Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Hannula-Katrin Pandis

Valimi paigutamise meetodid lihtsa juhusliku kihtvaliku korral

Matemaatilise statistika eriala

Bakalaureusetöö (6 EAP)

Juhendaja Natalja Lepik

Tartu 2016

Valimi paigutamise meetodid lihtsa juhusliku kihtvaliku korral

Bakalaureusetöö

Hannula-Katrin Pandis

Käesoleva bakalaureusetöö eesmärk on uurida ning kirjeldada valimi paigutamise meetodeid lihtsa juhusliku kihtvaliku korral juhul kui üldine valimimaht on fikseeritud. Teoreetilises osas tutvustatakse tõenäosusliku valikuuringu olulisi mõisteid, tuuakse välja hinnanguteks vajalikud valemid. Töös kirjeldatakse viit erinevat valimi paigutamise meetodit kihtvaliku korral: lihtne juhuslik valik, võrdeline paigutus, Neymani paigutus, astmeline paigutus ja Costa paigutus.

Töö praktilises osas viiakse läbi simulatsioon iga valimipaigutuse jaoks, millega leitakse hinnangud nii üldkogumi keskmisele kui ka keskmistele kihtides, nende suhtelised vead ja standardvead. Saadud hinnanguid võrreldakse omavahel ja analüüsitakse, milline valimi paigutus viib kõige paremate tulemusteni.

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

Märksõnad: kihtvalik, valikuuringud, valikuteooria, juhuvalik, valimi optimaalne paigutus

Different sampling allocation methods in stratified simple random sampling

Bachelor's thesis

Hannula-Katrin Pandis

The aim of this thesis is to study and describe the sample allocation methods in case of fixed sample size and stratified simple random sampling. The theoretical part introduces the key concepts of sample survey theory, brings out formulas for estimation. Five different sample allocations are described and used in stratified sampling: simp-

le random sampling, proportional allocation, Neyman allocation, power allocation and Costa allocation.

In the practical part, the simulation is carried out for each allocation. The estimates of the population mean and strata means are found. Their relative errors and standard errors are calculated through the simulation. Obtained estimates are compared and analysed.

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

Keywords: stratified sampling, sample survey, sampling theory, random sampling, allocation of sample

Sisukord

Sissejuhatus

Käesoleva bakalaureusetöö eesmärk on uurida valimi paigutuse meetodeid lihtsa juhusliku kihtvaliku korral, kui üldine valimimaht on fikseeritud. Valikuuringus on teada, et Neymani valimipaigutus kihtidesse minimiseerib üldkogumi kogusumma hinnangu, kuid võib viia mõnes kihis väga kehva hinnanguni. Antud töös keskendutakse ühe tunnuse uurimisele, st valimipaigutused leitakse ning hinnanguid võrreldakse ühe uuritava tunnuse suhtes.

Uuritavaks tunnuseks sai valitud kehalise aktiivsuse näitaja. Ühest küljest peegeldab tunnuse valik töö autori isiklikku huvi spordi vastu ning teisest küljest on inimeste füüsiline aktiivsus alati aktuaalne teema. Kehaline aktiivsus on tervisliku eluviisi üks tähtsamaid tegureid ning see ei ole vaid eesmärgistatud treenimine ja võistlustel käimine, vaid tähendab igasugu füüsilise liikumisega seotuid tegevusi [1].

Kihtvaliku üks eesmärk on tõsta hinnangu täpsust, mis aga sõltub valimimahust igas kihis. Valimimahu ja valikudisaini väljatöötamise etapil tuleb leida ka optimaalne valimijaotus kihtide vahel. Kui valimit oskuslikult paigutada, võib suurendada hinnangu täpsust ja samal ajal vähendada uuringu maksumust. Vastavalt uuritava tunnuse jaotusele üldkogumis annavad erinevad paigutused erineva täpsusega hinnanguid.

Töö esimeses kolmes osas on antud referatiivselt ülevaade käesolevas töös kasutatavatest statistilistest näitajatest ning tutvustatakse erinevaid valimipaigutusi, mida kihtvaliku korral kasutatakse. Töö neljandas osas on kirjeldatud kasutatavat Euroopa Sotsiaaluuringu andmestikku, kihistavat tunnust ning uuritavat tunnust. Lisaks on läbi viidud simulatsioon hindamaks keskmist, suhtelist viga ning standardviga töö teoreetilises osas kirjeldatud valimipaigutuste korral. Saadud valimimahtusid ning hinnanguid on võrreldud omavahel.

Bakalaureusetöö on kirjutatud ja vormistatud programmiga LaTeX. Andmete analüüs on läbi viidud statistikaprogrammiga SAS 9.4, vastav programmi kood on toodud Lisas 2. Joonised on tehtud programmiga Microsoft Excel 2013.

Autor tänab bakalaureusetöö juhendajat Natalja Lepikut rohkete nõuannete ning paranduste eest.

1 Valikuuring

Kõikne uuring on statistiline uuring, mille korral uuritakse üldkogumi igat objekti, et saada täpset informatsiooni üldkogumi kohta. **Valikuuring** on statistiline uuring, mille korral tehakse üldkogumi kohta järeldused valimi baasil. Valikuuringul on mitmeid eeliseid kõikse uuringu ees: väiksem maksumus, paindlikkus, suurem kiirus, kergem rakendatavus. Antud peatükis on kirjeldatud erinevaid oluliseid mõisteid ja karakteristikuid valikuuringu teoorias, mida töös edaspidi kasutatakse.

Peatükk põhineb loengukonspektil [2], kui ei ole märgitud teisiti.

Üldkogum ehk populatsioon on objektide hulk (lõplik hulk), mille kohta soovitakse informatsiooni saada ning järeldusi teha vastavalt püstitatud probleemülesandele.

Osakogum on üldkogumi alamhulk, mis on fikseeritud tausttunnuse või uuritava tunnuse väärtuste järgi ja mida soovitakse eraldi uurida. Osakogumi objektid on üldkogumi objektidega sama tüüpi.

Olgu meil lõplik üldkogum $U = \{1, 2, \ldots, i, \ldots, N\}$ jagatud H mittelõikuvaks osaks, edaspidi **kihiks** [3] $U_1, \ldots, U_h, \ldots, U_H$. Olgu kihi U_h maht N_h. Tähistagu U_h kihi h kõigi objektide hulka, kusjuures

$$U = \bigcup_{h=1}^{H} U_h,\ U_h \cap U_g = \emptyset \text{ kui } g \neq h,$$

$$N_h = \sum_{U_h} 1,$$

$$N = \sum_{h=1}^{H} N_h.$$

1.1 Valikumeetodid

Järgnev alapeatükk põhineb loengukonspektil [2].

Valikuuringud jagunevad tõenäosuslikeks ja empiirilisteks. Tõenäosusliku valikumeetodi

korral on iga üldkogumi elemendi jaoks teada tema valimisse sattumise tõenäosus. Sageli on vaja teada ka kahe elemendi koos valimisse sattumise tõenäosust, kuid see ei ole alati leitav. Empiirilise valiku korral pole elemendi valimisse sattumise tõenäosus teada.

Tõenäosuslikud meetodid jagunevad kaheks:

- tagasipanekuta valik - TTA. Kui üldkogumi element osutub valituks, siis eemaldatakse ta üldkogumist, st iga element saab valimisse sattuda ainult ühe korra;
- tagasipanekuga valik - TGA. Kui üldkogumi element osutub valituks, siis ei võeta teda üldkogumist välja, st iga element võib sattuda valimisse rohkem kui üks kord.

Levinumad tõenäosuslikud valikumeetodid on lihtne juhuslik valik, Poissoni valik, süstemaatiline valik, suurusega võrdelise tõenäosusega valik, kihtvalik, klastervalik, kaheastmeline valik. Antud töös keskendutakse nn lihtsale juhuslikule kihtvalikule tagasipanekuta (LJKV TTA), mida on põhjalikumalt kirjeldatud punktis 2.3.

1.2 Valikudisaini karakteristikud ja hindamine

Järgnev alapeatükk põhineb õpikul [3], kui ei ole märgitud teisiti.

Tunnus on näitaja, mida mõõdetakse või vaadeldakse. Tunnus võib omandada erinevatel objektidel erinevaid väärtusi. Ühel objektil võib mõõta mitmeid tunnuseid.

Valim on üldkogumi osahulk, mis määratakse statistiliste meetoditega. Objektide võtmine valimisse toimub loendi abil. Edaspidine uurimine toimub valimiga. Antud töös on valimi suurus fikseeritud.

Olgu meil valim s mahuga n, $s \subset U$. Objektil $i \in s$ mõõdetakse nii mitut tunnust kui vaja, antud töös mõõdetakse objektil ühte tunnust.

Tõenäosuslik valik on valik üldkogumist, mille korral

- on võimalik defineerida kõigi võimalike valimite hulga

$$S = \{s_1, s_2, \ldots, s_M\};$$

- iga valimi $s \in S$ jaoks on teada tema valikutõenäosus $p(s)$;
- iga üldkogumi objekti valimisse sattumise tõenäosus on teada ja on positiivne;
- valimi võtmiseks kasutatav juhuslik mehhanism tagab, et valimi s valikutõenäosus on $p(s)$.

Valikudisain, mis on tõenäosusjaotus $p(s)$ kõigi antud valiku jaoks võimalike valimite hulgal S, omab valikuteoorias väga suurt tähtsust, kuna sellega on määratud kõigi hinnangute statistilised omadused. Disanile optimaalse hinnangu konstrueerimiseks ja statistiliste omaduste esitamiseks kasutatakse kaasamistõenäosusi π_i ja valikutõenäosusi p_i. Valikutõenäosusi p_i vaadeldakse TGA disainide korral ning seetõttu siin töös neid ei kasutata.

Üldkogumi objekti i ($i = 1, 2, \ldots, N$) **kaasamistõenäosuseks** π_i nimetatakse tõenäosust, millega see objekt kaasatakse valimisse antud disaini $p(s)$ korral.

Üldkogumi objekti i kaasamistõenäosust saab vaadelda ka kui

$$\pi_i = P(i \in s) = \sum_{i \in s} p(s).$$

Analoogiliselt saab vaadelda kahe üldkogumi elemendi i ja j üheaegset kaasamistõenäosust ehk **teist järku kaasamistõenäosust**

$$\pi_{ij} = P(i, j \in s) = \sum_{i,j \in s} p(s).$$

Tõenäosusliku valiku korral eeldatakse, et kõik esimest järku kaasamistõenäosused on rangelt positiivsed, mis tagab igale objektile võimaluse valimisse sattuda, $\pi_i > 0$ iga $i \in U$ jaoks.

Valikuteoorias kasutatakse valemite esitamisel sageli valikuindikaatorit I_i.

Valikuindikaator I_i on TTA disainide korral iga üldkogumi objekti i ($i = 1, 2, \ldots N$) jaoks määratud binaarne juhuslik suurus, mis iseloomustab objekti kaasamist valimisse:

$$I_i = \begin{cases} 1, & \text{kui } i \text{ kaasatakse valimisse,} \\ 0, & \text{muidu.} \end{cases}$$

Järgnevad karakteristikud on vajalikud selleks, et uurida parameetrite hinnangute omadusi [2]:

$E(I_i)$ - esimest järku moment (ehk objekti i oodatav valikute arv);

$E(I_i I_j)$ - teist järku moment;

$D(I_i) = \triangle_{ii}$ - valikuindikaatori I_i dispersioon;

$Cov(I_i, I_j) = \triangle_{ij}$ - valikuindikaatorite I_i ja I_j vaheline kovariatsioon.

Kuna tagasipanekuta disainide korral $I_i \sim Be(\pi_i)$, saame otse mõned omadused [2]:

$E(I_i) = \pi_i = P(I_i = 1)$;

$D(I_i) = \pi_i(1 - \pi_i) = \triangle_{ii}$;

$Cov(I_i, I_j) = \triangle_{ij} = \pi_{ij} - \pi_i \pi_j$.

Edaspidi kasutatakse järgmisi lühendeid:

$$\sum_U y_i = \sum_{i \in U} y_i \quad \text{ja} \quad \sum_s y_i = \sum_{i \in s} y_i.$$

1.3 Üldkogumi summa hindamine

Olgu uuritava tunnuse y väärtused üldkogumis $y_1, y_2, \ldots, y_N$. Enamus huvipakkuvatest parameetritest (keskväärtus, osakaal, suhe jne) on võimalik esitada üldkogumi summa $Y = \sum_U y_i$ kaudu. Seetõttu on valikuuringutes just selle parameetri hindamine väga

tähtis. Järgnevalt on toodud kaks teoreemi parameetri Y hindamisest, mille tõestused on leitavad õpikus [3] lk 70-71 ja lk 71-72.

