Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kärttu Põrk

PISA andmeanalüüsi metoodika tutvustamine ja selle rakendamine

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: MSc Hannes Jukk

Tartu 2026

PISA ANDMEANALÜÜSI METOODIKA TUTVUSTAMINE JA SELLE RAKENDAMINE

Bakalaureusetöö

Kärttu Põrk

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on kirjeldada PISA (Programme for International Student Assessment) andmete analüüsimise protsessi ning töötada välja korduvkasutatav analüüsiraamistik, mis toetab metoodiliselt korrektset andmeanalüüsi. PISA uuringu keeruka valimidisaini tõttu on andmete statistiline analüüs nõudlik ning praktikas võidakse kasutada lihtsustatud lähenemist, mis võib viia ebatäpsete tulemusteni.

Töös kirjeldatakse PISA uuringu andmestruktuuri ning rakendatavaid statistilisi meetodeid, pöörates erilist tähelepanu tõepäraväärtuste ja replikatsioonide kaalude korrektsele kombineerimisele. Selle tulemusena töötatakse *R* tarkvara abil välja funktsioonide kogum, mis võimaldab läbi viia kirjeldavat statistikat ning korrelatsiooni- ja regressioonianalüüse vastavalt PISA metoodilistele nõuetele. Loodud lahendus on suunatud eeskätt tudengitele ja teistele huvilistele, kes soovivad PISA andmeid korrektselt analüüsida. Töö loob aluse edaspidiseks metoodiliselt keerukamate analüüside laiendamiseks.

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

Märksõnad: PISA, tõepäraväärtused, replikatsioonide kaalud, IRT, andmeanalüüs, *R* tarkvara, kompleksandmed.

INTRODUCTION TO PISA DATA ANALYSIS METHODOLOGY AND ITS APPLICATION

Bachelor thesis

Author

Abstract

The aim of this bachelor's thesis is to describe the process of analysing PISA (Programme for International Sudent Assesment) data and to develop a reusable analytical framework that supports methodologically sound data analysis. Due to the complex sampling design of the PISA study, statistical analyses of the data is demanding, and simplified approaches are sometimes applied in practice, which may lead to inaccurate results.

The thesis describes the structure of PISA data and the statistical methods used in their analysis, with particular emphasis on the correct combination of plausible values and replication weights. As a result, a set of functions is developed using the *R* software environment, enabling the application of descriptive statistics as well as correlation and regression analyses in accordance with PISA methodological requirements. The developed solution is primarily intended for students and other users who lack formal statistical training but require a methodologically correct approach to analysing PISA data. The proposed framework provides a foundation for the future development of more methodologically advanced analyses.

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

Key Words: PISA, plausible values, replication weights, IRT, data analysis, *R* software, complex data.

Sisukord

Sissejuhatus

Eesti osaleb 2006. aastast õpilaste teadmiste kontrolli uuringus PISA (Programme for International Student Assessment). PISA on OECD (Organisation for Economic Co-operation and Development) poolt loodud noorte oskuste uuring, mis hindab noorte teadmisi funktsionaalse lugemise, matemaatika ja loodusteaduste valdkondades. Peamine eesmärk on seejuures hinnata õpilaste võimekust rakendada oskusi ja teadmisi elulistes situatsioonides. Uuringu läbiviimiseks moodustatakse valim osalevate riikide või majanduspiirkondade 15-aastastest (täpsemalt 15 aastat ja 3 kuud kuni 16 aastat ja 2 kuud) õpilastest, kes sooritavad kolmes teadmiste valdkonnas testi ja täidavad taustinfo kohta käiva küsimustiku. Lühike küsimustik on loodud ka koolijuhtidele koolide kohta. (HARNO, 2025; OECD, [b])

Uuringu disaini tõttu on PISA testiga kogutud andmete analüüsimine piisavalt nõudlik ning praktikas võidakse lihtsuse huvides läheneda analüüsile lihtsustatult (ühe tõepäraväärtuse kasutamine või üle tõepäraväärtuste keskmistamine) (Jewsbury, J. ja Gonzalez, 2024). Inglise keeles on avaldatud analüüsi läbi viimist lihtsustavaid materjale. Eestikeelne käsitlus PISA andmete analüüsimise kohta on puudulik. Käesoleva bakalaureusetöö eesmärk on kirjeldada PISA andmete analüüsimise protsessi ning luua korduvkasutatav analüüsi raamistik, mis aitab vähendada lihtsustatud analüüsivõtete kasutamisest tulenevate vigade riski ning toetab läbipaistvat arvutusloogikat, mida on võimalik edaspidi laiendada keerukamate analüüside tarbeks. Töös tuuakse välja vajalikud meetodid ja eestikeelsed väljendid.

Töö koosneb teoreetilisest osast ja praktilisest näitest. Teoreetilises osas tutvustatakse PISA kogutud andmete analüüsis kasutatavaid meetodeid ja valemeid. Praktilise näitena luuakse vabavaralist *R* tarkvara kasutades kood, mis

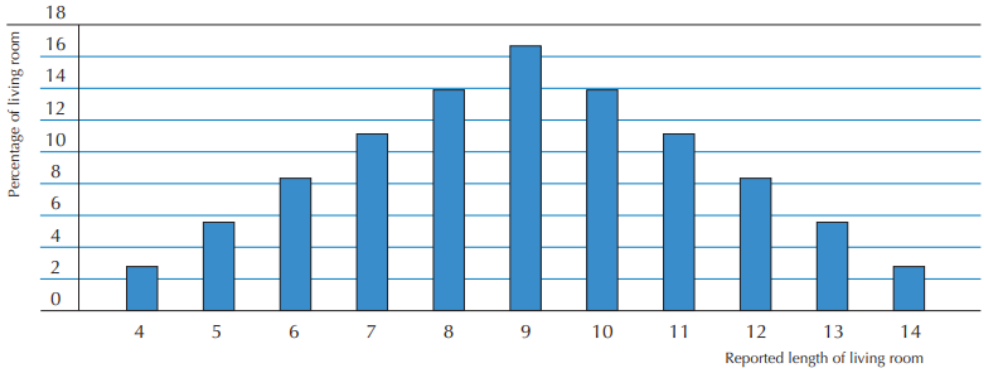
aitab mõista, kuidas teostada algelist statistilist analüüsi Eesti õpilaste andmete põhjal. Toodud kood võib olla aluseks kasutajale edaspidi keerulisemate analüüsi koodide moodustamisel.

1 Tõepäraväärtused

Selles peatükis tutvustatakse PISA testi disaini ja andmete analüüsimiseks loodud tõepäraväärtuste (ingl *plausible values*, lühend PV; nimetatakse ka prognoositud väärtused või tõepärased väärtused) leidmise ja kasutamise põhimõtteid. Järgnev näide põhineb PISA SPSS-i andmeanalüüsi juhendil (OECD, 2009).

Näide tõepäraväärtuste kontseptsiooni mõistmiseks: Oletame, et kehtestatakse hoonemaks, mis on proportsionaalne perekonna elutoa pikkusega. Inspektorid mõõdavad kõiki linna elutubasid, kusjuures pikkus tuleb neil üles märkida täisarvudes (st 1 meeter, 2 meetrit jne).

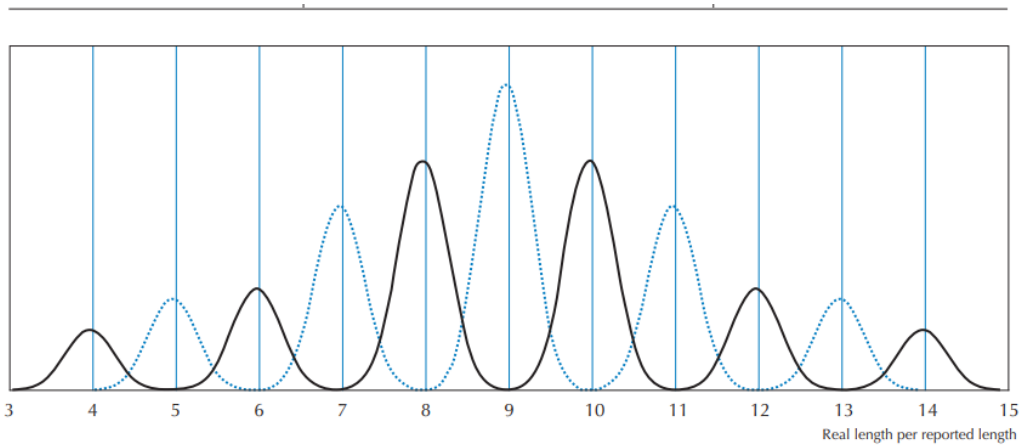
Joonisel 1 on kujutatud näite mõõtmistulemused, umbes 16% elutubadest on registreeritud 9 meetri pikkusega, ligikaudu 2% pikkusega 4 meetrit. Pikkus on pidev suurus, mille korral võivad vaatlused, erinevalt diskreetsest suurusest, võtta ükskõik, mis väärtuse miinimumi ja maksimumi vahel.



Joonis 1: Elutoa pikkused. Allikas: OECD, *PISA Data Analysis Manual: SPSS Second Edition* (lk 94).

