





Sotsiaalse analüüsi meetodid:  
kvantitatiivne lähenemine

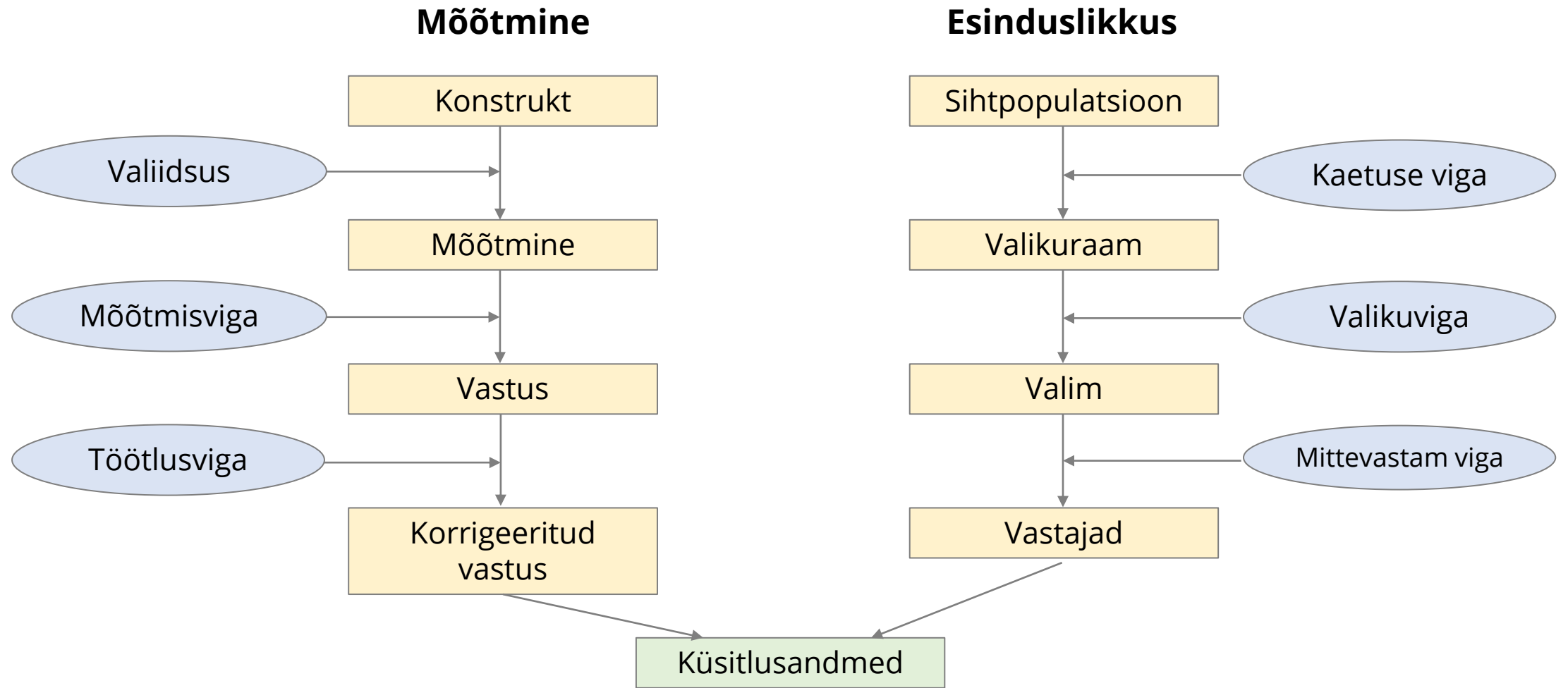
Andmete kvaliteedi hindamine,  
andmelüngad, erindid

Indrek Soidla

## *So now we've got our dataset, what are we gonna do with it?*

- Millele peaksime andmeid analüüsima hakates kõigepealt tähelepanu pöörama?
- Milline on andmete kvaliteet? Kas andmed on usaldusväärsed?
- Täpsemalt nt:
  - Kuidas on andmed kogutud?  küsimus esinduslikkusest, üldistatavusest
  - Kas andmed mõõdavad seda, mida me eeldame, et nad mõõdavad?  küsimus valiidsusest ja reliaablusest
  - Kas analüüsiühik on sobiv?  küsimus mõõtmistasandist
  - Kas andmed vastavad analüüsimeetodite eeldustele?  küsimus analüüsimeetoditest
- Põhiline, mida teada tahame:
  - Kui me neid andmeid analüüsime, kui kindlad saame olla tulemuste täpsuses?

