TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Jekaterina Prostakova

Mittevastamine ja selle kompenseerimine

Bakalaureusetöö

Juhendaja: Natalja Lepik,
doktorant

Tartu 2007

Sisukord

Sissejuhatus

Käesolev bakalaureusetöö on kirjutatud kõikse uuringu ja valikuuringute teooria alal. Uuritud probleemiks on mittevastamine ja meetodid selle kompenseerimiseks. Mittevastamine tähendab seda, et tunnuse väärtuseid pole võimalik saada mõningatelt valimi objektidelt, mis on valitud uuringu teostamise jaoks. Tavaliselt eristatakse kahte liiki kadu: tunnuse väärtuse kadu, st informatsioon puudub väärtuse tasemel, ja objekti kadu, milleks on terve objekti puudumine. Käesoleva töö ettevalmistamisel on kasutatud materjale erinevatest infoallikatest. Töö põhiteksti võib loogiliselt jaotada kaheks osaks: teoreetiliseks ja praktiliseks.

Teoreetiline osa koosneb neljast peatükist. Esimeses peatükis on antud lühike ülevaade mittevastamisest ja selle erinevatest liikidest. Peatükkides 2 ja 3 tutvustatakse olemasolevaid kao kompenseerimise meetodeid kõikses ja valik- uuringus. On märkimisväärne, et kolmas peatükk annab lugejale võimaluse tutvuda nüüdisaegsete arengutega valikuuringute teoorias, mis on esitatud C.-E. Särndali ja S. Lundströmi töödes [2]. Neljandas peatükis käib juttu valimi võtmise meetoditest valikuuringutes.

Praktiline osa on tellitud Sotsiaalministeeriumi Terviseinfo- ja analüüsiosakonna poolt. Uuring on tehtud andmetega, mis on saadud 2004. aasta statistilise tervishoiualase majandustegevuse aruannete põhjal. Sotsiaalministeerium püstitas uurimiseks kaks probleemi: esiteks, kuna andmed pole täielikud ja sisaldavad suurt mittevastamise protsenti, siis leida kõige sobivamaid meetodeid kao kompenseerimiseks, ja teiseks, kuna senini seda tüüpi uuringud on iga aasta olnud kõiksed ja väga kulukad, siis tekkis vajadus üle minna valikuuringu peale. Seega teiseks probleemiks on sobiva valikudisaini koostamine ning kao kompenseerimise meetodite kirjeldamine tulevase valikuuringu jaoks. Praktilises osas kõigepealt esitatakse andmestiku kirjeldust, seejärel rakendatakse teoreetilises osas kirjeldatud meetodid tervele üldkogumile ja üldkogumist võetud valimile.

1 Mittevastamine ja selle liigid

Mittevastamine on küllaltki tähtis probleem tänapäevastes uuringutes. Mittevastamine tähendab seda, et tunnuste väärtuseid pole võimalik saada mõningatelt valimi objektidelt, mis on valitud uuringu teostamise jaoks. Sellel on palju erinevaid põhjuseid, näiteks, inimest pole kodus küsitluse ajal või on ta hoopis kolinud mujale ja teda on raske üles leida. Võib ka juhtuda nii, et küsimustik ei ole hästi läbi mõeldud ning küsitletav ei leia sobilikku vastust või keeldub vastamisest seoses ükskõiksusega või privaatsuse rikkumisega. Posti teel läbiviidavate küsitluste korral on peamiseks mittevastamise põhjuseks tagastamata küsimustikud.

Mittevastamise põhjuseid on palju ja nad põhjustavad eri liiki kadu andmetes [3]:

- tunnuse väärtuse kadu (ehk kadu väärtuse tasemel) - sel juhul puudub vaadeldaval objektil mõne tunnuse väärtus (vt. tabel 1.1), näiteks jätavad küsitletavad sageli vastamata tundlikele küsimustele;
- objekti kadu (ehk kadu objekti tasemel) - sellel korral puudub andmestikust terve objekt, st puuduvad selle objekti kõigi tunnuste väärtused.

Mittevastamise tagajärjed võivad erineda. Üks näide on hinnangute suurem dispersioon, mis on tegeliku valimimahu vähendamise tulemus. Teine näide on nihkega hinnangud, juhul kui vastanute ja mittevastanute karakteristikud erinevad palju. Nihe aga tähendab seda, et valitud objektid ei saa enam esindada tervet üldkogumit. Veelgi enam, tekib risk tunduvalt alahinnata tegelikku dispersiooni, kui omistamisega tekitatud andmed on vaadeldavad nagu oleks nad saadud küsitletud inimeste käest.

Tabel 1.1 Väärtuse kao illustratsioon hüpoteetilises andmestikus. Mittevastamine on tähistatud mv, vastamine aga ×.

Registritunnused			Uuritavad tunnused		
ID	1	2	1	2	3
1	×	×	×	×	×
2	×	×	×	×	mv
3	×	×	×	mv	×
4	×	×	×	×	×
5	×	×	×	×	×
6	×	×	mv	×	mv
7	×	×	mv	mv	mv

On olemas mitmeid suundi objekti kao kompenseerimiseks. Israel (1992) oma artiklis pakub

järgmisi võimalusi [4]:

1) Valede andmete väljajätmine uuringust. See meetod tundub olevat mõistlik, kui uurija loeb andmed kehtetuks.
2) Tulemuste üldistus ainult vastanute grupile, kus püütakse vältida järelduste tegemist tervele üldkogumile.
3) Tulemuste üldistus tervele üldkogumile oletusel, et mittevastamisest tekkinud nihe on väga väike või ei eksisteeri üldse. Kui uurija teab üldkogumit hästi ja andmete uurimisel ei avasta silmnähtavaid nihkeid, siis sobib see strateegia hästi.
4) Mittevastanute kordne küsitlus, kus mittevastanuid vaadeldakse uue kogumina ning sellest võetakse uus valim. Seejärel püütakse saada kõiki vastuseid uues valimis ning võrreldakse karakteristikute käitumist vastanutel ja mittevastanutel.
5) Juhul kui on avastatud märkimisväärseid erinevusi varem- ja hiljem-vastanute vahel, siis eelistatakse hiljem-vastanuid.
6) Kontakti leidmise katsete arvu ja jõupingutuse suurendamine. Kuigi see võtab aega ja toob kaasa lisakulutusi, on vastanute osakaalu suurenemine parim vahend nihke vältimiseks.
7) Mittevastanute käitumise modelleerimine, kus püütakse leida mudel, mis võimaldab uurida, millest sõltub mittevastamine.

Nüüd vaatleme eraldi väärtuse ja objekti kao kompenseerimist.

2 Väärtuse kao kompenseerimine

Kui tegemist on kaoga väärtuse tasemel, siis selle kompenseerimiseks sobivad nn *omistusmeetodid*. Omistusmeetodite üldine eesmärk on lünkadeta andmestiku saamine, mis on vajalik paljude andmetöötlusprogrammide kasutamiseks [3]. Sageli jäetakse lünkadega objektid andmestikust välja (seda teeb enamus tarkvaradest). Kuid selle tulemuseks võib aga olla liiga väike valimimaht. Et seda vältida, võib asendada puuduvad väärtused andmestikus hinnanguliste väärtustega, kasutades selleks sobivat omistusmeetodit. Omistust kasutatakse tavaliselt objektide üksikute puuduvate väärtuste asendamiseks. Ka objekti puudumisel tervikuna valimist on tema kõiki väärtusi võimalik omistusmeetoditega prognoosida. Kuid seda pole soovituslik teha, kuna mittevastanute grupp võib väga erineda vastanute omast. Kui puuduvate objektide arv on suur, kasutatakse omistusmeetodite asemel nn *kaalumismeetodeid* [2].

Tavaliselt eristatakse kahte omistusmeetodite liiki: *deterministlikud* ja *stohhastilised* [5]. Deterministlikud meetodid annavad alati ühte ja sama väärtust objektidele ühesuguste karakteristikutega. Stohhastiliste meetodite rakendamisel mittevastanud objektid saavad omandada erinevaid väärtusi.

Omistusmeetodite üheks puuduseks on see, et kuna puuduvate väärtuste arvutamisel kasutatakse ainult neid andmeid, mis on kogutud vastanutelt ja mis võivad erineda kao hulga näitajatest, siis see viib nihkele hinnangutes. Teiseks, hinnangute dispersioon on sageli alahinnatud. Kuna puuduvad väärtused on tihti erandlikud, mida küsitletavad inimesed ei taha avaldada ühel või teisel põhjusel, siis nende asendamine teiste väärtustega, mis on vähem erandlikud ja mis on lähedasemad keskmisele juhule, muudab dispersiooni väiksemaks.

Traat ja Inno [3] ning Durrant [5] eristavad järgmisi omistusmeetodeid.

1. *Üldine keskmine omistus*. Selle meetodi korral vaadeldava tunnuse kõigile puuduvatele väärtustele omistatakse selle tunnuse vastanute hulga keskmine, mis silmnähtavalt ei sobi nominaaltunnuste jaoks. Meetod on väga lihtne, kuid selline lähenemine viib sageli kohatutele hinnangutele ja dispersiooni alahindamisele.
2. *Klassi keskmise omistus*. Kui on teada mõned tunnused kõigi objektide korral, siis saab nende tunnuste sarnaste väärtuste alusel jagada valim omistusklassideks. Igas klassis omistatakse vaadeldava tunnuse puuduvatele väärtustele klassi vastanute keskmine. Omistuse efektiivsus sõltub klasside moodustamise õnnestumisest, sest uuritava tunnusega tugevalt korreleeritud

tausttunnuste kasutamisel saab vähendada kaost põhjustatud nihet hinnangutes. Meetodi puudusteks on see, et dispersiooni alahindamise probleem jääb, ja korrelatsioon mõnede tunnuste vahel võib olla moondunud.

3. *Hot-deck ja Cold-deck omistus*. Hot-deck omistuse korral omistatakse puuduvale väärtusele (ingl *recipient*) sama andmestiku mõne teise objekti (doonori) väärtus. Selle meetodi kasutamisel tuleb võimalikult hästi valida doonorit. Cold-deck omistuse korral leitakse omistusväärtus muudest allikatest.

