Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kristi Tüli

# Empiirilise tõepära meetod valikuuringutes

Matemaatilise statistika eriala
Magistritöö (30 EAP)

Juhendajad:
Imbi Traat, PhD
Natalja Lepik, PhD

Tartu, 2016

**Empiirilise tõepära meetod valikuuringutes**

Magistritöö

Kristi Tüli

**Lühikokkuvõte.** Magistritöö eesmärk on anda ülevaade empiirilise tõepäraga lähenemisest suurusega võrdelise tagasipanekuga valiku näitel. Antud lähenemise korral saab moodustada disainipõhised usaldusintervallid üldkogumi keskmisele, kogusummale või kvantiilidele ning nende leidmisel pole vaja teada dispersiooni hinnanguid. Lisaks võrreldakse simuleerimisülesande abil saadud tulemusi varasemalt tuntud valemitega valikuuringutes. Uue meetodi tutvustamisel on aluseks võetud Berger ja De La Riva Torrese (2016) artikkel "Empirical Likelihood confidence interval for complex sampling designs".

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

**Märksõnad:** lõplik üldkogum, valikudisain, tõepärafunktsioon, hindav võrrand, usaldusvahemik

**Empirical Likelihood Method in Sample Surveys**

Master's thesis

Kristi Tüli

**Abstract.**The aim of this master's thesis is to introduce an empirical likelihood approach under unequal probability sampling with replacement. With this approach it is possible to construct design-based confidence intervals for population mean, total or quantiles which can be calculated without the need of variance estimates. In addition, simulation results obtained by the new method are compared to commonly used formulas in sample surveys. Introduction of the new method is based on the article "Empirical likelihood confidence interval for complex sampling designs" written by Berger and De La Riva Torres (2016).

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

**Keywords:** finite population, sampling design, likelihood function, estimating equation, confidence interval

# Sisukord

# 1 Sissejuhatus

Käesolev töö kuulub valikuteooria valdkonda, kus üldkogum on lõplik ja objektid satuvad valimisse läbi tõenäosusliku valikumehhanismi. Klassikalised valimi eeldused, nagu sõltumatus ja pärinemine samast jaotusest, ei ole enamasti täidetud. Seetõttu on valikuteoorias välja arendatud omad meetodid hinnangute ja nende täpsusnäitajate leidmiseks. Ometi on aegajalt astutud samme selles suunas, et klassikalise statistika vahendeid sisse tuua. Klassikalises statistikas on tähtsal kohal tõepäral põhinevad meetodid.

Antud magistritöö eesmärk on tutvustada uut meetodit valikuteooria jaoks. Kasutades empiirilise tõepäraga lähenemist saab moodustada disainipõhised usaldusintervallid ning nende leidmisel pole vaja teada dispersiooni hinnangut ega pea eeldama hinnangu normaaljaotust. Selle meetodi korral saab moodustada näiteks üldkogumi kogusummale, keskmisele ja kvantiilidele usaldusintervallid ka siis, kui üldkogumi maht on teadmata. Käesolevas magistritöös kasutatakse empiirilise tõepäraga lähenemist suurusega võrdelise tagasipanekuga valiku korral.

Töö esimeses pooles antakse lühike ülevaade tõenäosuslikest valikuuringutest ja käesoleva töö jaoks vajalikest mõistetest. Räägitakse suurusega võrdelisest tagasipanekuga disainist ning selle konkreetse disaini korral parameetrite hindamisest. Tuuakse välja valemid üldkogumi keskmise hindamiseks ning sellele usalduspiiride leidmiseks.

Töö teises osas tutvustatakse usalduspiiride leidmist empiirilise tõepära abil. Esitatakse vajalikud valemid üldkogumi keskmise ja selle usaldusintervallide leidmiseks. Tuuakse erinevaid näited, mis lihtsustavad uudsest lähenemisest arusaamist. Empiirilise tõepäraga lähenemist on palju uurinud ning oma artiklites kirjeldanud Y. G. Berger ning O. De La Riva Torres. Käesoleva töö teise osa definitsioonid ning uudse meetodi kirjeldus põhineb suures osas Bergeri ja De La Riva Torrese (2016) artiklil. Selle lähenemise vastu on viimasel ajal olnud suur huvi. On arendatud ka lähedalt seotud meetodeid, üheks selliseks on pseudo empiirilise tõepäraga lähenemine (Wu, Rao 2006).

Töö viimases osas teostatakse simuleerimisülesanne, mille käigus võetakse 1000 valimit kasutades suurusega võrdelist tagasipanekuga disaini. Leitakse nelja

erineva tunnuse üldkogumi keskmistele hinnangud ning võrreldakse eespool kirjeldatud meetodite abil leitud usaldusintervalle tegeliku vahemikuga, mis on määratud vaadeldud hinnangute jaotusega.

Magistritöö vormistamisel on kasutatud tekstitöötlusprogrammi LaTeX. Simuleerimisülesande lahendamiseks ja jooniste tegemiseks on kasutatud statistikapaketi R. Programmi kood on esitatud Lisas 3.

Autor tänab käesoleva magistritöö juhendajaid Imbi Traati ja Natalja Lepikut rohkete nõuannete ja näpunäidete eest.

# 2 Valikuuring ja selle põhimõisted

Valikuuringud on tänapäeval väga levinud uuringute teostamise viis. Tihti kasutavad majandusteadlased ja sotsioloogid oma töös just valikuuringuid. Nad teevad üldkogumi kohta järeldusi valimi (üldkogumi mingi osa) abil. Täpset infot saadakse ainult valimisse kuuluvatelt objektidelt ning selle põhjal leitakse üldkogumi huvipakkuvatele parameetritele hinnangud. Valikuuringuid eelistatakse tihti nende odavuse ning väiksema uuringu läbiviimiseks kulunud aja tõttu.

Valikuuringud saab jagada kahte rühma - tõenäosuslikud ja empiirilised valikud. Käesolevas töös vaadeldakse tõenäosuslikku valikumeetodit. Tõenäosuslik valik põhineb sellel, et iga objekti kohta on teada tema valimisse sattumise (kaasamise) tõenäosus.

Edasi defineerime töös vajaminevad mõisted (Traat, Inno (1997), Lepik, Traat (2013)).

**Definitsioon 2.1** *Valikuindikaator $I_i$ on juhuslik suurus, mis on määratud iga üldkogumi objekti $i$ $(i = 1, ..., N)$ jaoks ja näitab objekti $i$ valikute arvu. Tähistame realiseerunud valikute arvu $k_i$.*

**Definitsioon 2.2** *Juhuslikku vektorit $\boldsymbol{I} = (I_1, ..., I_N)$, kus $I_i$ tähistab objekti $i$ valikute arvu üldkogumist, nimetatakse valikuvektoriks ja selle jaotust*

$$p(k) = P(I = k),$$

*kus $k = (k_1, ..., k_N)$, nimetakase valikudisainiks.*

**Definitsioon 2.3** *Üldkogumi objekti $i$ $(i = 1, ..., N)$ kaasamistõenäosuseks $\pi_i$ nimetatakse tõenäosust, millega see objekt kaastakse valimisse $s$ antud disaini korral:*

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

*kus $k$ on vektorvalim (vektori $I$ realisatsioon).*

Objekti sattumine valimisse on juhuslik. Objekti sattumist valimisse saab mõjutada valikudisainiga ja samuti sellega, kas kasutatakse tagasipanekuta või tagasipanekuga valikut. Käesolevas töös kasutatakse tagasipanekuga (TGA) valikut, st iga objekt võib sattuda valimisse rohkem kui üks kord. Tagasipanekuga disainide korral räägitakse valikutõenäosusest, mis on defineeritud järgmiselt.

**Definitsioon 2.4** *Üldkogumi objekti* $i$ $(i = 1, 2, ..., N)$ *valikutõenäosuseks* $p_i$ *nimetatakse tõenäosust, millega seda objekti võidakse valida antud disaini ühel sammul.*

Käesolevas töös keskendume ühele tagasipanekuga ebavõrdsete valikutõenäosustega disainile, nimelt multinomiaaldisainile:

$$I \sim M(n, p_1, ..., p_N); \sum_U p_i = 1,$$

kus $n$ on valimimaht ning $\sum_U a_i$ tähistab summat üle kogu üldkogumi, $\sum_{i \in U} a_i$. Sellist lühendatud varianti kasutatakse edaspidi kogu töös. Multinomiaaldisaini korral on valikuindikaator $I_i \sim B(n, p_i)$. See tähendab, et objekti $i$ saab valida $k_i = 1, ..., n$ korda.

Multinomiaaldisaini tõenäosusfunktsioon on järgmine:

$$p(k) = \frac{n!}{\prod_U k_i!} \prod_U {p_i}^{k_i},$$

kui $\sum_U k_i = n$ ning $\prod_U b_i$ tähistab korrutist, mille indeks muutub üle terve üldkogumi, $\prod_{i \in U} b_i$

Valimi võtmiseks multinomiaaldisaini abil leitakse tausttunnus $x$, mis on teada kõigi üldkogumi objektide jaoks. Tihti valitakse tausttunnuseks objekti suurust iseloomustav näitaja, seega nimetataks multinomiaaldisaini ka suurusega võrdelise tõenäosusega disainiks (PPS ehk *probability proportional to size sampling*). PPS tagasipanekuga valiku korral satub valimisse suuremaid objekte rohkem. Valikutõenäosused leitakse järgmise valemi abil:

$$p_i = \frac{x_i}{\sum_U x_j}. \tag{1}$$

## 2.1 Parameetrite hindamine disainipõhise meetodi järgi

PPS tagasipanekuga valiku korral avalduvad disainikarakteristikud vastavalt multinomiaaldisaini omadustele järgmiselt:

$$E(I_i) = np_i, \tag{2}$$

$$E(I_iI_j) = n(n-1)p_ip_j, \tag{3}$$

$$V(I_i) = np_i(1-p_i), \tag{4}$$

$$Cov(I_i, I_j) = -np_ip_j, \tag{5}$$

$$E(I_i^2) = np_i(1-p_i+np_i), \tag{6}$$

$$w_i = \frac{I_i}{np_i} - valikukaal. \tag{7}$$

Kasutades ülaltoodud karakteristikuid ning üldist hindamisteoreemi (Lisa 1), saame esitada üldkogumi kogusummale $t = \sum_U y_i$ nihketa hinnangu. Rõhutame, et hinnangu nihketust ja dispersiooni mõistame multinomiaaldisaini suhtes.

**Teoreem 2.1** *Multinomiaaldisaini korral on nihketa hinnang üldkogumi kogusummale t järgmine:*

$$\hat{t} = \sum_U \frac{I_iy_i}{np_i}. \tag{8}$$

*Hinnangu $\hat{t}$ dispersioon on:*

$$V(\hat{t}) = \frac{1}{n}\left(\sum_U \frac{y_i^2}{p_i} - t^2\right). \tag{9}$$

Dispersioonile nihketa hinnangu saab esitada kahel erineval viisil (Lepik, Traat (2013)):

$$\hat{V}(\hat{t}) = \frac{1}{n-1}\left[\sum_U \frac{n}{1-p_i+np_i}\left(\frac{I_iy_i}{np_i}\right)^2 - \hat{t}^2\right], \tag{10}$$

$$\hat{V}(\hat{t}) = \frac{1}{n(n-1)}\left(\sum_U I_i\frac{y_i^2}{p_i^2} - n\hat{t}^2\right). \tag{11}$$

Käesolevas töös kasutatakse dispersiooni hinnanguna valemit (11), kuna see on stabiilsem ja lihtsama kujuga, kui dispersiooni hinnang (10). Kasutatavat hinnangut nimetatakse ka Sen-Yates-Grundy dispersiooni hinnanguks.

Üldkogumi keskmisele $\bar{Y} = \frac{1}{N}\sum_U y_i$ saame nihketa hinnangu, kui kasutame valemis (8) toodud kogusumma hinnangut, see avaldub kujul:

$$\hat{\bar{Y}} = \frac{1}{N}\hat{t} = \frac{1}{nN}\sum_U \frac{I_i y_i}{p_i}. \tag{12}$$

Antud valem kehtib, kui meil on teada üldkogumi maht $N$. Juhul, kui $N$ on teadmata, tuleb ka see hinnata. Nihketa hinnang üldkogumi mahule avaldub järgmiselt:

$$\hat{N} = \sum_U \frac{I_i}{np_i}. \tag{13}$$

Seega saame kirja panna ka alternatiivse hinnangu üldkogumi keskmisele:

$$\hat{\bar{Y}}_{alt} = \frac{\hat{t}}{\hat{N}}. \tag{14}$$

Kasutades valemeid (8) ja (13), saame $\hat{\bar{Y}}_{alt}$ viia kujule:

$$\hat{\bar{Y}}_{alt} = \frac{\sum_U \frac{I_i y_i}{np_i}}{\sum_U \frac{I_i}{np_i}} = \frac{\sum_s \frac{y_i}{p_i}}{\sum_s \frac{1}{p_i}}, \tag{15}$$

kus $s$ tähistab nn järjestusvalimit, st. valimi elemendid on esitatud elemendi võtmise järjekorras ja võivad esineda ka kordused.