Teoreem 1. Üldkogumi kogusumma $Y = \sum_U y_i$ nihketa hinnang on

$$\hat{Y} = \sum_U I_i \breve{y}_i \text{ või } \hat{Y} = \sum_U \omega_i y_i,$$

kus

$$\breve{y}_i = \frac{y_i}{E(I_i)} \text{ ja } \omega_i = \frac{I_i}{E(I_i)}.$$

Selle disainipõhine dispersioon on

$$D\hat{Y} = \sum \sum_U \triangle_{ij} \breve{y}_i \breve{y}_j,$$

kus $\triangle_{ij} = Cov(I_i, I_j)$. Dispersiooni nihketa hinnanguks $E(I_i I_j) > 0$ korral on

$$\hat{D}\hat{Y} = \sum \sum_U \breve{\triangle}_{ij} \breve{y}_i \breve{y}_j I_i I_j \text{ või } \hat{D}\hat{Y} = \sum \sum_U \breve{\triangle}_{ij} \omega_i y_i \omega_j y_j,$$

kus

$$\breve{\triangle}_{ij} = \frac{\triangle_{ij}}{E(I_i I_j)}.$$

Teoreem 1 kehtib iga valikudisaini korral, nii TTA kui ka TGA. Selleks on vaja vaid teada disainikarakteristikuid $E(I_i)$, $E(I_i I_j)$, $\triangle_{ij}$ [2].

Olgu valikudisain $p(k)$ fikseeritud valimimahuga n, st $\sum_U I_i \equiv n$.

Teoreem 2. Fikseeritud mahuga disaini $p(k)$ korral saab hinnangu $\hat{Y} = \sum_U I_i \breve{y}_i$ dispersiooni esitada alternatiivsel kujul

$$D\hat{Y} = -\frac{1}{2} \sum \sum_U \triangle_{ij} (\breve{y}_i - \breve{y}_j)^2,$$

ja eeldusel, et $E(I_i I_j) > 0 \; \forall \; i \neq j \in U$, on dispersiooni $D\hat{Y}$ nihketa hinnanguks

$$D\hat{Y} = -\frac{1}{2} \sum \sum_U I_i I_j \breve{\triangle}_{ij} (\breve{y}_i - \breve{y}_j)^2.$$

2 Kihtvalik ja seda iseloomustavad karakteristikud

Antud töös uuritakse valimi paigutamist nn lihtsa juhusliku kihtvaliku korral. Järgnevas peatükis esitatakse nii lihtsa juhusliku valiku kui ka kihtvaliku kohta vajalikke karakteristikuid.

2.1 Lihtne juhuslik valik

Järgnev alapeatükk põhineb õpikul [3], kui ei ole märgitud teisiti.

Lihtne juhuslik valik (LJV) on valikumeetod, kus kõikidel objektidel on võrdne tõenäosus valmisse sattuda, kõik valimid on samuti võrdse esinemistõenäosusega ning LJV kasutamisel saadakse valim etteantud mahuga n.

Kõigi n mahuliste hulkade arv, mida saab üldkogumist lõpliku mahuga N moodustada on $\binom{N}{n}$. Nende hulkade hulk on lihtsa juhusliku valiku kõigi võimalike valimite hulk $S = \{s_1, \ldots, s_H\}, H = \binom{N}{n}$.

Lihtsa juhusliku valiku disainiks nimetatakse diskreetset ühtlast jaotust $p(s)$ kõigi valimite hulgal S, kus

$$p(s) = 1/\binom{N}{n}, s \in S.$$

Võttes arvesse $p(s)$ kuju ja et valimite $s \in S$ arv, mis sisaldavad objekti i on $\binom{N-1}{n-1}$, saame esimest järku kaasamistõenäosuse

$$\pi_i = n/N, \forall\, i.$$

Teist järku kaasamistõenäosus:

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}, \forall\, i \neq j.$$

Lihtsa juhuvaliku kaasamistõenäosused on konstantsed, kuid objektide valik ei toimu sõltumatult, st $\triangle_{ij} \neq 0$ kui $i \neq j$.

Suhet $f = n/N$ nimetatakse valikusuhteks. See näitab, kui suur osa üldkogumist võeti valimisse. Lihtsa juhusliku valiku puhul kehtib $\pi_i = f \ \forall \ i$ korral.

Valikuindikaatori I_i dispersioon ja kovariatsioon avalduvad järgmiselt [2]:

$$\triangle_{ii} = D(I_i) = \frac{n}{N}\Big(1 - \frac{n}{N}\Big);$$

$$\triangle_{ij} = Cov(I_i, I_j) = -f(1-f)\frac{1}{N-1}.$$

Teoreem 3. Lihtsa juhuvaliku korral on hinnang $\hat{Y} = N\bar{y}$ nihketa hinnang üldkogumi kogusumma Y hindamiseks, dispersiooniga

$$D\hat{Y} = N^2(1-f)S^2_{yU}/n$$

ja dispersiooni hinnanguga

$$\hat{D}\hat{Y} = N^2(1-f)S^2_{ys}/n.$$

Siinjuures $f = n/N$ on valikusuhe,

$$S^2_{yU} = \frac{1}{N-1}\sum_U (y_i - \bar{Y})^2$$

on tunnuse y dispersioon üldkogumis ja

$$S^2_{ys} = \frac{1}{n-1}\sum_s (y_i - \bar{y})^2$$

on tunnuse y dispersioon valimis, kus

$$\bar{Y} = \frac{1}{N}\sum_U y_i$$

on üldkogumikeskmine ja

$$\bar{y} = \frac{1}{n}\sum_s y_i$$

on valimikeskmine.

Teoreemi 3 tõestus on toodud õpikus [3] lk 92-93.

2.1.1 Hinnang üldkogumi keskmisele

Käesolev alapeatükk põhineb õpikul [3].

Kuna $\bar{Y} = Y/N$, siis saab üldkogumi keskmise $\bar{Y}$ hinnangut moodustada kogusumma Y hinnangu abil. Kui kasutame Y hinnagu jaoks nihketa hinnangut ja teadaolevat üldkogumi mahtu N, saab $\bar{Y}$ hinnangu esitada järgmiselt:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{1}{N}\sum_s \frac{y_i I_i}{E(I_i)} = \frac{1}{N}\sum_s \omega_i y_i, \tag{1}$$

kusjuures annab see üldkogumi keskmise nihketa hinnangu mistahes valikudisaini korral.

Kasutades Teoreemi 1 tulemusi on näha, et üldkogumi keskmise $\bar{Y}$ hinnang $\hat{\bar{Y}}$ on nihketa, dispersiooniga

$$D\hat{\bar{Y}} = \frac{1}{N^2}\sum\sum_U \triangle_{ij}\, \breve{y}_i \breve{y}_j,$$

mille nihketa hinnanguks, kui $\pi_{ij} > 0$ iga $i, j \in U$ korral, on

$$\hat{D}\hat{\bar{Y}} = \frac{1}{N^2}\sum\sum_s \breve{\triangle}_{ij} \breve{y}_i \breve{y}_j.$$

Juhul kui üldkogumi maht N pole teada, saab kasutada nn alternatiivset hinnangut üldkogumi keskmisele. Üldjuhul annab alternatiivne hinnang täpsema (väiksema dispersiooniga) hinnangu üldkogumi keskmisele $\bar{Y}$, mistõttu eelistatakse alternatiivset hinnangut ka siis kui N on teada.

Keskmise hinnang on seega defineeritud järgmiselt:

$$\hat{\bar{Y}}_{alt} = \frac{\hat{Y}}{\hat{N}} = \frac{\sum_s y_i \omega_i}{\sum_s 1\omega_i}. \tag{2}$$

Selle abil on võimalik esitada ka alternatiivne hinnang üldkogumi summale:

$$\hat{Y}_{alt} = \frac{\hat{Y}}{\hat{N}} N.$$

Hinnangu $\hat{N}$ omadused tulenevad Teoreemist 1 erijuhul kui $y_i \equiv 1$. Sel juhul $Y = N = \sum_U 1$ annab objektide summa üldkogumis, mille nihketa hinnanguks on $\hat{N} = \sum_s \frac{1}{E(I_i)}$.

Näeme, et hinnang $\hat{Y}_{alt}$ moodustub kahe kogusumma $\hat{Y}$ ja $\hat{N}$ hinnangu suhte abil. Vastavad suurused on enamasti positiivselt korreleeritud, ehk valim mis annab suurema $\hat{Y}$ väärtuse annab ka suurema $\hat{N}$ väärtuse ning vastupidi, mille tulemusena suhte $\hat{Y}/\hat{N}$ varieeruvus valimist valimisse väheneb.

2.2 Kihtvalik

Kihtvalik on praktikas väga levinud valikudisain, mis võimaldab rakendada üldkogumi eri osades erinevaid valikudisaine. Kihtvaliku korral jaotatakse objektid üldkogumis esmalt mõne tausttunnuse väärtuse järgi osadesse, mida käsitletakse iseseisvate, üksteisest sõltumatute kogumitena ja milledes võidakse rakendada erisuguseid valikumeetodeid. Olgu kihistavaks tunnuseks näiteks sugu, siis saab jagada üldkogumi meeste ja naiste kihiks, ning kummaski saab teha sõltumatu juhusliku valiku [3].

Praktikas kasutatakse kihtvalikut järgmistel põhjustel [2],[3]:

- hinnangu täpsuse tõstmiseks - lähedaste y väärtustega tunnused tagavad valimihinnangule väikese varieeruvuse kihis, tunnuse y konstantsuse puhul saaks kihtvaliku abil konstrueerida varieeruvuseta hinnangu;
- osakogumite hindamiseks - väikeste valimite korral on kasulik osakogumite parameetritele etteantud täpsusega hinnangu leidmiseks esitada osakogumid erinevate kihtidena, kusjuures igas kihis rakendada neile sobivat optimaalset disaini (valida sobilik valimimaht või efektiivsem valikudisain);
- erinevat käsitlust vajavate kihtide hindamine - kui on teada, et mingite objekti-

de valim on tunduvalt kulukam ülejäänutest, saab vähendada kalliste objektide valimit; või kui on teada, et mingite objektide valimil on suur kaoprotsent, mida ette teades saab suurendada vastavat valimit;

- uuringu korraldamine - näiteks intervjuude puhul, kui on vaja kihte moodustada vastavalt intervjueerija keskusele, saab suunata valimi paigutust. Sellega saab ka uuringu maksumust vähendada.

Tähistades kihis h objektide arvu N_h ja valimi objektide arvu sellest kihist n_h, avaldub üldkogumi maht summana $N = \sum_{h=1}^{H} N_h$ ja kogu valimi maht summana $n = \sum_{h=1}^{H} n_h$.

Järgnev osa põhineb õpikul [3]. Arvestades üldkogumi U kihtstruktuuri, saab tunnuse y kogusumma esitada kujul

$$Y = \sum_{h=1}^{H} Y_h, \tag{3}$$

kus Y_h on kogusumma h-ndas kihis,

$$Y_h = \sum_{U_h} y_i.$$

Tunnuse y keskmine on avaldatav kujul

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N} \sum_{h=1}^{H} Y_h = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{Y}_h = \sum_{h=1}^{H} W_h \bar{Y}_h,$$

kus $\bar{Y}_h = Y_h/N_h$ on h-nda kihi keskmine ja $W_h = N_h/N$ on kihi kaal.

Kogusumma hinnang. Vastavalt seosele (3) saab kihtvaliku korral üldkogumi kogusumma hinnangu $\hat{Y}$ avaldada kihtide kogusumma hinnangute $\hat{Y}_h$ kaudu. Arvestades, et valikud toimuvad igas kihis sõltumatult, siis avaldub hinnangu $\hat{Y}$ dispersioon kihtide dispersioonide summana.

