Valim, üldkogum, usaldusvahemik

Ühiskonna ja infoprotsesside analüüsi metodoloogia

Indrek Soidla

19 11 2021

Valim ja populatsioon (ehk üldkogum)

- Kui analüüsime valimiandmeid, arvutame mingi näitaja – parameetri
  - Nt eluga rahulolu *aritmeetiline keskmine*
    - Skaalal 0-10 $m = 6{,}84$
  - Erakondade reitingud – *osakaal (%)* vastanutest, kes valiksid teatud erakonda
    - Nt Reformierakonda toetab 26% valimiseelistust omavatest vastajatest
- Valimi alusel arvutatud parameeter (nt keskmine) kehtib täpselt valimi kohta
- Kas saame eeldada, et populatsioonis on see näitaja täpselt sama?
- Kui mitte, siis kui täpselt kirjeldab see parameeter populatsiooni?
- Sõltub sellest, kuivõrd on valim populatsioonis suhtes esinduslik
  - n-ö täpne väike koopia populatsioonist
- Mis valimi koostamisel tagab, et valim on võimalikult esinduslik populatsiooni suhtes?
- Põhimõtteliselt saaksime seada valimile seada kvoodid
  - kui palju peab valimis olema nt mehi ja naisi
  - kui palju peab valimis olema nt erinevate vanusrühmade esindajaid
- Saame valimi, mis on täpne koopia populatsioonist nende tunnuste lõikes
- Samas ei taga see esinduslikkust teiste võimalike tunnuste suhtes
  - Nt kui arvamused, hinnangud, mida teada tahame, ei ole seotud soo ja vanusega
  - St arvamuste, hinnangute jaotus on samasugune meeste ja naiste seas

Valim ja populatsioon (ehk üldkogum)

- Selleks, et valim oleks populatsiooni suhtes esinduslik kõikvõimalike tunnuste osas
  - st tunnuste osas, mille suhtes pole kvoote rakendatud
- on parim garantii valimi juhuslikkus (tõenäosuslik valim)
- Kõigil populatsiooni liikmetel nullist erinev tõenäosus sattuda valimisse
- Lihtne juhuvalim – tõenäosus valimisse sattuda on kõigil populatsiooni liikmetel võrdne
- Reeglina eeldab valikuraami (populatsiooni liikmete loendi) olemasolu
- Tänu valimi juhuslikkusele on valim võimalikult sarnane populatsioonile
- Samas võib just tänu juhuslikkusele valim siiski mingil määral populatsioonist erineda
- Mitte palju, aga teatud määral siiski
  - Nt populatsioonis on 55% naisi ja 45% mehi
  - Puhtalt juhuslikkuse tõttu võivad need osakaalud valimis natuke varieeruda, nt 57% naisi ja 43% mehi
  - Selle tõttu võivad erineda ka muud valimi alusel arvutatud parameetrid
  - Nt eluga rahulolu aritmeetiline keskmine populatsioonis on 6,77, aga meie saame valimi alusel 6,84

Kui täpselt kirjeldab valimi alusel arvutatud parameeter populatsiooni?

