T A R T U   Ü L I K O O L
MATEMAATIKA-INFORMAATIKATEADUSKOND

Matemaatilise statistika instituut

Anna Leontjeva

Tõenäosuslik valikuuring 2004. aasta perearstikeskuste majandusaruande näitel

Bakalaureusetöö (4 AP)

Juhendaja: Natalja Lepik, M. Sc.

TARTU 2007

Sisukord

Sissejuhatus

Käesoleva töö eesmärk on välja töötada sobilik disain tõenäosusliku valikuuringu läbiviimiseks. Töö on tellitud Sotsiaalministeeriumi Terviseinfo- ja analüüsiosakonna poolt ning uurimuse objektideks on Eesti meditsiinilised asutused, mis pakkuvad perearstide teenust. Aastatel 2002 kuni 2006 teostas Sotsiaalministeerium kõikseid uuringuid analüüsimaks perearstikeskuste majandustegevusest tulenevaid kulude liike. Sellised uuringud on olnud kulukad ja aeganõudvad ning otsustati üle minna valikuuringutele juhul, kui need annavad tulevikus häid hinnanguid huvipakkuvatele näitajatele (hinnangu nihke ja varieeruvuse mõttes).

Valikudisaini väljatöötamiseks on kasutada kõikse uuringu andmeid 2004. aasta eest, mis võimaldab võrrelda saadud hinnangud üldkogumi tegelikke parameetritega.

Antud töö koosneb kahest osast – teoreetilisest ning praktilisest. Töö teoreetilises osas seletatakse tõenäosusliku valikuuringu põhimõtted, mida kasutatakse hiljem praktilise ülesande lahendamiseks. Samuti antakse lühiülevaade levinumatest valikuuringu disainidest ning kasutatud metoodikast.

Töö praktilises osas tutvustatakse kõigepealt andmestikku, mille maht on 474 objekti ja mis sisaldab 116 tunnust. Seejärel kirjeldatakse ning leitakse valikudisaini jaoks vajalike suurusi (kihistav ja tausttunnused, valimimaht, jms). Antud töö raames keskendatakse kahe tunnuse hindamisele, milleks on pidev tunnus „kogukulu 2004. aasta eest“ ja kodeeritud tunnus „majandusliku asutuse õiguslik vorm“. Hinnanguid leitakse nii terve Eesti kohta kui ka maakonniti. Valikumeetodi tõhususe hindamise kriteeriumiks on hinnangute suhteline viga.

Valikuuringute rakendamises pikemas perspektiivis võib tekkida olukord, kus ühed ja samad perearstide keskused satuvad valimisse mitu aastat järjest ning sel juhul on nende vastamiskoormus liiga suur ja võib mõjutada vastuste täpsust. Selle probleemi vältimiseks on mitmeid meetodeid, mis võimaldavad igaaastaseid valimeid koordineerida. Üks nendest on nn *püsijuhuarvude meetod*. Töös kirjeldatakse selle meetodi põhimõtted, objektide valimisse võtmise algoritm ning rakendamise viis. Tähtsaks aspektiks on ka see, et püsijuhuarvude meetod aitab leida hinnanguid nii uute objektide tekkimisel (*sündide*), kuid ka oma tegevuse lõpetatud objektide (*surmade*) korral.

Peatükk I
Lühiülevaade tõenäosusliku valikuuringu teooriast

1 Tõenäosuslik valikuuring

1.1 Valikuuringu ajalugu

Valikuuringute aluste ajalugu algab 19.sajandi lõpust - 20.sajandi algusest, kuigi vanim viide valiku kasutamisele pärineb juba India eeposest Mahabharata. Suur osa valikuuringu teooriast oli ajendatud praktikas tekkinud probleemidest disainis ja valikuuringu analüüsis. Ka vastupidi: valikuuringu teooria avaldas mõju praktikale ja tihti viis vapustavatele tulemustele. Norra statistik A.N. Kaier (1897) oli esimene, kes kirjeldas valimi võtmist lõplikust loendist: see hiljem sai nimeks *representatiivne meetod*. Selle meetodi põhimõte seisnes selles, et valim võib peegeldada üldkogumi kas juhusliku või tasakaalustatud valikuga. 1920tes representatiivne meetod oli olnud laial kasutusel. Aastal 1924, tänu Rahvusvahelise Statistika Instituudile (ISI), loodi komisjon representatiivse meetodi uurimise eesmärgil. Selle komisjoni 1925. aasta raportis olid vaatluse alla võetud juhusliku valiku meetodi nii praktilised, kui ka teoreetilised aspektid. Antud raporti tähtsamaks osaks oli Bowley'i fundamentaalne töö juhuslikust kihtvalikust võrdelise paigutusega (Rao, 2005).

Oluliseks momendiks valikuuringu ajaloos sai Neymanni töö disainil põhinevatest meetoditest. Neymann tõestas nii teoreetiliselt kui ka praktiliste näidete abil, et juhuslikul kihtvalikul on eelis tasakaalustatud valiku ees, kuna see võimaldab kaasamistõenäosuste varieerimise abil reguleerida uuringu maksumust ja täpsust. Neymann näitas ka oma tulemused efektiivsuse ja optimaalse paigutuse kohta (Rao, 2005).

1930tes tekkis suur informatsioonivajadus ning see omakorda tõi esile valikuuringu selliseid eeliseid, nagu väiksem maksumus ja suurem kiirus. Tõenäosuslik valikuuring sai üleüldiselt aktsepteeritud ja laiali rakendatud praktikas.

1.2 Mõisted

Valikuuringute teooria kuulub statistiliste meetodite hulka, mis võimaldab valimi põhjal teha järeldusi üldkogumi kohta (järeldav statistika). Valikuuringud jagunevad empiirilisteks ja tõenäosuslikeks. Tõenäosusliku valiku korral on iga üldkogumi objekti puhul teada tema kaasamistõenäosus ehk tõenäosus sattuda valimisse. Empiirilise valiku korral kaasamistõenäosust ei ole teada. Kuna antud töö käigus tegeldakse vaid tõenäosusliku valikuga, siis edaspidi valimi all mõistetakse tõenäosusliku valimit.

Definitsioon 1. *Tõenäosuslikuks valikuks* nimetatakse niisugust valikut üldkogumist, mille korral:

1) saab defineerida kõigi võimalike valimite hulga, $S = \{...\}$;

2) iga valimi s jaoks on teada tema valikutõenäosus $p(s)$;

3) iga üldkogumi objekti jaoks on teada valimisse sattumise tõenäosus, mis on ka positiivne; (Särndal *et al.*, 1992).

Eristatakse *tagasipanekuga* (*TGA*) ja *tagasipanekuta* (*TTA*) valikuid. Esimesel juhul võib iga objekt sattuda valimisse mitmekordselt, viimasel aga mitte.

Definitsioon 2. *Loendiks* nimetatakse vahendit, mis võimaldab pääseda üldkogumi (ÜK) objektide juurde. Loend peab:

- identifitseerima igat ÜK objekti ning võimaldama valima neid vastavalt fikseeritud valikudisainile;
- võimaldama saama kontakti valitud ÜK elementidega; (Särndal *et al.*, 1992)

Selleks, et saavutada juhuslikkuse valimi võtmise korral, teostatakse tihti üldises kasutuses oleva ning lihtsalt rakendatava algoritmi. Levinuim algoritm on selline, mille korral viiakse läbi juhuslik eksperiment iga loendis oleva elemendi jaoks, otsustades kas element kaasatakse valimisse või mitte.

Definitsioonis 1 punktis kolm mainitud valimisse sattumise tõenäosust, millega see objekt kaasatakse valimisse antud disaini $p(s)$ korral, nimetatakse üldkogumi i-nda objekti *kaasamistõenäosuseks* π_i.

Teine oluline disaini karakteristik on üldkogumi i-nda objekti *valikutõenäosus* p_i, mis on defineeritud kui tõenäosus, millega seda objekti võidakse valida antud disaini ühel valikusammul (Traat ja Inno, 1997).

Objekti sattumist valimisse iseloomustab nn. *valikuindikaator*, mis on defineeritud iga üldkogumi objekti jaoks ja mis on oma loomu poolest juhuslik suurus järgmiste väärtustega (TTA valik):

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

Tagasipanekuga valiku korral defineeritakse valikuindikaatorit teisiti: I_i iseloomustab mitu korda on objekt i valimisse kaasatud. Vektorit $I = (I_1, I_2, ..., I_n)^T$ nimetatakse *valikuvektoriks*.

1.3 Valikudisain ja selle karakteristikud

Olgu $I = (I_1, I_2, ..., I_n)^T$ valikuvektor. Valikuvektori jaotust kirjeldab valikudisain, $I : p(s)$. Igat jaotust on tavaks iseloomustada karakteristikute abil. Valikuuringutes nimetatakse neid *disainikarakteristikuteks*. Kõige tähtsamad nendest on järgmised:

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

$E(I_i I_j)$ – 2. järku moment

$V(I_i) = E(I_i)^2 - (EI_i)^2$ – valikuindikaatori I_i dispersioon

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

ÜK objekti *i* kaasamistõenäosust π_i võib vaadelda, kui

$$\pi_i = P(i \in s) = P(I_i \geq 1) = \sum_{I_i \geq 1} p(s).$$

Kui tegemist on TTA disainiga, siis $\pi_i = P(I_i = 1) = E(I_i)$, s.t. 1. järku moment ehk objekti *i* oodatav valikute arv on võrdne kaasamistõenäosusega π_i, dispersioon avaldub kujul $V(I_i) = \Delta_{ii} = \pi_i(1 - \pi_i)$, ning kovariatsioon kujul $Cov(I_i, I_j) = \Delta_{ij} = E(I_i I_j) - E(I_i)E(I_j) = \pi_{ij} - \pi_i \pi_j$.