Hot-deck meetodi eeliseks on reaalselt eksisteerivate väärtuste kasutamine omistuse korral. Sellepärast on selline omistus laialt kasutatav praktikas, isegi siis, kui on tegemist kvalitatiivsete andmetega.

Hot-deck meetodit realiseeritakse üldise juhusliku omistuse, klassides juhusliku omistuse, järjestikuse Hot-deck omistuse ja kaugusfunktsioonijärgse omistuse abil. Tuleb silmas pidada, et niisuguse omistamisega säilitatakse olemasoleva andmestiku varieeruvus, milles aga ei kajastu kao võimalikult suurem varieeruvus.

4. *Üldine juhuslik omistus*. Puuduvale väärtusele omistatakse vastanute seast juhuslikult valitud objekti väärtus.

5. *Juhuslik omistus klassis*. Moodustatakse omistusklassid. Puuduv tunnuseväärtus asendatakse selle tunnuse juhuslikult valitud väärtusega samas klassis.

6. *Järjestikune Hot-deck omistus*. Kõik valimi objektid järjestatakse ja seejärel läbitakse. Puuduva väärtuse korral omistatakse sellele järjekorras eelneva samasse klassi kuuluva objekti olemasolev väärtus.

7. *Kaugusfunktsioonijärgne omistus*. Tausttunnuste abil defineeritakse objektidevahelised kaugused. Puuduv tunnusväärtus asendatakse talle kauguse poolest lähima vastanud objekti väärtusega. Üheks eeliseks on see, et selle meetodi rakendamisel kasutatakse reaalselt eksisteerivaid andmeid. Samal ajal mõned väärtused saavad olla kasutatud mitu korda omistuse jaoks, kui reas esineb rohkem kui üks puuduv väärtus. Seega tekib variatsiooni täispuhumise probleem.

8. *Regressioonijärgne omistus*. Regressioonijärgse omistuse korral kasutatakse tuunustevahelist sõltuvust. Kasutades vastanud objektide andmeid moodustatakse regressioonivõrrand ning selle abil prognoositakse puuduvad väärtused. Arvtunnuse väärtuse prognoosimiseks kasutatakse tihti lineaarset regressiooni, samal ajal, kui tegemist on kvalitatiivse tunnusega, siis kasutatakse rohkem logistilist regressiooni. Regressioonijärgse omistuse potentsiaalseks puuduseks on see,

et dispersiooni suurus ja korrelatsioon tunnuste vahel, mis ei ole kasutatud regressioonimudelis, on moondunud.

9. *Mitmene omistus*. Mitmese omistuse korral omistatakse igale puuduvale väärtusele mitu, näiteks *m* väärtust. Omistatud väärtused järjestatakse ja saadakse *m* andmestikku, mida analüüsitakse standardsete vahenditega. Selle tulemusena leitakse kombineeritud hinnang koos dispersioonihinnanguga. Selle meetodi oluliseks puuduseks on see, et töö maht on küllaltki suur.

10. *Objekti asendus*. Üheks omistusmeetodiks võib pidada ka objekti asendust, kus mittevastanu asemel võetakse uuringusse uus objekt. Asenduse puhul on tähtis, et kao objekt asendatakse talle võimalikult sarnase objektiga (kõigi uuritavate tunnuste osas).

3 Kao kompenseerimine valikuuringutes

3.1 Objekti kadu valikuuringutes

Valikuuring on statistiline uuring, milles otsustused üldkogumi kohta tehakse valimi baasil. Andmeid kogutakse siin üksnes valimilt. Valikuuringul on võrreldes kõikse uuringuga rida eeliseid, sh väiksem maksumus, suurem kiirus, laiem rakendatavus ja suurem täpsus.

Järgmisena vaatleme objekti kao kompenseerimist valikuuringutes. Särndal ja Lundström [2] nimetavad neid meetodeid *kaalumismeetoditeks*, mis tähendab seda, et vastanutelt saadud uuritava tunnuse väärtustele omistatakse kaalusid, mis peavad kompenseerima mittevastamist. Üks tuntuim tehnika kaalude arvutamiseks on kalibreerimine. On olemas mitu lähenemist kalibreeritud kaalude arvutamisele, kuid peaaegu kõik need lähenemisviisid kasutavad teadaolevat abiinformatsiooni üldkogumi kohta, selleks on näiteks riikide registrid või muud allikad. Abiinformatsiooni efektiivne kasutamine on hea võimalus usaldusväärse hinnangu saamiseks kao olemasolu korral.

3.2 Vastamistõenäosustega mudel

Olgu s valim üldkogumist, mis on saadud teatava tõenäosusliku valikumeetodiga (valikumeetoditest on põhjalikumalt kirjutatud peatükis 4). Tähistame vastanute hulka r. Oletame, et igal objektil on kindel vastamistõenäosus θ_k, kusjuures $0 \le \theta_k \le 1$ iga k jaoks. Kuna vastamise mehhanismi pole teada, siis ka θ_k pole võimalik välja arvutada. Tähistame $d_k = 1/\pi_k$ objekti k kaalu (sõltub valikudisainist) ja $\phi_k = 1/\theta_k$ on objekti k vastamise mõju. Siis saame objektile uue kaalu, korrutades disaini kaalu objekti vastamise mõjuga

$$d_k \cdot \phi_k = \frac{1}{\pi_k \theta_k}. \tag{3.2.1}$$

Siis avaldub kogusumma hinnang järgmiselt:

$$\hat{t}_\theta = \sum_r d_k \phi_k y_k , \tag{3.2.2}$$

ja see on nihketa hinnang tunnuse y kogusummale. Mudelit (3.2.1)-(3.2.2) nimetatakse vastamistõenäosustega mudeliks. Mudeli puuduseks on aga see, et vastamistõenäosused ei ole tegelikult teada, ja siis saab kasutada üksnes hinnangulisi vastamistõenäosusi. Praktikas

kasutatakse võimalikult lihtsaid mudeleid. Üheks niisuguseks mudeliks on homogeensete vastamisgruppidega mudel [3], mille korral jagatakse kõiki valimi objekte gruppideks nii, et vastamistõenäosused oleksid konstantsed grupi sees. Seda tehakse tavaliselt abiinformatsiooni kasutades, näiteks moodustatakse vanusegrupid ja eeldatakse, et vastamistõenäosus on igas grupis sama, kuid varieerub grupiti.

Eeldame, et konkreetse objekti vastamine ei sõltu teiste objektide vastamisest. Tähistame grupist h saadud vastanute hulka r_h. Olgu grupi h valimi ja vastanute hulga mahud vastavalt n_h ja m_h. Konstantse vastamistõenäosuse korral on vastanute hulga r_h tekkimine vaadeldav Bernoulli valikuna kaasamistõenäosusega θ_h. Vastanute hulga maht m_h on binoomjaotusega juhuslik suurus parameetritega n_h ja θ_h, seega vastamistõenäosuse nihketa hinnanguks on $\hat{\theta}_h = m_h / n_h$. Järelikult kogusumma hinnang on kujul

$$\hat{t}_\theta = \sum_{h=1}^{H} \frac{n_h}{m_h} \sum_{r_h} d_k y_k \ . \tag{3.2.3}$$

On olemas ka teisi vastamistõenäosuse mudelite meetodeid.

3.3 Kalibreerimishinnang

Särndali ja Lundströmi [2] poolt on pakutud lähenemine kalibreerimisele, mille korral abiinformatsioon on esitatud abitunnuste vektori abil. Selle vektori väärtused on teada iga elemendi $k \in r$ jaoks, kuid lisaks sellele on teada ka informatsioon suuremal kui r hulgal (näiteks kogusummade näol).

Olenevalt uuringust võib esineda kolm erinevat tüüpi abitunnuste vektoreid, mida Särndal ja Lundström (2005) [3] tähistavad InfoU, InfoS, InfoUS.

1) InfoU. Informatsioon on saadav kogu üldkogumi U tasemel. Tähistame vastavat abivektorit $\mathbf{x}_k^*$, olgu selle dimensiooniks $J^* \geq 1$. Antud vektori jaoks kehtib:
   i. iga $k \in r$ korral on teada vektori $\mathbf{x}_k^*$ väärtust;
   ii. on teada $\sum_{\mathbf{U}} \mathbf{x}_k^*$.

2) InfoS. See on informatsioon, mis on kättesaadav valimi s tasemel, kuid mitte populatsiooni U tasemel. Olgu $\mathbf{x}_k^\circ$ niisugune abitunnuste vektor dimensiooniga $J^\circ \geq 1$, mille korral:
   i. iga $k \in s$ korral on teada vektori $\mathbf{x}_k^\circ$ väärtust, kuigi samal ajal $\sum_{\mathbf{U}} \mathbf{x}_k^\circ$ ei ole teada;
   ii. iga $k \in r$ korral on teada $\mathbf{x}_k^\circ$.
3) InfoUS. Informatsioon on saadaval mõlemal tasemel. InfoU ja InfoS'i kombinatsiooni saab kasutada kaalude väljaarvutamisel. Abiinformatsioon on kujul $\mathbf{x_k} = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$.

Tingimused i. ja ii. InfoU-s ja InfoS-s on minimaalsed tarvilikud tingimused punkthinnangu arvutamise protseduuride jaoks, mis on selgeks tehtud allpool.

Kogusumma

$$t = \sum_U y_k \tag{3.3.1}$$

kalibreerimishinnangu üldised soovitud omadused on järgmised:

i. väike nihe;
ii. väike dispersioon;
iii. kaalude süsteem, mis annab abimuutujale rakendades teadaoleva summaarse info;
iv. kaalude süsteem, mis on sama hea iga y-tunnuse kogusumma hindamisel suurest kogumist.

Kui tähistada kalibreeritud kaalu $w_k = d_k\, v_k$ ($k \in r$), siis avaldub kalibreerimishinnang $\hat{t}_y$ järgmiselt:

$$\hat{t}_y = \sum_r w_k\, y_k\,, \tag{3.3.2}$$

sealjuures v_k on kordaja, millega disaini kaal kalibreeritakse.

Kordaja v_k võimaldab

- määrata disaini kaalu iga $k \in r$ jaoks;
- viia sisse abiinformatsiooni;
- võimalikult vähendada mittevastamisest põhjustatud nihet.