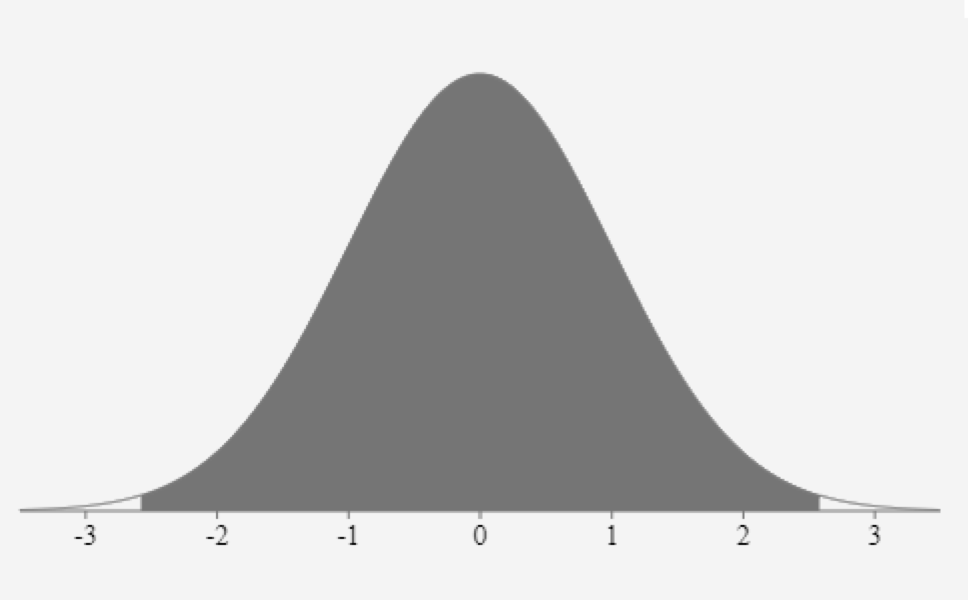
- Parameetri väärtus populatsioonis – parameetri tegelik väärtus *(true value)*
  - Ei ole võimalik teada saada muidu, kui uurides kogu populatsiooni (nt rahvaloendus)
  - Teame selle alusel, kui palju on populatsioonis nt mehi ja naisi
  - Ei tea ega olegi võimalik teada saada paljusid muid näitajaid (nt keskmist eluga rahulolu)
  - Seega ei saa me ka teada, kui palju *täpselt* erineb valimi alusel saadud parameeter tegelikust väärtusest
  - Küll aga saame hinnata vahemikku, kui palju valimi alusel saadud parameeter erineb tegelikust väärtusest
- Usaldusvahemik – valimi alusel arvutatav vahemik, mis katab parameetri tegeliku väärtuse teatud tõenäosusega
  - täpsemalt: teatud läve ületava tõenäosusega,
  - st tõenäosusega, mis ei ole väiksem kui usaldusnivoo (selgitame täpsemalt hiljem)

Kui täpselt kirjeldab valimi alusel arvutatud parameeter populatsiooni?

- Nt Euroopa Sotsiaaluuringu andmetel saame öelda, et 2016. aastal oli Eestis skaalal 0-10 eluga rahulolu aritmeetiline keskmine 6,84 usaldusvahemikuga 6,75 (alumine usalduspiir) kuni 6,93 (ülemine usalduspiir)
- St eluga rahulolu aritmeetilise keskmise tegelik väärtus võis olla mitte 6,84, vaid nii väike kui 6,75 või nii suur kui 6,93
- Lihtne!
- Siiski: tuleb kindlasti lisada, et eelnimetatud usaldusvahemik oli arvutatud usaldusnivool 95%
- Mida see tähendab?
- Ja kuidas sellised usalduspiiride väärtused üldse saadi?
- Püüamegi järgnevalt seda mõtestada
- Enne kordame üle standarhälbe ja normaaljaotuse olulised omadused

Normaaljaotus ja standardhälve

- Tuletame meelde: standardhälve näitab tüüpilist erinevust aritmeetilisest keskmisest
- Kui tunnus on normaaljaotusega, siis on teada, mitme standardhälbe kaugusel on teatud osa tunnuse väärtustest
  - 90% tunnuse väärtustest jääb vahemikku 1,64 standardhälvet keskmisest
  - 95% tunnuse väärtustest jääb 1,96 standardhälbe kaugusele
  - 99% tunnuse väärtustest jääb 2,58 standardhälbe kaugusele



Teeme mõtteharjutuse

- Võtame populatsioonist 100 juhuvalimit
- Viime samade meetoditega läbi 100 uuringut
- Saame sama tunnuse kohta 100 aritmeetilist keskmist
- Kuna tegu on juhuvalimitega, siis
  - enamiku valimite keskmised on populatsiooni keskmise lähedal
  - osad keskmised on populatsiooni keskmisest natuke kaugemal
  - valimite alusel saadud keskmiste keskmine = populatsiooni keskmine
  - valimite keskmised jaotuvad normaaljaotuse kohaselt =>
  - sajast valimist vähemalt
    - 90 valimi keskmised jäävad vahemikku 1,64 standardhälvet populatsiooni keskmisest
    - 95 valimi keskmised jäävad vahemikku 1,96 standardhälvet populatsiooni keskmisest
    - 99 valimi keskmised jäävad vahemikku 2,58 standardhälvet populatsiooni keskmisest
- Valimite keskmiste standardhälvet nimetatakse standardveaks

