TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Kaidi Jõgi

Tervishoiutöötajate keskmise tunnipalga hindamine süstemaatilise klastervaliku ja lihtsa juhusliku kihtvaliku korral

Bakalaureusetöö (6EAP)

Juhendaja:

Natalja Lepik

TARTU 2014

Tervishoiutöötajate keskmise tunnipalga hindamine süstemaatilise klastervaliku ja lihtsa juhusliku kihtvaliku korral

Käesoleva töö eesmärgiks on võrrelda kahte meetodit valikuuringu teostamiseks tervishoiutöötajate põhitunnipalga arvutamiseks. Andmestiku moodustavad Tervise Arengu Instituudi poolt kogutud aruande „Tervishoiutöötajate tunnipalk“ 2013. aasta andmed. Vaadeldavad disainid peaksid tagama rotatsiooni 70%. Meetodeid võrreldi simulatsiooni põhjal. Hinnangute täpsuse tõstmiseks rakendati ka regressioonhinnangut.

Esimeseks disainiks valiti lihtne juhuslik kihtvalik (LJKV), mille puhul võrreldi hinnanguid võrdelise ja Neymani paigutuse korral. Rotatsiooni arvestamiseks kasutati püsijuhuarvude meetodit. Teiseks disainiks konstrueeriti uus süstemaatilisele klaster-valikule baseeruv disain.

Parimaks osutus LJKV Neymani paigutusega. Regressioonhinnang parandas kõige rohkem süstemaatilisele klastervalikule konstrueeritud disaini.
Märksõnad: tervishoiutöötaja, keskmine palk, juhuväljavõtt, süsteemväljavõtt

Estimating the Basic Hourly Wages of Health Workers in Case of Systematic Cluster Sampling and Stratified Simple Random Sampling

The aim of this study is to compare two designes of carring out sample surveys to calculate the basic hourly wages of health workers. Used data is from the report “Health workers’ hourly wages“ colected by the National Institute for Health Development in 2013. These designes should ensure 70% rotation. Our methods were compared by simulation. Regression estimation was used to increase accuracy of estimates.

Stratified simple random sampling was chosen as the first design. In this case estimates were compared between proportional and Neyman allocation. To ensure the rotation permanent random number method was used. As for the second design a new design based on systematic cluster sampling was constructed.

Stratified simple random sampling Neyman allocation proved to be the best. Regression estimation improved the most the desing constructed on the systematic cluster sampling.
Keywords: health care professional, average wages, random sampling, systematic sampling.

Sisukord

Sissejuhatus

Käesoleva töö eesmärgiks on võrrelda kahte meetodit valikuuringu teostamiseks tervishoiutöötajate palgaandmete uurimiseks. Töö on tellitud Tervise Arengu Instituudi (edaspidi TAI) poolt ning ka kasutatvad meetodid on nende poolt välja pakutud. Seni on TAI teostanud kõikset uuringut, kus andmed kogutakse tervishoiuasutustelt nende töötajate kohta.

Kuna uuring viiakse läbi igal aastal, siis on TAI sooviks välja töötada meetod valikuuringu teostamiseks, mis tagaks valimite osalise kattumise aastate lõikes (rotatsioon). Kahte võimalikku disaini uuritaksegi käesolevas töös.

Töö koosneb kahest osast – teoreetilisest ning praktilisest. Teoreetilises osas tutvustatakse tõenäosusliku valikuuringu põhimõisteid, mida kasutatakse hiljem praktilise ülesande lahendamisel. Teoreetiline osa on referatiivne ning suures osas põhineb raamatul (Traat ja Inno, 1997).

Töö praktilises osas viiakse läbi simulatsioon etteantud meetodite võrdlemiseks. Selleks kasutatakse tervishoiuasutustelt kogutud aruande „Tervishoiutöötajate tunnipalk“ 2013. aasta andmeid. Antud töö raames keskendutakse vaid põhitunnipalga uurimisele. Valikumeetodite paremust võrreldakse hinnangute suhteliste vigade kaudu. Samuti leitakse uuritavale tunnusele regressioonhinnang mõlema meetodi jaoks.

Esimene disain on valitud nii, et valim moodustuks isikupõhiselt. Selleks disainiks on valitud lihtne juhuslik kihtvalik, kus on oluline leida sobivaim paigutus kihtide vahel, et minimiseerida hinnangute dispersioonid. Olemasoleva info põhjal moodustasid kihid 3 ametigruppi: arstid, õed ja ämmaemandad. Lihtsa juhusliku valiku korral on olemas metoodika rotatsiooniga arvestava valimi võtmise jaoks: püsijuhuarvude meetod.

Teine disain sooviti moodustada asutusepõhine. Selleks konstrueeriti süstemaatilisele valikule baseeruv disain. Sobiva freimi järjestuse korral annab süstemaatiline valik täpsema hinnangu uuritava tunnuse keskmisele kui lihtne juhuslik valik. Süstemaatilise valiku eripära tõttu ei saa valimid osaliselt kattuda ja tekib probleem rotatsiooni kattuvusega. Seepärast moodustati uus valimi võtmise meetod, mis koosneb freimist võetud süstemaatilisest valimist ja osast eelmisel aastal kaasatud valimist. Selle uue disaini põhjal tuli leida ka hinnangute arvutamise meetodid.

1. Tõenäosuslik valikuuring

Järgnevas peatükis on välja toodud olulisimad mõisted tõenäosusliku valikuuringu teooriast, mida töös hiljem kasutatakse.

1.1. Andmekogumise meetodid

Andmete kogumisel vaadeldakse põhiliselt kolme meetodit: kõikset uuringut, andmete kogumist registrisse ja valikuuringut.

Kõikse uuringu puhul kogutakse andmed üldkogumi kõigilt objektidelt, et saada täpset informatsiooni üldkogumi kohta kindlal ajahetkel.

Registrid on andmebaasid mitmesuguste üldkogumite kohta: rahvastikuregister, meditsiinitöötajateregister jne. Registritesse kantakse pidevalt regulaarsete aruannete andmed. Valikuurngutes kasutatakse registreid sageli abiinformatsiooni allikatena nii uuringu planeerimise, valimi võtmise kui ka tulemuste hindamise faasis.

Valikuuring on statistiline uuring, milles otsused üldkogumi kohta tehakse üldkogumi ühe osa (valimi) põhjal. Andmeid kogutakse ainult valimilt. Valikuuringul on kõikse uuringu ees mitmeid eeliseid, näiteks väiksem maksumus, suurem kiirus, paindlikkus, laiem rakendatavus.

1.2. Valikuuringu põhimõisted

Valikumeetodid jagunevad empiirilisteks ja tõenäosuslikeks. Tõenäosusliku valiku korral on iga objekti jaoks fikseeritud tema kaasamistõenäosus ehk tema tõenäosus valimisse sattuda. Empiirilise valiku korral kaasamistõenäosusi teada ei ole. Antud töö raames kasutatakse vaid tõenäosuslikke valikumeetodeid, seega mõeldakse edaspidi valimi all tõenäosuslikku valimit.

Definitsioon 1.2.1. *Tõenäosuslikuks valikuks* nimetatakse niisugust valikut üldkogumist, mille korral:

- saab defineerida kõigi võimalike valimite hulga
$$S = \{s_1, s_2, \dots, s_M\};$$
- 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 $p(s)$.

Tõenäosuslikud meetodid jagunevad tagasipanekuga (TGA) ja tagasipanekuta (TTA) valikuteks. Esimesel juhul võib iga üldkogumi objekt sattuda valimisse rohkem kui üks kord, teisel juhul saab iga objekt valimisse sattuda vaid ühe korra.

Definitsioon 1.2.2. *Valikudisainiks* nimetatakse tõenäosusjaotust $p(s)$ kõigi antud valiku jaoks võimalike valimite hulgal S.

Definitsioon 1.2.3. *Loendiks* ehk *freimiks* nimetatakse vahendit (nimekiri, register, andmebaas jms), mis võimaldab pääseda üldkogumi objektide juurde. Loend peab:

- identifitseerima igat üldkogumi objekti ning võimaldama neid valimisse kaasata vastavalt valikudisainile;
- võimaldama kontakti saamist valitud üldkogumi objektidega (telefonitsi, koduvisiit, elektronpostiga saadetud küsimustik jms). (Särndal et al., 1992)

1.3. Valikudisaini karakteristikud

Valikudisain on fundamentaalse tähtsusega mõiste valikuteooriast. Valikudisainiga $p(s)$ on määratud kõigi hinnangute statistilised omadused. Disainile optimaalse hinnangu konstrueerimiseks ja tema statistiliste omaduste esitamiseks ei kasutata otseselt disaini ennast vaid selle karakteristikuid: kaasamis- ja valikutõenäosust.

Definitsioon 1.3.1. Üldkogumi objekti i $(i = 1, 2, \dots, N)$ *kaasamistõenäosuseks* π_i nimetatakse tõenäosust, millega see objekt kaasatakse valimisse antud disaini $p(s)$ korral.

Üldkogumi objekti i kaasamistõenäosust π_i võib vaadelda, kui

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

Analoogselt avaldub kahe üldkogumi elemedi i ja j üheaegne kaasamistõenäosus ehk teistjärku kaasamistõenäosus

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

Definitsioon 1.3.2. *Kaasamisindikaator* I_i on iga üldkogumi objekti i $(i = 1, 2, \dots, N)$ jaoks määratud binaarne juhuslik suurus, mis iseloomustab objekti kaasamist valimisse

$$I_i = \begin{cases} 1, & \text{kui objekt } i \text{ on valimis} \\ 0, & \text{muidu.} \end{cases}$$

Mõned tähtsamad disainikarakteristikud on järgmised:

$E(I_i)$ – 1. järku moment, objekti *i* oodatav valikute arv;

$E(I_i I_j)$ – 2. järku moment;

$V(I_i) = \Delta_{ii}$ – valikuindikaatori I_i dispersioon;

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

TTA disainide puhul kehtivad seosed:

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

$V(I_i) = \Delta_{ii} = \pi_i(1 - \pi_i)$;

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

1.4. Üldkogumi parameetrite hindamine

Olgu antud üldkogum $U = \{1, .., N\}$ ja selle tunnuse y väärtused $y_1, y_2, \dots, y_N$. Üheks tähtsaks parameetriks, mida hinnatakse valikuuringute teoorias on üldkogumi summa $Y = \sum_U y_i$. Üldkogumi keskmine avaldub summa kaudu $\bar{Y} = Y/N$. Käesolev alapeatükk põhineb kospektil (Traat ja Lepik, 2013).