Tegelik pikkus varieerub ümber registreeritud pikkuse keskmise (Joonis 2). Erinevuse põhjuseks on ümardamine ja mõõtmisviga. Kui elutoa pikkuseks

on 4,15, siis ümardamisveana võib olla registreeritud väärtuseks 5. Jaotuse kattuvus tuleneb mõõtmisveast.



Joonis 2: Tegelikud pikkused registreeritud pikkuste suhtes. Allikas: OECD, *PISA Data Analysis Manual: SPSS Second Edition* (lk 95).

Näites on elutubade tegelikud pikkused normaaljaotusega ümber keskmise (registreeritud pikkuse). Tõenäosus registreerida pikkus selle lähima täisarvuna sõltub tegeliku pikkuse ja lähima täisarvu erinevusest. Kui erinevus on väike, on tõenäosus suur ja kui erinevus suur, siis vastupidiselt tõenäosus väike. Näiteks pikkus 4,50 meetrit registreeritakse võrdselt nii 4 kui ka 5 meetrina, kuid 4,95 tõenäoliselt 5 meetrina. Tõepäraväärtusi saab määratleda kui apostorerioorsest jaotusest (järeljaotus) valitud juhuslikke väärtusi. Näites 7-meetrisena märgitud elutuba võib saada normaaljaotusest väärtuseks 7,45, 6,55 või 6,95. Seetõttu ei tohiks tõepäraväärtusi kasutada individuaalsete hinnangute tegemiseks.

1.1 PISA testi disain

PISA on valimipõhine uuring, kus üldkogumi (riigi või majanduspiirkonna haridussüsteemi) kohta tehakse hinnanguid. Leitud hinnangud sisaldavad teatud määral ebakindlust. Hinnangu täpsust mõjutab uuringu testimisdisain, mille kohaselt ei lahenda kõik õpilased samu ülesandeid. Ülesanded on koostatud maatriksdisaini alusel ning ülesannete komplektid kattuvad tsüklilises rotatsioonis pooles ulatuses eelmise ja järgmise komplektiga. Õpilaste oskustaseme mõõtmiseks kasutatakse psühhomeetrilisi mudeleid, mis võimaldavad hinnata ka vastamata jäänud ülesannete sooritust. (Tire *et al.*, 2023)

Võime eristada PISA testimise ajaloos kolme ajajärku. Esialgu viidi PISA hindamine läbi pabertestide abil. Alates 2015. aastast toimus oluline muutus, hakati üle minema arvutipõhisele testimisele. Arvutipõhise testimisega ei läinud kaasa kõik riigid, osades riikides jäädi sel perioodil endiselt kasutama paberteste. Arvuti- ja pabertestide paralleelne kasutamine tõi kaasa testi läbiviimise viisi mõju tulemusele (ingl *mode effect*), mistõttu muudeti tulemuste skaleerimise metoodikat, et tagada tulemuste võrreldavus varasemate uuringutsüklitega. (OECD, 2017)

Elektrooniliste testide kasutuselevõtt võimaldab koguda täpsemat infot õpilaste lahendusprotsesside kohta ning loob eeldused adaptiivsemate ja mõõtmistäpsemate hindamisvahendite arendamiseks (OECD, 2017). Alates 2025. aastast viiakse PISA uuring läbi täielikult veebipõhise e-testina, kasutades OECD keskset digihindamise platvormi (HARNO, 2025).

1.2 PISA andmete analüüs

PISA andmete analüüs hõlmab oma testi disaini ning valimi struktuuri tõttu mitmeid samme. Peamised sammud on järgmised (OECD, 2009):

- valimile lõplike kaalude (põhikaal) rakendamine;
- üksikvastuste teooria (ingl *Item Response Theory*, lühend IRT) mudeli kasutamine õpilaste latentsete oskuste (nt matemaatiline kirjaoskus või funktsionaalne lugemisoskus) hindamiseks;
- latentse (varjatud; pole otseselt mõõdetav) tunnusega regressiooni rakendamine taustamuutujate mõju analüüsimiseks;
- Bayesi teoreemi kasutamine tõepäraväärtuste saamiseks;
- tõepäraväärtuste genereerimine;
- iga tõepäraväärtuse korral eraldi analüüsi läbiviimine;
- replikatsioonide kaalude rakendamine;
- Rubini reegleid rakendades tõepäraväärtuste tulemuste kombineerimine lõplikeks hinnanguteks ja standardvigadeks.

Selline lähenemine tagab, et PISA andmete analüüsi käigus arvestatakse nii latentse tunnuse mõõtevea kui ka valimidisainist tuleva ebakindlusega (ingl *uncertainty*). PISA andmestikus on antud tõepäraväärtused ja kaalud, kuid edasised sammud andmete analüüsimisel tuleb ise rakendada. (OECD, 2009)

1.2.1 Koolide andmed

Valim koolidest on kavandatud eelkõige õpilaste valimi optimeerimiseks, mitte optimaalse valimi moodustamiseks koolidest. Seetõttu on koolitaseme andmeid soovitatav analüüsida õpilase tasemel. (Gonzalez ja Kennedy, 2003)

PISA sihtrühm põhineb õpilaste vanusel, mitte klassil, mistõttu on koolitaseme andmete analüüs õpilaste omadustena eriti oluline. Pärast õpilaste ja

kooliandmete ühendamist saab neid andmeid käsitleda nagu mistahes õpilast iseloomustavaid muutujaid, kuid sel juhul tuleb kindlasti kasutada replikatsioonide kaalusid, et vältida eksitavaid tulemusi. (OECD, 2024)

1.3 Traditsiooniline lähenemine PISA andmete analüüsimisel

PISA andmete analüüsiks loodud töövahendid erinevad märkimisväärselt nii kasutusmugavuse kui ka metoodilise läbipaistvuse poolest. Analüüsiks soovitatakse OECD ametlikes kanalites (OECD, [a]) kasutada *SAS*, *SPSS*, *Stata* või *R* tarkvara. Sealjuures koodi kirjutamise asemel on võimalus kasutada OECD poolt koostatud makrosid või *IDB Analyzer* abiprogrammi, mis võimaldab saada metodoloogiliselt korrektseid tulemusi ilma otsese programmeerimiseta.

Abiprogrammis *IDB Analyzer* on võimalik väljastada *SAS*, *SPSS* või *R* kood, täpsemalt saab abiprogrammiga *IDB Analyzer* tutvuda vastavst käsiraamatust (IEA, 2022). Kuigi *IDB Analyzer* (versioon 5.0) lihtsustab PISA andmete kasutamist ning vähendab eksimuse riski valimi kaalude ja tõepäraväärtuste käsitlemisel, jääb kasutajale tulemuste saamise loogika kohati varjatuks ja sellega on ka analüüsi võimalused piiratud.

Traditsiooniliselt PISA andmete töötlemisel kasutatud *SAS*, *SPSS* ja *Stata* on muutunud halvemini ligipääsetavaks litsentsikulude tõttu. Tarkvara Stata nõuab PISA andmete analüüsiks porgrammeerimise oskust. Kasutamise lihtsustamiseks on OECD välja töötanud repest mooduli (Avvisati ja Keslair, 2014). Vabavaraline alternatiiv eelnevalt mainitud tarkvaradele on *JASP* (MacDougall, 2024). Tarkvara jaoks ei ole arendadud abivahendit, mis toetaks tõepäraväärtuste korrektset kombineerimist. Liskas esineb *JASP*is prak-

tilisi piiranguid suurte andmestike käsitlemisel, mis on seotud mälukasutusega ning võivad takistada PISA analüüsiks vajalike mahukate andmefailide töötlemist.

Kuigi *R* tarkvara (versioon R 4.5.2) iseseisev (sõltumatult abiprogrammist *IDB Analyzer*) kasutamine pakub suurimat paindlikkust ja läbipaistvust, eeldab selle kasutamine programmeerimise oskust, mis võib probleemiks osutuda. Seega tuleb PISA andmete analüüsimisel teha valik lihtsasti kasutatavate, kuid väiksema läbipaistvusega töövahendite ning tehniliselt nõudlikuma, kuid metoodiliselt paremini kontrollitava lähenemise vahel.

1.4 Tõepäraväärtuse meetodi olulisus

PISA testide analüüsi juures kasutatakse tõepäraväärtuse metoodikat, mis kujutab suurtes rahvusvahelistes võrdlusuuringutes (PISA, TIMSS, PIAAC jt) standardset lähenemist õpilaste oskuste hindamiseks (Wu, 2005). Akadeemik Robert J. Mislevy (1989) hinnangul ei ole klassikaline testiteooria piisav õppijate tegeliku võimekuse mõistmiseks, sest see tugineb liigselt ühele koondhinnangule ega arvesta teadmiste mitmekesisust, mõtlemisstrateegiaid ega õppimise käigus toimuvat muutust. Sellest lähtuvalt võeti rahvusvahelistes haridusuuringutes kasutusele 1990. aastate keskpaigast tõepäraväärtuse meetod (OECD, 2009).