Vaatleme kõige üldisemat juhtu, InfoUS, kus abitunnuste vektor on kujul $\mathbf{x_k} = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$ dimensiooniga $J^* + J^\circ$.

Kalibreerimise võrrand, mille abil on leitud kaalude süsteem w_k iga $k \in r$ on järgmine:

$$\sum_r w_k \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s \mathbf{x}_k^\circ \end{pmatrix}. \tag{3.3.3}$$

Kui $\mathbf{x}_k$ ja v_k vahel eksisteerib lineaarne sõltuvus, nimelt $v_k = 1 + \boldsymbol{\lambda}'\mathbf{x}_k$, siis vektori $\boldsymbol{\lambda}$ arvutamise ülesanne muutub küllaltki lihtsaks. Kui paneme võrrandi (3.3.3) sisse, saame, et sellistel tingimustel

$$\boldsymbol{\lambda}' = \boldsymbol{\lambda}_r' = \left[\begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s \mathbf{x}_k^\circ \end{pmatrix} - \sum_r d_k \mathbf{x}_k \right]' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1}. \tag{3.3.4}$$

Selle meetodi eelised:

1. Saab näidata, et kui uuritava tunnuse ja abitunnuse vahel on ideaalne lineaarne seos, siis kalibreerimishinnang $\hat{t}_y$ annab õige parameetri kätte. See tähendab seda, et mida tugevam on seos $\mathbf{x}_k$ ja v_k vahel, seda parem on kogusumma hinnang $\hat{t}_y$.
2. $\hat{t}_y$ avaldis esindab suurt hinnangute peret, vastavalt erinevatele $\mathbf{x}_k$ esitusviisidele. Paljud nendest hinnangutest on väga hästi uuritud ja on sageli kasutatavad praktikas.
3. Märkimisväärne on ka see, et kalibreerimise kasutamisel praktikas algebralise hinnangu valemi tuletus konkreetse $\mathbf{x}_k$ vektori jaoks pole vajalik. Tähelepanu pööratakse kõigepealt arvutamistele, mis on teostatud olemasoleva arvutitarkvaraga. Kasutajal tuleb täpselt määratleda vektor $\mathbf{x}_k$ ja vastav informatsiooni sisend. Tarkvara arvutab kalibreeritud kaalud ja vastava kalibreerimishinnangu $\hat{t}_y$ välja.

Selle meetodi puudused:

1. Praktikas on tavaliselt tegemist mitme uuritava tunnusega ühe asemel. Mitte alati pole lihtne leida võimsat abiinformatsiooni, mis seletaks kõiki uuritavaid tunnuseid.
2. Mitte igas riigis pole olemas häid registreid, kust saab abiinformatsiooni võtta.
3. Võib tekkida probleem kaalude v_k interpreteerimisel, kui v_k juhtuvad olla negatiivsed.

Eespool kirjeldatud hinnang ning meetodid abiinformatsiooni valimiseks on uuritud Toompere (2006) bakalaureusetöös.

Praktilistes uuringutes leidub sageli kaks või rohkem kvalitatiivset abitunnust ehk faktorit. Vaatleme juhtu, kui tegemist on kolmefaktorilise klassifikatsiooniga. Oletame, et esimesel faktoril on P taset, mis on tähistatud $p = 1, ..., P$, teisel faktoril H taset, kus $h = 1, ..., H$, ja kolmandal faktoril on Q taset, $q = 1, ..., Q$. Olgu $(\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk}, \delta_{1k}, ..., \delta_{hk}, ..., \delta_{Hk}, \zeta_{1k}, ..., \zeta_{qk}, ..., \zeta_{Qk})'$ indikaatorvektor väärtustega 0 ja 1, kusjuures $\gamma_{pk} = 1$, kui objekt k kuulub gruppi h, ja 0 vastasel juhul (δ_{hk} ja ζ_{qk} on defineeritud samamoodi). Juhul InfoU abivektor avaldub kujul

$$\mathbf{x}_k = \mathbf{x}_k^* = (\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{P-1,k}, \delta_{1k}, ..., \delta_{hk}, ..., \delta_{H-1,k}, \zeta_{1k}, ..., \zeta_{qk}, ..., \zeta_{Q-1,k})'$$

dimensiooniga $P + H + Q - 3$. Abivektori dimensiooni tuleb vähendada, et pöördmaatriks valemis (3.3.4) eksisteeriks (viimased väärtused igas grupis on avaldatavad eelmiste väärtuste kaudu). Siinjuures $\sum_U \mathbf{x}_k^*$ on vektor, mis koosneb marginaalsetest sagedustest,

$$N_{p\bullet\bullet} = \sum_{h=1}^{H}\sum_{q=1}^{Q} N_{phq}, p = 1,...,P, \; N_{\bullet h\bullet} = \sum_{p=1}^{P}\sum_{q=1}^{Q} N_{phq}, h = 1,...,H, \; N_{\bullet\bullet q} = \sum_{p=1}^{P}\sum_{h=1}^{H} N_{phq}, q = 1,...,Q-1,$$

kus N_{phq} on gruppi phq kuuluvate objektide arv.

3.4 Järelkihistamine kui meetod kao kompenseerimiseks

See meetod on laialt tuntud ja on kirjeldatud nt Traat ja Inno [3] poolt. Meetodi järgi tuleb moodustada järelkihid nii, et

- nad oleksid uuritava tunnuse suhtes võimalikult homogeensed ehk lahutaksid üldkogumit võimalikult erinevatesse gruppidesse;
- nad võimaldaksid jagada kadu nendesse kihtidesse (st kaol peab olema kihti määrav tunnus);
- oleksid teada abitunnuse kogusummad üldkogumi tasemel.

Järelkihistamise hinnang näeb välja järgmiselt:

$$\hat{t} = \sum_{h=1}^{H} t_{hx} \cdot \frac{\hat{t}_{hy}}{\hat{t}_{hx}} = \sum_r w_i' y_i, \tag{3.4.1}$$

kus

$$w_i' = w_i \cdot \frac{t_{hx}}{\hat{t}_{hx}}, \tag{3.4.2}$$

kus hinnangud on arvutatud vastanute pealt. Raskeim probleem siinkohal on saada info kao kohta, et teda kihtidesse jagada.

3.5 Regressiooni ja suhte tüüpi hinnangud

Särndal ja Lundström [2] toovad eraldi välja kaks hinnangut, nn *vabaliikmega* (ehk regressiooni-) ja *vabaliikmeta* (ehk suhte-) hinnangut, mida on võimalik tuletada eespool esitatud kalibreerimishinnangust. Teiselt poolt on need kaks hinnangut vaadeldavad nn *üldistatud regressioonihinnangu* (GREG) erijuhtudena. Juhul, kui on tegemist täieliku vastamisega, st $r = s$, on GREG hinnang identne eespool esitatud kalibreerimishinnanguga. See omadus võimaldab tõmmata paralleeli GREG hinnanguga, mille nihe on väga lähedane nullile ja millel on väike varieeruvus.

Olgu antud üks kvantitatiivne abitunnus $x = (x_1,...,x_N)'$ ja abiinfo olgu teada üldkogumi tasemel (InfoU). Sellisel juhul on kogusumma hinnanguks järgmine suhtehinnang [2, lk. 72]:

$$\hat{t}_{RA} = \left(\sum_U x_k\right).\frac{\sum_r d_k y_k}{\sum_r d_k x_k}. \tag{3.5.1}$$

InfoS juhul kogusumma hinnanguks on regressiooni tüüpi hinnang [2, lk. 72]

$$\hat{t}_{REG} = N\{\bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d})B_{r;d}\}, \tag{3.5.2}$$

kus

$$\bar{x}_{r;d} = \sum_r d_k x_k / \sum_r d_k\,, \tag{3.5.3}$$

$\bar{y}_{r;d}$ analoogiliselt ja

$$B_{r;d} = \frac{\sum_r d_k (x_k - \bar{x}_{r;d})(y_k - \bar{y}_{r;d})}{\sum_r d_k (x_k - \bar{x}_{r;d})^2}. \tag{3.5.4}$$

Nihke suhtes, mis on tingitud mittevastamise poolt, on regressiooni tüüpi hinnang (3.5.2) parem kui suhtehinnang (3.5.1). Suhtehinnang on valikuuringutes väga populaarne oma lihtsuse tõttu. Selle hinnangu omadusi on samuti palju uuritud. Kuid suhtehinnang on üles ehitatud nii, et see kasutab valimi kõiki objekte ehk täielikku vastamist. Mittevastamise korral on suhtehinnangu kasutamine mõnevõrra riskantne. Et $\hat{t}_{RA}$ oleks peaaegu nihketa, peavad vastamistõenäosused rahuldama väga ranget tingimust: nad peavad olema võrdsed kogu populatsiooni ulatuses.

Siiani oleme valemites kasutanud mõistet *disaini kaal* (d_k), kuid pole andnud valemeid selle leidmiseks. Järgmises peatükis vaatleme erinevaid valikuuringute liike ja seda, kuidas avaldub kaal konkreetse valikudisaini korral.

4 Valimi võtmise meetodeid valikuuringutes

Valikuuring on statistiline uuring, milles otsustused üldkogumi kohta tehakse valimi baasil. Andmeid kogutakse üksnes valimilt. Kogutud andmeid kasutatakse hinnangute leidmisel. Tavaliselt tehakse seda, kasutades *disaini kaalu* d_k. Disaini kaal näitab, mitmele üldkogumi objektile valimiväärtus y_k laiendatakse, et üle valimi summeerimisel saada üldkogumi kogusummat.

Valikuuringus kasutatakse valimi määramiseks erinevaid statistilisi meetodeid. Antud töö praktilises osas on kasutatud lihtne juhuslik kihtvalik, seepärast vaatleme seda disaini põhjalikumalt.