Teoreem 1.4.1. *(Üldine hindamisteoreem)* Üldkogumi kogusumma $Y = \sum_U y_i$ nihketa hinnang on

$$\hat{Y} = \sum_U I_i \check{y}_i \;\; (\text{või } \hat{Y} = \sum_U w_i y_i),$$

kus

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

Selle disainipõhine dispersioon on

$$V(\hat{Y}) = \sum\sum_U \Delta_{ij}\breve{y}_i\breve{y}_j,$$

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

$$\hat{V}(\hat{Y}) = \sum\sum_U \breve{\Delta}_{ij}\,\breve{y}_i\breve{y}_j I_iI_j \text{ (või } \hat{V}(\hat{Y}) = \sum\sum_U \breve{\Delta}_{ij} w_i\, y_i w_j y_j),$$

kus

$$\breve{\Delta}_{ij} = \frac{\Delta_{ij}}{E(I_iI_j)}.$$

Teoreem 1.4.2. Fikseeritud mahuga disaini $p(k)$ korral saab hinnangu $\hat{Y} = \sum_U I_i\breve{y}_i$ dispersiooni esitada alternatiivsel kujul:

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

ja eeldusel, et $E(I_iI_j) > 0 \quad \forall i \neq j \in U$, on dispersiooni $V(\hat{t})$ nihketa hinnanguks:

$$\hat{V}(\hat{Y}) = -\frac{1}{2}\sum\sum_U I_iI_j\breve{\Delta}_{ij}\,(\breve{y}_i - \breve{y}_j)^2.$$

Üldkogumi keskmine on defineeritud järgmiselt:

$$\bar{Y} = \frac{1}{N}\sum\nolimits_U y_i = \frac{\mathrm{Y}}{N}.$$

Kui üldkogumi maht *N* on teada, siis piisab keskmise nihketa hinnangu saamiseks kogusumma hindamisest:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N}. \tag{1}$$

Dispersiooni ja dispersioonihinnang avalduvad $\hat{Y}$ dispersiooni kaudu:

$$V\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}V(\hat{Y}), \tag{2}$$

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\hat{V}(\hat{Y}). \tag{3}$$

2. Lihtne juhuslik kihtvalik ja seda iseloomustavad karakteristikud

Lihtne juhuslik kihtvalik põhineb kahel valikudisainil: kihtvalik ja lihtne juhuslik valik.

2.1. Lihtne juhuslik valik (LJV)

Lihtsat juhuslikku valikut on võimalik teostada nii tagasipanekuga (TGA) kui ka tagasipanekuta (TTA) disainina. Käesolevas töös LJV kasutades mõeldakse LJV TTA ning edaspidine teooria kehtib TTA disaini kohta.

Olgu üldkogum $U = \{1, \dots, N\}$. Sellest on võimalik moodustada $M=\binom{N}{n}$ hulka suurusega n. Need hulgad moodustavad LJV kõigi võimalike valimite hulga $S = \{s_1, \dots, s_M\}$, kus igal valimil on võrdne tõenäosus realiseeruda.

Definistioon 2.1.1. *Lihtsa juhuvaliku disainiks* nimetatakse jaotust $p(s)$ kõigi valimite hulgal *S*, kus

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

Parameeterhinnangute leidmiseks on vaja teada disainikarakteristikuid. Need on järgmised:

$f = \frac{n}{N}$ − valikusuhe;

$\pi_i = f$ − esimest järku kaasamistõenäosus;

$\pi_{ij} = f\frac{n-1}{N-1}, i \neq j$ − teist järku kaasamistõenäosus;

$\Delta_{ii} = f(1-f) - I_i$ dispersioon;

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

(Traat ja Lepik, 2013)

Teoreemidest 1.4.1. ja 1.4.2. avaldub järgmine teoreem:

Teoreem 2.1.1. Lihtsa juhuvaliku TTA korral nihketa hinnang ÜK summale $Y = \sum_U y_i$ avaldub järgmiselt:

$$\hat{Y} = \frac{N}{n}\sum_U I_i y_i = \frac{N}{n}\sum_s y_i,$$

ehk alternatiivselt

$$\hat{Y} = N\bar{y}.$$

Hinnangu dispersioon on järgmine:

$$V(\hat{Y}) = N^2(1-f)\frac{S_y^2}{n}$$

Ja dispersiooni hinnang

$$\hat{V}(\hat{Y}) = N^2(1-f)\frac{s_y^2}{n},$$

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

$S_y^2 = \frac{1}{N-1}\sum_U (y_i - \bar{Y})^2$ tunnuse *y* ÜK dispersioon,

$s_y^2 = \frac{1}{n-1}\sum_s (y_i - \bar{y})^2$ tunnuse *y* valimi dispersioon.

(Traat ja Lepik, 2013)

Eelmise teoreemi ja valemite (1)-(3) põhjal avalduvad LJV nihketa hinnangud keskmisele kujul:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{N\bar{y}}{N} = \bar{y}, \tag{4}$$

$$V(\hat{\bar{Y}}) = \frac{1}{N^2}V(\hat{Y}) = (1-f)\frac{S_y^2}{n}, \tag{5}$$

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2}\hat{V}(\hat{Y}) = (1-f)\frac{s_y^2}{n}. \tag{6}$$

2.2. Kihtvalik

Kihtvaliku teostamisel jagatakse üldkogum mingi kihistava tunnuse alusel mittekattuvateks osakogumiteks ehk kihtideks. Kihid on üksteisest sõltumatud ning nendes võib rakendada erinevaid valikumeetodeid.

Kihtvalikut kasutatakse:

- hinnangu täpsuse tõstmiseks – tunnuse *y* suhtes homogeensed kihid tagavad valimihinnangu väikese varieeruvuse;
- osakogumite hindamiseks – osakogumite kasutamisel kihtidena saame täpsema hinnangu isegi väikese valimimahu juures;
- erinevat käsitlust vajavate kihtide hindamine – kallimalt uuritavate objektide valimimahtu vähendatakse või suure kaoprotsendiga valimit suurendatakse;
- uuringu administreerimine – kihid moodustatakse intervjueerijate keskuste järgi, et vähendada uuringu kulusid.

Olgu ÜK *U* jagatud osakogumiteks U_h, kus $h \in \{1, \dots, H\}$, $U_h \cup U_h' = \emptyset$, $U = \bigcup_{h=1}^{H} U_h$, ning N_h on h-nda kihi maht. Olgu h-nda kihi kogusumma $Y_h = \sum_{U_h} y_i$ ning hinnangu h-nda kihi kogusummale $\hat{Y}_h$.

Teoreem 2.2.1. 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 hinnangute $\hat{Y}_h$ dispersioonide summana

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

ning dispersiooni hinnang

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

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

h-nda kihi osakaalu tähistatakse $W_h = \frac{N_h}{N}$, $h \in \{1, \dots, H\}$.

Järeldus 2.2.1. Kihtvaliku korral avaldub hinnang ÜK keskmisele kihikeskmiste hinnangute kaalutud keskmisena,

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

mille dispersioon on

$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{H} W_{\mathrm{h}}^2 V\left(\hat{\bar{Y}}_h\right).$$

Kui kihtides kasutatakse nihketa hinnanguid dispersioonidele $\hat{V}\left(\hat{\bar{Y}}_h\right)$, siis nihketa hinnang dispersioonile on

$$\hat{V}\left(\hat{\bar{\mathrm{Y}}}\right) = \sum_{h=1}^{H} W_{\mathrm{h}}^2 \hat{V}\left(\hat{\bar{Y}}_h\right).$$

(Traat ja Lepik, 2013)

2.3. Lihtne juhuslik kihtvalik (LJKV)

Kui kihtvaliku kõikides kihtides kasutatakse lihtsat juhuslikku valikut tagasipanekuta, siis nimetatakse sellist valikumeetodit *lihtsaks juhuslikuks kihtvalikuks.* Erinevates kihtides võib kasutada erinevaid valikusuhteid

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

Kasutades järeldust 2.2.1 ja valemeid (4)-(6), avalduvad keskmise hinnangud järgmised:

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h, \tag{7}$$

$$V\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1 - f_h) S_{yh}^2 / n_h,$$

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1 - f_h) s_{yh}^2 / n_h, \tag{8}$$

kus $S_{yh}^2 = \frac{1}{N_h - 1}\sum_{U_h}(y_i - \bar{Y}_h)^2$,

$s_{yh}^2 = \frac{1}{n_h - 1}\sum_{s_h}(y_i - \bar{y}_h)^2$,

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

2.4. Valimi paigutamine kihtidesse

Lihtsa juhusliku kihtvaliku korral on oluline määrata valimi suurus igas kihis, sest sellest sõltub hinnangu täpsus. Käesolevas töös vaadatakse kahte valimi paigutamise meetodit: võrdeline paigutus ja Neymani paigutus.

Neymani paigutuse korral määratakse valmimahud nii, et hinnangu dispersioon oleks minimaalne:

$$n_h = n\frac{N_h S_{yh}}{\sum_{g=1}^{H} N_g S_{yg}} \tag{9}$$

Valemist on näha, et suuremast kihist võetakse valimisse rohkem objekte. Samuti sõltub n_h uuritava tunnuse standardhälbest S_{yh} ehk mida rohkem varieeruvad y_i väärtused kihis U_h, seda seda rohkem objekte võetakse antud kihi valimisse.

Võrdsete maksumuste korral kõikides kihtides annab valem (9) sellised kihtide valimimahud $(n_h, h = 1, \dots, H)$, mille korral $V(\hat{Y})$ (ja ka $V\left(\hat{\bar{Y}}\right)$) on minimaalsed (Traat ja Inno, 1997). Seetõttu nimetatakse sageli sellist valimi paigutust ka *optimaalseks paigutuseks*.

Meetodi puuduseks on see, et suurused S_{yh} pole sageli teada (need asendatakse näiteks pilootuuringu väärtustega). Samuti on see paigutus optimaalne ainult ühe uuritava tunnuse jaoks. Enamasti on uuringu all mitu uuritavat tunnust ja see paigutus ei pruugi teistele sobida.

Võrdelise valimi paigutuse korral on vastavate kihtide osakaalud valimis ja üldkogumis võrdsed ehk

$$n_h = n\frac{N_h}{N} \tag{10}$$

Selle paigutuse korral võetakse suuremast kihist suurem valim. Antud valimimahu leidmise valem ei sõltu uuritavast tunnusest ja on „ühtemoodi hea“ kõikide uuritavate tunnuste jaoks.

3. Süstemaatiline klastervalik ja seda iseloomustavad karakteristikud

Järgnevas peatükis on välja toodud süstemaatilise klastervaliku põhimõte ja keskmise hindamise valem.

3.1. Klastervalik

Peaaegu alati on üldkogumi objektid grupeeritud mingisugustesse rühmadesse ehk klastritesse ja üldkogumil esineb mingi loomulik struktuur. Näiteks kuuluvad riigielanikud selle haldusüksustesse: valdadesse. *Klastervaliku* korral ei võeta valimisse mitte üksikuid objekte, vaid valitakse klastreid, millest igaüks kaasab valimisse kõik enda objektid. See tähendab, et iga valitud vald kaasab valimisse kõik oma elanikud.

Klastervalikut kasutatakse kulude kokkuhoidmiseks või siis, kui objektide tasemel freim pole kättesaadav (näiteks koolide loend on olemas, kuid õpilaste oma puudub). Antud töös moodustavad klastrid tervishoiu asutused. Kui mingi asutus satub valimisse, siis kaasab see valimisse kõik oma tervishoiutöötajad. Selline valikuprotseduur on teostamise mõttes mugav ja lihtne.Siiski näitavad varasemad uuringud, et klastervalik ei ole tavaliselt efektiivsem kui LJV TTA hinnangu täpsuse mõttes.

Olgu üldkogum $U = (1, \dots, N)$ jagatud M klastriks $U_1, U_2, \dots, U_M$. Olgu klastrite indeksite kogum $U_I = \{1, \dots, k, \dots, M\}$. Sel juhul

$$U = \bigcup_{k \in U_I} U_k\,, \qquad N = \sum\nolimits_{k \in U_I} N_k,$$

kus N_k on klastrisse U_k kuuluvate objektide arv.

