Andmete kaalumine

Indrek Soidla

2025-01-20

Sisukord

[Disainikaalude põhimõte ja vajalikkus 1](#_Toc188271857)

[Disainikaalude arvutamine 2](#_Toc188271858)

[Millal on tarvis disainikaale kasutada? 3](#_Toc188271859)

[Kao kompenseerimine 4](#_Toc188271860)

[Järelkihistamine 6](#_Toc188271861)

[Andmete kaalumine ei ole imerohi 6](#_Toc188271862)

[Andmete kaalumine kirjeldavas analüüsis 8](#_Toc188271863)

[Andmete kaalumine järeldavas analüüsis 9](#_Toc188271864)

[Viited 15](#_Toc188271865)

Laias laastus on küsitlusandmete kaalumise eesmärk arvutada võimalikult täpsed hinnangud (üldisemalt öeldes mingid statistilised näitajad, olgu need aritmeetilised keskmised või protsentnäitajad või muud keerulisemad näitajad) populatsiooni kohta. Kaalumisel kaalutakse iga indiviidi n-ö üles või alla, nii et ta esindaks kas suuremat või väiksemat hulka populatsiooni liikmetest, vastavalt sellele, kas grupp, mida see indiviid esindab, on küsitlusandmetes ala- või üleesindatud. Teisiti sõnastades on kaalumise eesmärk tõsta küsitlusandmete esinduslikkust populatsiooni suhtes.

Eelnevast nähtub, et kuna esinduslikkust on tarvis tõsta, siis peitub siin varjatud eeldus, et küsitlusandmed ei esinda populatsiooni piisavalt täpselt, st vastajate kogum ei ole täpne vähendatud koopia populatsioonist. Täpsemalt öeldes, kui uurida erinevate vastajaid kirjeldavate tunnuste jaotust vastajate seas ja populatsioonis (olgu need siis vanus, sugu, haridustase vms), siis praktikas esineb ka parimat praktikat järgides võetud valimi korral nendes jaotustes suuremaid või väiksemaid erinevusi. See on praktikas paratamatu, sel on erinevaid põhjuseid ja seetõttu on ka erinevaid võtteid ja meetodeid andmete kaalumiseks. Käesolevas peatükis vaatlemegi neist mõningaid, et selgitada, miks ja mis juhtudel on tarvis andmeid kaaluda ja kuidas seda teha.

## Disainikaalude põhimõte ja vajalikkus

Vabalt kasutada olevates küsitlusandmestikes on reeglina kaalutunnused, mille abil andmeid kaaluda, juba olemas, st kaale ei tule ise arvutada. Sellegipoolest eeldab ka olemasolevate kaalude kasutamine reeglina kursis olemist sellega, kuidas need andmed on kogutud, sh kuidas on koostatud küsitluse aluseks olnud valim.

Näiteks lihtsa juhuvalimi puhul on kõigil populatsiooni liikmetel on võrdne tõenäosus valimisse sattuda. Kiht-, klaster- või mitmeastmelise valimi puhul võivad need tõenäosused – neid nimetatakse **kaasamistõenäosusteks** – aga erineda. Kihtvalimi puhul jaotatakse populatsioon valitud tunnus(t)e alusel kihtideks ja tehakse juhuvalik iga kihi sees eraldi, see tõstab valimi esinduslikkust populatsiooni suhtes. Proportsionaalse kihtvalimi puhul on kihtide proportsioonid valimis samad, mis populatsioonis. Vahel võidakse aga otsustada mitteproportsionaalse kihtvalimi kasuks, kus mõne grupi proportsioon valimis võib olla suurem või väiksem kui populatsioonis. Näiteks Eesti Integratsiooni Monitooringu puhul kasutatakse valimit, kus eestlasi ja mitte-eestlasi on sama palju, olgugi, et Eesti rahvastikus ehk populatsioonis on mitte-eestlasi oluliselt vähem. Selle eesmärk on, võrreldes küsitlusandmete põhjal neid kaht gruppi, teha mitte-eestlaste kohta samavõrd täpseid järeldusi kui eestlaste kohta. Kui aga soovitakse neid kaht gruppi mitte analüüsida võrdlevalt, vaid koos (nt mitte arvutada mingi tunnuse aritmeetilised keskmised eestlaste ja mitte-eestlaste kohta eraldi, vaid hinnata selle tunnuse keskmist Eesti rahvastikus), on tarvis andmeid kaaluda, et saada mitteproportsionaalse kihtvalimi andmed proportsionaalseks populatsiooniga. Sellisel juhul kasutatakse kaale, mida nimetatakse disainikaaludeks. **Disainikaalude** eesmärk ongi tõsta küsitlusandmete esinduslikkust populatsiooni suhtes, võttes arvesse valimiliikmete erinevaid kaasamistõenäosusi. Disainikaalud arvutatakse nii, et iga indiviidi disainikaalu väärtus on pöördvõrdeline selle indiviidi kaasamistõenäosusega valimisse. Selgitame seda konkreetse näite varal lähemalt järgmises osas.

Samamoodi on reeglina tarvis andmeid disainikaalude abil kaaluda, kui andmed on kogutud klaster- või mitmeastmelise valimi abil. Mitmeastmelise valimitüübi puhul tehakse valik mitmel tasandil, näiteks õpilaste küsitlemise puhul lihtsustab küsitlustööd valim, kus esmalt valitakse koolid (esmane valikutasand on kool), seejärel koolidest klassid (teisene valikutasand) ja klassidest kas kõik õpilased või osad neist (kolmas valikutasand). Selliselt valimi koostamine küll lihtsustab andmekogumist ja nõuab vähem ressursse, kuid eri koolide ja klasside õpilasi tervikuna uurides tuleb arvestada, et õpilaste kaasamistõenäosused erinevad – teisiti öeldes, eri õpilaste tõenäosus valimisse sattuda on erinev, sõltuvalt sellest, kui palju on koolis klasse ja kui palju on klassis õpilasi. Et teha kogu õpilaskonna ehk populatsiooni suhtes esinduslikke järeldusi, tuleb andmeid kaaluda disainikaaludega, nii et õpilased, kellel oli väiksem võimalus valimisse sattuda, omandaksid analüüsis suurema kaaluga (n-ö esindaksid rohkem õpilasi), ja õpilased, kellel oli suurem võimalus valimisse sattuda, omandaksid väiksema kaalu.