Tõepäraväärtuseid saab vaadelda kui mitmese imputatsiooni tehnika kasutamisega moodustatud hinnangute hulka (Davier, Gonzalez ja Mislevy, 2009). PISA mõistes tähendab see, et iga õpilase tegeliku oskuse kohta ei anta ühte kindlat punktiskoori, vaid luuakse mitu (alates 2015. aastast 10) tõenäolist hinnangut, mis õpilane oleks võinud saada, tuginedes tema vastustele ja taustmuutujatele. Vastavad väärtused aitavad paremini arvestada mõõtmis-

vigadega. Selline lähenemine aitab vähendada õpilaste ebaõiglast järjestamist ning sellest tulenevalt valede tulemuste interpreteerimist. (OECD, 2009)

Tõepäraväärtuse leidmine põhineb mitmese imputatsiooni tehnikal ja üksikvastuste teoorial, mis püüab selgitada seost latentsete tunnuste ja vaadeldavate tulemuste vahel (Hambleton, 1990). Tõepäraväärtuseid eelistatakse klassikalistele üksikvastuste teooria hinnangutele (suurima tõepära hinnangud ja kaalutud suurima tõepära hinnangud) kuna tõepäraväärtustel esinevad metoodilised eelised nagu näiteks õpilaste tulemuste pidev jaotus, nihketa hinnangud soorituse ja taustmuutujate vahelistele seostele ning üldkogumi parameetritele (keskmine, standardhälve) (OECD, 2009).

Tõepäraväärtuste põhjal saab teha täpseid järeldusi grupi tasandil, kuid neid ei tohiks kasutada individuaalsete punkthinnangutena (OECD, 2024). Seda seetõttu, et iga õpilase tõepäraväärtused on statistiliselt genereeritud väärtused, mis peegeldavad pigem üldkogumit ja alarühmade omadusi kui konkreetse õpilase tegelikku võimekust. Ühe õppija tõepäraväärtused võivad varieeruda ning ei ole mõeldud stabiilse või täpse isikliku tulemusena, vaid sisendina mudelitele, mis hindavad üldisi trende, rühmade erinevusi ja seoseid taustmuutujatega. (OECD, 2009)

1.4.1 Üksikvastuste teooria

Üksikvastuste teooria on mõõtmismudelite kogum, mille eesmärk on selgitada testitavate isikute vastuseid testülesannetele nende latentsete omaduste või võimekuste kaudu. (Hambleton, 1990). Järgnev alapeatükk põhineb PISA tehnilisel raportil (OECD, 2024).

PISA testide tulemuste modelleerimisel kasutatakse erinevaid üksikvastuste teooria mudeleid, sõltuvalt vastamise tüübist, mis kirjeldavad vastuse tõe-

näosust latentse tunnuse funktsioonina. Vastuste ja latentse tunnuse vahelist seost kirjeldatakse IRT-mudelitega, mis määravad vastuse tõenäosuse tingimusel θ_v (õpilase latentne oskustase). Käesolevas töös esitatakse seose illustreerimiseks kaheparameetriline logistiline mudel (ingl *two-parameter logistic model*, lühend 2PL). Kaheparameetriline logisiline mudel on Raschi mudeli üldistus, kus lisaks raskusparameetrile eeldatakse ülesande spetsiifilist eristusparameetrit (a_i). Olgu $\mathbf{x}_v$ õpilase v vastusvektor, mis koosneb kõikide testülesannete vastustest, siis vastuse tõenäosus on antud kujul:

$$P_i(x_{vi} = 1 \mid \theta_v, b_i, a_i) = \frac{e^{Da_i(\theta_v - b_i)}}{1 + e^{Da_i(\theta_v - b_i)}},$$

kus $P_i(x_{vi} = 1|\theta_v, b_i, a_i)$ on tõenäosus, et õpilane tasemega θ_v vastab ülesandele i õigesti, see väljendub logistilise kõverana, mille väärtused jäävad 0 ja 1 vahele. Ülesande i raskust iseloomustav parameeter on b_i. Skaleerimiskonstandi D eesmärk on muuta logistiline funktsioon võimalikult sarnaseks normaalse kumulatiivse sagedus funktsiooniga. Empiirilised uuringud on näidanud, et kui valida $D = 1{,}7$, siis kaheparameetrilise logistilise mudeli ja kaheparameetrilise normaalse kumulatiivse mudeli abil saadud tõenäosushinnangute erinevus jääb kõigi parameetri θ väärtuste korral väga väikeseks, kusjuures absoluutne erinevus ei ületa 0,01. Skaleerimiskonstant tuleb korrutada eristusparameetriga a_i, mis kirjeldab, kui hästi ülesanne suudab eristada erineva võimekusega testitavaid. See parameeter on seotud ülesande karakteristikkõvera tõusuga raskusparameetri b_i piirkonnas. Mida suurem on kõvera tõus selles punktis, seda paremini suudab ülesanne eristada testitavaid, kelle võimekus jääb vastava taseme lähedusse.

Kõigi ülesannete vastuste põhjal moodustub vastuste tõepärafunktsioon

$$P(\mathbf{x}_v|\theta_v, \boldsymbol{\beta}) = \prod_{i=1}^{n} P(x_{vi}|\theta_v, \beta_i). \tag{1}$$

See kirjeldab, kui tõenäolised on õpilase vastused $\mathbf{x}_v$ erinevate latentse oskustasemete θ_v korral kui testis on n ülesannet ning β_i tähistab ülesande i kohta käivate parameetrite vektorit, mis on saadud vastavast IRT mudelist. Käesolevas töös loetakse ülesandeparameetrid eelnevalt hinnatuks ning järgnevas analüüsis neid ei hinnata.

1.4.2 Latentne regressioonimudel

PISA testis eeldatakse, et õpilaste latentne tunnus sõltub taustmuutujatest (nt sotsiaalmajanduslik staatus, sugu jne). Selle seose kirjeldamiseks kasutatakse latentset regressioonimudelit (ingl *latent regression*), milles eeldatakse, et latentne tunnus järgib mitmemõõtmelist (D-mõõtmelist) normaaljaotust :

$$\boldsymbol{\theta_v} \sim N_D(\Gamma'\mathbf{y}_v, \Sigma), \tag{2}$$

kus Γ on regressioonikordajate maatriks ($K \times D$), K on latentse regressioonimudeli taustmuutujate arv (sealhulgas taustmuutujate põhjal saadud peakomponendid ning regressioonimudeli vabaliige), D näitab mitu latentset tunnust mudel kirjelab, Σ on kovariatsioonimaatriks ($D \times D$) ning $\mathbf{y}_v$ on õplase v taustmuutujate (kovariantide) vektor (nt sotsiaalmajanduslik staatus, kooli tüüp vms). (OECD, 2024)

Latentse regressioonimudeli abil kirjeldatakse õpilaste latentsete tunnuste jaotust sarnaste taustmuutujate korral nagu on eeldatud valemis (2) . Mudeli parameetrid (Γ ja Σ) hinnatakse PISA uuringus maksimaalse tõepä-

ra meetodil, kasutades ootuse maksimeerimise algoritmi (ingl *Expectation-maximization algorithm*, lühend EM) (Dempster, Laird ja Rubin, 1977). Mitmemõõtmeliste oskustunnuste korral kasutatakse latentse regressioonimudeli hindamisel Laplace'i lähendust (OECD, 2024; Thomas, 1993).

1.4.3 Bayesi teoreem

Järgnev alapeatükk põhineb Traat ja Lepik (2013) e-kursuse *Bayesi statistika Markovi ahelatega* materjalidel. Bayesi statistika võimaldab erinevalt klassikalisest statistikast mudelisse lisada eelteadmisi mudeliparameetrite α ja β kohta. Seejuures käsitletakse parameetreid α ja β juhuslike suurustena ning nende tulemus saadakse jaotusena. Bayesi hindamisülesande põhikomponendid on:

- eeljaotus (α ja β aprioorne jaotus);
- mudel (andmete ja parameetrite vaheline seos);
- valimi andmed.

Nende põhjal tuletatakse apostorerioorne jaotus, mis annab kogu info parameetrite kohta.

1.4.4 Aposterioorse jaotuse moodustamine ja tõepäraväärtuste genereerimine

Järgnev alapeatükk põhineb Traat ja Lepik (2013) e-kursuse *Bayesi statistika Markovi ahelatega* materjalidel ja PISA tehnilisel raportil (OECD, 2024). PISA metoodikas moodustatakse õpilase latentse oskuse aposterioorne jaotus

vastuste tõepärafunktsiooni ja latentse regressioonimudeli kombineerimisel Bayesi teoreemi alusel.