Paneme tähele, et $\hat{\bar{Y}} \neq \hat{\bar{Y}}_{alt}$. Üldjuhul on alternatiivne hinnang (15) paremate omadustega kui hinnang (12) ning seetõttu laialdaselt kasutatav praktikas. Seega kasutame ka meie oma töös just alternatiivset hinnangut.

Keskmise alternatiivne hinnang (15) on mittelineaarne, mistõttu on selle hinnangu täpseid statistilisi omadusi keeruline uurida. Arendades seda Taylori ritta punkti $(t, N)$ ümbruses ja võttes lineaarse osa, saab leida ligikaudse dispersiooni avaldise ja näidata, et hinnang (15) on asümptootiliselt nihketa. Tuginedes materjalile Lepik, Traat, (2013) tuletame dispersioonihinnangu avaldise oma multinomiaaldisaini korral.

**Lause 2.1** *Alternatiivse hinnangu* $\hat{\bar{Y}}_{alt} = \frac{\sum_s \frac{y_i}{p_i}}{\sum_s \frac{1}{p_i}}$ *ligikaudne dispersiooni hinnang on järgmine:*

$$\hat{V}(\hat{\bar{Y}}_{alt}) = \frac{1}{\hat{N}^2 n(n-1)} \sum_s \frac{(y_i - \hat{\bar{Y}}_{alt})^2}{p_i^2}. \tag{16}$$

***Tõestus.***

*Arendame* $\hat{\bar{Y}}_{alt} = \frac{\hat{t}}{\hat{N}}$ *Taylori ritta punkti* $(t, N)$ *ümbruses. Selleks võtame osatuletised* $\hat{t}$ *ja* $\hat{N}$ *järgi punktis* $(t, N)$*:*

$$\left.\frac{\partial \hat{\bar{Y}}_{alt}}{\partial \hat{t}}\right|_{(t,N)} = \frac{1}{N},$$

$$\left.\frac{\partial \hat{\bar{Y}}_{alt}}{\partial \hat{N}}\right|_{(t,N)} = -\frac{t}{N^2}.$$

*Seega*

$$\hat{\bar{Y}}_{alt} \approx \frac{t}{N} + \frac{1}{N}(\hat{t} - t) - \frac{t}{N^2}(\hat{N} - N).$$

*Lihtsustades viimast avaldist, saame*

$$\hat{\bar{Y}}_{alt} \approx \bar{Y} + \frac{1}{N}(\hat{t} - \bar{Y}\hat{N}). \tag{17}$$

*Saadud tulemuse (17) paneme kirja PPS disaini korrl. Selleks asendame* $\hat{t}$ *ja* $\hat{N}$ *valemite (8) ja (13) abil, kasutades summasid üle järjestusvalimi* $s$*.*

$$\hat{\bar{Y}}_{alt} \approx \bar{Y} + \frac{1}{N}\left(\sum_s \frac{y_i}{np_i} - \bar{Y}\sum_s \frac{1}{np_i}\right) = \bar{Y} + \frac{1}{N}\left(\sum_s \frac{1}{np_i}(y_i - \bar{Y})\right).$$

*Viimane saadud summa on nihketa hinnang kogusummale* $\sum_U (y_i - \bar{Y})$*, tähistame* $u_i = y_i - \bar{Y}$*. Valemist (9) saame kogusummale* $t_u = \sum_U u_i$ *kirja panna dispersiooni:*

$$V(\hat{t}_u) = \frac{1}{n}\left(\sum_U \frac{u_i^2}{p_i} - t_u^2\right). \tag{18}$$

*Dispersiooni hinnangu valem järeldub valemist (11), kui vaid tundmatu $u_i$ asendada selle hinnanguga $\hat{u}_i = y_i - \hat{\bar{Y}}_{alt}$.*

$$\hat{V}(\hat{t}_u) = \frac{1}{n(n-1)} \left( \sum_s \frac{\hat{u}_i^2}{p_i^2} - n\hat{t}_u^2 \right), \tag{19}$$

*kus* $\hat{t}_u = \frac{y_i - \hat{\bar{Y}}_{alt}}{np_i}$.

*Kuna vastavalt valemitele (13) ja (14),*

$$\hat{t}_u = \sum_s \frac{y_i - \hat{\bar{Y}}_{alt}}{np_i} = \sum_s \frac{y_i}{np_i} - \hat{\bar{Y}}_{alt} \sum_s \frac{1}{np_i} = 0,$$

*siis lihtsustub valem (19) kujule:*

$$\hat{V}(\hat{t}_u) = \frac{1}{n(n-1)} \sum_s \frac{(y_i - \hat{\bar{Y}}_{alt})^2}{p_i^2} \tag{20}$$

*Seega dispersiooni hinnang üldkogumi keskmisele on järgmine:*

$$\hat{V}(\hat{\bar{Y}}_{alt}) = \frac{1}{\hat{N}^2 n(n-1)} \sum_s \frac{(y_i - \hat{\bar{Y}}_{alt})^2}{p_i^2}.$$

Kasutades valemit (16) avaldub ligikaudne usaldusintervall usaldusnivool $1-\alpha$ üldkogumi keskmisele järgmiselt:

$$I_{\bar{Y}} = \hat{\bar{Y}}_{alt} \pm \bar{z}_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\bar{Y}}_{alt})}, \tag{21}$$

kusjuures ülemise usalduspiiri määrab summa ja alumise vahe.

Usalduspiiride arvutamisel kasutatud $\bar{z}_{\frac{\alpha}{2}}$ on standardnormaaljaotuse $\frac{\alpha}{2}$- täiendkvantiil.

Klassikaliste usaldusvahemike leidmine põhineb tsentraalsel piirteoreemil ning kui statistiku jaotus erineb normaaljaotusest, ei saa vastavaid valemeid kasutada. Võib ka esineda olukordi, kus alumine usalduspiir tuleb negatiivne isegi siis, kui huvialune parameeter saab olla ainult positiivsete väärtustega.

# 3 Lähenemine empiirilise tõepäraga

Empiirilise tõepäraga lähenemine on alternatiivne meetod usaldusintervalli $I_{\bar{Y}}$ leidmiseks. Berger ja De La Riva Torres (2016) väidavad, et selline lähenemisviis võib anda parema tulemuse, kui (21) ning nn pseudo empiirilise tõepäraga leitud usaldusvahemik. Seda eriti olukorras, kus uuritav tunnus on asümeetriline või sisaldab palju nulle, mis on sagedased olukorrad valikuuringute praktikas.

## 3.1 Üldkogumi parameetrite üldine esitus

Empiirilise tõepära meetod võimaldab leida usaldusintervalli lõpliku fikseeritud mahuga üldkogumi korral, kus üldkogumi maht $N$ võib olla ka tundmatu. Eeldame, et meile huvipakkuv üldkogumi parameeter, tähistame $\theta_0$, on ühene lahend järgnevale võrrandile (Godambe 1960):

$$G(\theta) = 0, \tag{22}$$

kus

$$G(\theta) = \sum_U g_i(\theta) \tag{23}$$

ja $g_i(\theta)$ on funktsioon, mis sõltub argumendist $\theta$ ja objektist $i \in U$. Võrrandit (22) nimetame edaspidi hindavaks võrrandiks.

Osutub, et väga paljud huvipakkuvad parameetrid on esitatavad funktsiooni $G(\theta)$ nullkohana.

**Näide 3.1** *Olgu $g_i(\theta) = y_i - \theta$. Sellisel juhul saame*

$$G(\theta) = \textstyle\sum_U (y_i - \theta) = \sum_U y_i - N\theta = N\bar{Y} - N\theta = N(\bar{Y} - \theta),$$

*kus $\bar{Y} = \frac{1}{N}\sum_U y_i$ on üldkogumi keskmine.*

*Võrrandist (22) saame, et ühene lahend parameetrile $\theta$ on $\theta_0 = \bar{Y}$.*

*Juhul kui uuritav tunnus $y$ on pidev (näiteks sissetulek), siis on $\theta_0$ keskmine (sissetulek) üldkogumis.*

*Kui aga $y$ on binaarne tunnus, näiteks*

$$y_i = \begin{cases} 1 & , \quad \textit{kui sissetulek on alla vaesuspiiri} \\ 0 & , \quad \textit{muidu,} \end{cases};$$

*siis $\theta_0$ on vaesusmäär.*

**Näide 3.2** *Keskmine osakogumites. Olgu üldkogum $U$ jagatud $D$ lõikumatuks osakogumiks $U_1, ..., U_d, ..., U_D$. Defineerime osakogumi indikaatori järgmiselt:*

$$\delta_{di} = \begin{cases} 1 & , \quad \textit{kui } i \in U_d; \\ 0 & , \quad \textit{muidu.} \end{cases}$$

*Siis valime $g_i(\theta) = (y_i - \theta)\delta_{di}$ ja seega*

$$G(\theta) = \sum_U (y_i - \theta)\delta_{di} = \sum_U y_i\delta_{di} - \theta \sum_U \delta_{di} = \sum_{U_d} y_i - \theta N_d,$$

*kus $N_d$ on osakogumi $U_d$ maht.*

*Hindavast võrrandist (22) saame, et $\theta_0 = \frac{1}{N_d}\sum_{U_d} y_i$. Analoogselt eelmisele näitele, kui uuritav tunnus $y$ on pidev ja näiteks sissetulek, siis on $\theta_0$ keskmine sissetulek osakogumis $U_d$.*

**Näide 3.3** *Regressioonihinnang.*

*a) Vabaliikmeta mudel $y_i = \beta_1 x_i + \epsilon_i$.*

*Valime $\theta = \beta_1$ ning $g_i(\theta) = x_i(y_i - x_i\beta_1)$. Sel juhul*

$$G(\theta) = \sum_U x_i(y_i - x_i\beta_1) = \sum_U x_i y_i - \beta_1 \sum_U x_i^2.$$

*Hindavast võrrandist (22) saame, et $\theta_0 = \hat{\beta}_1 = \frac{\sum_U x_i y_i}{\sum_U x_i^2} = \frac{t_{xy}}{t_{x^2}}$ .*

*b) Vabaliikmega mudel $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$,*

*siis $\boldsymbol{x_i} = \begin{pmatrix} 1 \\ x_i \end{pmatrix}$, $\boldsymbol{\theta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ ja valime*

$$g_i(\theta) = \boldsymbol{x_i}(y_i - \boldsymbol{x_i^T\theta}) = \begin{pmatrix} 1 \\ x_i \end{pmatrix}(y_i - (\beta_0 + \beta_1 x_i)).$$

*Sel juhul saame*

$$G(\theta) = \sum_U \begin{pmatrix} 1 \\ x_i \end{pmatrix} (y_i - (\beta_0 + \beta_1 x_i)).$$

*Võrrandist (22) saame süsteemi:*

$$\begin{cases} \sum_U (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_U (y_i x_i - \beta_0 x_i - \beta_1 x_i^2) = 0 \end{cases}$$

*Lihtsustades esimest võrrandit, näeme, et*

$$0 = \sum_U y_i - N\beta_0 - \beta_1 \sum_U x_i = N\bar{Y} - N\beta_0 - N\beta_1 \bar{X}.$$

*Seega* ${\theta_0}^{(1)} = \hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$.

*Asendame saadud tulemuse võrrandisüsteemi teise võrrandisse, saame:*

$$0 = \sum_U y_i x_i - (\bar{Y} - \beta_1 \bar{X}) \sum_U x_i - \beta_1 \sum_U x_i^2 = t_{xy} - \frac{t_x t_y}{N} + \beta_1 \frac{t_x^2}{N} - \beta_1 t_{x^2},$$

*millest*

$${\theta_0}^{(2)} = \hat{\beta}_1 = \frac{t_{xy} - \frac{t_x t_y}{N}}{t_{x^2} - \frac{t_x^2}{N}}.$$

Valemist (23) näeme, et $G(\theta)$ on üldkogumi kogusumma, mille nihketa hinnanguks on $\hat{G}(\theta) = \sum_s \frac{g_i(\theta)}{\pi_i}$. Võrrandist $\hat{G}(\theta) = 0$ saadakse $\theta$ hinnang.

