgime objekti i jaoks abitunnuste vektori tähisega $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})^T$, st i. objekti j. abitunnuse väärtus on x_{ij}, $i \in U$, $i = 1, \ldots, N$, $j = 1, \ldots, J$.

1.1 Juhusliku vea komponendid

Praktikas on keeruline leida valikuuringut, kus hinnangu juhuslikkus tuleneb ainult kasutatud valikudisainist. Üldjuhul kaasneb leitud hinnanguga juhuslik viga, mille mõõt sisaldab erinevaid veakomponente. Järgnev alapeatükk annab ülevaate põhilistest veakomponentidest, seejuures tuginetakse raamatule Traat *et al.* (1997).

Olgu $\hat{\theta}$ valimi põhjal leitud hinnang parameetrile θ. Vahet $\hat{\theta}-\theta$ nimetatakse hinnangu θ veaks. Vea mõõtmiseks kasutatakse vea ruudu keskväärtust ehk ruutkeskmist viga (MSE), mis avaldub valemiga

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = B^2 + D\hat{\theta},$$

kus $B = E\hat{\theta} - \theta$ ja $D\hat{\theta}$ tähistavad vastavalt hinnangu nihet ning dispersiooni. Igasugune mõju ruutkeskmisele veale mõjub seega omakorda kas hinnangu nihkele, dispersioonile või mõlemale.

Allpool on kirjeldatud põhilisi veakomponente.

- **Valikuviga** on valikudisainist põhjustatud varieeruvus hinnangus (üle kõikide võimalike valimite disaini $p(s)$ korral). Kui andmeid kogutakse üldkogumi asemel valimist, sisaldab leitud hinnang alati valikuviga.

- **Loendiviga** on ebakorrektsest loendist põhjustatud varieeruvus hinnangus. Uuringus kasutatavas loendis on ülekaetus, kui see sisaldab üldkogumisse mittekuuluvaid objekte. Teistpidi, kui loend ei sisalda kõiki üldkogumi objekte, on tegu alakaetusega. Ülekaetus ei ohusta hinnangu nihketust, ent alakaetud loendi korral võivad leitud hinnangud olla juba olulise nihkega.

- **Kaoviga** on tingitud mittevastamisest ehk kaost. Vastanute hulga põhjal arvutatud hinnangud on üldkogumi suhtes nihkega, kui kao ning vastanute hulgas on näitajad oluliselt erinevad. Kadu põhjustab hinnangu dispersiooni suurenemist.

- **Mõõtmisviga** tekib uuritava tunnuse väärtuse mõõtmisel (nt mõõtmisvahendi konstruktsioon on vigane, objektide mõõtmisolukorrad varieeruvad, intervjueerija märgib vastuse valesti üles). Mõõtmisviga suurendab hinnangu nihet ja dispersiooni.

- **Töötlusviga** on põhjustatud andmete ebakorrektsest kodeerimisest, sisestamisest, analüüsimisest ning tabuleerimisest.

Erinevatest vealiikidest põhjustatud nihete kõrvaldamiseks või vähendamiseks kasutatakse kompenseerimismeetodeid. Levinud meetod on järelkihistamine, mille abil saab vähendada loendist ja kaost põhjustatud nihet. Kadu kompenseeritakse ka kalibreerimismeetoditega, mis põhinevad teadaolevate abitunnuste alusel uute kaalude arvutamises pärast andmete kogumise etappi.

Vastanute ja kao hulga erinevuse tuvastamiseks ei pruugi alati olla piisavalt informatsiooni. Mittevastamisest tingitud nihet saab kõige kindlamalt vältida kao osakaalu vähendades, st suunates olemasolevad ressursid andmete kättesaamisele kao objektidelt.

2 Adaptiivne uuringu disain

Kadu on valikuuringutes üldlevinud probleem. Seetõttu tuntakse huvi uuringu disainide vastu, mis võimaldavad jooksvalt arvestada mittevastanute hulga mahu ning omadustega. Igasuguse valikudisaini korral on esialgselt tarvis uuringu põhinäitajate kohta seada mõningad eeldused. Adaptiivne lähenemine lubab püstitatud eelduseid ja kogutud andmeid võrrelda juba küsitlusperioodi vältel, et vajadusel teha disainis muudatusi. Kohandamise eesmärgiks ei ole üksnes soovitud vastanute arvuni jõudmine, vaid ka hinnangutes piisava täpsuse saavutamine juhul, kui uuringus esineb mittevastamist. (Wagner, 2008)

Üldjuhul rakendatakse valimi igale objektile ühesugust küsitlusstrateegiat. AUD keskseks põhimõtteks on eeldus, et olemasolevat informatsiooni kasutades saab leibkondade või isikute küsitlemisele läheneda erinevalt, parandades seeläbi vastanute hulga kvaliteeti. Abiinformatsiooni olemuse põhjal saab adaptiivsed disainid jagada staatilisteks ja dünaamilisteks. Staatiliste disainide aluseks on küsitlemise eelselt kättesaadavad – näiteks registritest ja varasematest uuringutest pärinevad – abitunnused, dünaamilised disainid kasutavad (lisaks eelnevale) küsitluse läbiviimisel tekkinud paraandmeid. (Schouten *et al.* 2013)

Dünaamilise disaini illustreerimiseks toovad Schouten *et al.* (2013) näite küsitlusest, kus vaadeldakse osalejate nn koostöösoodumust. Kogutud andmete põhjal määratakse küsitlejad uuesti intervjueerima valimi objekte, kes esialgselt keeldusid küsitlusele vastamast. Kirjeldatud strateegia võimaldab hinnata valimi objekti tõenäosust uuringus osaleda korduvküsitluse korral. Staatilise AUD näitena vaadeldakse Hollandi tarbijate rahulolu uuringut, mille pealt simuleeritav disain toetub peamiselt küsitleja resultatiivsusele. Disaini strateegia kohaselt määratakse oskuslikematele küsitlejatele keerukamad valimi objektid, intervjueerija resultatiivsust käsitletakse eelnevalt teadaoleva abitunnusena.