Eeldame, et $f(\mathbf{x}|\theta)$ on tinglik tihedusfunktsioon ehk tõepärafunktsioon. Sel juhul tähistatakse tõepärafunktsiooni kujul

$$\ell(\theta) = f(\mathbf{x}|\theta).$$

Kui $p(\theta)$ on θ eeljaotus saame parameetri θ aposterioorse jaotuse leida Bayesi teoreemi abil:

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)p(\theta)}{f(\mathbf{x})}, \tag{3}$$

kus $f(\mathbf{x})$ on vektori $\mathbf{x}$ marginaalne tihedusfunktsioon :

$$f(\mathbf{x}) = \int f(\mathbf{x}|\theta)p(\theta)\, d\theta$$

ning $f(\mathbf{x}|\theta)p(\theta)$ väljendab $\mathbf{x}$ ja θ ühistihedusfunktsiooni. Funktsioon $f(\mathbf{x})$ valemis (3) on normeeriv konstant, mis ei sõltu parameetrist θ. Seetõttu saab Bayesi teoreemi kirjutada ka kujul

$$\pi(\theta) \propto \ell(\theta) \cdot p(\theta), \tag{4}$$

kus $\pi(\theta) = p(\theta|\mathbf{x})$ tähistab parameetri aposterioorset jaotust.

PISA uuringus rakendatakse Bayesi teoreemi, et leida aposterioorne jaotus, mis kirjeldab latentse tunnuse tõenäosusjaotust, arvestades nii õpilase testivastuseid kui ka taustmuutujaid. Valemi (4) näide. Olgu $\boldsymbol{\theta}_v$ õpilase v latentne tunnus, $\mathbf{x}_v$ testivastused ning $\mathbf{y}_v$ taustmuutujad (nt sotsiaalmajanduslik staatus või kooli tüüp). Sellisel juhul moodustatakse latentse tunnuse aposterioorne jaotus kujul:

$$P(\boldsymbol{\theta}_v|\mathbf{x}_v, y_v, \Gamma, \Sigma) \propto P(\mathbf{x}_v|\boldsymbol{\theta}_v, \mathbf{y}_v, \Gamma, \Sigma) \cdot P(\boldsymbol{\theta}_v|\mathbf{y}_v, \Gamma, \Sigma) = P(\mathbf{x}_v|\boldsymbol{\theta}_v) \cdot P(\boldsymbol{\theta}_v|\mathbf{y}_v, \Gamma, \Sigma),$$

kus parameetrid Γ ja Σ kirjeldavad latentse regressiooni koefitsiente ning skaaladevahelist variatsiooni ja korrelatsiooni. Tinglik tõenäosus $P(\mathbf{x}_v|\boldsymbol{\theta}_v)$ peegeldab IRT-mudeli põhist tõepärafunktsiooni (valem (1)), mis hindab vastuste tõenäosust antud latentse tunnuse korral ning eeljaotus $P(\boldsymbol{\theta}_v|\mathbf{y}_v, \Gamma, \Sigma)$ tuleneb latentse regressioonimudeli eeldusest (valem (2)) ja lisab taustmuutujatel põhineva eelinfo.

Vastavast järeljaotusest juhusliku väärtuse genereerimise protseduuri korduvalt rakendades saadakse iga õpilase kohta mitu tõepäraväärtust. Alates 2015. aastast genereeritakse PISA uuringutes iga kognitiivse valdkonna jaoks 10 tõepäraväärtust, mis kajastavad mõõtmisest tulenevat ebakindlust õpilase tegeliku oskustaseme hindamisel.

1.5 Rubini reeglid

Üldkogumi parameetrite hindamiseks arvutatakse huvipakkuv statistik iga tõepäraväärtuse põhjal eraldi ning lõplik hinnang on nende statistikute aritmeetiline keskmine (OECD, 2009). Sellise lähenemiseni jõudis Donald B.Rubin (2009), kes esitas 1978. aastal mitmese imputeerimise teooria puuduvate andmete käsitlemiseks. Seepärast nimetatakse järgnevaid valemeid ka Rubini reegliteks. Järgnevad valemid põhinevad PISA tehnilisel raportil ja PISA SPSS-i andmeanalüüsi juhendil (OECD, 2024; OECD, 2009). Üldkogumi pa-

rameetri T hinnang:

$$\bar{T} = \frac{1}{U} \sum_{u=1}^{U} T_u, \tag{5}$$

kus T_u tähistab uuritava statistiku väärtust ühe tõepäraväärtuse puhul ja U on tõepära väärtuste arv. PISA testis oli kuni 2015 aastani $U = 5$, praegu on $U = 10$.

Mõõtmise hajuvus, imputeerimisest tulenev hajuvus, on leitav valemiga:

$$B_U = \frac{1}{U-1} \sum_{u=1}^{U} \left(T_u - \bar{T}\right)^2. \tag{6}$$

Üldkogumi parameetri dispersiooni hinnangu leidmiseks tuleb leida hinnangu T_u valimi dispersioon iga u korral ning seda tähistatakse $V(T_u)$. Selle abil leiame valimi ebakindluse:

$$W = \frac{1}{U} \sum_{u=1}^{U} V(T_u).$$

Hinnangu T dispersiooni hinnang:

$$V(\bar{T}) = W + \left(1 + \frac{1}{U}\right) B_U. \tag{7}$$

Selline lähenemine pakub praktilist ja teoreetiliselt põhjendatud meetodit puuduvate andmete käsitlemiseks, tagades, et kombineeritud hinnangud ja variatsioonid peegeldavad puuduvatest väärtustest tulenevaid määramatusi (Rubin, 2009).

2 Kaalud

Käesolev peatükk põhineb PISA SPSS-i andmeanalüüsi juhendil, kui ei ole märgitud teisiti (OECD, 2009). Kaalude kasutamine PISA andmetes tuleneb valimi kujundamise ja andmete kogumise eripäradest. Kuna õpilaste ja koolide valikusse sattumise tõenäosused riigis ei ole võrdsed ning osa valimikihte on riikliku aruandluse jaoks valimis üleesindatud. Arvesse võetakse kaalumisel ka osalemismäärade erinevused, mis on seotud koolide ja õpilaste teatud tunnustega (kooli õppekeel, sugu jne) ning nõuavad omakorda mitteosalemise korrigeerimist. Analüüside kõigis etappides, sõltumata sellest, kas tegemist on andmete esialgse uurimise või lõpliku analüüsiga, tuleks alati kasutada kaalusid.

2.1 Kaalud lihtsa juhuvalimi puhul

Lihtsa juhuvalimi moodustamisel valitakse valimi elemendid juhuslikult nii, et kõikidel üldkogumi liikmetel on võrdsed võimalused valimisse kuuluda (Beilmann ja Rämmer, 2025).

Tõenäosus, et i-s liige satub valimisse:

$$p_i = \frac{n}{N},$$

kus n vastab valimi suurusele ja N üldkogumile. Valimi kaalud põhinevad tavaliselt üksuse valikusse kaasamise tõenäosusel, täpsemalt kaal arvutatakse selle tõenäosuse pöördväärtusena:

$$w_i = \frac{1}{p_i} = \frac{N}{n}.$$

2.2 Kaalud PISA uuringus

Lihtsa juhuvalimi moodustamine on kallis ja ajakulukas, seetõttu kasutatakse PISA valimi moodustamiseks kahte sammu. Esimeseks sammuks on koolide nimekirjast valimi moodustamine. Kooli i valimisse sattumise tõenäosus on leitav valemiga:

$$p_{1_i} = \frac{N_i \cdot n_{sc}}{N},$$

kus n_{sc} on valimi suurus (koolide arv valimis), N_i koolis õppivate õpilaste arv ja N on õpilaste arv üldkogumis. Selline meetod on vajalik vältimaks koolide suurusest tulenevat lõpptõenäosuse erinevust. Sellest tulenevalt vastav kaal:

$$w_{1_i} = \frac{1}{p_{1_i}}.$$

Järgnevalt valitakse n_i õpilast lihtsa juhuvalimi abil eelnevalt valituks osutunud koolist i. Õpilase j valimisse sattumise tõenäosus on:

$$p_{2_{ij}} = \frac{n_i}{N_i}$$

ning vastav kaal:

$$w_{2_{ij}} = \frac{1}{p_{2_{ij}}}.$$

Seega lõplik tõenäosus, et õpilane satub valimisse on:

$$p_{ij} = p_{1_i} \cdot p_{2_{ij}}$$

ning lõplik kaal on leitav valemiga:

$$w_{ij} = \frac{1}{p_{ij}}.$$

Parameetrite nihketa hinnangu saamiseks tuleb analüüsis kasutada lõplikke kaale.

2.3 Replikatsioonide kaalud

PISA valimi sõltuvate vaatluste ja keeruka valimidisaini tõttu ei ole tavapärased dispersiooni hindamise meetodid sobivad ning seetõttu kasutatakse replikatsioonide kaale. PISA uuringus konstrueeritakse 80 replikatsioonide kaalu kasutades *BRR* meetodit (ingl *balanced repeated replication method*) koos Fay modifikatsiooniga.

Dispersioon hinnangule $\hat{\theta}$ hinnatakse järgneva valemiga

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G}\sum_{i=1}^{G}(\hat{\theta}_{(i)} - \hat{\theta})^2.$$

Kõigepealt on leitud kogu valimi hinnang $\hat{\theta}$, kasutades lõppkaalu. Seejärel arvutatakse replikatsioonide hinnangud, kus i-nda replikatsiooni hinnang $\hat{\theta}_{(i)}$ saadakse, kui samale valimile rakendatakse vastavaid replikatsioonide kaale. G tähistab replikatsioonide arvu (meil $G = 80$).