Teoreem 4. [3] Kihtvaliku korral on hinnang

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h$$

nihketa kogusumma Y jaoks, kui $E\hat{Y}_h = Y_h$. Hinnangu $\hat{Y}$ dispersioon avaldub hinnan-

gute $\hat{Y}_h$ dispersioonide summana

$$D\hat{Y} = \sum_{h=1}^{H} D\hat{Y}_h$$

ning dispersiooni hinnang

$$\hat{D}\hat{Y} = \sum_{h=1}^{H} \hat{D}\hat{Y}_h$$

on nihketa dispersiooni $D\hat{Y}$ jaoks, kui $E(D\hat{Y}_h) = \hat{D}\hat{Y}_h$.

Järeldus 1. [3] Seosest (3) lähtudes avaldub kihtvaliku korral üldkogumi keskmise hinnang kujul

$$\hat{\bar{Y}} = \sum_{h=1}^{H} W_h \hat{\bar{Y}}_h, \tag{4}$$

mis on kihikeskmiste $\hat{\bar{Y}}_h$ kaalutud keskmine. Hinnangu $\hat{\bar{Y}}$ dispersioon järeldub otseselt valemist (4)

$$D\hat{\bar{Y}} = \sum_{h=1}^{H} (W_h)^2 D\hat{\bar{Y}}_h.$$

Dispersioonihinnang on nihketa $D\hat{\bar{Y}}_h$ jaoks, kui liidetavateks on nihketa dispersioonihinnangud $\hat{D}\hat{\bar{Y}}_h$ ning avaldub järgmiselt:

$$\hat{D}\hat{\bar{Y}} = \sum_{h=1}^{H} (W_h)^2 \hat{D}\hat{\bar{Y}}_h.$$

2.3 Lihtne juhuslik kihtvalik

Lihtsa juhusliku kihtvaliku korral rakendatakse igas kihis lihtsat juhuvalikut. Tänu kihtide autonoomsusele võib neis kasutada mitmesuguse valikusuhetega lihtsaid juhuvalikuid, kus f_h on kihi valimi kaal ja on defineeritud järgmiselt:

$$f_h = \frac{n_h}{N_h}, h = 1, 2, \ldots, H.$$

Kuna lihtsa juhusliku valiku korral

$$\pi_h = f_h = const, \forall\ h \in U_h,$$

siis avaldub kogusumma nihketa hinnang h-ndas kihis kujul

$$\hat{Y}_h = \sum_{s_h} \frac{y_i}{f_h} = N_h \bar{y}_{s_h},$$

kus $\bar{y}_{s_h}$ on h-nda kihi valimikeskmine,

$$\bar{y}_{s_h} = \frac{1}{n_h} \sum_{s_h} y_i.$$

Teoreemi 3 ja teoreemi 4 põhjal saab esitada järgmise tulemuse.

Teoreem 5. [3] Lihtsa juhusliku kihtvaliku korral avaldub kogusumma $Y = \sum_U y_i$ nihketa hinnang kujul

$$\hat{Y} = \sum_{h=1}^{H} N_h \bar{y}_{s_h},$$

dispersiooniga

$$D\hat{Y} = \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{yU_h}^2 / n_h$$

ja nihketa dispersiooni hinnanguga

$$\hat{D}\hat{Y} = \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{ys_h}^2 / n_h,$$

kus $S_{yU_h}^2$ on tunnuse y dispersioon kihis U_h

$$S_{yU_h}^2 = \frac{1}{N_h - 1} \sum_{U_h} (y_i - \bar{Y}_h)^2 \tag{5}$$

ja $S_{ys_h}^2$ on tunnuse y dispersioon h-ndas valimis

$$S_{ys_h}^2 = \frac{1}{n_h - 1} \sum_{s_h} (y_i - \bar{y}_{s_h})^2.$$

3 Valimi erinevad paigutused kihtidesse

Etteantud valimimahu n korral tuleb otsustada, mitu objekti igast kihist võtta. Valimimahust n sõltub üldkogumi hinnangu dispersioon: mida rohkem objekte on valimis, seda väiksem on hinnangu varieeruvus. Samas ei saa valimimahtu n liiga suureks ajada, kuna sellega kaasneb ka uuringu maksumuse suurenemine, mis on aga tavaliselt fikseeritud etteantud summaga. Osutub, et kui valimimahtu n oskuslikult kihtidesse paigutada, saab vähendada nii hinnangu dispersiooni kui ka uuringu maksumust. Selleks on mitmeid meetodeid, mida järgnevalt kirjeldatakse.

3.1 Neymani paigutus

Neymani paigutus [3] on selline valimipaigutus, mille korral valimimahtusid h-ndas kihis leitakse järgmiselt:

$$n_h = n\frac{N_h S_{yU_h}}{\sum_{g=1}^{H} N_g S_{yU_g}}, h = 1, 2, \ldots, H. \tag{6}$$

Fikseeritud valimimahu korral minimiseerib Neymani paigutus üldkogumi summa hinnangu $\hat{Y}$ dispersiooni.

Valemist (6) on näha, et Neymani paigutuses sõltub n_h väärtus uuritava tunnuse standardhälbest S_{yU_h} kihis U_h, ehk mida rohkem varieeruvad y_i väärtused kihis U_h, seda rohkem elemente võetakse vastava kihi valimisse. Lisaks võetakse suuremast kihist valimisse enam elemente.

Neymani paigutus minimiseerib küll kogusumma hinnangu $\hat{Y}$ dispersiooni (sest on $D\bar{Y}$ suhtes optimaalne), kuid see paigutus võib viia mõnikord väga kehva hinnanguni mõnes kihis. Lisaks muudab Neymani paigutuse praktilise kasutamise raskemaks asjaolu, et igas kihis h peab teada olema uuritava tunnuse standardhälve S_{yU_h}, mis aga on üldjuhul enne uuringut teadmata [3]. Tavaliselt kasutatakse S_{yU_h} hindamiseks pilootuuringut või võetakse see samalaadsest läbiviidud uuringust.

Mõned paigutused sõltuvad uuritavast tunnusest y ja seetõttu kasutatakse tihti selle asemel abitunnust x, mille väärtused on teada terves üldkogumis (nt registritunnus, **register** on kirjete kogum erinevate üldkogumite kohta, nt rahvastikuregister) ning mis on tugevalt ja positiivselt korreleeritud uuritava tunnusega.

3.2 Võrdeline paigutus

Võrdeline paigutus [3] on selline valimipaigutus, kus vastavate kihtide osakaalud valimis ja üldkogumis on võrdsed ehk

$$n_h = n\frac{N_h}{N}, h = 1, 2, \ldots, H. \tag{7}$$

Näitame, et võrdeline paigutus (7) on optimaalne ehk langeb kokku valemiga (6), kui dispersioonid S_{yU_h} kõikides kihtides on võrdsed. Lõpptulemus ei sõltu uuritavast tunnusest y.

Olgu $S_{yU_h} \equiv S$ iga $h = 1, 2, \ldots, H$ korral. Siis lihtsustub Neymani paigutus järgmiselt:

$$n_h = n\frac{N_h S_{yU_h}}{\sum_{g=1}^{H} N_g S_{yU_h}} = n\frac{SN_h}{S\sum_{g=1}^{H} N_g} = n\frac{N_h}{\sum_{g=1}^{H} N_g}, h = 1, 2, \ldots, H.$$

Kuna $\sum_{g=1}^{H} N_g = N$, siis saamegi võrdelise paigutuse (7).

Mitme tunnuse uurimise korral võib osutuda Neymani paigutus kehvaks mõnede tunnuste jaoks, kuna ta on optimaalne vaid ühe tunnuse suhtes. Võrdeline paigutus aga uuritavast tunnusest ei sõltu. Võrdeline paigutus (7) annab siiski üsna hea hinnangu $\hat{Y}_h$, kuid ei pruugi viia aksepteeritava dispersioonini kogusumma $\hat{Y}$ jaoks.

3.3 Astmeline paigutus

Järgnev alapeatükk põhineb artiklil [4], kui ei ole märgitud teisiti.

Astmelist paigutust võib vaadelda kui kompromissi Neymani paigutuse ja võrdse paigu-

tuse ($n_h = n/H,\ h = 1, 2, \ldots, H$) vahel, st üritab tagada võimalikult väikest dispersiooni üldkogumi summa hinnangule, samal ajal kasutades võimalikult häid hinnanguid ka kihtides. Eespool kirjeldatud paigutused ei pruugi anda soovitud hinnanguid kui üldkogumi kihid erinevad üksteisest väga palju, näiteks suuruses või tähtsuses [5].

Olgu $C_h = S_{yU_h}/\bar{Y}_h$ kihi U_h variatsioonikordaja. **Astmeline paigutus** on defineeritud järgmiselt:

$$n_h = n\frac{C_h X_h^q}{\sum_{h=1}^{H} C_h X_h^q},\ h = 1, 2, \ldots, H, \tag{8}$$

kus X_h on kihi h mingisugune suuruse või tähtsuse mõõt ja aste q on nn tuunimise parameeter, kusjuures $0 \leq q \leq 1$.

Paneme tähele, et kui $q = 1$ ja $X_h = N_h\bar{Y}_h = Y_h$, siis

$$n_h = n\frac{S_{yU_h}N_h\bar{Y}_h}{\bar{Y}_h\sum_{g=1}^{H}\frac{S_{yU_g}}{\bar{Y}_g}N_g\bar{Y}_g} = n\frac{N_h S_{yU_h}}{\sum_{g=1}^{H}N_g S_{yU_g}}$$

ja astmeline paigutus langeb kokku Neymani paigutusega (6).

Seejuures kui valida $q = 0$, $X_h = N_h\bar{Y}_h = Y_h$ ja võtta, et $C_h \equiv C$ iga $h = 1, 2, \ldots, H$ korral, siis

$$n_h = n\frac{CN_h\bar{Y}_h}{\sum_{g=1}^{H}CN_g\bar{Y}_g} = n\frac{N_h\frac{Y_h}{N_h}}{\sum_{g=1}^{H}N_g\frac{Y_g}{N_g}} = n\frac{Y_h}{Y},$$

mis on võrdne võrdelise paigutusega juhul kui uuritav tunnus on lihtsalt $(N \times 1)$ vektor elementidega 1.

Valides astmelises paigutuses (8) $q = 1/2$, $X_h = N_h$ ja $C_h \equiv C$, iga $h = 1, 2, \ldots, H$ korral, saame nn **ruutjuure paigutuse**:

$$n_h = n\frac{C\sqrt{N_h}}{\sum_h C\sqrt{N_h}} = n\frac{\sqrt{N_h}}{\sum_h\sqrt{N_h}},\ h = 1, 2, \ldots, H. \tag{9}$$

Astmelist paigutust on kasutatud näiteks Kanada Statistikaametis tulumaksu deklaratsioonide valikuuringus [5]. Kanada provintsid varieeruvad väga palju oma pindalalt, rikkuselt, elanikkonna arvu poolest.

3.4 Costa paigutus

Costa paigutus [4] on defineeritud järgmiselt:

$$n_h = k(nW_h) + (1-k)(n/H),\ h = 1, 2, \ldots, H, \tag{10}$$

kus k on mingi konstant $0 \leq k \leq 1$, $W_h = N_h/N$ on kihi h osakaal, n on valimimaht ning H on kihtide arv üldkogumis U.

Paneme tähele, et Costa paigutus viib võrdse valimimahuni kihtides juhul kui $k = 0$:

$$n_h = 0 \cdot (nW_h) + (1-0)(n/H) = n/H,\ h = 1, 2, \ldots, H$$

ja võrdeliseks paigutuseks juhul kui $k = 1$:

$$n_h = 1 \cdot (nW_h) + (1-1)(n/H) = nW_h = n\frac{N_h}{N},\ h = 1, 2, \ldots, H.$$

Juhul kui mõne kihi U_h korral on $n/H > N_h$ ehk kihi valimisse peab sattuma rohkem objekte kui on kihi maht N_h, siis tuleb Costa paigutust muuta. Olgu A kõikide selliste kihtide kogum. Sel juhul uuendatud Costa paigutus on järgmine:

$$n_h = k(nW_h) + (1-k)n_h^0,$$

kus $n_h^0 = N_h$ kui $h \in A$ ja $n_h^0 = (n - \sum_{h \in A} N_h)/(H - m)$ mujal. Muutuja m on kihtide arv kogumis A.