Edaspidi vaadatakse objektidena klastreid, mille üldkogumit tähistatakse U_I. Indeks I lisatakse kõikidele tähistustele, mis on seotud klastrite kui objektidega.

Klastervaliku korral võetakse üldkogumist U_I klastrite valim s_I klastrite arvuga n_I ja uuritavasse valimisse s kaasatakse kõik valitud klastritesse kuuluvad objektid

$$s = \bigcup_{k \in s_I} U_k\,, \qquad n_s = \sum\nolimits_{k \in s_I} N_k.$$

Klastervaliku disainiks võib olla ükskõik mis disain. Disain $p_I(\cdot)$ määrab klastrite esimest ja teist järku kaasamistõenäosused vastavalt seostega

$$\pi_{Ik} = \sum_{k \in s_I} p_I(s_I), \qquad \pi_{Ikl} = \sum_{k,l \in s_I} p_I(s_I),$$

kus summeerimispiirkond $k \in s_I$ tähendab summeerimist üle valimite s_I, mis sisaldavad klastrit k.

Objekti esimest järku kaasamistõenäosus on võrdne tema klastri kaasamistõenäosusega

$$\pi_i = P(i \in s) = P(k \in s_I) = \pi_{Ik}.$$

Olgu klastri U_k kogusumma $Y_k = \sum_{U_k} y_i$. Seega on üldkogumi U kogusumma esitatav kujul

$$Y = \sum_U Y_i = \sum_{U_I} Y_k$$

ning keskmine

$$\bar{Y} = \frac{1}{N} \sum_{U_I} Y_k.$$

Teoreemi 1.4.1 põhjal on kogusumma nihketa hinnang TTA klastervaliku korral kujul

$$\hat{Y} = \sum_{s_I} \breve{Y}_k = \sum_{s_I} \frac{Y_k}{\pi_{Ik}} = \sum_{k \in s_I} \sum_{U_k} \frac{y_i}{\pi_i} = \sum_s \frac{y_i}{\pi_i}$$

ning valemi (1) kohaselt on keskmise hinnanguks

$$\hat{\bar{Y}} = \frac{\hat{\bar{Y}}}{N} = \frac{1}{N} \sum_s \frac{y_i}{\pi_i}. \qquad (11)$$

3.2. Süstemaatiline valik

Süstemaatilise valiku korral võetakse valimisse järjestatud loendist kõik üksteisest fikseeritud sammu a kaugusel asuvad objektid, alustades juhuslikult leitud objektist r. Fikseeritud üldkogumimahu N korral määrab a valimisuuruse.

Tähistades $n = \left[\frac{N}{a}\right]$, kus nurksulud tähistavad täisosa võtmist, saame kirjutada

$$N = na + c,$$

kus täisarv c on valikujääk $0 \leq c \leq a$. Süstemaatilise valiku alguspunkt r määratakse diskreetse juhusliku suuruse $\boldsymbol{r}$ abil, mille korral $P(\boldsymbol{r} = r) = 1/a$ iga $r = 1,2, \dots, a$ korral.

Valimimaht võib ühe ja sama sammu korral omandada 2 erinevat väärtust, sõltuvalt realiseerunud alguspunktist *r*:

$$n_s = \begin{cases} n, & \text{kui } c \leq r \leq a, \\ n+1, & muidu. \end{cases}$$

Kuna süstemaatilise valiku korral kuulub iga üldkogumi objekt parajasti ühte valimisse, siis on kõikidel üldkogumi objektidel sama kaasamistõenäosus

$$\pi_i = \sum_{\substack{i \in s \\ s \in S}} p(s) = \frac{1}{a}. \qquad (12)$$

Süstemaatilise valiku korral ei saa kogusumma hinnangu dispersiooni vaid meile teadaoleva valimi põhjal hinnata. On teada seos, et süstemaatilise valiku kogusumma hinnangu dispersiooni hinnang on väiksem kui LJV dispersiooni hinnag. Seega kasutatakse praktikas siin sageli LJV dispersiooni hinnangut.

Süstemaatilise valiku algoritmi järgi ei ole võimalik võtta iga suurusega valimit. Näiteks olgu üldkogumi maht 20. Võttes sellest valim sammuga $a = 3$, saadakse valim mahuga 6 või 7. Sammu $a = 2$ korral saadakse valim mahuga 10. Seega ei ole võimalik võtta valimit mahuga 8 või 9.

Selle probleemi lahendamiseks võib kasutada muudetud algoritmi, mille korral arvutatakse samm a etteantud valimimahu järgi: $a = N/n$. Seejärel leitakse juhuslik element $r \in [1, a]$. Valimi esimeseks elemendiks on ÜK *r*'s element, teiseks $[r + a]$, kus nurksulud tähistavad täisosa. Valimi *n*-ndaks elemendiks on ÜK element $[r + (n-1)a]$. See algoritm on kasutusel ka tarkvara SAS valimi võtmise protseduuris *surveyselect*.

3.3. Süstemaatiline klastervalik

Käesolevas töös kasutatakse klastervaliku teostamiseks süstemaatilist valikut. Asendades valemis (11) esimest järku kaasamistõenäosuse süstemaatilise valiku kaasamistõenäosusega (12), saadakse antud disaini keskmise hinnanguks

$$\hat{\bar{Y}} = \frac{1}{N}\sum_s \frac{y_i}{\pi_i} = \frac{a}{N}\sum_s y_i.$$

Kuna muudetud algoritmiga võetud valimi korral $a = N/n$, sii on keskmise hinnang:

$$\hat{\bar{Y}} = \frac{a}{N}\sum_s y_i = \frac{N}{nN}\sum_s y_i = \frac{1}{n}\sum_s y_i. \quad (13)$$

4. Dispersiooni hindamine *Jackknife* meetodil

Tihtipeale ei ole keerulisematel valikudisainidel valemeid, millega soovitud statistikut ja selle dispersiooni hinnata. Sellistel juhtudel on üheks hindamise võimaluseks *jackknife* hinnang. Järgnev peatükk põhineb artiklil (Bruch et al., 2011) ning tutvustab dispersiooni hindamist *jackknife* meetodiga. Artiklis toodud valemid kohandatakse selles töös kasutatud keskmise hindamisele.

Olgu hinnang üldkogumi keskmisele $\hat{\bar{Y}} = \hat{\bar{Y}}(y_1, \dots, y_n)$ ning olgu $\hat{\bar{Y}}_{-i} = \hat{\bar{Y}}(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$, mis on samuti hinnang keskmisele, kuid hinnangu leidmisel on valimist eemaldatud üks element.

Jackknife hinnang keskmisele on:

$$\hat{\bar{Y}}_{JK} = n\hat{\bar{Y}} - \frac{n-1}{n}\sum_{i=1}^{n} \hat{\bar{Y}}_{-i}.$$

Seda hinnangut võib kirjutada ka kujul:

$$\hat{\bar{Y}}_{JK} = \frac{1}{n}\sum_{i=1}^{n} \hat{\bar{Y}}_i',$$

kus $\hat{\bar{Y}}_i' = n\hat{\bar{Y}} - (n-1)\,\hat{\bar{Y}}_{-i}$. Suurusi $\hat{\bar{Y}}_i'$ $(i = 1, \dots, n)$ nimetatakse *jackknife*'i pseudoväärtusteks. Eeldades, et pseudoväärtused on sõltumatud ühtlase jaotusega ning sama dispersiooniga kui $\sqrt{n}\hat{\bar{Y}}$, on keskmise hinnangu *jackknife* dispersiooni hinnang kujul:

$$\hat{V}_{d1JK}\left(\hat{\bar{Y}}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\bar{Y}}_i' - \frac{1}{n}\sum_{j=1}^{n}\hat{\bar{Y}}_j'\right)^2.$$

$$\hat{V}_{d1JK}\left(\hat{\bar{Y}}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(n\hat{\bar{Y}} - (n-1)\,\hat{\bar{Y}}_{-i} - \frac{1}{n}\sum_{j=1}^{n}\left(n\hat{\bar{Y}} - (n-1)\,\hat{\bar{Y}}_{-i}\right)\right)^2 \qquad (14)$$

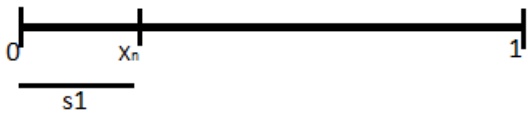
Antud hinnangut nimetatakse kustuta-1 (ingl *delete-1*) *jackknife* hinnanguks valimi keskmisele. Nagu valemist (14) on näha, eemaldatakse võetud valimist igat elementi ühe korra ning saadud n hinnangu põhjal hinnatakse dispersioon üldkogumis.

Klastervaliku korral on objektideks klastrid ning uuritava tunnuse väärtusteks on klastersummad. Seetõttu tuleks *jackknife* meetodi korral eemaldada arvutustes terveid klastreid.

5. Valimi rotatsioon

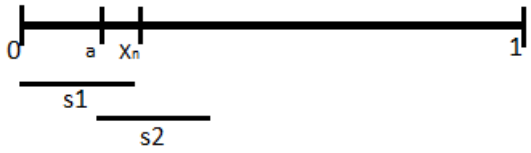
Korduvatel uuringutel on sageli probleemiks ühtede ja samade objektide sattumine valimisse mitmel järjestikusel uuringul. Samas mõni teine objekt ei pruugi pikka aega valimisse sattuda. Korduvatel uuringutel uute objektide kaasamist valimisse nimetatakse *rotatsiooniks*. Mõnes uuringus on valimite kattumine taotuslik: see võimaldab võrrelda uuritavaid tunnuseid järjestikustel perioodidel samadel objektidel. Käesolev peatükk põhineb artiklil (Cox et al., 1995) ning kirjeldab püsijuhuarvude meetodit (ingl *permanent random number* ehk *PRN)*. Antud meetodit saab rakendada LJV korral ning seega sobib eelnevalt kirjeldatud LJKV teostamiseks, kus kattuvus tekitatakse kihtides.

Olgu soovitud valimimaht n. Iga üldkogumi objekt seotakse ühe sõltumatu juhusliku arvuga ühtlasest jaotusest vahemikus nullist üheni, $X_i \sim U(0,1), i = 1, \dots, N$. Seejärel sorteeritakse freim X_i'de järgi näiteks kahanevalt. Valimi moodustavad n esimest freimi objekti (Joonis 1).



Joonis 1. Esimese valimi moodustamine

Igal järgneval valimi võtmise korral jäävad üldkogumi objektidele omistatud juhuslikud arvud samaks. Oletame, et teisel aastal soovitakse rotatsiooni p protsenti (ehk uus valim moodustub p protsendi ulatuses uutest objektidest ja $q = 100 - p$ protsenti on eelmise valimi objekte). Sel juhul leitakse punkt a, millest valimi s_1 objektidele vastavatest juhuslikest arvudest X_i on q protsenti suuremad kui a. Teise aasta valimi s_2 moodustavad a'le järgnevad n objekti (Joonis 2).

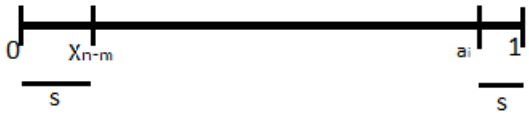


Joonis 2. Teise valimi moodustamine

Uue objekti lisamisel freimi genereeritakse sellele uus juhuslik arv ning see salvestatakse sorteeritud freimi vastavasse kohta. Freimi objekti kadumisel, kustutatakse

see koos tema juhusliku arvuga. Sedasi püsib freim ajakohasena ning samasid juhuslikke arve saab kasutada ka järgmisel valimi võtmisel.