## 3.2 Empiiriline tõepära valimi korral

Olgu antud valim $s$ fikseeritud mahuga $n$.

Empiirilise tõepära funktsioon on defineeritud järgmiselt (Owen, 2001):

$$L(m) = \prod_{i=1}^{n} \frac{m_i}{N}, \tag{24}$$

kus $m = (m_1, ..., m_n)$ ja $m_i$ on objekti $i$ mass üldkogumis. Sageli on $\frac{m_i}{N}$ objekti $i$ valikutõenäosus.

Järgmises peatükis on näidatud, kuidas saab empiirilist tõepära funktsiooni (24) kasutada punkthinnangute leidmiseks ja usaldusintervalli konstrueerimiseks fikseeritud mahuga valikudisainide korral. Idee seisneb selles, et lisaks tõepärafunktsioonile tuuakse läbi kitsenduste sisse muu teadaolev informatsioon valimi kohta.

## 3.3 Logaritmiline empiiriline tõepärafunktsioon

Seosest (24) saame logaritmilise empiirilise tõepärafunktsiooni:

$$\ln L(m) = \ln(\textstyle\prod_s \frac{m_i}{N}) = \ln(\frac{1}{N^n} \prod_s m_i) = \ln 1 - n \ln N + \sum_s \ln(m_i),$$

mille maksimiseerimiseks $m$ järgi on vaja edaspidi üksnes suurustest $m_i$ sõltuvat osa.

Tähistame logaritmilise empiirilise tõepärafunktsiooni järgmiselt:

$$l(m) = \sum_s \ln(m_i). \tag{25}$$

Funktsiooni $l(m)$ maksimiseeritakse järgmiste kitsenduste olemasolul:

$$m_i \geq 0, \tag{26}$$

$$\sum_s m_i \boldsymbol{c}_i = \mathbf{C}, \tag{27}$$

kus $\boldsymbol{c}_i$ on objektiga $i$ seotud $Q$x1 vetor ja $\mathbf{C}$ on $Q$x1 vektor.

Vektor $\boldsymbol{c}_i$ sõltub valikudisainist ning kui kasutatakse abitunnuseid, siis ka nendest. Nii $\boldsymbol{c}_i$ kui $\mathbf{C}$ on spetsiaalsel viisil valitud. Võimalikud $\boldsymbol{c_i}$ ja $\boldsymbol{C}$ valikud on toodud artiklis Berger ja De La Riva Torres (2016) (Tabel 5, Peatükk 8). Kitsendus (27) on valikuuringutes üsna loomulik ning on rahuldatud enamuse disainide poolt.

Eeldame, et suurused $\boldsymbol{c_i}$ sisaldavad suurusi $\pi_i$, see tähendab, et leidub sobivalt valitud vektor $\boldsymbol{t}$ nii, et $\boldsymbol{t^T c_i} = \pi_i$ ja $\boldsymbol{t^T C} = \sum_U \pi_i$.

Kitsendusest (27) saame kirjutada $\sum_s m_i \boldsymbol{t^T c_i} = \boldsymbol{t^T C}$, mis on samaväärne sellega, et

$$\sum_s m_i \pi_i = \sum_U \pi_i = n. \tag{28}$$

Teisisõnu kitsendus (27) tagab, et kitsendus (28) on täidetud.

Olgu näiteks $\boldsymbol{c}_i$ esimeseks komponendiks $Nn^{-1}\pi_i$ ja vektori $\mathbf{C}$ esimeseks komponendiks $N$. Sellisel juhul saame valida $\mathbf{t} = (nN^{-1}, 0, \ldots, 0)^T$ ja kitsendus (28) on täidetud.

Paneme tähele, et kitsenduses (28) on valimimaht fiskeeritud, sest kehtib $\sum_U \pi_i = n$. Seega võrrand (28) on ühtlasi ka kitsendus disaini jaoks.

Kitsendustega maksimiseerimisülesande lahendamiseks moodustame Lagrange'i funktsiooni lisades kitsendused (27) ja (28) logaritmilisse empiirilise tõepära funktsiooni (25):

$$Lag(m, \boldsymbol{\lambda}) = \sum_{i \in s} ln(m_i) - \boldsymbol{\lambda}^T (\sum_{i \in s} m_i \boldsymbol{c_i} - \boldsymbol{C}) - (\sum_{i \in s} m_i \pi_i - n).$$

Kitsendus (28) on $\boldsymbol{c_i}$ valikuga rahuldatud, mistõttu Lagrange'i kordajat ei ole vaja. Võttes saadud Lagrange'i funktsioonist osatuletised $m_i$ ja $\boldsymbol{\lambda}$ järgi, saame:

$$\frac{\partial Lag(m, \boldsymbol{\lambda})}{\partial m_i} = \frac{1}{m_i} - \boldsymbol{\lambda}^T \boldsymbol{c_i} - \pi_i, \quad i \in s, \tag{29}$$

$$\frac{\partial Lag(m, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \sum_{i \in s} m_i \boldsymbol{c_i} - \boldsymbol{C}. \tag{30}$$

Võrdsustades osatuletise (29) nulliga, saame

$$\hat{m}_i = (\pi_i + \boldsymbol{\lambda^T c_i})^{-1}. \tag{31}$$

Võrdsustades (30) nulliga ja asendades sellesse $m_i = \hat{m}_i$, saame võrrandi $\boldsymbol{\lambda}$ leidmiseks:

$$\boldsymbol{C} = \sum_{i \in s} \hat{m}_i \boldsymbol{c_i} = \sum_{i \in s} (\pi_i + \boldsymbol{\lambda}^T \boldsymbol{c_i})^{-1} \boldsymbol{c_i}. \tag{32}$$

Artiklis (Chen, Sitter, Wu. 2002) on toodud algoritm $\boldsymbol{\lambda}$ ja $\hat{m}_i$ leidmiseks.

## 3.4 Suurima empiirilise tõepära hinnang

Olgu $\hat{m}_i$ suurus (31), mis on saadud võrrandi (25) maksimiseerimisel ning arvestades kitsendusi (26) ja (27) etteantud $\boldsymbol{c_i}$ ja $\mathbf{C}$ korral. Seega funktsiooni (25) maksimaalne väärtus on:

$$l(\hat{m}) = \sum_s \ln(\hat{m}_i). \tag{33}$$

Kuna antud valem ei sisalda parameetrit $\theta$, aga meie soovime hinnata $\theta$ ja selle usaldusvahemikke, siis olgu $\hat{m}_i^*(\theta)$ teine suurus, mis maksimiseerib funktsiooni (25) fikseeritud $\theta$ ja järgnevate kitsenduste korral:

$$m_i \geq 0 \tag{34}$$

$$\sum_s m_i \boldsymbol{c_i^*} = \boldsymbol{C^*}, \tag{35}$$

kus $\boldsymbol{c_i^*} = (\boldsymbol{c_i^T}, g_i(\theta))^T$ ja $\boldsymbol{C^*} = (\boldsymbol{C^T}, 0)^T$.

Paneme tähele, et uued kitsendused sisaldavad eelmisi (26) - (27), kusjuures lisaks on kitsendus $G(\theta) = \sum_s m_i g_i(\theta) = 0$.

Taaskord moodustame Lagrange'i funktsiooni kitsendustega maksimiseerimisülesande lahendamiseks. Selleks lisame empiirilise logaritmilise tõepära funktsioonile kitsendused (35) ja (28):

$$Lag(m^*, \boldsymbol{\lambda}) = \textstyle\sum_s \ln(m_i) - \boldsymbol{\lambda}^T \left[\sum_s m_i \boldsymbol{c_i^*} - \boldsymbol{C^*}\right] - \left[\sum_s m_i \pi_i - n\right].$$

Analoogiliselt eelmise peatüki lõpus esitatud tuletuskäigule, saame:

$$\hat{m}_i^* = (\pi_i + \boldsymbol{\lambda^T c_i^*})^{-1}, \quad i \in s. \tag{36}$$

Suuruse $\boldsymbol{\lambda} = \binom{\lambda_1}{\lambda_2}$ leidmiseks nõuame kitsenduste (35) rahuldamist.

$$\boldsymbol{C^*} = \sum_s \hat{m_i^*} \boldsymbol{c_i^*} = \sum_s (\pi_i + \boldsymbol{\lambda}^T \boldsymbol{c_i^*})^{-1} \boldsymbol{c_i^*}. \tag{37}$$

Sellisel juhul tähistame funktsiooni (25) maksimaalse väärtuse järgmiselt:

$$l(\hat{m}^*, \theta) = \sum_s ln(\hat{m}_i^*(\theta)). \tag{38}$$

Kasutades ülaltoodud tähistusi (33) ja (38) saame kirja panna logaritmilise empiirilise tõepärasuhte funktsiooni.

**Definitsioon 3.1** *Logaritmiline empiirilise tõepärasuhte funktsioon on $\theta$ funktsioon, mis on esitatud valemiga:*

$$\hat{r}(\theta) = 2[l(\hat{m}) - l(\hat{m}^*, \theta)]. \tag{39}$$

Suurima empiirilise tõepära hinnang üldkogumi parameetrile $\theta_0$ saadakse minimiseerides $\hat{r}(\theta)$ valemis (39) ning see tähistatakse $\hat{\theta}$.

Paneme tähele, et $\hat{r}(\theta) \geq 0$ iga $\theta$ korral, sest teine liige maksimiseeritakse rohkemate kitsenduste korral. Seega $\hat{\theta}$ on võrrandi $\hat{r}(\theta) = 0$ lahend. Eeldame, et $g_i(\theta)$ on selline, et hindaval võrrandil:

$$\hat{G}(\theta) = 0, \tag{40}$$

kus

$$\hat{G}(\theta) = \sum_s \hat{m}_i g_i(\theta), \tag{41}$$

on ühene lahend. Suurus $\hat{\theta}$ on võrrandi (41) lahend, kuna $\hat{r}(\hat{\theta}) = 0$ toob kaasa selle, et $\hat{m}_i^*(\hat{\theta}) = \hat{m}_i$ iga $i$ korral.

**Näide 3.4** *Kui $\boldsymbol{c_i} = \pi_i$ ja $\boldsymbol{C} = n$ (1-dimensionaalsed suurused), siis saame seosest (32):*

$$n = \textstyle\sum_s (\pi_i + \lambda\pi_i)^{-1}\pi_i.$$

*Järelikult $\lambda$ peab võrduma 0 ja seosest (31), et $\hat{m}_i = \pi_i^{-1}$. Asendades $\hat{m}_i$ seosesse (41), saame*

$$\hat{G}(\theta) = \sum_s \hat{m}_i g_i(\theta) = \sum_s \frac{g_i(\theta)}{\pi_i}, \tag{42}$$

*mis on Horvitz ja Thompsoni hinnang nulliga võrduvale funktsioonile $G(\theta)$ seoses (22). Võrdsustades seose (42) nulliga, $\hat{G}(\theta) = 0$, saame hinnangu parameetrile $\theta$.*

*a) Kasutades näites 2.1 toodud $g_i(\theta) = y_i - \theta$, saame*

$$\hat{G}(\theta) = \sum_s \frac{y_i - \theta}{\pi_i} = \sum_s \frac{y_i}{\pi_i} - \theta \sum_s \frac{1}{\pi_i}. \tag{43}$$

*Võrdsustades saadud tulemuse nulliga, saame*

$$\hat{\theta} = \frac{\sum_s \frac{y_i}{\pi_i}}{\sum_s \pi_i^{-1}}, \tag{44}$$

*mis on tuntud ka kui Hajeki hinnang üldkogumi keskmisele.*

*b) Kui $\theta_0$ on vaesusmäär ja $g_i(\theta)$ on antud nagu näites 2.2, siis saame*

$$\hat{G}(\theta) = \sum_s \frac{(y_i - \theta)\delta_{di}}{\pi_i} = \sum_s \frac{y_i \delta_{di}}{\pi_i} - \theta \sum_s \frac{\delta_{di}}{\pi_i}. \tag{45}$$