## Disainikaalude arvutamine

Selgitame disainikaalude arvutamist mitmeastmelise valimi puhul Dillmani jt (2014:87-88) näite varal. Uuringus, mille populatsiooniks on teatud regiooni elanikud, koostatakse kaheastmeline valim: esmasel valikutasandil valitakse 10000 leibkonna seast juhuslikkuse alusel 1000 leibkonda, teisel valikutasandil valitakse igast leibkonnast juhuslikkuse alusel üks leibkonnaliige. Esimesel valikutasandil on iga leibkonna kaasamistõenäosus ehk tõenäosus sattuda valimisse 1/10 ehk 0.1. Iga inimese kaasamistõenäosus seevastu sõltub sellest, kui palju on tema leibkonnas liikmeid: kaheliikmelises leibkonnas on iga liikme tõenäosus sattuda valimisse 1/10 (leibkonna kaasamistõenäosus) \* 1/2 (leibkonnaliikme kaasamistõenäosus) = 1/20 ehk 0.05. Sellise indiviidi disainikaalu väärtus on pöördvõrdeline tema kaasamistõenäosusega ehk 1 / 0.05 = 20. Viieliikmelises leibkonnas on iga liikme kaasamistõenäosus 1/10 \* 1/5 = 1/50 ehk 0.02; sellise indiviidi disainikaal on 1 / 0.02 = 50.

Tabel 1. Disainikaalude arvutamine kaasamistõenäosuste põhjal mitmeastmelise valimi näitel (Dillman, Smyth, ja Christian 2014: 88).

| Indiviid, kelle leibkonnas on… | Leibkonna kaasamistõenäosus | Indiviidi kaasamistõenäosus leibkonnast | Indiviidi kaasamistõenäosus populatsioonist | Indiviidi disainikaal |
| --- | --- | --- | --- | --- |
| üks liige | 1/10 | 1/1 | 1/10 | 10 |
| kaks liiget | 1/10 | 1/2 | 1/20 | 20 |
| kolm liiget | 1/10 | 1/3 | 1/30 | 30 |
| neli liiget | 1/10 | 1/4 | 1/40 | 40 |
| viis liiget | 1/10 | 1/5 | 1/50 | 50 |

Tabelis 1 arvutatud disainikaalud näitavad, kui mitut populatsiooni liiget mingi valimiliige esindab. See tähendab, et kui arvutaksime selles näites kõigile 1000-le valimiliikmele disainikaalud, siis nende summa oleks sama, mis inimeste arv selles regioonis ehk populatsioonis. Praktikas võime andmestikes seevastu kohata disainikaale, kus suur osa kaalude väärtustest jääb 1 ligidale. Sellisel juhul on eelneva näite kohaselt saadud disainikaalud (ehk kaasamistõenäosuste pöördväärtused) jagatud läbi kaalude aritmeetilise keskmisega. Selle tulemusena on disainikaalude aritmeetiline keskmine 1 ning kaalude summa võrdub indiviidide arvuga valimis. Kasutades selliselt arvutatud kaale analüüsis, on kaalutud indiviidide arv võrdne mitte indiviidide arvuga populatsioonis, vaid indiviidide arvuga valimis. Sellisel kujul on disainikaalud ka näiteks Euroopa Sotsiaaluuringu andmestikes.

## Millal on tarvis disainikaale kasutada?

Disainikaale on reeglina tarvis kasutada, kui andmed on kogutud tõenäosusliku valimi alusel, kasutades keerukamat valikutüüpi kui lihtne juhuvalim (st valik tehakse kihistatult või klastrite alusel). “Reeglina” ei tähenda siin seda, et kaalumise vallas ennast mitte eriti kindlalt tundes saab otsustada jätta andmed kaalumata, lootes, et ehk on mul tegu ühega neist ülejäänud, ebareeglipärastest juhtumitest. Otsus andmeid kaaluda või kaalumata jätta peaks põhinema konkreetsetel kaalutlustel, võttes lisaks valimi koostamise eripäradele arvesse, milline on analüüsiühik ehk mis on konkreetses olukorras kogum, millele soovime analüüsitulemusi üldistada.

Disainikaalude arvutamise näites, kus valik tehti esmalt leibkondade vahel ja seejärel leibkonnasiseselt, on andmeid kindlasti tarvis disainikaaludega kaaluda, kui soovime teha üldistusi selle regiooni *elanike* kohta (st kogum, millele soovime tulemusi üldistada on regiooni elanikkond, mitte regiooni leibkonnad). Kui jätaksime sel juhul andmed disainikaaludega kaalumata, mõjutaksid väiksematest leibkondadest pärit vastajad tulemusi ebaproportsionaalselt rohkem. Kui aga soovime teha järeldusi regiooni *leibkondade* kohta – näiteks küsiti küsitluses andmeid ka leibkondade kohta nagu leibkonna sissetulek –, siis pole andmeid disainikaaludega kaaluda vaja, sest valimi koostamisel oli esimesel valikutasandil kõigil leibkondadel samasugune tõenäosus valimisse sattuda (sest esimesel valikutasandil oli tegu ainult lihtsa juhuvalikuga). Esimeses, Integratsiooni Monitooringu näites on vaja andmeid kaaluda, kui analüüsime eestlaste ja mitte-eestlaste ehk erinevate mitteproportsionaalsete kihtide andmeid koos. Küll aga pole otsest vajadust andmeid kaaluda, kui analüüsime eestlasi ja mitte-eestlasi võrdlevalt, st erinevate kihtide andmeid eraldi.

