Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Laura Himma

# Kas kohvi joomine pikendab eluiga? Põhjuslike mõjude analüüs TÜ Eesti geenivaramu andmete põhjal

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Krista Fischer

Tartu 2025

**KAS KOHVI JOOMINE PIKENDAB ELUIGA? PÕHJUSLIKE MÕJUDE ANALÜÜS TÜ EESTI GEENIVARAMU ANDMETE PÕHJAL**

Bakalaureusetöö

Laura Himma

**Lühikokkuvõte**

Kohv on maailmas üks enamlevinumaid jooke. Lisaks kohvi ergutavale toimele on palju uuritud ka selle seost inimeste tervisele, sealhulgas suremusele. Mitmed varasemad uuringud on aga leidnud, et kohvi joovate inimeste hulgas on suremus madalam kui nende seas, kes kohvi ei joo.

Käesoleva bakalaureusetöö eesmärk on esmalt tutvuda põhjusliku mõju hindamise metoodikaga ning seejärel püüda hinnata kohvi joomise põhjuslikku mõju suremusele. Analüüs tugineb Eesti geenivaramu andmetel, mis hõlmab aastatel 2002 kuni 2013 liitunud $50-69$-aastaseid inimesi.

Kõigepealt antakse ülevaade põhjusliku mõju mõistest ning selle hindamiseks kasutatavatest meetoditest nagu pöördtõenäosuse kaalumine ja standardiseerimine. Käsitletakse nii mitteparameetrilisi kui ka mudelipõhiseid lähenemisi, millest viimased põhinevad logistilisel regressioonmudelil.

Töö praktilises osas rakendatakse nimetatud meetodeid geenivaramu andmestikul, et hinnata kohvi tarbimise mõju suremusele, arvestades seejuures mitmesuguseid segavaid tegureid nagu sugu, vanus, haridus ja erinevaid tervisega seotud näitajaid.

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

**Märksõnad:** Põhjuslik mõju, suremuse analüüs.

**DOES COFFEE CONSUMPTION INCREASE LONGEVITY? APPLICATION OF CAUSAL INFERENCE METHODOLOGY USING DATA FROM THE ESTONIAN BIOBANK**

Bachelor thesis

Laura Himma

**Abstract**

Coffee is one of the most widely consumed beverages in the world. A lot of research has been done on its effects on human health, including mortality. Several previous studies have found that mortality tends to be lower among coffee drinkers than among those who do not drink coffee.

The aim of this bachelor's thesis is to explore the methodology of causal effect estimation and then to assess the causal effect of coffee consumption on mortality. The analysis is based on the data from the Estonian Biobank, which includes individuals aged 50 to 69 who joined the biobank between 2002 and 2013.

Firstly, an overview is given on the concept of causal effect and the methods used to assess it, such as inverse probability weighting and standardization. Both non-parametric and model-based approaches are considered, with the model-based methods implemented through logistic regression.

In the practical part of this thesis, these methods are applied to the data of the Estonian Biobank to estimate the effect of coffee consumption on mortality, accounting for various confounding factors such as gender, age, education level, and a range of health-related indicators.

**CERCS research specialisation:** P160 Statistics, operations research, programming, financial and actuarial mathematics.

**Key Words:** Causal inference, mortality analysis.

# Sisukord

# Sissejuhatus

Kohvi joomine kuulub suure osa inimeste igapäevaharjumuste hulka. Seda juuakse nii kuumalt kui ka külmalt ning lai valik erinevaid kohvijooke võimaldab igal tarbijal leida oma maitse-eelistustele sobiva joogi. Kohvi populaarsuse tõttu on palju uuritud ka selle joogi mõju inimeste tervisele ja suremusele. Üks sellistest töödest on 2017. aastal avaldatud kohortuuring, mis analüüsis 521 330 mehe ja naise andmeid kümnest Euroopa riigist, uurides nende kohvitarbimise harjumuste seost suremusega. Uuring võttis arvesse riikidevahelisi erinevusi tarbimisharjumustes ning jaotas osalejad kohvitarbimise järgi nelja rühma. Tulemused näitasid, et kõige rohkem kohvi joonud inimeste üldsuremus oli madalam võrreldes nendega, kes kohvi üldse ei joonud. Uuring kinnitas ka varasemate tööde tulemusi, mille kohaselt on kohvijoojatel väiksem suremusrisk kui mittejoojatel. (Gunter *et al.*, 2017)

Käesoleva töö raames uurime kohvi tarbimise põhjuslikku mõju suremusele kasutades Tartu Ülikooli Eesti geenivaramu andmestikku, kus on andmed 14 999 inimese kohta vanuses 50 kuni 69 aastat, kes liitusid geenivaramuga vahemikus 2002 kuni 2013. Andmed oleme saanud väljastuse 6-7/GI/11519 raames (väljastusluba U31), samuti oleme saanud Eesti bioeetika inimuuringute kooskõlastuse (20.09.2023, nr 1.1-12/3455). Tegemist on vaatlusandmetega, mis sisaldavad teavet nii soo, vanuse kui ka diagnoositud haiguste ja tervisekäitumise kohta. Nende põhjal soovime hinnata kohvi tarbimise põhjuslikku mõju inimeste 10-aastasele suremusele alates geenivaramuga liitumise hetkest.

Töö esimeses osa anname teoreetilise ülevaate põhjusliku mõju defineerimise viisidest ja lihtsamatest selle hindamiseks kasutatavatest meetoditest. Töös käsitletud meetodid on pöördtõenäosuse kaalumine ja standardiseerimine. Töö teises osas kirjeldame praktiliselt läbiviidud põhjusliku mõju hindamist geenivaramu andmestikul. Analüüsimiseks oleme kasutanud statistikatarkvara R ning töös kasutatud R-i koodid pärinevad Hernán'i ja Robins'i raamatust „Causal inference: What IF".

# 1 Põhjuslik mõju ja selle hindamise meetodid

Järgnevalt anname ülevaate põhjusliku mõju mõistest ning selle hindamiseks kasutatavatest meetoditest. Ülevaade on koostatud Hernán'i ja Robins'i raamatu „Causal inference: What IF" põhjal, kui ei ole viidatud teisiti (Hernán ja Robins, 2020).

## 1.1 Põhjuslik mõju

Vaatleme olukorda, kus eesmärgiks on hinnata tunnuse A põhjuslikku mõju tunnusele Y. Käesolevas arutelus nimetame tunnust $A$ kui „ravi" ning tunnust $Y$ kui „suremus", kuid reaalsuses võivad $A$ ja $Y$ tähistada ka mõnd teistsugust mõjutegurit ja lõpptulemust.

Olgu defineeritud juhuslikud suurused $A$ ja $Y$, kus $A$ on kahe tasemega suurus, mis näitab ravi saamist ($A = 1$ tähendab, et saadi ravi, $A = 0$, et ravi ei saadud) ning binaarne juhuslik suurus $Y$ näitab lõpptulemust, kas inimene suri ($Y = 1$) või mitte ($Y = 0$). Juhuslik suurus $Y^{a=1}$ näitab lõpptulemust (surma või ellu jäämist) juhul, kui isik oleks saanud ravi ning $Y^{a=0}$ lõpptulemust juhul, kui isik ravi saanud ei oleks.

$Y^{a=0}$ ja $Y^{a=1}$ nimetatakse kontrafaktuaalseteks ehk potentsiaalseteks tunnusteks. Tegelikkuses on kontrafaktuaalsetest tunnustest võimalik vaadelda vaid ühte, näiteks, kui vaadeldav isik saab ravi ($A = 1$), siis tema puhul on võimalik vaadelda ainult tunnust $Y^{a=1}$.

**Definitsioon 1.** Öeldakse, et ravil $A$ on põhjuslik mõju indiviidi lõpptulemusele $Y$, kui antud isiku korral kehtib $Y^{a=1} \neq Y^{a=0}$.

Seega eksisteerib põhjuslik mõju, kui ravi saamise ning ravi mittesaamise korral on isiku lõpptulemus (näiteks, kas inimene suri või jäi ellu) erinev.

Vaadeldavas üldkogumis esineb keskmine põhjuslik mõju lõpptulemuse $Y$ ja ravi $A$ vahel, kui $P(Y^{a=1} = 1) \neq P(Y^{a=0} = 1)$ või $E(Y^{a=1}) \neq E(Y^{a=0})$.

Põhjuslikule mõjule vastavat parameetrit saab binaarse $A$ ja $Y$ korral defineerida kolmel moel. Esimene võimalik parameeter on põhjuslik riskide vahe (ingl *causal risk difference*)

$$RD_c = P(Y^{a=1} = 1) - P(Y^{a=0} = 1). \tag{1}$$

Põhjuslik riskide vahe jääb alati $-1$ ja 1 vahele ning kui see võrdub nulliga, siis üldkogumis põhjuslik mõju puudub. Teiseks võimalikuks põhjuslikuks parameetriks on riskisuhe (ingl *risk ratio*):

$$RR_c = \frac{P(Y^{a=1} = 1)}{P(Y^{a=0} = 1)}. \tag{2}$$

Valimis puudub põhjuslik mõju, kui antud jagatis võrdub ühega. Kolmas parameeter on šansside suhe (ingl *odds ratio*) kujul

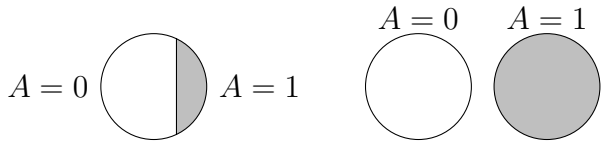
$$OR_c = \frac{P(Y^{a=1} = 1)/P(Y^{a=1} = 0)}{P(Y^{a=0} = 1)/P(Y^{a=0} = 0)}. \tag{3}$$

Šansside suhte korral saame samuti öelda, et põhjuslikku mõju ei leidu, kui avaldise väärtus võrdub ühega.