*Saadud tulemuse võrdsustamisel nulliga, saame*

$$\hat{\theta} = \frac{\sum_s y_i \delta_{di} \pi_i^{-1}}{\sum_s \delta_{di} \pi_i^{-1}}, \tag{46}$$

*mis on Hajeki hinnang vaesusmäärale osakogumis $U_d$.*

*c) Valides $g_i(\theta) = y_i - \theta N^{-1}$, saame*

$$\hat{G}(\theta) = \sum_s \frac{(y_i - \theta N^{-1})}{\pi_i} = \sum_s \frac{y_i}{\pi_i} - \theta N^{-1} \sum_s \frac{1}{\pi_i},$$

*Avaldades $\theta$, saame Hajeki suhtehinnangu üldkogumi kogusummale*

$$\hat{\theta} = \frac{N \sum_s \frac{y_i}{\pi_i}}{\sum_s \frac{1}{\pi_i}}.$$

## 3.5 Empiirilise tõepära usaldusintervall

Logaritmilist empiirilise tõepärasuhte funktsiooni (39) saab kasutada empiirilise tõepära usaldusintervallide konstrueerimiseks. On näidatud (Berger, De La Riva Torres (2016)), et küllaltki üldistel eeldustel $\hat{r}(\theta_0)$ on $\chi^2$ -jaotusega vabadusastmete arvuga üks. Sel juhul $(1-\alpha)$ empiirilise tõepära usaldusintervall üldkogumi parameetrile $\theta_0$ avaldub järgmiselt (Wilks, 1938):

$$\{\theta : \hat{r}(\theta) \leq \chi_1^2(\alpha)\} = [min\{\theta|\hat{r}(\theta) \leq \chi_1^2(\alpha)\}; max\{\theta|\hat{r}(\theta) \leq \chi_1^2(\alpha)\}], \quad (47)$$

kus $\chi_1^2(\alpha)$ on $\chi^2$ -jaotuse vabadusastmete arvuga üks ülemine $\alpha$-kvantiil.

Intervalli (47) võib pidada tõepoolest $\theta_0$ usaldusintervalliks, sest see sisaldab kõiki $\theta$ väärtusi, mis on sellised, et hüpoteesi $\theta_0 = \theta$ ei saa ümber lükata kasutades teststatistikut $\hat{r}(\theta)$. Paneme tähele, et $\hat{r}(\theta)$ on kumer ebasümmeetrilise jaotusega funktsioon, mis saavutab oma miinimumi, kui $\theta$ on suurima
empiirilise tõepära hinnang ($\theta = \hat{\theta}$). Usaldusintervalli (47) saab leida kasutades lõigu poolitamise meetodit, millest on lähemalt räägitud artiklis Wu (2005). See toob kaasa $\hat{r}(\theta)$ arvutamise mitmete $\theta$ väärtuste korral.

Kui $g_i(\theta)$ ja $\theta$ on $R$ x 1 vektorid, siis osutub, et suurus $\hat{r}(\theta_0)$ läheneb $\chi^2$ -jaotusele vabadusastemete arvuga $R$ (Ogus-Alper ja Berger, 2014).

### 3.5.1 Empiirilise tõepära usaldusintervall PPS TGA valiku korral

Suurusega võrdelise tagasipanekuga (PPS TGA) disaini korral kasutatakse kitsendustes (27) ja (35) järgmisi suurusi:

$$\boldsymbol{c_i} = Nn^{-1}\pi_i,$$
$$\boldsymbol{C} = N,$$
$$\boldsymbol{c_i^*} = (\boldsymbol{c_i^T}, g_i(\theta_0))^T = (Nn^{-1}\pi_i, g_i(\theta_0))^T,$$
$$\boldsymbol{C^*} = (\boldsymbol{C}^T, 0)^T = (N, 0)^T,$$

kusjuures $\pi_i = np_i$ tagasipanekuga disainide korral.

Berger ja De La Riva Torres (2016) on näidanud, et kui valim on võetud PPS TGA disaini abil ja on kasutatud eeltoodud kitsendusi, siis suurus $\hat{r}(\theta_0)$ valemis (39) on asümptootiliselt $\chi^2$ -jaotusega vabadusastmete arvuga üks. Järelikult saab usaldusintervallide leidmisel kasutada valemit (47).

Paneme tähele, et kitsendused $\sum_s m_i\pi_i = n$ ja $\sum_s m_i(Nn^{-1}\pi_i) = N$ on ekvivalentsed, sest:

$$\sum_s m_i(Nn^{-1}\pi_i) = Nn^{-1}\sum_s m_i\pi_i,$$

ning see saab olla võrdne üldkogumi mahuga $N$ vaid siis, kui $\sum_s m_i\pi_i = n$.

Seega kasutades kitsendustes suurusi $\boldsymbol{c_i} = \pi_i$ ja $\boldsymbol{C} = n$ või $\boldsymbol{c_i} = Nn^{-1}\pi_i$ ja $\boldsymbol{C} = N$, saame mõlemal juhul suurustele $\hat{m}_i$ ja $\hat{m}_i^*$ samad väärtused. See omakorda tähendab seda, et me saame leida suurused $\hat{m}_i$ , $\hat{m}_i^*$ ja $\hat{r}(\theta)$ isegi siis, kui üldkogumi maht $N$ on tundmatu.

Antud töös pakub huvi üldkogumi keskmise $\bar{Y} = \frac{1}{N}\sum_U y_i$ hindamine. Seetõttu vaatame põhjalikumalt juhtu, kus huvialuseks üldkogumi parameetriks on keskmine ning valimi võtmisel kasutatakse suurusega võrdelise tõenäosusega tagasipanekuga disaini. Sel juhul $g_i(\theta) = y_i - \theta$. Eespool oleme näidanud, et sellises olukorras avaldub $m_i$ hinnang järgmiselt (vt. näide 3.4): $\hat{m}_i = \frac{1}{np_i}$.

Logaritmilise empiirilise tõepärasuhte funktisooni (39) leidmiseks on meil vaja leida $\hat{m}^*$. Nagu ka eespool kirjas, peame selleks maksimiseerima logaritmilise

empiirilise tõepära funktsiooni (25). Lisaks tuleb arvestada kitsendusega (27), mis saab $c_i = \pi_i$ ja $C = n$ korral kuju:

$$\sum_s m_i \pi_i = n \tag{48}$$

ja ka kitsendusega

$$\sum_s m_i g_i(\theta) = \sum_s m_i (y_i - \theta) = 0. \tag{49}$$

Kitsenduse (48) asemel saame kasutada kitsendust $\sum_s m_i p_i = 1$ , sest $\pi_i = np_i$.

Empiirilise logaritmilise tõepära funktsiooni maksimumi leidmiseks antud erijuhul kasutame uuesti Lagrange'i kordajate meetodit. Lagrange'i funktsioon avaldub kujul:

$$Lag(m_1, ..., m_n, \lambda_1, \lambda_2) = \sum_s log(m_i) - \lambda_1 \left[ \sum_s m_i p_i - 1 \right] - \lambda_2 \left[ \sum_s m_i (y_i - \theta) \right].$$

Edasi tuleb leida Lagrange'i funktsioonist esimest järku osatuletised $m_i$, $\lambda_1$ ja $\lambda_2$ järgi. Saadud osatuletised on:

$$\frac{\partial Lag}{\partial m_i} = \frac{1}{m_i} - \lambda_1 p_i - \lambda_2 (y_i - \theta), \quad i \in s, \tag{50}$$

$$\frac{\partial Lag}{\partial \lambda_1} = \sum_s m_i p_i - 1, \tag{51}$$

$$\frac{\partial Lag}{\partial \lambda_2} = \sum_s m_i (y_i - \theta). \tag{52}$$

Paneme tähele, et kui võrdsustame saadud osatuletised (51) ja (52) nulliga, siis on need tegelikult meie poolt seatud kitsendused.

Võrdsustame $m_i$ järgi leitud osatuletise (50) nulliga

$$\frac{1}{m_i} - \lambda_1 p_i - \lambda_2 (y_i - \theta) = 0, \tag{53}$$

Saadud võrrandi korrutame mõlemalt poolt läbi suurusega $m_i$ ning seejärel võtame summa üle valimi.

Saame

$$\sum_s 1 - \lambda_1 \sum_s m_i p_i - \lambda_2 \sum_s m_i (y_i - \theta) = 0.$$

Kasutades etteantud kitsendusi ja teadmist $\sum_s 1 = n$, saame kirjutada:

$$n - \lambda_1 = 0.$$

Korrutame antud võrduse läbi suurusega $p_i$, saame $np_i - \lambda_1 p_i = 0$. Pannes saadud tulemuse võrduma tulemusega (53), saame:

$$np_i - \lambda_1 p_i = \tfrac{1}{m_i} - \lambda_1 p_i - \lambda_2 (y_i - \theta).$$

Antud seosest saame $m_i$ hinnanguks:

$$\hat{m_i^*} = \frac{1}{np_i + \lambda_2 (y_i - \theta)}. \tag{54}$$

Asendades saadud hinnangu valemisse (49), saame $\lambda_2$ leidmiseks järgmise võrrandi:

$$\sum_s \frac{(y_i - \theta)}{np_i + \lambda_2 (y_i - \theta)} = 0. \tag{55}$$

Sellisel juhul saab logaritmiline empiiriline tõepärasuhte funktsioon (39) kuju:

$$\hat{r}(\theta) = 2 \left[ \sum_s ln \frac{1}{np_i} - \sum_s ln \frac{1}{np_i + \lambda_2 (y_i - \theta)} \right],$$

mida lihtsustades saame:

$$\hat{r}(\theta) = 2 \sum_s \ln \frac{np_i + \lambda_2 (y_i - \theta)}{np_i}, \tag{56}$$

kus $\lambda_2$ leitakse seosest (55).

Saadud tulemusi kasutame ka käesoleva töö praktilises osas üldkogumi keskmisele usaldusintervalli leidmiseks.

# 4 Simuleerimisülesanne

Käesolevas peatükis rakendame eelnevalt kirjeldatud teooriat praktilisel andmestikul. Eesmärk on võrrelda empiirilise tõepära abil leitud usadusintervalle klassikalise meetodiga leitud usaldusintervallidega. Huvialuseks hinnatavaks parameetriks on üldkogumi keskmine. Uuritavad tunnused on valitud kolme erinevat tüüpi - pidev, binaarne ja diskreetne. Lisaks on valitud üks tunnus, mis sisaldab palju nulliga võrdseid väärtuseid. Kõigi tunnuste korral leitakse üldkogumi keskmisele hinnang ja usaldusintervall.

Simuleerimisülesandes võetakse üldkogumist 1000 korda valimit, mis saadakse suurusega võrdelise tõenäosusega tagasipanekuga valiku abil. Tausttunnusena kasutati leibkonna liikmete arvu, mis tähendab seda, et suurematel perekondadel on suurem tõenäosus osutuda valituks. Valimi suuruseks on võetud 30% üldkogumi mahust, milleks tuli 542 isikut.

Usaldusintervalli leiame kolmel erineval viisil:

1. **Tegelik**: tuhandest üldkogumi keskmise hinnangust on moodustatud variatsioonrida ning sealt on võetud usalduspiirideks 25. ja 975. element. Hinnang üldkogumi keskmisele leitakse valemi (15) abil.

2. **Klassikaline**: Usaldusintervalli leidmiseks on igal sammul kasutatud valemit (21), kus $\alpha = 0,05$ ja dispersioon $\hat{V}(\hat{\bar{Y}}_{alt})$ on hinnatud valemi (16) abil.

3. **Empiiriline tõepära (EL)**: Usaldusvahemike leidmiseks igal sammul kasutame valemit (56) ja usaldusvahemiku definitsiooni (47).

Nii teise kui kolmanda variandi korral on leitud 1000 korda usalduspiirid, millest on seejärel võetud keskmine. Lisaks on vaadeldud mõlema meetodi korral usalduspiiride varieeruvust üle simulatsioonide. Vastav programmi kood on toodud Lisas 3.

## 4.1 Andmestiku kirjeldus

Käesolevas töös kasutatav andmestik on pärit *European Social Survey* (ESS) kodulehelt (www.europeansocialsurvey.org). ESS on orienteeritud akadeemiliselt mitme riigi uuringutele. Uuring hõlmab rohkem kui 30 riigi andmeid. Põhiliselt on sellel kolm eesmärki:

1. jälgida ja tõlgendada avaliku arvamuse muutumisi ja väärtusi Euroopas,

2. edendada ja tugevdada täiustatud meetodeid riikidevahelistes uuringute mõõtmiseks Euroopas ja mujal,

3. töötada välja mitmeid Euroopa sotsiaalseid näitajaid, sealhulgas suhtumise näitajaid.