Valimisse ei pea võtma just esimest n elementi. Võib fikseerida suvalise arvu $a_i \in [0; 1]$ ning moodustada valimi sellele järgnevast n elemendist. Kui a'le järgneb $m < n$ elementi, siis võetakse valimisse need m elementi ja üldkogumi esimesed $n - m$ elementi (Joonis 3).



Joonis 3. Liitega valimi moodustamine

6. Regresioonhinnang (GREG)

Regresioonhinnang on klassikaline lisainformatisooni kasutav kogusumma hinnang, mis võimaldab parandada disainil põhineva hinnangu täpsust. Järgnev peatükk põhineb konspektil (Traat, 2012) ning toob välja regresioonhinnangu valemid keskmisele ja dispersioonile.

6.1. Mudeli eeldused

Olgu uuritav tunnus y_i mõõdetud objektil $i,\ i \in U$ ja $x_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{Ji} \end{pmatrix}$ $J \times 1$ abitunnuse vektor objektil $i,\ i \in U$.

Kogusummale regressioonhinnangu teostamiseks eeldatakse regressioonimudelit üldkogumis:

1. $y_i,\ i \in U$ väärtused on juhuslikud väärtused $Y_i \sim \xi$ (jaotusega ξ);
2. ξ keskväärtus ja dispersioon avalduvad järgmiselt:

$$E_\xi(Y_i|x_i) = x_i'\beta = \sum_{j=1}^{J} \beta_j x_{ji},$$

$$V_\xi(Y_i|x_i) = \sigma_i^2,$$

Y_i on sõltumatud;

3. x_i ei ole juhuslik.

Siin on $\beta' = (\beta_1, \dots, \beta_J)$ regressioonikordajate vektor.

Kui regressioonimudel hinnata üle terve populatsiooni, kus y_i väärtused on teada, siis kaalutud vähimruutude hinnanguga saab hinnata β järgmiselt:

$$\hat{\beta} = B = \sum_U \left[\frac{x_i x_i'}{\sigma_i^2}\right]^{-1} \sum_U \frac{x_i y_i}{\sigma_i^2} \ : J \times 1. \tag{15}$$

Üldkogumijäägid avalduvad valemiga:

$$E_i = y_i - x_i'B, \qquad i \in U. \tag{16}$$

Suurused B, E_i ei ole teada ning need tuleb hinnata valimi järgi. Olgu valim võetud TTA disainiga ($E(I_i) = \pi_i, E(I_i I_j) = \pi_{ij}$). Suurus B sisaldab kahte kogusummat:

$T = \sum_U \frac{x_i x_i'}{\sigma_i^2}$ – maatrtksite summa (koosneb $J \times J$ summast);

$t_{xy} = \sum_U \frac{x_i y_i}{\sigma_i^2}$ – vektorite summa (koosneb J summast).

Neid summasid hinnatakse disainipõhiselt

$$\hat{T} = \sum_s \frac{x_i x_i'}{\sigma_i^2 \pi_i}$$

$$\hat{t}_{xy} = \sum_s \frac{x_i y_i}{\sigma_i^2 \pi_i}$$

ning saadakse B hinnang:

$$\hat{B} = \sum_s \left[\frac{x_i x_i'}{\sigma_i^2 \pi_i}\right]^{-1} \sum_s \frac{x_i y_i}{\sigma_i^2 \pi_i} = \hat{T}^{-1} \hat{t}_{xy}.$$

Vektor $\hat{B}$ on arvutatav valimi põhjal, selle põhjal prognoosid y väärtustele on

$$\hat{y}_i = x_i' \hat{B}, \qquad i \in U$$

ja valimijäägid

$$e_i = y_i - \hat{y}_i, \qquad i \in s. \tag{17}$$

Kogusumma hinnangut Y saab teisendada järgmiselt:

$$Y = \sum_U y_i = \sum_u \hat{y}_i + \sum_u (y_i - \hat{y}_i), \tag{18}$$

kus $\hat{y}_i$ on teada iga $i \in U$ korral, kuid y_i on teada vaid valimis.

6.2. Regressioonhinnang

Regressioonhinnangu (GREG) saamiseks hinnatakse nihketult teist liiget valemis (18):

$$\hat{Y}_{greg} = \sum_u \hat{y}_i + \sum_s \frac{y_i - \hat{y}_i}{\pi_i}. \tag{19}$$

Valemi (1) kohaselt avaldub keskmise hinnang kujul:

$$\hat{\bar{Y}}_{greg} = \frac{\hat{Y}_{greg}}{N} = \frac{1}{N}\sum\nolimits_{u} \hat{y}_i + \frac{1}{N}\sum\nolimits_{s} \frac{y_i - \hat{y}_i}{\pi_i}. \quad (20)$$

Teoreem 6.2.1. (Regressioonhinnang) Üldkogumi kogusumma $Y = \sum_U y_i$ regressioon-hinnang on antud valemiga (19), dispersiooniga

$$V(\hat{Y}_{greg}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij}(w_i E_i)(w_j E_j)$$

ja dispersiooni hinnanguga

$$\hat{V}(\hat{Y}_{greg}) = \sum \sum_{i,j\in s} \frac{\Delta_{ij}}{\pi_{ij}}(w_i g_{is} e_i)(w_j g_{js} e_j),$$

kus üldkogumi jäägid E_i on antud valemis (16), valimi jäägid e_i valemis (17) , g-kaalud avalduvad valemiga $g_{is} = 1 + \left(X - \hat{X}\right)' \hat{T}^{-1} \frac{x_i}{\sigma_i^2}$ ning $w_i = \frac{1}{\pi_i}$ on disainikaalud.

Valemite (2) ja (3) kohaselt avalduvad Teoreem 6.2.1 valemid üldkogumi keskmisele kujul:

$$V\left(\hat{\bar{Y}}_{greg}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij}(w_i E_i)(w_j E_j),$$

$$\hat{V}\left(\hat{\bar{Y}}_{greg}\right) = \frac{1}{N^2}\sum \sum_{i,j\in s} \frac{\Delta_{ij}}{\pi_{ij}}(w_i g_{is} e_i)(w_j g_{js} e_j). \quad (21)$$

7. Praktiline näide tervishoiutöötajate tunnipalga andmetel

Käesoleva peatüki eesmärgiks on kahe meetodi rakendamine konkreetsele andmestikule ja hinnangute täpsuse võrdlemine. Valimite võtmine ja hinnangute leidmine viidi läbi statistika paketiga SAS ning töö lõppu on lisatud vastav programmi kood (Lisa 1).

7.1. Andmestiku kirjeldus

Töö aluseks on 2013. aasta tervishoiutöötajate palgaandmed, mida Tervisearengu Instituut (TAI) kogub aruandega „Tervishoiutöötajate tunnipalk“ igal aastal märtsikuu kohta. Seda uuringut on läbi viidud alates 2002. aastast. Esmalt koguti andmeid vaid haiglatelt, alates 2006. aastast on uuringusse kaasatud kõik tervishoiuteenuse osutajad (hambaravid, perearstikeskused, Kaitsevägi jne). TAI tegeleb tervisestatistika kogumise, analüüsimise ja avaldamisega alates 2008. aastast.

Uuringus kajastatud tunnipalga andmed sisaldavad nii täis- kui osalise tööajaga töötajate andmeid ehk töötajaid, kes töötavad kas täis- või osalise koormusega, ja ka neid, kes märtsis osaliselt töölt puudusid (olid kas puhkusel, töövõimetus- või hooldusleh el). (Poolakese, 2013) Andmestikku tööga kaasas ei ole, sest palgaandmete puhul on tegemist delikaatsete isikuandmetega ning nende avalikustamine on keelatud. Andmestiku kokkuvõte on kättesaadav TAI andmebaasist (www.tai.ee/tstua).

Andmestiku kontrollimisel leiti mitmeid vigu. Näiteks selgus, et mõnedel ridadel, kus asutuse nimi ja isiku järjekorra number on samad, on isiku vanus, sugu või haridustase erinevad. Samuti leidus isik, kelle töökoormuseks oli märgitud 160 (täiskohaga töötades on töökoormus 1), kusjuures tema märtsikuu töötundide arv oli 160 tundi. TAI’d on antud vigadest teavitatud ning nad lubasid omalt poolt järgnevatel aastatel andmeid põhjalikumalt kontrollida. Leitud vead käesoleva töö tulemusi ei mõjuta.

Töös kasutatud andmestikus on 39 tunnust, millest käesolevas töös kasutati järgmiseid:

- asutuse nimi;
- töötaja järjekorra number – iga asutuse andmetel määratakse igale isikule unikaalne järjekorra number. Kui isik töötab samas asutuses mitmel ametikohal, siis märgitakse tema andmeid sisaldavatele ridadele sama number;
- vanus – isiku vanus täisaastates 31. märtsi 2013 seisuga;
- ametiala – kodeeritakse vastavalt sellele ametile, millel nad töötavad (Lisa 2);

- ametigrupp – kodeeritud tunnus, meie andmestikus 1-arst, 3-õde, 5-hooldaja
- põhitunnipalk valvega – (edaspidi põhitunnipalk) lepingulise põhipalga ja valvetundide eest makstud keskmine tunnitasu;
- välja põhitunnipalk – indikaatortunnus, mis näitab, kas antud isik on põhitunnipalga arvestuses sees.

Terviseamet on koostanud meditsiinitöötajate registri, mis sisaldab arstide, hambaarstide, õdede ja ämmaemandate töökoode, hariduse andmeid ja töökohti. Sellest lähtuvalt kodeeriti uus tunnus 'ametigrupp1', kus eraldati õdedest ämmaemandad nende ametiala koodi alusel (ämmaemandate ametikood on 2222). Kuna antud register ei sisalda andmeid hooldajate kohta, siis käesolevas töös neid ei uurita. Hooldajate andmetega ridade kustutamisel (hooldajate ametigrupp on 5) jäi andmestikku 12747 andmerida (Lisa 1.1).

TAI poolt läbiviidavas uuringus hinnatakse kolme tunnust: põhitunnipalk, kogutunnipalk ja kuupalk. Antud töös uuriti neist vaid põhitunnipalka. Indikaatortunnuse 'välja põhitunnipalk' põhjal jäeti andmestikus olnud 12 747 väärtusest keskmise arvutamiseks alles 12588 rida (Lisa 1.2), mis moodustab 98,75% kõikidest andmetest. Põhitunnipalga arvutustes kasutatud andmete põhitunnipalga karakteristikud on välja toodud Tabelis 1.

Tabel 1. Põhitunnipalga karakteristikud (eurodes)

Miinimum	Maksimum	Keskmine	Standardhälve
2,0000	71,4300	6,5884	3,5656

7.2. Isikupõhine vs asutusepõhine valikudisain

Töö eesmärgiks on võrrelda kahte valimi võtmise meetodit: isikupõhist, kus valim moodustatakse meditsiinitöötajate registri põhjal, ja asutusepõhist, kus valik tehakse asutuste registri põhjal ning iga asutus kaasab valimisse kõik oma tervishoiutöötajad. Samuti sooviti arvestada rotatsiooni 70%, et valimid oleksid kahe aasta lõikes paremini võrreldavad ning välistada samade objektide pidevat valimisse kaasamist.