4 Simuleerimisülesanne

Antud peatükis on võrrdeldud erinevaid paigutusi (lihtne juhuslik valik, võrdeline paigutus, Neymani paigutus, astmeline paigutus, Costa paigutus) keskmiste hinnangute täpsuse suhtes. Keskmisi on vaadeldud nii terves üldkogumis kui ka igas kihis eraldi. Valimite moodustamiseks ja hinnangute leidmiseks kasutati statistikapaketti SAS ning vastav kood on toodud Lisas 2.

Hinnangute leidmiseks kasutati iga paigutuse korral simuleerimise meetodit. Igas kihis võeti 1000 valimit kasutades LJV TTA disaini. Edasi uuriti keskmiste hinnanguid, suhtelist viga ja standardhälbe hinnanguid üle simulatsioonide Monte-Carlo meetodil. Lisaks kihtvalikule on uuritud ka tavalise lihtsa juhuvaliku tulemusi, kus valik toimub tervest üldkogumist korraga ilma kihte arvestamata.

Lihtsa juhusliku valiku korral kasutati keskmise hindamiseks valemit (1).

Kuna LJV TTA korral langevad N ja $\bar{N}$ kokku

$$\bar{N} = \sum_s \omega_i 1 = \sum_s \frac{N}{n} = \frac{N}{n} \sum_s 1 = \frac{N}{n} n = N,$$

siis ei vaadeldud alternatiivset hinnangut (2) eraldi.

Võrdelise, Neymani, astmelise, nn ruutjuure ja Costa paigutuste hinnang keskmisele, suhtelisele veale ja standardhälbele leiti kasutades Monte-Carlo meetodit. Järgmised näitajad illustreerivad hinnanguid üle M simulatsiooni. Antud töös on võetud $M = 1000$.

Töös kasutati järgmiseid Monte-Carlo hinnanguid:

- hinnang üldkogumi keskmisele:

$$\hat{\bar{Y}}_{MC} = \frac{1}{M} \sum_{m=1}^{M} \hat{\bar{Y}}^{(m)}, \tag{11}$$

kus $\hat{\bar{Y}}^{(m)}$ on hinnang üldkogumi keskmisele (4) sammul m, $m = 1, 2, \ldots M$;

- hinnangu $\hat{\hat{Y}}^{(m)}$ standardviga, mis on statistiku standardhälbe hinnang:

$$\sqrt{D_{MC}(\hat{\hat{Y}}_{MC})} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\hat{Y}}^{(m)} - \hat{\hat{Y}}_{MC})^2}; \tag{12}$$

- hinnangu $\hat{\hat{Y}}_{MC}$ suhteline viga (CV), mis näitab kui suure osa moodustab hinnangu standardviga hinnangust endast:

$$CV_{MC}(\hat{\hat{Y}}_{MC}) = \frac{\sqrt{D_{MC}(\hat{\hat{Y}}_{MC})}}{\hat{\hat{Y}}_{MC}}. \tag{13}$$

4.1 Andmestiku üldkirjeldus

Töös kasutati Euroopa sotsiaaluuringu (*ESS*) andmestikku 2014. aasta uuringust. *ESS* on korraldanud uuringuid rohkem kui 30 Euroopa riigis, seejuures esimene uuring korraldati aastal 2002 ning järgnevad uuringud on tehtud iga kahe aasta tagant [6]. Tänaseks on kokku korraldatud seitse uuringut.

Uuringul on kolm peaeesmärki: jälgida ja tõlgendada avaliku elu arvamuse muutumist ja väärtuseid Euroopas, edendada ja tugevdada meetodeid riikidevaheliste uuringute mõõtmisteks Euroopas, töötada välja erinevaid Euroopa sotsiaalseid näitajaid [6].

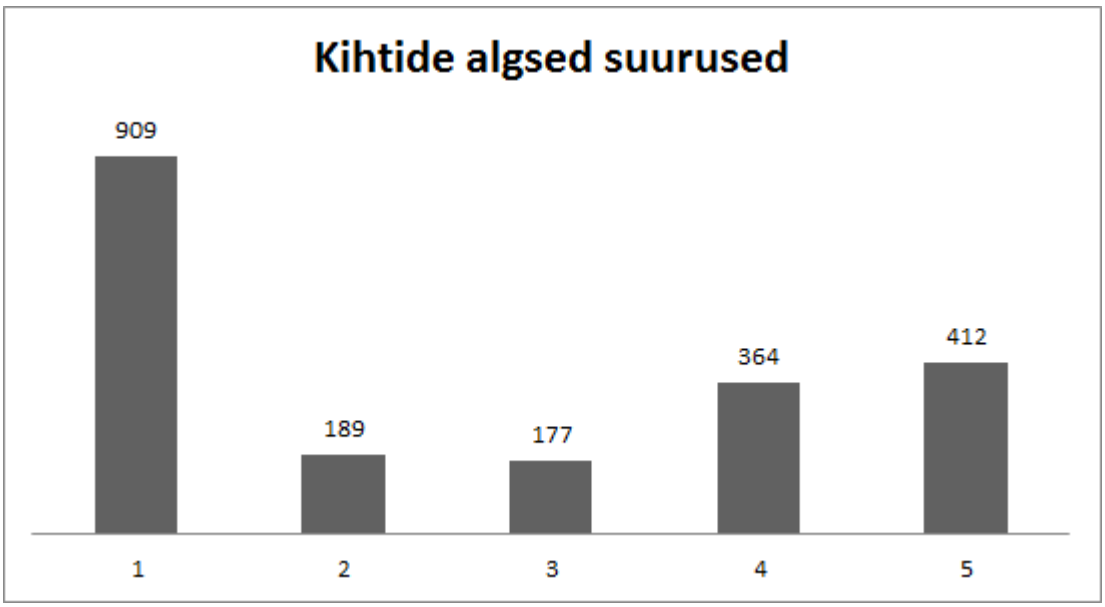
2014. aastal oli vastajaid 22-st Euroopa riigist. Uuringus kasutati juhuslikku valimit, vastanute osakaaluga valimis ehk vastamismääraga 70%. Uuring korraldati intervjuuvormis, Eestis toimusid intervjuud ajavahemikul 07.09.14 - 29.12.14 nii eesti kui ka vene keeles.

Antud töös tehti suur kitsendus andmetele - nimelt kasutati vaid Eesti elanikkonna vastuseid. Eestis oli plaanitud küsitleda 3620 inimest, kusjuures ühest leibkonnast küsitleti vaid ühte isikut. Kasutusele sai lõpuks võtta 2051 vastust, mis on ka kõik antud töö algses üldkogumis esindatud [7].

Andmestikus oli kokku 2051 andmerida ja 571 tunnust, millest antud töös on kasutatud kahte tunnust:

- vastaja elukoha piirkond (tunnus "region2", millel on viis erinevat väärtust);
- spordi või muu füüsilise tegevuse harrastamine viimase seitsme päeva jooksul (tunnus "dosprt").

Andmestikku kihistades saadi viis kihti: Põhja-Eesti (kodeeritud 1), Lääne-Eesti (2), Kesk-Eesti (3), Kirde-Eesti (4) ja Lõuna-Eesti (5).



Joonis 1: Kihtide suurused

Algselt oli üldkogumis kokku 2051 tulemust, mis jaotusid viieks kihiks nagu on näha Jooniselt 1.

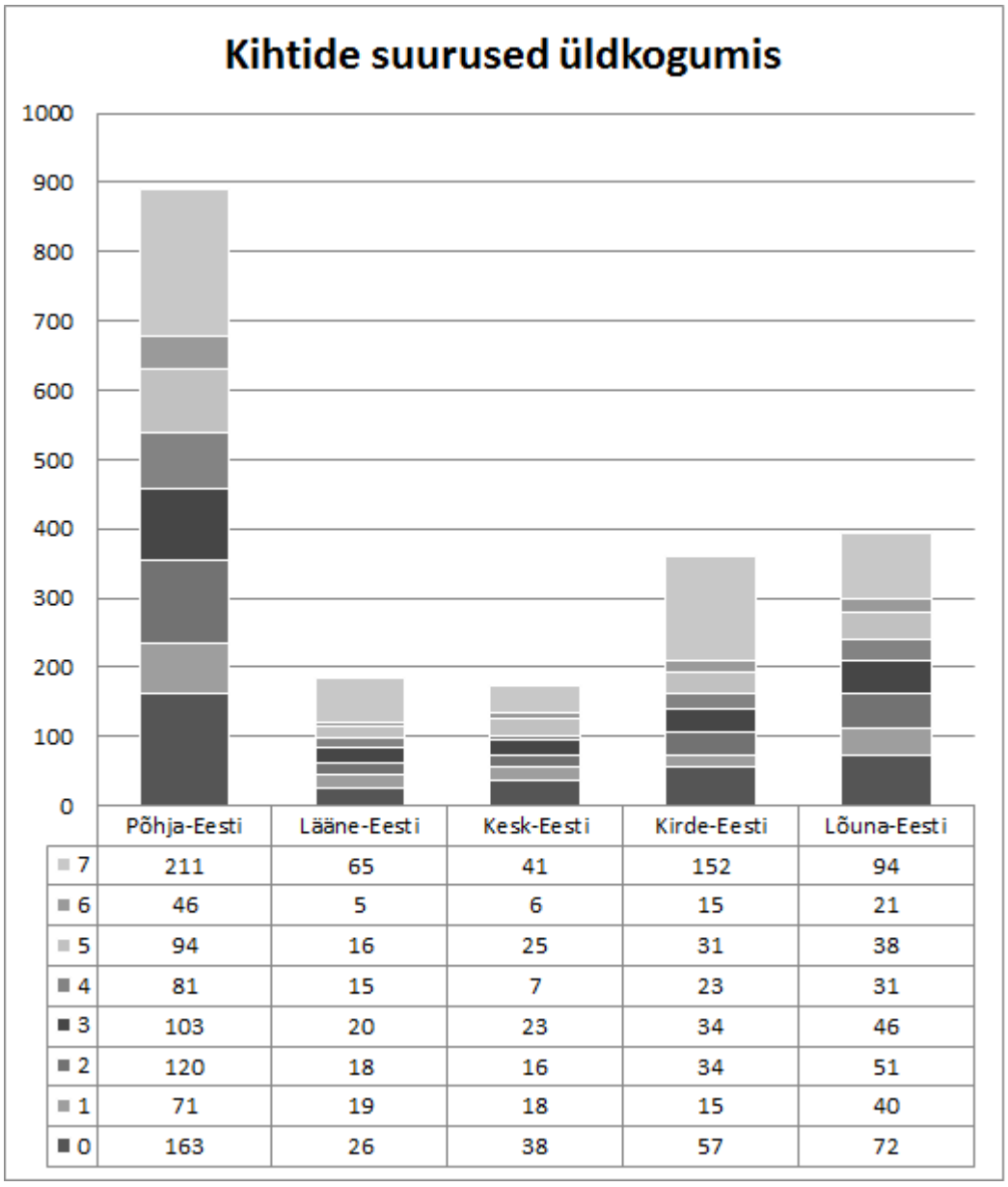
Uuritava tunnuse, milleks oli kehaline aktiivsus, väärtusteks olid täisarvud nullist seitsmeni, mis näitasid vastavalt mitmel erineval päeval sporditi või liigutati aktiivselt. Uuringut korraldades oli defineeritud kehaline aktiivsus järgmiselt: "kiire jalutus, spordi või muu kehalise aktiivsuse tegemine vähemalt 30 minutit päevas, kusjuures liikumine ei pea olema toimunud ühe korraga"[6].

Osutus et 50 uuritava tunnuse väärtust on kodeeritud arvuga 88, mis tähendas vastusevarianti "ei tea". Need vastuseread eemaldati andmestikust, mistõttu lõplikku andmestikku jäi järele 97,56% kogu Eestist vastanute andmestikust ehk 2001 andmerida.

4.2 Lõpliku andmestiku kihtide kirjeldus

Analüüsis võeti kasutusele andmestik 2001 andmereaga, mida edaspidi käsitletakse üldkogumina.

Kõige suurem kiht oli Põhja-Eesti (kiht nr 1), mahuga 889, mis moodustas 44,43% üldkogumist. Suuruselt teiseks kihiks oli Lõuna-Eesti (kiht nr 5), vastavalt mahuga 393 moodustades 19,64% üldkogumist. Kirde-Eesti (kiht nr 4) oli vaid natuke väiksem kui Lõuna-Eesti ehk mahuga 361 ja 18,04% üldkogumist. Lääne-Eesti (kiht nr 2) ja Kesk-Eesti (kiht nr 3) olid juba tunduvalt väiksemad, vastavalt 184 ja 174 ning 9,2% ja 8,7% üldkogumist. Kihtide suuruste erinevus on küllaltki suur, ligikaudu viiekordne erinevus suurima ja väikseima kihi vahel.