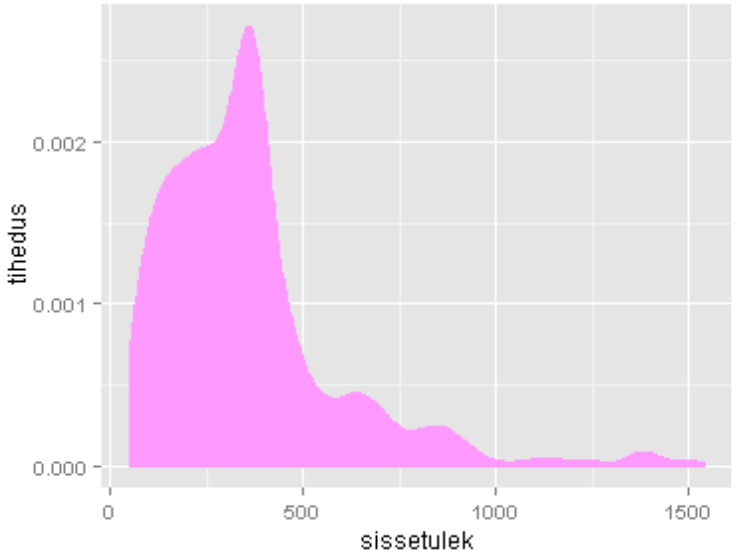
Kuna Eesti osaleb ESS uuringutes, siis üldkogumiks ongi võetud Eesti elanike andmestik. Andmed on kogutud intervjueerimise teel ning perioodil 07.09.2014 - 29.12.2014. Intervjuud viidi läbi nii eesti kui vene keeles ning vastajate vanus jäi vahemikku 15 - 99. Infot on kogutud väga paljude näitajate kohta, mille põhjal sooviti näiteks teha järeldusi sotsiaalse ebavõrdsuse, tervise ja seda mõjutavate tegurite ning hoiakutest sisserändajate ja nende eelkäijate kohta. Eesti andmestikus on ridu 2051 inimese kohta.

Käesoleva töö uuritavateks tunnusteks võetakse antud andmestikust neli tunnust - sissetuleku detsiilid, televiisori vaatamine, õnnelikkus ning organisatsiooni mõjutavates otsustes kaasarääkimine. 1808 inimest oli vastanud korrektselt antud tunnuste korral, seega moodustub meie üldkogum just nendest inimestest.

### 4.1.1 Sissetuleku detsiilid

Andmestikus tunnus HINCTNEE on leibkonna kogu netosissetulek (kõikidest allikatest). Inimestele anti ette kaart, mis sisaldas infot sissetuleku detsiilide kohta ning paluti hinnata, millisesse detsiili kuulub nende leibkonna kogusissetulek pärast maksude ja kohustuslike kinnipidamiste maha arvamist ehk nad pidid hindama leibkonna netosissetulekut. Kui inimene ei teadnud täpset numbrit, siis paluti anda ligikaudne hinnang, kaartil oli info toodud nädala, kuu ja aasta sissetuleku kohta.

Kuna on sooviks kasutada pidevat tunnust, siis kogutud andmete põhjal tekitati uus tunnuse - SISSETULEK. Tunnuse sissetuleku leidmiseks on kasutatud Eesti Statistikaameti 2014. aasta andmeid leibkonnaliikme netosissetulek kuus (stat.ee, tabel ST10) ning ühtlast jaotust. Näiteks, kui inimene oli märkinud, et tema sissetulek jääb 3. detsiili, siis uus tunnus sissetulek sai väärtuse, mis on genereeritud vahemikust (277,4 ; 336) kasutades ühtlast jaotust. Lisas 2 on toodud tabel leibkonnaliikme netosissetuleku detsiilide kohta. Tunnuse sissetulek jaotus on toodud joonisel 1.




Joonis 1: Tunnuse sissetulek jaotus

Üldkogumi keskmine sissetulek ühe leibkonna liikme kohta on 359,70 eurot. Sissetulekud jäävad vahemikku (50,95; 1538,40).

### 4.1.2 Televiisori vaatamine

Andmestikus tunnus TVTOT on televiisori vaatamine. Inimesel paluti hinnata, kui palju ta kulutab päevas keskmiselt aega televiisori vaatamisele. Võimalikud vastuse variandid olid:

0 - ei vaata üldse,

1 - vähem kui poolt tundi,

2 - 0,5 - 1 tund,

3 - 1 - 1,5 tundi,

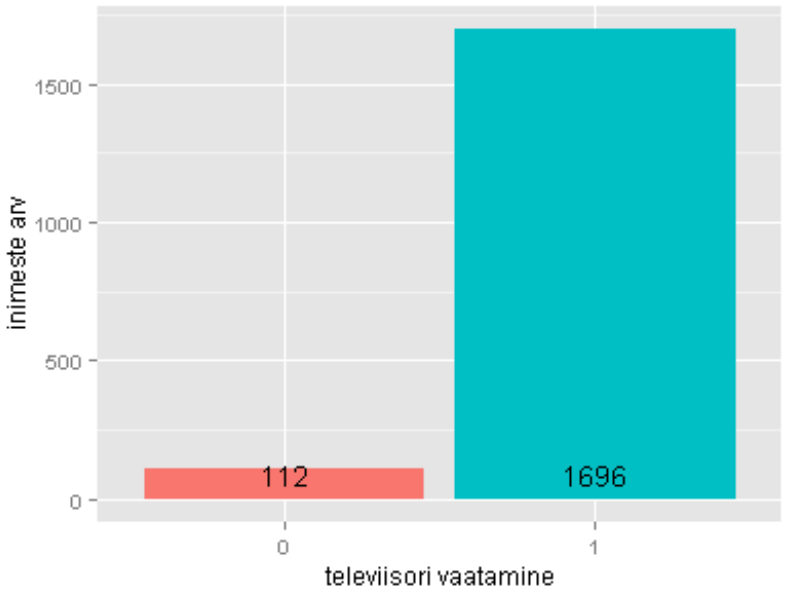
4 - 1,5 - 2 tundi,

5 - 2 - 2,5 tundi,

6 - 2,5 - 3 tundi,

7 - rohkem kui 3 tundi.

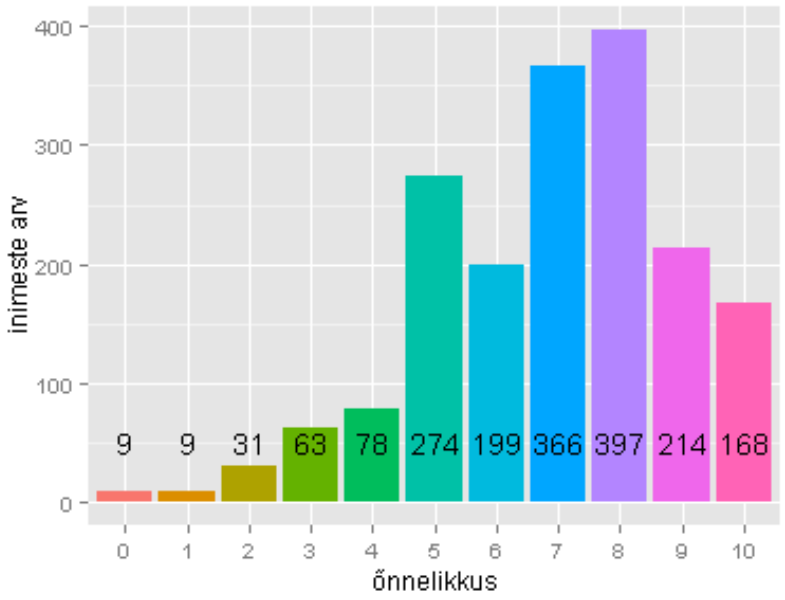
Kuna sooviks oli, et üks tunnus oleks binaarne, siis sai tekitatud uus tunnus TEL. Tunnus TEL sai väärtuse 0, kui inimene ei vaata üldse televiisorit, ning väärtuse 1 said kõik ülejäänud, kes vaatasid vähem kui pool tundi kuni rohkem kui kolm tundi televiisorit.




Joonis 2: Televiisori vaatamine, 0 -ei vaata, 1-vaatab

### 4.1.3 Õnnelikkus

Andmestikus tunnus HAPPY kirjeldab õnnelikkust. Inimesel paluti hinnata, kui õnnelikuks ta end peab. Vastuste skaala jäi vahemikku 0 - 10, kus 0 vastab väga õnnetule seisundile ning 10 väga õnnelikule. Inimene pidi hindama oma õnnelikkust antud skaalal, kuid väärtustele 1 - 9 sõnalist vastet ei olnud. Inimeste keskmine õnnelikkus üldkogumis oli 6,9. Andmestikus oli inimesi, kes pidasid end väga õnnetuks või väga õnnelikuks ning joonisel 3 on toodud vastanute õnnelikkuse väärtuste jagunemine.

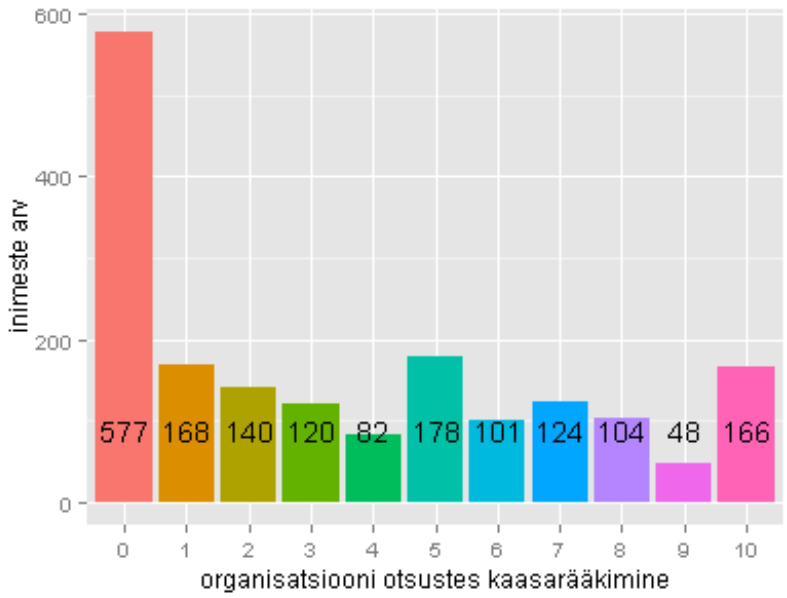



Joonis 3: Inimese õnnelikkuse hinnang

### 4.1.4 Organisatsiooni otsustes kaasarääkimine

Andmestikus tunnus IORGACT kirjeldab oma organisatsiooni otsustes kaasarääkimise võimalikkust. Inimesele loeti ette väiteid, mis käisid tööelu kohta ning nende hulgas olid ka väited, kus isik pidi hindama, kui palju lubatakse tal organisatsiooni otsuste tegemisel kaasa rääkida. Võimalikud vastuse variandid jäid vahemikku 0 - "ei saa mõjutada otsuste tegemist" kuni 10 - "saan täielikult kaasa rääkida".

Organisatsioonis otsuste tegemisel kaasarääkimise üldkogumi keskmine oli 3,6. Joonisel 4 on toodud vastanute arvamuste jagunemine. Näeme, et ligi kolmandik inimestest ei saa üldse kaasa rääkida oma organisatsiooni otsuste tegemisel.