Adaptiivse disaini täpne ülesehitus sõltub lisaks abiinformatsioonile ka uuringu prioriteetidest. Vastamismäärade vähenemise aktuaalsuse tõttu on adaptiivse uurin-

gu disaini teooria ja rakendused keskendunud eeskätt kaost tingitud nihkele. Ent AUD ei piirne vaid kao ega ka nihke adresseerimisega. Schouten *et al.* (2017) kirjeldavad nelja põhilist eesmärgitüüpi, millest adaptiivsed uuringu disainid lähtuvad:

- **Kulud.** Uuringu eelarvest ja selle fikseeritusest sõltub suuresti, millist metoodikat on adaptiivses disainis võimalik kasutada. Praktikas on levinud fikseeritud eelarvega valikuuringud, mis vajavad näiteks valimimahu või läbiviidavate intervjuude arvu optimeerimist.

- **Dispersioon.** Prioriteetne on hinnangute hajuvuse kontrollimine. Dispersiooni suurust võivad mõjutada kõik alapeatükis (1.1) kirjeldatud vealiigid.

- **Nihe.** Eesmärgiks seatakse hinnangute nihketus. Ka siin saab kitsamalt keskenduda kindlat tüüpi veakomponendile – nihe võib olla põhjustatud nii loendi-, kao-, mõõtmis- kui töötlusveast.

- **Muud kvaliteedimõõtmed.** Uuring võib olla ajaliselt piiratud, st lõplikud andmed ning hinnangud tuleb leida kindlaksmääratud aja jooksul. Ka vastamismäära loetakse eraldi kvaliteedimõõtmeks.

Eesmärkide seadmisel tasub arvesse võtta, et sageli tähendab prioriteedi valik kompromissi teiste uuringus kehtestatud piirangutega. Illustreerivaks näiteks on uuring, mille põhieesmärk on kaost põhjustatud nihke vähendamine. Eesmärgi täitmiseks küsitletakse kao objekte efektiivsema, aga kulukama küsitlusstrateegia alusel. Mittevastanute kättesaamise suuremad kulud võivad aga kokkuvõttes viia uuringu valimimahu vähendamiseni. (Schouten *et al.* 2017)

2.1 Kao indikaatorid

Üldjuhul ei ole võimalik otse leida, kui suur on kaost tingitud nihe uuritava tunnuse hinnangus. Lisaks, et valdavalt on uuringutes vaatluse all mitmed üldkogumi parameetrid, tekib ka hulgaliselt erinevaid kaost põhjustatud nihkeid, mida uurida (Schouten *et al.* 2012). Kao mõjude analüüsimiseks on välja töötatud kaudsed indikaatorid (*proxy indicators*), millest juhinduda andmete kogumisel küsitlusetapil.

Mittevastamisega seotud kvaliteediindikaatorid saab laias laastus jagada kaheks. Esimest tüüpi näitajad kasutavad ainult abitunnustest saadavat informatsiooni (nn *kovariandipõhised* indikaatorid). Teist tüüpi indikaatorid rakendavad lisaks abitunnustele ka uuritavaid tunnuseid, st põhinevad vektoris $\mathbf{x}_i$ sisalduvate abitunnuste ning uuritava(te) tunnus(t)e suhtel (nn *kovariandi- ja tunnusepõhised* indikaatorid). (Schouten *et al.* 2017)

Töös vaadeldakse lähemalt R-indikaatorit, mida loetakse vastanute hulga esinduslikkuse mõõduks (tulenevalt ingliskeelsest terminist *representativeness*). R-indikaator on olemuselt kovariandipõhine, st selle hinnang lähtub ainult teadaolevast abiinformatsioonist. Lisaks üldisele R-indikaatorile on välja töötatud ka osalised R-indikaatorid, mis võimaldavad abitunnuste alusel detailsemalt tuvastada gruppe, kes on küsitlusperioodi vältel vastanute hulgas ala- või üleesindatud. Antud töös tutvustatakse ainult üldist R-indikaatorit, osalistest R-indikaatoritest annab täpsema ülevaate artikkel Schouten *et al.* (2012).

R-indikaator põhineb üldkogumi objektide hinnatud vastamistõenäosustel. Seega tuuakse käesolevas alapeatükis esmalt välja vastamistõenäosuse definitsioon ja kirjeldatakse metoodikat selle hindamiseks. Seejärel selgitatakse esinduslikkuse mõistet ja defineeritakse töö teises pooles kasutatav R-indikaator.

2.1.1 Vastamistõenäosuste leidmine

Vastamistõenäosuste kirjeldamisel tuginetakse artiklitele Schouten *et al.* (2009) ja Bethlehem (2020).

Olgu antud üldkogum U, millest on võetud valim s. Esinegu uuringus mittevastamist, mille tulemusel on lõplikuks vastanute hulgaks r. Eeldame esmalt, et iga üldkogumi objekti kohta on teada vastamistõenäosus ρ_i, mis näitab, kui tõenäoliselt objekt i valimisse sattumise korral küsitlusele vastab:

$$\rho_i = P(r_i = 1 \mid s_i = 1), \quad i = 1, \ldots, N. \tag{1}$$

Olgu $\mathbf{p}$ vastamistõenäosuste vektor, st $\mathbf{p} = (\rho_1, \rho_2, \ldots, \rho_N)^T$. Vastamistõenäosuste standardhälve on arvutatav valemiga

$$S(\mathbf{p}) = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\rho_i - \bar{\rho})^2}, \tag{2}$$

kus $\bar{\rho} = 1/N \sum_{i=1}^{N} \rho_i$ tähistab vastamistõenäosuste keskmist üldkogumis. Võrdsete vastamistõenäosuste korral $S(\mathbf{p}) = 0$, tõenäosuste varieeruvuse suurenedes $S(\mathbf{p})$ väärtus kasvab.