## Kao kompenseerimine

Disainikaalud võimaldavad elimineerida valimivõtu eripäradest tulenevat esinduslikkuse kadu. Samas esineb valikküsitluste puhul teisigi tegureid, mis võivad esinduslikkust populatsiooni suhtes vähendada, mille tõttu disainikaaludest ei pruugi kaalumisel piisata. Üks neist teguritest on **mittevastamine** ehk valimiliikmete kadu – kaugeltki mitte kõik valimiliikmed ei soovi või ei saa küsitluses osaleda.

Kui vastamismäär on valimi eri gruppides samasugune, siis ei vähenda mittevastamine veel andmete esinduslikkust – vastajate hulk on sel juhul küll mittevastamise tõttu väiksem ja seepärast ei pruugi olla võimalik nõrgemaid populatsioonis esinevaid tunnustevahelisi seoseid tuvastada, kuid andmete esinduslikkust see veel ei mõjuta. Reeglina ei ole siiski vastamismäär valimi eri gruppides ühetaoline, vaid erineb teatud määral. Kui vastamismäär eri gruppides, st mingite tunnuste lõikes, erineb (näiteks on vastamismäär nooremates vanusegruppides madalam), saab kannatada esinduslikkus nende tunnuste suhtes ja lisaks selliste tunnuste suhtes, mis on omakorda nende tunnustega seotud. Sellest järeldub, et antud vanuse näites ei tekiks esinduslikkuse osas probleemi näiteks juhul, kui mõõdaksime küsitlusega mingeid väärtushinnanguid, mis ei sõltu vanusest. Üldiselt ei ole see aga realistlik – väärtushinnangud, olgu need poliitilised, ühiskondlikud, religioossed või muud, ikkagi teatud määral vanuse lõikes varieeruvad. Kui vastamismäär valimi eri gruppides varieerub, ei saa reeglina eeldada, et see tulemuste esinduslikkust ei mõjuta, mistõttu võib olla vaja kaalumisel lisaks arvesse võtta mittevastamisest tulenevat esinduslikkuse kadu.

Selle jaoks on erinevaid võimalusi. Üks võimalus mittevastamise ehk **kao kompenseerimiseks** on leida tunnused, mille lõikes vastamismäär erineb ja mis on seotud muude tunnustega andmestikus, ja nende gruppides leida vastamistõenäosused ehk arvutada igas grupis

* respondentide (ehk uuringus osalenud valimiliikmete) disainikaalude summa,
* valimiliikmete (ehk uuringus osalenud ja mitteosalenud liikmete) disainikaalude summa,
* ja leida nende summade suhe ehk jagatis.

Iga respondendi disainikaal tuleb seejärel läbi korrutada eelnevalt kirjeldatud jagatise pöördväärtusega. (Lihtsa juhuvalimi puhul, kus kõigi valimiliikmete kaasamistõenäosused valimisse on samad, saab nimetatud jagatise asemel kasutada lihtsalt vastamismäära grupis.) Tabeli 2 näites tähendaks see, et kõigi 15-24-aastaste respondentide disainikaal tuleks läbi korrutada 1.622-ga.

Tabel 2. Disainikaalude kalibreerimine vastamistõenäosuste alusel (Lohr 2019: 341).

|  | 15-24 | 25-34 | 35-44 | 45-64 | 65+ | Kokku |
| --- | --- | --- | --- | --- | --- | --- |
| Valimimaht | 202 | 220 | 180 | 195 | 203 | 1000 |
| Vastajate arv | 124 | 187 | 162 | 187 | 203 | 863 |
| Disainikaalude summa valimis | 30322 | 33013 | 27046 | 29272 | 30451 | 150104 |
| Disainikaalude summa vastajate seas | 18693 | 28143 | 24371 | 28138 | 30451 |  |
| Vastamistõenäosus | 0.6165 | 0.8525 | 0.9011 | 0.9613 | 1.0000 |  |
| Vastamistõenäosuse pöördväärtus | 1.622 | 1.173 | 1.110 | 1.040 | 1.000 |  |
|  |  |  |  |  |  |  |

Eelnevalt kirjeldatud meetodil saame kohandada disainikaale nii, et andmete kaalumine vähendaks ka mittevastamisest tulenevat esinduslikkuse kadu (inglise keeles kutsutakse seda meetodit *weighting class adjustment*). Selle eeldus on, et gruppides, mille kohta vastamistõenäosused leiti, vastamistõenäosus ei varieeru. See tähendab, et kui disainikaalude kohandamiseks kasutatakse ainult vanusegrupiti arvutatud vastamistõenäosusi, siis vanusegrupi sees ei tohiks erineda näiteks meeste ja naiste vastamistõenäosused või vastamistõenäosused haridustasemeti vms. See on tugev eeldus, mida püütakse rahuldada vastamistõenäosuste arvutamisega rohkem kui ühe tunnuse lõikes, st rohkemates gruppides. Siiski ei saa gruppe olla ülemäära palju, sest osadesse gruppidesse jääks siis liiga vähe vastajaid, mille tõttu võivad osade indiviidide kaalud olla väga suured, mis omakorda muudaks kaalutud analüüsitulemused n-ö ebastabiilseks.

Silmas tuleb pidada ka seda, et grupikuuluvus peab olema teada kõigi valimiliikmete kohta, st ka nende kohta, kes olid valimis, aga ei osalenud uuringus. See piirab kohandamiseks kasutatavate tunnuste hulka oluliselt, praktikas on nende tunnuste hulk piiratud valikuraamis olevate andmetega.