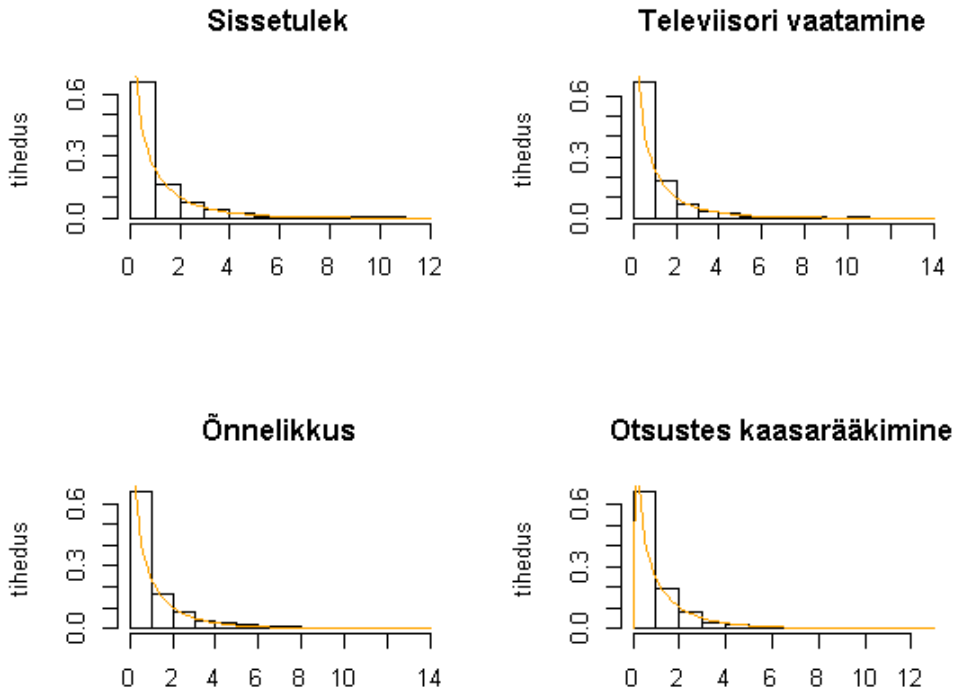



Joonis 4: Luba mõjutada organisatsiooni otsuseid

Antud tunnus võeti vaatluse alla, kuna sooviks oli uurida tunnust, mis sisaldab nulle mõnevõrra rohkem, kui teisi vastuste variante.

## 4.2 Logaritmilise empiirilise tõepärasuhte jaotus

Empiirilise tõepärasuhte meetod eeldab $\chi_1^2$ - jaotuse kasutamist. Seetõttu kontrollisime oma andmestiku korral, kas see eeldus on täidetud logaritmilise empiirilise tõepärasuhte korral. Saadud graafikud on toodud alljärgneval joonisel.




Joonis 5: Logaritmilise empiirilise tõepärasuhte jaotused uuritavate tunnuste korral

Jooniselt 5 on näha, et kõigi nelja uuritava tunnuse korral on logaritmiline empiiriline tõepärasuhe $\chi^2$ - jaotusega vabadusastmete arvuga üks.

## 4.3 Hinnangud ja usaldusvahemikud

Pideva tunnusena on vaadeldud sissetulekut ühe leibkonna liikme kohta. Saadud hinnang üldkogumi keskmisele on 359,40. Binaarseks tunnuseks on võetud televiisori vaatamine. Keskmine hinnang televiisori vaatamise osakaalule on 0,94 ehk 94% inimestest vaatavad iga päev televiisorit. Keskmisele õnnelikkusele on saadud hinnanguks 6,9 ning otsustes kaasarääkimisele 3,6. Kõigi tunnuste usaldusintervallid saadud hinnangutele on toodud Tabelis 1.

Tabel 1: Üldkogumi keskmiste usaldusvahemikud

| Uuritav tunnus | Meetod | Alumine usalduspiir | Ülemine usalduspiir | Usaldusintervalli laius |
|---|---|---|---|---|
| Sissetulek | Tegelik | 338,34 | 380,02 | 41,68 |
| | Klassikaline | 338,13 | 380,75 | 42,63 |
| | EL | 338,84 | 381,76 | 42,82 |
| Televiisor | Tegelik | 0,91 | 0,96 | 0,05 |
| | Klassikaline | 0,91 | 0,96 | 0,05 |
| | EL | 0,91 | 0,96 | 0,05 |
| Õnnelikkus | Tegelik | 6,70 | 7,09 | 0,39 |
| | Klassikaline | 6,70 | 7,10 | 0,41 |
| | EL | 6,69 | 7,10 | 0,41 |
| Otsus | Tegelik | 3,23 | 3,89 | 0,66 |
| | Klassikaline | 3,23 | 3,88 | 0,65 |
| | EL | 3,24 | 3,89 | 0,65 |

Tabelist 1 näeme, et nii klassikalise meetodi kui empiirilise tõepära abil leitud usaldusvahemikud üldkogumi keskmisele on üsna sarnased, nii mõnelgi juhul tuli tegelikult erinevus alles kolmandas või neljandas komakohas. Seega võime öelda, et üldiselt töötavad mõlemad meetodid sama hästi.

Kuna praktikas on meil üks valim ja ka nii alumist kui ülemist usalduspiiri leiame selle ühe konkreetse valimi korral, siis vaatame antud olukorras liasks ka usalduspiiride varieeruvust.

Olgu alumise usalduspiiri simulatsioonid $a_1, ..., a_{1000}$, alumise usalduspiiri keskmine on $\frac{1}{1000}\sum_{i=1}^{1000} a_i$ ja kvantiilvahemik $q_{0,975} - q_{0,025} = a_{(975)} - a_{(25)}$, kus $a_{(i)}$ on variatsioonirea $i$. element. Antud olukorras näitab kvantiilvahemik alumise usalduspiiri varieeruvust. Analoogiliselt saab leida ka ülemise usalduspiiri varieeruvust.

Järgmises tabelis esitame kõigi tunnuste alumiste ja ülemiste usalduspiiride keskmised ning neile vastavad kvantiilvahemikud klassikalise ning EL meetodite korral.

Tabel 2: Usalduspiiride keskmised ja kvantiilvahemikud klassikalise ja EL meetodi korral

| Tunnus | Meetod | Alumise usalduspiiri keskmine | $q_{0,975} - q_{0,025}$ | Ülemise usalduspiiri keskmine | $q_{0,975} - q_{0,025}$ |
|---|---|---|---|---|---|
| Sissetulek | Klassikaline | 338,13 | 38,69 | 380,75 | 44,91 |
| | EL | 338,84 | 38,80 | 381,76 | 45,78 |
| Televiisor | Klassikaline | 0,91 | 0,06 | 0,96 | 0,04 |
| | EL | 0,91 | 0,06 | 0,96 | 0,04 |
| Õnnelikkus | Klassikaline | 6,70 | 0,42 | 7,10 | 0,37 |
| | EL | 6,69 | 0,42 | 7,10 | 0,38 |
| Otsus | Klassikaline | 3,23 | 0,63 | 3,88 | 0,67 |
| | EL | 3,24 | 0,63 | 3,89 | 0,67 |

Tabelist 2 näeme, et nii klassikalise kui empiirilise tõepära abil leitud valemite korral tulevad usalduspiiride laiused üsna sarnased.

Saadud tulemused kinnitavad, et EL meetodit saab kasutada alternatiivina klassikalisele meetodile. Sellest võiks olla abi siis, kui huvialuse parameetri hinnang omab keerulist mittelineaarset kuju, kus klassikalise meetodi abil dispersiooni hinnanguid on raske või isegi võimatu leida.

# 5 Kokkuvõte

Käesolevas magistritöös leiti üldkogumi keskmise hinnangule usalduspiirid klassikalisel meetodil ning uudsel meetodil, mille korral leitakse usaldusintervallid mitteparameetrilisel viisil empiirilise tõepära abil. Üldkogumi keskmise hinnang ja selle usaldusintervallid leiti nelja erineva tunnuse korral, millest üks oli pidev, üks binaarne ning kaks diskreetset tunnust. Ühel diskreetsel tunnusel esines nulle mõnevõrra rohkem. Kõigi tunnuste korral andsid nii klassikalised valemid kui uudne meetod üsna sarnsed tulemused.

Kokkuvõtteks võib öelda, et antud töö täitis oma eesmärgi. Tuletati valemid empiirilise tõepäraga lähenemisel suurusega võrdelise tagasipanekuga valiku korral. Saadud valemeid sai proovida reaalsel andmestikul ning tulemusi võrrelda varasemalt tuntud valemitega. Kontrolliti jooniste abil, kas uuritavate tunnuste korral tuli $\hat{r}(\theta_0)$ valemis (56) $\chi^2$ -jaotusega vabadusastmete arvuga üks.

# 6 Kasutatud kirjandus


Berger, Y. G., De La Riva Torres, O. (2016) Empirical likelihood confidence intervals for complex sampling designs. Journal of the Royal Statistical Society, Series B.

Chen, J., Sitter, R.R., Wu, C. (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. Biometrika, 89, pp. 230-237.

Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics, 31, pp. 1208-1211.

Lepik, N., Traat, I. (2013) Valikuuringute teooria I. Ainekonspekt.

Oguz-Alper, M., Berger, Y. G. (2014) Empirical likelihood confidence intervals and significance test for regression parameters under complex sampling designs. Survey Research Methods Section,American Statistical Association, pp. 2070–2079, Boston.

Owen. A.B. (2001) Empirical Likelihood. New York, Chapman and Hall.

Traat, I., Inno. J. (1997) Tõenäosuslik valikuuring. Tartu, TÜ Kirjastus

Wilks, S. S. (1938) Shortest average confidence intervals from large samples. The Annals of Mathematical Statistics, 9, pp. 166-175.

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. Survey Methodology, 31 (2), pp. 239-243.

Wu, C., Rao, J. N. K. (2006) Pseudo-Empirical Likelihood Ratio Confidence Intervals for Complex Surveys. The Canadian Journal of Statistics, vol. 34, no. 3, pp. 359-375.

# Lisad

## Lisa 1. Üldine hindamisteoreem

Antud teoreem pärineb aine "Valikuuringute teooria I" konspektist (Lepik, Traat, 2013).

**Teoreem 6.1** *Üldine hindamisteoreem.*

*Üldkogumi kogusumma* $t = \sum_U y_i$ *nihketa hinnang on*

$$\hat{t} = \sum_U I_i \check{y}_i,$$

*kus* $\check{y}_i = \frac{y_i}{E(I_i)}$.

*Selle disainipõhine dispersioon on*

$$V(\hat{t}) = \sum_U \sum_U \Delta_{ij} \check{y}_i \check{y}_j,$$

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

*Dispersiooni nihketa hinnanguks* $E(I_i I_j) > 0$ *korral on*

$$\hat{V}(\hat{t}) = \sum_U \sum_U \check{\Delta}_{ij} \check{y}_i \check{y}_j I_i I_j,$$

*kus* $\check{\Delta}_{ij} = \frac{\Delta_{ij}}{E(I_i I_j)}$.

Üldine hindamisteoreem kehtib nii tagasipanekuga kui ka tagasipanekuta disainide korral. Vaja on teada disainikarakteristikuid.

## Lisa 2. Sissetuleku detsiilid

Tabel 3: Sissetuleku detsiilid 2014. aastal

| 2014 | Netosissetulek kokku |
|---|---|
| I tuludetsiil | 154,4 |
| II tuludetsiil | 277,3 |
| III tuledetsiil | 336,0 |
| IV tuludetsiil | 36,4 |
| V tuludetsiil | 410,0 |
| VI tuludetsiil | 482,4 |
| VII tuludetsiil | 581,4 |
| VIII tuludetsiil | 717,1 |
| IX tuludetsiil | 938,5 |
| X tuludetsiil | 1540,4 |