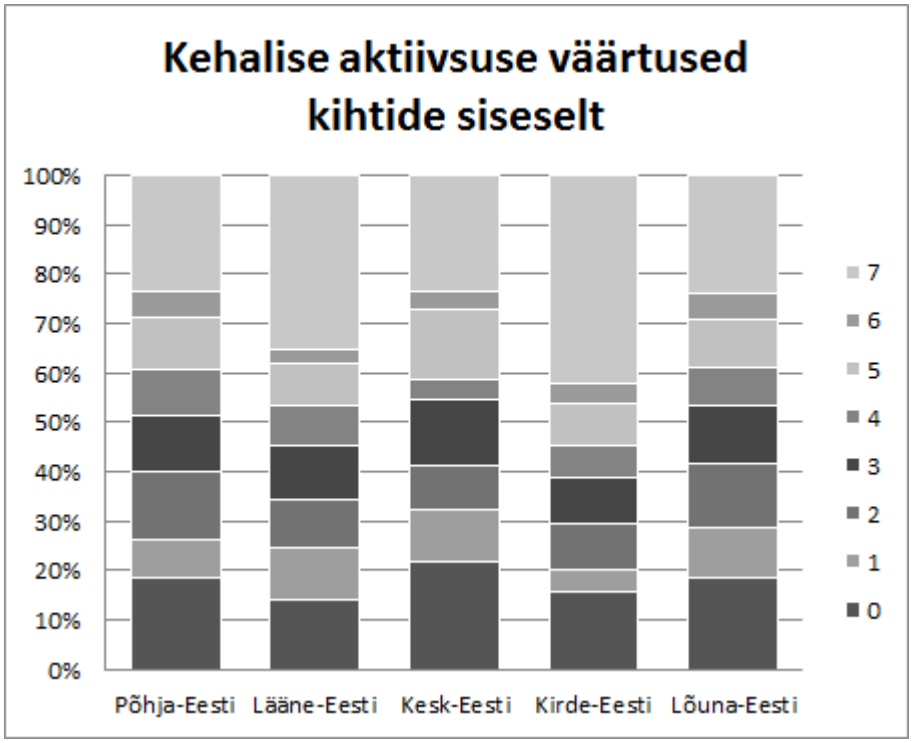


	Põhja-Eesti	Lääne-Eesti	Kesk-Eesti	Kirde-Eesti	Lõuna-Eesti
7	211	65	41	152	94
6	46	5	6	15	21
5	94	16	25	31	38
4	81	15	7	23	31
3	103	20	23	34	46
2	120	18	16	34	51
1	71	19	18	15	40
0	163	26	38	57	72

Joonis 2: Kihtide suurused, osadeks jaotatud uuritava tunnuse väärtuste järgi

Jooniselt 2 on näha, millised olid kihid ning Jooniselt 3 on näha, kuidas on jaotunud uuritava tunnuse väärtused kihtide siseselt. Kuigi oleks võinud arvata, et suurim hulk vastuseid füüsilise aktiivsuse harrastamise kohta tulevad keskmiste väärtuste seast, st kehaliselt aktiivne oldi kolmel-neljal päeval nädalas, siis antud andmetest tuleb välja, et kõige enam oldi aktiivsed iga päev, täpsemalt 563 vastust ehk 28,14% kogu tulemustest. Suuruselt teine tulemus oli mitte ühelgi päeval füüsilise aktiivsuse harrastamine viimase

seitsme päeva jooksul, so 356 vastust ehk 17,79% üldkogumist. See näitab seda, et kõige äärmuslikumad vastusevariandid (0 ja 7) on kihtides suurima osakaaluga.



Joonis 3: Kehalise aktiivsuse väärtused kihtide siseselt

Joonisel 3 on võetud iga kihi maht kui 100% ning uuritava tunnuse erinevad väärtused selgelt välja toodud. On näha, kuidas jaotuvad kehalisuse aktiivsuse erinevad väärtused protsentuaalselt kihtide kaupa.

Tabelites kasutatakse edaspidi kihtide nimesid järgmiselt: Põhja-Eesti ehk Põhja, Lääne-Eesti ehk Lääne, Kesk-Eesti ehk Kesk, Kirde-Eesti ehk Kirde, Lõuna-Eesti ehk Lõuna.

Tabel 1: Sportimise või muu füüsilise tegevuse keskmised kihiti

	Põhja	Lääne	Kesk	Kirde	Lõuna	üldkogum
keskmine sportimine	3,562	4,022	3,420	4,393	3,506	3,731

Edaspidi keskenduti uuritava tunnuse keskmise hinnangu, suhtelise vea ning standard-

vea uurimisele. Selleks leiti alustuseks ka iga kihi tegelik keskmine füüsilise aktiivsuse harrastamise kohta, et edaspidi võrrelda hinnangutega. Täpsed tulemused on toodud Tabelis 1. Kasutatud andmetele ja kihtidele toetudes spordivad või teevad muud füüsilist tegevust keskmiselt kõige tihedamini Kirde-Eesti inimesed, neile järgnevad Lääne-Eesti elanikud. Kõige vähem on füüsiliselt aktiivsed Kesk-Eesti inimesed.

4.3 Valimi erinevad paigutused

Käesolevas töös rakendati erinevaid valimi paigutusi kihtvaliku korral ning leiti sportimise või muu füüsilise tegevuse keskmise hinnangud, suhtelised vead ning standardvead. Kasutati lihtsa juhusliku kihtvaliku meetodit fikseeritud valimimahu korral. Valimimahuks valiti 30% üldkogumist ehk 600. Lisaks kihtvalikule viidi läbi eksperiment, kus korduvaid valimeid võeti lihtsa juhuvaliku abil tervest üldkogumist. Sellega sooviti kontrollida, kas kihtvalik annab võrreldes tavalise lihtsa juhusliku valikuga paremaid tulemusi või mitte.

Valimimahtude arvutamiseks kihtides on rakendatud järgnevaid meetodeid:

- **Neymani paigutus** (6) peab andma optimaalse paigutuse üldkogumi keskmisele, kuid võib viia mõnes kihis väga kehva hinnanguni. Neymani paigutus sõltub ka uuritava tunnuse standardhälbest, milleks antud töös võeti igas kihis kogu kihi väärtuste põhjal leitud S_{yU_h} (5) ning kasutati saadud väärtusi. Kasutatud standardhälbed on välja toodud Lisas 1 Tabelis 5.
- **Võrdelise paigutuse** (7) korral võetakse igast kihist valimimaht nii, et vastavate kihtide osakaalud on valimis ja üldkogumis võrdsed.
- **Astmelises paigutuses** (8) sõltub kihi h valimimaht vastava kihi U_h variatsioonikordajast C_h ja kihi tähtsuse või suuruse mõõdust, mis on omakorda astendatud nn tuunimise parameetriga q, X_h^q. Tuunimise parameeter q võib muutuda vahemikus $0 \leq q \leq 1$.
  - Esmalt valiti tuunimise parameetri väärtuseks $q = 0,5$. See andis nn **ruutjuure paigutuse** (9). Kasutati eeldust, et variatsioonikordaja C_h on konstantne igas kihis h ning kihi tähtsuse mõõt võeti $X_h = N_h$.

– Astmelise paigutuse puhul genereeriti ka teine valimimahu n paigutus kihtidesse, kus anti muutujatele erinevaid väärtuseid, et tekiks võrdlusmoment ruutjuure paigutusega.

Variatsioonikordaja võeti $C_h = S_{yU_h}/\bar{Y}_h$, kus $\bar{Y}_h = Y_h/N_h$. Selle jaoks kasutatud standardhälbed ja kogusummad ning arvutatud variatsioonikordajad on välja toodud Lisas 1 Tabelis 5.

Kihi tähtsuse mõõt X_h valiti $X_h = N_h\bar{Y}_h = Y_h$. Tuunimise parameetri q väärtus võeti 0,75.

Saadi valem

$$n_h = n\frac{S_{yU_h}(Y_h)^{0,75}}{\bar{Y}_h \sum_{h=1}^{H} \frac{S_{yU_h}}{\bar{Y}_h}(Y_h)^{0,75}}$$

ning muutujate vastavaid väärtusi rakendati valemisse. Tulemusena saadi astmelise paigutuse valimimahud kihtide jaoks, kui $q = 0,75$.

- **Costa paigutuses** (10) valiti konstandi k väärtuseks 0,5.

  Kuna ühegi kihi U_h korral ei olnud juhust kus $n/H > N_h$, ehk kihi h maht N_h oli iga kihi korral suurem kui kihi valimimaht, siis sai rakendada muutmata Costa paigutust (10).

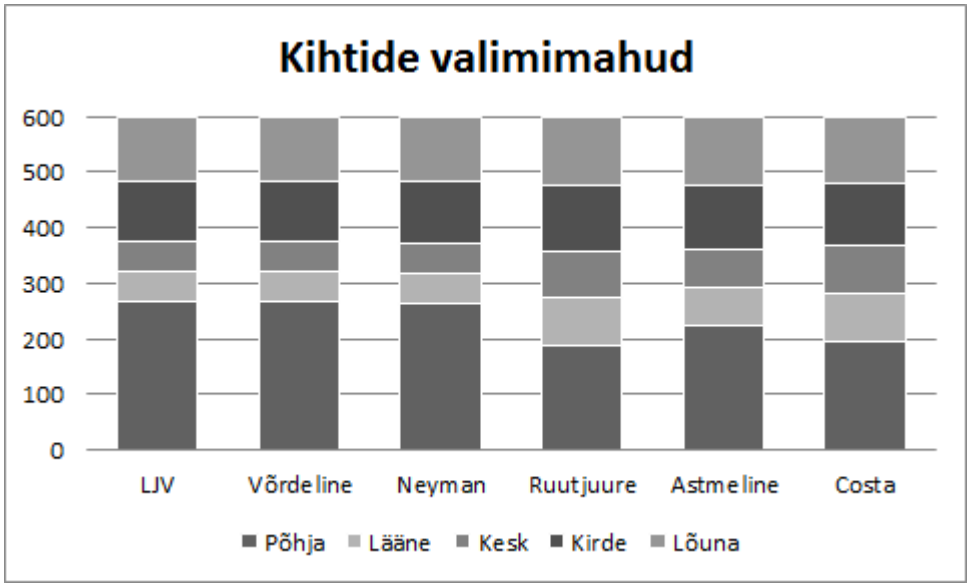
4.3.1 Valimimahud

Rakendades kirjeldatud paigutuste valemeid saadi iga paigutuse jaoks eraldi valimimahud kihtides. Kõik valimimahud on toodud välja Tabelis 2, kusjuures üldine valimimaht oli fikseeritud, $n = 600$ ehk 30% üldkogumist.

Tabel 2: Valimimahud kirjeldatud paigutuste korral

paigutus	Põhja	Lääne	Kesk	Kirde	Lõuna
LJV (reas on keskmised)	266,388	55,293	52,168	108,475	117,676
Neyman	262	56	53	112	117
Võrdeline	267	55	52	108	118
"Ruutjuur", $q = 0.5$	188	85	83	119	125
Astmeline, $q = 0.75$	224	69	69	114	124
Costa	193	88	86	114	119

Näeme, et võrdelise paigutuse tulemused langevad praktiliselt kokku LJV keskmiste tulemustega, mis on ka loomulik. Siiski peab arvestama, et igal konkreetsel simuleerimise sammul on võrdeline paigutus sama, kuid LJV paigutus varieerub, mille tõttu võib juhtuda realisatsioon väga ebasoodsa paigutuse variandiga.



Joonis 4: Kihtide mahud paigutuste kaupa

Tabelist 2 ja Jooniselt 4 on näha, kuidas fikseeritud valimimaht 600 jaotus kihtide vahel. Valimimahtusid kihtides uurides jääb veel silma, et ka Neymani paigutuse korral tulid kihtide suurused küllaltki sarnased võrdelise paigutusega. Analoogiliselt tulid omavahel sarnaste kihtide suurustega nn ruutjuure ja Costa paigutused. Teistest täiesti eristuv

oli astmeline paigutus juhul kui $q = 0,75$.