# Uuringu koguviga



# Uuringu koguviga

- Kaetuse viga – valikuraam ei kata sihtpopulatsiooni täpselt
  - Valikuraam – loend sihtpopulatsiooni liikmetest
  - Sõltub valikuraami katvusest, kvaliteedist
- Valikuviga – kõrvalekalle tegelikust väärtusest populatsioonis, mis tuleneb valimi juhuslikkusest
  - Sõltub valimi suurusest, konkreetse tunnuse hajuvusest
  - See on see, millel põhinevad usaldusvahemikud, olulisuse tõenäosus
  - Ei saa (täpselt) hinnata, kui valim pole juhuslik / juhuslikkus kannatab (vt eelm ja järgm punkt)
- Mittevastamise viga – vastamata jätmisest tulenev viga
  - Sõltub vastajate koostöövalmidusest (selle juhuslikkusest), välitöö kvaliteedist
- Valiidsus – kui täpselt instrument (uuringuküsimus) tegelikult mõõdab konstrukti, mida peaks mõõtma
  - Sõltub uuringuküsimuste koostamise põhjalikkusest, testimisest
- Mõõtmisviga – mõõtmisprotsessist tulenev viga (tekib, kui mõõtmismeetod mõjutab vastust)
  - Sõltub küsitlusviisist, välitöö kvaliteedist, küsimuste tundlikkusest / sotsiaalsest soovitatavusest
- Töötlusviga – erinevus vastaja antud vastuse ja analüüsidest kasutatava väärtuse vahel
  - Sõltub, kuivõrd on andmestiku järelkontrolliga vaeva nähtud

# Uuringu koguviga

- Kui me andmeid analüüsime, kui kindlad saame olla tulemuste täpsuses?
- Väga oluline uurida uuringu / andmete kogumise metoodikat!
- Võime näha palju vaeva andmete analüüsiga, aga...
- ...kui jätame tähelepanuta, kuidas andmed on tekkinud, siis
  - heal juhul me lihtsalt ei oska analüüsil olulistele nüanssidele tähelepanu pöörata
  - halval juhul lähevad eksijäreldused kellelegi palju maksma
- Halbade andmete äratundmiseks võib piisata vähesest

# Halbade andmete äratundmiseks võib piisata vähesest

- Millele peaks küsitlusuuringu andmete puhul tähelepanu pöörama?
- Lühike nimekiri olulistest küsimustest uuringu metoodika hindamisel
  - Millal ja kuidas uuring läbi viidi?
  - Keda küsitleti?
  - Kuidas nad välja valiti?
  - Kes rahastas uuringut?
  - Milliseid küsimusi küsiti?
- Millise info peaks uuringu korraldaja edastama:
  - <https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics/Disclosure-Standards.aspx>
- Põhjalikum nimekiri olulistest küsimustest:
  - <https://www.aapor.org/Education-Resources/Reports/Evaluating-Survey-Quality.aspx>
- Kuidas tagada võimalikult kvaliteetsed andmed:
  - <https://www.aapor.org/Standards-Ethics/Best-Practices.aspx>
- Kui uuringu korraldaja
  - ei oska/suuda anda selgeid vastuseid olulistele küsimustele või
  - ajab kesksete mõistete kohta villast (vastamismäär, veapiir, valimitüüp),
- siis pigem hoida nendest andmetest eemale