## Järelkihistamine

Eelnevalt kirjeldatud kao kompenseerimise meetod ongi lihtsasti rakendatav juhul, kui meil on ligipääs andmetele ka mittevastanute kohta (st kogu valimi andmetele), enamasti on see ligipääs vaid uuringu korraldajatel. Kui kasutada on ainult küsitlusandmed ehk andmed küsitletute kohta, on lihtsam kohandada disainikaale **järelkihistamise** meetodil. See ei eelda andmete olemasolu mittevastanute kohta, küll aga eeldab järelkihistamise aluseks olevate (reeglina sotsiaaldemograafiliste) tunnuste ühisjaotuste olemasolu vastanute ja populatsiooni kohta. Vastanute kohta on võimalik need ühisjaotused arvutada küsitlusandmete põhjal, populatsiooni kohta võivad (sõltuvalt populatsiooni ja tunnuste spetsiifilisusest) ühisjaotused leiduda muudes allikates, nt riiklikus statistikas. Järelkihistamise meetodil disainikaalude kohandamine toimub analoogselt eelnevalt kirjeldatud kao kompenseerimisega, ainuke erinevus on see, et arvutustes kasutatakse *valimiliikmete* disainikaalude summa asemel indiviidide arvu vastavas grupis *populatsioonis.* See muidugi eeldab, et respondentide disainikaalude summa põhineb disainikaaludel, mis ei ole läbi jagatud disainikaalude keskmisega, nagu küsitlusandmestikes võib disainikaalude puhul olla tehtud.

Erinevalt eelnevalt kirjeldatud kao kompenseerimise meetodist võimaldab järelkihistamine kompenseerida nii mittevastamisest kui kaetuse veast tulenevat esinduslikkuse kadu. Kaetuse viga tuleneb sellest, et valikuraam ehk loend populatsiooni liikmetest, mida kasutatakse valimi võtmisel, ei kata täpselt populatsiooni (näiteks võib valikuraamis esineda isikuid, kes on nüüdseks surnud või võivad mõne isiku andmed esineda topelt).

## Andmete kaalumine ei ole imerohi

Enne kaalumise praktilise osa juurde minekut väärib tähelepanu veel mõni tõsiasi. Kuigi kaalumine üldjuhul võimaldab tõsta andmete esinduslikkust ja saada populatsiooni kohta täpsemad hinnangud, ei saa kaalumist pidada mingiks imerohuks, mis muudab andmed täiesti esinduslikuks populatsiooni suhtes ja kõrvaldab andmetest kõik vead. Eelkõige käib see just kaalumismeetodite kohta, mis püüavad lahendada mittevastamisest tulenevaid probleeme (nagu eelnevalt kirjeldatud kao kompenseerimine ja järelkihistamine). Miks on see hoiatus oluline?

* Nagu eelnevalt kirjeldatud, on neil kaalumismeetoditel üsna tugevad eeldused, mida täielikult rahuldada on keeruline, kui mitte võimatu.
* Pealegi on nende meetodite rakendamises alati teatud määral subjektiivsust, näiteks kaalumise aluseks olevate tunnuste valikul.
* Mil määral disainikaalude kohandamise meetodid mittevastamisest ja muudest uuringuvigadest tulenevaid probleeme konkreetsel juhul lahendavad ja andmete esinduslikkust parandavad, sõltub paljuski sellest, kas tegu on juhuslike või süstemaatiliste vigadega (vt lähemalt uuringu koguvea kontseptsiooni kohta). Mittevastamise puhul tähendab see laias laastus seda, kas uuringus osalenud valimiliikmete kogum (respondendid) ja mitteosalenud valimiliikmete kogum (mittevastanud) on mõõdetavate tunnuste poolest samasugused või mitte.
  + Kui on, töötavad disainikaalude kohandamise meetodid hästi (õieti pole neid otseselt vaja rakendadagi, sest sel juhul on vastanute kogum juba esinduslik ka populatsiooni suhtes). See on paraku ideaaljuhtum, mida esineb harva.
  + Kui ei ole samasugused, sõltub kaalumise edukus sellest, kuivõrd on need erinevused seletatavad kaalumise aluseks olevate tunnustega. Näiteks võib juhtuda, et vastanud ja mittevastanud erinevad mingite mõõdetud väärtushinnangute osas. Kui kohandame sellisel juhul disainikaale, kompenseerides mittevastamist näiteks vanuse, soo ja elukoha alusel, aga mõõdetud väärtushinnangud on nimetatud tunnustega seotud ainult nõrgalt, ei paranda see väärtushinnangute analüüsil saadavate tulemuste esinduslikkust kuigivõrd. Ka juhul, kui kaalumise aluseks olevad tunnused on mõõdetavate tunnustega seotud, ei ole need seosed üks-ühesed; kõik mõõdetavad tunnused ei ole kaalumise aluseks olevate tunnustega ühtmoodi ja samavõrd tugevalt seotud, osad mõõdetavad tunnused ei pruugi nendega üldse olla seotud.
* Kui tegu on süstemaatiliste vigadega, võib andmete kaalumine teatud juhtudel tulemuste esinduslikkust isegi halvendada. Näiteks on tänapäeval telefoniküsitlustes raskem vastama saada nooremaealisi. Võib eeldada, et need noored, kes võõrale numbrile vastavad, on üldiselt usaldavamad kui nende eakaaslased üldiselt. Seega, kui uuringuküsimustega mõõdetakse usaldustaset ühiskonnas, on küsitletute seas alaesindatud madalama usaldustasemega vastajad ning kaalumata tulemused näitavad usaldustaset kõrgemana kui see tegelikult on. Kui sel juhul andmeid veel vanuse alusel kaaluda, st üles kaaluda nooremaid vastajaid, kes on niigi kõrgema usaldustasemega, näitaksid analüüsitulemused usaldustaset ekslikult veelgi kõrgemana. Seega on disainikaalude kohandamise jaoks tunnuste valikul (ja kohandatud kaalude kasutamisel) oluline arvesse võtta mitte ainult seda, millised tunnused on seotud mittevastamise ja teiste mõõdetavate tunnustega, vaid ka seda, millised need seosed on.
* Kaalud arvutatakse kompenseerima n-ö objekti kadu ehk nende valimiliikmete kadu, kes uuringus üldse ei osalenud. Samas esineb küsitlusandmetes ka küsimusele mittevastamist, kus muidu koostööaldis respondent ei soovi või ei oska mingile konkreetsele ankeediküsimusele vastust anda. Olenevalt küsimusest võib selline andmete kadu olla märkimisväärne – sellist kadu kaalumine ei kompenseeri (kui soovime, et kompenseeriks, tuleks arvutada uued, konkreetse küsimuse mittevastamist arvestavad kaalud).