Tervishoiutöötajate põhipalka on keeruline võrrelda, sest paljud inimesed töötavad osakoormusega. Samuti töötatakse osakoormustega mitmel ametikohal nii sama asutuse piires kui ka erinevates asutustes. Praeguse andmestiku ülesehitusega on võimalik

ühendada sama asutuse piires olevaid isikuid (seda juhul, kui vigased töötajate järjekorranumbrid ümber kodeerida), kuid ei ole võimalik liita sama isiku erinevates asutustes saadud palkasid ja töökoormuseid. Isikupõhise disainiga oleks filtritega lihtsam sama isiku erinevates asutustes teenitud palkasid liita. Selle disaini miinuseks võib osutuda mittevastamine, sest palgaandmed on delikaatsed isikuandmed.

Näide 7.2.1. Töötagu vaadeldav inimene koormusega 0,5 hooldajana ja koormusega 0,5 kiirabiõena. Seega on ta andmestikus esindatud kahel real. Kui ta töötab samas asutuses, siis on tema töötaja järjekorra number sama ning tema palgaandmeid on võimalik liita. Kui ta töötaks erinevates asutustes, siis ei oleks võimalik tema erinevaid ameteid ühendada, sest andmestikus puuduvad isikukoodid. Seega esindavad tema andmed justkui kahe erineva inimese, kes töötavad osakoormusega, palkasid. Teades registri põhjal inimese erinevaid töökohti ja ameteid, saab enamasti üheselt kätte tema kohta käivad read asutuse nime, ametiala, soo, vanuse ja haridustaseme järgi.

Isikupõhise valimi teostamiseks valiti käesolevas töös LJKV, kus üldkogum jaotati töötajate ametigruppide kaupa kihtideks. See kuulub abiinformatsiooni hulka ja on kättesaadav iga töötaja kohta meditsiinitöötajate registrist. Töös käsitletakse võrdelist ja Neymani paigutust eraldi ja võrreldakse neid omavahel. Kihtvaliku kasuks otsustati eesmärgiga saada valim üldkogumiga võimalikult sarnase ülesehitusega (võrdeline paigutus). Hajuvuse vähendamiseks uuriti ka Neymani paigutust. Samuti tagab kihtvalik väiksemates kihtides asuvate isikute (nt ämmaemandad) esindatuse valimis. Rotatsiooni saamiseks kasutati püsijuhuarvude meetodit igas kihis eraldi.

Asutusepõhise disaini loomisel lähtuti mõttekäigust, et kui asutus juba mõne töötaja kohta andmed esitab, siis on ta juba valimis ning võiks esitada aruande kõikide töötajate kohta. Pealegi on sellist meetodit lihtsam organiseerida ning kogutud andmeid kontrollida. Meetodile tuleb kasuks veel asjaolu, et mõned asutused ei pea selle disaini põhjal aruannet igal aastal esitama. Samuti ei ole selle disaini puhul vaja registrit kõikide meditsiinitöötajate kohta, vaid piisab asutuste nimekirjast.

Asutusepõhise valiku teostamiseks kasutati süstemaatilisele klastervalikule põhinevat meetodit, kus arvestati rotatsiooniga. Esmalt järjestati asutused töötajate arvu järgi kahanevalt ning moodustati asutuste valim s_{I_0} süstemaatilise valikuga kui nö eelmise aasta valim mahuga 70% soovitud valimimahust. Seejärel valim s_{I_1} kui selle aasta

valim suurusega 70% soovitud valimimahust ning kattuvuse saamiseks võeti s_{I_0} asutustest 43% kui valim $s_{I_{01}}$ süstemaatilise valikuga. Lõplik käesoleva aasta valim moodustus valimite s_{I_1} ja $s_{I_{01}}$ liitmisel. Teise aasta valimi jaoks tuleks võtta uus valim s_{I_2} ning lisada sellele 43% s_{I_1} asutustest.

Näide 7.2.2. Olgu soovitud valimimaht 100 asutust. Et saada rotatisooni 70%, peab meie valimis olema 70 uut asutust ja 30 asutust eelmise aasta valimist. Kui eelmise aasta valim oli sama moodi üles ehitatud, siis ei soovi me enam valimisse neid 30 asutust, mis olid valimis üleeelmisel aastal. Seega valime 30 asutust eelmise aasta 70'st asutusest ehk $\frac{30}{70} \simeq 43\%$. Lõplik valim on meil seega ikka 70+30=100 asutust.

Lõplik valim moodustus isikutest, sest klastervaliku tõttu esitavad kõik valimisse sattunud asutused andmed oma kõikide meditsiinitöötajate kohta ehk kaasavad nad valimisse. Kuna asutused freimis on järjestatud töötajate arvu järgi ning asutusi on palju ($N_I = 742$), siis saadakse igal aastal ligikaudu sama struktuuriga valim. Sarnaselt püsijuhuarvude meetodile võiks ka siin kaaluda freimist valimite järjestikust võtmist. Nt kui esimesel aastal alustati valimi võtmist juhuslikust objektist $r = 3$, siis järgmisel aastal fikseeritakse $r = 4$. Selle meetodi juhuslikkus vajab edaspidi põhjalikumat uurimist.

Ehkki viimane disain võib tunduda keeruline, on sellele keskmise hinnangu arvutamine üpriski lihtne:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{\sum_s w_i y_i}{N} = \frac{\sum_{s_1} w_i^1 y_i + \sum_{s_{01}} w_i^{01} y_i}{N}, \qquad (22)$$

kus s_1 ja s_{01} on isikutest koosnevad valimid ning w_i^1 ja w_i^{01} neile vastavad disainikaalud.

7.3. Valimi- ja kihtide mahtude määramine

Iga valikuuringu korral tuleb otsustada, kui suurt valimit tuleb võtta. TAI poolt määrati, et disainide sobivuse testimisel oleks valimimahuks 20% üldkogumist. Tunnuste 'ID' ja 'haigla nimi' järgi selgus, et osad isikud töötavad samas asutuses mitmel ametikohal. Kuna TAI arvestab ühe isiku erinevaid ametikohti eraldi objektidena, siis sedasi käsitleti neid ka käesolevas töös. Kuna valimi võtmisel ei ole arvutustesse kaasatavate ridade (st indikaatortunnuse välja põhitunnipalk väärtuste) koguarv teada, siis tuleks

arvutustest välja jäetavaid väärtusi käsitleda kui mittevastamist. Käesolevas töös eemaldati lihtsuse huvides need read enne valimi võtmist.

Võrdelise ja Neymani paigutuse valimimahud leiti valemitega (9)-(10) ning on välja toodud Tabelis 2. Valemis (9) kasutatud standardhälbed on välja toodud Lisas 3.

Tabel 2. Valimimahud võrdelise ja Neymani paigutusega

	Kokku	Valimimaht võrdelise paigutusega	Valimimaht Neymani paigutusega
Üldkogum	12588	2518	2518
Arstid	4366	873	1803
Õed	7794	1559	674
Ämmaemandad	428	86	41

Tabelist on näha, et Neymani paigutuse korral suureneb arstide osakaal valimis. Selle põhjuseks on uuritava tunnuse suur hajuvus arstide üldkogumis.

Süstemaatilisele klastervalikule baseeruva disaini korral kaasati valimisse 20% asutustest. Kokku oli meie andmestikus 742 asutust, seega kaasati valimisse 148 asutust. Arvestades eelmise aasta valimist võetava osaga, moodustus valim 104 nn „uuest asutusest“ ja 44 „eelmisel aastal kaasatud“ asutusest.

7.4. Meetodite rakendamine andmestikule

LJKV rakendamiseks omistati esmalt igale üldkogumi objektile ühtlasest jaotusest juhuslik suurus ning jagati kihtideks (Lisa 1.3). Seejärel sorteeriti kihid juhuslike arvude järgi ning võeti igas kihis valim püsijuhuarvude meetodiga. Kasutatud valimimahud on Tabelis 1. Lõpliku valimi saamiseks ühendati kihtide valimid. Keskmise ja dispersiooni hindamiseks kasutati valemeid (7) ja (8). Sedasi talitati nii võrdelise (Lisa 1.4) kui ka Neymani paigutuse korral (Lisa 1.5).

Süstemaatilisel klastervalikul baseeruva disainiga valimi võtmiseks moodustati asutuste freim ning märgiti ära ka tervishoiutöötajate arv igas asutuses (mitmel ametikohal töötavad isikud on loetud mitmekordselt) (Lisa 1.6). Seejärel järjestati asutused nende töötajate arvu järgi ning võeti valim punktis 7.2 kirjeldatud süstemaatilise klastervalikuga, mis arvestab rotatsiooni (Lisa 1.7). Keskmine hinnati valemiga (22) ning dispersioon *jackknife* meetodiga (valem (14)).

Ühe valimi põhjal saadud tulemused on välja toodud Tabelis 3.

Tabel 3. Põhitunnipalga hinnangud (tegelik keskmine on 6,5884)

Valiku meetod	Keskmise hinnang, $\hat{\bar{Y}}$	Standardhälbe hinnang, $\sqrt{\hat{V}\left(\hat{\bar{Y}}\right)}$	Hinnangu suhteline viga $\frac{\sqrt{\hat{V}(\hat{\bar{Y}})}}{\hat{\bar{Y}}}$
LJKV võrdelise paigutusega	6,9549	0,0544	0,0078
LJKV Neymani paigutusega	7,0592	0,0511	0,0072
Süstemaatilisel klastervalikul baseeruv disain (valim 1)	5,8458	0,2598	0,0444
Süstemaatilisel klastervalikul baseeruv disain (valim 2)	9,0588	0.0964	0,0106

Süstemaatilisel klastervalikul baseeruva disaini valimite 1 ja 2 põhjal on näha, et selle disaini keskmise hinnang võib väga palju varieeruda. Hinnangute täpsuse parandamiseks raknedati valimis regressioonhinnangut (20), kus abitunnustena kasutati töötajate ametigruppe ja vanuseid. Hinnangu dispersioon on leitud valemiga (21) . Saadud tulemused on välja toodud Tabelis 4, kood on Lisas 1.8.

Tabel 4. Põhitunnipalga regressioonhinnangud (tegelik keskmine on 6,5884)

Valiku meetod	Keskmise hinnang, $\hat{\bar{Y}}_{greg}$	Standardhälbe hinnang, $\sqrt{\hat{V}\left(\hat{\bar{Y}}_{greg}\right)}$	Hinnangu suhteline viga $\frac{\sqrt{\hat{V}(\hat{\bar{Y}}_{greg})}}{\hat{\bar{Y}}_{greg}}$
LJKV võrdelise paigutusega	6,9480	0,0539	0,0078
LJKV Neymani paigutusega	7,0834	0,0514	0,0073
Süstemaatilisel klastervalikul baseeruv disain (valim 1)	6,6674	0,0449	0,0067
Süstemaatilisel klastervalikul baseeruv disain (valim 1)	6,4759	0,0315	0,0049

Saadud tulemuste põhjal ei saa öelda, milline disain on parim, sest saadud hinnang sõltub realiseerunud valimist. Küll aga on näha, et hinnangu täpsus (suhteline viga) LJKV nii võrdelise kui ka Neymani paigutusega jäi ligikaudu samaks kui sellele

rakendati regressioonhinnangut. Süstemaatilisel klastervalikul baseeruva disaini hinnang paranes aga märgatavalt.