# Lisa 3. Simuleerimisülesandes kasutatud kood

```
#failist EestiLyhend on eemaldatud read, kus sissetulekut ei oldud märgitud,
#alles jäi 1808 rida.

install.packages("pps")
library(pps)
library(ggplot2)

andmed=read.csv("C:/Users/kristity/Downloads/EEandmed.csv",
                header=T, sep=";")
N=1808

#Tunnuse HINCTNEE pidevaks tegemine.
sisset=rep(NA,1808)
sis_pidev = function(detsiil){
  for( i in 1:1808){
    if (detsiil[i] == 1 ) {
      sisset[i]=runif(1,50,145.4)
    } else if(detsiil[i]==2){ sisset[i]=runif(1,145.5,277.3) }
    else if(detsiil[i]==3){ sisset[i]=runif(1,277.4,336) }
    else if(detsiil[i]==4){ sisset[i]=runif(1,336.1,368.4) }
    else if(detsiil[i]==5){ sisset[i]=runif(1,368.5,410) }
    else if(detsiil[i]==6){ sisset[i]=runif(1,410.1,482.4) }
    else if(detsiil[i]==7){ sisset[i]=runif(1,482.5,581.4) }
    else if(detsiil[i]==8){ sisset[i]=runif(1,581.5,717.1) }
    else if(detsiil[i]==9){ sisset[i]=runif(1,717.2,938.5) }
    else {sisset[i] = runif(1,938.6,1540.4)}}
  return(sisset[])
}

sis_detsiil=andmed$HINCTNEE
sissetulek=sis_pidev(sis_detsiil)

#---tunnuse televiisori vaatamine binaarseks tegemine
tel=rep(NA,1808)
for( i in 1:1808) {
  if (andmed$TVTOT[i] == 0){tel[i]=0}
  else {tel[i]=1}
}

a=data.frame(andmed, sissetulek, tel)

#--sissetuleku jaotus
ggplot(a, aes(sissetulek)) +
  geom_density(colour="#ff99ff", fill="#ff99ff") +
  ylab("tihedus")

#--televiisori vaatamine
ggplot(a, aes(factor(tel), fill=factor(tel)))  + geom_bar() +
  xlab("televiisori vaatamine")+
```

```
  ylab("inimeste arv")+
  theme(legend.position="none") +
  geom_text(stat="bin",
            aes(label = ..count..,y=(..count..-..count..)+90))

#--õnnelikkus
ggplot(a, aes(factor(HAPPY), fill=factor(HAPPY)))  + geom_bar() +
  xlab("õnnelikkus")+
  ylab("inimeste arv")+
  theme(legend.position="none") +
  geom_text(stat="bin",
            aes(label = ..count..,y=(..count..-..count..)+50))

#---organisatsiooni otsustes kaasarääkimine
otsus=andmed$IORGACT

ggplot(a, aes(factor(otsus), fill=factor(otsus)))  +
  geom_bar() +
  xlab("organisatsiooni otsustes kaasarääkimine")+
  ylab("inimeste arv")+
  theme(legend.position="none") +
  geom_text(stat="bin",
            aes(label = ..count..,y=(..count..-..count..)+90))


#--üldkogumi keskmised
mean(a$HAPPY) #6.902655
mean(a$sissetulek) #3359.7449
mean(a$tel) #0.9380531
mean(a$IORGACT) #3.553097

min(a$sissetulek) #50.94863
max(a$sissetulek) # 1538.42

#Jätame andmestikku alles vaid meile huvipakkuvad tunnused.
#Lisaks leibkonna suurus, mida kasutame
#tausttunnusega PPS korral.

UK1=data.frame(liikmed=andmed$HHMMB,tel, sissetulek,
               onnelikkus=andmed$HAPPY, otsus=andmed$IORGACT,
               valikutn = (andmed$HHMMB)/sum(andmed$HHMMB))
valikutn = (andmed$HHMMB)/sum(andmed$HHMMB)

#---valimi võtmine
m=1000 #valimite arv
n=542 #valimi suurus, 30% 1808-st
valimi_jrk=matrix(NA, n, m) #igas veerus on valimi jrk numbrid

for (i in 1:m){
  valimi_jrk[,i]=ppswr(UK1$liikmed,n)
}
```

```
#Kasutame alternatiivset hinnangut, kus N asemel kasutame N hinnangut.

kesk_hinnang_sissetulek=matrix(NA,m,1)
disp_hinnang_sissetulek=matrix(NA,m,1)
kesk_hinnang_tel=matrix(NA,m,1)
disp_hinnang_tel=matrix(NA,m,1)
kesk_hinnang_onnelikkus=matrix(NA,m,1)
disp_hinnang_onnelikkus=matrix(NA,m,1)
kesk_hinnang_otsus=matrix(NA,m,1)
disp_hinnang_otsus=matrix(NA,m,1)
t_hinnang_tel=matrix(NA,m,1)
t_hinnang_onnelikkus=matrix(NA,m,1)
t_hinnang_sissetulek=matrix(NA,m,1)
t_hinnang_otsus=matrix(NA,m,1)
for (i in 1:m){
  kesk_hinnang_sissetulek[i,] =
    (sum(UK1$sissetulek[valimi_jrk[,i]]/n/valikutn[valimi_jrk[,i]]))/(sum((1/n
       /valikutn[valimi_jrk[,i]])))
  disp_hinnang_sissetulek[i,] = 1/((sum(1/n/valikutn[valimi_jrk[,i]]))^2*n*(n
      -1))*sum((sissetulek[valimi_jrk[,i]]-kesk_hinnang_sissetulek[i,])^2/(
      valikutn[valimi_jrk[,i]])^2)
  kesk_hinnang_tel[i,] = sum(UK1$tel[valimi_jrk[,i]] / n/ valikutn[valimi_jrk
      [,i]])/ sum((1/n/valikutn[valimi_jrk[,i]]))
  kesk_hinnang_onnelikkus[i,] = sum(UK1$onnelikkus[valimi_jrk[,i]] / n/
      valikutn[valimi_jrk[,i]])/ sum((1/n/valikutn[valimi_jrk[,i]]))
  t_hinnang_tel[i,] =sum(tel[valimi_jrk[,i]]/n/valikutn[valimi_jrk[,i]])
  t_hinnang_onnelikkus[i,] =sum(UK1$onnelikkus[valimi_jrk[,i]]/n/valikutn[
      valimi_jrk[,i]])
  disp_hinnang_tel[i,] = 1/((sum(1/n/valikutn[valimi_jrk[,i]]))^2*n*(n-1))*sum
      ((((tel[valimi_jrk[,i]]-kesk_hinnang_tel[i,])/valikutn[valimi_jrk[,i]]))
      ^2)
  disp_hinnang_onnelikkus[i,] = 1/((sum(1/n/valikutn[valimi_jrk[,i]]))^2*n*(n
      -1))*sum((((UK1$onnelikkus[valimi_jrk[,i]]-kesk_hinnang_onnelikkus[i])/
      valikutn[valimi_jrk[,i]]))^2)
  kesk_hinnang_otsus[i,] = sum(UK1$otsus[valimi_jrk[,i]] / n/ valikutn[valimi_
      jrk[,i]])/ sum((1/n/valikutn[valimi_jrk[,i]]))
  t_hinnang_otsus[i,] =sum(UK1$otsus[valimi_jrk[,i]]/n/valikutn[valimi_jrk[,i
      ]])
  disp_hinnang_otsus[i,] = 1/((sum(1/n/valikutn[valimi_jrk[,i]]))^2*n*(n-1))*
      sum((((UK1$otsus[valimi_jrk[,i]]-kesk_hinnang_otsus[i])/valikutn[valimi_
      jrk[,i]]))^2)
}

mean(disp_hinnang_sissetulek)

#####----Keskmise sissetuleku ja dispersiooni hinnang
mean(kesk_hinnang_sissetulek) # 359.4378

UI_al_sissetulek=kesk_hinnang_sissetulek-1.96*sqrt(disp_hinnang_sissetulek)
UI_yl_sissetulek=kesk_hinnang_sissetulek+1.96*sqrt(disp_hinnang_sissetulek)
```

```
mean(UI_al_sissetulek) # 338.1251
mean(UI_yl_sissetulek) # 380.7505
mean(UI_yl_sissetulek) - mean(UI_al_sissetulek) # 42.62538
sort(UI_al_sissetulek)[975]-sort(UI_al_sissetulek)[25]
# 38.68926
sort(UI_yl_sissetulek)[975] -sort(UI_yl_sissetulek)[25]
# 44.91179

#--keskmise sissetuleku variatsioonirea 25. ja 975. element.

sort(kesk_hinnang_sissetulek)[25] # 338.3391
sort(kesk_hinnang_sissetulek)[975] # 380.0168
sort(kesk_hinnang_sissetulek)[975] - sort(kesk_hinnang_sissetulek)[25]
# 41.67767

####----Keskmine teleka vaatamise hinnang, 1000 keskmise
#usalduspiirid ja nende variatsioonirea 25. ja 975. element

UI_al_tel=kesk_hinnang_tel-1.96*sqrt(disp_hinnang_tel)
UI_yl_tel=kesk_hinnang_tel+1.96*sqrt(disp_hinnang_tel)
mean(UI_al_tel) # 0.9145103
mean(UI_yl_tel) # 0.961487
mean(UI_yl_tel)-mean(UI_al_tel) # 0.04697671
sort(UI_al_tel)[975] - sort(UI_al_tel)[25] # 0.0572379
sort(UI_yl_tel)[975] - sort(UI_yl_tel)[25] # 0.03611794

#--telekavaatamise osakaalu variatsioonirea 25. ja 975. element.

mean(kesk_hinnang_tel) # 0.9379987 ehk 93,8 % vaatavad telekat
sort(kesk_hinnang_tel)[25] # 0.912833
sort(kesk_hinnang_tel)[975] # 0.9593978
sort(kesk_hinnang_tel)[975] - sort(kesk_hinnang_tel)[25]
#  0.04656483

####----Keskmine Õnnelikkuse hinnang, tuhat usalduspiiri
#ja nende variatsioonirea 25. ja 975. element

mean(kesk_hinnang_onnelikkus) # 6.89934

UI_al_onnelikkus_kesk=kesk_hinnang_onnelikkus-1.96*sqrt(disp_hinnang_
    onnelikkus)
UI_yl_onnelikkus_kesk=kesk_hinnang_onnelikkus+1.96*sqrt(disp_hinnang_
    onnelikkus)
mean(UI_al_onnelikkus_kesk) # 6.696704
mean(UI_yl_onnelikkus_kesk) # 7.101975
mean(UI_yl_onnelikkus_kesk) - mean(UI_al_onnelikkus_kesk)
# 0.4052712

sort(UI_al_onnelikkus_kesk)[975] - sort(UI_al_onnelikkus_kesk)[25]
# 0.4191792
sort(UI_yl_onnelikkus_kesk)[975] - sort(UI_yl_onnelikkus_kesk)[25]
```

```
# 0.374115

#--keskmise õnnelikkuse hinnangu variatsioonirea 25. ja 975. element.
mean(kesk_hinnang_onnelikkus) # 6.89934
sort(kesk_hinnang_onnelikkus)[25] # 6.699691
sort(kesk_hinnang_onnelikkus)[975] # 7.088938
sort(kesk_hinnang_onnelikkus)[975] - sort(kesk_hinnang_onnelikkus)[25]
# 0.3892472

#-Keskmine hinnang organisatsiooni otsustes kaasarääkimisele ja 1000 UI

mean(kesk_hinnang_otsus) # 3.555154

UI_al_otsus=kesk_hinnang_otsus-1.96*sqrt(disp_hinnang_otsus)
UI_yl_otsus=kesk_hinnang_otsus+1.96*sqrt(disp_hinnang_otsus)
mean(UI_al_otsus) # 3.228501
mean(UI_yl_otsus) # 3.881807
mean(UI_yl_otsus) - mean(UI_al_otsus) #0.6533063

sort(UI_al_otsus)[975] - sort(UI_al_otsus)[25] # 0.6338368
sort(UI_yl_otsus)[975] - sort(UI_yl_otsus)[25] # 0.6713756


#--keskmise otsus variatsioonirea 25. ja 975. element.

sort(kesk_hinnang_otsus)[25] # 3.230001
sort(kesk_hinnang_otsus)[975] # 3.88907
sort(kesk_hinnang_otsus)[975] - sort(kesk_hinnang_otsus)[25]
# 0.6590685


############----------Empiirilise tõepära kasutamine

###--punkthinnang keskmisele ÜK-s
#Hajeki hinnang sum(y_i/np_i)/sum(1/np_i)

onnelikkus_EL=matrix(NA,m,1)
sissetulek_EL=matrix(NA,m,1)
tel_EL=matrix(NA,m,1)
otsus_EL=matrix(NA,m,1)

for (i in 1:m){
  onnelikkus_EL[i,]=sum(UK1$onnelikkus[valimi_jrk[,i]]/(n*valikutn[valimi_jrk
      [,i]]))/sum(1/(n*valikutn[valimi_jrk[,i]]))
  sissetulek_EL[i,]=sum(UK1$sissetulek[valimi_jrk[,i]]/(n*valikutn[valimi_jrk
      [,i]]))/sum(1/(n*valikutn[valimi_jrk[,i]]))
  tel_EL[i,]=sum(UK1$tel[valimi_jrk[,i]]/(n*valikutn[valimi_jrk[,i]]))/sum(1/(
      n*valikutn[valimi_jrk[,i]]))
  otsus_EL[i,]=sum(UK1$otsus[valimi_jrk[,i]]/(n*valikutn[valimi_jrk[,i]]))/sum
      (1/(n*valikutn[valimi_jrk[,i]]))
}
```

```
mean(onnelikkus_EL) # 6.89934
mean(sissetulek_EL) # 359.4378