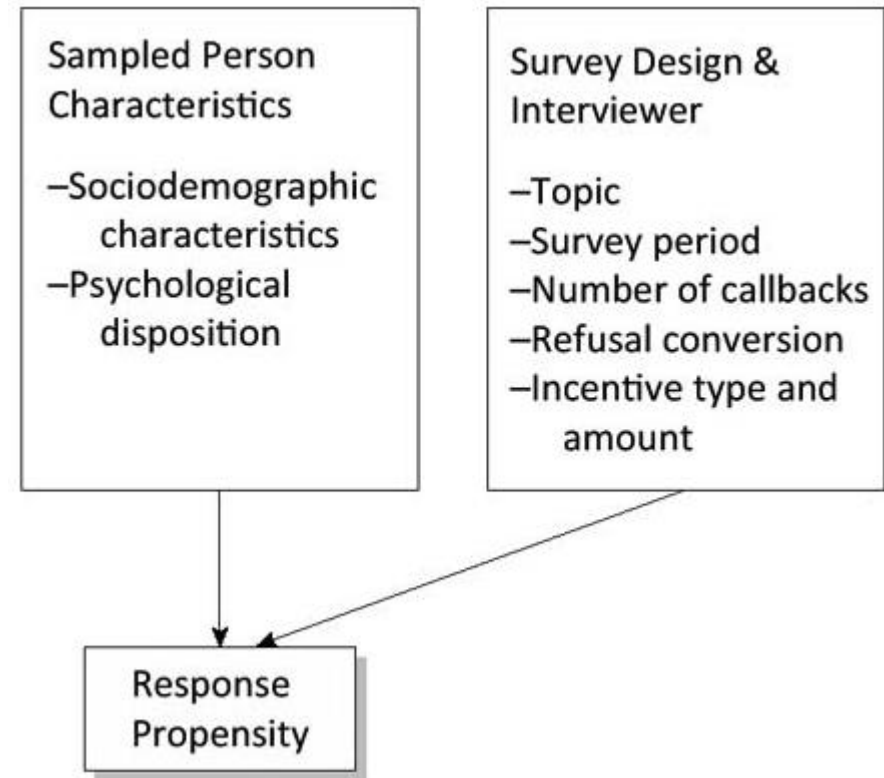
# Andmete esinduslikkus, tulemuste üldistatavus

- Põhiline on küsimus andmete esinduslikkusest
- Kui meil ei ole andmeid kogu sihtpopulatsiooni kohta, kas tulemused on sellele üldistatavad?
- Kas seda on võimalik kuidagi hinnata?
- Kui oluliste tunnuste lõikes esineb lahknevusi sihtpopulatsioonist, kas on võimalik parandada?
- On võimalik (kaalumise teel), aga oluline on siiski valimi juhuslikkus
- Pole küsimus ainult valimiandmete kohta
- Ka registri- või suurandmete puhul küsimus andmete täielikkusest
- Millest võivad andmelüngad tekkida?
  - Kutsele mittevastamine ehk täielik mittevastamine (*unit nonresponse, total nonresponse*) – vastused puuduvad kõigile küsimustele
  - Küsimusele mittevastamine ehk osaline mittevastamine (*item nonresponse, partial nonresponse*) – vastused puuduvad vähemalt ühele küsimusele
  - Täielike andmete ühendamisel mittetäielikega
  - Tehnilistest probleemidest (nt andmekogumisel, andmete töötlemisel)



# Kas madal vastamismäär ja andmelüngad on alati probleem?

- Mitte alati
- Ainult siis, kui esineb mittevastamise nihe (*nonresponse bias*)
- Sõltub sellest, kas esineb seos mõõdetavate tunnuste ja vastamiskäitumise vahel
- Üldisemalt andmelünkade kontekstis: kas andmelüngad esinevad juhuslikult või süstemaatiliselt





# Andmelünkade (mittevastamise) liigid

- *Missing completely at random (MCAR)* ehk täiesti juhuslikud lüngad
- Andmelünkade esinemises pole midagi süstemaatilist
- Andmelünkade esinemine ei sõltu uuritavast tunnusest ega teistest tunnustest
- Kui
  - $\varphi_i$  – respondent  $i$  tõenäosus vastata
  - $x_i$  – mingi andmestikus olev tunnus
  - $y_i$  – vaadeldav tunnus (kas esineb andmelünk või mitte)
- siis  $\varphi_i$  ei ole seotud  $x_i$ ,  $y_i$  ega ühegi teise tunnusega
- Näide kehakaalu tunnuses esinevate lünkade kohta:
  - sugu ei ole seotud lünkade esinemisega kaalu tunnuses, st naiste ja meeste hulgas ei erine kaalu ütlejate jätjate osakaal,
  - ka kergemate ja raskemate inimeste tõenäosuses kaal öelda või ütlejate jätjate erinevust ei ole
- Populatsiooni kohta tehtavad järeldused on nihketa (muidugi eeldusel, et tegu on tõenäosusliku valimisega => esineb siiski valikuviga)
- Kui mittevastamise viga ignoreeritakse, siis sisuliselt eeldatakse MCAR