Kihtvalik on praktikas enim kasutatud disain, mis sobib valikuuringu kohandamiseks praktilises elus esinevatele, sageli üpris keerulistele tingimustele. Tähistades kihis *h* objektide arvu N_h ja valimi objektide arvu sellest kihist n_h, avalduvad üldkogumi ja kogu valimi mahud järgmiste summade kaudu:

$$N = \sum_{h=1}^{H} N_h \text{ ja } n = \sum_{h=1}^{H} n_h .$$

Lihtsa juhusliku kihtvaliku korral teostatakse igas kihis lihtne juhuvalik. Olgu $f_h = \frac{n_h}{N_h}$ *h*-nda kihi valikusuhe. Siis kehtib järgmine teoreem [3].

Teoreem 4.1. *Lihtsa juhusliku kihtvaliku korral avaldub kogusumma hinnang kujul*

$$\hat{t}_y = \sum_{h=1}^{H} N_h \bar{y}_{s_h} , \tag{4.1}$$

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

$$\bar{y}_{s_h} = \frac{1}{n_h} \sum\nolimits_{s_h} y_i . \tag{4.2}$$

Kogusumma dispersioon on kujul

$$D\hat{t} = \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{yU_h}^2 / n_h \tag{4.3}$$

dispersiooni hinnanguga

$$\hat{D}\hat{t} = \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{ys_h}^2 / n_h , \tag{4.4}$$

kus

$$S_{yU_h}^2 = \frac{1}{N_h - 1} \sum\nolimits_{U_h} \left(y_i - \overline{Y}_h \right)^2 \tag{4.5}$$

ja

$$s_{ys_h}^2 = \frac{1}{n_h - 1} \sum\nolimits_{s_h} \left(y_i - \overline{y}_{s_h} \right)^2 . \tag{4.6}$$

Tõestus. Nagu on teada, avaldub lihtsa juhusliku kihtvaliku korral kogusumma hinnang kujul

$$\hat{t} = N\overline{y} , \tag{4.7}$$

kus $\overline{y}$ on valimikeskmine. Seega on *h*-nda kihi kogusumma hinnang kujul

$$\hat{t}_h = N_h \overline{y}_{s_h} . \tag{4.8}$$

Nihketa hinnang uuritava tunnuse kogusummale on kihtvaliku korral

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_h , \tag{4.9}$$

millest saame lihtsa juhusliku kihtvaliku hinnangu järgmisel kujul

$$\hat{t}_y = \sum_{h=1}^{H} N_h \overline{y}_{s_h} . \tag{4.10}$$

Samuti on teada, et lihtsa juhusliku valiku korral on kogusumma hinnangu dispersiooni ja dispersiooni hinnangu avaldised

$$D\hat{t} = N^2 (1 - f) S_{yU}^2 / n \tag{4.11}$$

ja

$$\hat{D}\hat{t} = N^2 (1 - f) s_{ys}^2 / n . \tag{4.12}$$

Siinjuures

$$S_{yU}^2 = \frac{1}{N - 1} \sum\nolimits_{U} \left(y_i - \overline{Y} \right)^2 \tag{4.13}$$

on tunnuse *y* dispersioon üldkogumis ja

$$s_{ys}^2 = \frac{1}{n - 1} \sum\nolimits_{s} \left(y_i - \overline{y} \right)^2 \tag{4.14}$$

on tunnuse y dispersioon valimis, kus $\overline{Y}$ ja $\overline{y}$ on vastavalt üldkogumi ja valimi keskmine. Rakendades valemeid (4.11 ja 4.12) h-nda kihi dispersioonile ja dispersiooni nihketa hinnangule ning arvestades seda, et kihtvaliku korral avaldub hinnangu $\hat{t}_y$ dispersioon kujul

$$D\hat{t} = \sum_{h=1}^{H} D\hat{t}_h \tag{4.15}$$

ja dispersiooni hinnang on

$$\hat{D}\hat{t} = \sum_{h=1}^{H} \hat{D}\hat{t}_h \,, \tag{4.16}$$

kus $D\hat{t}_h$ ja $\hat{D}\hat{t}_h$ on h-nda kihi kogusumma hinnangu dispersioon ja dispersiooni hinnang, saame kogusumma hinnangu dispersiooni ja dispersiooni hinnangu valemid lihtsa juhusliku kihtvaliku korral:

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

ja

$$\hat{D}\hat{t} = \sum_{h=1}^{H} N_h^2 (1 - f_h) s_{ys_h}^2 / n_h .$$

5 Praktiline ülesanne

5.1 Andmete kirjeldus

Antud bakalaureusetöö praktiline osa on tellitud Sotsiaalministeeriumi Terviseinfo- ja analüüsiosakonna poolt ja sisaldab andmeid, mis on saadud 2004. aasta statistilise tervishoiualase majandustegevuse aruannete põhjal. „Tervishoiualane majandustegevus" on aruanne, mida esitavad kõik tervishoiuasutused (näiteks haiglad, perearstikeskused, hambaravid jms). Aruandekohustuslased on kõik asutused, mis osutavad tervishoiuteenuseid isegi siis, kui nende põhitegevus pole tervishoid.

Aruande vorm koos juhendiga on koostatud Sotsiaalministeeriumi poolt (vt. Lisa 1).

Tervishoiualase majandustegevuse aruanne koosneb:

1) tervishoiuasutuse tuludest;
2) tervishoiuasutuse kuludest;
3) majandusaasta tulemist.

Tulu – aruandeperioodi sissetulekud, mille tekkimisega kaasneb vara suurenemine või kohustuste vähenemine.

Kulu - tulu tekkimiseks vajalikud väljaminekud aruandeperioodi jooksul, millega kaasneb vara vähenemine või kohustuste suurenemine.

Tulem - puhaskasum (-kahjum) - kõigi tulusummade ja kulusummade vahe.

Kuna andmestik on küllalt mahukas (kokku 1541 objekti ning 286 tunnust), on antud töös keskendutud rohkem tervishoiuasutuste tuludele. Mittevastamise ja selle kompenseerimise analüüs puudutab tunnuseid, mis kirjeldavad tulusid. Erilist huvi pakuvad seitse tunnust:

- tulud riigieelarvest,
- laekumised Eesti Haigekassale osutatud teenustest,
- toetused asutustelt,
- toestused juriidilistelt ja füüsilistelt isikutelt,
- muud tegevustulud (äritulud),
- finantstulud,
- erakorralised tulud.

5.2 Ülesande püstitus

Sotsiaalministeerium püstitas uurimiseks kaks probleemi:

1. Kuna andmed pole täielikud, st sisaldavad suurt mittevastamise protsenti, siis tuleb leida kõige sobivamad meetodid kao kompenseerimiseks.
2. Kuna senini on seda tüüpi uuringud iga aasta olnud kõiksed ja väga kulukad, siis tekkis vajadus üle minna valikuuringu peale. Seega on teiseks probleemiks sobiva valikudisaini koostamine ning kao kompenseerimise meetodite kirjeldamine tulevase valikuuringu jaoks.

5.3 Kao kompenseerimine terves andmestikus

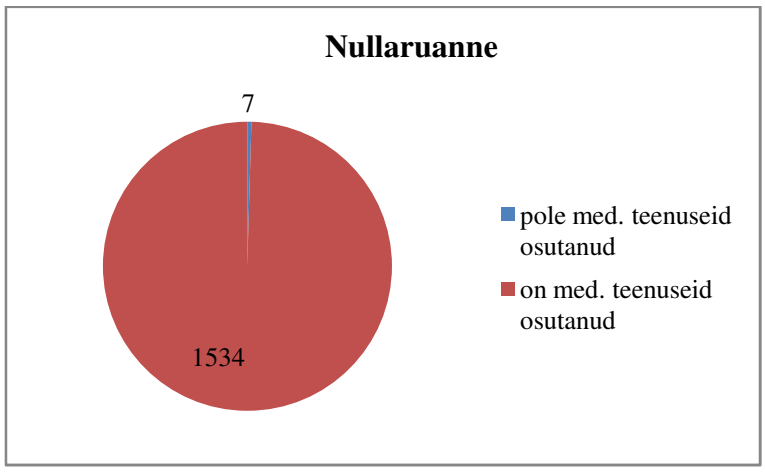
Statistikas on heaks tavaks jätta andmestikus mittevastamise korral vastav lahter tühjaks. Kuid selle andmestiku raskus seisnes selles, et mittevastamine oli peamiselt kodeeritud nulliga (mõned lahtrid olid ka tühjad), ja samal ajal oli küllaltki palju selliseid nulle, mis ei olnud seotud mittevastamisega, vaid tähendasidki väärtust 0.

Andmete töötlemise esimesel etapil tuligi seda mittevastamist esialgses andmestikus kindlaks määrata. See aga nõudis kõigi tunnuste analüüsimist ja tunnustevahelisi seoste leidmist.

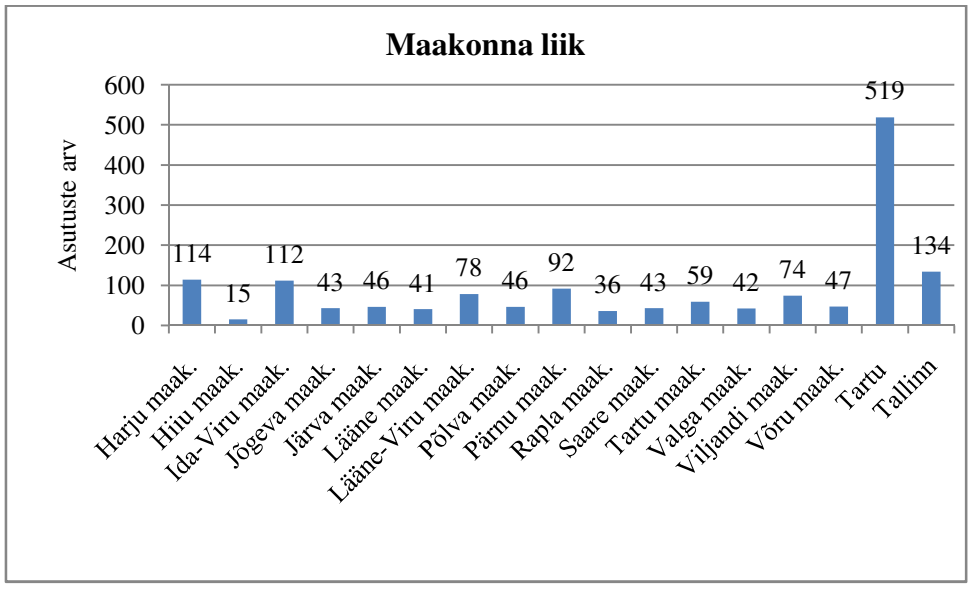
5.3.1 Tunnuste kirjeldus

Kokku on andmestikus 286 uuritavat tunnust ning kolm tausttunnust, mille korral on kõikide objektide väärtused teada. Tausttunnusteks on **Nullaruanne**, **Maakonna liik** ja **Teenuse tüüp** (vt. joonised 5.3.1, 5.3.2 ja 5.3.3). Tunnus **Nullaruanne** on binaarne tunnus, mille väärtusteks on 1 – kui 2004. aasta jooksul pole asutus meditsiinilisi teenuseid osutanud ja 0 – muidu.