Praktikas on raske leida uuringut, kus vastamistõenäosused oleks eelnevalt teada kogu üldkogumi kohta. Sel juhul hinnatakse need olemasoleva abiinformatsiooni põhjal. See tähendab, et tõenäosuse (1) asemel huvitume vastamise tõenäosusest tingimusel, et objekti i korral on teada abitunnuste vektori $\mathbf{x}_i$ väärtused:

$$\rho_i(\mathbf{x}_i) = P(r_i = 1 \mid \mathbf{x}_i), \quad i = 1, \ldots, N. \tag{3}$$

Eeldusel, et abitunnused on mõõdetud nii vastanute kui mittevastanute jaoks, saab vastamistõenäosuse (1) asemel kasutada tõenäosuse (3) mudelipõhist hinnangut. Üheks levinuimaks meetodiks on logistilise regressiooni kasutamine, mille kohaselt

esitub abitunnuste ja vastamistõenäosuse $\rho_i(\mathbf{x}_i)$ seos kujul

$$logit(\rho_i) = logit(\rho_i(\mathbf{x}_i)) = log\left(\frac{\rho_i(\mathbf{x}_i)}{1-\rho_i(\mathbf{x}_i)}\right) = \sum_{j=1}^{J} x_{ij}\beta_j, \tag{4}$$

kus $\beta = (\beta_1, \beta_2, \ldots, \beta_J)^T$ on J regressioonikordajat sisaldav vektor. Seosest (4) saab avaldada vastamistõenäosuse hinnangu:

$$\hat{\rho}_i = \hat{\rho}_i(\mathbf{x}_i) = \frac{e^{logit(\rho_i)}}{1+e^{logit(\rho_i)}}.$$

Mudelipõhise lähenemise korral kasutame standardhälbe arvutamisel valemi (2) asemel hinnangut

$$\hat{S}(\hat{\mathbf{p}}) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} s_i d_i \left(\hat{\rho}_i - \hat{\bar{\rho}}\right)^2},$$

kus $\hat{\mathbf{p}} = (\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_N)^T$ tähistab hinnatud vastamistõenäosuste vektorit ja $\hat{\bar{\rho}}$ hinnangut vastamistõenäosuste kaalutud keskmisele:

$$\hat{\bar{\rho}} = \frac{1}{N}\sum_{i=1}^{N} s_i d_i \hat{\rho}_i.$$

2.1.2 R-indikaator

Alljärgnevalt selgitatakse esinduslikkuse mõistet ning defineeritakse R-indikaator, lähtudes seejuures allikatest Muusikus (2015) ja Schouten *et al.* (2009).

Esinduslikkuse tugeva definitsiooni kohaselt on vastanute hulk valimi suhtes esinduslik, kui üldkogumi objektide vastamistõenäosused (1) on võrdsed ning objektide vastamine on üksteisest sõltumatu. Praktikas on tugevat definitsiooni keeruline kontrollida, mistõttu on esinduslikkusele konstrueeritud ka nõrk definitsioon:

Definitsioon (nõrk)**.** Olgu antud kvalitatiivne tunnus X, mille on H kihti. Vastanute hulk on tunnuse X suhtes esinduslik, kui keskmine vastamistõenäosus alamklassides on konstantne, st

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho, \quad h = 1, \ldots, H,$$

kus N_h on kihi h suurus, ρ_{hi} on h. kihi k. objekti vastamistõenäosus ja summeeritakse üle kõigi objektide kihis h.

Vastavalt tugevale ja nõrgale definitsioonile defineeritakse R-indikaator kahel viisil. Esmalt eeldame, et vastamistõenäosused (1) on eelnevalt teada. Siis taandub esinduslikkus vastamistõenäosuste hajuvusele – mida rohkem on varieeruvust vastamistõenäosustes, seda vähem esinduslik on vastanute hulk tugeva definitsiooni kohaselt. R-indikaator on sel juhul defineeritav kujul

$$R(\mathbf{p}) = 1 - 2S(\mathbf{p}),$$

mis võrdub ühega juhul, kui vastamistõenäosused on konstantsed ehk tegemist on täielikult esindusliku vastanute hulgaga. Kuna standardhälve $S(\mathbf{p})$ jääb vahemikku $[0; 0,5]$, on R-indikaatori väärtuste vahemikuks $[0; 1]$.

Tasub tähele panna, et indikaatori minimaalne väärtus sõltub vastamistõenäosuste keskmisest üldkogumis. Keskmise $\bar{\rho} = 1/2$ korral on $R(\mathbf{p})$ alumiseks piiriks 0. Madala $\bar{\rho}$ korral ei saa vastamistõenäosused olla tugevalt varieeruvad, mistõttu keskmise kahanemisel piirkonnas $(0; 0,5)$ R-indikaatori alumine piir kasvab. Keskmise $\bar{\rho} = 0$ korral on indikaatori minimaalne väärtus taas võrdne ühega vastamistõenäosuste konstantsuse tõttu.

Kui vastamistõenäosused ei ole eelnevalt teada ja need tuleb mudeli põhjal hinnata, on R-indikaatori hinnanguks

$$\hat{R}(\hat{\mathbf{p}}) = 1 - 2\hat{S}(\hat{\mathbf{p}}) = 1 - 2\sqrt{\frac{1}{N-1} \sum_{i=1}^{N} s_i d_i \left(\hat{\rho}_i - \hat{\bar{\rho}}\right)^2}. \tag{5}$$

Kui hinnangute $\hat{\rho}_i$ leidmiseks kasutatavad abitunnused $\{X_1, \ldots, X_J\}$ on kvali-

tatiivsed, saab R-indikaatorit tõlgendada kui (vektori X põhjal) moodustuvate alamgruppide vastamismäärade hajuvuse mõõtu. Vastamismäärade varieeruvuse kasvades R-indikaatori väärtus langeb.

Avaldis (5) on niisiis seotud kahe juhusliku protsessiga: esmalt hinnatakse valimi põhjal objektide vastamistõenäosused, seejärel leitakse hinnang vastamistõenäosuste standardhälbele. Seega tekib küsimus R-indikaatori standardhälbe kohta, mille abil konstrueerida indikaatori hinnangu usaldusintervall.

Käesolevas töös R-indikaatori standardhälbe hindamist ei käsitleta. Varasemas kirjanduses on standardhälvet hinnatud näiteks mitteparameetrilise *bootstrap* meetodiga (Schouten *et al.* 2009).

3 Eesti tööjõu-uuring

ETU üldkogumi moodustavad vähemalt 15-aastased Eesti elanikud (välja arvatud ajateenijad ja institutsioonides viibijad) (Statistikaamet, 2021). Uuringus käsitletakse alalist rahvastikku, st isikuid, kes on riigis elanud üks või enam aastat (Statistikaamet, 2012).