Seega tuleks igasugustesse väidetesse, mis käsitlevad kaalutud andmeid kui esinduslike ja kallutamata tulemuste garantiid, suhtuda eluterve skepsisega. Üldjuhul kehtib reegel, et kehva või olematu metoodikaga kogutud andmete puhul ei aita ka kaalumine esinduslikkust parandada. Usaldusväärse uuringuraporti puhul peaks kaalumise läbipaistvuse tagamiseks olema välja toodud, kas ja kuidas andmeid kaaluti, kuidas olid kaalud koostatud, sh milliseid tunnuseid kaalude arvutamisel kasutati ja millistest allikatest saadi andmed populatsiooni kohta (nt järelkihistamise jaoks) (AAPOR 2025).

## Andmete kaalumine kirjeldavas analüüsis

Vaatame Euroopa Sotsiaaluuringu (ESS) näitel, kuidas kaaluda andmeid R-s. ESS-i andmestikes on disainikaalud tunnuses dweight, n-ö järelkihistuskaalud ehk järelkihistamise alusel kohandatud disainikaalud on tunnuses pspwght. ESS-i andmeid analüüsides soovitatakse kasutada kaale, mis on järelkihistamisega kohandatud, sest need võimaldavad korrigeerida mittevastamisest jm vigadest tulenevaid nihkeid (European Social Survey 2023). Järgnevates näidetes ongi kasutatud kaalumiseks tunnust pspwght.

Avame R-s ESS-i 10. küsitluslaine ehk 2020/21. aasta andmestiku, mis on alla laetud [ESS-i andmeportaalist](https://ess.sikt.no/en/). Võtame näite aluseks Kreeka andmed. ESS-i andmeportaali [10. küsitluslaine leheküljelt](https://ess.sikt.no/en/study/172ac431-2a06-41df-9dab-c1fd8f3877e7) (*Country documentation* => *Greece*) leiame taustainfo, et Kreekas kasutati kolmeastmelist valimit (esimesel astmel valikuühikuks rahvaloenduse piirkond, teisel astmel sisuliselt leibkond, kolmandal leibkonnaliige), kus esimesel astmel kasutati ka kihistamist.

library(haven)  
  
ess10 <- read\_sav("data/ess10.sav")  
  
library(tidyverse)  
  
gr10 <- ess10 |>   
 filter(cntry == "GR")

Kui soovime arvutada kirjeldavaid kaalutud näitajaid, st punkthinnanguid nagu kaalutud aritmeetiline keskmine, kaalutud standardhälve või tunnuse kaalutud jaotus, siis saame selleks kasutada R-i sisseehitatud pakettide (n-ö *base* R), paketi dplyr või muude pakettide võimalusi. Näiteks tunnuse netusoft jaotuse saame funktsiooniga dplyr::count. Kui soovime kaalutud jaotust, tuleb käsku lisada argument wt, mille väärtuseks on kaalutunnus. Siinses näites arvutatakse tunnuse netusoft, mis mõõdab vastaja internetikasutuse sagedust, kaalutud jaotus; funktsiooni mutate abil arvutatakse ka suhteline kaalutud jaotus ehk protsentnäitajad. Pakett dplyr on osa tidyverse-st, mille eelnevalt laadisime library-käsuga mällu, seetõttu pole tarvis dplyr’it eraldi laadida.

gr10 |>   
 count(netusoft, wt = pspwght) |>   
 mutate(protsent = n / sum(n) \* 100)

## # A tibble: 6 × 3  
## netusoft n protsent  
## <dbl+lbl> <dbl> <dbl>  
## 1 1 [Never] 514. 18.4   
## 2 2 [Only occasionally] 117. 4.18   
## 3 3 [A few times a week] 177. 6.31   
## 4 4 [Most days] 309. 11.0   
## 5 5 [Every day] 1676. 59.9   
## 6 NA 5.55 0.198

Funktsioonil mean, millega saab R-s arvutada aritmeetilise keskmise, seevastu pole kaalumiseks eraldi argumenti (käivita funktsiooni nimi koos küsimärgiga nime ees, et näha, milliseid argumente funktsioon võimaldab kasutada). R-i põhipaketis on kaalutud keskmise arvutamiseks eraldi funktsioon weighted.mean, millel on andmete kaalumiseks argument w. Järgnevas näites arvutatakse tervishoiusüsteemiga rahulolu (tunnus stfhlth) kaalutud keskmine, võttes abiks funktsiooni dplyr::summarise.

gr10 |>   
 summarise(stfhlth\_kesk = weighted.mean(stfhlth, w = pspwght, na.rm = TRUE))

## # A tibble: 1 × 1  
## stfhlth\_kesk  
## <dbl>  
## 1 4.43

Mitmeid erinevaid jaotusparameetreid saab arvutada näiteks paketis TAM olevate funktsioonide abil (vt lähemalt käsuga ?TAM::weighted\_mean)

library(TAM)  
gr10 |>   
 summarise(stfhlth\_kesk = weighted\_mean(stfhlth, w = pspwght),  
 stfhlth\_sthälve = weighted\_sd(stfhlth, w = pspwght),  
 stfhlth\_asümmeetria = weighted\_skewness(stfhlth, w = pspwght),  
 stfhlth\_järskus = weighted\_kurtosis(stfhlth, w = pspwght))

## # A tibble: 1 × 4  
## stfhlth\_kesk stfhlth\_sthälve stfhlth\_asümmeetria stfhlth\_järskus  
## <dbl> <dbl> <dbl> <dbl>  
## 1 4.43 2.31 -0.00704 -0.989