## 1.2 Põhjusliku mõju hinnatavus lihtsatel meetoditel

Olgu huvipakkuv üldkogum selline, kus mingi osa on ravitud ($A = 1$) ja ülejäänud on ravita jäänud isikud. Ravil $A$ on põhjuslik mõju lõpptulemusele $Y$, kui $E(Y^{a=1}) \neq E(Y^{a=0})$ ehk esineb erinevus keskmises tulemuses kui kogu vaadeldav üldkogum oleks ravitud, võrreldes sellega, kui keegi poleks ravitud. Tavapärase statistilise seose hindamisel võrdleme omavahel aga huvipakkuva üldkogumi kahte osa, see tähendab, võrdleme riski üldkogumis olevate ravi saanud ning ravi mittesaanud isikute vahel. Seega matemaatiliselt kirjeldatuna on ravil ja lõpptulemusel

seos, kui $E(Y|A=1) \neq E(Y|A=0)$. Üldjuhul on statistiline ja põhjuslik seos erinevad, nende erisus on välja toodud joonisel 1.




Statistiline seos $E(Y|A=1) - E(Y|A=0)$

Põhjuslik seos $E(Y^{a=1}) - E(Y^{a=0})$

Joonis 1: Statistilise ja põhjusliku seose erinevus.

Nagu eelnevalt mainisime, ei ole tegelikkuses mõlemad indiviidi kontrafaktuaalsed tunnused vaadeldud. Samas on vaja põhjusliku mõju hindamiseks (kasutades riskide vahet, suhtelist riski või šansside suhet) mõlemat kontrafaktuaalset väärtust. Marginaalselt randomiseeritud katse korral valitakse ravigruppi kuuluvad inimesed juhuslikult. Seega eeldame, et ravigrupis olevate isikute suremus on võrdne suremusega juhul, kui need, kes algselt ei saanud ravi, nüüd ravi saaksid. Lihtsamalt öeldes vahetame ravi saajad ning mittesaajad omavahel ära. Seda nimetatakse vahetatavuseks.

**Definitsioon 2.** Öeldakse, et binaarsete juhuslike suuruste $Y^a$ ja $A$ korral kehtib vahetatavus, kui iga $a \in (0,1)$ korral on täidetud tingimus:

$$P(Y^a = 1|A = 1) = P(Y^a = 1|A = 0) = P(Y^a = 1). \tag{4}$$

Vahetatavuse korral on $A$ ja $Y^a$ omavahel sõltumatud. Randomiseeritud katse korral on vahetatavus tagatud ning võrdused (4) kehtivad. Seega, kui katse on randomiseeritud ja seal kehtib vahetatavus, siis saame hinnata põhjuslikku mõju põhjusliku riskide vahena kasutades valemit

$$P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = P(Y = 1|A = 1) - P(Y = 1|A = 0).$$

Samuti saame riskisuhet esitada järgmiselt:

$$\frac{P(Y^{a=1}=1)}{P(Y^{a=0}=0)} = \frac{P(Y=1|A=1)}{P(Y=1|A=0)}$$

ning šansside suhet kujul

$$\frac{P(Y^{a=1}=1)/P(Y^{a=1}=0)}{P(Y^{a=0}=1)/P(Y^{a=0}=0)} = \frac{P(Y=1|A=1)/P(Y=0|A=1)}{P(Y=1|A=0)/P(Y=0|A=0)}.$$

Olgu peale ravitaseme $A$ ja lõpptulemuse $Y$ vaadeldud ka tunnus $L$, mis on mõõdetud enne ravi määramist. Ravi määramisel on arvesse võetud $L$-i tasemed: erinevatel faktori $L$ tasemetel võib ravi saamise tõenäosus olla erinev ning igal tasemel on ravi määramine randomiseeritud. Seega saame katset vaadelda kui kahest või enamast marginaalselt randomiseeritud katsest koosnevat eksperimenti. Kuna randomiseerimine toimub erinevatel $L$ tasemetel eraldi, siis enamasti tavaline vahetatavus (Definitsioon 2) ei kehti. Samas kehtib vahetatavus igal $L = l$ tasemel eraldi. Seda nimetatakse tinglikuks vahetatavuseks.

**Definitsioon 3.** Öeldakse, et binaarsete juhuslike suuruste $Y^a$ ja $A$ ning diskreetse suuruse $L$ korral kehtib tinglik vahetatavus, kui iga $a \in (0,1)$ korral on täidetud tingimus:

$$P(Y^a = 1|A = 1, L = l) = P(Y^a = 1|A = 0, L = l). \tag{5}$$

Tinglikult randomiseeritud eksperimendi korral võime põhjuslikku mõju hinnata igal tasemel eraldi. See tähendab, et põhjuslikku mõju saab hinnata sellisel juhul põhjusliku riskide vahena igal $L$ tasemel $l$ kujul:

$$P(Y = 1|A = 1,\, L = l) - P(Y = 1|A = 0, L = l).$$

Samuti saame tasemeti hinnata suhtelist riski ja šansside suhet. Kirjeldatud meetodit nimetatakse stratifitseerimiseks (ingl *stratification*).

### 1.2.1 Põhjusliku mõju hindamine standardiseerimise abil

Kui kehtib tinglik vahetatavus (Definitsioon 3), siis saame üldkogumi põhjusliku riskisuhte hindamiseks kasutada standardiseerimist. Standardiseerimisel leitakse soovitud tõenäosus $P(Y^a = 1)$ liites iga taseme $L = l$ korral tõenäosuse $P(Y^a = 1|L = l)$ ning taseme tõenäosuse $P(L = l)$ korrutised. Kuna iga kihi korral on tegemist marginaalselt randomiseeritud katsega ning sellisel juhul kehtib võrdus $P(Y^a = 1|L = l) = P(Y = 1|L = l, A = a)$, saame tõenäosuse $P(Y^a = 1)$ arvutada standardiseerimise teel järgmiselt:

$$\begin{aligned} P(Y^a = 1) &= \sum_l P(Y^a = 1|L = l) \cdot P(L = l) \\ &= \sum_l P(Y = 1|L = l, A = a) \cdot P(L = l). \end{aligned} \tag{6}$$

Põhjuslik riskide vahe on seega hinnatav kujul:

$$\begin{aligned} P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = &\sum_l P(Y = 1|L = l, A = 1) \cdot P(L = l) \\ &- \sum_l P(Y = 1|L = l, A = 0) \cdot P(L = l). \end{aligned}$$

Analoogiliselt saame leida ka põhjusliku riskisuhte ning šansside suhte.

### 1.2.2 Põhjusliku mõju hindamine pöördtõenäosuse kaalude abil

Teine võimalus, matemaatiliselt ekvivalentne standardiseerimisega, on kasutada pöördtõenäosuse kaalumise meetodit (ingl *inverse probability weighting*). Binaarse juhusliku suuruse $A$ ja diskreetse suuruse $L$ korral leitakse tõenäosuse $P(Y^a = 1)$ arvutamiseks kõigepealt indiviidide kaalud. Ravi saajate $(A = 1)$ kaal avaldub kujul:

$$w = \frac{1}{P(A = 1|L = l)}$$

ning ravi mittesaajate kaal on arvutatav järgmiselt:

$$w = \frac{1}{P(A=0|L=l)} = \frac{1}{1-P(A=1|L=l)}.$$

Edasi korrutame ravi saajate lõpptulemused $Y$ läbi vastava kaaluga. Nii tekib pseudo-valim, kus ravi ja taustmuutujate $L$ seos on eemaldatud ning ravi saajate andmed on kaalutud nii, et need esindaksid kogu valimit.

Otsitava tõenäosuse $P(Y^{a=1} = 1)$ leidmiseks liidame kõik kaalutud ravi saajate lõpptulemused $Y = 1$ ning jagame vaadeldava valimi suurusega. Analoogiliselt saab leida ka tõenäosuse $P(Y^{a=0} = 1)$. Meetodit kasutades tekib algsest valimist kaks korda suurem pseudo-valim, kus iga indiviid on arvesse võetud kahel korral: nii ravi saajana kui ka ravi mittesaajana. Kui algses üldkogumis kehtib tinglik vahetatavus $Y \perp A|L$, siis pöördtõenäosuse kaalumise teel saadud valim esindab pseudo-üldkogumit, kus kehtib tavaline ehk mitte tinglik vahetatavus ravitute ja mitteravitute hulgas $Y \perp A$, sest pseudo-üldkogumis on $L$ ja $A$ omavahel sõltumatud.

Näiteks on mingi $L = l$ korral 3-liikmelises valimis 2 isikut, kes said ravi ($A = 1$) ning 1 isik, kes ravi ei saanud ($A = 0$). Tõenäosused $P(A = 1|L = l)$ ja $P(A = 0|L = l)$ tulevad vastavalt $\frac{2}{3}$ ning $\frac{1}{3}$, seega $A = 1$ korral kasutame kaalu $\frac{3}{2}$ ning $A = 0$ korral kaalu 3. Saadud pseudo-valimis on kokku kuus isikut, kolm, kes said ravi ($A = 1$) ning kolm, kes ravi ei saanud ($A = 0$).