1.4 Üldkogumi parameetrite hindamine

Üheks tähtsaks parameetriks, mida hinnatakse valikuuringute teoorias on üldkogumi summa $t_y = \sum_U y_i$, kus $y = (y_1, y_2, ..., y_N)^T$ on uuritav tunnus. Mitmeid teisi huvipakkuvaid parameetreid saab esitada ÜK summa kaudu, näiteks ÜK keskmine $\bar{Y} = t_y / N$. Sageli ühe ja sama uuringu põhjal tahetakse hinnata erinevate osakogumite parameetreid (meeste ja naiste keskmine sissetulek, töötute arv maakonniti jms). Sellisel juhul võib defineerida igas osakogumis binaarset tunnust *z*, mis iseloomustab objekti kuuluvust osakogumisse:

$$z_i = \begin{cases} 1, kui \quad i \in U_d, \\ 0, vastasel \quad juhul. \end{cases}$$

Osakogumi U_d maht avaldub sellisel juhul järgmiselt:

$$t_z = \sum_U z_i.$$

Kui soovitakse teada saada osakogumi osakaalu ÜK-s, siis on selleks tunnuse z keskmine:

$$\bar{Z} = t_z/N.$$

Üheks parameetriks, mida samuti tehetakse sageli hinnata, on kahe kogusumma suhe:

$$R = t_y/t_u.$$

Ülalpool öeldust järeldub, et ÜK parameetrite hindamiseks peab oskama hindama ÜK summat. Järgmine teoreem kasutab selleks disainipõhist nihketa hinnangut (Traat *et al.*, 2001).

Teoreem (Üldine hindamisteoreem):

Nihketa hinnangufunktsiooniks ÜK summale $t_y = \sum_U y_i$ on järgmine statistik:

$$\hat{t}_y = \sum_U \frac{I_i y_i}{E(I_i)}, \tag{1}$$

millele vastav valimi hinnang on järgmine:

$$\hat{t}_y = \sum_s w_i y_i, \tag{2}$$

kus $w_i = \dfrac{k_i}{E(I_i)}$ on i -nda objekti kaal valimis, kus k_i valikuvektori I realisatsioon.

Hinnangufunktsiooni $\hat{t}_y$ dispersioon avaldub järgmiselt:

$$V(\hat{t}_y) = \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij} \frac{y_i}{E(I_i)} \frac{y_j}{E(I_j)}, \tag{3}$$

kus $\Delta_{ij} = Cov(I_i, I_j)$.

Dispersiooni nihketa hinnangufunktsioon on järgmine:

$$\hat{V}(\hat{t}_y) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\Delta_{ij}}{E(I_i I_j)} \frac{I_i y_i}{E(I_i)} \frac{I_j y_j}{E(I_j)}, \tag{4}$$

valimihinnanguga:

$$\hat{V}(\hat{t}_y) = \sum_s \sum_s \frac{\Delta_{ij}}{E(I_i I_j)} w_i y_i w_j y_j. \tag{5}$$

Tõestus:

Esiteks näitame, et statistik $\hat{t}_y$ on nihketa. Tõepoolest,

$$E[\hat{t}_y] = E[\sum_U \frac{I_i y_i}{E(I_i)}] = \sum_U \frac{E(I_i) y_i}{E(I_i)} = \sum_U y_i$$

ning sellele vastav valimi hinnang on:

$$\hat{t}_y = \sum_s \frac{k_i y_i}{E(I_i)},$$

kus valim $s = \{i : i \in U, \ \ mille \ \ jaoks \ \ k_i > 0\}$.

Nõuame, et $E(I_i) > 0$ (igal objektil oleks võimalus olla valitud).

Tähistades *i*-nda objekti kaalu $w_i = \dfrac{k_i}{E(I_i)}$ ja arvestades, et

$$V[\sum_{i=1}^{N} c_i X_i] = \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j Cov(X_i, X_j), \tag{6}$$

kus X_i on juhuslik suurus, saame rakendada (6) hinnangule $\hat{t}_y = \sum_{i=1}^{N} \dfrac{I_i y_i}{E(I_i)}$.

Tulemuseks on:

$$V(\hat{t}_y) = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{y_i}{E(I_i)}\frac{y_j}{E(I_j)}\underbrace{Cov(I_iI_j)}_{\Delta_{ij}}$$

$$\begin{cases} \hat{V}(\hat{t}_y) = & \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij}\frac{I_iy_i}{E(I_i)}\frac{I_jy_j}{E(I_j)}Cov(I_i,I_j) \\ E(\hat{V}(\hat{t}_y)) = & V(\hat{t}_y) \\ E(\hat{V}(\hat{t}_y)) = & \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij}\frac{y_i}{E(I_i)}\frac{y_j}{E(I_j)}Cov(I_i,I_j)E(I_iI_j) \end{cases} \Rightarrow$$

$$c_{ij} = \frac{1}{E(I_iI_j)} \Rightarrow$$

$$\hat{V}(\hat{t}_y) = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{Cov(I_i,I_j)}{E(I_iI_j)}\frac{I_iy_i}{E(I_i)}\frac{I_jy_j}{E(I_j)},$$

ja vastav hinnang

$$\hat{V}(\hat{t}) = \sum_s\sum_s\frac{Cov(I_iI_j)}{E(I_iI_j)}w_iy_iw_jy_j, \text{kus} \quad w_i = \frac{k_i}{E(I_i)}.$$

□

Märkus 1: TTA disainide korral nimetatakse hinnangut (1) Horvits-Thompsoni hinnanguks ja TGA disainide korral – Hansen-Hurwitz hinnanguks.

1.5 Valikudisainide näiteid

On välja töötatud palju erinevaid valikudisaine, millega saab tutvuda näiteks, Traat ja Inno (1997) ning Särndal *et al.* (1992) raamatutes. Siin toome vaid mõningad näited.

Üks enamkasutatavatest on *süstemaatiline valik*, mille korral objekti valitakse valimisse mingi fikseeritud sammu tagant, alustades esimesest juhuslikult valitud objektist. Antud disaini kasutatakse väga tihti praktikas, mille põhjuseks on valikuprotseduuri lihtsus. Samuti valides õiget sammu ning sorteerides objekte uuritava tunnusega tugevalt korreleeritud taustunnuse järgi, võib saavutada hea täpsusega hinnanguid.

Suurusega võrdelise tõenäosusega valikut kasutatakse siis, kui on teada mingit objekti suurust iseloomustavat abitunnust. Objekti valimisse sattumise tõenäosus sõltub sellest taustunnusest ning seega on objektidel erinevad kaasamistõenäosused. Selle valikudisaini kasutamine on õigustatud, kui suuremaid objekte on üldkogumis vähe (ja mõnda teise disaini korral nad ei pruugi valimisse üldse sattuda), kuid nende mõju hinnangu väärtusele on küllalt suur.

Klastervaliku korral on ÜK jagatud klastriteks (näiteks asulad). Antud juhul ei valita objekte, vaid klastreid vastavalt mingile disainile ning kõik objektid valitud klastrist sattuvad valimisse.

Antud töö praktilises osas kasutatakse nn. *lihtsat juhuslikku kihtvalikut*, seepärast vajavad sellega seotud disainid põhjalikumat iseloomustamist.

2 Lihtne juhuslik kihtvalik ja selle iseloomustavad karakteristikud

Nagu oli juba eespool mainitud põhineb lihtne juhuslik kihtvalik kahel disainil: teostatakse kihtvalik ning igas kihis rakendatakse lihtne juhuslik valik.

2.1 Lihtne juhuslik valik (LJV)

Lihtsa juhusliku valiku disain on teoreetiliselt kõige põhjalikumalt läbi uuritud, samuti LJV on praktikas laialt kasutatav. LJV on kahte liiki: tagasipanekuta (TTA), kui valitud ja mõõdetud objekti eemaldatakse ÜKst, ning tagasipanekuga (TGA), kui objekti mõõdetakse, andmed kantakse valimisse ja ÜKst ei eemaldata, seega sama objekti võib valida korduvalt.

Selleks, et leida parameeterhinnangud LJV korral, on vaja teada selle disaini karakteristikuid. Need on järgmised:

$f = \frac{n}{N}$, valiku suhe,

$\pi_i = f$, 1. järku moment (*i*-nda objekti kaasamistõenäosus),

$\pi_{ij} = f\frac{n-1}{N-1}, \quad i \neq j$, 2. järku moment,

$\Delta_{ii} = f(1-f)$, *i*-nda objekti valikuindikaatori I_i dispersioon,

$\Delta_{ij} = -f(1-f)\frac{1}{N-1}$, kovariatsioon kahe valikuindikaatori I_i ja I_j vahel,

Rakendades üldise hindamisteoreemi LJV korral on nihketa hinnangufunktsioon tunnuse kogusummale ÜK-s $t_y = \sum_U y_i$ järgmine:

$$\hat{t}_y = \frac{N}{n}\sum_U I_i y_i \tag{7}$$

ja temale vastav valimi hinnang on $\hat{t} = N\overline{y}$ (Särndal *et al.*, 1992, lk.66-68). Dispersiooniks on $V(\hat{t}) = N^2(1-f)\frac{S_y^2}{n}$ ja sellele vastav hinnang on $V(\hat{t}) = N^2(1-f)\frac{s_y^2}{n}$, kus $\overline{y} = \frac{1}{n}\sum_s y_i$ on valimi keskmine, $S^2 = \frac{1}{N-1}\sum_U (y_i - \overline{Y})^2$ on tunnuse *y* dispersioon ÜK-s, $s^2 = \frac{1}{n-1}\sum_U (y_i - \overline{y})^2$ on valimi dispersioon.

Nihketa hinnangfunktsioon ÜK keskmisele: $\overline{Y} = \frac{1}{N} t_y$ on:

$$\hat{\overline{Y}} = \frac{1}{N} \hat{t} = \overline{y}, \tag{8}$$

$$V(\hat{\overline{Y}}) = (1 - f) S_y^2 / n, \tag{9}$$

$$\hat{V}(\hat{\overline{Y}}) = (1 - f) s_y^2 / n. \tag{10}$$

Ülaltoodud valemite tõestusi võib leida töös Särndal et al. (1992, lk.66-69).

2.2 Kihtvalik

Kihtvalikus on ÜK jagatud mittekattuvateks alaosadeks ehk kihtideks mingi tausttunnuse järgi. Tõenäosuslik valik on rakendatud igas kihis ning võib erineda kihiti. See on võimas ja paindlik meetod, mida sageli kasutatakse praktikas. Selle meetodi populaarsuse põhjuste seas on näiteks järgmised:

- igas kihis on võimalik rakendada optimaalset disaini teistest kihtidest sõltumatult;
- praktilised aspektid, mis on seotud vastamise ja mittevastamise, mõõdu ja lisatunnustega, võivad oluliselt erineda (nt. piirkonniti). Kihistamine ja erinevate disainide rakendamine aitavad tõsta hinnangute täpsust.
- andmed võivad olla kihtide kaupa jaotatud administratiivsete põhjuste tõttu (näiteks erinevad uuringukeskused võivad asuda erinevates kohtades).