4.4 Tulemused

Igal sammul (kokku 1000) arvutati hinnangud nii üldkogumi keskmisele kui ka keskmistele kihtides. Hinnangute kokku võtmiseks ja kirjeldamiseks kasutati valemeid (11), (12) ja (13). Tulemused on koondatud Tabelisse 3 ja Tabelisse 4. Tegelikud parameetrid on kättesaadavad Tabelis 1.

Tabel 3: Sportimise või muu füüsilise tegevuse hinnangud kihiti

meetod	kiht	keskmise hinnang	suhteline viga	standardviga
LJV	Põhja	3,563	5,20	0,187
	Lääne	4,004	12,95	0,521
	Kesk	3,419	14,06	0,481
	Kirde	4,403	8,70	0,383
	Lõuna	3,502	9,17	0,321
	kogu valim	3,732	2,35	0,08786
Neyman	Põhja	3,561	3,54	0,134
	Lääne	3,997	7,45	0,298
	Kesk	3,405	8,95	0,305
	Kirde	4,397	4,80	0,212
	Lõuna	3,509	5,76	0,195
	kogu valim	3,728	2,42	0,09039
võrdeline	Põhja	3,560	3,71	0,130
	Lääne	4,019	7,65	0,308
	Kesk	3,439	8,71	0,300
	Kirde	4,397	4,93	0,217
	Lõuna	3,498	5,71	0,200
	kogu valim	3,731	2,34	0,08712

Tabel 4: Sportimise või muu füüsilise tegevuse hinnangud kihiti

meetod	kiht	keskmise hinnang	suhteline viga	standardviga
ruutjuure, $q = 0.5$	Põhja	3,556	4,76	0,170
	Lääne	4,012	5,19	0,207
	Kesk	3,428	6,18	0,212
	Kirde	4,404	4,80	0,212
	Lõuna	3,506	5,46	0,192
	kogu valim	3,730	2,53	0,09429
astmeline, $q = 0.75$	Põhja	3,566	4,15	0,148
	Lääne	4,024	6,16	0,247
	Kesk	3,421	7,03	0,241
	Kirde	4,395	4,78	0,210
	Lõuna	3,517	5,29	0,187
	kogu valim	3,736	2,35	0,08758
Costa	Põhja	3,567	4,44	0,158
	Lääne	4,022	5,14	0,207
	Kesk	3,416	6,07	0,207
	Kirde	4,390	4,60	0,202
	Lõuna	3,503	5,74	0,200
	kogu valim	3,733	2,41	0,08994

Näeme, et kõikide valimipaigutuste korral jäid hinnangud üldkogumi keskmisele üpriski lähedale tegelikele parameetritele, kõige suurem erinevus oli astmelisel paigutusel.

Kihis nr 1 ehk Põhja-Eesti kihis, mis oli kõige suurem kiht, tulid kõikide kihtide keskmiste hinnangud väga lähedased kihi tegelikule keskmisele 3,562, kusjuures võrdelise paigutuse korral oli keskmise hinnangu väärtus sama. Kihis nr 2 ehk Lääne-Eesti kihis oli kihi tegelik keskmine 4,022 ning Neymani paigutuse hinnang 3,997 oli kõige kehvem. Sama oli ka järgmises kihis, kihis nr 3, kus tulid kehvad hinnangud nii Neymanil kui ka võrdelisel paigutusel. Kihi tegelik keskmine oli 3,420, Neymani paigutuse korral vastavalt 3,405 ning võrdelise paigutuse korral 3,439. Kihis nr 4 ehk Kirde-Eesti ja nr 5 ehk Lõuna-Eesti ei tulnud ühegi valimipaigutuse korral väga kehva hinnangut. Kirde-Eesti

kihi puhul, kus tegelik keskmine oli 4,393, tuli kõige parem hinnang astmelisel paigutusel vastavalt 4,395 ning Lõuna-Eesti kihi puhul, kus tegelik keskmine oli 3,506, tuli tegeliku väärtusega võrdne hinnang nn ruutjuure paigutuse korral.

Hinnangu standardviga valimile oli väikseim võrdelisel paigutusel, väärtusega 0,08712. Kõige kehvem standardviga oli nn ruutjuure paigutusel, vastavalt 0,09429.

Kihis nr 1 tulid kõikidel paigutustel väikesed standardvead, selle põhjuseks võib olla kihi suurus, mis oli uuritud kihtidest suurim. Kihis nr 2 olid kõige parema standardveaga nn ruutjuure paigutus ja Costa paigutus, mõlema väärtused 0,207. Ka kihis nr 3 oli Costa paigutuse korral standardveaks 0,207, nn ruutjuure paigutuse korral vastavalt 0,212. Kahes väikseimas kihis, nr 2 ja nr 3, olid LJV standardvead väga kehvad, samuti jäid võrdelise ja Neymani hinnangute väärtused 0,298-0,308 vahele, mis on samuti kehvad. Kihtides nr 4 ja nr 5 tulid kõikide paigutuste korral (va LJV) standardvead päris head, jäädes 0,187-0,217 vahele.

Kõige stabiilsemalt väikeseid standardvigu kihtides oli Costa ja nn ruutjuure paigutuse korral, kuid kummalgi juhul ei tulnud üle valimi leitud standardvead kõige paremad. Ühe väikseima standardvea üle valimi saanud LJV korral tulid kihisisesed standardvead väga halvad. See näitab, et antud andmestikule tuginedes saab väita, et LJV hinnang sobib valimi hindamiseks, kus valimimaht on suur. Mida väiksemaks muutus kihi maht, seda kehvemaks muutus hinnang.

Kõige stabiilsemalt head keskmise hinnangud ja standardvead tulid Costa valimipaigutuse korral. Kogu valimi korral oli standardviga 0,08994 ning kihiti eraldi jäid kõik standardvead väiksemaks või võrdseks väärtusega 0,207.

Teooria kohaselt oleks pidanud Neymani paigutuse korral tulema väikseim standardviga, kuid antud andmetega ja fikseeritud valimimahu n korral tuli väikseim standardviga hoopis võrdelise paigutuse korral. Erinevus ei ole väga suur, standardvead vastavalt 0,09039 Neymani paigutusel ning 0,08712 võrdelisel paigutusel. Selline tulemus võis tekkida kuna kõikides kihtides tulid dispersioonid, mis on välja toodud Lisas 1 Tabelis 5, väga lähedaste väärtustega ning valimimahtusid leides ümardati vastavalt ümardamisreeglitele valimimahtude väärtused kihtides täisarvudeks.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli tutvuda ning kirjeldada valimi paigutamise meetodeid lihtsa juhusliku kihtvaliku korral. Kirjeldati kihtvaliku statistikuid, erinevaid valimi paigutamise meetodeid. Valimi paigutamist prooviti Euroopa Sotsiaaluuringu korraldatud iga kahe aasta taguse uuringu 2014. aasta tulemuste andmestikul. Kihid moodustati vastaja elukoha regiooni järgi, saadi viis kihti. Uuritavaks tunnuseks valiti kehalise aktiivsuse näitaja, seejuures valimimahud leiti ja hinnanguid võrreldi ühe tunnuse suhtes.

Kihid olid oma suuruse poolest erinevad. Leiti iga paigutuse jaoks kihtide mahud, kusjuures üldine valimimaht oli fikseeritud, $n = 600$. Simuleeriti 1000 valimit iga valimi paigutamise disaini jaoks ning analüüsiti saadud hinnanguid keskmistele ja standardhälvetele valimis ning iga kihi kohta eraldi. Saadi kuus erinevat hindamise võimalust: lihtne juhuslik valik, võrdeline paigutus, Neymani paigutus, astmeline paigutus (tuunimise parameetriga $q = 0,5$, millest tekkis nn ruutjuure paigutus, ja $q = 0,75$) ning Costa paigutus (konstandiga $k = 0,5$).

Parimaid tulemusi üldiselt andis kasutatud andmestikule tuginedes Costa paigutus: standardviga üle kogu valimi oli hea ning kihiti olid standardvead stabiilselt head. Samuti olid keskmiste hinnangud väga lähedased kihtide tegelike keskmistega.

Astmelise paigutuse korral tuli väike standardviga valimile, täpsemalt 0,08758. Sarnaselt Costa paigutusele olid kihtide kaupa hinnangu standardvead head, jäädes Costast kehvemaks ainult väikeste valimimahtudega kihtide korral. Hinnangud keskmistele olid samuti üsna head, välja arvatud Lõuna-Eesti kihis.

Neymani ehk optimaalse paigutuse korral tuli standardhälbe hinnang pigem kehvapoolne ning kihtidesiseselt oli nii häid kui ka halbu hinnanguid. Üllatusena tuli kõige väiksem standardviga ning kõige täpsem keskmise hinnang valimile võrdelisel paigutusel.

Antud tööst saaks edasi uurida astmelises paigutuses erinevate tuunimisparameetrite q väärtustega tulenevaid paigutusi ning Costa paigutuse puhul muutes konstandi k väärtust ning leida, milliste väärtuste puhul tulevad kõige täpsemad hinnangud.

Kasutatud kirjandus

[1] Kehaline aktiivsus. Wikipedia, URL (vaadatud 14.05.2016)

https://et.wikipedia.org/wiki/Kehaline_aktiivsus

[2] Lepik, N.; Traat, I. (2013). E-kursuse "Valikuuringute teooria I" materjalid. Tartu Ülikool. Loengukonspekt, URL (vaadatud 17.05.2016)

http://dspace.ut.ee/bitstream/handle/10062/30680/ValiuurI.pdf?sequence=1&isAllowed=y

[3] Traat, I.; Inno, J. (1997). Tõenäosuslik valikuuring. Tartu Ülikool. Matemaatilise statistika instituut.

[4] Choudhry, G. H.; Rao, J. N. K.; Hidiroglou, M. A. (2012). On sample allocation for efficient domain estimation. Statistics Canada. Business Survey Methods Division. Catalogue No. 12-001-X. Vol 38, No. 1, pp. 23-24.

http://www.statcan.gc.ca/pub/12-001-x/2012001/article/11682-eng.pdf

[5] Bankier, M. D. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. The American Statistician, Vol 42, No. 3, pp. 174-177.

http://www.jstor.org/stable/pdf/2684995.pdf?_=1463560465046

[6] Euroopa sotsiaaluuringu info. URL (vaadatud 14.05.2016)

Pealeht -> ESS7-2014, ed.1.0 -> Metadata -> Study Description

Pealeht -> ESS7-2014, ed.1.0 -> Variable Description -> Health inequalities -> Do sports or other physical activity, how many of last 7 days

http://nesstar.ess.nsd.uib.no/webview/

[7] Euroopa sotsiaaluuringu vastamise andmed. European Social Survey, URL (vaadatud 18.05.2016)

http://www.europeansocialsurvey.org/essdoc/doc.html?ddi=2.3.3.1&year=2014&land=233

Lisad

Lisa 1 - Kihtide dispersioonid, standardhälbed, kogusummad ja variatsioonikordajad

Tabel 5: Kihtide dispersioonid, standardhälbed, kogusummad ja variatsioonikordajad

	dispersioon	standardhälve	kogusumma	variatsioonikordaja
Põhja	6,573	2,564	3167	0,7197
Lääne	7,103	2,665	740	0,6626
Kesk	7,031	2,652	595	0,7755
Kirde	7,278	2,698	1586	0,6141
Lõuna	6,720	2,592	1378	0,7392