Standardviga

- Analoogselt eelnevalt kirjeldatuga saab öelda, et populatsiooni keskmisest on kuni 1,96 standardvea kaugusel vähemalt 95 valimi andmete alusel arvutatud keskmised
- Standardviga saaks arvutada standardhälbe valemiga
  - Selleks oleks tarvis kõigi valimite põhjal arvutatud keskmisi
  - Õnneks saab standardvea arvutada lihtsamini valemiga $\frac{\sigma}{\sqrt{n}}$, kus
    - σ – populatsiooni keskmise standardhälve (suure valimi puhul ligilähedane valimi alusel arvutatud keskmise standardhälbega)
    - n – indiviidide arv

Usaldusvahemik

- Seega, tähistades
  - μ – keskmine populatsioonis (keskmise tegelik väärtus)
  - m – keskmine valimis,
- saab öelda, et vähemalt 95 valimis kehtib tingimus:

$$\mu - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \leq m \leq \mu + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}$$

- Sellest tulenevalt saab ka öelda, et 95 valimis kehtivad tingimused:

$$\mu \geq m - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}$$

$$\mu \leq m + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}$$

- Olemegi jõudnud valimikeskmise alumise ja ülemise usalduspiiri ehk usaldusvahemiku valemiteni (täpsemalt: usaldusnivool 95%)
- Tuletame meelde usaldusvahemiku mõiste: vahemik, mis katab parameetri tegeliku väärtuse teatud läve ületava tõenäosusega
- Antud juhul usaldusnivoo ongi see lävi
- Mida see sisuliselt tähendab? Kuidas neid näitajaid tõlgendada? Vaatame konkreetse näite varal

Usaldusvahemiku näide

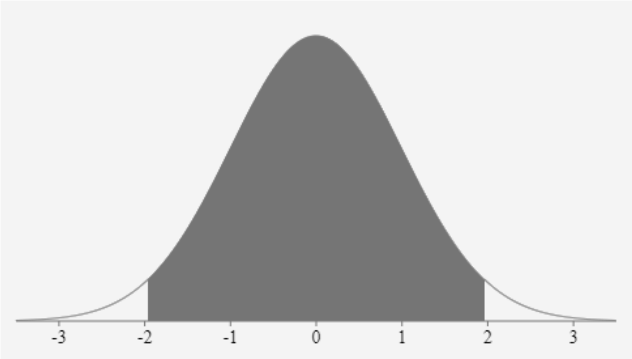
- ESS 2016 eluga rahulolu
- Valimi keskmine (skaalal 0-10) $m = 6{,}84$
- Arvutame keskmise usalduspiirid usaldusnivool 95%, kui
  - $n = 2019$
  - $\sigma \approx s = 2{,}02$
  - Alumine usalduspiir $m - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} = 6{,}84 - 1{,}96 \cdot \frac{2{,}02}{\sqrt{2019}} = 6{,}84 - 0{,}09 = 6{,}75$
  - Ülemine usalduspiir $m - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} = 6{,}84 + 1{,}96 \cdot \frac{2{,}02}{\sqrt{2019}} = 6{,}84 + 0{,}09 = 6{,}93$
- Vaatame, kuidas sama R-s teha
- Mida saame selle põhjal öelda?
- Võttes aluseks usaldusnivoo 95%, saame öelda, et 2016. aastal oli Eesti 15+ elanikkonnas eluga rahulolu tegelik keskmine vahemikus 6,75 ja 6,93
- Mida näitab siin usaldusnivoo 95%?