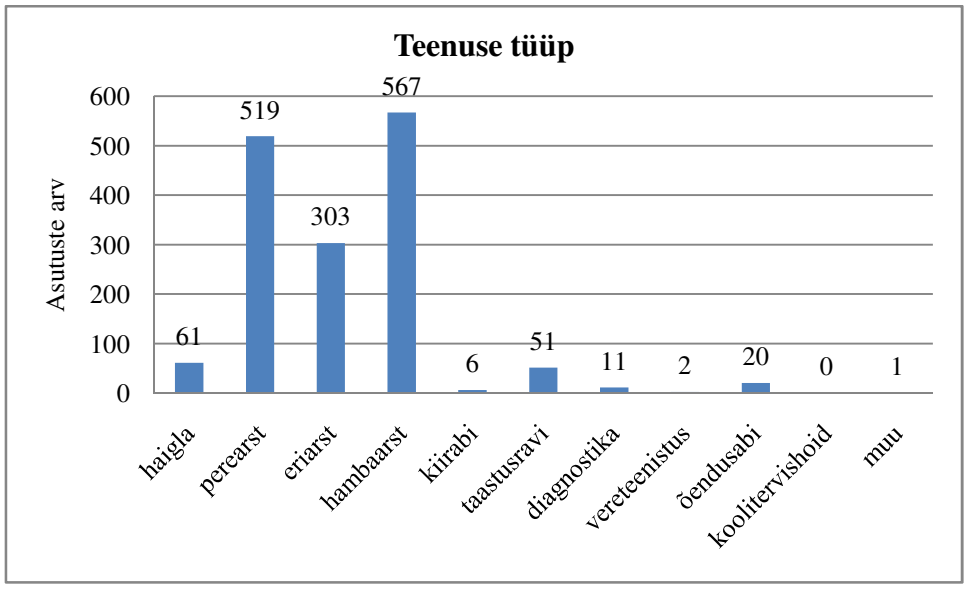
Joonis 5.3.1 Tunnuse **Nullaruanne** väärtuste jaotus



Joonis 5.3.2 Tunnuse **Maakonna liik** väärtuste jaotus



Joonis 5.3.3 Tunnuse **Teenuse tüüp** väärtuste jaotus



Tunnuse **Nullaruanne** analüüsimisel selgus välja, et peaaegu kõiki asutusi saab jagada kolmeks grupiks. Esimesse gruppi kuuluvad need asutused, mis 2004. aastal pole osutanud meditsiinilisi teenuseid (**Nullaruanne** = 1), seega teiste tunnuste väärtuste puudumine sellistel objektidel on loomulik ja ei viita mittevastamisele. Teisse gruppi kuuluvad asutused, mille **Nullaruande** väärtuseks on null (meditsiinilisi teenuseid osutasid), kuid teiste tunnuste väärtused on samuti nullid (seega ka tunnuse **Tulud kokku** väärtus võrdub nulliga). Selline olukord pole aga tegelikult võimalik, järelikult selliste asutuste jaoks on tegemist mittevastamisega terve objekti tasemel. Viimasesse, kolmandasse gruppi kuuluvad kõik need asutused, mis 2004. aastal

osutasid meditsiinilisi teenuseid (**Nullaruanne** = 0). Selles grupis saab esineda väärtuse kadu. Näiteks, kui teenuse tüübiks on „kiirabi", siis sellise objekti jaoks peab kindlasti olema täidetud tunnus **Kiirabi sihtfinantseerimine** (JB01006 andmestikus, vt. Lisa 1). Teenuse tüüp „haigla" eeldab tunnuse **Voodipäeva tasu** (JB01030 andmestikus) täitmist. Kui on tegemist teenuse tüübiga „hambaarst", siis peab kindlasti olema ära toodud **Visiiditasu** (JB01029 andmestikus). Haiglatel, pere- ja hambaarstidel on niisuguseks kohustuslikuks tunnuseks **Raviteenus** (JB01010), perearstidel on ka lisaks **Laekumised Eesti Haigekassale osutatud teenustest**. **Raviteenus** peab olema täidetud ka sel juhul, kui mitte ükski tunnustest **Hooldusravi**, **Taastusravi**, **Ennetusravi** pole täidetud.

Lisaks toodud näidetele andmete õigsuse kontrollimiseks tuli kasutada ka eelmise aasta (2003) andmestikku. Mitme objekti korral tekkis situatsioon, kui tunnuse **Eelmise aasta tulem** (JB01001 andmestikus) väärtus oli 0 (st tulude summa langeb kokku kulude summaga), mis on tavalises elus väga harva esinev situatsioon, seega peame seda mittevastamiseks. Nagu oli juba eespool mainitud, on tulem puhaskasum (-kahjum), st kõigi tulusummade ja kulusummade vahe. Seega tunnuse **Eelmise aasta tulem** väärtus saab olla ka negatiivne. **Raviteenus** peab olema täidetud ka sel juhul, kui mitte ükski tunnustest **Hooldusravi**, **Taastusravi**, **Ennetusravi** pole täidetud.

Tabel 5.3.1 Uuritavate tunnuste mittevastamise protsent

Tunnus	Mittevastamine(%)	Vastamine(%)
Tulud riigieelarvest	79.4	20.6
Laekumised EH os.teenustest	78.8	21.2
Toetused asutustelt	79.4	20.6
Toetused jur. ja füüs.isikutelt	79.4	20.6
Muud tegevustulud	79.4	20.6
Finantstulud	79.4	20.6
Erakorralised tulud	79.4	20.6

5.3.2 Kao kompenseerimise meetodite analüüs

Rakendame punktis 2 kirjeldatud omistusmeetodeid tervele andmestikule (st kõikse uuringu kontekstis).

Kõigepealt kontrollime klassi keskmise omistuse efektiivsust. Igas klassis omistatakse vaadeldava tunnuse puuduvatele väärtustele klassi vastanute keskmine. Klasside määramisel

saame kasutada kolme tausttunnust.

Esmalt võtame ühtainsat tausttunnust **Teenuse tüüp** (vt. Lisa 1). Programmi käivitamisel (vt. Lisa 2) selgus, et üks objekt (viimane andmestikus) jääb ikkagi tühjaks, st, et antud juhul pole sellele objektile hinnangulisi väärtusi võimalik omistada, kuna vastav klass on moodustatud ainult ühest objektist.

Vaatleme kõigepealt hinnanguid tunnuse **Tulud riigieelarvest** (JB01003) kogusummale ja dispersioonile saadud meetodil. Tulemused on koondatud tabelisse 5.3.2. Nagu näha, on keskmised klasside järgi enne ja pärast meetodi rakendust samad, mis oli ka oodatav, kuna puuduvad väärtused on asendatud kihi keskmistega. Samal põhjusel on ka tunnuse dispersioon vähenenud. Teiste tunnuste hinnangud on toodud Lisas 3 ja saadud tulemused on analoogilised eespool kirjeldatuga.

Tabel 5.3.2 Klasside keskmise omistus tunnuse **Teenuse tüüp** järgi: tunnuse **Tulud riigieelarvest** hinnangud

Teenuse tüüp	Kogusumma	Keskmine	Dispersioon	Kokku
haigla	230400326.9	3777054.5	7.1539917E+13	61
perearst	17934199.4	34555.3	8.8931363E+10	519
eriarst	4278018.1	14118.9	1.2269535E+10	303
hambaarst	45253.4	79.8	4.3999241E+05	567
kiirabi	83495049.6	13915841.6	1.6213163E+14	6
taastusravi	5521103.7	108256.9	1.1309086E+11	51
diagnostika	0	0	0	11
vereteenistus	0	0	0	2
õendusabi	0	0	0	20
muu	.	.	.	0
Kokku	341673951.2	221866.2	4.6238403E+12	1540

Järgnevalt uurime, kuidas muutuvad tulemused, kui klasside määramisel kasutame kahte tunnust: **Teenuse tüüp** ja **Maakonna liik**. Selle meetodi tulemuseks on viiel objektil lüngad. Põhjuseks on vastamise puudumine vastavates klassides ja hinnanguliste väärtuste prognoosimine pole võimalik. Analoogiliselt ülalpool kirjeldatuga vaatame hinnanguid tunnuse **Tulud riigieelarvest** (JB01003) kogusummale ja dispersioonile. Tabelist 5.3.3 on näha, et keskmised tunnuse **Teenuse tüüp** klasside järgi on natuke muutunud, mis on täiesti ootuspärane tulemus, kuna klasside koostamisel on arvesse võetud tunnus **Maakonna liik**. Kogusumma hinnanguks on nüüd 343678333.01. Dispersioon on läinud suuremaks, mille saab seletada

klasside arvu suurenemisega.