Lisa 2 - SASi kood

```
/*Andmetabeli sisselugemine*/
proc import out=baka.andmed
 datafile="C:\Users\Teele\Desktop\HANNULA\EestiLyhendatud2"
 DBMS = XLSX;
run;
proc sort data=baka.andmed;
by region;
run;
proc freq data=baka.andmed;
tables region;
run;

/*Kihistava tunnuse "region" kodeerimine*/
data baka.andmed; set baka.andmed;
if region="EE001" then region=1;
if region="EE004" then region=2;
if region="EE006" then region=3;
if region="EE007" then region=4;
if region="EE008" then region=5;
run;

/*Muudan "region"->"region2" character->numeric*/
data baka.andmed2;
set baka.andmed;
region2 = input(region,?? best12.);
drop region;
run;

/*Kustutan uuritava tunnuse dosprt vastused "ei tea" ehk 88*/
data baka.andmed2;
set baka.andmed2;
if dosprt=88 then delete;
run;
```

```
/*Leian kihtide suurused*/
proc sql;
create table baka.kihtidesum
as select region2, count(region2) as _total_ from baka.andmed2
group by region2;
quit;

/*Jagan yldkogumi kihtideks ja leian kihtidele keskmise,
suhtelise vea, dispersiooni, kogusumma*/
proc sql;
create table baka.pohja
as select * from baka.andmed2
where region2 = 1;
quit;
proc means data=baka.pohja mean cv var sum;
var dosprt; /*dosprt on uuritav tunnus*/
run;
proc sql;
create table baka.laane
as select * from baka.andmed2
where region2 = 2;
quit;
proc means data=baka.laane mean cv var sum;
var dosprt;
run;
proc sql;
create table baka.kesk
as select * from baka.andmed2
where region2 = 3;
quit;
proc means data=baka.kesk mean cv var sum;
var dosprt;
run;
```

```
proc sql;
create table baka.kirde
as select * from baka.andmed2
where region2 = 4;
quit;
proc means data=baka.kirde mean cv var sum;
var dosprt;
run;
proc sql;
create table baka.louna
as select * from baka.andmed2
where region2 = 5;
quit;
proc means data=baka.louna mean cv var sum;
var dosprt;
run;

/*Leian v6rdelise paigutuse valimimahud kihtides,
30% igast kihist*/
proc sql;
create table baka.LJKV_vordeline
as select region2, _total_*0.3 as valimimaht
from baka.kihtidesum;
quit;
/*LJKV v6rdeliste valimimahtudega kihtides*/
proc surveyselect data=baka.andmed2
rep=1000 /*Simulatsioon 1000 korda*/
method=srs /*meetod on LJV*/
n=(267 55 52 108 118) /*kihtide suurused*/
out=baka.LJKV;
strata region2; /*kihi tunnus*/
run;
/*Hinnangud LJKV v6rdeliste valimimahtudega, hinnangud kihtides*/
/*Pohja-Eesti piirkond ehk region2 = 1*/
```

```
data baka.pe;
set baka.LJKV;
if region2=1 then output;
run;
proc sql;
create table baka.pe_hinnang as
select sum(dosprt*SamplingWeight)/889 as hinnang_sport1
from baka.pe
group by Replicate; /*grupeerida simulatsioonide kaupa*/
run;
quit;
proc means data=baka.pe_hinnang mean cv var;
var hinnang_sport1;
run;
/*Laane-Eesti piirkond ehk region2 = 2*/
data baka.lae;
set baka.LJKV;
if region2=2 then output;
run;
proc sql;
create table baka.lae_hinnang as
select sum(dosprt*SamplingWeight)/184 as hinnang_sport2
from baka.lae
group by Replicate;
run;
quit;
proc means data=baka.lae_hinnang mean cv var;
var hinnang_sport2;
run;
/*Kesk-Eesti piirkond ehk region2 = 3*/
data baka.kee;
set baka.LJKV;
if region2=3 then output;
run;
```

```
proc sql;
create table baka.kee_hinnang as
select sum(dosprt*SamplingWeight)/174 as hinnang_sport3
from baka.kee
group by Replicate;
run;
quit;
proc means data=baka.kee_hinnang mean cv var;
var hinnang_sport3;
run;
/*Kirde-Eesti piirkond ehk region2 = 4*/
data baka.kie;
set baka.LJKV;
if region2=4 then output;
run;
proc sql;
create table baka.kie_hinnang as
select sum(dosprt*SamplingWeight)/361 as hinnang_sport4
from baka.kie
group by Replicate;
run;
quit;
proc means data=baka.kie_hinnang mean cv var;
var hinnang_sport4;
run;
/*Louna-Eesti piirkond ehk region2 = 5*/
data baka.loe;
set baka.LJKV;
if region2=5 then output;
run;
proc sql;
create table baka.loe_hinnang as
select sum(dosprt*SamplingWeight)/393 as hinnang_sport5
from baka.loe
```

```
group by Replicate;
run;
quit;
proc means data=baka.loe_hinnang mean cv var;
var hinnang_sport5;
run;
/*V6rdelise paigutuse valimi hinnangud*/
data baka.vordeline_kokku;
merge baka.pe_hinnang baka.lae_hinnang
baka.kee_hinnang baka.kie_hinnang baka.loe_hinnang;
run;
proc sql;
create table baka.vordeline_yldkogum as
select (hinnang_sport1*889+hinnang_sport2*184+hinnang_sport3*
174+hinnang_sport4*361+hinnang_sport5*393) / 2001 as hinnang
from baka.vordeline_kokku;
run;
proc means data=baka.vordeline_yldkogum mean cv var;
var hinnang;
run;

/*LJV valimimaht 30%*/
proc surveyselect data=baka.andmed2
method=srs
sampsize=600
reps=1000
out = baka.ljkv_valim;
run;
/*Lisan yldkogumile kaalu*/
data baka.ljkv_valim;
set baka.ljkv_valim;
kaal=3.335;
run;
/*Pohja-Eesti piirkond ehk region2 = 1*/
```

```
data baka.ljkv_valim2;
set baka.ljkv_valim;
if region2 = 1 then r1 = 1;
else r1 = 0;
run;
proc sql;
create table baka.ljkv_valim_hinnang as
select kaal*sum(dosprt*r1) / 889 as hinnang1
from baka.ljkv_valim2
group by Replicate;
run;
proc means data=baka.ljkv_valim_hinnang mean cv var;
var hinnang1;
run;
/*Laane-Eesti piirkond ehk region2 = 2*/
data baka.ljkv_valim3;
set baka.ljkv_valim;
if region2 = 2 then r1 = 1;
else r1 = 0;
run;
proc sql;
create table baka.ljkv_valim_hinnang2 as
select kaal*sum(dosprt*r1) / 184 as hinnang1
from baka.ljkv_valim3
group by Replicate;
run;
proc means data=baka.ljkv_valim_hinnang2 mean cv var;
var hinnang1;
run;
/*Kesk-Eesti piirkond ehk region2 = 3*/
data baka.ljkv_valim4;
set baka.ljkv_valim;
if region2 = 3 then r1 = 1;
else r1 = 0;
```

```
run;
proc sql;
create table baka.ljkv_valim_hinnang3 as
select kaal*sum(dosprt*r1) / 174 as hinnang1
from baka.ljkv_valim4
group by Replicate;
run;
quit;
proc means data=baka.ljkv_valim_hinnang3 mean cv var;
var hinnang1;
run;
/*Kirde-Eesti piirkond ehk region2 = 4*/
data baka.ljkv_valim5;
set baka.ljkv_valim;
if region2 = 4 then r1 = 1;
else r1 = 0;
run;
proc sql;
create table baka.ljkv_valim_hinnang4 as
select kaal*sum(dosprt*r1) / 361 as hinnang1
from baka.ljkv_valim5
group by Replicate;
run;
proc means data=baka.ljkv_valim_hinnang4 mean cv var;
var hinnang1;
run;
/*Louna-Eesti piirkond ehk region2 = 5*/
data baka.ljkv_valim6;
set baka.ljkv_valim;
if region2 = 5 then r1 = 1;
else r1 = 0;
run;
proc sql;
create table baka.ljkv_valim_hinnang5 as
```

```
select kaal*sum(dosprt*r1) / 393 as hinnang1
from baka.ljkv_valim6
group by Replicate;
run;
quit;
proc means data=baka.ljkv_valim_hinnang5 mean cv var;
var hinnang1;
run;
/*Kogu valimi hinnang*/
data baka.ljkv_valim_kogujaoks;
set baka.ljkv_valim;
arv=1;
run;
proc sql;
create table baka.ljkv_kokku as
select kaal*sum(dosprt)/2001 as hinnang
from baka.ljkv_valim_kogujaoks
group by Replicate;
run;
proc means data=baka.ljkv_kokku mean cv var;
var hinnang;
run;

/*Neymani valimipaigutuse leidmine*/
proc sql;
create table baka.neyman_maht as select
600*889*sqrt(6.573)/(889*sqrt(6.573)+184*sqrt(7.103)+174*sqrt(7.031)
+361*sqrt(7.278)+393*sqrt(6.720)) as kiht1,
600*184*sqrt(7.103)/(889*sqrt(6.573)+184*sqrt(7.103)+174*sqrt(7.031)
+361*sqrt(7.278)+393*sqrt(6.720)) as kiht2,
600*174*sqrt(7.031)/(889*sqrt(6.573)+184*sqrt(7.103)+174*sqrt(7.031)
+361*sqrt(7.278)+393*sqrt(6.720)) as kiht3,
600*361*sqrt(7.278)/(889*sqrt(6.573)+184*sqrt(7.103)+174*sqrt(7.031)
+361*sqrt(7.278)+393*sqrt(6.720)) as kiht4,
```

```
600*393*sqrt(6.720)/(889*sqrt(6.573)+184*sqrt(7.103)+174*sqrt(7.031)
+361*sqrt(7.278)+393*sqrt(6.720)) as kiht5
from baka.andmed2;
quit;
/*Neymani paigutus*/
proc surveyselect data=baka.andmed2
reps=1000
method=srs
n=(262 56 53 112 117)
out=baka.neyman;
strata region2;
run;
/*Pohja-Eesti piirkond ehk region2 = 1*/
data baka.pe_n;
set baka.neyman;
if region2=1 then output;
run;
proc sql;
create table baka.pe_n_hinnang as
select sum(dosprt*SamplingWeight)/889 as hinnang_sport1
from baka.pe_n
group by Replicate;
run;
quit;
proc means data=baka.pe_n_hinnang mean cv var;
var hinnang_sport1;
run;
/*Laane-Eesti piirkond ehk region2 = 2*/
data baka.lae_n;
set baka.neyman;
if region2=2 then output;
run;
proc sql;
create table baka.lae_n_hinnang as
```

```
select sum(dosprt*SamplingWeight)/184 as hinnang_sport2
from baka.lae_n
group by Replicate;
run;
quit;
proc means data=baka.lae_n_hinnang mean cv var;
var hinnang_sport2;
run;
/*Kesk-Eesti piirkond ehk region2 = 3*/
data baka.kee_n;
set baka.neyman;
if region2=3 then output;
run;
proc sql;
create table baka.kee_n_hinnang as
select sum(dosprt*SamplingWeight)/174 as hinnang_sport3
from baka.kee_n
group by Replicate;
run;
quit;
proc means data=baka.kee_n_hinnang mean cv var;
var hinnang_sport3;
run;
/*Kirde-Eesti piirkond ehk region2 = 4*/
data baka.kie_n;
set baka.neyman;
if region2=4 then output;
run;
proc sql;
create table baka.kie_n_hinnang as
select sum(dosprt*SamplingWeight)/361 as hinnang_sport4
from baka.kie_n
group by Replicate;
run;
```

```
quit;
proc means data=baka.kie_n_hinnang mean cv var;
var hinnang_sport4;
run;
/*Louna-Eesti piirkond ehk region2 = 5*/
data baka.loe_n;
set baka.neyman;
if region2=5 then output;
run;
proc sql;
create table baka.loe_n_hinnang as
select sum(dosprt*SamplingWeight)/393 as hinnang_sport5
from baka.loe_n
group by Replicate;
run;
quit;
proc means data=baka.loe_n_hinnang mean cv var;
var hinnang_sport5;
run;
/*Neymani valimipaigutuse korral hinnangud valimile*/
data baka.neyman_kokku;
merge baka.pe_n_hinnang baka.lae_n_hinnang
baka.kee_n_hinnang baka.kie_n_hinnang baka.loe_n_hinnang;
run;
proc sql;
create table baka.n_yldkogum as
select (hinnang_sport1*889+hinnang_sport2*184+hinnang_sport3*
174+hinnang_sport4*361+hinnang_sport5*393) / 2001 as hinnang
from baka.neyman_kokku;
run;
proc means data=baka.n_yldkogum mean cv var;
var hinnang;