## Andmete kaalumine järeldavas analüüsis

Eelnevad andmete kaalumise näited olid punkthinnangute kohta. Tulenevalt asjaolust, et analüüsime valimi andmeid, mitte kogu populatsiooni, ei pruugi siiski punkthinnang populatsiooni täpselt kirjeldada (ka kaalumisest hoolimata). Ka parimate valimi koostamise ja kaalumise praktikate puhul esineb enamasti teatud viga selle tõttu, et juhuslikkuse alusel võetud valim võib populatsioonist mõnevõrra erineda – indiviidide jaotus (nt sotsiaaldemografiliste või mis tahes muude tunnuste alusel) võib valimis teatud määral erineda jaotusest populatsioonis. Et saada populatsiooni kohta täpsemad hinnangud, peaksime kasutama lisaks kirjeldavatele näitajatele ka **järeldavat analüüsi**, näiteks arvutama lisaks punkthinnangutele ka vahemikhinnangud (usalduspiirid) ja seoste või erinevuste esinemise hindamiseks populatsioonis läbi viima statistilisi teste. R-s on ka erinevates pakettides funktsioone statistiliste testide jaoks, mis võimaldavad andmeid kaalutunnuse alusel kaaluda, nt weights::wtd.t.test kaalutud *t*-testi jaoks. Selline, ainult kaalutunnuse arvestamine andmete kaalumisel, on järeldavas analüüsis adekvaatne lähenemine juhul, kui andmed on saadud lihtsa juhuvaliku alusel. Kui aga on kasutatud keerukamat tõenäosuslikku valimit (ja tihtilugu on küsitlusuuringutes seda tehtud), ei pruugi lihtsalt kaalutunnusega kaalumisest järeldaval analüüsil täpsete tulemuste saamiseks piisata – siin on kaalumisel tarvis täpsemalt arvesse võtta ka valimivõtu eripärasid ehk valikudisaini, mida on valimi koostamisel kasutatud. Lihtsustatult öeldes tähendab see mitte ainult kaasamistõenäosuste erinevuste arvesse võtmist, vaid ka selle arvestamist, millised indiviidid asuvad millistes kihtides ja/või klastrites.

Miks piisab järeldavas analüüsis pelgalt kaalumisest lihtsa juhuvalimi korral, aga keerulisema valikuskeemi puhul tuleb arvestada valikudisainiga? Kihtvalimi puhul jaotatakse populatsioon kihtideks ja juhuslik valik tehakse kihtide sees. See vähendab valimi hajuvust ning reeglina parandab valimi esinduslikkust populatsiooni suhtes, mille tõttu on näiteks punkthinnangute standardvead mõnevõrra väiksemad kui tavalised, lihtsa juhuvalimi jaoks mõeldud standarvea arvutamise meetodid võimaldaksid järeldada. Klastrite puhul ei ole valimiliikmed üksteisest sõltumatud – kuna klastrid moodustatakse mingi tunnuse alusel, on see siis elukoht või õppimine mingis koolis ja klassis, on ühes klastris olevad liikmed üksteisele reeglina mõnevõrra sarnasemad kui muude klastrite liikmetele. See suurendab mõõdetavate tunnuste hajuvust ja omakorda punkthinnangute standardvigu. Andmete kaalumisel seda klastritesisest homogeensust ja klastritevahelist heterogeensust ehk valikudisaini mitte arvestades hindaksime järeldaval analüüsil andmete täpsust üle ehk saaksime näiteks punkthinnangutele kitsamad usaldusvahemikud kui oleks tegelikult kohane.

R-s võimaldavad andmete kaalumisel valikudisaini arvestada näiteks paketid survey (Lumley 2010) ja selle põhjal koostatud pakett srvyr (Freedman Ellis ja Schneider 2024). Pakett survey on vanem ja täielikum, samas on srvyr lihtsam kasutada neile, kes on R-s harjunud tidyverse’i süntaksiga, ja võimaldab käskudesse integreerida dplyr’i funktsioone.

Mõlema paketi võimaluste kasutamiseks tuleb esmalt seadistada andmete kaalumine ja valikudisain, paketis srvyr on selleks funktsioon as\_survey\_design. Selle funktsiooni põhilised argumendid on

* .data ehk andmestik;
* ids; selle väärtuseks tuleb omistada andmestikust tunnus(ed), mis sisaldab/-vad infot selle kohta, millisesse klastrisse vastaval andmereal olev indiviid kuulub. ESS-i andmestikes on taoline info tunnuses psu;
* strata, mille väärtuseks saab antud näites stratum ehk andmestikust tunnus, milles on kirjas, millisesse valikukihti mingi indiviid kuulus (juhul, kui valimi võtmisel kasutati kihte ehk populatsioon jaotati kihtideks).
* weights ehk kaalutunnus.

Paketid survey ja srvyr eeldavad vaikimisi, et andmeid kaalutakse disainikaaludega. Et järelkihistamist või muud disainikaalude kohandamist järeldaval analüüsil täpselt arvesse võtta, tuleks kohandatud kaalud uuesti kalibreerida. See on võimalik (vt täpsemalt [siit](https://cran.r-project.org/web/packages/survey/vignettes/precalibrated.pdf)), aga ESS-i andmete puhul keerulisem, sest ESS-i andmestikes on kaalud kujul, kus kaalude summa on ligikaudu võrdne indiviidide arvuga andmestikus, mitte indiviidide arvuga populatsioonis. Alternatiiv, mida praktikas tihtipeale kasutatakse, on käsitleda kohandatud kaale justkui oleks tegu tavaliste disainikaaludega. Sellega saadakse küll mõnevõrra konservatiivsemad hinnangud (nt mõnevõrra laiemad usaldusvahemikud kui järelkihistatud disainikaalud tegelikult eeldaks), aga erinevus ei ole enamasti suur (Lumley 2021). Ka ESS-i andmete kaalumise juhend pakub välja sellise lähenemise, seda on ka järgnevas käsus kasutatud, kus argumendi weights väärtuseks on pspwght.