Olgu U_h h–nda kihi kõigi objektide hulk ning N_h kihi U_h maht, $h = 1, \dots H$. Kihtvaliku korral nihketa ÜK kogusumma hinnang on:

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h, \tag{11}$$

kui $E(\hat{t}_h) = t_h$ ja t_h on kogusumma kihis U_h. Hinnangufunktsiooni $\hat{t}$ dispersioon on:

$$V(\hat{t}) = \sum_{h=1}^{H} V(\hat{t}_h), \tag{12}$$

kuna hinnangud $\hat{t}_h$ on sõltumatud. Dispersiooni nihketa hinnangufunktsioon on

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_h), \tag{13}$$

kui $E(\hat{V}(\hat{t}_h)) = V(\hat{t}_h)$.

Antud valemite keskpunktis on hinnang kihi sees. Rakendades üldise hindamisteoreemi, saab avaldada hinnangu t_h jaoks järgmiselt:

$$\hat{t}_h = \sum_{U_h} \frac{I_i y_i}{E(I_i)}, \tag{14}$$

millele vastab hinnang:

$$\hat{t}_h = \sum_{s_h} w_i y_i, \tag{15}$$

kus $w_i = \dfrac{k_i}{E(I_i)}$ - i – nda objekti kaal, $s_h = \{i : i \in U_h, k_i > 0\}$.

Kui kogusumma *t* on hinnatud, siis ÜK keskmine on ka hinnatud:

$$\hat{\bar{Y}} = \frac{\hat{t}}{N}, \qquad V(\hat{\bar{Y}}) = \frac{V(\hat{t})}{N^2}, \qquad \hat{V}(\hat{\bar{Y}}) = \frac{\hat{V}(\hat{t})}{N^2}.$$

Kasutades teadmist, et $\hat{t} = \sum_{h=1}^{H} \hat{t}_h$, valemeid võib kirjutada alternatiivselt järgmiselt:

$$\hat{\bar{Y}} = \sum_{h=1}^{H} W_h \hat{\bar{Y}}_h, \tag{16}$$

$$V(\hat{\bar{Y}}) = \sum_{h=1}^{H} W_h^2 V(\hat{\bar{Y}}_h), \tag{17}$$

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} W_h^2 \hat{V}(\hat{\bar{Y}}_h), \tag{18}$$

kus $W_h = \dfrac{N_h}{N}$, $h = 1...H$.

2.3 Lihtne juhuslik kihtvalik

Antud töö teises osas on andmete analüüsimiseks rakendatud just *lihtne juhuslik kihtvalik*, kus igas kihis kasutatakse lihtsa juhusliku valikut. See on TTA disain, seega $E(I_i) = \pi_i$ ja on võrdne n_h / N_h (st konstantne kõikide objektide jaoks kihis *h*, kus $h = 1...H$.

Rakendades üldist hindamisteoreemi lihtsale juhuslikule kihtvalikule, saame et nihketa hinnangufunktsioon ÜK kogusummale avaldub järgmiselt:

$$\hat{t} \overset{(11)}{=} \sum_{h=1}^{H} \hat{t}_h \overset{(2)}{=} \sum_{h=1}^{H} (\sum_i w_i y_i) = \sum_{h=1}^{H} (\sum_i \frac{N_i}{n_i} y_i) = \sum_{h=1}^{H} N_i \frac{(\sum_i y_i)}{n_i} = \sum_{h=1}^{H} N_h \overline{y}_h, \tag{19}$$

tema dispersioon on

$$V(\hat{t}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{yh}^2 / n_h, \tag{20}$$

dispersiooni nihketa hinnangufunktsiooniga

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) s_{yh}^2 / n_h, \tag{21}$$

kus $S_{yh}^2 = \frac{1}{N_h - 1} \sum_{U_h} (y_i - \overline{Y}_h)^2$ ja $s_{yh}^2 = \frac{1}{n_h - 1} \sum_{S_h} (y_i - \overline{y}_h)^2$.

Hinnangud ÜK keskmisele on toodud valemites (16), (17), (18).

3 Paigutus kihtidesse

Kihtvaliku korral on väga oluline määrata valimimahu igas kihis, sest sellest sõltub hinnangu täpsus. Täpsuse tõstmiseks on olemas mitu erinevat valimipaigutust kihtides, näiteks, *optimaalne*, *võrdeline* ning *y-kogusummaga võrdeline paigutus*.

Optimaalse paigutuse korral on vaja teada uuringu maksumust. On eeldatud, et kogu uuringu maksumus koosneb üldkuludest ehk püsikuludest, ja igas kihis andmete saamise kuludest ehk muutuvatest kuludest. Loomulik järeldada, et kogu uuringu maksumus avaldub järgmiselt:

$$C = c_0 + \sum_{h=1}^{H} n_h c_h \ , \tag{22}$$

kus c_0 on püsikulud, c_h on kihis h andmete saamiseks vajalikud kulud ja n_h on h-nda kihi valimimaht.

Optimaalse paigutuse korral võib lähtuda kahest aspektist: kogu maksumuse minimiseerimine või hinnangu dispersiooni minimiseerimine. Optimaalse paigutuse leidmiseks kasutame järgmist teoreemi

Teoreem.

Kihtvaliku korral, kus uuringu maksumus on esitatud kujul (22) ja dispersiooni avaldatakse hajuvuste komponentide A_h ja B kaudu kui

$$V = V(\hat{Y}) = \sum_{h=1}^{H} A_h / n_h + B \,, \tag{23}$$

valimi optimaalne paigutus saavutatakse, kui

$$n_h \propto \left[\frac{A_h}{c_h} \right]^{\frac{1}{2}}, h = 1, \ldots .H \ .$$

Teoreemi tõestust võib leida raamatust Särndal *et al.*, 1992 (lk.104-105).

LJV korral esineb dispersioon kujul:

$$V(\hat{t}) = \sum_{h=1}^{H} \frac{N_h^2 (1 - \frac{n_h}{N_h}) S_{yh}^2}{n_h} = \sum_{h=1}^{H} \frac{N_h^2 S_{yh}^2}{n_h} - \sum_{h=1}^{H} N_h S_{yh}^2 \ ,$$

kust on näha, et see on kujul (23) ning komponendid A ja B avalduvad kui:

$$A_h = N_h^2 S_{yh}^2,\ B = -\sum_{h=1}^{H} N_h S_{yh}^2 \ . \tag{24}$$

Ühendades teoreemi tulemust ja saadud seost (24), võib järeldada, et optimaalne paigutus on selline, mille korral:

$$n_h \propto \frac{N_h S_{yh}}{\sqrt{c_h}}, h = 1, ..., H \ . \tag{25}$$

Nüüd eeldame, et uuringu maksumus C on fikseeritud, siis

$$n_h = \frac{N_h S_{yh}}{\sqrt{c_h}} \frac{(C - c_0)}{\sum_{h=1}^{H} N_h S_{yh} \sqrt{c_h}} \ . \tag{26}$$

Antud tulemus on samuti kirjutatud raamatus Inno ja Traat, 1997.

Kui nüüd vaadelda olukorda, kui kihimaksumus on konstantne, ehk $c_h = c = const, h = 1, ..., H$, siis valemit (26) võib lihtsustada, kasutades seost $C - c_0 = \sum_{h=1}^{H} c n_h = cn,$ kus n on kogu valimi maht. Tulemuseks on nn Neymanni paigutus:

$$n_h = n \frac{N_h S_{yh}}{\sum_{h=1}^{H} N_h S_{yh}}, h = 1, ..., H \ . \tag{27}$$

Optimaalne ja Neymani paigutused on optimaalsed siis, kui on tegemist ühe ainsa uuritava tunnusega ja kui uuritava tunnuse dispersioonid kihtides on teada (või on teada sellega tugevalt korreleeritud tunnuse dispersioonid). Sageli see nii aga ei ole. Üheks alternatiiviks on nn võrdeline paigutus kihtides, mis tähendab seda, et kihtide osakaalud nii valimis, kui ka üldkogumis on võrdsed:

$$n_h = n \frac{N_h}{N} \ . \tag{28}$$

Y kogusummaga võrdeline paigutus avaldub järgmiselt:

$$n_h = n \frac{\sum_{U_h} y_i}{\sum_{U} y_i}, \tag{29}$$

kus y on mittenegatiivne tunnus.

4 Koondatud ning püsijuhuarvud

Püsijuhuarvude (*Permanent Random Numbers*) ehk PRN ja koondatud juhuarvude (*Collocated Random Numbers*) ehk CRN valik on meetodid, mis võimaldavad koordineerida valimeid ning kontrollida ülekattumist erinevate valimite vahel. Selline meetod on kasulik siis, kui on vaja ühest ja samast üldkogumist võtta mitu valimit, näiteks kui erinevatel perioodidel hinnatakse samu näitajaid või ühel perioodil võrreldakse näitajad erinevate valimite vahel.

Samuti kasutatakse meetodit tihti siis, kui üldkogum võib muutuda nn objektide „surma" või „sünni" tõttu ehk teiste sõnadega, kui tekkivad uued objektid või vanad enam ei eksisteeri. Tihtipeale erinevatel valikutel on kasutusel võetud sama loend ning PRN ja CRN on tõhusad meetodid valimi koordineerimiseks.

4.1 Meetodite kirjeldus

Olgu F_0 esialgne loend, mis koosneb N_0 objektist (ajahetkel 0). Eeldatakse, et hilisematel perioodidel toimuvad objektide „surmad" ning uute objektide „sünnid". Tähistagu u_i lõigul $\left[0,1\right]$ ühtlase jaotusega juhusliku tunnust ning olgu n – soovitava valimi suurus. Vaatleme püsijuhuarvude meetodit võrdsete valiku tõenäosuste korral.

(1) igale objektile ÜK-st omistatakse sõltumatult $u_i \leftarrow U\left[0,1\right]$;

(2) ÜK objektid järjestatakse vastavalt u_i -ga;

(3) Alustades suvalisest punktist $a_0 \in \left[0,1\right]$ võetakse valimisse esimest n elementi, mille korral $u_i > a_0$. Kui n elementi ei ole intervallist $\left[a_0,1\right]$ valitud, siis jätkatakse 0-st.

Antud meetodi nimetatakse järjestikuseks lihtsaks juhuslikuks valikuks tagasipanekuta (Ernst *et al*, 1998). Valimi uuendamiseks igale uuele üksusele, mis sattub üldkogumisse (toimub „sünd") genereeritakse püsijuhuarv sõltumata juba olemasolevatest arvudest, mis jääb seotuks antud üksusega. Kui ettevõte enam ei eksisteeri (toimub „surm"), siis kustutades üksust ei mõjuta see teisti püsijuhuarve.