## 1.3 Põhjuslikud mudelid

Eelnevates peatükkides kasutasime põhjusliku mõju hindamiseks mitteparameetrilisi ehk mudeliteta meetodeid. Neid saab kasutada olukordades, kus andmed on esitatavad suhteliselt lihtsa sagedustabeli kujul. See on võimalik juhul, kui segajaid ehk $L$ komponente on vähe ning nende seas ei leidu pidevaid tunnuseid. Suures ja paljude tunnustega andmestikus muutub põhjusliku mõju hindamine mitteparа-

meetrilisel viisil keeruliseks. Sellistel juhtudel kasutatakse mõju hindamisel mudeleid.

### 1.3.1 Pöördtõenäosuse kaalumist kasutavad mudelid

Vaatleme ravi $A$ mõju lõpptulemusele $Y$. Tunnus $Y$ võib olla ka pidev arvuline tunnus, mitteparameetriliste meetodite korral vaatlesime $Y$-t kui binaarset tunnust. Olgu $E(Y^{a=1})$ keskmine tunnuse $Y$ väärtus, kui kõik inimesed vaadeldavas üldkogumis oleksid saanud ravi ning $E(Y^{a=0})$ keskmine tunnuse $Y$ väärtus, kui keegi ei oleks saanud ravi. Binaarse $Y$ korral kehtivad $E(Y^{a=1}) = P(Y^{a=1})$ ja $E(Y^{a=0}) = P(Y^{a=0})$. Keskmise põhjusliku mõju saame seega defineerida kui $E(Y^{a=1}) - E(Y^{a=0})$.

Pöördtõenäosuse kaalumise korral peame iga indiviidi põhiselt leidma kaalud. Binaarse tunnuse $A$ korral saab need defineerida järgmiselt:

$$w = \begin{cases} \frac{1}{P(A=1|L)}, \text{ kui } A = 1; \\ \frac{1}{P(A=0|L)} = \frac{1}{1-P(A=1|L)}, \text{ kui } A = 0. \end{cases} \tag{7}$$

Tõenäosuse $P(A = 1|L)$ hindamiseks sobitame logistilise regressioonmudeli, kus $A$ on sõltuv tunnus ning kõik segajad $L$ võtame mudelis arvesse argumenttunnustena. $K$ argumenttunnusega logistilise regressiooni mudel avaldub kujul:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K, \tag{8}$$

kus $\beta_0, \beta_1, \ldots, \beta_K$ on parameetrid ning $X_1$,$X_2, \ldots, X_K$ argumenttunnused (Nahhas, 2025). Logistiline regressioonmudel annab tulemuseks logaritmitud šansside suhte $\ln\left(\frac{p}{1-p}\right)$. Huvipakkuva sündmuse tõenäosuse saab logistilise regressioonmudeli kaudu avaldada järgmiselt (Nahhas, 2025):

$$p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_K X_K}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_K X_K}}.$$

Logistilise regressioonmudeli kaudu saadud tõenäosuste $P(A = 1|L)$ ja $P(A = 0|L) = 1 - P(A = 1|L)$ abil saame leida vajalikud kaalud (7). Järgmiseks hindame uue logistilise regressiooni mudeli, kus ainus argumenttunnus on $A$ ja kasutame eelnevalt leitud kaale (7). Kaalutud logistilise regressioonmudeli hindamisel kasutame üldistatud hinnanguvõrrandite (lühemalt GEE) meetodit (Højsgaard, Halekoh ja Yan, 2005).

Statistikatarkvaras R kasutame logistilise regressioonmudeli hindamiseks funktsiooni *glm()*. Edasi saame kasutades *predict()* funktsiooni leida igale indiviidile kaalude arvutamiseks vajalikud tõenäosused $P(A = 1|L)$ ja $P(A = 0|L)$. Nüüd saame andmestikku lisada uue veeru, kus on igale indiviidile vastavad kaalud (7). Põhjusliku mõju hindamiseks hindame uue logistilise regressioonmudeli kasutades paketis *geepack* olevat funktsiooni *geeglm()*. Mudelis on uuritavaks tunnuseks lõpptulemus $Y$ ning argumenttunnuseks ravi $A$, mudelis võtame arvesse ka eelnevalt leitud kaalud. Hinnatav mudel avaldub kujul:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 A, \tag{9}$$

kus $A$ tähistab ravi ning $p = P(Y = 1|A)$. Kaalude tõttu saame hinnangu parameetrile $\beta_1$ pseudo-üldkogumis, kus kehtib vahetatavus (Definitsioon 2), eeldusel, et $L$ sisaldab kõiki segajaid ehk tunnuseid, mis samaaegselt mõjutavad tunnuseid $A$ ja $Y$. Seega saame parameetrit $\beta_1$ tõlgendada kui põhjuslikku parameetrit. GEE meetodi kasutamine garanteerib, et parameetri standardvea hinnang oleks korrektne ehk see arvestaks tegeliku, mitte pseudo-üldkogumi suurusega. Kasutades saadud standardvea hinnangut, leiame 95%-se usaldusintervalli parameetrile $\beta_1$. Kui usaldusintervall sisaldab nulli, siis saame öelda, et ravil $A$ puudub mõju lõpptulemusele $Y$. Usaldusintervalli piirid leiame valemiga $\beta_1 \pm z_{0,975} \cdot se(\beta_1)$, kus $z$ ja $se(\beta_1)$ on vastavalt standardse normaaljaotuse kriitiline väärtus ja standardviga.

### 1.3.2 Standardiseerimist kasutavad mudelid

Standardiseerimist kasutades saame põhjusliku mõju hindamiseks leida hinnangud suurustele $E(Y^{a=1})$ ja $E(Y^{a=0})$ ning riskide vahe avaldub nende põhjal kujul $E(Y^{a=1}) - E(Y^{a=0})$. Samuti saame leida riskisuhte (2) ja šansside suhte (3). Kõigepealt hindame mudeli, mille põhjal leiame hinnangu suurusele $E(Y|L, A)$. Binaarse $Y$ korral saame hinnata logistilise regressioonmudeli (8), mis võib sisaldada ka $A$ ja $L$ omavahelisi koosmõjusid. Seejärel kasutame saadud mudeli parameetreid, et hinnata suurused $E(Y^{a=1}|L)$ ning $E(Y^{a=0}|L)$. Olgu vastavad hinnangud $\hat{Y}^{a=1}$ ja $\hat{Y}^{a=0}$. Nende suuruste põhjal saamegi põhjusliku parameetri hinnangu leida näiteks kujul $\hat{RD}_C = \hat{E}(\hat{Y}^{a=1}) - \hat{E}(\hat{Y}^{a=0})$.

Statistikatarkvaras R loome standardiseerimise rakendamiseks algsest andmestikust kolm koopiat. Esimene neist on identne päris andmestikuga. Teises ja kolmandas koopias muudame kõik $A$ väärtused vastavalt 0-ks ja 1-ks ning lõpptulemuse $Y$ väärtused puuduvateks. Koondame need kolm andmestikuplokki üheks ning hindame sellel logistilise regressioonmudeli (8), kus $Y$ on sõltuv tunnus ning arvesse on võetud nii $A$ kui ka kõik segajad $L$. Kuna $Y$ tunnused eksisteerivad vaid esimeses andmestikus, kasutame regressioonmudeli hindamiseks vaid esimest koopiat. Mudeli põhjal saame prognoosida teisele ja kolmandale koopiale puuduolevad $Y$ väärtused. Kasutades põhjusliku riski vahet, saame leida põhjusliku mõju kui lahutame kolmanda koopia ennustatud $Y$ väärtuste keskmisest teise koopia $Y$-i keskmise.

Usaldusintervalli leiame samuti kujul $\hat{RD}_c \pm z_{0,975} \cdot se(\hat{RD}_C)$, kus $se(\hat{RD}_C)$ on bootstrap meetodit kasutades leitud standardviga. Bootstrap meetod kujutab endast juhuvalimite võtmist algsest andmestikust ning nende valimite peal keskmise ja standardvea leidmist (Waples, 2024). R-is kasutame bootstrap meetodi kasutamiseks paketi *boot* funktsiooni *boot()*, mis võtab argumentideks algse andmestiku, funktsiooni, mille oleme loonud standardiseerimise läbiviimiseks ning hinnangu leidmiseks, ja täisarvu R, mis määrab, mitu korda juhuvalimit võtta (Canty, Ripley

ja Brazzale, 2024). Funktsioon tagastab keskmise, hinnangu nihke (ingl *bias*) ja standardvea. Kasutades bootstrapi abil leitud keskmist ja standardviga, saame arvutada soovitud usaldusintervalli.

# 2 Põhjusliku mõju hindamine TÜ Eesti geenivaramu andmestikul

Järnevalt rakendame uuritud põhjusliku mõju hindamise meetodeid Tartu Ülikooli Eesti geenivaramu andmestiku peal.