library(srvyr)  
  
gr10w <- as\_survey\_design(.data = gr10, ids = psu, strata = stratum, weights = pspwght)

Valikudisaini objekti ee10w ehk objekti, kus on koos küsitlusandmestik koos valikudisaini ja kaalumise seadistusega, saab kasutada sisendina järeldavas analüüsis, nt standardvigade/usalduspiiride arvutamiseks või statistiliste testide jaoks. Selle jaoks tuleb kasutada vastavaid srvyr või survey pakettide funktsioone, nt aritmeetilise keskmise ja selle standardvea/usalduspiiride arvutamiseks ei saa kasutada lihtsalt funktsiooni weighted.mean vms, vaid tuleb kasutada funktsioone srvyr::survey\_mean või survey::svymean. Siinkohal on toodud kaalutud keskmiste arvutamise näide neist esimesega; argumendiga vartype saab selle funktsiooni puhul seadistada usalduspiiride vm näitajate arvutamise ("ci"), vaikeseadena arvutatakse muidu ainult keskmise standardviga ("se"). Allolevas näites arvutatakse keskmine rahulolu tervishoiusüsteemiga ja selle standardviga ja usalduspiirid tervisehinnangu lõikes (eelnevalt kodeeritakse väga halva tervisehinnanguga vastajad, keda on väga vähe, kokku halva tervisehinnanguga vastajatega), lisaks arvutatakse summarise-käsus kaalutud vastajate arv. Usalduspiirid arvutatakse vaikimisi usaldusnivool 95%, vajadusel saab seda muuta survey\_mean argumendi level abil (kõigi argumentide vaikeseadeid ja võimalike sätteid saab uurida käsuga ?survey\_mean).

gr10w |>   
 mutate(health4 = case\_match(health,  
 5 ~ 4,  
 .default = health)) |>   
 group\_by(health4) |>   
 summarise(stfhlth\_kesk = survey\_mean(stfhlth, na.rm = TRUE, vartype = c("se", "ci")),  
 n = sum(pspwght))

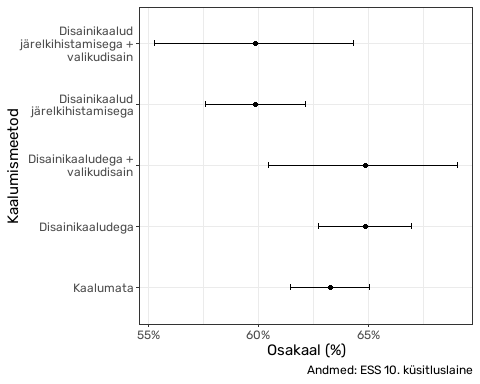
## # A tibble: 5 × 6  
## health4 stfhlth\_kesk stfhlth\_kesk\_se stfhlth\_kesk\_low stfhlth\_kesk\_upp n  
## <dbl+lb> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 [Ver… 4.35 0.142 4.07 4.63 1.25e3  
## 2 2 [Goo… 4.64 0.168 4.30 4.97 9.99e2  
## 3 3 [Fai… 4.47 0.198 4.08 4.86 4.31e2  
## 4 4 [Bad] 3.48 0.334 2.82 4.14 1.15e2  
## 5 NA 1 0 1 1 4.00e0

Funktsiooniga srvyr::survey\_prop saab arvutada tunnuse kaalutud jaotuse, grupeerides eelnevalt andmestiku selle tunnuse alusel, mille jaotust soovime saada. Arvutame siinkohal internetikasutuse sageduse kaalutud jaotuse ning osakaalude standardvead ja usalduspiirid.

gr10w |>   
 group\_by(netusoft) |>   
 summarise(prop = survey\_prop(vartype = c("se", "ci")))

## # A tibble: 6 × 5  
## netusoft prop prop\_se prop\_low prop\_upp  
## <dbl+lbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 [Never] 0.184 0.0128 0.160 0.210   
## 2 2 [Only occasionally] 0.0418 0.00595 0.0315 0.0552   
## 3 3 [A few times a week] 0.0631 0.0102 0.0458 0.0865   
## 4 4 [Most days] 0.110 0.0118 0.0891 0.136   
## 5 5 [Every day] 0.599 0.0229 0.553 0.643   
## 6 NA 0.00198 0.00102 0.000717 0.00548

Järgnev näide ilmestab andmete kaalumise ja valikudisaini arvesse võtmise mõju analüüsitulemustele. Arvutatakse tunnuse netusoft kaalutud jaotus ja osakaalude usalduspiirid usaldusnivool 95%, kasutades erinevaid kaalumise seadeid. Joonisel esitatakse suurima osakaalu (internetti igapäevaselt kasutavate inimeste protsent) punkt- ja vahemikhinnang (usaldusnivool 95%) eri kaalumise seadete korral.



Joonis 1. Kreekas internetti igapäevaselt kasutavate inimeste osakaalu punkt- ja vahemikhinnang (usaldusnivool 95%) eri kaalumise seadete korral.

Jooniselt ilmneb, et kui kasutada osakaalu arvutamisel ainult disainikaale, on punkthinnang antud juhul paari protsendipunkti võrra kõrgem ja selle usaldusvahemik natuke laiem kui andmed kaalumata jättes. Üldjuhul ongi nii, et kaalumisel võib usaldusvahemik muutuda mõnevõrra laiemaks, sest kaalumine suurendab variatiivsust andmetes. Kui lisaks disainikaaludega kaalumisele võtta arvesse ka valikudisaini, jääb punkthinnang võrreldes ainult kaalumisega samaks, ent märkimisväärselt suureneb usaldusvahemik. See tuleneb asjaolust, et nüüd arvestatakse usalduspiiride arvutamisel ka valimi võtmisel kasutatud mitmeastmelise valikuga (ühes piirkonnas asuvad leibkonnad on keskmiselt üksteisele sarnasemad kui teiste piirkondade leibkondadele).