Üheks probleemiks selle meetodi juures on olukord, kus $\hat{\theta}_{(i)}$ ei pruugi olla defineeritud, kuigi kogu valimi hinnang $(\hat{\theta})$ on määratud. Sellise olukorra lahenduseks on Fay meetod, mille puhul konstrueeritakse replikatsioonide kaalud nii, et osa kaaludest suurendatakse k-võrra ja teisi vähendatakse $1 - k$ võrra (Judkins, 1990). Kasutatavat faktorit k nimetatakse Fay faktoriks. PISA kasutab Fay faktorit suurusega 0,5. Vastava meetodiga leitav dispersioon esitub kujul:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}(\hat{\theta}_{(i)} - \hat{\theta})^2.$$

Asendame eelnevasse valimisse PISA replikatsioonide arvu $G = 80$ ja $k = 0{,}5$, seega on dispersioon kujul:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{20} \sum_{i=1}^{80} (\hat{\theta}_{(i)} - \hat{\theta})^2.$$

Seega võimaldab replikatsioonikaalude kasutamine hinnata parameetri dispersiooni, rakendades samu arvutusreegleid mis tahes huvipakkuva statistiku puhul.

3 Metoodika

Käesoleva bakalaureusetöö eesmärgi rakenduseks oli arendada korduvkasutatav metoodiline raamistik PISA andmete statistiliseks analüüsimiseks kasutades vabavaralist statistika tarkvara *R* (versioon R 4.5.2). *R*-i realiseerimiseks kasutati *RStudio* töökeskkonda ning raamistik loodi *R Markdown*i HTML vormingus. Raamistik järgib PISA uuringu metoodilisi nõudeid ning loob aluse edasiste analüüsi koodide korrektseks väljatöötamiseks.

3.1 Analüüsi raamistik

Analüüsi raamistiku puhul pöörati erilist tähelepanu tõepäraväärtuste kasutamisele, kompleksvalimi disainile ning replikatsioonikaalude korrektsele käsitlemisele. PISA andmete analüüsi keerukuse tõttu võidakse kasutada praktikas lihtsustatud võtteid (nt ainult ühe tõepäraväärtuse kasutamine ja/või kaalude ignoreerimine), mis võivad viia kallutatud hinnangute või alahinnatud standardvigadeni. Käesolev töö pakub metoodiliselt korrektset ja kasutajasõbralikku lahendust, mis aitab selliseid vigu ennetada ning toimib alusstruktuurina ka keerukamate analüüside jaoks koodi iseseisval loomisel ja analüüsi teostamisel.

Loodud funktsioonid kasutavad süstemaatiliselt kõiki tõepäraväärtusi ja hindavad statistilised mudelid iga tõepäraväärtuse põhjal eraldi. Valimi dispersioonid hinnatakse Fay modifikatsiooniga BRR replikatsioonimeetodi abil ning lõplike hinnangute ja standardvigade saamiseks kombineeritakse tõepäraväärtuste põhjal leitud hinnangud ning dispersioonid Rubini reeglite kohaselt. Abivahendina kasutatakse eelnevalt välja töötatud funkstioone. Selline lähenemine tagab, et analüüs võtab arvesse kogu PISA andmete keerukust.

3.2 Koodi arendamine

Arenduse lähtepunktiks oli PISA metoodiliste juhendite ja analüüsinõuete läbitöötamine, mille käigus kaardistati peamised metoodilised ohukohad: tõepäraväärtuste kasutamine, replikatsioonikaalude rakendamine ning Rubini reeglitega kombineerimine. Ohukohti silmas pidades koostati plaan, millised üldisemad funktsioonid luua, millised funktsioonid teha manuaalselt ja kus kasutada eelnevalt välja töötatud funktsioone pakettidest *intsvy* ja *survey*.

Analüüsiks on olemas OECD loodud pakett *Rrepest*, mida saab rakendada analüüside läbi viimisel. See ei ole mõeldud ainult PISA jaoks, vaid kasutatakse lisaks teiste suurte uuringute (nt TIMSS, PIRLS) analüüsides. Probleemiks on see, et suur osa arvutusprotsessist peidetakse sisemisse abstraktsiooni. Samalaadseid piiranguid esineb ka teistes suuremahuliste analüüside jaoks loodud pakettides (nt *RALSA*). Käesolevas töös eelistati läbipaistvamat lähenemist, kus analüüsi peamised etapid (tõepäraväärtuste ning Rubini reeglite kasutamine) on kasutajale nähtavad ja kontseptuaalselt jälgitavad.

Otsatarbekaks ei osutunud hakata nullist programmeerima Fay modifikatsiooniga BRR replikatsioonimeetodit ega kompleksvalimi disaini arvutusi. Nende protseduuride korrektne matemaatiline teostus on tehniliselt detailne ning vigadele äärmiselt tundlik. Seetõttu kasutati loodud funktsioonides *R*-i paketti *survey*, mis on laialdaselt kasutatud töövahend kompleksvalimi analüüsiks (Lumley, 2024). Pakett toetab nii PISA valimidisaini, ametlikke BRR replikatsioonikaale kui ka Fay modifikatsiooni rakendamist. Paketi *survey* kasutamine tagab metoodilise täpsuse ja kooskõla PISA tehniliste juhenditega.

Paketti *intsvy* kasutati analüüsi ajakulu vähendamiseks ja OECD poolt heaks kiidetud standardlahenduse tutvustamiseks. Täpsemalt saab *intsvy* paketiga tutvuda Caro ja Biecek (2024) loodud juhendis. Selline hübriidne lahendus

ühendab läbipaistvuse ning metoodilise korrektsuse, vältides vajadust dubleerida juba usaldusväärselt rakendatud statistilisi algoritme.

Koodi arendamist raskendas tõepäraväärtuste ja replikatsioonide kaalude korrektne käsitlemine, kuna tulemuste saamiseks tuli õigesti kombineerida kahte tsüklit. Lisaks osutus keerukaks koodi hoidmine piisavalt lihtsa ja arusaadavana. Arendusprotsessi käigus esines süntaksvigu ja kasutatavates pakettides ilmnesid spetsiifilised piirangud. Eraldi väljakutseks kujunes koodi korduv testimine, mis oli ajakulukas. Koodi optimeerimisel ja kirjutamise käigus tekkinud vigade tuvastamisel kasutati tehisintellektil põhinevat abivahendit Microsoft CoPilot (Microsoft 365 Copilot Chat).

3.3 Loodud funktsioonide kirjeldus

Loodud funktsioonide struktuur võimaldab nende rakendamist eri PISA uuringutsüklites, eeldusel et valimidisaini põhimõtted ei muutu. Praktilise rakendatavuse ja reprodutseeritavuse huvides piirdub töö tsüklitega alates 2015. aastast kuna varasemate andmete töötlemine tekstifailide kujul nõuab täiendavat tehnilist eeltööd.

Analüüsi raamistiku alguses on olemas juhend PISA andmete allalaadimiseks ning loetelu tööks vajalikest *R* pakettidest, mis tagab kasutajale selge ülevaate analüüsi ettevalmistavatest sammudest. *R* ja *RStudio* kasutamisega saab lähemalt tutvuda vastavas juhendis (Kolnes, 2020).

3.3.1 Tõepäraväärtuste tuvastamine

Kõigi analüüside aluseks on abifunktsioon `pv_veerud()`, mille ülesandeks on tuvastada andmestikust vastava valdkonna tõepäraväärtuste veerud ning

tagastada need korrektses järjekorras. Automaatne tõepäraväärtuste tuvastamine võimaldab vältida käsitsi veergude määramist ning tagab funktsioonide rakendatavuse erinevatele PISA uuringutsüklitele (juhul kui muutub tõepäraväärtuste arv). Kui sobivaid veerge ei tuvastata, katkestab funktsioon töö veateatega, mis aitab vältida analüüsi läbiviimist valede sisenditega. Funktsiooni ülesehitust ja keskset loogikat illustreerib joonis 3.

```
pv_veerud = function(andmestik, aine_pv) {
# aine_pv on näiteks "MATH", "READ", "SCIE"

# Märgime, millise kujuga on tõepäraväärtused andmestikus
  kuju = paste0("^PV(\\d+)", aine_pv, "$")

# Leiame andmestikust eelnevalt antud kujuga väärtused
  leitud = names(andmestik)[grepl(kuju, names(andmestik))]

# Kui leitud väärtuste pikkus on 0, siis annab veateate
  if (length(leitud) == 0) stop("Tõepära väärtuseid ei leitud vastava aine jaoks: ", aine_pv)

# Leiame veeru nimest numbri ja loome numbritest vektori
  pv_nr = as.integer(sub(kuju, "\\1", leitud))

# Sorteerime veerunimed numbrite järgi kasvavasse järjekorda.
  leitud[order(pv_nr)]
}
```

Joonis 3: Kuvatõmmis funktsiooni tõepäraväärtuste tuvastamise koodist.