Valikudisainina kasutatakse süstemaatilist juhuslikku kihtvalikut. Isiku elukoha piirkonna põhjal moodustatakse neli kihti: 1) Tallinn; 2) neli suuremat maakonda (Harju (v.a Tallinn), Ida-Viru, Pärnu ja Tartu maakond); 3) 10 väiksemat maakonda; 4) Hiiu maakond. Kõige tõenäolisemalt kaasatakse valimisse Hiiu maakonna leibkonnad. (Statistikaamet, 2012; Statistikaamet, 2024)

Valim on isikuviisiline, valimisse sattunud isik vastab nii leibkonna- kui ka isikuküsitlusele. Samuti vastavad isikuküsitlusele valimiisiku kõik vähemalt 15-aastased leibkonnaliikmed. Küsitlusviisina kasutatakse uuringunädala järgselt 1.–4. päeval veebiküsitlust (CAWI), 5. päevast alates telefoniintervjuud (CATI). Erandiks on teadaoleva e-posti aadressita isikud, kellele rakendatakse CATI meetodit esimesest päevast alates. Uuringunädala järgne küsitlusperiood kestab kokku 20 päeva, kuu viimasel nädalal kaasatud leibkondi küsitletakse 13 päeva jooksul alates küsitlusperioodi algusest.

Uuringusse kaasatud leibkonda küsitletakse neljal korral. Lisaks esmakordsele küsitlemisele võetakse leibkond valimisse ka järgnevas kvartalis, seejärel küsitletakse sama leibkonda järgmise aasta samades kvartalites. Igas kvartalis moodustavad 36% valimist esmakordselt küsitletud leibkonnad. (Statistikaamet, 2012)

Andmeid kogutakse igakuiselt, sõltuvalt andmetabelist avaldatakse uuringu tulemused iga kuu, kvartali või kalendriaasta kohta.

4 Praktiline ülesanne

Käesolevas peatükis rakendatakse eelnevalt kirjeldatud metoodikat ETU vastanute hulga esinduslikkuse analüüsimiseks R-indikaatori abil. Analüüsi teostamiseks loodud programmikood indikaatorite arvutamise ning tabelite ja jooniste koostamisega on leitav lisast 2.

4.1 Andmestiku kirjeldus

Ülesandes analüüsiti tööjõu-uuringu 2023. aasta andmeid (edaspidi ka ETU 2023). Kuna andmed on konfidentsiaalsed, ei ole töös kasutatud andmestik avalikult kättesaadav.

Disainikaalude kasutamise lihtsustamiseks vaadeldi ainult esimest korda küsitletud leibkondi ja nende vastamist leibkonnaküsitlusele. Kasutatud andmestik koosneb 6 509 leibkonnast, kelle kohta on andmestikus olemas väärtused järgnevatele tunnustele:

- KV – kvartali number (1-4);
- DKAAL – leibkonna disainikaal;
- SUGU – valimiisiku sugu (1 – mees, 2 – naine);
- VANUSGR – valimiisiku vanusegrupp (15–29, 30–49, 50–69, 70+);
- MLINN – linnastumise aste (1 – linn; 2 – maa);
- KYSPAEV – küsitlusele vastamise päev perioodis (1-20);
- VASTUS – vastamise indikaator (1 – vastas; 0 – ei vastanud).

Enne andmete analüüsimist arvutati esialgse disainikaalu põhjal korrigeeritud disainikaal tunnusena DKAAL_UUS, mis üle kvartali summeerituna võrdub üldkogumi

suurusega. Vastanute hulga esinduslikkust analüüsitakse edaspidises valimiisiku soost ja vanusegrupist ning leibkonna elukohast (linn või maa lähtuvalt.

4.2 Valim ja vastanute hulk

Ebaõnnestunud kontaktide ja keeldumiste tõttu oli lõplikuks vastanute arvuks 3 511 esmaküsitletud leibkonda. Tabelis 1 on toodud abitunnuste väärtuste sagedused esimest korda küsitletud leibkondade seas nii valimi kui vastanute hulga lõikes. Kuigi valimisse sattus 47% mehi, oli vastanute hulgas meeste osakaaluks 43, 9%. Analoogne erinevus ilmnes ka linnastumise tasemete lõikes: võrreldes valimiga (67, 2%) oli linnapiirkonnas elavate isikute osakaal vastanute seas 2,8% võrra madalam. Kuni 50-aastaste osakaal üldiselt langes, st vanemaealised inimesed moodustasid vastanutest oodatust suurema osa.

Tabel 1: Abitunnuste jaotumine ETU 2023 valimis ja vastanute hulgas (esimest korda küsitletud leibkonnad)

	Sagedused		**Osakaalud (%)**	
	Valim	**Vastanud**	**Valim**	**Vastanud**
Sugu				
Mehed	3056	1543	47,0	43,9
Naised	3453	1968	53,0	56,1
Vanusegrupp				
15–29	1083	475	16,6	13,5
30–49	2247	1132	34,5	32,2
50–69	1993	1172	30,6	33,4
70+	1186	732	18,2	20,8
Linnastumise aste				
Linn	4377	2262	67,2	64,4
Maa	2132	1249	32,8	35,6

Valimi ja vastanute hulga jaotus kvartalite lõikes on leitav lisa 1 tabelitest 4 ja 5.

4.3 Vastanute hulga esinduslikkus

Et võrrelda nelja uuringukvartali lõpliku vastanute hulga esinduslikkust valimi suhtes, leiti R-indikaatori hinnang kvartalite viimase küsitluspäeva seisuga. Arvutustes hinnati vastamistõenäosused logistilise regressiooniga, kasutades *logit* mudeli argumenttunnustena valimiisiku sugu, vanusegruppi ja linnastumise astet:

$$logit(\rho_i) = \beta_0 + \beta_1 \cdot SUGU + \beta_2 \cdot VANUSGR + \beta_3 \cdot MLINN.$$

Kvartali lõppseisu R-indikaatori hinnangud koos vastanute arvu ja vastamistõenäosuste kaalutud keskmisega on toodud tabelis 2. Igas kvartalis vastas küsitlusele umbes 870 leibkonda, seejuures püsis uuringukvartali R-indikaator aasta jooksul võrdlemisi stabiilsena. Hinnangud vastamistõenäosuste keskmisele on ligilähedased Statistikaameti kodulehel avaldatud vastamismääradega (Statistikaamet, 2024). R-indikaatori hinnangu põhjal saavutati sarnane esinduslikkus I ja III kvartalis ning II ja IV kvartalis.