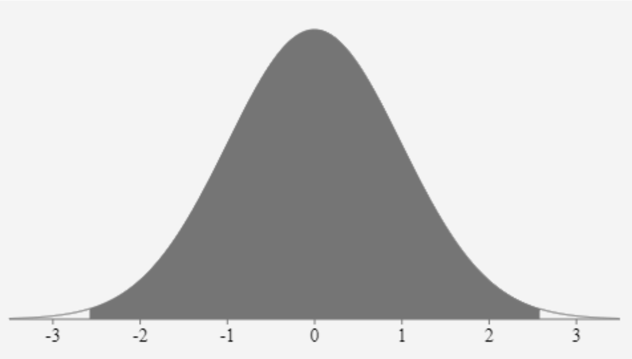
Mida näitab siin usaldusnivoo 95%?

- Lihtsalt sõnastades: teatavat täpsuse astet
- Rakendades eelnevat 100 valimi hüpoteetilist näidet, siis eeldades, et meie valim on üks neist 95-st valimist, milles arvutatud keskmine jääb 1,96 standardvea piiresse, saab öelda, et leitud usaldusvahemik katab eluga rahulolu tegelikku keskmise populatsioonis
- Kust me teame, et meie valim on üks neist 95-st?
- Ei teagi :)
- Lihtsalt eeldame, et 95% on piisavalt kõrge usaldusnivoo
- St et meil peab väga kehv õnn olema, et juhuse tahtel oleks meil kasutada olev valim kõikvõimalike valimite hulgast just nende 5% valimite seas, mis on populatsioonist kõige erinevamad
- Võime valida ka kõrgema usaldusnivoo, et olla veel kindlam, et saadav usaldusvahemik tõesti katab tegeliku keskmise (sel juhul on usaldusvahemik laiem)

Mida näitab usaldusnivoo?

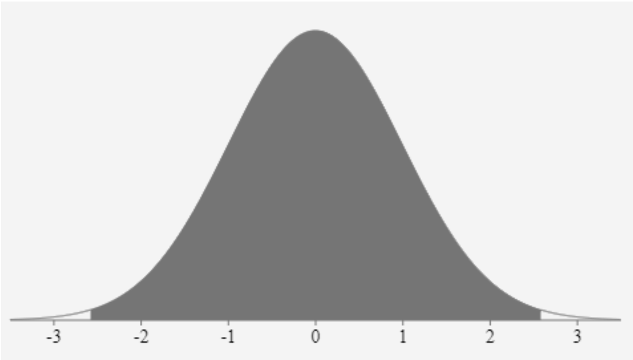
- Teisiti öeldes,
  - kui võtame usaldusnivoo 95%,
  - arvestame võimalusega, et
  - meie valim on (meie analüüsitava tunnuse jaotuse seisukohalt) 95 protsendi populatsiooniga kõige sarnasemate valimite seas;
  - me ei arvesta võimalusega, et meie valim võib olla 5 protsendi populatsioonist kõige rohkem erineva valimi seas
- Samalaadselt,
  - kui võtame rangema usaldusnivoo, 95% asemel 99%,
  - arvestame usaldusvahemiku arvutamisel võimalusega, et
  - meie valim ei pruugi olla 95 protsendi populatsiooniga kõige sarnasemate valimite seas,
  - st võib populatsioonist veel rohkem erineda;
  - samas ei arvesta me võimalusega, et meie valim võib olla 1 protsendi populatsioonist kõige rohkem erineva valimi seas





Mida näitab usaldusnivoo?

- Mida rangema usaldusnivoo valime,
  - seda suuremat erinevust populatsioonist peame võimalikuks ja
  - seda laiem on usaldusvahemik
  - sest arvestame võimalusega, et meie valim erineb populatsioonist rohkem



Kuidas oleks õige usaldusvahemikku tõlgendada

- Kuidas tekstis sõnastada:
- 2016. aastal oli Eestis eluga rahulolu keskmine skaalal 0-10 6,84 palli, CI 95% [6,75; 6,93]
- Tihti tõlgendatakse nii, et tõenäosusega 95% on keskmise tegelik väärtus vahemikus 6,75 kuni 6,93
- Ei ole õige!
- Oluline vahet teha, mis on kindel ja mille kohta saab rakendada tõenäosuse (juhuslikkuse) mõistet
- Tegelik väärtus populatsioonis on kindel väärtus
    - Me ei tea selle väärtust, aga lähtume sellest, et ta on olemas
    - Võib ajas varieeruda, aga mõõtmise hetkel on konkreetne väärtus
- Valim, mille populatsioonist võtame, võib mingi tõenäosusega erineda populatsioonist
- 95% – tõenäosus saada valimit, kus eluga rahulolu keskmise usaldusvahemik (mis arvutatakse usaldusnivool 95%) katab keskmise tegeliku väärtuse
- „Usaldusvahemik katab keskmise tegeliku väärtuse 95 juhul 100-st“
    - Õige, aga oluline teada, et „juhud“ ei viita indiviididele, vaid valimitele!
    - Sealjuures hüpoteetilistele valimitele (st tegelikkuses eksisteerib meil ainult üks valim)