## 2.1 Ülevaade andmestikust

Töö praktilises osas kasutame Tartu Ülikooli Eesti geenivaramu andmestikku, kus on andmed $50-69$-aastaste inimeste kohta, kes liitusid geenivaramuga vahemikus 2002 kuni 2013. Kokku on andmeid 14 999 inimese kohta. Põhjusliku mõju hindamisel mudelitega jätsime välja isikud, kellel esines puuduvaid väärtuseid ning peale väljajätmist jäi andmestikku alles 14 826 isikut.

Analüüsis võtsime lisaks isikuid iseloomustavatele tunnustele ja tervisekäitumistele arvesse ka mitmeid haiguseid, millel on seos kohvi tarbimisega. Sellisteks haigusteks on südamepuudulikkus või südamekahjustusega hüpertooniatõbi (lühemalt $H_{SV}$), seedeelundite haigused nagu gastriit, düspepsia ja maohaavandid (lühemalt $H_{SE}$), kõrgvererõhutõbi (lühemalt $H_{KVR}$), kopsuhaigused (lühemalt $H_K$) ning kõik halvaloomulised kasvajad (lühemalt $H_V$). Lisaks oleme arvesse võtnud, kas inimesel oli liitumise hetkel süstoolne vererõhk üle 150 mm/hg (lühemalt vererõhk). Tabelis 1 on välja toodud iga andmestikus oleva tunnuse jaotus nii kohvijoojate kui ka mittejoojate seas.

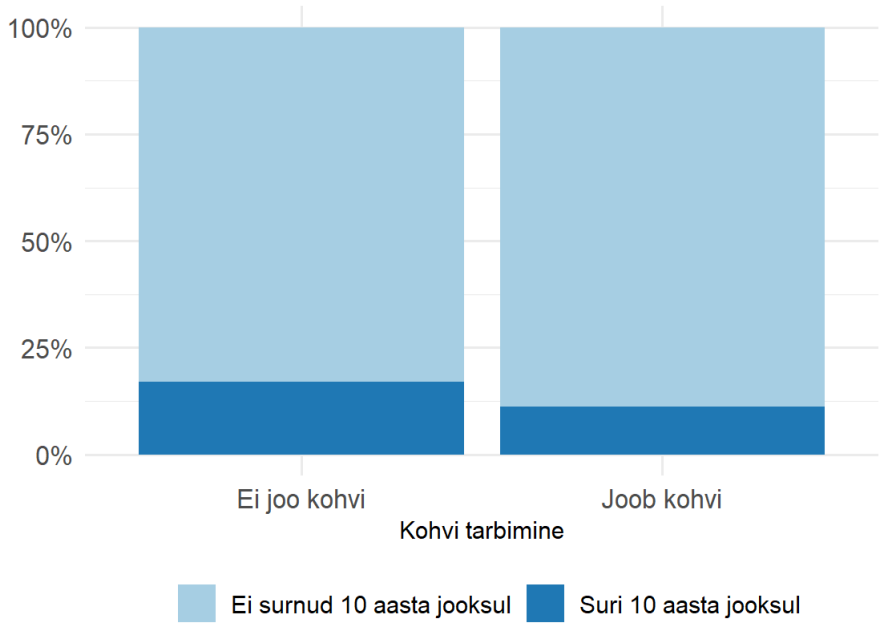
Andmestikus olevatest inimestest oli kohvijoojaid 87,9% ning mittejoojaid 12,1%. Kohvi tarbijatest suri kümne aasta jooksul peale geenivaramuga liitumist 11,2% ning kohvi mittetarbijatest suri sama aja jookul 17,1%.

Tabel 1: Tunnuste ülevaade kohvijoojate ja mittejoojate seas.

| Tunnus | Kohvijoojad (13 036) n (%) | Ei joo kohvi (1790) n (%) |
|---|---|---|
| Keskmine vanus aastates | 58,4 | 59,0 |
| Naiste arv | 9049 (69,4%) | 919 (51,3%) |
| Suitsetamine: | | |
| endine | 2313 (17,7%) | 376 (21,0%) |
| praegune | 3020 (23,2%) | 334 (18,7%) |
| ei suitseta | 7703 (59,1%) | 1080 (60,3%) |
| $H_{SV}$ | 3036 (23,3%) | 598 (33,4%) |
| $H_{SE}$ | 1544 (11,8%) | 267 (14,9%) |
| Rahvuselt eestlane | 10 475 (80,4%) | 1107 (61,8%) |
| $H_{KVR}$ | 5860 (45,0%) | 1025 (57,3%) |
| $H_V$ | 732 (5,6%) | 111 (6,2%) |
| Vererõhk 150 | 2561 (19,6%) | 458 (25,6%) |
| Haridus: | | |
| põhiharidus või madalam | 2250 (17,3%) | 390 (21,8%) |
| keskharidus | 7493 (57,5%) | 988 (55,2%) |
| kõrgharidus | 3293 (25,3%) | 412 (23,0%) |
| $H_K$ | 7447 (57,1%) | 1004 (56,1%) |
| Laste arv: | | |
| lastetu | 2633 (20,2%) | 338 (18,9%) |
| 1 laps | 2275 (17,5%) | 403 (22,5%) |
| 2 last | 5142 (39,4%) | 680 (38,0%) |
| 3 või enam last | 2986 (22,9%) | 369 (20,6%) |
| Saiaviilude arv päevas (lühemalt sai): | | |
| ei sööö | 2535 (19,4%) | 389 (21,7%) |
| 1 viil | 1995 (15,3%) | 244 (13,6%) |
| 2 viilu | 3698 (28,4%) | 476 (26,6%) |
| 3 või enam viilu | 4808 (36,9%) | 681 (38,0%) |
| Trennile kuluv tundide arv nädalas (lühemalt trenn): | | |
| 0 tundi | 9505 (72,9%) | 1383 (77,3%) |
| 1 tund | 514 (3,9%) | 41 (2,3%) |
| 2 tundi | 921 (7,1%) | 89 (5,0%) |
| 3 või enam tundi | 2096 (16,1%) | 277 (15,5%) |

## 2.2 Kohvi ja suremuse põhjuslik mõju mitteparameetrilistel meetoditel

Järgnevalt hindame geenivaramu andmete peal kohvi põhjuslikku mõju suremusele, kasutades selleks mitteparameetrilisi meetodeid. Selles peatükis on võimalikeks segavateks teguriteks kolm kahe tasemega tunnust: sugu, vanusegrupp ja haridustase.




Joonis 2: Suremus kohvijoojate ja mittejoojate seas.

Joonis 2 kujutab, kui suur osa kohvi tarbivatest ning kohvi mittetarbivatest inimestest suri kümne aasta jooksul peale geenivaramuga liitumist. Näeme, et kohvi mitte joovate inimeste seas on suremus kõrgem võrreldes kohvijoojatega.

Tabel 2: Kohvijoojate ja mittejoojate jaotus.

| A(1-joob, 0-ei joo) | Kokku | Surnud 10 aasta pärast ($Y = 1$) | Elus 10 aasta pärast ($Y = 0$) |
|---|---|---|---|
| A=1 | 13 187 | 1469 | 11 718 |
| A=0 | 1812 | 308 | 1504 |

Arvutame tabeli 2 põhjal kohvi ja suremuse põhjusliku mõju. Olgu $A$ binaarne tunnus, mis näitab, kas inimene joob ($A = 1$) või ei joo ($A = 0$) kohvi. Binaarne tunnus $Y$ näitab, kas isik on 10 aasta pärast peale geenivaramuga liitumist elus ($Y = 0$) või mitte ($Y = 1$).

Leiame kõigepealt tõenäosused $P(Y = 1|A = 1)$ ning $P(Y = 1|A = 0)$. Kohvi joovaid inimesi on kokku 13 187, kellest kümne aasta jooksul suri 1469 ning kohvi mittejoovaid 1812, kelles suri 308. Seega

$$P(Y = 1|A = 1) = \frac{1469}{13\,187} \approx 0{,}111 \text{ ja}$$
$$P(Y = 1|A = 0) = \frac{308}{1812} \approx 0{,}170.$$

Kuna andmed on saadud vaatlusuuringu tulemusena, peab põhjusliku mõju hindamisel eeldama, et vahetatavus kehtib, see tähendab

$$P(Y^{a=1} = 1) = P(Y = 1|A = 1) \text{ ning}$$
$$P(Y^{a=0} = 1) = P(Y = 1|A = 0).$$