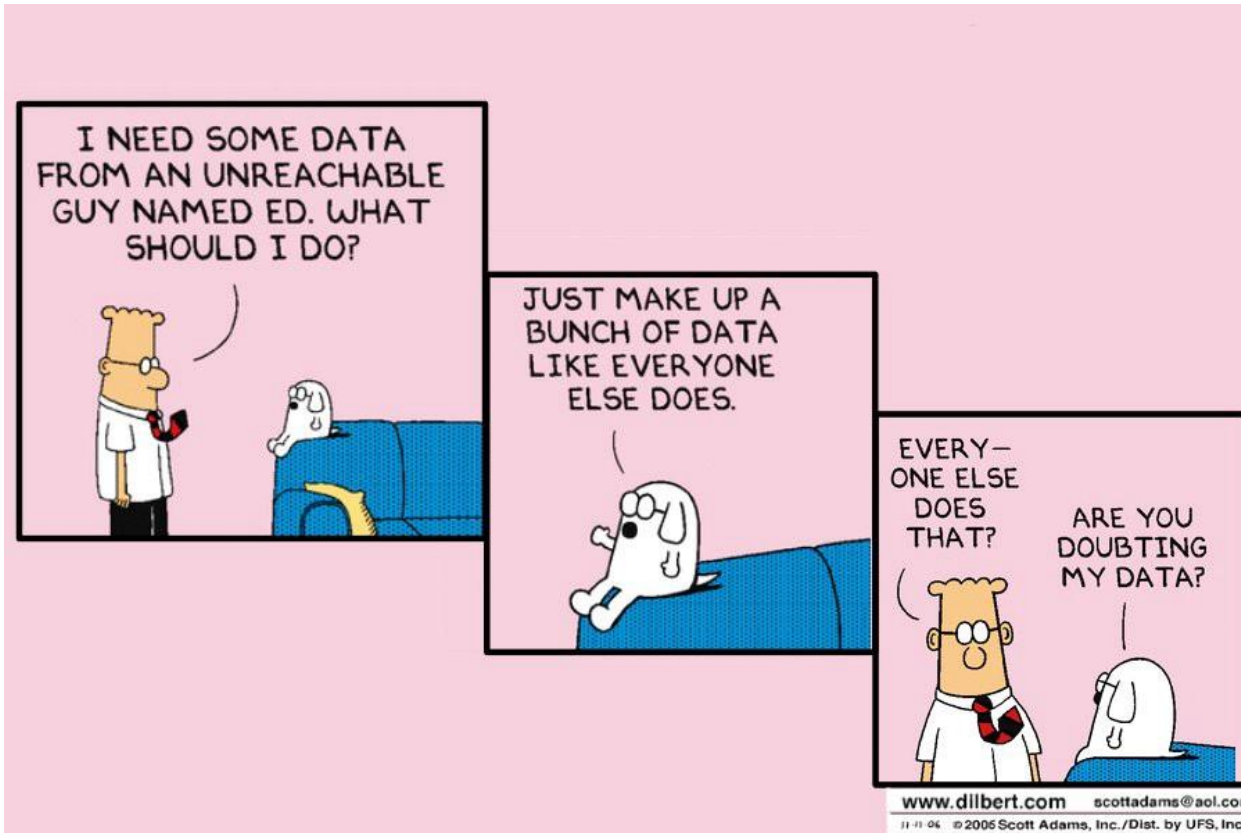
# Andmelünkade (mittevastamise) liigid

- *Missing at random (MAR)* ehk juhuslikud lüngad
- Andmelünkade esinemine ei sõltu uuritavast tunnusest
- Kui
  - $\varphi_i$  – respondent  $i$  tõenäosus vastata
  - $x_i$  – mingi andmestikus olev tunnus
  - $y_i$  – vaadeldav tunnus (kas esineb andmelünk või mitte),
- siis  $\varphi_i$  (respondent  $i$  tõenäosus vastata) on seotud  $x_i$ -ga, aga mitte  $y_i$ -ga
- Näide kehakaalu tunnuses esinevate lünkade kohta:
  - naised jätavad uuringus oma kaalu sagedamini ütlemata kui mehed (seos  $x_i$ -ga),
  - samas kergemate ja raskemate inimeste tõenäosuses kaal öelda või ütlemata jätta erinevust ei ole ( $y_i$ -ga ehk tunnuse endaga seost pole)