Usaldusnivoo valik

- Usaldusnivoo 95% on enim levinud usaldusnivoo
  - Andmeanalüüsiprogrammides tihti vaikeseadena määratud
- Kui tahame kindlamad olla, et meie arvutatav keskmise usaldusvahemik katab populatsiooni keskmise =>
- Arvutame keskmise usaldusvahemiku usaldusnivool 99%
- Kui meil pole nii suurt kindlust vaja =>
- Arvutame keskmise usaldusvahemiku nt usaldusnivool 90%
- Millest lähtuda usaldusnivoo valikul?
  - Kui kriitiline on tagada, et usaldusvahemik katab tegeliku väärtuse
  - Kui pole otsest vajadust tavalisest rangema/leebema usaldusnivoo järele, võib võtta 95%

Usalduspiiride arvutamise valem

- Üldisem aritmeetilise keskmise usalduspiiride arvutamise valem (usaldusnivoo väärtus valemis fikseerimata)

$$\mu \geq m - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\mu \leq m + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

- μ – keskmine populatsioonis (keskmise tegelik väärtus)
- m – keskmine valimis
- $z_{1-\frac{\alpha}{2}}$ – standardiseeritud normaaljaotuse $1-\frac{\alpha}{2}$ kvantiil, kus α on vea tõenäosuse piir
  - nt usaldusnivool 90% $\alpha = 10\%$, vastav $z_{1-\frac{\alpha}{2}} = 1{,}64$
  - nt usaldusnivool 95% $\alpha = 5\%$, vastav $z_{1-\frac{\alpha}{2}} = 1{,}96$
  - nt usaldusnivool 99% $\alpha = 1\%$, vastav $z_{1-\frac{\alpha}{2}} = 2{,}58$
- $\frac{\sigma}{\sqrt{n}}$ – keskmise standardviga:
  - σ – populatsiooni keskmise standardhälve (suure valimi puhul ligilähedane valimikeskmise standardhälbega)
  - n – valimimaht

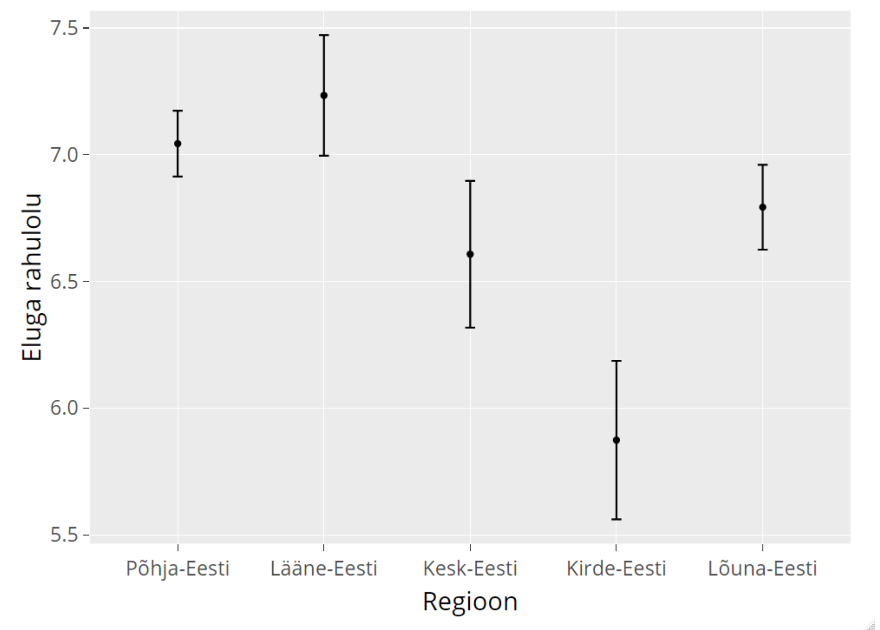
Usaldusvahemikud erinevatel usaldusnivoodel