Tabel 2: Vastanute arv, vastamistõenäosuste kaalutud keskmine ja R-indikaator ETU 2023 uuringukvartalite lõppseisuga (esimest korda küsitletud leibkonnad)

	I	**II**	**III**	**IV**
m	897	861	876	877
$\hat{\bar{\rho}}$	0,550	0,528	0,536	0,537
$\hat{R}$	0,821	0,861	0,829	0,871

Lisaks kvartali lõppseisule uuriti ka R-indikaatori muutumist üle kvartali küsitlusperioodide, et analüüsida, kuivõrd esinduslik on vastanute hulk küsitlusperioodi vältel. Esmalt tehti iga kvartali jaoks arvutused läbi perioodi kolmes erinevas punktis, mille valikul lähtuti ETU-s kasutatavate küsitlusviiside etappidest:

1. 4. päev (CAWI – veebiküsitluse etapi lõpp);
2. 13. päev (CATI_{13} – telefoniküsitluste lõpp kuu viimasel nädalal kaasatud leibkondade jaoks);
3. 20. päev (CATI_{20} – telefoniküsitluste lõpp ülejäänud leibkondade jaoks).

Küsitlusperioodi lõikes leitud hinnangud iga kvartali kohta on toodud tabelis 3. Igas kvartalis oli üksikuid leibkondi, keda intervjueeriti erandjuhul ka pärast 20. küsitluspäeva. Seega on 20. päeva hinnangud mõnevõrra erinevad tabelis 2 toodud kvartali lõppseisu hinnangutest.

Tabel 3: Vastanute arv, vastamistõenäosuste kaalutud keskmine ja R-indikaator ETU 2023 kvartalite 4., 13. ja 20. päeva seisuga (esimest korda küsitletud leibkonnad)

	I			**II**		
	CAWI	**CATI_{13}**	**CATI_{20}**	**CAWI**	**CATI_{13}**	**CATI_{20}**
m	136	771	892	105	687	856
$\hat{\bar{\rho}}$	0,083	0,472	0,547	0,064	0,421	0,525
$\hat{R}$	0,908	0,806	0,819	0,915	0,820	0,857
	III			**IV**		
	CAWI	**CATI_{13}**	**CATI_{20}**	**CAWI**	**CATI_{13}**	**CATI_{20}**
m	126	726	859	130	702	863
$\hat{\bar{\rho}}$	0,077	0,443	0,525	0,080	0,429	0,528
$\hat{R}$	0,902	0,803	0,829	0,896	0,829	0,867

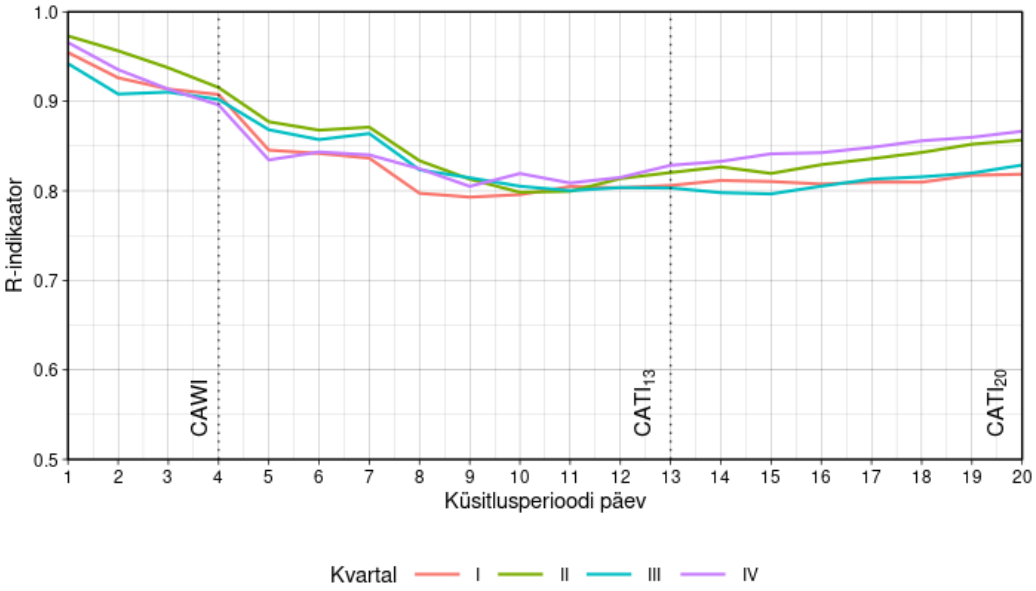
Tabelist 3 ilmneb, et kuigi R-indikaatori hinnang oli kõrgeim CAWI etapi lõpuks, on see pigem tingitud vähesest vastanute arvust perioodi alguspäevadel. Vastamistõenäosuste keskmise hinnang 4. päeva seisuga oli igas kvartalis alla 10%. Punktis CATI_{13} oli vastanud leibkondi juba tunduvalt rohkem ning vastamistõenäosuste keskmine ligikaudu 45%. Iga kvartali puhul on näha, et küsitlusperioodi 3. nädalaga

(14.-20. päevaga) õnnestus esinduslikkust mõnevõrra parandada, märgatavaim tõus R-indikaatori hinnangus toimus seejuures II ja IV kvartalis.