Meetodi puuduseks on aga see, et püsijuhuarvude konkreetse kihi sees ei pruugi olla hästi jaotatud. Meetod on eriti tundlik juhul, kui eesmärgiks on koordineerida kaks valimit nii, et ülekattumine nende vahel oleks minimaalne. Probleem tekkib siis, kui palju u_i-dest kuhjub ühte osasse intervallis $[0,1]$. Selle tulemuseks on valikute ebasoovitav ülekattumine. Tihti juhtub see siis, kui kihtide üldkogumid on väiksed. Illustreerimiseks on toodud järgmine näide (Ernst, L.R., *et al*, 1998):

Näide 1.

Olgu planeeritud 3 uuringut ühes ja samas üldkogumis, kus $N = 10$, $n_1 = n_2 = 2$, $n_3 = 4$. Nende alguspunktid on: $a_1 = 0$, $a_2 = 0.25$, $a_3 = 0.5$

Kui juhtub nii, et kõik u_i, $i = 1,...,12$ juhuslikult sattuvad intervalli $[0;0.25)$, siis teostades PRN valikut, 1. ja 2. elemendid sorteeritud andmestikus saavad valitud kõigis kolmes uuringus.

Selle probleemi üheks lahenduseks on nn. *koondatud juhuarvude* kasutamine. Selle meetodi algoritm on järgmine:

(1) igale objektile loendist omistatakse juhuslik arv $u_i \sim U[0,1]$;

(2) loendi järjestatakse kasvavalt vastavalt u_i-dele ning omistatakse igale objektile järku R_i;

(3) genereeritakse teistest sõltumata juhuslik arv $\varepsilon \sim U[0,1]$ ning arvutatakse uus u_i iga objekti jaoks järgmiselt: $u_i = (R_i - \varepsilon) / N_0$

Antud meetod võimaldab jaotada u_i- d ühtlaselt lõigul [0,1] ning vältida probleemi, mis oli kirjeldatud ülalpool toodud näides (Ernst, *et al*, 1998).

Kui soovitatakse võtta sellised valimid, kus elemendid erinevate perioodide lõikes ei kattu, siis alguspunktid leitakse valimi ja üldkogumi mahu osakaalust lähtudes.

Näide 2.

Lihtsaks näiteks on kaks valimit, mis on võetud samas üldkogumis. Olgu alguspunktideks $a_0 = 0$ ja $a_1 = 0.5$. Kui valimite valikusuhted n_1 / N ja n_2 / N on väiksemad kui 0.5, siis valimite vahel ühisosa puudub. Pärast püsijuhuarvude omistamist ning järjestamist, esimesse valimisse sattuvad need elemendid, mille püsijuhuarvud on vahemikus 0-0.5 ja need, mis on vahemikus 0.5-1 sattuvad teisesse. Kui üldkogumist on võetud mitu valimit, on ilmne, et elementide ühisosa tekkib siis, kui valimite mahtude summa on võrdne üldkogumi mahuga või sellest suurem. Kui ei ole võimalik saada paarikaupa ühisosata valimeid, siis üritatakse ülekattumist vähendada (Tamm, E., 2003).

Kui valimite mahtude summa on oluliselt väiksem, kui üldkogumimaht, siis on võimalik leida sellised alguspunktid, mille korral saavutatakse olukord, kus valimid omavahel ei kattu.

Püsijuhuarvude meetod on lihtne ja tõhus meetod valimite koordineerimiseks. Selle meetodi kasutamise peamine põhjus on soov vähendada kao määra, ehk objektidelt saamata andmete osakaalu. See võimaldab vähendada uuringus osalevate küsitlevate vastamiskoormust, mis omavahel toob kaasa kao vähenemist.

Peatükk II

Valikuuring perearstikeskuste majandusaruande näitel

1 Andmete kirjeldus

Töö aluseks on perearstikeskuste poolt esitatud majandustegevuse aruanded 2004 aasta eest (kokkuvõtvad tulemused on samuti kajastatud Eesti tervishoiustatistika aastaraamatus 2004. aastal). Andmed on esitatud Eesti Sotsiaalministeeriumi Terviseinfo- ja analüüsiosakonna poolt. Tervishoiu kogukulude arvestuse aluseks olevad andmed pärinevad sellistest administratiivsetest allikatest nagu Rahandusministeeriumi riigieelarve ja kohalike omavalitsuste eelarvete täitmise aastaaruanne, Riigikassa andmebaas ning Vabariigi Valitsuse reservfondist tehtud tervishoiukulutused, ministeeriumide haldusala tervishoiukulutused, Haigekassa finants- ja majandusaruanded.

Tervishoiu kogukuludega mõõdetakse majanduslikke ressursse, mida kulutati aasta jooksul kaupadele ja teenustele. See summa sisaldab lisaks raviteenustele administreerimis- ja kapitalikulu. Haigushüvitisi ega koolituskulusid ei arvestata (Eesti tervisehoiustatistika aastaraamat 2004).

Töös kasutatud andmestik sisaldab järgmisi näitajaid:

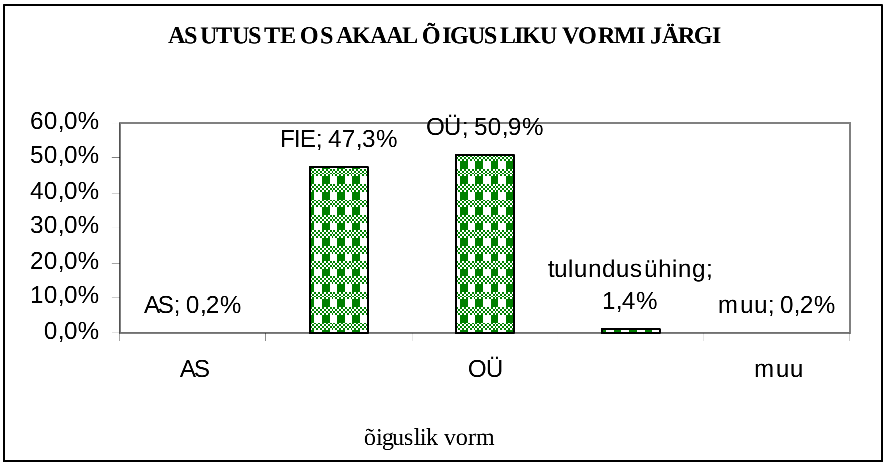
- asutuse järjekorranumber;
- omaniku liik (antud asutuse omanik, näiteks riiklik, kohalik, jne);
- õiguslik vorm (näiteks FIE, OÜ, MTÜ, jne);
- maakond/linn;
- kuupäev, millal asutus alustas oma tegevuse;
- kuupäev, millal asutus lõpetas oma tegevuse (fikseeritud ainult nende asutuste kohta, mis aasta 2004 alguses veel töötasid, kuid 2004. aasta jooksul oma tegevuse lõpetasid);
- töötajate aastakeskmine arv ning nn „normeeritud“ aastakeskmine töötajate arv;
- erinevad kululiigid, kogutulud (2003), kogukulud (2004).

Sotsiaalministeeriumi peamiseks huviobjektideks on järgmised parameetrid: asutuste koguarvud õigusliku vormi järgi ning asutuste kulu. Teostatav valikuuring on suunatud nende näitajate hindamisele. Lisaks soovib Sotsiaalministeerium neid hinnanguid ka maakonniti.

1.1. Põhinäitajate kirjeldav analüüs

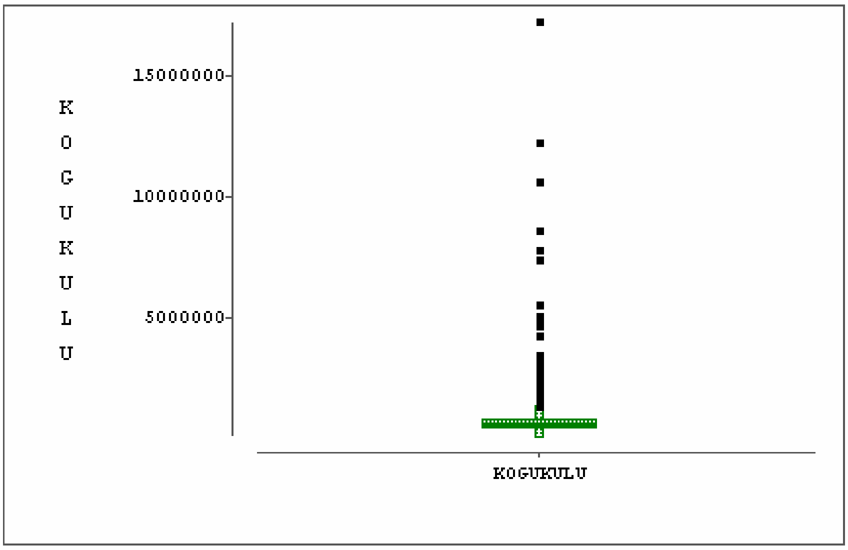
Vaatleme huvipakkuvaid tunnuseid põhjalikumalt. Selleks on kasutada 2004. aasta kõikse uuringu andmeid.

a) **Õiguslik vorm** võib omandada 10 erinevat väärtust (vaata lisa 1), kuid 46,8% kõikidest andmetest on füüsilisest isikust ettevõtted (FIE) ja 51,3% on osaühingud. Ülejäänute vormide osakaal on liiga väike (MTÜ, KOV ja usaldusühinguid puuduvad üldse), et valikuuringu põhjal saada hiljem hea hinnangu nendele. Seepärast ühendame need kokku ja anname uue väärtuse „muu õiguslik vorm“.



Joonis 1. Asutuste osakaal õigusliku vormi järgi

b) Asutuste **kogukulu** on jaotanud samuti mittesümmeetriliselt (Tabel 2 ja joonis 2). Keskväärtuseks on 967 405 EEK. Standardhälve on küllalt suur (1 435 927 EEK) ning perearstide keskuste kogukulude summa üle Eesti moodustab 429 528 018 EEK. Nagu tabelist 2 on näha, $N = 444$, kuid Eestis on kokku 474 perearstikeskust. Seega meil on tegu mittevastamisega, mis moodustab umbes 6% kõikidest keskustest.



Joonis 2. Kogukulu 2004.a

Moments			
N	444.0000	Sum Wgts	444.0000
Mean	967405.446	Sum	429528018
Std Dev	1435927.96	Variance	2.062E+12
Skewness	6.3058	Kurtosis	53.5049
USS	1.329E+15	CSS	9.134E+14
CV	148.4308	Std Mean	68146.1251

Tabel 2. Kogukulu 2004.a

2 Valimi- ning kihtide mahtude määramine

Kihtvaliku korral tuleb kõigepealt otsustada, kui suur on valimimaht. Valimimahu määramise korral antakse hinnangute soovitav täpsus suhtelise vea fikseerimisega.