Tabel 5.3.3 Klassi keskmise omistus tunnuste **Teenuse tüüp** ja **Maakonna liik** järgi: tunnuse **Tulud riigieelarvest** hinnangud

Teenuse tüüp	Kogusumma	Keskmine	Dispersioon	Kokku
haigla	242967521.5	3983074.1	7.5266264E+13	61
perearst	19537926.2	37645.3	8.9949898E+10	519
eriarst	3411978.3	11297.9	1.2365439E+10	302
hambaarst	43555.4	76.8	4.4596073E+05	567
kiirabi	72948976	12158162.7	1.8066825E+14	6
taastusravi	4768375.7	93497.6	1.1808393E+11	51
diagnostika	0	0	0	10
vereteenistus	0	0	.	1
õendusabi	0	0	0	19
muu	.	.	.	0
Kokku	343678333.0	223748.9	4.7260867E+12	1536

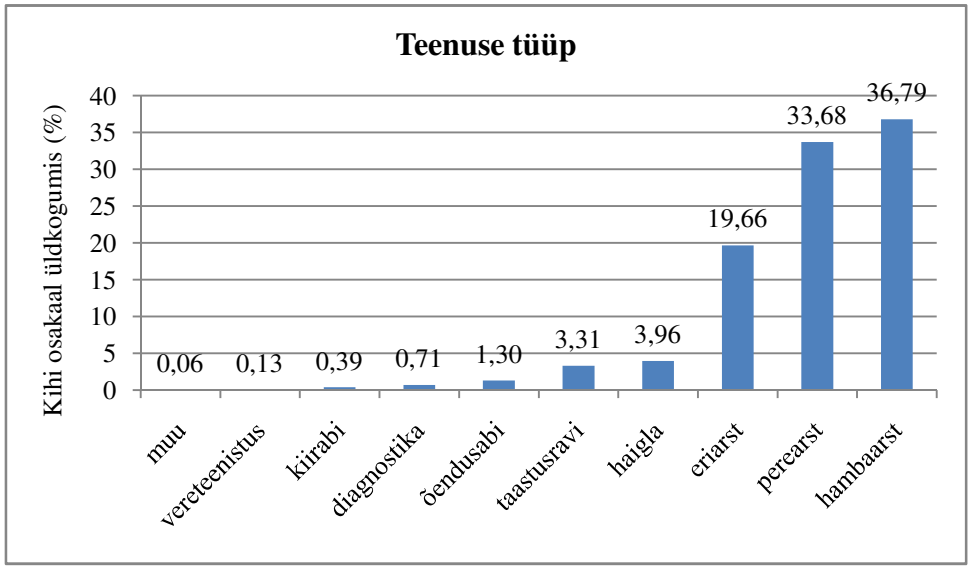
Teiste tunnuste kogusumma ja dispersiooni hinnangud on toodud Lisas 3.

5.4 Valikuuringu läbiviimine perspektiivis

5.4.1 Disaini kirjeldus

Terviseinfo- ja analüüsiosakonna ülesandeks on anda hinnanguid näitajatele nii terve Eesti kohta kui ka gruppide järgi, st eraldi haiglates, perearstikeskustes jne ehk pakutava teenuse tüübi järgi (vt. joonis 5.4.1). Seda on aga mugav teostada kasutades kihtvalikut (kihiks on **Teenuse tüüp**), mis oli kirjeldatud punktis 4. Samuti oli ka otsustatud lihtsa juhuvaliku kasuks igas kihis.

Joonis 5.4.1 Tunnuse **Teenuse tüüp** jaotus



Järgmisel etapil tuli määrata üldine valimimaht *n*, kus $n = \sum_{h=1}^{10} n_h$ ja n_h on *h*-nda kihi valimimaht. Selleks on kasutatud Terviseinfo- ja analüüsiosakonna poolt püstitatud soovitust, et uuritavate parameetrite hinnangute suhteline viga $s.v.(\hat{\theta})$ ei oleks suurem, kui 0.03, st

$$s.v.(\hat{\theta}) = \frac{\sqrt{\hat{D}\hat{\theta}}}{\hat{\theta}} \leq 0.03 \qquad (5.4.1)$$

Kihtvaliku korral saab kasutada mitmeid variante valimimahu määramiseks kihtides. Üks parimaid on nn *optimaalne* paigutus [3], kus kihi valimimahu määramisel kasutatakse uuritava tunnuse standardhälvet selles kihis üldkogumis S_h. Meetodi puuduseks on aga see, et mitme uuritava tunnuse korral tulevad head hinnangud ainult nendel tunnustel, mis on tugevalt korreleeritud valimi määramiseks kasutatud tunnusega. Meie uuritavate tunnuste vahelised korrelatsioonid on toodud tabelis 5.4.1. Tabelist on näha, et mitte ükski uuritavatest tunnustest ei ole tugevalt korreleeritud valimi määramiseks kasutatud tunnusega **Tulud riigieelarvest**.

Tabel 5.4.1 Tunnuse **Tulud riigieelarvest** korrelatsioon teiste tunnustega

Tunnus	Tulud riigieelarvest
Tulud riigieelarvest	1.0000
Laekumised EH-le osutatud teenustest	0.6236
Toetused asutustelt	0.2920
Toetused jur. ja füüs. isikutelt	0.3346
Muud tegevustulud	0.3219
Finantstulud	0.4316
Erakorralised tulud	0.2439

Teine alternatiiv valimite mahtude määramiseks kihtides on nn *võrdeline* paigutus kihtides, mis tähendab, et vastavate kihtide osakaalud valimis ja üldkogumis on võrdsed ehk

$$\frac{n_h}{n} = \frac{N_h}{N}, \tag{5.4.2}$$

kus N_h on objektide koguarv kihis *h*.

Lihtsa juhusliku kihtvaliku korral (vt. punkt 4)

$$\hat{\theta} = \hat{t}_y = \sum_{h=1}^{H} N_h \bar{y}_{s_h} \tag{5.4.3}$$

ja

$$\hat{D}\hat{\theta} = \hat{D}\hat{t} = \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{ys_h}^2 / n_h . \tag{5.4.4}$$

Valemitest (5.4.1)-(5.4.4) saame järgmise võrratuse:

$$\frac{\sqrt{\sum_{h=1}^{H} \frac{N_h \left(1 - \frac{n}{N}\right) S_{ys_h}^2}{\frac{n}{N}}}}{\sum_{h=1}^{H} N_h \bar{y}_{s_h}} < 0.03 . \tag{5.4.5}$$

Võrratuses (5.4.5) on aga mitu suurust tundmatud – valimidispersioonid ja keskmised kihiti. Lihtsa juhusliku valiku korral on aga $S_{ys_h}^2$ hinnanguks S_h^2-le, kus S_h^2 on üldkogumi dispersioon kihis *h*, ning $\bar{y}_{s_h}$ on hinnanguks $\bar{Y}_h$-le, kus $\bar{Y}_h$ on üldkogumi keskmine kihis *h*. Seda saame kasutada hinnangulise valimimahu leidmiseks:

$$\frac{\sqrt{\frac{1}{n} \sum_{h=1}^{H} N N_h \left(1 - \frac{n}{N}\right) S_h^2}}{t_y} < 0.03 , \tag{5.4.6}$$

kust

$$n > \frac{N \sum_{h=1}^{H} N_h S_h^2}{\sum_{h=1}^{H} N_h S_h^2 + 0.03^2 t_y^2} . \tag{5.4.7}$$

Olgu uuritavaks tunnuseks **Tulud riigieelarvest**. Jooniselt 5.4.1 on näha, et kokku on 10 kihti. Mõnes kihis on aga objektide arv väike (näiteks kaks). Sageli väikseid kihte ühendatakse suuremate sisu poolest sarnaste kihtidega, kuid antud juhul on Terviseinfo- ja analüüsiosakond

huvitatud hinnangutest igas kihis. Seega ei saa me neid väikseid kihte mõne teisega ühendada. Järelikult tuleb nendes kihtides teostada kõikset uuringut.

Defineerime kihi väikseks, kui selle kihi maht ei ületa 20 asutust. Nendeks tulid: 40 (kiirabi), 51 (diagnostika), 52 (vereteenistus), 53 (õendusabi) ja 90 (muu). Jäetud kihtide tegelikud mahud (N_h) arvestades mittevastamist on 50, 474, 218, 431 ja 31.

Kokku saame uue üldkogumi mahuks N = 1204. Uuritava tunnuse tegelik kogusumma t_y = 211700217. Rakendades võrratuse (5.4.7) ülejäänutele kihtidele saame, et n > 1195. Liites sellele arvule 19 objekti väikestest kihtidest, kus on planeeritud teostada kõikset uuringut, ja veel 318 mittevastanud objekti, saame valimimahu lõppväärtuseks 1533! Selline tulemus aga ei vasta meie ootustele, kuna see on peaaegu võrdne üldkogumi mahuga.

Sel juhul kasutame alternatiivset sobiva valimimahu arvutamisviisi, valimimahu simuleerimist, mis on kooskõlas nõutava suhtelise veaga. SAS-programmi vastavas lõigus (vt. Lisa 2) on kihtides ette antud valimimahud, mis on arvutatud järgmise skeemi järgi: kui lähtuda sellest, et kogu valimi maht on 600, st peaaegu pool üldkogumi mahust, siis kuna kihtide 40 (kiirabi), 51 (diagnostika), 52 (vereteenistus), 53 (õendusabi) ja 90 (muu) mahud ja osakaalud üldkogumis on liiga väiksed, siis jätame neid arvutamisest välja, võttes nende kihtide valimimahtudeks 6, 11, 2, 20 ja 1. Teiste kihtide mahud valimis on arvutatud vastavalt võrdelisele paigutusele. Esialgne andmestik on sorteeritud tunnuse **Teenuse tüüp** järgi ja siis 1000 valimit on võetud üldkogumist kasutades lihtsat juhuslikku kihtvalikut.

Igas sellises simulatsioonis (valimis) on arvutatud tunnuse **Tulud riigieelarvest** kogusumma hinnang $\hat{t}_{y,r}$, r=1,...,1000. Arvutades neid hinnanguid korduvalt saame nn Monte-Carlo hinnangu suhtelisele veale:

$$s.v.(\hat{t}_y)_{MC} = \frac{\sqrt{(\hat{V}\hat{t}_y)_{MC}}}{(\hat{t}_y)_{MC}}, \tag{5.4.8}$$

kus

$$(\hat{t}_y)_{MC} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{t}_{y,r} \tag{5.4.9}$$

ja

$$\left(V\hat{t}_y\right)_{MC} = \frac{1}{999}\sum_{r=1}^{1000}\left(\hat{t}_{y,r} - (\hat{t}_y)_{MC}\right)^2. \qquad (5.4.10)$$

Selliselt leitud suhteline viga tuli 0.22.