3.3.2 Kirjeldav statistika

Kirjeldava statistika arvutamiseks töötati välja kaks funktsiooni, mis mõlemad arvestavad tõepäraväärtuste olemasolu, kuid erinevad teostuse ja eesmärgi poolest. Lisaks loodi visualiseerimiseks funktsioonid (`plot_keskmised()` ja `plot_protsendid()`), mis võimaldavad esitada keskmisi usaldusvahemikega ja osakaale rühmade kaupa. Graafikud põhinevad eelnevalt arvutatud kaalutud hinnangutel. Vaikimisi leiti 95%-usaldusvahemikud, kuid koodis saab lihtsasti määrata muu usaldusnivoo.

Funktsioon `pv_kokkuvote_intsvy()` kasutab paketi *intsvy* sisseehitatud funkt-

siooni. Valitud tõepäraväärtustega tunnuse (näites "MATH") kohta arvutatud näitajad on kuvatud tabelis 1. Tabelis on toodud valimi vaatluste arv (N), kaalutud keskmine, keskmise standardviga (s.e.), standardhälbe (SD), standardviga standardhälbele (s.e) ning alumine ja ülemine usalduspiir. Kasutajal tuleb funktsiooni argumentideks sisestada andmestik, tõepäraväärtustega tunnus.

Tabel 1: Kirjeldav statistika matemaatika tulemusele.

N	keskmine	s.e.	SD	s.e	alumine	ülemine
6392	509.95	1.98	84.95	1.1	506.07	513.83

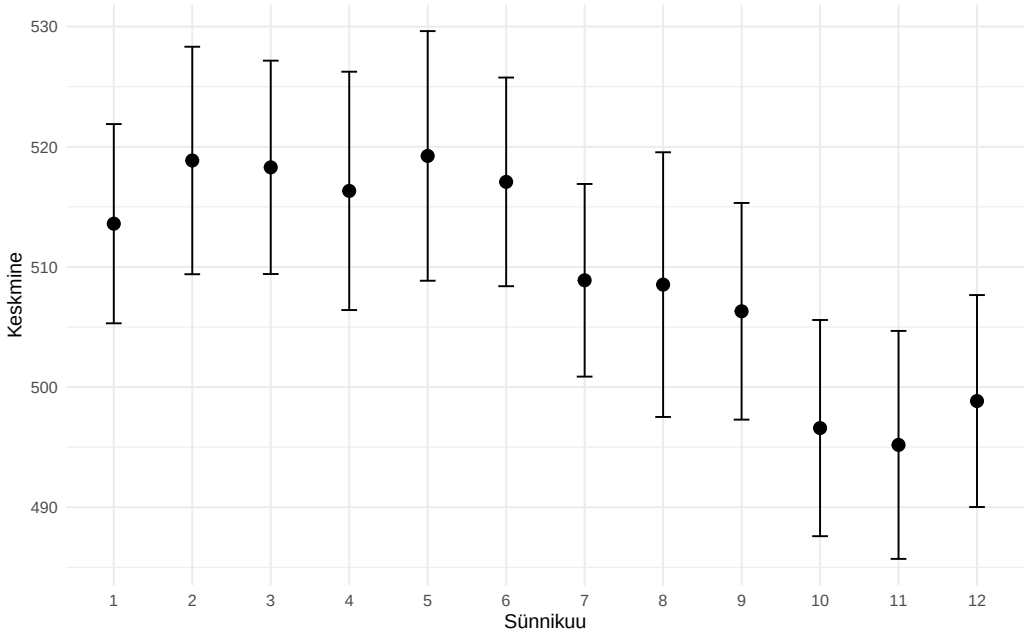
Võimalik on ka saada rühmade lõikes vastavad näitajad, kui lisada sisendisse grupeeriv tunnus. Sel juhul leitakse eelnevale lisaks rühmade osakaalud protsentides ning arvutatakse kahepoolne usaldusvahemik (usaldusnivool $1-\alpha$). Funktsioon on mõeldud praktiliseks kasutamiseks ja võimaldab kiiresti ja optimaalselt saada PISA metoodikale vastavaid kirjeldavaid näitajaid.

Funktsioon `pv_kokkuvote_manuaalne()` on loodud illustratiivsel eesmärgil, et demonstreerida tõepäraväärtuste käsitlemist kirjeldava statistika arvutamisel. Funktsioon arvutab valitud tunnuse vaatluste arvu (N), kaalutud keskmise, standardvea keskmisele (s.e.), standardhälve (SD) ja standardvea standardhälbele (s.e) (vt Tabel 2). Analüüsi on võimalik teostada nii kogu valimi kui ka rühmade lõikes. Funktsiooni sisendiks on andmestik, tõepäraväärtustega tunnus (näites "READ") ja soovi korral grupeeriv tunnus.

Tabel 2: Kirjeldav statistika funktsionaalse lugemisoskuse tulemusele.

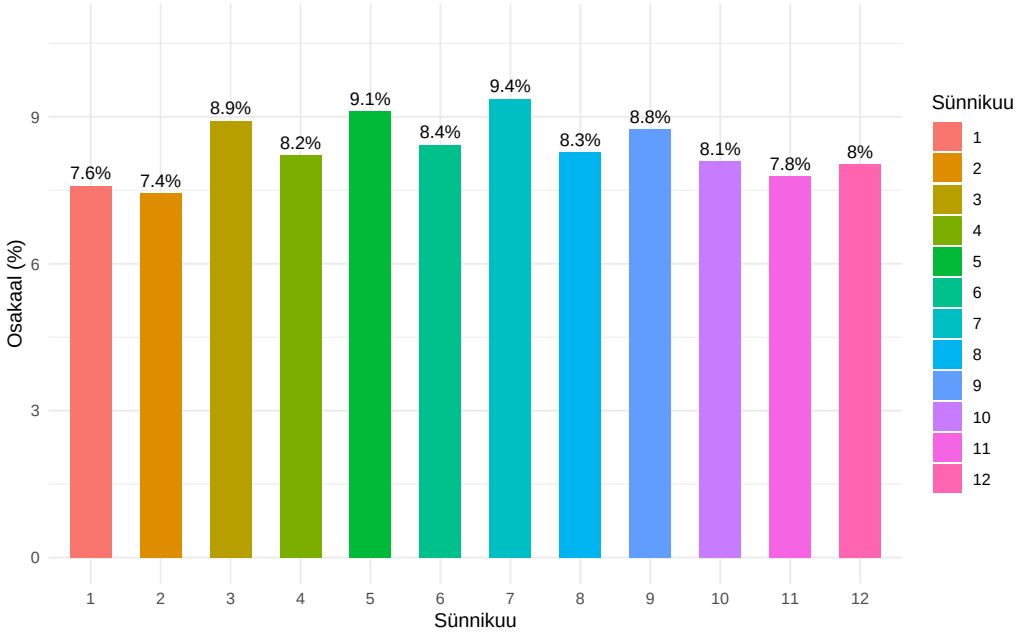
keskmine	s.e.	SD	s.e	N
511.03	2.36	92.48	1.11	6392

Funktsiooni `pv_kokkuvote_intsvy()` keskmiste ja usaldusvahemike visualiseerimiseks loodi funktstioon `plot_keskmised()`, mille väljund on kujutatud joonisel 4. Kuvatud on Eesti õpilaste matemaatika oskuse ("MATH") keskmine tulemus grupeeritult sünnikuu ("ST003D02T") järgi.



Joonis 4: Matemaatika tulemuste keskmine ja usaldusvahemik sõltuvalt sünnikuust.

Funktsioon `plot_protsendid()` visualiseerib grupeeritava tunnuse osakaale. Joonisel 5 on näidatud Eesti õpilaste osakaalud sünnikuu järgi.



Joonis 5: Eesti õpilaste osakaalud sünnikuu järgi.

3.3.3 Logistiline regressioon

Logistilise regressiooni jaoks loodi funktsioon `log_reg()`. Funktsioon põhineb *intsvy* poolt PISA andmetele loodud logistilise regressiooni funktsioonil. Loodud funktsiooni sisendiks on tõepäraväärtustega tunnus, millest konstrueeritakse binaarne sõltuv muutuja kasutaja poolt etteantud lävendi (cutoff) alusel (vt Joonis 6). Lisaks tuleb kasutajal sisestada andmestik, tõepäraväärtustega tunnus ning kirjeldavad tunnused. Lävend mõjutab klasside jaotust, mistõttu on oluline tagada piisav arv vaatlusi mõlemas kategoorias. Soovi korral on võimalik lisada ka grupeeriv tunnus.

```
log_reg = function(andmestik, aine_pv, kirjeldav_tunnus, grupeering= NULL, cutoff){

# Leiame tõepäraväärtuse tunnusele vastad nimed kasutades eelnevalt defineeritud funkstiooni pv_veerud()
pv_nimed = pv_veerud(andmestik, aine_pv)

# Vaatame kas grupeeritav tunnus on antud
   if(is.null(grupeering)){
# Defineerime mudeli
    mudel= pisa.log.pv(
# Määrame regressioonis kasutatavad tõepäraväärtuse veerud
      pvlabel=pv_nimed,

# Määrame seletavad tunnused
      x=kirjeldav_tunnus,
# Määrame andmestiku
      data=andmestik,
# Määrame cutoffi
      cutoff = cutoff)

# Kui grupeeritav tunnus on antud, vaatame logistilist regressiooni rühmiti
  }else{
    mudel= pisa.log.pv(
     pvlabel=pv_nimed,
     x=kirjeldav_tunnus,
# Määrame rühmitava tunnuse
     by = grupeering,
     data=andmestik,
     cutoff = cutoff)
  }
  return(mudel)
}
```

Joonis 6: Ekraanitõmmis logistilise regressiooni koodist.