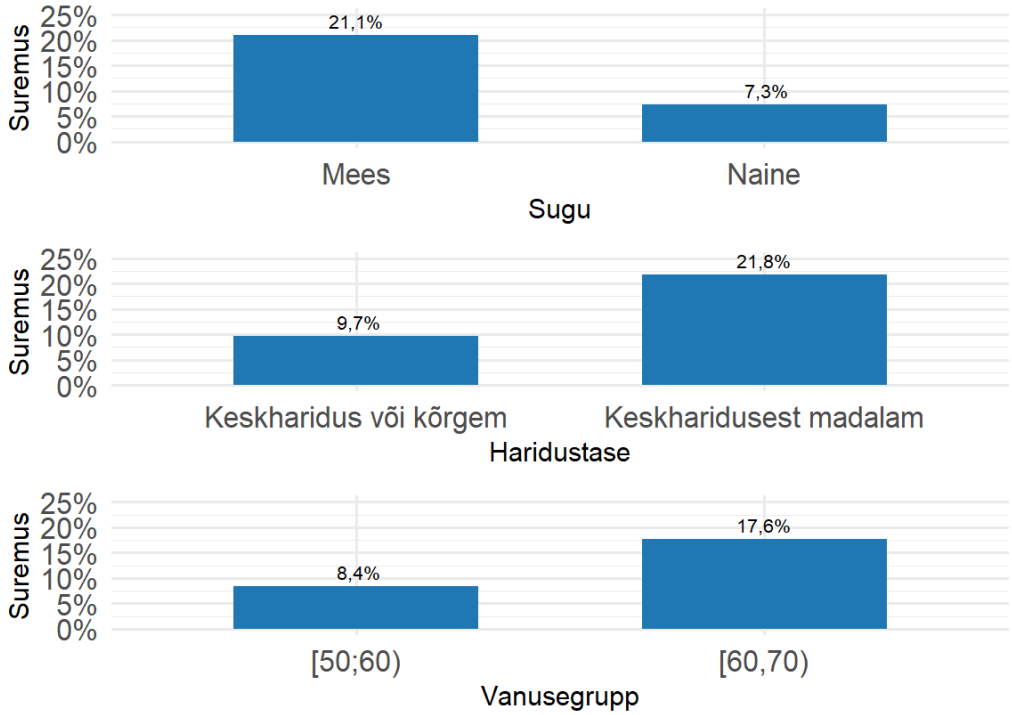
Hindame põhjuslikku mõju kohvi ja suremuse vahel, kasutades põhjusliku riski vahe leidmist:

$$P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = \frac{1469}{13\,187} - \frac{308}{1812} \approx -0{,}059.$$

Põhjusliku riski vahe ei võrdu nulliga, mis viitab sellele, et kohvi joomisel on põhjuslik mõju suremusele. Eeldusel, et kehtib vahetatavus järeldub, et kohvi joomine vähendab 10 aasta sees suremise tõenäosust 5,9% võrra võrreldes kohvi mittejoomisega. Vahetatavuse kehtivus ei ole aga antud andmestikus realistlik.

### 2.2.1 Kohvi ja suremuse põhjuslik mõju ühe segaja korral

Joonisel 3 on välja toodud suremus kolme segava tunnuse, sugu, haridus ja vanusegrupp, korral. Järgnevalt uurime, kuidas hinnata põhjuslikku mõju ühe segava tunnuse $L$ arvesse võtmisel.

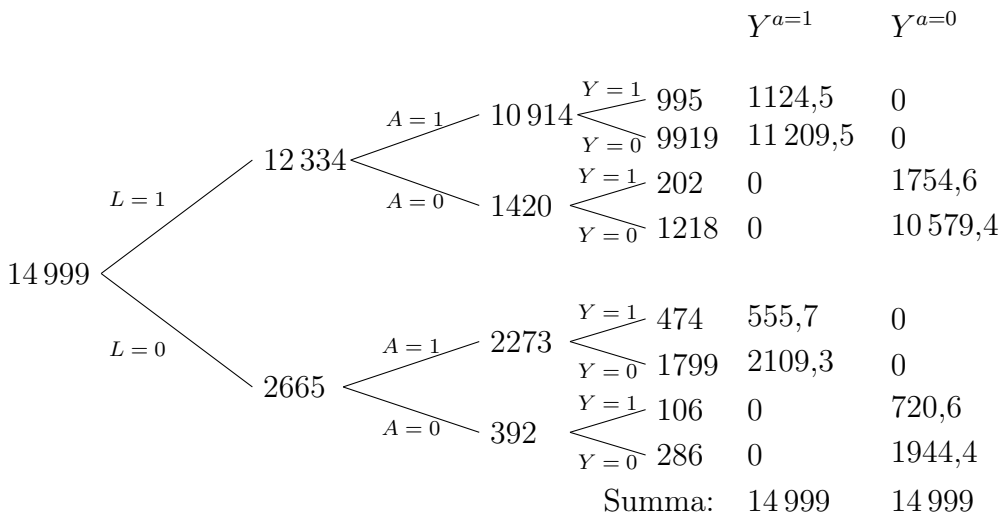



Joonis 3: 10 aasta suremus soo, haridustaseme ja vanusegrupi lõikes.

Segavaks tunnuseks $L$ valime hariduse. Tunnus näitab, kas inimesel on vähemalt keskharidus ($L = 1$) või on tal sellest madalam haridustase ($L = 0$). Nagu joonilt 3 näha, on keskharidusest madalamat haridust omavate inimeste seas suremus 12,1% võrra kõrgem kui keskhariduse või kõrgema haridusega isikute seas. Põhjusliku mõju hindamiseks saame kasutada kas pöördtõenäosuse kaalumise või standardiseerimise meetodit. Antud meetodite korral peab kehtima tinglik vahetatavus $Y^a \perp A|L$. Andmete põhjal koostame tõenäosuspuu, mis on välja toodud joonisel 4. Jooniselt näeme, et andmestikus on vähemalt keskharidusega 12 334 ning sellest madalama haridustasemega 2665 inimest. Vähemalt keskharidusega inimeste seas

on kohvijoojaid 10 914 (88,5%) ning mittejoojaid 1420. Keskharidusest madalama haridustasemega kohvi tarbijaid on 2273 (85,3%) ja mittetarbijaid 392.

Oletame, et tinglik vahetatavus (Definitsioon 3) kehtib ning hindame tõenäosused $P(Y^{a=1} = 1)$ ja $P(Y^{a=0} = 1)$ kasutades pöördtõenäosuse kaalumise meetodit.




Joonis 4: Tõenäosuspuu pöördtõenäosuse kaalumise meetodi jaoks.

Joonisel 4 tähistab veerg $Y^{a=1}$, kui palju inimesi sureks ning palju jääks ellu, kui kõik inimesed jooksid kohvi. Nagu jooniselt näha, on veerus $Y^{a=1}$ nullid nendes ridades, kus $A = 0$ ehk inimesed kohvi ei tarbi. Teistes ridades on arvud leitud järgmiselt: vähemalt keskharidusega inimesi on kokku 12 334, kellest 10 914 jõid kohvi ning kohvijoojatest suri kümne aasta jooksul 995 inimest. Kui kõik vähemalt keskharidusega inimesed oleksid joonud kohvi, siis oleks kümne aasta jooksul surnud $\frac{12\,334}{10\,914} \cdot 995 \approx 1124{,}5$ inimest, kus

$$\frac{12\,334}{10\,914}$$

on vähemalt keskharidusega kohvi joovate inimeste kaal ehk

$$\frac{1}{P(A = 1|L = 1)} = \frac{1}{\frac{10\,914}{12\,334}} = \frac{12\,334}{10\,914} \approx 1{,}13.$$

Analoogiliselt, kui kõik vähemalt keskharidusega inimesed oleksid tarbinud kohvi, oleks peale kümmet aastat ellu jäänud $\frac{12\,334}{10\,914} \cdot 9919 \approx 11\,209{,}5$ inimest. Samamoodi saame leida kümne aasta jooksul surnud ja ellu jäänud inimeste arvud ka madalama haridustasemega inimeste seas, eeldades, et kõik oleksid joonud kohvi. Sama ideed kasutades saame leida arvud joonisel veergu $Y^{a=0}$, see tähendab olukorra jaoks, kui mitte keegi ei jooks kohvi.

Kasutades joonisel veergudes $Y^{a=1}$ ja $Y^{a=0}$ olevaid arve, saame hinnata vajalikud tõenäosused $P(Y^{a=1} = 1)$ ja $P(Y^{a=0} = 1)$ järgmiselt:

$$P(Y^{a=1} = 1) = \frac{1124{,}5 + 555{,}7}{14\,999} \approx 0{,}112$$
$$P(Y^{a=0} = 1) = \frac{1754{,}6 + 720{,}6}{14\,999} \approx 0{,}165.$$

Põhjuslikku mõju saame nüüd hinnata põhjusliku riski vahega

$$P(Y^{a=1} = 1) - P(Y^{a=0} = 1) = 0{,}112 - 0{,}165 \approx -0{,}053.$$

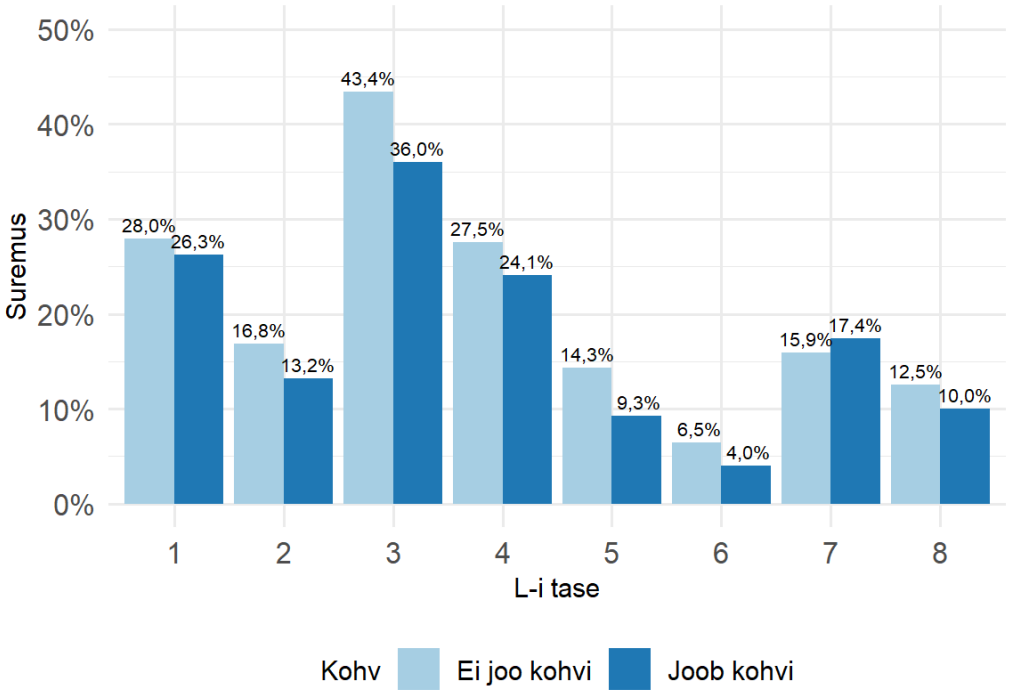
Saadud väärtus $-0{,}053$ tähendab, et suremise risk oleks ligikaudu 5,3% madalam juhul, kui kõik oleksid kohvijoojad, võrreldes juhuga, kui keegi ei jooks kohvi.