Nagu oli juba mainitud, on valimi võtmiseks üldkogumist kasutatud lihtne juhuslik kihtvalik. Valimimaht on 600 objekti. SAS–tarkvaraga on võimalik valida üldkogumist valimiobjekte ptotseduuri **surveyselect** abil, kusjuures tuleb kindlasti ära märkida iga kihi valimimahtu. On kokku lepitud kasutada võrdelist paigutust kihtides, seega kihtide valimimahud on järgmised: 23, 194, 113, 211, 6, 19, 11, 2, 20, 1. Täieliku vastamise juhule vastavad kogusumma ja dispersiooni hinnangud olid juba kirjeldatud punktis 4.

5.4.2 Mittevastamise käsitlemine saadud valimis

Peale valimimahtude määramist on tähtis arvestada ka seda, et saadud valim pole täielik, ehk sisaldab mittevastamist. Rakendame järgnevalt punktides (3.2) ja (3.3) kirjeldatud meetodeid, et kompenseerida kadu andmetes.

Kalibreerimine. Kalibreerimishinnang avaldub valemina (3.3.2). Vaatleme kolmefaktorilist klassifikatsiooni, kus faktoriteks on tausttunnused **Teenuse tüüp**, **Nullaruanne** ja **Maakonna liik**. Esimesel tunnusel on 10 taset, teisel on neid 17 ja kolmandal 2. Seega abivektor on dimensiooniga 26. Vastav programm on toodud Lisas 2. Tulemusi saab vaadata tabelist 5.4.2. Suhteline nihe näitab, kui palju kalibreerimise abil saadud hinnangud erinevad klassi keskmise omistusest, ja on arvutatud, kui $(\hat{t}_{kalibr.} - \hat{t}_{kesk.om.})/\hat{t}_{kesk.om.}$.

Tabel 5.4.2 Tunnuste kogusumma hinnangud kalibreerimise abil

Uuritav tunnus	Kogusumma	Suhteline nihe
Tulud riigieelarvest	594870855	0.7410
Laekumised EH-le osutatud teenustest	7966400000	0.5877
Toetused asutustelt	211386948	0.4668
Toetused jur. ja füüs. isikutelt	1460190000	0.0422
Muud tegevustulud	104007090	0.0552
Finantstulud	20759455	0.1204
Erakorralised tulud	286988.36	-0.5180

Homogeensete vastamisgruppide mudel. See meetod oli kirjeldatud punktis 3.2. Nagu oli juba eelpool mainitud, võtab kogusumma hinnang kuju (3.2.3). Sisuliselt toimub vastanute

laiendamine kogu valimile. Leiame tunnuse **Tulud riigieelarvest** kogusumma selle valemiga, kus homogeenseteks vastamistõenäosuste gruppideks on võetud grupid tunnuse **Teenuse tüüp** järgi. Valimi võtmine üldkogumist toimub täpselt sama skeemi järgi, mis oli juba kasutatud simuleerimisel. Kogu valimi mahuks on 600, mis on proportsionaalselt jagunenud gruppide vahel (vastav kood on toodud Lisas 2). Saadud valim asub failis ValimKiht. Järgmiste arvutuste teostamiseks tuleb jagada see fail kaheks osaks vast1 ja kadu1, et vastanute valim asuks eraldi kaost, ja seejärel kasutame vastanute valimit failist vast1 kaalutud summade ja vastanute arvu leidmiseks. Kao failist saab leida kao arvud gruppides. Kogusumma hinnangu arvutamiseks vajaminevad arvud on esitatud tabelis 5.4.3.

Tabel 5.4.3 Kaalutud summad, vastanute ja kao arvud gruppides

Teenuse tüüp	Kaalutud summa	Vastanute arv	Kao arv
haigla	309393130	17	6
perearst	35451003.9	173	21
eriarst	2904257.68	79	34
hambaarst	49957.8057	156	55
kiirabi	69579208.0	5	1
taastusravi	5684194.26	11	8
diagnostika	0	5	6
vereteenistus	0	1	1
õendusabi	0	8	12
muu	0	0	1

Kogusumma hinnanguks saame 555879976.1. Teiste tunnuste hinnangud on esitatud tabelis 5.4.4.

Tabel 5.4.4 Tunnuste hinnangud homogeensete vastamistõenäosuste mudeli kasutamisel

Uuritav tunnus	Kogusumma	Suhteline nihe
Tulud riigieelarvest	555879976.1	0.6269
Laekumised EH-le osutatud teenustest	7654153773	0.5254
Toetused asutustelt	212675171.2	0.4758
Toetused jur. ja füüs. isikutelt	1442193600	0.0294
Muud tegevustulud	102741131.9	0.0423
Finantstulud	20148632.3	0.0874
Erakorralised tulud	300219	-0.4957

Tabelist 5.4.4 on näha, et hinnangud, mis on saadud homogeensete vastamistõenäosuste mudeli kasutamisel, on küllaltki lähedased kalibreerimishinnangutele, kuigi enamjaolt

kalibreerimishinnangutest väiksemad. Kui võrrelda neid klassi keskmise omistusega, siis klassi keskmine omistus annab väiksemaid hinnanguid ja erinevus hinnangute vahel on enamikul juhtudel suur.

Kokkuvõte

Käesolevas bakalaureusetöös uuritud probleemiks on mittevastamine ja selle kompenseerimine. Uuring on tehtud andmetega, mis on saadud 2004. aasta statistilise tervishoiualase majandustegevuse aruannete põhjal. Püstitatud eesmärgid on täidetud. Kao kompenseerimiseks terves andmestikus on valitud klassi keskmise omistus. Tulevase valikuuringu jaoks sobib lihtne kihtvalik. Punktides (3.2) ja (3.3) kirjeldatud meetodite rakendmine saadud valimis näitab, et hinnangud on omavahel lähedased, kuigi homogeensete vastamistõenäosuste mudeli kasutamisel saadud hinnangud on peaasjalikult väiksemad, kui kalibreerimishinnangud. Võrreldes neid klassi keskmise omistusega, tuleb märkida, et klassi keskmine omistus annab väiksemaid hinnanguid ja suhteline nihe on enamikul juhtudel suur.

Dealing with nonresponse

Jekaterina Prostakova

Summary

This bachelor thesis studies dealing with nonresponse. What is meant by nonresponse is that the required data are not obtained for all elements which are selected for observation. Generally, a distinction is made between unit nonresponse, i.e. the failure of a selected sample member to respond, and item nonresponse where it is failed to obtain some required information from individual sample members. Normally, weighting methods are applied for unit nonresponse. To compensate for item nonresponse a range of imputation methods exist.

Today, nonresponse is a normal but undesirable feature of the survey undertaking. There is complete agreement that nonresponse can severely harm the quality of the statistics computed and published in a survey. In praxis, different methods can be used to solve this problem. The results differ depending on what method is used. It is shown that not every method can predict all the missing values and the choice of an appropriate method may strongly depend on the data available, the application and purpose of the analysis.

This bachelor thesis is structured as follows. In the first chapter of the thesis, we give an overview of nonresponse and its different types. Chapters 2 reviews various imputation methods used within the social and other sciences to compensate for item nonresponse. Chapter 3 introduces different approaches to handling missing data in sample surveys. It provides the reader familiar with newer developments in this field made by C.-E. Särndal and S. Lundström [3]. In the fourth chapter, one of several ways of sample selecting is presented briefly. Chapter 5 is the practical part of the thesis which is ordered by the Ministry of social affairs of Estonia. The analysis of the data is made according to the previously mentioned methods' description. The Ministry of social affairs of Estonia posed two main problems: firstly, there are missing values in the data that must be predicted using appropriate methods and secondly, there is a strong need to use these methods in future sample survey researches.

Kasutatud kirjandus

[1] Cochran, W. G. (1977) Sampling Techniques. New York: Wiley.

[2] Särndal, C.-E., Lundström, S. (2005) Estimation in surveys with nonresponse. England: Wiley.

[3] Traat, I., Inno, J. (1997) Tõenäosuslik valikuuring. Tartu: Tartu Ülikool.

[4] http://edis.ifas.ufl.edu/PD008 (viimati kasutatud 30.05.2007): Israel G. D. (1992) Sampling the evidence of extension program impact. Determining sample size. University of Florida.

[5] http://www.ncrm.ac.uk/publications/methodsreview/MethodsReviewPaperNCRM002.pdf (viimati kasutatud 30.05.2007): Durrant, G. B. (2005) Imputation methods for handling item-nonresponse in the social sciences: a methodological review. University of Southampton.

Lisad

Lisa 1

Aruande vorm koos juhendiga. Asub CD kettal.