3.3.4 Korrelatsioon

Korrelatsioonianalüüsi jaoks töötati välja kaks funktsiooni:

- `corr_pv_taust()`, mis hindab korrelatsiooni ühe tõepäraväärtustega tunnuse ja tausttunnuse vahel;
- `corr_pv_pv()`, mis võimaldab hinnata kahe tõepäraväärtustega tunnuse vahelist korrelatsiooni.

Taustunnuse ja tõepäraväärtustega tunnuse vahelise korrelatsiooni leidmiseks loodud funktsioon tagastab kaalutud Pearsoni korrelatsioonikordaja, stan-

dardvea hinnangule ning teststatistiku R^2 (vt Tabel 3). Tabelis 3 on leitud korrelatsioonikordaja matemaatika tulemuse ("MATH") ja sotsiaalmajandusliku tausta ("ESCS") vahel.

Tabel 3: Korrelatsioon matemaatika tulemuse ja sotsiaalmajandusliku tausta vahel.

korrelatsioonikordaja	se	R^2
0.37	0.02	0.13

Kahe tõepäraväärtustega tunnuse vahelise korrelatsiooni leidmiseks loodud funktsioon väljastab kaalutud Pearsoni korrelatsioonikordaja, standardvea ning lisastatistikud nagu teststatistik (z), p-väärtus ja usaldusvahemik (alumine ja ülemine) (vt Tabel 4). Tabelis 4 on leitud korrelatsioon matemaatika tulemuse ("MATH") ja funktsionaalse lugemisoskuse ("READ") vahel.

Tabel 4: Korrelatsioon matemaatika tulemuse ja funktsionaalse lugemisoskuse vahel.

korrelatsioonikordaja	se	p-väärtus	alumine	ülemine
0.77	0.01	< 0.000001	0.76	0.79

Kokkuvõttes pakuvad loodud funktsioonid paindliku võimaluse hinnata korrelatsiooni tunnuste vahel, toetudes PISA andmete jaoks sobivatele analüüsi meetoditele.

3.3.5 Lineaarne regressioon

Lineaarse regressiooni analüüsimiseks loodi kaks funktsiooni:

- `lin_reg_intsvy()`, mis võimaldab hinnata regressioonimudelit, kus sõltuvaks muutujaks on tõepäraväärtustega tunnus ning seletavateks muutujateks tausttunnused;
- `lin_reg_pv()`, mis võimaldab hinnata regressioonimudelit kahe tõepäraväärtustega tunnuse ja tausttunnuste vahel.

Funktsioon `lin_reg_intsvy()` väljastab kasutab paketi *intsvy* sisseehitatud funkstiooni. Funktsioon väljastab mudeli kordajad, vastavad standardvead ja teststatistikud. Sisenditeks on andmestik, tõepäraväärtustega aine, kirjeldavad tunnused ja soovi korral grupeeriv tunnus.

Funktsioon `lin_reg_pv()` kasutab paketi *survey* võimalusi kompleksvalimi disaini arvestamiseks. Funktsioon väljastab mudeli kordajad, vastavad standardvead (se), teststatistikud (t), usaldusvahemikud (alumine ja ülemine) ning kokkuvõtlikud näitajad nagu vaatluste arv (n) ja determinatsioonikordaja (R^2) (vt Tabel 5). Funktsiooni sisenditeks on andmestik, tõepäraväärtustega aine, kirjeldavad tunnused ja soovi korral grupeeriv tunnus. Tabelis 5 on linearse regressiooni muutuvaks tunnuseks matemaatika tulemus ("MATH") ja kirjeldavateks tunnusteks funkstionaalne lugemisoskus ("READ") ja sotsiaalmajanduslik taust ("ESCS").

Tabel 5: Lineaarse regressiooni tulemused matemaatika tulemuse, funktsionaalse lugemisoskuse ja sotsiaalmajandusliku tausta vahel.

tunnused	kordajad	se	t	alumine	ülemine	n	R^2
(Intercept)	164.67	7.76	21.22	149.46	179.88	6291	0.62
XPV	0.67	0.01	45.51	0.64	0.70		
ESCS	15.45	1.47	10.51	12.56	18.33		

3.4 Koodi kontrollimine

Loodud funktsioonide toimimise kontrollimiseks viidi läbi näidisanalüüs OECD PISA 2022 õpilaste andmestiku põhjal (CY08MSP_STU_QQQ.SAV). Analüüs teostati Eesti valimi kohta (CNT == "EST"). Näidisanalüüsi eesmärk oli hinnata, kas bakalaureusetöö raames arendatud funktsioonid annavad tulemusi, mis on kooskõlas OECD poolt soovitatud analüüsiprotseduuriga. Selleks võrreldi saadud hinnanguid paketi *Rrepest* abil arvutatud tulemustega.

Analüüs hõlmas järgmisi etappe. Kirjeldav statistika, mille käigus arvutati kaalutud keskmised, standardvead ja 95% usaldusvahemikud nii kogu valimi kui ka rühmade lõikes. Lineaarne regressioon, mille abil hinnati seoseid tõepäraväärtustega tulemuse ja valitud tausttunnuste vahel. Logistiline regressioon, kus tõepäraväärtuse põhjal konstrueeriti binaarne tunnus etteantud lävendi (500) alusel. Korrelatsioonianalüüs, milles hinnati nii tõepäraväärtuse ja taustmuutuja vahelist kui ka kahe tõepäraväärtustega tunnuse vahelist seost.

Lineaarse regressiooni puhul arvutati täiendavalt kaalutud determinatsioonikordaja R^2, mida käsitletakse käesolevas töös kirjeldava lisainfona, kuna OECD PISA raportites ei ole komplekse valimidisaini korral determinatsioonikordaja peamine sobivusnäitaja. Võrdlustulemuste põhjal võib järeldada, et loodud funktsioonid annavad PISA 2022 Eesti andmetel samaväärsed hinnangud OECD poolt pakutavate *Rrepest* lahendustega. Eriti oluline on kooskõla standardvigade tasemel, kuna need sõltuvad otseselt korrektse replikatsioonidisaini ja tõepäraväärtuste kombineerimise rakendamisest. Kasutatud programmikood on toodud lisades (vt lisa 1).

Seega kinnitab näidisanalüüs, et realiseeritud funktsioonid järgivad PISA metoodilisi juhiseid ning võimaldavad analüüse läbi viia reprodutseeritavalt

ja metoodiliselt korrektselt. Lisaks pakub loodud lahendus abi olukordades, kus valmisfunktsioonid ei kata konkreetset analüüsivajadust ning tuleb luua oma funktsioonid.

Kokkuvõte

Bakalaureusetöö eesmärk oli kirjeldada PISA andmete analüüsi protsessi ning tuua välja analüüsi läbiviimiseks vajalikud meetodid ja põhimõtted, kasutades eestikeelseid mõisteid ja selgitusi. Töö idee kujunes vajadusest käsitleda PISA uuringu keerukat valimidisaini, mille tõttu on andmete korrekktne analüüs metoodiliselt nõudlik ning eksimisvõimalus suur. Kuigi inglisekeelset juhendmaterjali PISA andmete analüüsimiseks leidub, on eestikeelne PISA metoodikat süstemaatiliselt käsitlev juhend analüüsi iseseisvaks läbiviimiseks seni olnud puudulik.

Töö teoreetilises osas anti ülevaade PISA uuringu ülesehitusest ning andmete analüüsimisel kasutatavatest peamistest meetoditest. Erilist tähelepanu pöörati tõepäraväärtuste leidmisele ja korrektsele käsitlemisele, et vältida ebausaldusväärsete tulemuste tekkimist.

Praktilises osas loodi vabavaralist *R* tarkvara kasutades kood, mis võimaldab teostada PISA andmete põhjal lihtsamaid analüüse ning on võimalusel aluseks keerukamate analüüside koostamisel. Käesolevas töös loodud funktsioonid moodustavad tervikliku ja korduvkasutatava analüüsiraamistiku, mis võimaldab teostada PISA andmetel metoodiliselt korrektselt kirjeldavat statistikat, regressiooni- ja korrelatsioonianalüüse.

Lahendus on eelkõige on suunatud tudengitele ja teistele huvilistele, kellel on soov teha PISA andemetel põhinev analüüs. Väljatöötatud raamistik aitab vähendada lihtsustatud analüüsivõtete kasutamisest tulenevate vigade riski ning toetab läbipaistvat arvutusloogikat, mida on võimalik edaspidi laiendada keerukamate analüüside tarbeks.