### 2.2.2 Kohvi ja suremuse põhjuslik mõju kolme segaja korral

Leiame põhjusliku mõju juhul, kui arvesse on võetud kõik joonisel 3 kujutatud segajad: inimese sugu, vanusegrupp ning haridustase. Kõik $L$-i taseme väärtused, nende kirjeldused, vastavasse tasemesse kuuluvate inimeste ja kohvijoojate arvud on välja toodud tabelis 3.

Tabel 3: Segaja $L$ tasemete selgitused.

| Väärtus | Sugu | Vanusegrupp | Vähemalt keskharidus | Inimeste arv | Kohvijoojate arv (%) |
|---|---|---|---|---|---|
| L=1 | mees | [50; 60) | ei | 565 | 472 (83,5%) |
| L=2 | mees | [50; 60) | jah | 2440 | 2012 (82,5%) |
| L=3 | mees | [60; 70) | ei | 589 | 467 (79,3%) |
| L=4 | mees | [60; 70) | jah | 1329 | 1093 (82,2%) |
| L=5 | naine | [50; 60) | ei | 693 | 623 (89,9%) |
| L=6 | naine | [50; 60) | jah | 5744 | 5251 (91,4%) |
| L=7 | naine | [60; 70) | ei | 818 | 711 (86,9%) |
| L=8 | naine | [60; 70) | jah | 2821 | 2558 (90,7%) |




Joonis 5: Suremus kohvijoojate ja mittejoojate seas $L$ tasemeti.

Joonisel 5 on välja toodud suremus kohvi tarbijate ja mittetarbijate seas kõikidel tabelis 3 kirjeldatud segajate $L$ tasemetel. Nagu jooniselt näha, on kõige kõrgem suremuse $L$-i tasemel 3, mis vastab keskharidusest madalama haridustasemega meestele vanuses 60 kuni 70 eluaastat. Kõige madalam on suremus tasemel 6, mis iseloomustab naisi vanuses 50 kuni 60, kellel on vähemalt keskharidus. Samuti näeme jooniselt, et peaaegu kõigis gruppides on kohvi joovate inimeste suremus

madalam kohvi mittejoojatest, erandiks on tase 7, kuhu kuuluvad $60-70$-aastased keskharidusest madalama haridusega naised.

Kasutame põhjusliku mõju hindamiseks standardiseerimise meetodit. Meetodi korral on eeldus, et tinglik vahetatavus kehtib ning üldine arvutuseeskiri on järgmine:

$$P(Y^a = 1) = \sum_l P(Y = 1|L = l, A = a) \cdot P(L = l).$$

Tabel 4: Kohvijoojate ja mittejoojate jaotus $L$-i tasemeti.

| L väärtus | $A$ (kohv) | Inimesi kokku | $Y = 1$ (suri) | $Y = 0$ (ei surnud) |
|---|---|---|---|---|
| 1 | 1 | 472 | 124 | 348 |
| 1 | 0 | 93 | 26 | 67 |
| 2 | 1 | 2012 | 266 | 1746 |
| 2 | 0 | 428 | 72 | 356 |
| 3 | 1 | 467 | 168 | 299 |
| 3 | 0 | 122 | 53 | 69 |
| 4 | 1 | 1093 | 263 | 830 |
| 4 | 0 | 236 | 65 | 171 |
| 5 | 1 | 623 | 58 | 565 |
| 5 | 0 | 70 | 10 | 60 |
| 6 | 1 | 5251 | 209 | 5042 |
| 6 | 0 | 493 | 32 | 461 |
| 7 | 1 | 711 | 124 | 587 |
| 7 | 0 | 107 | 17 | 90 |
| 8 | 1 | 2558 | 257 | 2301 |
| 8 | 0 | 263 | 33 | 230 |

Tabelis 4 toodud andmete põhjal saab standardiseerimist kasutades hinnata tõenäosused $P(Y^{a=1}=1)$ ning $P(Y^{a=0}=1)$ järgmiselt:

$$\begin{aligned}
P(Y^{a=1}=1) &= \frac{124}{472}\cdot\frac{565}{14\,999}+\frac{266}{2012}\cdot\frac{2440}{14\,999}+\frac{168}{467}\cdot\frac{589}{14\,999}+\frac{263}{1093}\cdot\frac{1329}{14\,999}\\
&\quad+\frac{58}{623}\cdot\frac{693}{14\,999}+\frac{209}{5251}\cdot\frac{5744}{14\,999}+\frac{124}{711}\cdot\frac{818}{14\,999}+\frac{257}{2558}\cdot\frac{2821}{14\,999}\\
&\approx 0{,}115\\
P(Y^{a=0}=1) &= \frac{26}{93}\cdot\frac{565}{14\,999}+\frac{72}{428}\cdot\frac{2440}{14\,999}+\frac{53}{122}\cdot\frac{589}{14\,999}+\frac{65}{236}\cdot\frac{1329}{14\,999}\\
&\quad+\frac{10}{70}\cdot\frac{693}{14\,999}+\frac{32}{493}\cdot\frac{5744}{14\,999}+\frac{17}{107}\cdot\frac{818}{14\,999}+\frac{33}{263}\cdot\frac{2821}{14\,999}\\
&\approx 0{,}143.
\end{aligned}$$

Põhjuslik mõju avaldub seega riskide vahena $P(Y^{a=1}=1)-P(Y^{a=0}=1)=0{,}115-0{,}143\approx -0{,}028$. Saadud vahe ei võrdu nulliga, mis tähendab, et eksisteerib põhjuslik mõju kohvi tarbimise ja suremuse vahel: kui kõik inimesed jooksid kohvi, oleks suremus 2,8% madalam võrreldes juhuga, kui keegi ei jooks kohvi.

Tabel 5: Põhjuslikud parameetrid mitteparameetriliste meetodite korral.

| Arvesse võetud segajad | $P(Y^{a=1}=1)$ | $P(Y^{a=0}=1)$ | riskide vahe |
|---|---|---|---|
| $\emptyset$ | 0,111 | 0,170 | −0,059 |
| haridus | 0,112 | 0,165 | −0,053 |
| haridus, sugu, vanusegrupp | 0,115 | 0,143 | −0,028 |

Ülalolevas tabelis on välja toodud kõik eelnevalt leitud tulemused. Märkame, et tõenäosus juhul, kui kõik joovad kohvi, on kõikide segajate korral sarnane. Rohkem varieerub erinevaid segajaid arvesse võttes suremistõenäosus olukorras, kus keegi ei jooks kohvi. Samuti näeme, et rohkemate segajate korral väheneb riskide vahe absoluutväärtus. Seega võime arvata, et võttes arvesse kõikvõimalikud segajad, kaob kohvi ja suremuse vaheline põhjuslik mõju ära.

## 2.3 Kohvi ja suremuse põhjuslik mõju mudelitega

Antud alapeatükis hindame kohvi ja suremuse põhjuslikku mõju, kasutades mudelipõhiseid pöördtõenäosuse kaalumise ja standardiseerimise meetodeid.