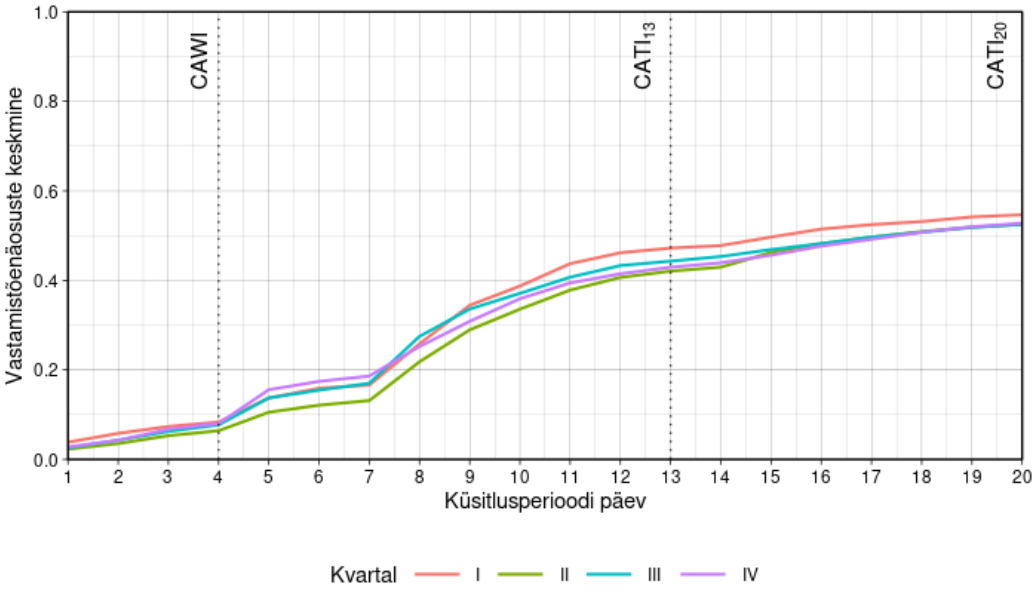
Et esinduslikkust küsitlusperioodi jooksul veelgi tihedamini jälgida, võib R-indikaatori hinnangu leida ka iga küsitluspäeva seisuga. Detailsema vaate abil saab analüüsida, kuivõrd märgatavalt mõjutavad küsitlusperioodi viimastel päevadel intervjueeritud leibkonnad vastanute hulga esinduslikkust. ETU 2023 kvartalite põhjal leitud tulemused on esitatud joonisel 1, kus on vertikaalse katkendjoonega ära märgitud eelnevalt kirjeldatud punktid küsitlusperioodi jooksul. Et kvartalite erinevused oleksid paremini nähtavad, on y-telje minimaalseks väärtuseks määratud 0,5.

Jooniselt 1 on samuti märgata teatavat sarnasust I ja III ning II ja IV kvartali vahel, nagu sai eelnevalt täheldatud tabeli 2 põhjal. Küsitlusperioodi viimasel nädalal on II ja IV kvartali R-indikaatori muutust indikeerivad jooned peaaegu paralleelsed. Jooniselt paistab, et küsitlusperioodi 11. päevaks on kvartalite R-indikaatori hinnangud jõudnud ligikaudselt samale tasemele.

Nagu tabelist 3 nähtus, võib küsitlusperioodi esimestel päevadel olla liialt vähe vastajaid, et üksnes R-indikaatorist lähtuvalt vastanute hulga kvaliteedi kohta põhjapanevaid järeldusi teha. Seega on paralleelselt R-indikaatoriga otstarbekas jälgida küsitluspäevaks saavutatud vastamismäära või hinnatud vastamistõenäosuste keskmist. Jooniselt 2 on näha sama muster, mis ilmnes tabeli 3 juures – kuni 4. päevani on keskmine igas kvartalis väga madal, mistõttu on R-indikaatori hinnang CAWI päevadel ühelähedane (vt joonis 1). Vastamistõenäosuste keskmine hakkab märgatavamalt kasvama alates 7. küsitluspäevast.



Joonis 1: ETU 2023 kvartalite R-indikaator
küsitluspäevade lõikes (esimest korda küsitletud leibkonnad)



Joonis 2: ETU 2023 kvartalite keskmine vastamistõenäosus
küsitluspäevade lõikes (esimest korda küsitletud leibkonnad)

Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks oli anda põgus ülevaade adaptiivsest uuringu disainist ning võimalustest hinnata vastanute hulga kvaliteeti uuringu andmekogumisetapil. Töö teoreetilises osas käsitleti põhjalikumalt vastanute hulga esinduslikkusega seotud R-indikaatorit, mis põhineb teadaolevate abitunnuste alusel hinnatud vastamistõenäosustel.

Töö praktilises osas rakendati R-indikaatori metoodikat, et analüüsida vastanute hulga esinduslikkust Eesti tööjõu-uuringu andmetel. Vastamistõenäosuste hindamisel lähtuti seejuures valimiisiku soost, vanusegrupist ning linnastumise tasemest. Tulemused näitasid, et 2023. aastal oli uuringukvartalite lõppseisu vaates R-indikaatori hinnang võrdlemisi stabiilne, jäädes kõigi nelja kvartali puhul ligikaudu 80% juurde.

Analüüsides esinduslikkust üle kvartali küsitlusperioodide, ilmnes probleem R-indikaatori hindamisega veebiküsitluse etapi lõpul. Nimelt põhjustab 4. küsitluspäeva lõpuks saavutatud madal vastamismäär vähest varieeruvust vastamistõenäosustes, mis viib aga ekslikult kõrge R-indikaatori väärtuseni. Seega tasub vastanute hulga kvaliteedi analüüsimisel koguda informatsiooni nii R-indikaatori kui vastamismäära kohta, eriti küsitlusperioodi alguspäevadel.

Töös jäi katmata teooria R-indikaatori hinnangu standardvea kohta. Indikaatori punkthinnangus põhjustab ebatäpsust juhuslikkus, mis tuleneb vastamistõenäosuste ja nende standardhälbe hindamisest valimi põhjal. Erinevate uuringute ja andmekogumismeetodite võrdlemisel tuleks valiidsemate järelduste tegemiseks R-indikaatori hinnangud esitada koos usaldusintervallidega.

Edaspidistes analüüsides tasub lisaks üldisele R-indikaatorile keskenduda ka artiklis Schouten *et al.* (2012) käsitletud osalistele R-indikaatoritele, mille abil detailsemalt tuvastada alamgruppe, kes on küsitlusetapi vältel vastanute hulgas ala- või üleesindatud.

Kasutatud allikad

Bethlehem, J. (2020). "Working with Response Probabilities". *Journal of Official Statistics* 36.3, lk. 647–674. DOI: 10.2478/jos-2020-0033.

Muusikus, M. (2015). "Valikuuringutes vastanute hulga kvaliteeti mõõtvad indikaatorid". Bakalaureusetöö. Tartu Ülikool.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. ja Skinner, C. (2012). "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators". *International Statistical Review* 80.3, lk. 382–399. ISSN: 0306-7734, 1751-5823. DOI: 10.1111/j.1751-5823.2012.00189.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2012.00189.x.

Schouten, B., Calinescu, M. ja Luiten, A. (2013). "Optimizing quality of response through adaptive survey designs". *Survey Methodology* 39.1, lk. 29–58. URL: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X201300111824.