# Andmelünkade (mittevastamise) liigid

- *Not missing at random (NMAR)* ehk mittejuhuslikud lüngad
- $\varphi_i$  (respondent  $i$  tõenäosus vastata) on seotud  $y_i$ -ga, seda ei ole võimalik täielikult seletada  $x_i$  abil
- Lüngad on seotud tunnuse enda (esile tulemata jäänud) väärtustega ja teiste tunnustega
- Lünkade tekkemehhanism ei ole olemasolevate tunnuste varal kirjeldatav
- Näide kehakaalu tunnuses esinevate lünkade kohta:
  - naised jätavad uuringus oma kaalu sagedamini ütlemata kui mehed ja
  - raskemad inimesed jätavad oma kaalu sagedamini ütlemata
- Mittevastamist ei saa ignoreerida
- Longituudsetes andmetes / aegridades probleem tõsisem
- NMAR on keeruline või isegi võimatu tuvastada, põhimõtteliselt ainult populatsioonistatistika, muu kvaliteetse uuringu või kordusuuringu abil (Valliant et al 2013)

# Mida andmelünkadega teha?

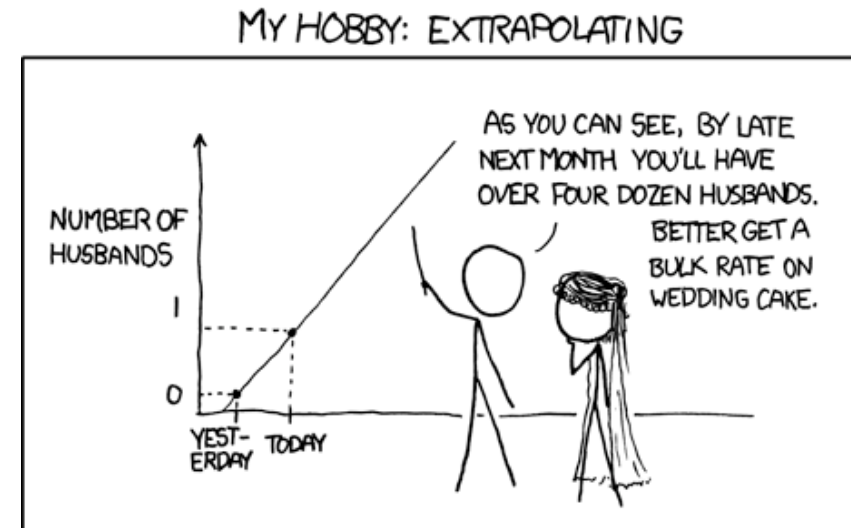


# Andmete kaalumine

- Oluline:
  - Andmete kaalumine võimaldab esinduslikkuse kadu vähendada, kui kehtib MCAR või MAR
  - Kui kehtib NMAR, võib kaalumine esinduslikkuse kadu vähendada, aga võib ka suurendada
  - Aga mida ütles just eelmise slaidi viimane punkt?
  - Võiks öelda, et laias laastus (subjektiivselt) saab siiski hinnata, kuivõrd kaalumisele saab lootma jääda
- NB! Esinduslikkusest saab rääkida
  - sihtpopulatsiooni suhtes
  - mingite tunnuste lõikes
- Kui konkreetseid tunnuseid ei mainita, eeldatakse esinduslikkust üleüldiselt
  - st kõikvõimalike tunnuste suhtes
- Kaalumisjärgselt saab andmete esinduslikkust kindlalt väita vaid tunnuste kohta, mida kasutati kaalumisel
- Esinduslikkus teiste tunnuste suhtes:
  - mida suurem on mittevastamise nihe kaalumisel kasutatud tunnustes, seda tõenäolisemalt püsib (või suureneb) esinduslikkuse kadu teistes tunnustes
- Kaalumine mõeldud vähendama kutsele mittevastamisest tulenevad esinduslikkuse kadu
- Küsimusele mittevastamisest tulenev esinduslikkuse kadu ikkagi probleem