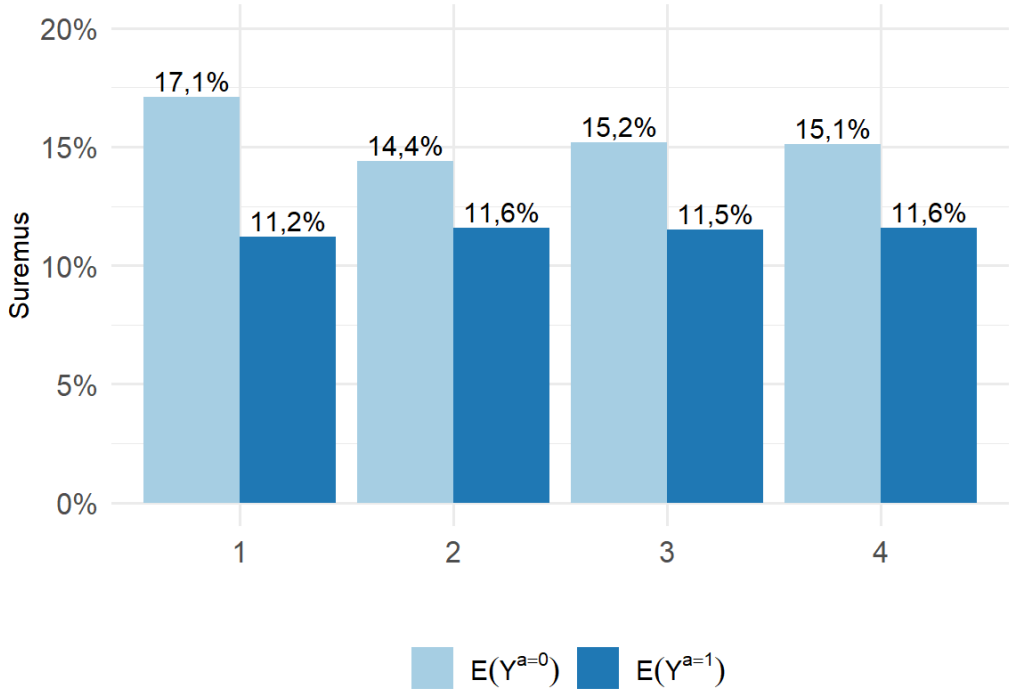
### 2.3.1 Pöördtõenäosuse kaalumist kasutavad mudelid

Vaatame kohvi mõju suremusele, kasutades mudelipõhist pöördtõenäosuse kaalumist. Põhjusliku mõju hinnangu $\beta_1$ saame, kui kõigepealt hindame logistilise regressioonmudeli kohvijoomisele, kus argumenttunnuseks on $L$ komponendid. Saadud mudeli abil saame leida tõenäosused $P(A = 0|L)$ ja $P(A = 1|L)$ ning nendega igale indiviidile vastavad kaalud. Leitud kaalusid kasutades teeme uue logistilise regressioonmudeli, mis vaatab suremuse ja kohvi seost, võttes arvesse eelnevalt leitud kaalud. Saadud mudeli (9) põhjal saame öelda, et kohvil on põhjuslik mõju suremusele, kui $\beta_1 \neq 0$ ning $\beta_1$ on statistiliselt oluline (usaldusintervall ei sisalda nulli).

Tabelis 6 on toodud erinevate mudelite $\beta_1$ hinnangud, nende standardvead, šansside suhted koos 95%−te usaldusintervallidega ning riskide vahe hinnangud. Näeme, et kaasates mudelisse peale soo ja vanuse ka hariduse, väheneb riskide vahe absoluutväärtus ehk kohvi tarbimise mõju suremusele väheneb. Samas, lisades suitsetamise, riskide vahe absoluutväärtus jällegi kasvab: kui kõik jooksid kohvi, oleks suremisrisk ligikaudu 5% võrra madalam, kui keegi ei jooks kohvi. Südame- ja seedeelundite haiguseid segajana arvestades riskide vahe absoluutväärtus jällegi väheneb, kuid kui kaasata kõik andmestikus olevad tunnused, saame riskide vahe väärtuseks −0,047, mis tähendab, et kohvi üldine tarbimine vähendaks suremust 4,7% võrra võrreldes olukorraga, kus kohvi ei tarbitaks üldse. Šansside suhte 95%-ne usaldusintervall ei sisalda ühte, seega saame öelda, et leitud põhjuslik mõju on statistiliselt oluline.

Tabel 6: Põhjuslikud parameetrid pöördtõenäosuse kaalumise meetodil.

| Arvesse võetud segajad $L$ | $\beta_1$ (standardviga) | Šansside suhe (95%$CI$) | Riskide vahe |
|---|---|---|---|
| sugu, vanus | −0,250 (0,072) | 0,779 (0,676; 0,896) | −0,037 |
| sugu, vanus, haridus | −0,228 (0,072) | 0,796 (0,692; 0,917) | −0,034 |
| sugu, vanus, haridus, suitsetamine | −0,325 (0,074) | 0,723 (0,626; 0,834) | −0,050 |
| sugu, vanus, haridus, suitsetamine, $H_{SV}$ | −0,295 (0,074) | 0,745 (0,644; 0,862) | −0,044 |
| sugu, vanus, haridus, suitsetamine, $H_{SV}$, $H_{SE}$ | −0,294 (0,075) | 0,745 (0,644; 0,862) | −0,044 |
| sugu, vanus, haridus, suitsetamine, $H_{SV}$, $H_{SE}$, rahvus, $H_{KVR}$, $H_V$, $H_K$, vererõhk, lapsed, sai trenn | −0,305 (0,080) | 0,737 (0,631; 0,862) | −0,047 |




Joonis 6: Suremuse prognoos kohvijoojate ja mittejoojate seas erinevate mudelite korral. Joonisel tähistab 1 andmestiku põhjal leitud suremust. 2 ja 3 tähistavad mudelite prognoose, kus on segajatena arvestatud vastavalt sugu ja vanus (2) ning sugu, vanus, suitsetamine ja haridus (3). Väärtus 4 tähistab mudeli prognoosi, kuhu on kaasatud kõik segajad.

Joonisel 6 on kujutatud pöördtõenäosuse kaalumist kasutatavate mudelite prognoosid suremusele kohvijoojate ja mittejoojate seas. Nagu jooniselt näeme, on andmestiku peal leitud väärtuste korral kohvi mittetarbijate seas suremus kõige suurem ning kohvijoojate suremus kõige madalam. Samuti märkame, et kohvijoojate hulgas on kõikide mudelite korral suremus sarnane, esineb 0,4%-ne erinevus, kuid mittejoojate hulgas varieeruvad tõenäosused rohkem.

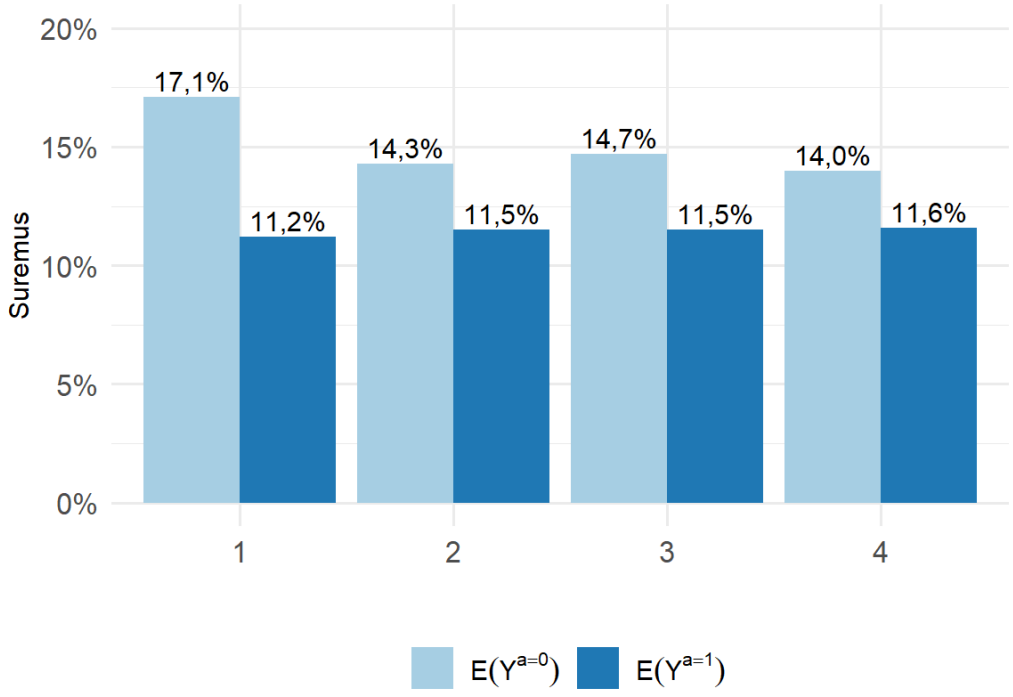
### 2.3.2 Standardiseerimist kasutavad mudelid

Mudelite abil saab põhjuslikku mõju leida ka kasutades standardiseerimist. Standardiseerimise abil saame leida $E(Y^{a=1})$ ja $E(Y^{a=0})$ ning põhjusliku mõju hindamiseks võime kasutada põhjusliku riski vahet $E(Y^{a=1}) - E(Y^{a=0})$. Põhjusliku mõju hindamiseks vajalike keskväärtuste leidmiseks peame tegema juba olemasolevast andmestikust kaks koopiat, ühele neist määrame kohvi väärtuse 1-ks (see tähendab, et kõik inimesed joovad kohvi) ning teise koopia kohvi väärtused on kõik võrdsed 0-ga (ükski inimene ei joo kohvi). Samuti on koopiana tehtud andmestike suremust näitavate veergude kõik väärtused puuduvad. Need kolm andmestikku teeme üheks andmestikuks ning hindame logistilise regressioonmudeli, kus on arvesse võetud ka valitud segajad $L$.

Tabelis 7 on välja toodud kovariandid $L$ ning standardiseerimise tulemusena saadud riskide vahe, selle standardviga ja bootstrap meetodit kasutades põhjusliku mõju hinnangu 95%−ne usaldusintervall. Näeme, et mitte ühegi $L$-i komplekti korral ei sisalda usaldusintervall nulli, seega eksisteerib statistiliselt oluline põhjuslik mõju kohvi tarbimise ning suremuse vahel. Samuti märkame, et suitsetamise lisamisel riskide vahe absoluutväärtus suureneb ning rohkemate tunnuste lisamisel riskide vahe absoluutväärtus väheneb. Kui võtta arvesse kõik andmestikus olevad tunnused, saame riskide vaheks $-0{,}024$, see tähendab, et kui kõik inimesed jooksid kohvi, sureks 2,4% vähem inimesi kui olukorras, mil ükski inimene ei jooks kohvi.