7.5. Tulemuste võrdlemine üle simulatsioonide

Erinevate disainide hinnagute täpsust kontrolliti simuleerimise teel. See tähendab, et võeti 1000 valimit iga vaadeldava disainiga ning leiti uuritava tunnuse keskmise hinnangud, hinnangute standard hälbed ja suhtelised vead üle 1000 valimi.

Simulatsioonide keskmine avaldub kujul:

$$\bar{\hat{Y}} = \frac{\sum_{i=1}^{m} \hat{\bar{Y}}_i}{m}, \tag{23}$$

kus m on erinevate valimite genereerimise arv ja $\hat{\bar{Y}}_i$ on i'nda valimi keskmise hinnang.

Üle simulatsioonide leitud standardhälve avaldub kujul:

$$\sqrt{\hat{V}\left(\bar{\hat{Y}}\right)} = \frac{1}{N}\sqrt{\hat{V}\left(\hat{Y}\right)} = \frac{1}{N}\sqrt{\frac{\sum_{i=1}^{m}\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}{m-1}}, \tag{24}$$

kus $\bar{\hat{Y}}$ on valimite kogusumma hinnangute keskmine üle m genereeritud valimi.

Kuna püsijuhuarvude meetod on sisuliselt lihtsalt üks meetod LJKV teostamiseks, siis lihtsuse huvides võeti simulatsiooni valimid SAS'i sisseehitatud meetodiga *surveyselect*. Süstemaatilisel klastervalikul baseeruva disaini simulatsioonil võeti valimid eelnevalt kirjeldatud meetodiga.

Keskmine üle simulatsioonide arvutati valemiga (23), kus valimi keskmise hinnang leiti LJKV puhul valemiga (7) ning süstemaatilisele klastervalikule konstrueeritud disaini puhul valemiga (22). Standardhälve leiti valemiga (24). Saadud tulemused on välja toodud Tabelis 5 ja kood Lisas 1.9.

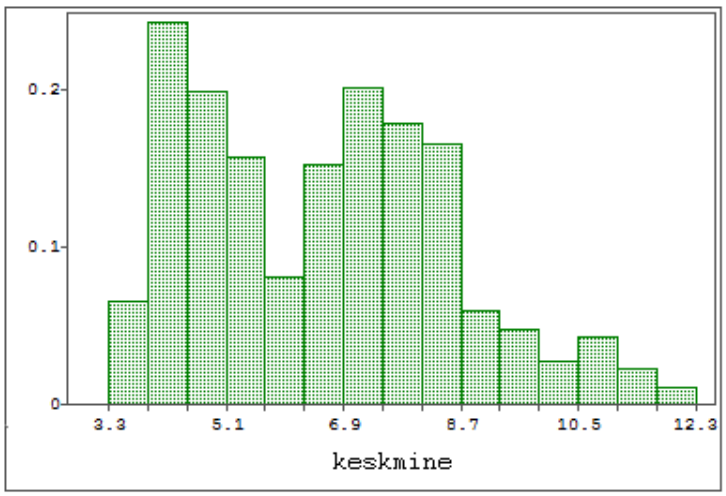
Tabel 5. Põhitunnipalga hinnangud üle 1000 valimi (tegelik keskmine on 6,5884)

Valiku meetod	Keskmise hinnang, $\bar{\hat{\bar{Y}}}$	Standardhälbe hinnang, $\sqrt{\hat{V}(\hat{\bar{Y}})}$	Hinnang suhtelisele veale, $\frac{\sqrt{\hat{V}(\hat{\bar{Y}})}}{\hat{\bar{Y}}}$
Võrdelise paigutusega LJKV üle simulatsoonide	6,5915	0,0467	0,0071
Neymani paigutusega LJKV üle simulatsoonide	6,5898	0,0337	0,0051
Süstemaatilisel klastervalikul baseeruv disain üle simulatsioonide	6,6134	1,9776	0,2990

Tabeli 5 tulemustest on näha, et hinnangud sellise valimimahu juures LJKV nii võrdelise kui ka Neymani paigutusega on hea täpsusega. Ka keskmise hinnangud on väga lähedased tegelikule väärtusele. Kuna suhtelise vea hinnangud on mõlemal juhul väga väikesed, siis ei ole põhjust eelistada Neymani paigutust võrdelisele paigutusele. Pealegi tuleb arvestada, et Neymani paigutus, mis on hea põhitunnipalga hindamiseks, ei pruugi olla sobiv teiste uuritavate tunnuste hindamiseks.

Süstemaatilisele klastervalikule põhineva disaini suhteline viga tuli kahjuks üpriski suur (0,2990). Kuna töötajate arvult suurima ja suuruselt viienda asutuse töötajate arvud on vastavalt 1799 ja 464 (Lisa 4), siis sõltuvalt juhusliku arvu k valikust, erinevad meie valimimahud erinevatel valimitel väga palju. Seda võib aidata ühtlustada suurte haiglate kaasamine igal aastal ning valimi moodustamine vaid väiksematest haiglatest, kuid käesoleva töö raames seda ei kontrollitud.

Üllatav on ka see, et süstemaatilisele klastervalikule põhineva disaini standardhälve üle simulatsioonide on palju suurem kui üle ühe valimi (vastavalt 1,9776 ja 0,2598). Selle valimimeetodi hinnangud varieeruvad tõepoolest väga palju. Selle kirjeldamiseks on välja toodud keskmiste histogramm (Joonis 4).



Joonis 4. Histogramm süstemaatilisele klastervalikule põhineva disaini keskmisele

Hinnanguid võib aidata parandada abiinformatsiooni kasutamine, mis on leitav kasutatavatest registritest. Seega leiti regresioonhinnang põhitunnipalga arvutamiseks. Nagu ennemgi, kasutati abiinformatsioonina töötajate ametigruppe ja vanust. Simulatsioonide keskmine arvutati valemiga (23), kus valimi keskmise hinnang ühel sammul leiti valemiga (20) ning standardhälve valemiga (24). Regressioonhinnangud simulatsioonidele on toodud Tabelis 6 ja kood Lisas 1.10.

Tabel 6. Regressioonhinnangud põhitunnipalgale (tegelik keskmine on 6,5884)

Valiku meetod	Keskmise hinnang, $\bar{\hat{\bar{Y}}}_{greg}$	Standardhälbe hinnang, $\sqrt{\hat{V}(\hat{\bar{Y}}_{greg})}$	Hinnang suhtelisele veale, $\frac{\sqrt{\hat{V}(\hat{\bar{Y}}_{greg})}}{\bar{\hat{\bar{Y}}}_{greg}}$
Regressioonhinnang võrdelise paigutusega LJKV'le üle simulatsioonide	6,5916	0,0466	0,0071
Regressioonhinnang Neymani paigutusega LJKV'le üle simulatsioonide	6,5900	0,0338	0,0051
Regressioonhinnang süstemaatilisel klastervalikul baseeruvale disainile üle simulatsioonide	6,6223	0,2486	0,0375

Võrreldes tabelite 5 ja 6 tulemusi, on näha, et süstemaatilise klastervaliku hinnang paranes abiinformatsiooni kasutamise tõttu märgatavalt (suhteline viga paranes 0,2990'lt 0,0375'le). LJKV hinnang peaaegu ei muutunud. See võib olla seetõttu, et kihtvaliku teostamisel on juba kasutatud osa abiinformatsioonist kihtide moodustamiseks (ametigruppe).

Kokkuvõte

Käesoleva töö eesmärgiks oli võrrelda kahte valimi võtmise meetodit valikuuringu teostamiseks tervishoiutöötajate palgaandmete uurimiseks. Nendeks olid lihtne juhuslik kihtvalik (LJKV), kus kihid moodustusid töötajate ametigruppidest, ja süstemaatiline klastervalik, kus klastriteks olid asutused, mis pakuvad tervishoiu teenust. Andmestikuna kasutati Tervise Arengu Instituudi (TAI) poolt kogutud aruande „Tervishoiutöötajate tunnipalk“ 2013. aasta andmed ning keskenduti vaid põhitunnipalga hindamisele.

Valimite võtmisel tuli arvestada rotatsiooniga 70%. Selle saavutamiseks kasutati LJKV korral püsijuhuarvude meetodit ning süstemaatilise kihtvaliku jaoks pakuti uus disain, mis koosnes 70% ulatuses uutest objektidest ning 30% valimist moodustas eelmisel aastal kaasatud objektidest võetud valim. Uue meetodi korral pakuti välja valemid keskmise hindamiseks ning hinnangu standardvea arvutamiseks.

Valimimaht määrati TAI poolt 20% üldkogumist. Meie andmestikul moodustus LJKV puhul valim 2518 isikust ning nende seas oli oluline määrata valimi paigutus kihtide vahel, et uuritava tunnuse standardhälve tuleks minimaalne. Selleks võrreldi kahte valimi paigutamise meetodit: võrdelist ja Neymani paigutust. Süstemaatilisel klastervalikul põhineva disaini valim moodistus 148 asutusest.

Hinnangute kontrollimiseks viidi läbi 1000 simulatsiooni hinnangute arvutamiseks ning leiti hinnangute suhtelised vead. Selgus, et meie valimi mahu juures LJKV võrdelisel paigutusel ja Neymani paigutusel väga suurt erinevust ei ole ning hinnangud on üpriski täpsed. Arvestades, et valimi võtmise ajal uuritava tunnuse standardhälvet teada ei ole (seda kasutab Neymani valem) ning põhitunnipalga suhtes hea paigutus ei pruugi olla hea teiste uuritavate tunnuste suhtes, võiks edaspidi kasutada pigem võrdelist paigutust. Süstemaatilisele klastervalikule põhineva disaini hinnang suhtelisele veale on märgatavalt halvem ja seda disaini ei saa soovitada. Küll võib seda proovida edaspidi modifitseerida kaasates alati valimisse suured asutused ning valimit võtta väiksemate asutuste hulgast. Kuid see meetod vajaks lisauurimist.

Regressioonhinnangu kasutamisel paranes kõige rohkem süstemaatilisele klastervalikul põhineva disaini hinnang. LJKV võrdelise paigutusega ja Neymani paigutusega

hinnangud jäid samadeks, sest nende valimi võtmisel on juba kasutatud regressioonhinnagu abitunnust ametigrupp.

Edaspidi tuleks uurida hinnangute täpsust ka teistele uuritavatele tunnustele. Antud töös uuritud tunnuse ’põhitunnipalk valvega’ põhjal soovitaksime edaspidi kasutada lihtsat juhuslikku kihtvalikut võrdelise paigutusega, sest väga suurt erinevust Neymani paigutusega ei ole. Regressioonhinnang vanust ja ametigruppi abiinformatsioonina kasutades hinnangu täpsust ei parandanud.

Kasutatud kirjandus

C. Bruch, R. Münnich, S. Zins (2011). „Variance *Estimation for Complex Surveys*“.

B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, P. S. Kott (1995). „*Business Survey Methods*“, John Wiley & sons. Inc.

A. Poolakese (2013). „Tervishoiutöötajate tunnipalk, märts 2013“.

C.-E. Särndal, B. Swensson, J. Wretman (1992). „*Model Assisted Survey Sampling*“, Springer-Verlag.