Edasiste arendustena võiks keskenduda visuaalsete väljundite täiendamisele, keerukamate analüüside jaoks funktsioonide väljatöötamisele ja rohkem

lahti seletada saadud tabelite tulemused, et aidata kasutajal paremini tulemusi interpreteerida. Kokkuvõttes pakkub töö struktureeritud ja praktilist ülevaadet PISA andemete analüüsiprotsessist ning toetab korrektse analüüsi iseseisvat läbiviimist.

Kasutatud allikad

Avvisati, F. ja F. Keslair (2014). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values*. URL: https://ideas.repec.org/c/boc/bocode/s457918.html (vaadatud 07.05.2026).

Beilmann, M. ja A. Rämmer (2025). *Valimi moodustamine*. URL: https://samm.ut.ee/valimi-moodustamine/ (vaadatud 01.05.2026).

Caro, D. ja P. Biecek (2024). *intsvy: International Assessment Data Manager*. R package version 2.9. URL: https://CRAN.R-project.org/package=intsvy (vaadatud 01.05.2026).

Davier, M. von, E. J. Gonzalez ja R. Mislevy (2009). "What are plausible values and why are they useful". Teoses: köide 2. 1, lk. 9–36. URL: https://ierinstitute.org/fileadmin/Documents/IERI_Monograph/Volume_2/IERI_Monograph_Volume_02_Chapter_01.pdf (vaadatud 07.05.2026).

Dempster, A. P., N. M. Laird ja D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x (vaadatud 11.04.2026).

Gonzalez, E. J. ja A. M. Kennedy (2003). *PIRLS 2001 User Guide for the International Database*. International Study Center Lynch School of Education, Boston College. URL: https://timssandpirls.bc.edu/pirls2001i/pdf/UserGuide.pdf (vaadatud 11.04.2026).

Hambleton, R. K. (1990). "Item response theory: introduction and bibliography". *Psicothema* 2.1, lk. 97–107. URL: https://www.redalyc.org/pdf/727/72702108.pdf (vaadatud 11.04.2026).

HARNO (2025). *PISA – noorte teadmiste ja oskuste uuring*. URL: https://harno.ee/pisa (vaadatud 11.04.2026).

IEA (2022). *Help Manual for the IEA IDB Analyzer (Version 5.0)*. Hamburg, Germany. URL: https://www.iea.nl/sites/default/files/2025-01/Manual-for-the-IDB-Analyzer-%28Version%205%200%29.pdf (vaadatud 07.05.2026).

Jewsbury, P. A., Yue J. ja E. J. Gonzalez (2024). "Considerations for the use of plausible values in large-scale assessments". *Large-scale Assessments in Education* 12.1. URL: https://link.springer.com/article/10.1186/s40536-024-00213-y (vaadatud 07.05.2026).

Judkins, D. R. (1990). "Fay's method for variance estimation". *Journal of Official Statistics* 6.3, lk. 223–239. URL: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/fay39s-method-for-variance-estimation.pdf (vaadatud 07.05.2026).

Kolnes, M. (2020). *Sissejuhatus R-i ja RStudio kasutamisse*. URL: https://samm.ut.ee/sissjuhatus-r-i-ja-rstudiosse/ (vaadatud 08.05.2026).

Lumley, T. (2024). *survey: Analysis of Complex Survey Samples*. R package. URL: https://CRAN.R-project.org/package=survey (vaadatud 11.05.2026).

MacDougall, J. (2024). *A User's Guide for JASP*. Versioon 2.0. JASP Team. URL: https://jasp-stats.org/wp-content/uploads/2024/08/MacDougall-Users-Guide-for-JASP-v-2.0.pdf (vaadatud 08.05.2026).

OECD ([a]). *PISA: How to prepare and analyse the PISA database*. URL: https://www.oecd.org/en/about/programmes/pisa/how-to-prepare-and-analyse-the-pisa-database.html (vaadatud 11.04.2026).

OECD ([b]). *PISA: Programme for International Student Assessment*. URL: https://www.oecd.org/en/about/programmes/pisa.html (vaadatud 11.04.2026).

OECD (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. OECD Publishing. URL: https://doi.org/10.1787/9789264056275-en (vaadatud 29.01.2026).

OECD (2017). *PISA 2015 Technical Report*. Paris: OECD. URL: https://www.oecd.org/pisa/data/2015-technical-report/ (vaadatud 05.05.2026).

OECD (2024). *PISA 2022 Technical Report*. URL: https://doi.org/10.1787/01820d6d-en (vaadatud 11.04.2026).

Rubin, D.B. (2009). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780470317365.

Thomas, N (1993). "Asymptotic corrections for multivariate posterior moments with factored likelihood functions". *Journal of Computational and Graphical Statistics* 2.3, lk. 309–322. URL: https://www.tandfonline.com/doi/abs/10.1080/10618600.1993.10474614 (vaadatud 11.04.2026).

Tire, G., H. Puksan, T. Kraav, H. Jukk, I. Henno, K. Lindmann, K. Täht, K. Konstabel, B. Lorenz ja M. Kitsing (2023). *PISA 2022 Eesti tulemused*. URL: https://harno.ee/sites/default/files/documents/2023-12/Pisa_tulemused_2022_veebi.pdf (vaadatud 11.04.2026).

Traat, I. ja N. Lepik (2013). *Bayesi statistika Markovi ahelatega*. URL: https://dspace.ut.ee/server/api/core/bitstreams/10307c60-db4d-49fa-bf2c-ff79e4441669/content (vaadatud 11.04.2026).

Wu, M. (2005). "The role of plausible values in large-scale surveys". *Studies in educational Evaluation* 31.2-3, lk. 114–128. URL: https://doi.org/10.1016/j.stueduc.2005.05.005 (vaadatud 10.05.2026).

Lisa 1. Programmikood

```
library(Rrepest)
library(foreign)
library(dplyr)

setwd("C:\\Lõputöö\\Lõputöö\\STU")

pisa = read.spss("CY08MSP_STU_QQQ.SAV",
use.value.labels=FALSE,to.data.frame=TRUE)

EST = pisa[pisa$CNT=="EST",]

# Kirjeldav statistika
Rrepest(
  data = EST,
  svy  = "PISA2015",
  est  = est(c("mean","std"), "pv@math"),
  n.pvs = 10
)

Rrepest(
  data = EST,
  svy  = "PISA2015",
  est  = est(c("mean","std"), "pv@read"),
  n.pvs = 10
)
```

```
# Rühmade lõikes
Rrepest(
  data = EST,
  svy  = "PISA2015",
  est  = est(c("mean","std"), "pv@math"), by = "ST004D01T",
  n.pvs = 10
)
Rrepest(
  data = EST,
  svy  = "PISA2015",
  est  = est(c("mean","std"), "pv@read"), by = "ST004D01T",
  n.pvs = 10
)

# Korrelatsioon taustmuutujaga corr_pv_taust()
# ESCS
Rrepest(
  data = EST,
  svy = "PISA2015",
  est("corr", c("pv@math", "ESCS")),
  n.pvs = 10
)

# Sugu
Rrepest(
  data = EST,
```

```
  svy = "PISA2015",
  est("corr", c("pv@math", "ST004D01T")),
  n.pvs = 10
)

# Korrelatsioon kahe tõepäraväärtuse vahel corr_pv_pv()
Rrepest(
  data = EST,
  svy = "PISA2015",
  est("corr", c("pv@math", "pv@read")),
  n.pvs = 10
)

# Lineaarne regressioon
Rrepest(
  data = EST,
  svy = "PISA2015",
  est= est(
    statistic = "lm",
    target = "pv@math",
    regressor = c("ESCS", "ST004D01T")),
  n.pvs = 10
)

# Tõepäraväärtustega kirjeldavate tunnustega
Rrepest(
  data = EST,
```

```
  svy = "PISA2015",
  est= est(
    statistic = "lm",
    target = "pv@math",
    regressor = "pv@read"),
  n.pvs = 10
)

Rrepest(
  data = EST,
  svy = "PISA2015",
  est= est(
    statistic = "lm",
    target = "pv@math",
    regressor = c("pv@read", "ESCS", "ST004D01T")),
  n.pvs = 10
)

# Logistiline regressioon
# Muudame et väärtused oleks binaarsed
f.t <- EST %>%
  mutate(
    PV1MATH500 = as.integer(PV1MATH >= 500),
    PV2MATH500 = as.integer(PV2MATH >= 500),
    PV3MATH500 = as.integer(PV3MATH >= 500),
    PV4MATH500 = as.integer(PV4MATH >= 500),
    PV5MATH500 = as.integer(PV5MATH >= 500),
```

```
    PV6MATH500 = as.integer(PV6MATH >= 500),
    PV7MATH500 = as.integer(PV7MATH >= 500),
    PV8MATH500 = as.integer(PV8MATH >= 500),
    PV9MATH500 = as.integer(PV9MATH >= 500),
    PV10MATH500 = as.integer(PV10MATH >= 500)
  )

Rrepest(data = f.t,
        svy = "PISA2015",
        est = est(statistic = "log",
        target = "PV@MATH500",
        regressor = "ESCS"),
        n.pvs = 10)
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kärttu Põrk,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose PISA andmeanalüüsi metoodika tutvustamine ja selle rakendamine, mille juhendaja on Hannes Jukk, 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.

Kärttu Põrk
11.05.2026