- ESS 2016 eluga rahulolu
- Valimi keskmine (skaalal 0-10) $m = 6{,}84$
- Arvutame keskmise usalduspiirid usaldusnivool 95%, kui
  - $n = 2019$
  - $\sigma \approx s = 2{,}02$
  - Alumine usalduspiir $m - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 6{,}84 - 1{,}96 \cdot \frac{2{,}02}{\sqrt{2019}} = 6{,}84 - 0{,}09 = 6{,}75$
  - Ülemine usalduspiir $m - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 6{,}84 + 1{,}96 \cdot \frac{2{,}02}{\sqrt{2019}} = 6{,}84 + 0{,}09 = 6{,}93$
- Usaldusnivool 99% ja 90% on kõik muud näitajad arvutustes samad, välja arvatud $z_{1-\frac{\alpha}{2}}$, mis on vastavalt 2,58 ja 1,64
- Eluga rahulolu keskmine hinnang oli 2016. aastal Eestis 6,84 palli,
  - 90% CI [6,77; 6,92]
  - 95% CI [6,75; 6,93]
  - 99% CI [6,73; 6,96]
- Usalduspiirid on keskmise suhtes sümmeetrilised, väikesed erinevused antud näites tulenevad ümardamisest
- Teeme sama R-s

Usaldusvahemike arvutamine grupiti

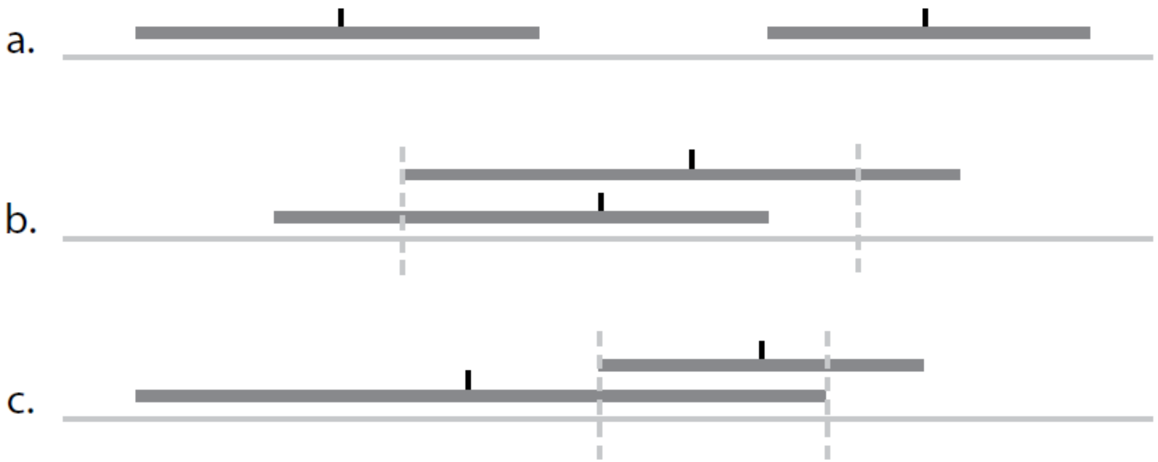
- Arvutame ESS 2016 põhjal eluga rahulolu regiooni lõikes usaldusnivool 95%
- Teeme selle läbi ka R-s
- Usaldusnivool 95% saab öelda, et
  - eluga rahulolu alusel eristuvad kolm regioonide rühma
  - Põhja- ja Lääne-Eestis oli eluga rahulolu keskmine üle 7 või 7 piirimail
  - Kesk- ja Lõuna-Eestis jäi eluga rahulolu keskmine alla 7 palli
  - Kirde-Eestis oli eluga rahulolu teistest regioonidest madalam ja keskmine jäi alla 6,2 palli
- Kas eluga rahulolu keskmised Põhja-Eestis ja Lõuna-Eestis erinevad?
- Usaldusvahemikud kattuvad, kuigi vähesel määral
- Keskmiste usaldusvahemike võrdlemisel seda väita ei saa
- Kui usaldusvahemikud kattuvad, siis keskmiste erinevuse osas saaks täpsema järelduse teha t-testi (keskmiste erinevuse usaldusvahemiku) põhjal – vaatame järgmine kord