Lisa 2

```
/* nullid on eemaldatud esialgsest andmestikust */
data diplom.nonresponse_ilma_nullid;
set diplom.nonresponse;
if (nullarua=0 and TOOTAJAT=0) then TOOTAJAT='.';
if (nullarua=0 and TEENUSET=40 and JB01006=0) then JB01006='.';
if (nullarua=0 and TEENUSET=10 and JB01030=0) then JB01030='.';
if (nullarua=0 and TEENUSET=30 and JB01029=0) then JB01029='.';
if (nullarua=0 and TEENUSET=20 and JB01009=0) then JB01009='.';
if (nullarua=0 and ((TEENUSET=10 or TEENUSET=20 or TEENUSET=30) and
JB01010=0)) then JB01010='.';
if (nullarua=0 and (JB01010=0 and JB01011=0 and JB01012=0 and JB01013=0))
then JB01010='.';
if nullarua=0 and JB01002=0 and JB01003=0 then JB01003='.';
if nullarua=0 and JB01002=0 and JB01009=0 then JB01009='.';
if nullarua=0 and JB01002=0 and JB01016=0 then JB01016='.';
if nullarua=0 and JB01002=0 and JB01022=0 then JB01022='.';
if nullarua=0 and JB01002=0 and JB01035=0 then JB01035='.';
if nullarua=0 and JB01002=0 and JB01044=0 then JB01044='.';
if nullarua=0 and JB01002=0 and JB01049=0 then JB01049='.';
if JB01001=0 then JB01001='.';
run;


/* 2 andmestikku (esialgne ja hiljem lisatud) on ühendatud, saadud
andmestikus on 1541 objekti */
data diplom.nonresponse_kokku;
set diplom.nonresponse_ilma_nullid diplom.lisa;
run;

/* tunnuse Teenuse tüüp sagedus- ja jaotustabel */
proc freq data=diplom.nonresponse_kokku;
tables teenuset;
run;

/* andmestik on sorteeritud tunnuse Teenuse tüüp järgi */
proc sort data=diplom.nonresponse_kokku;
by teenuset;
run;

/* klassi keskmine omistus */
/* uue andmestiku loomine, kuhu on lisatud klasside keskväärtused tunnuse
Teenuse tüüp järgi */
proc sql;
create table diplom.mean as
select a.*, avg(jb01003) as keskjb01003, avg(jb01009) as keskjb01009,
avg(jb01016) as keskjb01016, avg(jb01022) as keskjb01022, avg(jb01035) as
keskjb01035,
avg(jb01044) as keskjb01044, avg(jb01049) as keskjb01049
from diplom.nonresponse_kokku as a
group by teenuset;
```

```
quit;

/* puuduvad väärtused on asendatud keskmistega klassides */
data diplom.nonresponse_imput_mean;
set diplom.mean;
if jb01003='.' then jb01003=keskjb01003;
if jb01009='.' then jb01009=keskjb01009;
if jb01016='.' then jb01016=keskjb01016;
if jb01022='.' then jb01022=keskjb01022;
if jb01035='.' then jb01035=keskjb01035;
if jb01044='.' then jb01044=keskjb01044;
if jb01049='.' then jb01049=keskjb01049;
run;

/* tunnuse Tulud riigieelarvest arvkarakteristikute arvutamine pärast klassi
keskmise omistusmeetodi rakendamist */
PROC MEANS DATA=diplom.nonresponse_imput_mean;
CLASS teenuset;
VAR jb01003;
output out=nonresponse_kokkuv1a sum=Kogusumma mean=Keskmine var=Dispersioon
stddev=Standardh4lve N=Kokku Nmiss=Kadu;
run;

/* andmestik on sorteeritud tunnuste Teenuse tüüp ja Maakonna liik järgi */
proc sort data=diplom.nonresponse_kokku;
by teenuset maakondl;
run;

/* uue andmestiku loomine, kuhu on lisatud klasside keskväärtused tunnuste
Teenuse tüüp ja Maakonna liik järgi */
proc sql;
create table diplom.meanclass as
select a.*, avg(jb01003) as keskjb01003, avg(jb01009) as keskjb01009,
avg(jb01016) as keskjb01016, avg(jb01022) as keskjb01022, avg(jb01035) as
keskjb01035,
avg(jb01044) as keskjb01044, avg(jb01049) as keskjb01049
from diplom.nonresponse_kokku as a
group by teenuset*maakondl;
quit;

 /* puuduvad väärtused on asendatud keskmistega klassides */
data diplom.nonresponse_imput_meanclass;
set diplom.meanclass;
if jb01003='.' then jb01003=keskjb01003;
if jb01009='.' then jb01009=keskjb01009;
if jb01016='.' then jb01016=keskjb01016;
if jb01022='.' then jb01022=keskjb01022;
if jb01035='.' then jb01035=keskjb01035;
if jb01044='.' then jb01044=keskjb01044;
if jb01049='.' then jb01049=keskjb01049;
run;

/* tunnuse Tulud riigieelarvest arvkarakteristikute arvutamine pärast klassi
keskmise omistusmeetodi rakendamist */
PROC MEANS DATA=diplom.nonresponse_imput_meanclass;
CLASS teenuset;
VAR jb01003;
output out=nonresponse_kokkuv1b sum=Kogusumma mean=Keskmine var=Dispersioon
stddev=Standardh4lve N=Kokku Nmiss=Kadu;
```

```
run;

/* suhtelise vea arvutamine simuleerimise abil */
/* kihtide valimimahtude määramine */
data kokku;
input teenuset _nsize_;
datalines;
10 23
20 194
25 113
30 211
40 6
50 19
51 11
52 2
53 20
90 1
;
run;

/* andmestik on sorteeritud tunnuse Teenuse tüüp järgi */
proc sort data=diplom.nonresponse_kokku;
by teenuset;
run;

/* simuleerimisel saadud 1000 valimit sama disainiga */
proc surveyselect data=diplom.nonresponse_kokku rep=1000
method=srs
n=kokku
out=simuleerimine;
strata teenuset;
run;

/* iga simuleerimise puhul on arvutatud tunnuse Tulud riigielarvest
kogusumma*/
PROC MEANS DATA=simuleerimine ;
CLASS replicate;
VAR jb01003;
output out=simul_kokkuv sum=Kogusumma;
run;

data simul_kokkuv;
set simul_kokkuv;
if replicate='.' then delete;
run;

/* kogusummad on keskmistatud MC hinnangu keskmisele veale arvutamiseks */
PROC MEANS DATA=simul_kokkuv;
VAR kogusumma;
output out=kokkuv3 mean(kogusumma)=t;
run;

/* homogeensete vastamisgruppide mudeli kasutamine */
/* valimi võtmine üldkogumist */
proc surveyselect data=diplom.nonresponse_kokku
method=srs
n=kokku
out=diplom.ValimKiht;
strata teenuset;
```

```
run;

/* võetud valim on jagatud kaheks osaks tunnuse Tulud riigieelarvest järgi:
vastanute valim ja kao valim */
data diplom.vast1 diplom.kadu1;
set diplom.ValimKiht;
if jb01003='.' then output diplom.kadu1; else output diplom.vast1;
run;

/*kalibreerimishinnang*/
/*abitunnuste määramine*/
data diplom.viimane;
set diplom.viimane;
if teenuset=10 then abi1=1;
else abi1=0;
if teenuset=20 then abi2=1;
else abi2=0;
if teenuset=25 then abi3=1;
else abi3=0;
if teenuset=30 then abi4=1;
else abi4=0;
if teenuset=40 then abi5=1;
else abi5=0;
if teenuset=50 then abi6=1;
else abi6=0;
if teenuset=51 then abi7=1;
else abi7=0;
if teenuset=52 then abi8=1;
else abi8=0;
if teenuset=53 then abi9=1;
else abi9=0;
if maakondl=37 then abi10=1;
else abi10=0;
if maakondl=39 then abi11=1;
else abi11=0;
if maakondl=44 then abi12=1;
else abi12=0;
if maakondl=49 then abi13=1;
else abi13=0;
if maakondl=51 then abi14=1;
else abi14=0;
if maakondl=57 then abi15=1;
else abi15=0;
if maakondl=59 then abi16=1;
else abi16=0;
if maakondl=65 then abi17=1;
else abi17=0;
if maakondl=67 then abi18=1;
else abi18=0;
if maakondl=70 then abi19=1;
else abi19=0;
if maakondl=74 then abi20=1;
else abi20=0;
if maakondl=78 then abi21=1;
else abi21=0;
if maakondl=82 then abi22=1;
else abi22=0;
if maakondl=84 then abi23=1;
else abi23=0;
```

```
if maakondl=86 then abi24=1;
else abi24=0;
if maakondl=784 then abi25=1;
else abi25=0;
if nullarua=0 then abi26=1;
else abi26=0;
run;

/*disainikaalu määramine*/
data diplom.viimane;
set diplom.viimane;
if teenuset=10 then kaal= 61/23;
if teenuset=20 then kaal= 519/194;
if teenuset=25 then kaal= 303/113;
if teenuset=30 then kaal= 567/211;
if teenuset=40 then kaal= 1;
if teenuset=50 then kaal= 51/19;
if teenuset=51 then kaal= 1;
if teenuset=52 then kaal= 1;
if teenuset=53 then kaal= 1;
if teenuset=90 then kaal= 1;
output;
run;

/*on kustutatud read, kus tunnuse väärtus puudub*/
data reaalne_valim;
set diplom.viimane;
if jb01003='.' then delete;
run;

/*kogusumma kalibreerimishinnangu arvutamine*/
proc iml;
start;
use reaalne_valim;
read all var {abi1 abi2 abi3 abi4 abi5 abi6 abi7 abi8 abi9 abi10
abi11 abi12 abi13 abi14 abi15 abi16 abi17 abi18 abi19 abi20
abi21 abi22 abi23 abi24 abi25 abi26} into X;
read all var {kaal} into D;
read all var {jb01003} into Y;
n=nrow(D);
X_summa={61 519 303 567 6 51 11 2 20 114 15 112 43 46 41 78 46
92 36 43 59 42 74 47 519 1534};
r=nrow(D);
yks=j(r,1,1);
XL=X;
do i=1 to 26;
XL[,i]=D#X[,i];
end;
th=XL[+,];
v=yks+X*inv(t(XL)*X)*(X_summa-th)`;
create v_kaal from v;
append from v;
wy=Y#D#v;
t=sum(wy);
print 'kogusumma kalibreerimishinnang' t;
finish;
run;
quit;
```

Lisa 3

Tabel 1 Klasside keskmise omistus tunnuse **Teenuse tüüp** järgi: uuritavate tunnuste hinnangud

Tunnus	Kogusumma	Dispersioon
Tulud riigieelarvest	341673951	4.6238403E+12
Laekumised EH os.teenustest	5017716243	9.7137495E+14
Toetused asutustelt	144112890	1.2014091E+13
Toetused jur. ja füüs.isikutelt	1401000000	1.2014091E+13
Muud tegevustulud	98568212	308637775842
Finantstulud	18528693	10457923902
Erakorralised tulud	595376.2	5.1254558E+07

Tabel 2 Klasside keskmise omistus tunnuste **Teenuse tüüp** ja **Maakonna liik** järgi: uuritavate tunnuste hinnangud

Tunnus	Kogusumma	Dispersioon
Tulud riigieelarvest	343678333.0	4.7260867E+12
Laekumised EH os.teenustest	5233273077.4	1.0400167E+15
Toetused asutustelt	143012352	1.2729948E+12
Toetused jur. ja füüs.isikutelt	1461927045.9	1.2994879E+13
Muud tegevustulud	107244893.3	343763699068
Finantstulud	19210669.8	10821218260
Erakorralised tulud	540734.1	50119382.8