run;
```

```
/*Nn ruutjuure paigutuse kihtide mahud -- q = 0.5*/
proc sql;
create table baka.astm1_maht as select /*q=0.5*/
600*sqrt(889)/(sqrt(889)+sqrt(184)+sqrt(174)+sqrt(361)+sqrt(393))
as kiht1,
600*sqrt(184)/(sqrt(889)+sqrt(184)+sqrt(174)+sqrt(361)+sqrt(393))
as kiht2,
600*sqrt(174)/(sqrt(889)+sqrt(184)+sqrt(174)+sqrt(361)+sqrt(393))
as kiht3,
600*sqrt(361)/(sqrt(889)+sqrt(184)+sqrt(174)+sqrt(361)+sqrt(393))
as kiht4,
600*sqrt(393)/(sqrt(889)+sqrt(184)+sqrt(174)+sqrt(361)+sqrt(393))
as kiht5
from baka.andmed2;
quit;
/*Nn ruutjuure paigutus*/
proc surveyselect data=baka.andmed2
rep=1000
method=srs
n=(188 85 83 119 125)
out=baka.astm1;
strata region2;
run;
/*Pohja-Eesti piirkond ehk region2 = 1*/
data baka.pe_astm1;
set baka.astm1;
if region2=1 then output;
run;
proc sql;
create table baka.pe_astm1_hinnang as
select sum(dosprt*SamplingWeight)/889 as hinnang_sport1
from baka.pe_astm1
group by Replicate;
run;
```

```
quit;
proc means data=baka.pe_astm1_hinnang mean cv var;
var hinnang_sport1;
run;
/*Laane-Eesti piirkond ehk region2 = 2*/
data baka.lae_astm1;
set baka.astm1;
if region2=2 then output;
run;
proc sql;
create table baka.lae_astm1_hinnang as
select sum(dosprt*SamplingWeight)/184 as hinnang_sport2
from baka.lae_astm1
group by Replicate;
run;
quit;
proc means data=baka.lae_astm1_hinnang mean cv var;
var hinnang_sport2;
run;
/*Kesk-Eesti piirkond ehk region2 = 3*/
data baka.kee_astm1;
set baka.astm1;
if region2=3 then output;
run;
proc sql;
create table baka.kee_astm1_hinnang as
select sum(dosprt*SamplingWeight)/174 as hinnang_sport3
from baka.kee_astm1
group by Replicate;
run;
quit;
proc means data=baka.kee_astm1_hinnang mean cv var;
var hinnang_sport3;
run;
```

```
/*Kirde-Eesti piirkond ehk region2 = 4*/
data baka.kie_astm1;
set baka.astm1;
if region2=4 then output;
run;
proc sql;
create table baka.kie_astm1_hinnang as
select sum(dosprt*SamplingWeight)/361 as hinnang_sport4
from baka.kie_astm1
group by Replicate;
run;
quit;
proc means data=baka.kie_astm1_hinnang mean cv var;
var hinnang_sport4;
run;
/*Louna-Eesti piirkond ehk region2 = 5*/
data baka.loe_astm1;
set baka.astm1;
if region2=5 then output;
run;
proc sql;
create table baka.loe_astm1_hinnang as
select sum(dosprt*SamplingWeight)/393 as hinnang_sport5
from baka.loe_astm1
group by Replicate;
run;
quit;
proc means data=baka.loe_astm1_hinnang mean cv var;
var hinnang_sport5;
run;
/*Valimi hinnagud ruutjuure paigutuse korral*/
data baka.astm1_kokku;
merge baka.pe_astm1_hinnang baka.lae_astm1_hinnang
baka.kee_astm1_hinnang baka.kie_astm1_hinnang baka.loe_astm1_hinnang;
```

```
run;
proc sql;
create table baka.astm1_yldkogum as
select (hinnang_sport1*889+hinnang_sport2*184+hinnang_sport3*
174+hinnang_sport4*361+hinnang_sport5*393) / 2001 as hinnang
from baka.astm1_kokku;
run;
proc means data=baka.astm1_yldkogum mean cv var;
var hinnang;
run;

/*Astmelise paigutuse kihtide mahud -- q = 0.75*/
proc sql;
create table baka.astm2_maht as select /*q=0.75*/
600*(sqrt(6.573)/(3167/889))*3167**0.75/
((sqrt(6.573)/(3167/889))*3167**0.75+
(sqrt(7.103)/(740/184))*740**0.75+
(sqrt(7.031)/(595/174))*595**0.75+
(sqrt(7.278)/(1586/361))*1586**0.75+
(sqrt(6.720)/(1378/393))*1378**0.75) as kiht1,
600*(sqrt(7.103)/(740/184))*740**0.75/
((sqrt(6.573)/(3167/889))*3167**0.75+
(sqrt(7.103)/(740/184))*740**0.75+
(sqrt(7.031)/(595/174))*595**0.75+
(sqrt(7.278)/(1586/361))*1586**0.75+
(sqrt(6.720)/(1378/393))*1378**0.75) as kiht2,
600*(sqrt(7.031)/(595/174))*595**0.75/
((sqrt(6.573)/(3167/889))*3167**0.75+
(sqrt(7.103)/(740/184))*740**0.75+
(sqrt(7.031)/(595/174))*595**0.75+
(sqrt(7.278)/(1586/361))*1586**0.75+
(sqrt(6.720)/(1378/393))*1378**0.75) as kiht3,
600*(sqrt(7.278)/(1586/361))*1586**0.75/
((sqrt(6.573)/(3167/889))*3167**0.75+
```

```
(sqrt(7.103)/(740/184))*740**0.75+
(sqrt(7.031)/(595/174))*595**0.75+
(sqrt(7.278)/(1586/361))*1586**0.75+
(sqrt(6.720)/(1378/393))*1378**0.75) as kiht4,
600*(sqrt(6.720)/(1378/393))*1378**0.75/
((sqrt(6.573)/(3167/889))*3167**0.75+
(sqrt(7.103)/(740/184))*740**0.75+
(sqrt(7.031)/(595/174))*595**0.75+
(sqrt(7.278)/(1586/361))*1586**0.75+
(sqrt(6.720)/(1378/393))*1378**0.75) as kiht5
from baka.andmed2;
quit;
/*Astmeline paigutus*/
proc surveyselect data=baka.andmed2
rep=1000
method=srs
n=(224 69 69 114 124)
out=baka.astm2;
strata region2;
run;
/*Pohja-Eesti piirkond ehk region2 = 1*/
data baka.pe_astm2;
set baka.astm2;
if region2=1 then output;
run;
proc sql;
create table baka.pe_astm2_hinnang as
select sum(dosprt*SamplingWeight)/889 as hinnang_sport1
from baka.pe_astm2
group by Replicate;
run;
quit;
proc means data=baka.pe_astm2_hinnang mean cv var;
var hinnang_sport1;
```

```
run;
/*Laane-Eesti piirkond ehk region2 = 2*/
data baka.lae_astm2;
set baka.astm2;
if region2=2 then output;
run;
proc sql;
create table baka.lae_astm2_hinnang as
select sum(dosprt*SamplingWeight)/184 as hinnang_sport2
from baka.lae_astm2
group by Replicate;
run;
quit;
proc means data=baka.lae_astm2_hinnang mean cv var;
var hinnang_sport2;
run;
/*Kesk-Eesti piirkond ehk region2 = 3*/
data baka.kee_astm2;
set baka.astm2;
if region2=3 then output;
run;
proc sql;
create table baka.kee_astm2_hinnang as
select sum(dosprt*SamplingWeight)/174 as hinnang_sport3
from baka.kee_astm2
group by Replicate;
run;
quit;
proc means data=baka.kee_astm2_hinnang mean cv var;
var hinnang_sport3;
run;
/*Kirde-Eesti piirkond ehk region2 = 4*/
data baka.kie_astm2;
set baka.astm2;
```

```
if region2=4 then output;
run;
proc sql;
create table baka.kie_astm2_hinnang as
select sum(dosprt*SamplingWeight)/361 as hinnang_sport4
from baka.kie_astm2
group by Replicate;
run;
quit;
proc means data=baka.kie_astm2_hinnang mean cv var;
var hinnang_sport4;
run;
/*Louna-Eesti piirkond ehk region2 = 5*/
data baka.loe_astm2;
set baka.astm2;
if region2=5 then output;
run;
proc sql;
create table baka.loe_astm2_hinnang as
select sum(dosprt*SamplingWeight)/393 as hinnang_sport5
from baka.loe_astm2
group by Replicate;
run;
quit;
proc means data=baka.loe_astm2_hinnang mean cv var;
var hinnang_sport5;
run;
/*Astmelise paigutuse korral hinnangud valimile*/
data baka.astm2_kokku;
merge baka.pe_astm2_hinnang baka.lae_astm2_hinnang
baka.kee_astm2_hinnang baka.kie_astm2_hinnang baka.loe_astm2_hinnang;
run;
proc sql;
create table baka.astm2_yldkogum as
```

```
select (hinnang_sport1*889+hinnang_sport2*184+hinnang_sport3*
174+hinnang_sport4*361+hinnang_sport5*393) / 2001 as hinnang
from baka.astm2_kokku;
run;
proc means data=baka.astm2_yldkogum mean cv var;
var hinnang;
run;

/*Costa paigutuse kihtide mahud*/
proc sql;
create table baka.costa_maht as select
0.5*(600*889/2001)+(1-0.5)*(600/5) as kiht1,
0.5*(600*184/2001)+(1-0.5)*(600/5) as kiht2,
0.5*(600*174/2001)+(1-0.5)*(600/5) as kiht3,
0.5*(600*361/2001)+(1-0.5)*(600/5) as kiht4,
0.5*(600*393/2001)+(1-0.5)*(600/5) as kiht5
from baka.andmed2;
quit;
/*Costa paigutus*/
proc surveyselect data=baka.andmed2
rep=1000
method=srs
n=(193 88 86 114 119)
out=baka.costa;
strata region2;
run;
/*Pohja-Eesti piirkond ehk region2 = 1*/
data baka.pe_costa;
set baka.costa;
if region2=1 then output;
run;
proc sql;
create table baka.pe_costa_hinnang as
select sum(dosprt*SamplingWeight)/889 as hinnang_sport1
```

```
from baka.pe_costa
group by Replicate;
run;
quit;
proc means data=baka.pe_costa_hinnang mean cv var;
var hinnang_sport1;
run;
/*Laane-Eesti piirkond ehk region2 = 2*/
data baka.lae_costa;
set baka.costa;
if region2=2 then output;
run;
proc sql;
create table baka.lae_costa_hinnang as
select sum(dosprt*SamplingWeight)/184 as hinnang_sport2
from baka.lae_costa
group by Replicate;
run;
quit;
proc means data=baka.lae_costa_hinnang mean cv var;
var hinnang_sport2;
run;
/*Kesk-Eesti piirkond ehk region2 = 3*/
data baka.kee_costa;
set baka.costa;
if region2=3 then output;
run;
proc sql;
create table baka.kee_costa_hinnang as
select sum(dosprt*SamplingWeight)/174 as hinnang_sport3
from baka.kee_costa
group by Replicate;
run;
quit;
```

```
proc means data=baka.kee_costa_hinnang mean cv var;
var hinnang_sport3;
run;
/*Kirde-Eesti piirkond ehk region2 = 4*/
data baka.kie_costa;
set baka.costa;
if region2=4 then output;
run;
proc sql;
create table baka.kie_costa_hinnang as
select sum(dosprt*SamplingWeight)/361 as hinnang_sport4
from baka.kie_costa
group by Replicate;
run;
quit;
proc means data=baka.kie_costa_hinnang mean cv var;
var hinnang_sport4;
run;
/*Louna-Eesti piirkond ehk region2 = 5*/
data baka.loe_costa;
set baka.costa;
if region2=5 then output;
run;
proc sql;
create table baka.loe_costa_hinnang as
select sum(dosprt*SamplingWeight)/393 as hinnang_sport5
from baka.loe_costa
group by Replicate;
run;
quit;
proc means data=baka.loe_costa_hinnang mean cv var;
var hinnang_sport5;
run;
/*Costa paigutuse korral hinnangud valimile*/
```

```
data baka.costa_kokku;
merge baka.pe_costa_hinnang baka.lae_costa_hinnang
baka.kee_costa_hinnang baka.kie_costa_hinnang baka.loe_costa_hinnang;
run;
proc sql;
create table baka.costa_yldkogum as
select (hinnang_sport1*889+hinnang_sport2*184+hinnang_sport3*
174+hinnang_sport4*361+hinnang_sport5*393) / 2001 as hinnang
from baka.costa_kokku;
run;
proc means data=baka.costa_yldkogum mean cv var;
var hinnang;
run;
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Hannula-Katrin Pandis

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Valimi paigutamise meetodid lihtsa juhusliku kihtvaliku korral", mille juhendaja on Natalja Lepik,

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

   (b) üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguste 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, **20.05.2016**