Definitsioon 3: Suhteline viga näitab, kui suure osa moodustab hinnangu standardviga hinnangust endast (Traat ja Inno, 1997).

$$suhtv(\hat{\theta}) = \sqrt{\hat{V}\hat{\theta}(\boldsymbol{s})} / \hat{\theta}(s)\,. \tag{30}$$

Otsime paigutust, mille puhul suhteline viga on väiksem kui p. Selleks kasutame kogusumma hinnangufunktsiooni (19) ja dispersiooni hinnangut (21) kogukulu summale ning otsime paigutust, mis rahuldab järgmist võrratust:

$$\frac{\sqrt{\sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) s_{yh}^2 / n_h}}{\sum_{h=1}^{H} N_h \bar{y}_h} \le p, h = 1, \dots, H \tag{31}$$

Näiteks, selleks et leida sobiva suhtelise veaga valimimahu optimaalse jaotuse puhul, lahendame võrratust järgmiselt:

$$\sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) s_{yh}^2 / n_h \le p^2 \left(\sum_{h=1}^{H} N_h \bar{y}_h \right)^2,$$

$$\sum_{h=1}^{H} \frac{N_h^2 s_{yh}^2}{n_h} - \sum_{h=1}^{H} N_h s_{yh}^2 \le p^2 \left(\sum_{h=1}^{H} N_h \bar{y}_h \right)^2,$$

$$\sum_{h=1}^{H} \frac{N_h^2 s_{yh}^2}{n_h} \le p^2 \left(\sum_{h=1}^{H} N_h \bar{y}_h \right)^2 + \sum_{h=1}^{H} N_h s_{yh}^2,$$

Nüüd asendame ülevalolevas võrratuses n_h kasutades seost (27) ning saame:

$$\frac{1}{n} \sum_{h=1}^{H} N_h S_{yh} \sum_{h=1}^{H} \frac{N_h s_{yh}^2}{S_{yh}} \le p^2 \left(\sum_{h=1}^{H} N_h \bar{y}_h \right)^2 + \sum_{h=1}^{H} N_h s_{yh}^2,$$

kus saame asendada meile teadmata valimist sõltuv s_{yh}^2 ligikaudse hinnanguga S_{yh}^2 ning viia n ühele poole saamaks järgmise alumise piiri huvipakkuva valimimahu jaoks:

$$\frac{\left(\sum_{h=1}^{H} N_h S_{yh} \right)^2}{p^2 \left(\sum_{h=1}^{H} N_h \bar{y}_h \right)^2 + \sum_{h=1}^{H} N_h S_{yh}^2} \le n,$$

Kihtide mahtusid saab nüüd määrata kasutades uuesti valemit (27). Võrdelise ning y kogusummaga võrdelise paigutuse puhul saab leida nõutava täpsuse saavutamiseks vajalik valimimaht analoogiliselt.

Esialgseks suhteliseks veaks oli võetud 3%, kuid saadud tulemused olid vastuvõetamatud nii optimaalse, kui ka y kogusummaga võrdelise paigutuse korral (mõnede kihtide mahud olid võrdsed ÜK mahuga või valimimaht oli liiga suur ÜK suhtes). Selle probleemi üheks lahenduseks on vähendada täpsuse nõue, st. suhteliseks veaks võtta 5%. Sellisel juhul y kogusumma võrdelise paigutuse korral (31) oli saadud minimaalseks valimimahuks $n = 214$, võrdelise paigutuse korral (30) $n = 288$ ning optimaalne paigutus (29) andis kõige väiksema valimimahu $n = 191$, mida oligi antud juhul võetud kasutusele (lisas 3 on toodud kihtide mahud teiste paigutuste korral).

Järgmise sammuga said leitud kihtide suurused optimaalse paigutuse korral, kus kihistavaks tunnuseks on maakond, $h = 1,....,17$ ja y tunnuseks on kogukulu. Allpool on toodud tabel 3, kus on näidatud asutused kokku ning kihtide mahud maakoniti:

Maakond/Linn	**Kokku**	**Kihi maht**
Harju	35	10
Hiiu	6	2
Ida-Viru	53	32
Jõgeva	18	2
Järva	21	5
Lääne	12	2
Lääne-Viru	28	6
Põlva	19	2
Pärnu	27	24
Rapla	14	4
Saare	19	4
Tartu	24	5
Valga	18	2
Viljandi	33	3
Võru	17	4
Tallinn	75	68
Tartu linn	25	16

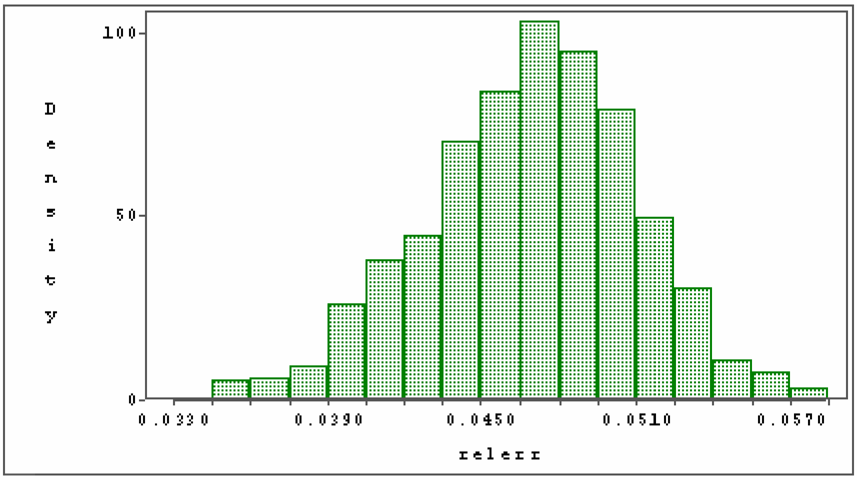
Tabel 3. Optimaalse paigutuse korral kihtide mahud maakonniti

Tabelist on näha, et mõne maakonna jaoks tuleb valimisse võtta suurema osa selle maakonna keskustest (nt. Tallinna ja Pärnu puhul). Selle põhjuseks on tunnuse suur varieeruvus nendes maakondades (vt lisa 12). Mida stabiilsemad on tunnuse näitajad maakonnas, seda vähem objekte on tarvis selle maakonna kirjeldamiseks.

3 Hinnangud ning nende kontroll suhtelise vea kaudu

Valikuuringu teostamiseks leitud valimimaht 191 oli saadud, kasutades hinnangulisi väärtuseid (ÜK dispersioon ja keskmine valimi omade asemel). Seepärast oleks kasulik kontrollida, milline on tegeliku vea suurus, võrreldes nõutava 5%-ga. Seda saab teostada simuleerimise teel. Võtame näiteks üldkogumist 1000 valimit vastavalt lihtsale juhuslikule kihtvalikule ning leiame iga kord kogusumma hinnangu suhtelist viga. Vaatame seda suhtelist viga nii üle kogu Eesti ning maakonniti (viimasel juhul see võib tugevasti varieeruda, kuna valimimaht ja optimaalne paigutus on seotud kogusumma hinnanguga üle Eesti, mitte maakonniti).

Suhtelise vea jaotus üle 1000 simulatsiooni on näha joonisel 3.



Joonis 3. Suhtelise vea jaotus üle Eesti kogukulu summa hinnangule

Moments			
N	1000.0000	Sum Wgts	1000.0000
Mean	0.0469	Sum	46.9366
Std Dev	0.0041	Variance	1.702E-05
Skewness	-0.2549	Kurtosis	0.0378
USS	2.2201	CSS	0.0170
CV	8.7893	Std Mean	0.0001

Tabel 4. Suhtelise vea parameetrid kogukulu summale üle Eesti

Histogrammist ja tabelist 4 on näha, et suhteline viga kogusumma hinnangule üle Eesti keskmiselt 0,047 ehk 4,7% ja maksimaalne viga on 0,057.

Nagu oli oodata, suhteline viga maakonniti on halvem, keskmiselt 18,5% (Lisa 13). Mõned maakonnad on esindatud üldkogumis vaeselt, näiteks maakonnas 39 (Hiiu maakond) on kokku vaid 6 meditsiinilist asutust, mis pakkuvad perearstide teenust. Optimaalse paigutuse rakendamise tulemusel on sealt võetud 2 väärtust ja suhteline viga on võrdne 16%. Kõige suurem suhteline viga on Rapla maakonnas – 29% ning kõige väiksem on Tallinnas – 5%.

Kui sooviks on tõsta hinnangute täpsust maakonniti, siis tuleb ka valimimahtu määrata iga maakonna jaoks eraldi.

Tulemuste illustreerimiseks võrdleme ÜK-i kogukulu summat (ehk tegelikku parameetrit) selle tunnuse Monte-Carlo hinnanguga $\hat{\theta}_{MC} = 1/1000\sum_{i=1}^{1000}\hat{\theta}_i$, kus θ_i on kogusumma hinnang leitud ühe juhusliku kihivaliku teel

ÜK kogukulu summa (mln kr)	**MC hinnang kogukulu summale**	**vahe osakaal ÜK-st %**
429,5	429,7	0,04%

Tabel 5. Kogukulu summa üldkogumi ja Monte-Carlo hinnangu võrdlus

Tegu on punktihinnanguga, mis üle simulatsioone annab keskmiselt üsna täpse hinnangu. Vahe osakaal on leitud järgmiselt: $\frac{|t_y - \hat{t}_y|}{t_y}$, kus t_y on ÜK kogusumma ja $\hat{t}_y$ on temale vastav hinnang. Lisas 4 on toodud hinnangud maakonniti ning võrdlemiseks üldkogumi tegelikult väärtused.

Leiame ka Monte-Carlo usaldusintervalli kogukulu summale $I_\theta = (\hat{\theta} \pm \lambda_{\frac{\alpha}{2}}\sqrt{\hat{V}\hat{\theta}})$, milleks tuli üle Eesti: $I^{\text{MC}}_{\text{kogusumma}} = (279{,}5;\ 579{,}9)$. Lisas 5 võib vaadata usaldusvahemikud maakonniti.