mean(tel_EL) # 0.9379987
mean(otsus_EL) # 3.555154

#lambda leidmine

lag=function(valim,kesk,p){
  L=1/max((valim-kesk)/(n*p))
  R=1/min((valim-kesk)/(n*p))
  dif=1
  tol=1e-07
  while(dif>tol){
    M=(L+R)/2
    glag=sum((valim-kesk)/((n*p)-(M*(valim-kesk))))
    if(glag>0) L=M
    if(glag<0) R=M
    dif=abs(glag)
  }
  return(M)
}

####----Keskmisele sissetulekule usaldusvahemike leidmine;
#----ülemise ülsaduspiiri leidmine

tol=1e-08
cut=qchisq(0.95,1) #hii-ruut jaotuse 0.05-kvantiil, kui vabadusastmete arv on
    1
el_R2=matrix(NA,m,1)
YL_sissetulek=matrix(NA,m,1)

for(i in 1:m){
  t1=sissetulek_EL[i]
  t2=max(UK1$sissetulek[valimi_jrk[,i]])
  vahe=t2-t1
  while(vahe>tol){
    tau=(t1+t2)/2
    M=lag(UK1$sissetulek[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
    el_R2[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-
                           M*((UK1$sissetulek[valimi_jrk[,i]])-tau))/(n*(
                               valikutn[valimi_jrk[,i]]))))))
    if(el_R2[i]>cut) t2=tau
    if(el_R2[i]<=cut) t1=tau
    vahe=t2-t1
  }

  YL_sissetulek[i]=(t1+t2)/2
}
```

```
mean(YL_sissetulek) # 381.7619
sort(YL_sissetulek)[25] #358.6503
sort(YL_sissetulek)[975] #404.4274
sort(YL_sissetulek)[975] - sort(YL_sissetulek)[25] #45.77709

#--alumise usalduspiiri leidmine

el_R2=matrix(NA,m,1)
AL_sissetulek=matrix(NA,m,1)

for(i in 1:m){
  t1=sissetulek_EL[i]
  t2=min(UK1$sissetulek[valimi_jrk[,i]])
  vahe=t1-t2
  while(vahe>tol){
    tau=(t1+t2)/2
    M=lag(UK1$sissetulek[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
    el_R2[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$sissetulek[
        valimi_jrk[,i]])-tau))/(n*(valikutn[valimi_jrk[,i]])))))
    if(el_R2[i]>cut) t2=tau
    if(el_R2[i]<=cut) t1=tau
    vahe=t1-t2
  }
    AL_sissetulek[i]=(t1+t2)/2
}

mean(AL_sissetulek) # 338.8377
mean(YL_sissetulek) - mean(AL_sissetulek) #42,9242
sort(AL_sissetulek)[25] # 319.4732
sort(AL_sissetulek)[975] # 358.2706
sort(AL_sissetulek)[975]-sort(AL_sissetulek)[25] # 38.79739

####----Televiisori vaatamise osakaalule usaldusvahemike leidmine;
#----ülemise ülsaduspiiri leidmine

el_R2=matrix(NA,m,1)
YL_tel=matrix(NA,m,1)

for(i in 1:m){
  t1=tel_EL[i]
  t2=max(UK1$tel[valimi_jrk[,i]])
  vahe=t2-t1
  while(vahe>tol){
    tau=(t1+t2)/2
    M=lag(UK1$tel[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
    el_R2[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$tel[valimi_jrk[,
        i]])-tau))/(n*(valikutn[valimi_jrk[,i]])))))
    if(el_R2[i]>cut) t2=tau
    if(el_R2[i]<=cut) t1=tau
    vahe=t2-t1
  }
```

```
  YL_tel[i]=(t1+t2)/2
}

mean(tel_EL) # 0.9379987
mean(YL_tel) # 0.958154
sort(YL_tel)[25]  # 0.9379249
sort(YL_tel)[975] # 0.9744693
sort(YL_tel)[975] - sort(YL_tel)[25] # 0.03654444

#--alumise usalduspiiri leidmine

el_R2=matrix(NA,m,1)
AL_tel=matrix(NA,m,1)

for(i in 1:m){
  t1=tel_EL[i]
  t2=min(UK1$tel[valimi_jrk[,i]])
  vahe=t1-t2
  while(vahe>tol){
    tau=(t1+t2)/2
    M=lag(UK1$tel[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
    el_R2[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$tel[valimi_jrk[,
        i]])-tau))/(n*(valikutn[valimi_jrk[,i]])))))
    if(el_R2[i]>cut) t2=tau
    if(el_R2[i]<=cut) t1=tau
    vahe=t1-t2
  }

  AL_tel[i]=(t1+t2)/2
}

mean(AL_tel) # 0.9106549
sort(AL_tel)[25] # 0.8801972
sort(AL_tel)[975]# 0.9381789
sort(AL_tel)[975] - sort(AL_tel)[25] # 0.0579817
mean(YL_tel) - mean(AL_tel) #0.04749902

#----Keskmisele õnnelikkusele usaldusvahemike leidmine
#----ülemise ülsaduspiiri leidmine
el_R2=matrix(NA,m,1)
YL_onnelikkus=matrix(NA,m,1)

for(i in 1:m){
t1=onnelikkus_EL[i]
t2=max(UK1$onnelikkus[valimi_jrk[,i]])
vahe=t2-t1
while(vahe>tol){
  tau=(t1+t2)/2
  M=lag(UK1$onnelikkus[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
  el_R2[i]=2*sum((log(((n*valikutn[valimi_jrk[,i]])-M*((UK1$onnelikkus[valimi_
      jrk[,i]])-tau))/(n*(valikutn[valimi_jrk[,i]])))))
```

```
  if(el_R2[i]>cut) t2=tau
  if(el_R2[i]<=cut) t1=tau
  vahe=t2-t1
}
YL_onnelikkus[i]=(t1+t2)/2
}

mean(onnelikkus_EL) # 6.89934
mean(YL_onnelikkus) # 7.095968
sort(YL_onnelikkus)[25] # 6.900095
sort(YL_onnelikkus)[975] #7.277359
sort(YL_onnelikkus)[975] - sort(YL_onnelikkus)[25]
# 0.3772648

#--alumise usalduspiiri leidmine

el_R2=matrix(NA,m,1)
AL_onnelikkus=matrix(NA,m,1)

for(i in 1:m){
t1=onnelikkus_EL[i]
t2=min(UK1$onnelikkus[valimi_jrk[,i]])
vahe=t1-t2
while(vahe>tol){
  tau=(t1+t2)/2
  M=lag(UK1$onnelikkus[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
  el_R2[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$onnelikkus[valimi_
      jrk[,i]])-tau))/(n*(valikutn[valimi_jrk[,i]])))))
  if(el_R2[i]>cut) t2=tau
  if(el_R2[i]<=cut) t1=tau
  vahe=t1-t2
}
AL_onnelikkus[i]=(t1+t2)/2
}

mean(AL_onnelikkus) #  6.688159
sort(AL_onnelikkus)[25] # 6.475397
sort(AL_onnelikkus)[975] # 6.899356
sort(AL_onnelikkus)[975] - sort(AL_onnelikkus)[25] # 0.4239584
mean(YL_onnelikkus) - mean(AL_onnelikkus)  # 0.4078092


#----Keskmisele organisatsiooni otsustes kaasarääkimisele usaldusvahemike
    leidmine;
#----ülemise ülsaduspiiri leidmine

el_R2=matrix(NA,m,1)
YL_otsus=matrix(NA,m,1)

for(i in 1:m){
  t1=otsus_EL[i]
```

```
  t2=max(UK1$otsus[valimi_jrk[,i]])
  vahe=t2-t1
  while(vahe>tol){
    tau=(t1+t2)/2
    M=lag(UK1$otsus[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
    el_R2[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$otsus[valimi_jrk
        [,i]])-tau))/(n*(valikutn[valimi_jrk[,i]])))))
    if(el_R2[i]>cut) t2=tau
    if(el_R2[i]<=cut) t1=tau
    vahe=t2-t1
  }

  YL_otsus[i]=(t1+t2)/2
}


mean(YL_otsus) #3.888143
sort(YL_otsus)[25] # 3.556788
sort(YL_otsus)[975] # 4.226897
sort(YL_otsus)[975] - sort(YL_otsus)[25] # 0.6701087

#--alumise usalduspiiri leidmine

el_R2=matrix(NA,m,1)
AL_otsus=matrix(NA,m,1)

for(i in 1:m){
  t1=otsus_EL[i]
  t2=min(UK1$otsus[valimi_jrk[,i]])
  vahe=t1-t2
  while(vahe>tol){
    tau=(t1+t2)/2
    M=lag(UK1$otsus[valimi_jrk[,i]],tau, valikutn[valimi_jrk[,i]])
    el_R2[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$otsus[valimi_jrk
        [,i]])-tau))/(n*(valikutn[valimi_jrk[,i]])))))
    if(el_R2[i]>cut) t2=tau
    if(el_R2[i]<=cut) t1=tau
    vahe=t1-t2
  }
  AL_otsus[i]=(t1+t2)/2
}

mean(AL_otsus) #   3.235211
sort(AL_otsus)[25] # 2.923864
sort(AL_otsus)[975] # 3.558416
sort(AL_otsus)[975] - sort(AL_otsus)[25] # 0.6345519
mean(YL_otsus) -mean(AL_otsus) # 0.6529318

########################################
####---Joonised, kas on hii-ruutjaaotusega.
#-õnnelikkus
```

```
onnelikkus_kesk=mean(onnelikkus_EL)
r_onnelikkus=matrix(NA,m,1)

for (i in 1:m){
  M=lag(UK1$onnelikkus[valimi_jrk[,i]],onnelikkus_kesk, valikutn[valimi_jrk[,i
      ]])
  r_onnelikkus[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*(UK1$onnelikkus[
      valimi_jrk[,i]]-onnelikkus_kesk))/
                              n/(valikutn[valimi_jrk[,i]])))))
}

#-sissetulek
sissetulek_kesk=mean(sissetulek_EL)
r_sissetulek=matrix(NA,m,1)

for (i in 1:m){
  M=lag(UK1$sissetulek[valimi_jrk[,i]],sissetulek_kesk, valikutn[valimi_jrk[,i
      ]])
  r_sissetulek[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$sissetulek[
      valimi_jrk[,i]])-sissetulek_kesk))/(n*(valikutn[valimi_jrk[,i]])))))
}

#-televiisori vaatamine
telekas_kesk=mean(tel_EL)
r_televiisor=matrix(NA,m,1)
for (i in 1:m){
  M=lag(UK1$tel[valimi_jrk[,i]],telekas_kesk, valikutn[valimi_jrk[,i]])
  r_televiisor[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$tel[valimi_
      jrk[,i]])-telekas_kesk))/(n*(valikutn[valimi_jrk[,i]])))))
}

#-organisatsiooni otsustes kaasarääkimine
otsus_kesk=mean(otsus_EL)
r_otsus=matrix(NA,m,1)
for (i in 1:m){
  M=lag(UK1$otsus[valimi_jrk[,i]], otsus_kesk, valikutn[valimi_jrk[,i]])
  r_otsus[i]=2*sum((log((n*(valikutn[valimi_jrk[,i]])-M*((UK1$otsus[valimi_jrk
      [,i]])-otsus_kesk))/(n*(valikutn[valimi_jrk[,i]])))))
}

op=par(mfrow=c(2,2))
hist(r_sissetulek, freq=FALSE, main="Sissetulek", ylab="tihedus", xlab=" " )
x=rchisq(1000,1)
curve( dchisq(x, df=1), col='orange', add=TRUE)
hist(r_televiisor, freq=FALSE, main="Televiisori vaatamine", ylab="tihedus",
    xlab=" " )
x=rchisq(1000,1)
curve( dchisq(x, df=1), col='orange', add=TRUE)
hist(r_onnelikkus, freq=FALSE, main="Õnnelikkus", ylab="tihedus", xlab=" " )
x=rchisq(1000,1)
curve( dchisq(x, df=1), col='orange', add=TRUE)
```

```
hist(r_otsus, freq=FALSE, main="Otsustes kaasarääkimine", ylab="tihedus", xlab
    =" " )
x=rchisq(1000,1)
curve( dchisq(x, df=1), col='orange', add=TRUE)
par(op)
```

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina, Kristi Tüli,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

"Empiirilise tõepära meetod valikuuringutes",

mille juhendajad on Imbi Traat ja Natalja Lepik.

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

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

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

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

Tartus, 12.05.2016