I. Traat (2012). „Valikuuringute teooria edasijõudnutele“.

I. Traat, J. Inno (1997). „Tõenäosuslik valikuuring“, TÜ kirjastus.

I. Traat, N. Lepik (2013). „Valikuuringute teooria I“.

Lisad

Lisa 1. SAS'i kood

Lisa 1.1. Ämmaemandate erialakoodi ümberkodeerimine

```
data Kodeeritud;
set Loputoo.Algandmed;
if Ametigrupp1=5 then delete;
if _5=2222 then Ametigrupp1=4;
else Ametigrupp1=Ametigrupp;
run;
```

Lisa 1.2. Põhitunnipalga arvutustes mittekasutatud andmeridade kustutamine

```
data YK;
set Kodeeritud;
if v2lja_pohitunnipalk=1 then delete;
run;
```

Lisa 1.3. Üldkogumi objektidele juhuslike arvude genereerimine ja kihtideks jagamine

```
proc sort data=YK;
by Asutuse_nimi Id;
run;
*Genereerin igale reale juhusliku arvu;
data Loputoo.YK;
set YK;
U=ranuni(123);
run;
proc sort data=Loputoo.YK;
by Ametigrupp1;
run;
*Leian uuritavate kihtide suurused;
proc sql;
create table Loputoo.Kihisumma as select
Ametigrupp1, count(u) as _total_
from Loputoo.YK
group by Ametigrupp1;
quit;
*Lagan üldkogumi kihtideks;
proc sql;
create table Arstid as
select * from Loputoo.YK
where ametigrupp1=1;
quit;
proc sql;
create table Oed as
select * from Loputoo.YK
where ametigrupp1=3;
quit;
proc sql;
create table Ammaemandad as
select * from Loputoo.YK
where ametigrupp1=4;
quit;
```

Lisa 1.4. Võrdelise paigutusega LJKV valimi võtmine püsijuhuarvude meetodiga

```
*Leian LJKV võrdelise paigutuse valimimahud kihtides;
proc sql;
create table Loputoo.LJKV_vordeline_valimimaht as select
Ametigrupp1, _total_*0.2 as Valimi_maht
from Loputoo.Kihisumma;
quit;
*Võtan võrdelise paigutusega valimi LJKV;
proc sort data=Loputoo.YK;
by Ametigrupp1 U;
run;
data Arstid_valim;
set Arstid;
if (_N_ LE 873) then output;
run;
data Oed_valim;
set Oed;
if (_N_ LE 1559) then output;
run;
data Ammaemandad_valim;
set Ammaemandad;
if (_N_ LE 86) then output;
run;
*Liidan kihid kokku lõplikuks valimiks ja lisan disainikaalud;
proc sql;
create table LJKV as
select * from Arstid_valim
union
select * from Oed_valim
union
select * from Ammaemandad_valim;
quit;
data loputoo.LJKV_vordeline;
set LJKV;
if Ametigrupp1=1 then W=4366/873;
if Ametigrupp1=3 then W=7794/1559;
if Ametigrupp1=4 then W=428/86;
run;
proc sort data=Loputoo.LJKV_vordeline;
by Ametigrupp1;
run;
*Leian keskmise ja dispersiooni;
proc surveymeans data=Loputoo.LJKV_vordeline
total=Loputoo.Kihisumma
mean var;
stratum Ametigrupp1;
var Pohitunnipalk_valvega;
weight W;
domain Ametigrupp1;
run;
```

Lisa 1.5. Neymani paigutusega LJKV valimi võtmine püsijuhuarvude meetodiga

```
*Leian Neymani paigutuse valimimahud kihtides;
proc sql;
create table Loputoo.LJKV_Neymani_valimimaht as select
2518*4366*4.3076/(4366*4.3076+7794*0.9020+428*0.9973) as Arst,
2518*7794*0.9020/(4366*4.3076+7794*0.9020+428*0.9973) as Ode,
2518* 428*0.9973/(4366*4.3076+7794*0.9020+428*0.9973) as Ammaemand
from Loputoo.YK;
quit;
```

```
*Valimi võtmine Neymani paigutusega;
data Arstid_valim_Neyman;
set Arstid;
if (_N_ LE 1803) then output;
run;
data Oed_valim_Neyman;
set Oed;
if (_N_ LE 674) then output;
run;
data Ammaemandad_valim_Neyman;
set Ammaemandad;
if (_N_ LE 41) then output;
run;
*Liidan kihid kokku lõplikuks valimiks ja lisan disainikaalud;
proc sql;
create table LJKV_Neyman as
select * from Arstid_valim_Neyman
union
select * from Oed_valim_Neyman
union
select * from Ammaemandad_valim_Neyman;
quit;
data Loputoo.LJKV_Neyman;
set LJKV_Neyman;
if Ametigrupp1=1 then w=4366/1803;
if Ametigrupp1=3 then w=7794/674;
if Ametigrupp1=4 then w=428/41;
run;
*Leian keskmise ja dispersiooni;
proc surveymeans data=Loputoo.LJKV_Neyman
total=Loputoo.Kihisumma mean var;
stratum Ametigrupp1;
var Pohitunnipalk_valvega;
weight W;
domain Ametigrupp1;
run;
```

Lisa 1.6. Asutuste freimi moodustamine

```
proc sort data=Loputoo.YK;
by Id Asutuse_nimi;
run;
proc sql;
create table Loputoo.Asutus as
select Asutuse_nimi, count(Asutuse_nimi) as Tootajaid,
sum(pohitunnipalk_valvega) as Summa
from Loputoo.YK
group by Asutuse_nimi;
quit;
```

Lisa 1.7. Süstemaatilisele klastervalikule konstrueeritud valimi võtmine

```
proc sort data=Loputoo.Asutus;
by descending Tootajaid  Asutuse_nimi;
run;
*Võtan valimi 0 kui eelmise aasta valimi (14% üldkogumist);
proc surveyselect
data=Loputoo.Asutus
method=sys
n=104
out=ValimSYS0;
run;
```

```
*Lisan disainikaalu;
data ValimSYS0_;
set ValimSYS0;
w=754/148;
run;
*Sellest 43% võtan ka sel aastal valimisse;
proc surveyselect
data=ValimSYS0_
method=sys
n=44
out=ValimSYS01;
run;
*Võtan 14% selle aasta valimi;
proc surveyselect
data=Loputoo.Asutus
method=sys
n=104
out=ValimSYS1;
run;
*Lisan disainikaalu;
data ValimSYS1_;
set ValimSYS1;
w=754/148;
run;
*Liidan valimid kui selle aasta valim;
proc sql;
create table Loputoo.Sys as
select * from ValimSYS01
union
select * from ValimSYS1_;
quit;
*Leian keskmise hinnangu;
proc sql;
create table Loputoo.SYS_keskmine as select
sum(Summa*w)/12588 as keskmine
from Loputoo.SYS;
quit;
*Leian dispersiooni hinnangud Jackknife meetodiga;
proc sql;
create table Loputoo.SYS_isikud as
select *
from Loputoo.YK as YK, Loputoo.SYS as SYS
where YK.Asutuse_nimi=SYS.Asutuse_nimi;
quit;
proc surveymeans data=Loputoo.SYS_isikud varmethod=jackknife;
cluster Asutuse_nimi;
weight w;
var pohitunnipalk_valvega;
run ;
```

Lisa 1.8. Regressioonhinnangu leidmine kogusummale

```
*Keskmise hinnangu saamiseks jagan kogusumma hinnangu üldkogumi mahuga
(12588);
*GREG hinnang LJKV võrdelise paigutusega;
proc surveyreg data=Loputoo.LJKV_vordeline
total=Loputoo.Kihisumma;
strata Ametigrupp1 / list;
class  Ametigrupp1;
model  pohitunnipalk_valvega = _4 Ametigrupp1 /solution;
weight w;
estimate 'Pohitunnipalk valvega'
```

```
INTERCEPT 12588 _4 598018 Ametigrupp1 4366 7794 428 ;
run;
*GREG hinnang LJKV Neymani paigutusega;
proc surveyreg data=Loputoo.LJKV_Neyman
total=Loputoo.Kihisumma;
strata Ametigrupp1 / list;
class  Ametigrupp1;
model  pohitunnipalk_valvega = _4 Ametigrupp1 /solution;
weight w;
estimate 'Pohitunnipalk valvega'
INTERCEPT 12588 _4 598018 Ametigrupp1 4366 7794 428 ;
run;
*GERG hinnang süstemaatilisele klastervalikult baseeruvale disainile;
proc surveyreg data=Loputoo.Sys_isikud
total=12588;
class  Ametigrupp1;
model  pohitunnipalk_valvega = _4 Ametigrupp1 /solution;
weight w;
estimate 'Pohitunnipalk valvega'
INTERCEPT 12588 _4 598018 Ametigrupp1 4366 7794 428 ;
run;
```

Lisa 1.9. Simulatsioon üle 1000 valimi

```
*#####################################################################;
*Simulatsiooni LJKV võrdelise paigutusega;
*#####################################################################;
*Võtan 1000 valimit;
proc sort data=Loputoo.YK;
by Ametigrupp1 U;
run;
proc surveyselect
data=Loputoo.YK
n=(873 1559 86)
reps=1000
out=Loputoo.LJKV_vordeline_simulatsioon;
strata Ametigrupp1;
run;
*Leian iga valimi keskmise;
proc sql;
create table Loputoo.LJKV_vord_sim_kesk As select
(4366/ 873*sum((Pohitunnipalk_valvega)*(Ametigrupp1=1))+
 7794/1559*sum((Pohitunnipalk_valvega)*(Ametigrupp1=3))+
  428/  86*sum((Pohitunnipalk_valvega)*(Ametigrupp1=4)))/(12588)as
Keskmine,
    1/873*sum((Pohitunnipalk_valvega)*(Ametigrupp1=1)) as Arst,
   1/1559*sum((Pohitunnipalk_valvega)*(Ametigrupp1=3)) as Ode,
     1/86*sum((Pohitunnipalk_valvega)*(Ametigrupp1=4)) as Ammaemand
from Loputoo.LJKV_vordeline_simulatsioon
group by Replicate;
quit;
*#####################################################################;
*Simulatsiooni LJKV Neymani paigutusega;
*#####################################################################;
*Võtan 1000 valimit;
proc surveyselect
data=Loputoo.YK
n=(1803 674 41)
reps=1000
out=Loputoo.LJKV_Neyman_simulatsioon;
strata Ametigrupp1;
run;
```

```
*Leian iga valimi keskmise;
proc sql;
create table Loputoo.LJKV_Neyman_sim_kesk As select
(4366/1803*sum((pohitunnipalk_valvega)*(Ametigrupp1=1))+
 7794/ 674*sum((pohitunnipalk_valvega)*(Ametigrupp1=3))+
  428/  41*sum((pohitunnipalk_valvega)*(Ametigrupp1=4)))/(12588)as
Keskmine,
    1/1803*sum((pohitunnipalk_valvega)*(Ametigrupp1=1)) as Arst,
    1/ 674*sum((pohitunnipalk_valvega)*(Ametigrupp1=3)) as Ode,
    1/  41*sum((pohitunnipalk_valvega)*(Ametigrupp1=4)) as Ammaemand
from Loputoo.LJKV_Neyman_simulatsioon
group by Replicate;
quit;
*########################################################################;
*Simulatsioon süstemaatilisele klastervalikule baseeruvale disainile;
*########################################################################;
*Võtan eelmiste aastate valimi;
proc surveyselect
data=Loputoo.Asutus
method=sys
n=104
reps=1000
out=sys_sim0;
run;
*Võtan neist järgmise aasta valimi;
proc surveyselect
data=sys_sim0
method=sys
n=44
out=sys_sim01;
stratum replicate;
run;
*Võtan käesoleva aasta valimid;
proc surveyselect
data=Loputoo.Asutus
method=sys
n=104
reps=1000
out=sys_sim1;
run;
*Liidan need kokku üheks andmestikuks;
proc sql;
create table Loputoo.Sys_sim as
select replicate, asutuse_nimi, tootajaid, summa from sys_sim01
union ALL
select replicate, asutuse_nimi, tootajaid, summa from sys_sim1;
quit;
*Lisan disainikaalud;
proc sql;
create table Loputoo.Sys_sim_isikud as
select *, 742/148 as w from Loputoo.YK as YK, Loputoo.Sys_sim as sim
where YK.Asutuse_nimi=sim.Asutuse_nimi;
quit;
*Leian keskmised;
proc sql;
create table Loputoo.SYS_sim_keskmine as select
1 as Meetod, sum(pohitunnipalk_valvega*w)/12588 as keskmine
from Loputoo.Sys_sim_isikud
group by Replicate;
quit;
```