Vaatame veel keskmist hinnangut keskmisele kogukulule (tulemused on koondatud tabelisse 6)

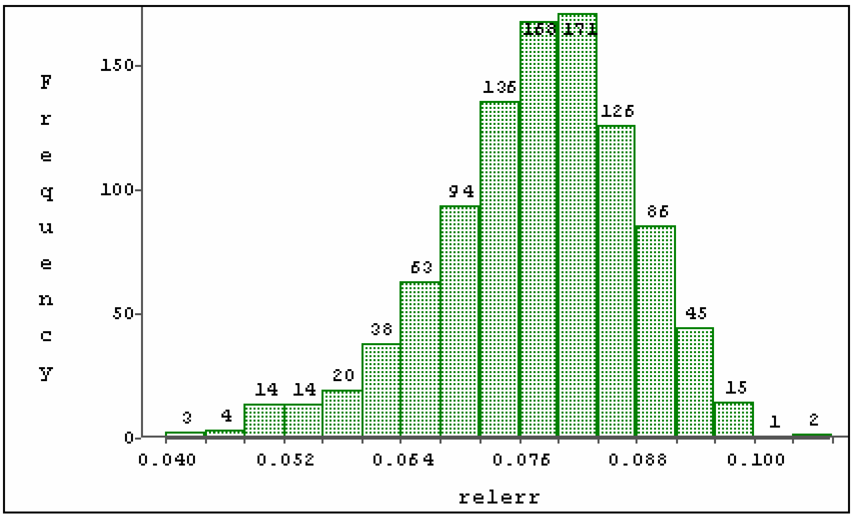
ÜK kogukulu keskmine	MC hinnang kogukulu keskmisele	vahe osakaal ÜK-st %
967,4	967,8	0,04%

Tabel 6. Kogukulu keskmise üldkogumi ja Monte-Carlo hinnangu võrdlus

On näha, et keskmise hinnang ei erine üldkogumi keskmisest. Kuid hinnangud maakonniti (Lisa 6) on täpsemad, ning vahe osakaal on 1% või väiksem. Suurte dispersioonide tõttu saadud usaldusintervallid nii maakonniti kui ka üldine tunduvad olevat laiad. Usaldusvahemikuks on $I^{MC}_{\text{keskmine}} = (0,44;\ 1,19)$. Usaldusintervalle maakonniti võib leida lisas 7.

Üheks Sotsiaalministeeriumi poolt püstitatud ülesandeks on hinnata kvalitatiivset tunnust „õiguslik vorm“. Punktis 5.1 oli kirjeldatud, et mõningaid õigusliku vormi liike üldkogumis ei olnud üldse või leidusid vaid mõned üksikud väärtused. Sel põhjusel oli tehtud gruppide liigendus binaarseteks tunnusteks: OÜ, FIE ning „muu õiguslik vorm“. Seejärel hindamisülesanne taandub osakogumite hindamisele. Allpool on teostatud analüüs OÜ puhul, kui kõige levinuma õigusliku vormiga meditsiiniliste asutuste seas. Teiste gruppide osakaalude hinnangud leitakse analoogiliselt.

Mis puudutab suhtelist viga üle Eesti (Joonis 4, Tabel 7), siis tunnuse „õigusliku vorm OÜ“ kogusumma jaoks see on suurem, kui sama näitaja tunnuse kogukulu jaoks, 7,8%. Võib öelda, et osakaalu hinnang osutus ebatäpsem ning 2,8% võrra suurem töös määratud täpsusest.



Joonis 4. Suhtelise vea jaotus üle Eesti tunnuse „OÜ“ kogusumma hinnangule

Moments			
N	1000.0000	Sum Wgts	1000.0000
Mean	0.0778	Sum	77.7618
Std Dev	0.0102	Variance	0.0001
Skewness	-0.5752	Kurtosis	0.5216
USS	6.1499	CSS	0.1030
CV	13.0568	Std Mean	0.0003

Tabel 7. Suhtelise vea parameetrid üle Eesti tunnuse „OÜ“ kogusumma hinnangule

Kuid tabelist 8 on näha, et Monte-Carlo simuleerimisel kogusumma üldkogumis ja saadud hinnangu vahe on suhteliselt väike – moodustab väiksem kui 1% üldkogumist.

OÜ ÜK summa	**MC hinnang OÜ summale**	**vahe osakaal ÜK-st %**
226	226,2	0,08%

Tabel 8. Tunnuse „OÜ“ kogusumma võrdlus üldkogumi ja Monte-Carlo hinnangu vahel

Usaldusvahemik, mis katab üldkogumi kogusumma parameetrit 95% tõenäosusega on $I_{kogusumma}^{MC} = (101;\, 351)$. Analoogilisi tulemusi maakonniti võib leida lisades 8 ja 9. Nende tulemuste põhjal võib öelda, et kogusumma hinnangud erinevates maakonnades on hea täpsusega ja suhteline erinevus üldkogumi omaga on mitte suurem kui 1 % .

Nagu ka teiste tunnuste korral keskmise kogukulu hinnang ei erine üldkogumi omast.

OÜ ÜK keskmine	**MC hinnang OÜ keskmisele**	**vahe osakaal ÜK-st %**
0,509	0,510	0,08%

Tabel 9. Tunnuse „OÜ“ keskmise võrdlus üldkogumi ja Monte-Carlo hinnangu vahel

Usaldusintervall $(1-\alpha) = 0.95$ osakaalu keskmisele on $I_{\mathrm{keskmine}}^{\mathrm{MC}} = (0{,}14;0{,}82)$, mis on küllalt lai. Keskmise hinnangud piirkonniti (lisa 10) jätavad positiivse mulje, sest erinevused tegelikkusest parameetrist on sageli väga väikesed. Samuti tuleb arvestada ka sellega, et mõnedel juhtudel hinnangud on leitud kasutades vaid üksikuid väärtusi, ehk teiste sõnadega asutusi oli väga vähe. Näiteks, Järve maakonnas on üldkogumis on vaid 4 asutust õigusliku vormiga OÜ, Lääne maakonnas vaid 2. Hiiu maakonnas aga 6-st asutusest kõik 6 olid osaühingud, seega nii hinnang, kui ka usaldusvahemik on võrdne ühega.   Usaldusvahemikud piirkonniti on toodud lisas 11.

Tehtud analüüsi põhjal võib kokkuvõtvalt järeldada, et lihtsa juhusliku kihtvaliku optimaalse paigutusega rakendamine on olnud edukas ning kirjeldamine üldkogumi hinnangute abil on täpne ja vastuvõetav (üldkogumi tasemel). Pideva tunnuse kogusumma standardviga üle Eesti jääb nõutavas piiris, mis oli ette antud 5%-ga. Maakonniti see nii täpne ei ole, kuid seda polnud esialgu nõutud, st suhteline viga oli fikseeritud vaid üle Eesti. Osakaalu hinnangud ei tulnud nõutud täpsusega ning keskmiselt on nad määratud normist 2,8% võrra suuremad . See on põhjendatud sellega, et valimimaht oli ette antud suhtelise vea määramisega tunnuse „kogukulu“ jaoks. Selleks, et suhteline viga tunnuse „OÜ“ jaoks oleks väiksem on soovitatav suurendada valimimahu, mis teisest küljest raskendab analüüsi teostamist.

Optimaalse paigutuse rakendamine võimaldab võtta suhteliselt väikest valimit – 191 objekti üldkogumist, mille suurus on 444. Positiivne on see, et leitud kogusumma hinnang on nn baashinnanguks, ehk selle abil on lihtne leida teisi huvipakkuvaid hinnanguid.

4 Püsijuhuarvude meetodi rakendus perearstikeskuste andmestiku näitel

Olgu seatud järgmised tingimused perearstikeskuste andmestikule: rotatsiooniperioodiks on 3 aastat, ning püsijuhuarvude meetod rakendatakse igas kihis (st maakonnas) eraldi. Vaadates tabelit 3, on näha, et ülekattumus tekkib Ida-Virumaal (44), Pärnus (67), Tallinnas (784) ja Tartus (795). Kuna tahetakse valimit vahetada kolme aasta tagant, siis alguspunktideks on $a_0 = 0,\ a_1 = 0.33,\ a_2 = 0.66$.

Tihti on vaja hinnata näitajate muutused kahel järjestikusel perioodil. Sel juhul on tähtis, et valimid oleksid suures osas kattuvad. Samuti peaks toimuma ka valimi uuendamine. Selliste tingimuste realiseerimiseks sobib nn *konstantse nihke meetod* (constant shift method) (Sõstra ja Puusepp, 2005). Oletame, et valimeid võetakse kord aastas, siis eelmise aasta valimi alguspunktidele lisatakse mingi arv, mis on igal aastal sama, ehk *nihe*. Seega, igal aastal valimisse satuvad uued ettevõtted. Näiteks, kui nihe on 0.01 ja valimisse on võetud 10% üldkogumist, võetakse juurde keskmiselt $0.01/0.1$ ehk üks kümnendik valimis on uued ettevõtted ja ettevõtte oodatav valimis oleku aeg on 10 aastat. Kuid tuleb meeles pidada, et toimub ka üldkogumi uuenemine, seega need arvud on ligikaudsed.

Olgu kasutusele võetud nihe on 0.02. Leitud uute ettevõtete osakaal üldkogumist on toodud tabelis 10.

Maakond	Kokku	Kihi maht	Valimi osakaal ÜKst (%)	Uued ettevõtted valimis (%)
Harju	35	10	29	7
Hiiu	6	2	33	6
Ida-Viru	53	32	60	3
Jõgeva	18	2	11	18
Järva	21	5	24	8
Lääne	12	2	17	12
Lääne-Viru	28	6	21	9
Põlva	19	2	11	19
Pärnu	27	24	89	2
Rapla	14	4	29	7
Saare	19	4	21	10
Tartu	24	5	21	10

Valga	18	2	11	18
Viljandi	33	3	9	22
Võru	17	4	24	9
Tallinn	75	68	91	2
Tartu linn	25	16	64	3

Tabel 10. Konstantse nihke meetodiga leitud uute ettevõtete osakaal

Kokkuvõte

Antud töö eesmärgiks oli välja töötada tõenäosusliku valikuuringut konkreetse andmestiku korral ja võrrelda saadud hinnanguid tegelikke parameetritega. Andmestikuna olid kasutatud perearstikeskuste poolt esitatud majandustegevuse aruanded 2004 aasta eest. Sobivaks disainiks sai valitud lihtne juhuslik kihtvalik, kus oli oluline leida õige paigutuse kihtide vahel hinnangu minimaalse dispersiooni ja mõistliku valimimahu mõttes. Tähtsaks kriteeriumiks selleks oli Sotsiaalministeeriumi poolt pakutud hinnangu suhteline viga 3%, kuid sel juhul ei õnnestunud saada ei optimaalse ega *y*-kogusummaga võrdelise paigutustega vastuvõetavaid kihtide mahtusid. Rakendades võrdelist paigutust kihtides, valimimahu suurus tuli lähedane üldkogumiga, mis samuti ei vastanud seatud eesmärkidele.