Samalaadseid erinevusi võrreldes kaalumata andmetelt tehtud arvutustega näeme ka juhul, kui disainikaale on kohandatud järelkihistamise alusel. Põhjus, miks punkthinnangud antud juhul eelnevaga võrreldes erinevad, tuleneb erinevast vastamismäärast eri interneti kasutamise sagedusega valimiliikmete seas (täpsemalt öeldes küll sellest, kuidas interneti kasutamise sagedus on seotud järelkihistamise aluseks olevate tunnustega ja vastamismäärast nende tunnuste lõikes).

Olulised järeldused siit on, et andmeid kaalumata võime eksida punkthinnangutes, valikudisaini arvestamata jättes võime alahinnata vahemikhinnangute laiust ehk määramatust valimi alusel leitavates hinnangutes. Kasutades disainikaale ilma järelkihistamiseta saame küll arvutada näitajad, mis arvestavad valimivõtu eripäradega, ent ei arvesta mittevastamisest lähtuva esinduslikkuse vähenemisega. Seega, eeldusel, et järelkihistamisel on kasutatud sobivaid tunnuseid, mis on seotud mittevastamise mustrite ja analüüsitavate tunnustega, võimaldab järelkihistamisega kohandatud disainikaalude rakendamine koos valikudisaini arvesse võtmisega leida täpsemad hinnangud.

Paketid srvyr ja survey võimaldavad kaalutud andmetega läbi viia ka statistilisi teste. Need paketid töötavad koos: funktsioon svyttest, millega saab teha *t*-testi, on paketis survey, kuid selle sisendiks sobib meie varasemalt srvyr-i abil koostatud valikudisaini objekt gr10w.

library(survey)  
  
svyttest(formula = stfhlth ~ gndr, design = gr10w)

##   
## Design-based t-test  
##   
## data: stfhlth ~ gndr  
## t = -0.035822, df = 175, p-value = 0.9715  
## alternative hypothesis: true difference in mean is not equal to 0  
## 95 percent confidence interval:  
## -0.2228339 0.2148890  
## sample estimates:  
## difference in mean   
## -0.003972449

Paketi survey abil saab kaalutud andmete põhjal koostada ka üldistatud lineaarseid regressioonimudeleid, Coxi regressioonimudeleid, loglineaarseid mudeleid ja kasutada muid analüüsimeetodeid. Vt lähemat infot paketi (Lumley 2024) veebilehelt.

Euroopa Sotsiaaluuringu andmete kaalumise kohta olgu veel öeldud, et andmestikus on lisaks kaalutunnus pweight. Tegu on populatsioonikaaludega, mis võtavad arvesse asjaolu, et elanike arv erineb riigiti kordades, samas kui vastajate arv on riigiti samas suurusjärgus. Populatsioonikaale oleks tarvis kasutada juhul, kui analüüsitakse mitme riigi andmeid koos, st kogum, millele tulemusi soovitakse üldistada, pole ühe riigi elanikkond, vaid mitme riigi elanikkond tervikuna. See tähendab, et soovime teha järeldusi riikideülese regiooni kui terviku kohta, näiteks soovime teada, milline on üldistatud usalduse tase Baltikumis tervikuna või milline on seos sisserändesse suhtumise ja vanuse vahel Põhjala regioonis üldiselt. Ilma populatsioonikaale kasutamata oleksid väiksemate riikide vastajad tulemustes üleesindatud.

Kui soovime populatsioonikaale kasutada, ei tohiks andmeid lihtsalt kaaluda tunnusega pweight, vaid arvestada tuleks ka disainikaale ja nende järelkihistamisega kohandamist. Seetõttu tuleks populatsioonikaalud läbi korrutada järelkihistamisega kohandatud disainikaaludega ja kaaluda andmestikku saadud korrutiste alusel:

gr10 <- gr10 |>   
 mutate(anweight = pspwght \* pweight)

Alates 9. küsitluslainest on see tehe juba andmete kasutaja eest ära tehtud, st nende lainete andmestikes on juba ka tunnus anweight. ESS-i kaalumise juhendites (European Social Survey 2025) üldiselt soovitataksegi igasugustes analüüsides kasutada seda kaalutunnust. Käesolevas SAMM-u peatükis, kus tegeletakse ainult ühe riigi andmete analüüsiga, on siiski läbivalt kasutatud kaalumiseks tunnust pspwght. Selle eelis on, et ühe riigi andmeid (või riike võrdlevalt, eri kogumitena) analüüsides on indiviidide kaalutud koguarv sama, mis indiviidide koguarv kaalumata andmestikus. See võimaldab andmetes paremini orienteeruda – tunnusega anweight andmeid kaaludes oleks nt Eesti vastajate kaalutud koguarv saja-paarisaja ringis, mis on informatiivne ainult juhul, kui analüüsime eri riikide andmeid ühe tervikuna.

## Viited

AAPOR. 2025. „Disclosure Standards“. <https://aapor.org/standards-and-ethics/disclosure-standards/>.

Dillman, Don A, Jolene D Smyth, ja Leah Melani Christian. 2014. *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

European Social Survey. 2023. „ESS Weighting Data“. <https://www.europeansocialsurvey.org/sites/default/files/2023-06/ESS_weighting_data_1_1.pdf>.

———. 2025. „ESS Methodology: Data Processing and Archiving - Weighting“. <https://www.europeansocialsurvey.org/methodology/ess-methodology/data-processing-and-archiving/weighting>.

Freedman Ellis, Greg, ja Ben Schneider. 2024. *srvyr: ’dplyr’-Like Syntax for Summary Statistics of Survey Data*. <http://gdfe.co/srvyr/>.

Lohr, Sharon L. 2019. *Sampling: design and analysis*. Chapman; Hall/CRC.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley; Sons.

———. 2021. „Using post-stratification weights in R survey package“. Cross Validated. <https://stats.stackexchange.com/q/540570>.

———. 2024. „survey: analysis of complex survey samples“.