Lisa 1.10. Regressioonhinnang simulatsioonile üle 1000 valimi

```
*########################################################################;
*LJKV võrdelise paigutusega;
*########################################################################;
*Defineerin abitunnused;
data Loputoo.LJKV_vordeline_reg;
set Loputoo.LJKV_vordeline_simulatsioon;
x1=1;
if Ametigrupp1=1 then x2=1;
else x2=0;
if Ametigrupp1=3 then x3=1;
else x3=0;
if Ametigrupp1=4 then x4=1;
else x4=0;
run;
*Leian kogusumma hinnangu;
proc iml;
start;
t={12588 598018 4366 7794}; *Abiinfo kogusummad;
a=1;
regh=(1:1000);
do while (a<=1000);
use Loputoo.LJKV_vordeline_reg where (Replicate=a);
read all var {x1 _4 x2 x3} into X;*Abitunnuste fail;
read all var {pohitunnipalk_valvega} into Y; *Uuritava tunnuse fail;
read all var {SamplingWeight} into W;
XL=W#X;*Korrutab W-ga kõik X veeerud;
th=XL[+,];*Veerusummad ongi kogusummade disain-kaalutud hinnangud;
g=1+X*inv(t(XL)*X)*t(t-th);
Y1=W#g#Y;
regh[a]=Y1[+,];
a=a+1;
end;
create Loputoo.Reg1 var {regh}; *Loon tulemuste faili;
append;        * Kirjutan tulemused faili;
close Loputoo.Reg1; *Sulen faili;
finish;
run;
quit;
*########################################################################;
*LJKV Neymani paigutusega;
*########################################################################;
data Loputoo.LJKV_Neyman_reg;
set Loputoo.LJKV_Neyman_simulatsioon;
x1=1;
if Ametigrupp1=1 then x2=1;
else x2=0;
if Ametigrupp1=3 then x3=1;
else x3=0;
if Ametigrupp1=4 then x4=1;
else x4=0;
run;
proc iml;
start;
t={12588 598018 4366 7794};
a=1;
regh=(1:1000);
do while (a<=1000);
use Loputoo.LJKV_Neyman_reg where (Replicate=a);
read all var {x1 _4 x2 x3} into X;
```

```
read all var {pohitunnipalk_valvega} into Y;
read all var {SamplingWeight} into W;
XL=W#X;
th=XL[+,];
g=1+X*inv(t(XL)*X)*t(t-th);
Y1=W#g#Y;
regh[a]=Y1[+,];
a=a+1;
end;
create Loputoo.Reg2 var {regh};
append;
close Loputoo.Reg2;
finish;
run;
quit;
*########################################################################;
*Süstemaatilisel disainil baseeruv disain;
*########################################################################;
data Loputoo.SYS_reg;
set Loputoo.SYS_sim_isikud;
x1=1;
if Ametigrupp1=1 then x2=1;
else x2=0;
if Ametigrupp1=3 then x3=1;
else x3=0;
if Ametigrupp1=4 then x4=1;
else x4=0;
run;
proc iml;
start;
t={12588 598018 4366 7794};
a=1;
regh=(1:1000);
do while (a<=1000);
use Loputoo.SYS_reg where (Replicate=a);
read all var {x1 _4 x2 x3} into X
read all var {pohitunnipalk_valvega} into Y;
read all var {w} into W;
XL=W#X;
th=XL[+,];
g=1+X*inv(t(XL)*X)*t(t-th);
Y1=W#g#Y;
regh[a]=Y1[+,];
a=a+1;
end;
create Loputoo.Reg3 var {regh};
append;
close Loputoo.Reg3;
finish;
run;
quit;
```

Lisa 2. Ametiala koodid

AMETIKOHA NIMETUS	KOOD	AMETIKOHA NIMETUS	KOOD
ARSTID JA ARST-RESIDENDID		Diabeediõde	222108
Erialase spetsialiseerumiseta arsti töö	221101	Geriaatriaõde	222109
Abiarst	22110101	Lasteõde	222110
Perearst	221102	Nakkustõrjeõde	222111
Kooliarst	221103	Onkoloogiaõde	222112
Anestesioloogia ja intensiivravi arst	221201	Operatsiooniõde	222113
Dermatoveneroloog	221202	Pulmonoloogiaõde	222114
Endokrinoloog	221203	Taastusraviõde	222115
Erakorralise meditsiini arst	221204	Koduõde	222116
Gastroenteroloog	221205	Kooliõde	222117
Günekoloog	221206	Töötervishoiuõde	222118
Hematoloog	221207	Pereõde	222119
Infektsioonhaiguste arst	221208	Psühhiaatriaõde	222120
Kardioloog	221209	Ämmaemand	2222
Kardiovaskulaarkirurg	221210	Abiämmaemand	222201
Kliiniline immunoloog	221211	**TEISED MEDITSIINILISED TÖÖTAJAD**	
Kliiniline mikrobioloog	221212	Haiglaapteeker farmatseudi kutsega	226201
Kohtuarst-ekspert	221213	Haiglaapteeker proviisori kutsega	226202
Laboriarst	221214	Tervisekaitse-, töötervishoiu- ja tööhügieenitippspetsialist	2263
Lastekirurg	221215	Füsioterapeut (meditsiinilise kõrgema eriharidusega)	2264
Meditsiinigeneetik	221216	Dieedi ja toitumise tippspetsialist	2265
Nefroloog	221217	Audioloog ja logopeed	2266
Neurokirurg	221218	Liikumisraviterapeut	226901
Neuroloog	221219	Liikumisravi spetsialist	226902
Oftalmoloog	221220	Loovterapeut	226903
Onkoloog (kiiritus ja keemiaravi)	221221	Muusikaterapeut	226904
Ortopeed	221222	Tegevusterapeut	226905
Otorinolarüngoloog	221223	Kliiniline psühholoog	226906
Patoloog	221224	Psühhoterapeut (kliinilise psühholoogi haridusega)	22690601
Pediaater	221225	Muu tervishoiu tippspetsialist	226907
Plastikakirurg	221226	Radioloogiatehnik	321101
Psühhiaater	221227	Abiradioloogiatehnik	32110101
Pulmonoloog	221228	Meditsiiniliste kuvamis- ja raviseadmete tehnik	321102
Radioloog	221229	Bioanalüütik	321201
Reumatoloog	221230	Meditsiini- ja patoloogialaborite tehnik	321202
Sisehaiguste arst	221231	Meditsiiniliste proteeside tehnik	321401
Taastusarst	221232	Hambaproteeside tehnik	321402
Torakaalkirurg	221233	Ämmaemanda abiline	3222
Töötervishoiuarst	221234	Hambaraviõde	325101
Uroloog	221235	Suuhügienist	325102

Üldkirurg	221236	Optik	325401
HAMBAARSTID JA HAMBAARST-RESIDENDID		Optometrist	325402
Hambaarst	226101	Füsioterapeut (TÜ kehalise kasvatuse haridusega)	325501
Ortodont	226102	Massöör	325502
Suu-ja näolõualuukirurg	226103	Hambaarsti eriala praktikant	325601
ÕENDUSTÖÖTAJAD		Meditsiiniülesvõtete protseduuride assistent	325602
Õde	222105	Kiirabitehnik	3258
Abiõde	22210501	Hooldaja	5321
Anesteesia-intensiivraviõde	222106	Koduõendus-hooldusteenus	5322
Erakorralise meditsiini õde	222107	Hooldustöötajad meditsiiniasutustes	5329

Lisa 3. Põhitunnipalga karakteristikud üldkogumi kihtides

Tabel 7. Arstide põhitunnipalga karakteristikud

Moments			
N	4380.0000	Sum Wgts	4380.0000
Mean	9.9053	Sum	43385.4300
Std Dev	5.2278	Variance	27.3301
Skewness	13.5310	Kurtosis	388.1476
USS	549426.332	CSS	119678.492
CV	52.7777	Std Mean	0.0790

Tabel 8. Õdede tunnipalga karakteristikud

Moments			
N	7827.0000	Sum Wgts	7827.0000
Mean	4.8413	Sum	37892.6500
Std Dev	1.1428	Variance	1.3060
Skewness	11.5961	Kurtosis	338.8542
USS	193669.360	CSS	10220.6666
CV	23.6053	Std Mean	0.0129

Tabel 9. Ämmaemandate tunnipalga karakteristikud

Moments			
N	430.0000	Sum Wgts	430.0000
Mean	4.8662	Sum	2092.4800
Std Dev	1.0493	Variance	1.1010
Skewness	3.4657	Kurtosis	23.6173
USS	10654.8110	CSS	472.3167
CV	21.5623	Std Mean	0.0506

Lisa 4. Tervishoiu teenust osutavate asutuste freim

Tabel 10. Tervishoiu teenust osutavate asutuste freim

	Asutuse nimi	Tervishoiutöötajate arv
1	Tartu Ülikooli Kliinikum, sihtasutus	1799
2	Põhja-Eesti Regionaalhaigla, sihtasutus	1486
3	Ida-Tallinna Keskhaigla, AS	1066
4	Lääne-Tallinna Keskhaigla, AS	795
5	Ida-Viru Keskhaigla, sihtasutus	464
6	Pärnu Haigla, sihtasutus	441
7	Tallinna Lastehaigla, sihtasutus	343
8	Narva Haigla, sihtasutus	328
9	Viljandi Haigla, sihtasutus	280
10	Rakvere Haigla, AS	212
...	...	...
733	Timmermann, AS	1
734	Töökeskkonna Haldus OÜ	1
735	Visus Pluss, OÜ	1
736	Veroniks Hermet OÜ	1
737	Voorman, OÜ	1
738	Vändra Tervis OÜ	1
739	Osaühing Liina Viitas	1
740	Õendusteenused OÜ	1
741	Õismäe Perearstikeskus OÜ	1
742	Üksikettevõtja Rahusoov Katrin	1

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kaidi Jõgi,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Lihtsa juhusliku kihtvaliku ja süstemaatilise klastervaliku võrdlemine tervishoiutöötajate põhitunnipalga uurimise näitel“, mille juhendaja on Natalja Lepik

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

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

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

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

Tartus, **02.06.2014**