Lahenduseks otsustati suurendada suhtelise vea 5%-ni. Proovides erinevaid paigutuse variante, leiti, et optimaalse paigutuse korral valimimaht on kõige väiksem (191 elementi). Järgmisel etapil sai läbi viidud 1000 simulatsiooni hinnangute arvutamisel, ning kontrollitud, et suhteline viga üle Eesti jääb tõesti 5% piiris.

Võrreldes uuritavate tunnuste hinnanguid üldkogumi omadega osutus, et keskmiselt on nad lähedased ja täpsed nii üle Eesti kui ka maakonniti. Lisaks pidevatele tunnustele sai uuritud ka binaarne tunnus „OÜ“ ehk osaühing, mis on moodustatud kodeeritud tunnusest „õiguslik vorm“ ning leitud hinnanguid selle tunnuse kogusummale nii üle Eesti kui ka maakonniti. Hinnangu suhteline viga üle Eesti oli keskmiselt 7,8%, kuid hinnang ise tuli üle simulatsioonide täpne nii üle Eesti, kui ka maakonniti.

Ühest püstitatud eesmärkidest oli ka valimite koordineerimine iga kolme aasta tagant, milleks kasutati püsijuhuarvude meetodit. Selle meetodi plussiks on küsitlevate vastamiskoormuse vähendamine. Juhul, kui ülekattuvus valimite vahel ei ole soovitav, elementide jaotus kihtide vahel optimaalse paigutuse korral sobib väga hästi. Analüüs näitas, et ülekattuvus tekkib vaid neljas piirkonnas, milleks on Tallinn, Pärnu, Ida-Viirumaa ja Tartu. Kui soovitatakse rakendada konstantse nihke meetodit, siis see on samuti võimalik ning vastav meetod on kirjeldatud.

Kokkuvõttes võib öelda, et üleminek valikuuringutele on võimalik ja annab häid tulemusi. Tekkinud on ka mõned probleemid (nagu näiteks hinnangute täpsuse tõstmine maakonniti), mida soovitatakse põhjalikumalt edaspidi uurida.

Survey Sampling: a Case Study on The GP Annual Report 2004

Anna Leontyeva

Summary

Survey sampling is a way of performing statistical data analysis by selecting a random sample from a finite population. It is widely used in the industry of public opinion polling and census activities because it allows to obtain scientifically reliable results more economically.

In this work survey sampling was applied to the analysis of the Estonian general practitioners annual report data in order to evaluate whether it is reliable enough to be more widely adopted by the Ministry of Social Affairs of Estonia, as this would allow to significantly decrease their costs.

We used the complete census of general practitioners of Estonia as the base dataset and were interested in estimating the population total and the population average of one interval variable ("total cost") and one nominal variable ("legal status"), the latter preliminarily transformed to a set of binary variables. The question of interest was, how good are the estimates obtained using the survey sampling methods in comparison to the values, estimated on the whole dataset, and what is the size of the sample required to achieve acceptable relative error.

Stratified random sampling without replacement was applied. The size of the sample was selected so that the relative standard error of the estimates would not significantly exceed a predefined threshold. Originally it was decided to take 3% as the maximum relative error, but it turned out there were no proper allocations that would satisfy this requirement. The choice of a slightly higher threshold of 5% solved the problem. A number of allocation techniques was tried (optimal, proportional, y-proportional), and the sample obtained by the optimal allocation was the smallest (191 objects).

We performed Monte-Carlo simulations to ensure that the chosen sample indeed ensured a relative error of approximately 5%, which turned out to be the case.

When we used the same sample to estimate the population total and the population average of the “legal status” variable we have obtained a slightly larger relative error of 7.5%.

Finally, we demonstrated the use of the Permanent Random Number method, which is useful for sample coordination and is meant to help to reduce the overlap between samples taken in consecutive years.

To conclude, we find the survey sampling methods to be reliable and usable in practice.

Kasutatud kirjandus

1. Egel, E., Edomskihh-Eigo, N., Karelson, K., *et al*, Eesti tervishoiustatistika aastaraamat 2004, Sotsiaalministeerium, Tallinn 2007
2. Ernst, R.L,, Valliant, R.,Casady, R.J., Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths, US Bureau of Labor Statistics, December 1998
3. Rao,J.N.K, Interplay Between Sample Survey Theory and Practice: An Appraisal, Statistics Canada, 12-001, December 2005
4. Särndal, C.-E., Swensson, B., Wretman, J., Model Assisted Survey Sampling, Springer-Verlag, 2003
5. Sõstra, K., Puusepp, V., Valimite koordineerimine, *„Kuidas valitakse ettevõtteid uuringutesse"*, Eesti Statitika 7/05, Tallinn, 2005
6. Tamm, E., Valimite koordineerimine, *„Ühest üldkogumist võetud valimite koordineerimine"*, Eesti Statistika 1/03, Tallinn, 2003
7. Traat, I., Inno, J., Tõenäosuslik valikuuring, TÜ kirjastus, Tartu, 1997

Lisad

Lisa 1. Tunnuste koodid

Maakond/Linn	**Õiguslik vorm**
37 - Harju maakond	11 - FIE
39 - Hiiu maakond	22 - usaldusühing
44 - Ida-Viru maakond	23 - OÜ
49 - Jõgeva maakond	24 - AS
51 - Järva maakond	25 - tulundusühing
57 - Lääne maakond	30 välismaa äriühingu filiaal
59 - Lääne-Viru maakond	40 - SA
65 - Põlva maakond	50 - MTÜ
67 - Pärnu maakond	60 riiklik üksus
70 - Rapla maakond	70 KOV üksus
74 - Saare maakond	
78 - Tartu maakond	
82 - Valga maakond	
84 - Viljandi maakond	
86 - Võru maakond	
795 - Tartu	
784 - Tallinn	

Lisa 2. Kasutatud programmikood

```
data tmp1.puhas_andmestik;
	/* Eemaldame andmestikust ebavajalikud read */
	set Tmp1.majandus_perearstid;
	if kogukulu='.' then delete;
	run;


proc surveyselect data=Tmp1.puhas_andmestik rep=1000
	/* Teeme igast kihist juhusliku valimi, kordame 1000 korda
*/
	method=srs   /* srs = Simple Random Sampling */
	/* Igast maakonnast valime niipalju */
	n=(10 2 32 2 5 2 6 2 24 4 4 5 2 3 4 68 16)
	out=Tmp1.simuleerimine;
	/* Kihid = maakonnad */
	strata maakondl;
	run;

proc means data=Tmp1.simuleerimine;
	/* Hinnangute leidmine:
	     arvutame statistikud iga 1000 simulatsiooni jaoks */
	class maakondl replicate;
	var Kogukulu ;
	output out=Tmp1.simul_kokkuv sum=kogusumma mean=keskmine
var=disp
	             N=kokku;
	run;


data Tmp1.simul_kokkuv;
	/* Kustutame ebavajalikud tulemused */
	set Tmp1.simul_kokkuv;
	if maakondl='.' or replicate='.' then delete;
	run;

data Tmp1.simul_kokkuv;
	/* Lisame andmestikule YK mahu, arvutame suhtelised vead */
	set Tmp1.simul_kokkuv;
	if maakondl=37 then N=35; /* UK math */
	if maakondl=39 then N=6;
	if maakondl=44 then N=53;
	if maakondl=49 then N=18;
	if maakondl=51 then N=21;
	if maakondl=57 then N=12;
	if maakondl=59 then N=28;
	if maakondl=65 then N=19;
	if maakondl=67 then N=27;
	if maakondl=70 then N=14;
	if maakondl=74 then N=19;
	if maakondl=78 then N=24;
	if maakondl=82 then N=18;
	if maakondl=84 then N=33;
	if maakondl=86 then N=17;
	if maakondl=784 then N=75;
	if maakondl=795 then N=25;
	Summa_hinnang=kogusumma*N/kokku;
	summa_hinnang_Disp=N*N*(1-Kokku/N)*disp/Kokku;
	/*suhtelised vead iga sim jaoks*/
```

```
      keskmine_hinnang_Disp=(1-Kokku/N)*disp/Kokku;
      summa_suhtv=sqrt(summa_hinnang_Disp)/summa_hinnang;
      keskm_suhtv=sqrt(keskmine_hinnang_Disp)/keskmine;
      run;

proc sql;
      /* Summeerime kihtide hinnangud iga simulatsiooni korral */
      create table Tmp1.result_eesti as
      select replicate,
               sum(summa_hinnang) as summa_hinnang,
               sum(summa_hinnang_Disp) as summa_hinnang_Disp,
               sqrt(sum(summa_hinnang_Disp))/sum(summa_hinnang) as
relerr
      from Tmp1.simul_kokkuv
      group by replicate;
      quit;

proc sql;
      /* Summeerime hinnangud iga maakonna jaoks eraldi */
      create table Tmp1.sql as
      select maakondl,
               sum(summa_hinnang) as summa_hinnang,
               sum(summa_hinnang_Disp) as summa_hinnang_Disp,
               count(*) as count,
               sum(summa_hinnang)/count(*) as summa_hin_maak,
               sum(summa_hinnang_Disp)/count(*) as disp_hinnang,
               sum(keskmine) as kesk,
               sum(keskmine)/count(*) as hinnang_kesk,
               sum(keskmine_hinnang_disp)/count(*) as keskm_disp
      from Tmp1.simul_kokkuv
      group by maakondl;
      quit;

data Tmp1.sql;
      /* Usaldusvahemikud maakonniti */
      set Tmp1.sql;
      lambda_sqrt=1.96*sqrt(disp_hinnang);
      Alumine=summa_hin_maak-lambda_sqrt;
      Ylemine=summa_hin_maak+lambda_sqrt;
      lambda_sqrt_kesk=1.96*sqrt(keskm_disp);
      Alumine_k=hinnang_kesk-lambda_sqrt_kesk;
      Ylemine_k=hinnang_kesk+lambda_sqrt_kesk;
      run;


data Tmp1.simul_disp;
      /* usaldusvahemik yle Eesti */
      set Tmp1.simul_kokkuv;
      disp_hinnang=N*N*(1-Kokku/N)*disp/Kokku;
      disphin_keskmisele=(1-Kokku/N)*disp/Kokku;
      output;
      run;
```