Schouten, B., Cobben, F. ja Bethlehem, J. (2009). "Indicators for the representativeness of survey response". *Survey Methodology* 35.1, lk. 101–113.

Schouten, B., Peytchev, A. ja Wagner, J. (2017). *Adaptive Survey Design*. CRC Press.

Shlomo, N., Skinner, C. ja Schouten, B. (2012). "Estimation of an indicator of the representativeness of survey response". *Journal of Statistical Planning and Inference* 142.1, lk. 201–211. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2011.07.008.

Statistikaamet (2012). *Eesti tööjõu-uuring. Metoodika*. URL: https://www.stat.ee/sites/default/files/2020-12/Eesti%20t%C3%B6%C3%B6j%C3%B5-uuring.pdf.

Statistikaamet (2021). *Eesti tööjõu-uuringu metoodika muudatused*. URL: https://www.stat.ee/et/eesti-toojou-uuringu-metoodika-muudatused.

Statistikaamet (2024). *TT54: EESTI TÖÖJÕU-UURINGU VASTAMISMÄÄR MAAKONNA JÄRGI (KVARTALID)*. URL: https://andmed.stat.ee/et/stat/sotsiaalelu__tooturg__tooturu-uldandmed__aastastatistika/TT54.

Traat, I. ja Inno, J. (1997). *Tõenäosuslik valikuuring*. Tartu: Tartu Ülikooli Kirjastus. ISBN: 978-9985-56-226-0.

Wagner, J. R. (2008). "Adaptive Survey Design to Reduce Nonresponse Bias". Doktoritöö. University of Michigan.

Lisa 1. Abitunnuste jaotus kvartalite lõikes

Tabel 4: ETU 2023 abitunnuste väärtuste osakaalud, I ja II kvartal (esimest korda küsitletud leibkonnad)

	I		II	
	Valim	**Vastanud**	**Valim**	**Vastanud**
Sugu				
Mehed	47,4	43,3	45,8	43,1
Naised	52,6	56,7	54,2	56,9
Vanusegrupp				
15–29	15,6	12,0	17,8	14,4
30–49	34,5	32,0	34,1	32,5
50–69	30,2	34,1	30,5	32,5
70+	19,7	21,9	17,6	20,6
Linnastumise aste				
Linn	67,2	64,0	67,2	64,9
Maa	32,8	36,0	32,8	35,1

Tabel 5: ETU 2023 abitunnuste väärtuste osakaalud, III ja IV kvartal (esimest korda küsitletud leibkonnad)

	III		IV	
	Valim	**Vastanud**	**Valim**	**Vastanud**
Sugu				
Mehed	46,0	42,8	48,5	46,6
Naised	54,0	57,2	51,5	53,4
Vanusegrupp				
15–29	15,8	12,2	17,4	15,5
30–49	34,6	32,3	34,9	32,2
50–69	31,2	35,2	30,5	31,7
70+	18,4	20,3	17,2	20,6
Linnastumise aste				
Linn	67,0	63,6	67,5	65,2
Maa	33,0	36,4	32,5	34,8