Tabel 7: Põhjusliku mõju hinnang standardiseerimise meetodil erinevate segajate korral.

| Arvesse võetud segajad $L$ | Riskide vahe (standardviga) | $(95\%CI)$ |
|---|---|---|
| sugu, vanus | -0,027 (0,009) | (-0,044;-0,010) |
| sugu, vanus, haridus | -0,024 (0,008) | (-0,040;-0,008) |
| sugu, vanus, haridus, suitsetamine | -0,032 (0,005) | (-0,042;-0,023) |
| sugu, vanus, haridus, suitsetamine, $H_{SV}$ | -0,028 (0,012) | (-0,051;-0,005) |
| sugu, vanus, haridus, suitsetamine, $H_{SV}$, $H_{SE}$ | -0,027 (0,007) | (-0,041;-0,014) |
| sugu, vanus, haridus, suitsetamine, $H_{SV}$, $H_{SE}$, rahvus, $H_{KVR}$, $H_V$, $H_K$, vererõhk, lapsed sai, trenn | -0,024 (0,008) | (-0,040;-0,008) |




Joonis 7: Suremuse prognoos kohvijoojate ja mittejoojate seas erinevate mudelite korral. Joonisel on 1-ga tähistatud tulpade väärtused leitud andmestiku peal. Numbritega 2, 3 ja 4 tähistatud tulbad iseloomustavad mudelite prognoose, kus on segajatena arvesse võetud vastavalt sugu ja vanus (2), sugu, vanus, suitsetamine ja haridus (3) ning kõik võimalikud segajad (4).

Joonisel 7 on esitatud standardiseerimist kasutavate mudelite suremuse prognoosid, kui kõik jooksid kohvi ja kui keegi ei jooks kohvi. Nagu jooniselt märgata, on ka standardiseerimist kasutavate mudelite korral suremuse hinnangud kohvijoojate korral üsna sarnased. Rohkem erineb suremus mudelite lõikes juhul, kui kohvi ei jooda.

# 3 Tulemuste arutelu

Praktilises osas käsitlesime põhjusliku mõju hindamisel erinevaid meetodeid. Järgnevalt toome välja ja võrdleme kasutatud meetoditega saadud hinnangud kohvi tarbimise põhjusliku mõju suurusele.

Mitteparameetriliste meetoditega saime absoluutväärtuselt kõige väiksema riskide vahe, kui segajatena kaasasime soo, vanuserühma ja haridustaseme. Põhjusliku mõju hinnanguks saime -0,028, mis tähendab, et võrreldes olukorraga, kus ükski inimene ei jooks kohvi, oleks suremus 2,8% madalam juhul, kui kõik oleksid kohvijoojad. Kõige suurema mõju hinnangu saime juhul, kui ühtegi segajat ei olnud kaasatud, riskide vahe oli siis $-0{,}059$, mis viitab sellele, et kohvi joomine vähendaks suremust 5,9% võrra. Selline tulemus on ootuspärane, sest ilma segajaid arvesse võtmata jäävad erinevused inimeste tausttegurites, mis võivad mõjutada nii kohvi joomist kui ka suremust, analüüsis arvestamata.

Mudelipõhise pöördtõenäosuse kaalumise meetodi puhul saime samuti kõige väiksema riskide vahe ($-0{,}034$), kui arvesse oli võetud sugu, vanus ja haridus. Võrreldes mitteparameetrilise lähenemisega, kus vanus ja haridustase olid jagatud kahte rühma, on mudelipõhisel meetodil arvestatud vanus täisaastates ning haridustase kolmes kategoorias. See võiks arvestada segajaid täpsemalt ning anda usaldusväärsema tulemuse. Mitteparameetrilise lähenemise korral tuli aga riskide vahe hinnang absoluutväärtuselt väiksem ($-0{,}028$). Kõige suurema riskide vahe hinnangu mudelipõhise pöördtõenäosuse kaalumise puhul ($-0{,}05$) saime siis, kui lisasime eelnevatele kovariantidele ka suitsetamise. Kaasates kõik andmestikus olevad tunnused, saime riskide vahe väärtuseks $-0{,}047$, mis on oma absoluutväärtuse poolest suuruselt teine. Tulemus on veidi mitteootuspärane, sest oleksime arvanud, et mida rohkem segajaid kaasata, seda väiksemaks kohvi joomise ja suremuse seos muutub.

Kõiki segajaid, mida kasutasime mudelipõhise pöördtõenäosuse kaalumise puhul, rakendasime ka standardiseerimisel. Ka siin saime suurima absoluutse riskide vahe, kui kaasasime mudelisse soo, vanuse, hariduse ning suitsetamise. Siiski oli stan-

dardiseerimist kasutades saadud hinnang 0,018 võrra väiksem kui pöördtõenäosuse kaalumise puhul. Standardiseerimisega saime madalaima riskide vahe hinnangu ($-0{,}024$) nii juhul, kui arvestasime segajatena ainult sugu, vanust ja haridust kui ka siis, kui kaasasime kõik andmestiku tunnused. Mitteparameetrilise meetodiga arvutades saime riskide vahe hinnanguks, juhul kui segajatena arvestasime sugu, vanusegruppi ja haridust, $-0{,}028$, mis on ligilähedane mudelipõhise standardiseerimisega saadud tulemusele, erinevus on vaid 0,004. Samuti märkame tabeleid 6 ja 7 vaadates, et pöördtõenäosuse kaalumise meetodil on kõikide mudelite korral riskide vahe aboluutväärtuselt suurem kui standardiseerimise mudelite puhul. Lisaks näeme joonistelt 6 ja 7, et võrreldes andmestiku peal leitud $E(Y|A = 1)$ ja $E(Y|A = 0)$ korral (joonistel tulbad 1), muutub kohvijoojate seas suremuse hinnang kõrgemaks ning mittejoojate seas madalamaks. Arvatavasti võib olla põhjuseks see, et algsed grupid on ka muudel põhjustel erinevate suremusriskidega.

# Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks on uurida kohvi joomise mõju inimese kümneaastasele suremusele. Analüüsi läbiviimiseks kasutame Tartu Ülikooli Eesti geenivaramu andmestikku. Põhjusliku mõju hindamiseks oleme rakendanud mitteparameetrilist ja mudelipõhist pöördtõenäosuse kaalumist ja standardiseerimist. Segajatena käsitlesime nii inimese sugu, vanust, diagnoositud haiguseid kui ka tervisekäitumist.

Saadud tulemused ühtivad varasemate uuringute tulemustega: kohvi joovatel inimestel on madalam suremusrisk kui neil, kes kohvi ei joo. See viitab kohvi joomise ja suremuse vahelisele põhjuslikule mõjule. Samas on käsitletud meetodite eelduseks vahetatavuse tingimus, mille saab tagada aga randomiseeritud katse tegemisega. Antud töös tegeleme aga vaatlusandmetega ning me ei saa kindlalt väita, et vahetatavus kehtib. Samuti võib kohvi joomist ja mittejoomist mõjutada veel mõni segaja, mida meie analüüsis käsitletud ei ole.

Töö võimalikuks edasiarenduseks võime põhjusliku mõju hindamisel arvesse võtta rohkem segajaid, sealhulgas ka geneetilisi tegureid, mis võivad mõjutada nii kohvi tarbimist kui ka suremust. See aitaks vähendada segajate tekitatud mõju ja parandada uuritava seose täpsust. Samas ei saa me ilma randomiseeritud katse tegemiseta olla kindlad, et vahetatavus kehtib ning leitud tulemused ei taga täielikku usaldusväärsust. Lisaks saab töös kasutatud meetodeid rakendada ka teiste tegurite põhjusliku mõju hindamisel.

# Kasutatud allikad


Canty, Angelo, Brian Ripley ja Alessandra R. Brazzale (2024). *Package 'boot'*. URL: https://cran.r-project.org/web/packages/boot/boot.pdf.

Gunter, Marc J., Neil Murphy, Amanda J. Cross, Laure Dossus *et al.* (juuli 2017). *Coffee Drinking and Mortality in Ten European Countries – the EPIC Study*. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5788283/.

Hernán, Miguel A. ja James M. Robins (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC Press, lk. 3–180. URL: https://miguelhernan.org/whatifbook.

Højsgaard, Søren, Ulrich Halekoh ja Jun Yan (2005). „The R Package Geepack for Generalized Estimating Equations". *Journal of Statistical Software*. URL: https://www.jstatsoft.org/article/view/v015i02.

Nahhas, Raamzi W. (2025). *Introduction to Regression Methods for Public Health Using R*. Boca Raton: Chapman & Hall/CRC Press, lk. 179–180. URL: https://bookdown.org/rwnahhas/RMPH/.

Waples, Josef (2024). *What is Bootstrapping in Statistics? A Deep Dive*. URL: https://www.datacamp.com/tutorial/bootstrapping?dc_referrer=https%3A%2F%2Fwww.google.com%2F (vaadatud 04.04.2025).

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Laura Himma,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Kas kohvi joomine pikendab eluiga? Põhjuslike mõjude analüüs TÜ Eesti geenivaramu andmete põhjal“, mille juhendaja on Krista Fischer, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Laura Himma
14.05.2025