Lisa 3. Kihtide mahud erinevate paigutuste korral suhtelise veaga mitte suurem kui 5%

Võrdeline paigutus: valimimaht N >= 288.3

Maakond	Kokku	Kihimaht
Harju	35	23
Hiiu	6	4
Ida-Viru	53	34
Jõgeva	18	12
Järva	21	14
Lääne	12	8
Lääne-Viru	28	18
Põlva	19	12
Pärnu	27	18
Rapla	14	9
Saare	19	12
Tartu	24	16
Valga	18	12
Viljandi	33	21
Võru	17	11
Tartu linn	25	16
Tallinn	75	49

Kogusummaga võrdeline paigutus: valimimaht N >= 214.09

Maakond	Kokku	Kihimaht
Harju	35	17
Hiiu	6	3
Ida-Viru	53	26
Jõgeva	18	9
Järva	21	10
Lääne	12	6
Lääne-Viru	28	14
Põlva	19	9
Pärnu	27	13
Rapla	14	7

Saare	19	9
Tartu	24	12
Valga	18	9
Viljandi	33	16
Võru	17	8
Tartu linn	25	12
Tallinn	75	36

Lisa 4. Tunnuse kogukulu summa võrdlus üldkogumi ja hinnangu vahel maakonniti

maakond	kogusumma ÜK-s	MC hinnang	vahe	vahe osakaal ÜK-st
Harju	37 063 193	36979377	0,083816	0,23%
Hiiu	3904802	3885268,7	19533,3	0,50%
Ida-Viru	51048158	51124009,1	75851,1	0,15%
Jõgeva	10506289	10503996,3	2292,7	0,02%
Järva	12699106	12671867,9	27238,1	0,21%
Lääne	8314641	8241058,7	73582,3	0,88%
Lääne-Viru	22721093	22876370,8	155277,8	0,68%
Põlva	8804062	8719526,9	84535,1	0,96%
Pärnu	31392360	31414934,5	22574,5	0,07%
Rapla	10852825	10888946,2	36121,2	0,33%
Saare	12033893	11979214,3	54678,7	0,45%
Tartu	15303451	15457390,2	153939,2	1,01%
Valga	9976499	9987412,5	10913,5	0,11%
Viljandi	18603502	18525482,1	78019,9	0,42%
Võru	11452781	11384373,5	68407,5	0,60%
Tallinn	132300333	132440548,4	140215,4	0,11%
Tartu linn	32551030	32634764,2	83734,2	0,26%

Lisa 5. Kogukulu summa usaldusvahemik

maakond	Usaldusvahemik	
	alumine piir	ülemine piir
Harju	23 475 225,57	50 483 528,44
Hiiu	979 642,58	6 790 894,87
Ida-Viru	32 406 755,90	69 841 262,20
Jõgeva	6 385 616,36	14 622 376,19
Järva	3 151 452,03	22 192 283,76
Lääne	1 609 310,94	14 872 806,45
Lääne-Viru	12 300 781,83	33 451 959,71
Põlva	4 999 968,04	12 439 085,86
Pärnu	22 975 300,03	39 854 568,95
Rapla	2 587 981,72	19 189 910,73
Saare	4 468 609,62	19 489 818,94
Tartu	5 252 057,69	25 662 722,74
Valga	5 302 637,76	14 672 187,15
Viljandi	10 653 093,71	26 397 870,57
Võru	3 301 839,86	19 466 907,07
Tallinn	119 289 477,94	145 591 618,80
Tartu linn	20 376 001,43	44 893 526,93

Lisa 6. Tunnuse kogukulu keskmise võrdlus üldkogumi ja hinnangu vahel maakonniti

maakond	ÜK keskmine	MC hinnang keskmisele	vahe osakaal ÜK-st %
Harju	1058948,37	1056553,63	0,23%
Hiiu	650800	647544,79	0,50%
Ida-Viru	963172,792	964603,94	0,15%
Jõgeva	583682,722	583555,35	0,02%
Järva	604719,333	603422,28	0,21%
Lääne	692886,75	686754,89	0,88%
Lääne-Viru	811467,607	817013,24	0,68%
Põlva	463371,684	458922,47	0,96%
Pärnu	1162680	1163516,09	0,07%
Rapla	775201,786	777781,87	0,33%
Saare	633362,789	630484,96	0,45%
Tartu	637643,792	644057,93	1,01%
Valga	554249,944	554856,25	0,11%
Viljandi	563742,485	561378,25	0,42%
Võru	673693	669669,03	0,60%
Tallinn	1764004,44	1765873,98	0,11%
Tartu linn	1302041,2	1305390,57	0,26%

Lisa 7. Tunnuse kogukulu keskmise usaldusvahemikud maakonniti

maakond	Usaldusvahemik	
	alumine piir	ülemine piir
Harju	670 720,73	1 442 386,53
Hiiu	163 273,76	1 131 815,81
Ida-Viru	611 448,22	1 317 759,66
Jõgeva	354 756,46	812 354,23
Järva	150 069,14	1 056 775,42
Lääne	134 109,24	1 239 400,54
Lääne-Viru	439 313,64	1 194 712,85
Põlva	263 156,21	654 688,73
Pärnu	850 937,04	1 476 095,15
Rapla	184 855,84	1 370 707,91
Saare	235 189,98	1 025 779,94
Tartu	218 835,74	1 069 280,11
Valga	294 590,99	815 121,51
Viljandi	322 821,02	799 935,47
Võru	194 225,87	1 145 112,18
Tallinn	1 590 526,37	1 941 221,58
Tartu linn	815 040,06	1 795 741,08

Lisa 8. Tunnuse OÜ summa võrdlus üldkogumi ja hinnangu vahel maakonniti

maakond	ÜK OÜ kogusumma	MC hinnang OÜ kogusummale	vahe osakaal ÜK-st %
Harju	19,00	19,03	0,14%
Hiiu	6,00	6,00	0,00%
Ida-Viru	22,00	21,88	0,53%
Jõgeva	7,00	6,92	1,13%
Järva	4,00	4,19	4,79%
Lääne	2,00	2,07	3,50%
Lääne-Viru	18,00	18,26	1,42%
Põlva	5,00	5,19	3,74%
Pärnu	10,00	10,03	0,31%
Rapla	7,00	6,73	3,80%
Saare	8,00	8,01	0,17%
Tartu	17,00	16,80	1,20%
Valga	5,00	4,84	3,16%
Viljandi	10,00	9,97	0,34%
Võru	6,00	6,16	2,71%
Tallinn	59,00	59,07	0,12%
Tartu linn	21,00	21,02	0,10%

Lisa 9. Tunnuse OÜ summa usaldusvahemikud maakonniti

maakond	Usaldusvahemik	
	Alumine	Ülemine
Harju	9,74	28,31
Hiiu	6,00	6,00
Ida-Viru	16,14	27,63
Jõgeva	-4,66	18,50
Järva	-2,41	10,80
Lääne	-4,01	8,15
Lääne-Viru	8,63	27,88
Põlva	-6,44	16,82
Pärnu	8,26	11,80
Rapla	0,76	12,71
Saare	-0,35	16,37
Tartu	8,01	25,58
Valga	-5,91	15,59
Viljandi	-6,80	26,73
Võru	-1,05	13,38
Tallinn	56,83	61,31
Tartu linn	18,28	23,76

Lisa 10. Tunnuse OÜ keskmise võrdlus ÜK ja hinnangute vahel maakonniti

maakond	ÜK OÜ keskmine	MC hinnang OÜ keskmisele	vahe osakaal ÜK-st %
Harju	0,54	0,54	0,67%
Hiiu	1	1,00	0,00%
Ida-Viru	0,42	0,41	0,53%
Jõgeva	0,39	0,38	1,13%
Järva	0,19	0,20	4,79%
Lääne	0,17	0,17	3,50%
Lääne-Viru	0,64	0,65	1,42%
Põlva	0,26	0,27	3,74%
Pärnu	0,37	0,37	0,31%
Rapla	0,5	0,48	3,80%
Saare	0,42	0,42	0,17%
Tartu	0,71	0,70	1,20%
Valga	0,28	0,27	3,16%
Viljandi	0,3	0,30	0,34%
Võru	0,35	0,36	2,71%
Tallinn	0,79	0,79	0,12%
Tartu linn	0,84	0,84	0,10%

Lisa 11. Tunnuse OÜ keskmise usaldusvahemikud maakonniti

maakond	Usaldusvahemik	
	alumine piir	ülemine piir
Harju	0,28	0,81
Hiiu	1,00	1,00
Ida-Viru	0,30	0,52
Jõgeva	-0,26	1,03
Järva	-0,11	0,51
Lääne	-0,33	0,68
Lääne-Viru	0,31	1,00
Põlva	-0,34	0,89
Pärnu	0,31	0,44
Rapla	0,05	0,91
Saare	-0,02	0,86
Tartu	0,33	1,07
Valga	-0,33	0,87
Viljandi	-0,21	0,81
Võru	-0,06	0,79
Tallinn	0,76	0,82
Tartu linn	0,73	0,95

Lisa 12. Kogukulu keskmine ja dispersioon maakonniti (EEK)

Maakond/Linn	Keskmine	Dispersioon
Harju	1058948,37	5,50E+11
Hiiu	650800,33	1,85E+11
Ida-Viru	963172,79	2,62E+12
Jõgeva	583682,72	3,20E+10
Järva	604719,33	3,51E+11
Lääne	692886,75	1,98E+11
Lääne-Viru	811467,60	2,82E+11
Põlva	463371,68	2,19E+10
Pärnu	1162680	5,47E+12
Rapla	775201,79	5,22E+11
Saare	633362,79	2,16E+11
Tartu	637643,79	2,88E+11
Valga	554249,94	3,77E+10
Viljandi	563742,49	4,83E+10
Võru	673693	3,08E+11
Tallinn	1764004,44	5,78E+12
Tartu linn	1302041,2	2,77E+12

Lisa 13. Monte-Carlo meetodi simuleerimise teel saadud keskmised kogukulu summa hinnangu suhtelised vead maakonniti

Maakond/Linn	Keskmine suhtelina viga
Harju	18%
Hiiu	16%
Ida-Viru	17%
Jõgeva	14%
Järva	22%
Lääne	25%
Lääne-Viru	20%
Põlva	18%
Pärnu	13%
Rapla	29%
Saare	22%
Tartu	24%
Valga	17%
Viljandi	15%
Võru	22%
Tallinn	5%
Tartu linn	18%