Lisa 2. Tarkvara R kood

```
# Paketid -----------------------------------------------------------------
library(dplyr)
library(ggplot2)
library(openxlsx)
# Funktsioonid ------------------------------------------------------------
# Vastamistõenäosuste leidmine
leiaVtn <- function(data) {
  vtn_glm <- glm(VASTUS1 ~ SUGU + VANUSGR + MLINN,
                 family = binomial(), data = data)
  prob <- predict(vtn_glm, data, type = "response")
  return(prob)
}
# Indikaatorite arvutamine
leiaInd <- function(data) {
  DKAAL <- data$DKAAL_UUS
  N <- sum(DKAAL)
  VTN <- leiaVtn(data)
  roo_bar <- (1/N)*sum(VTN*DKAAL)
  S_p <- sqrt((1/(N-1))*sum(DKAAL*(VTN - roo_bar)**2))
  R <- 1 - 2*S_p
  return(c(roo_bar, S_p, R))
}
# Andmete sisselugemine ---------------------------------------------------
response <- read.xlsx("etu_2023_kordsus1.xlsx",colNames=T,detectDates=T)
faktorid <- c("SUGU", "VANUSGR", "MLINN")
response[faktorid] <- lapply(response[faktorid], factor)
# Andmete töötlemine ------------------------------------------------------
# Uue disainikaalu arvutamine
response <- response %>%
  group_by(KV, KIHI_NR) %>%
  mutate(KIHT_VSUURUS_uus = n_distinct(VAATLUSPERIOOD, LEIBKOND)) %>%
  ungroup() %>%
  mutate(DKAAL_UUS = KIHT_FSUURUS_VALIM / KIHT_VSUURUS_uus)
# Mittevastanutel on küsitluspäeva info puudu.
# Määrame väärtuseks 0 (vajalik indikaatorite arvutamise tsüklis)
response <- response %>%
  mutate(KYSPAEV = ifelse(is.na(KYSPAEV), 0, KYSPAEV))
# Nelja kvartali eraldamine tabeliteks
response_list <- list()
for (i in 1:4) {
  df <- response %>% filter(KV == i)
  df_name <- paste("response", i, sep="_")
  response_list[[df_name]] <- df
}
```

```r
# Andmestiku kirjeldav analüüs ----------------------------------------
n <- nrow(response)
m <- nrow(response %>% filter(VASTUS == 1))
maatriksid <- c()
for (i in 1:3) {
  matr <- as.matrix(
    addmargins(table(response %>% select(faktorid[i], VASTUS))))
  maatriksid <- rbind(maatriksid, matr)
}
# Abitunnuste tasemed valimis ja vastanute hulgas (2023)
abi_2023 <- data.frame(maatriksid, TASE = rownames(maatriksid)) %>%
  filter(TASE != "Sum" & Sum != 0) %>%
  mutate("Valim_n" = as.numeric(Sum), "Vastanud_n" = as.numeric(X1),
         "Valim_ok" = round(as.numeric(Sum)/n,3)*100,
         "Vastanud_ok" = round(as.numeric(X1)/m,3)*100) %>%
  select(TASE, "Valim_n", "Vastanud_n", "Valim_ok", "Vastanud_ok")
rownames(abi_2023) <- NULL
# Abitunnuste osakaalud kvartalite lõikes
abi_kv <- data.frame()
for (i in 1:4) {
  kvartal <- response_list[[i]]
  n <- nrow(kvartal)
  m <- nrow(kvartal %>% filter(VASTUS == 1))
  maatriksid <- c()
  for (j in 1:3) {
    matr <- as.matrix(
      addmargins(table(kvartal %>% select(faktorid[j], VASTUS))))
    maatriksid <- rbind(maatriksid, matr)
  }
  tulem <-
    data.frame(maatriksid, TASE = rownames(maatriksid), KV = i) %>%
    filter(TASE != "Sum" & Sum != 0) %>%
    mutate("Valim_ok" = round(as.numeric(Sum)/n,3)*100,
           "Vastanud_ok" = round(as.numeric(X1)/m,3)*100) %>%
    select(KV, TASE, "Valim_ok", "Vastanud_ok")
  rownames(tulem) <- NULL
  abi_kv <- rbind(abi_kv, tulem)
}
# R-indikaator üle küsitlusperioodide ---------------------------------
ind_periood <- list()
# Tsükkel üle nelja kvartali
for (i in 1:4) {
  kvartal <- response_list[[i]]
  ind <- data.frame()
  ind_name <- paste("ind", i, sep="_")
  vastamine <- kvartal %>% mutate(VASTUS1 = 0)
```

```
  max_kyspaev <- max(vastamine$KYSPAEV, na.rm = T)
  # Tsükkel üle küsitluspäevade
  for (j in 1:max_kyspaev) {
    vastamine[
      vastamine$VASTUS == 1 & vastamine$KYSPAEV <= j, "VASTUS1"] <- 1
    m <- vastamine %>% filter(VASTUS1 == 1) %>% nrow()
    tulem <- leiaInd(vastamine)
    ind <- rbind(ind, data.frame(
      KV = i, PAEV = j, M_PAEV = m, RB_PAEV = tulem[1],
      SP_PAEV = tulem[2], R_PAEV = tulem[3]))
  }
  ind_periood[[ind_name]] <- ind
  rm(ind, kvartal, vastamine)
}
# Kõigi kvartalite tulemused ühe tabelina
ind_2023_periood <- bind_rows(ind_periood)
# Esinduslikkus kvartalite viimaseks küsitluspäevaks
kv_lopp <- ind_2023_periood %>%
  group_by(KV) %>%
  filter(PAEV == max(PAEV)) %>%
  select(KV, M_PAEV, RB_PAEV, R_PAEV)
# Esinduslikkus I kvartali kolmes punktis
kp_kolm <- ind_2023_periood %>%
  filter(KV == 1 & PAEV %in% c(4, 13, 20)) %>%
  select(PAEV, M_PAEV, RB_PAEV, R_PAEV)
# Joonised ------------------------------------------------------------------
# R-indikaator
R_2023 <-
  ind_2023_periood %>%
  filter(PAEV <= 20) %>%
  ggplot(aes(x = PAEV, y = R_PAEV)) +
  geom_line(size = 0.75, aes(color = factor(KV),
                             linetype = factor(KV))) +
  scale_y_continuous(limits = c(0.5, 1.0),
                     breaks = seq(0.5, 1.0, by = 0.1)) +
  scale_x_continuous(breaks = seq(1, 20, 1)) +
  scale_linetype_manual(values = c("solid", "longdash",
                                   "solid", "longdash"),
                        labels=c("I","II","III","IV")) +
  scale_color_discrete(labels=c("I","II","III","IV")) +
  labs(x = "Küsitlusperioodi päev",
       y = "R-indikaator",
       color = "Kvartal", linetype = "Kvartal") +
  expand_limits(x = 1, y = 0) +
  coord_cartesian(expand = FALSE, clip = "off") +
  geom_vline(xintercept = c(4, 13), linetype = "dotted",
             color = "black", size = 0.5, alpha = 0.7) +
```

```
  annotate("text", x = 4, y = 0.525,
           label = expression(CAWI), size = 4,
           angle = 90, vjust = -0.75, hjust = 0) +
  annotate("text", x = 13, y = 0.525,
           label = expression(CATI["13"]), size = 4,
           angle = 90, vjust = -0.75, hjust = 0) +
  annotate("text", x = 20, y = 0.525,
           label = expression(CATI["20"]), size = 4,
           angle = 90, vjust = -0.75, hjust = 0) +
  theme_linedraw() +
  theme(legend.position = "bottom",
        legend.key.size = unit(1, "cm"))
# Vastamistõenäosuste keskmine
RB_2023 <-
  ind_2023_periood %>%
  filter(PAEV <= 20) %>%
  ggplot(aes(x = PAEV, y = RB_PAEV)) +
  geom_line(size = 0.75, aes(color = factor(KV),
                             linetype = factor(KV))) +
  scale_y_continuous(limits = c(0.0, 1.0),
                     breaks = seq(0.0, 1.0, by = 0.2)) +
  scale_x_continuous(breaks = seq(1, 20, 1)) +
  scale_linetype_manual(values = c("solid", "longdash",
                                   "solid", "longdash"),
                        labels=c("I","II","III","IV")) +
  scale_color_discrete(labels=c("I","II","III","IV")) +
  labs(x = "Küsitlusperioodi päev",
       y = "Vastamistõenäosuste keskmine",
       color = "Kvartal", linetype = "Kvartal") +
  expand_limits(x = 8, y = 0) +
  coord_cartesian(expand = FALSE, clip = "off") +
  geom_vline(xintercept = c(4, 13), linetype = "dotted",
             color = "black", size = 0.5, alpha = 0.7) +
  annotate("text", x = 4, y = 0.825,
           label = expression(CAWI), size = 4,
           angle = 90, vjust = -0.75, hjust = 0) +
  annotate("text", x = 13, y = 0.825,
           label = expression(CATI["13"]), size = 4,
           angle = 90, vjust = -0.75, hjust = 0) +
  annotate("text", x = 20, y = 0.825,
           label = expression(CATI["20"]), size = 4,
           angle = 90, vjust = -0.75, hjust = 0) +
  theme_linedraw() +
  theme(legend.position = "bottom",
        legend.key.size = unit(1, "cm"))
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Mariliis Kütt,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Adaptiivne uuringu disain", mille juhendajad on Kristi Lehto ja Mare Vähi, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Mariliis Kütt
15.05.2024