Regioon	Keskmine	Standardviga	Alumine usalduspiir	Ülemine usalduspiir
Põhja-Eesti	7,04	0,07	6,91	7,17
Lääne-Eesti	7,23	0,12	7,00	7,47
Kesk-Eesti	6,61	0,15	6,32	6,90
Kirde-Eesti	5,87	0,16	5,56	6,19
Lõuna-Eesti	6,79	0,09	6,63	6,96



Usaldusvahemike võrdlemine grupiti

- Lihtne viis keskmiste usaldusvahemike alusel võrdlemise „usaldamiseks“



a. Saab erinevust väita
b. Erinevust ei saa väita
c. Kontrolli keskmiste erinevuse usaldusvahemikku või mõne statistilise testi alusel

Tooding (2015: 165)

Praktikumiülesanne

- ESS 8. laines küsiti vastajatelt, kui tõenäoliseks nad peavad, et piisaval hulgal valitsusi rakendab meetmeid kliimamuutuse vähendamiseks (gvsrdcc)
- Millised on selle arvamuse keskmise punkthinnangud ja vahemikhinnangud haridustasemeti? Arvutage usaldusvahemikud kõigil kolmel tavapärasel usaldusnivool.
- Pange keskmised ja nende usaldusvahemikud joonisele
- Mida saab väita nende usaldusvahemike võrdlemise põhjal?
- Milliste gruppide vahel saab populatsiooni tasandil väita erinevuste esinemist
    - usaldusnivool 90%?
    - usaldusnivool 95%?
    - usaldusnivool 99%?

Millest sõltub usaldusvahemiku laius?

$$\mu \geq m - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\mu \leq m + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

- Valimimaht
  - Suurem valimimaht => kitsam usaldusvahemik
- Tunnuse hajuvus
  - Väiksem hajuvus => kitsam usaldusvahemik
- Usaldusnivoo
  - Madalam usaldusnivoo => kitsam usaldusvahemik
- Kas selleks siis, et tulemuses maksimaalselt kindel olla, on vaja suuremat valimimahtu, tunnuse väiksemat hajuvust ja madalamat usaldusnivood?
- Madalam usaldusnivoo tähendab ka suuremat eksimisvõimalust
- Kompromiss:
  - tulemuse suuremat konkreetsust taotledes kaotame tulemuse tõsikindluses,
  - tõepärasuse astet suurendades väheneb tulemuse konkreetsus

Millal ei ole kohane arvutada usaldusvahemikke

- Kui tegu pole valimiandmetega
  - Nt rahvaloendus, registriandmed
- Kui tegu pole tõenäosusliku valimi andmetega
  - AAPOR: mittetõenäosusliku valimi puhul usaldusvahemike jt statistilisi järeldusi võimaldavate näitajate esitamine eksitav, ei tohiks teha
  - Nt mugavusvalim, ekspertvalim ei ole tõenäosuslik valim
  - Ei saa eeldada valimiliikmete juhuslikkust
  - Kvootvalim?
  - Erinevate mittetõenäosuslike valimite esinduslikkus populatsiooni suhtes erinev
  - Kui siiski arvutatakse usaldusvahemik nt kvootvalimi põhjal, tuleb
    - lugejat valimi mittetõenäosuslikkusest ja usaldusvahemike ebatäpsusest selgelt teavitada
    - Järelduste tegemisel olla ettevaatlik
- Püüame usaldusvahemike arvutamist / populatsiooni tegeliku väärtuse hindamist mõtestada mõne näite varal
  - Imikusuremuse protsent Balti riikides
  - Ettevõtte rahulolu-uuringu tulemused, nt keskmine rahulolu töötasuga, töökeskkonnaga