# Mida teha küsimusele mittevastamisest tuleneva esinduslikkuse kaoga?

- ...ehk andmelünkadega, mis meil olemasolevas andmestikus esinevad?
- On erinevaid imputeerimise (andmelünkade valiidsete väärtustega) asendamise viise
- Lühidalt:
  - traditsioonilised viisid enamasti liiga ebatäpsed
  - täpsed meetodid antud kursuse jaoks liiga keerulised (et neid asjatundlikult kasutada)
    - nt mitmene imputeerimine
    - Kui andmelüngad on NMAR, ei saa mitmest imputeerimist kasutada



# Mida teha küsimusele mittevastamisest tuleneva esinduslikkuse kaoga?

- ...ehk andmelünkadega, mis meil olemasolevas andmestikus esinevad?
- On erinevaid imputeerimise (andmelünkade valiidsete väärtustega) asendamise viise
- Lühidalt:
  - traditsioonilised viisid enamasti liiga ebatäpsed
  - täpsed meetodid antud kursuse jaoks liiga keerulised (et neid asjatundlikult kasutada)
    - nt mitmene imputeerimine
    - Kui andmelüngad on NMAR, ei saa kasutada mitmest imputeerimist
- Kas saab andmelünkadega indiviidid analüüsist lihtsalt välja jätta?
- Ainult juhul, kui andmelüngad esinevad täiesti juhuslikult (MCAR)
- Kui andmelüngad on MAR või NMAR, ei saa andmelünkadega indiviide lihtsalt analüüsist välja jätta
- Kui andmelünkade hulk on väga väike (2-3%, vb 5%), on tõenäolisem, et nad on (täiesti) juhuslikud või et nende mõju tulemustele on väike
- Kas esinevad andmelüngad on MAR või MCAR?
  - Võrrelda oluliste tunnuste jaotuseid andmelünkadega ja valiidsete väärtustega indiviidide seas
  - $\chi^2$ -test, sõltumatute kogumite t-test, Little'i test

# Mida teha küsimusele mittevastamisest tuleneva esinduslikkuse kaoga?

- Mida teha siis, kui ilmneb, et lüngad ei esine täiesti juhuslikult ( $\neq$ MCAR)?
- Kas tuleks andmetest loobuda?
- Mitte tingimata
- Mingi viga esineb andmetes alati, küsimus on, kui suurt viga oleme valmis lubama
- Teisiti öeldes, kui ettevaatlikud peaksime tulemuste tõlgendamisel olema?
- Oluline olla andmetes esinevatest probleemidest teadlik ja neid arvesse võtta
- ...ning anda neist lugejale teada!
- Veel oluline:
  - erinevad analüüsid, erinevad küsimused nõuavad erinevat täpsust
  - oskus viga hinnata tuleb aja ja analüüsikogemusega
  - saame kasutada erinevaid näitajaid vea hindamisel, aga teatud ulatuses otsus subjektiivne



# Erindid

- Erind – teatava kriteeriumi kohaselt skaala teistest väärtusest märkimisväärselt erinev väärtus
  - Milline on märkimisväärne erinevus?
  - Pole üheselt defineeritav
  - On erinevaid kriteeriume (nt erinevus 2, 2,5, 3 vms standardhälvet keskmisest)
  - Millise kriteeriumi ja lävendi kasuks otsustada – hinnanguline, sõltub kontekstist

# Erindid: miks oluline?

- Miks on oluline erindid tuvastada ja nende osas midagi ette võtta?
  - Erindite võimalikud põhjused
    - vead andmetes
      - juhuslik kõrvalekalle või süstemaatiline nihe
    - heterogeensus andmetes (erinevad distinktiivsed alamrühmad)
  - Võivad oluliselt mõjutada/kallutada tunnuse põhjal arvutatavaid näitajaid
    - Nt tunnuse keskmine, standardhälve, dispersioon
  - Võivad seoste uurimisel viia I või II tüüpi veani
  - Võivad viia uue sisulise teadmise jälile

# Erindid: liigid ja võimalused tuvastamiseks

- Ühemõõtmeline erind – märkimisväärselt erinev väärtus **ühe tunnuse poolest**
  - Nt individ, kelle kuusissetulek on 10 000 eurot
  - Tuvastatavad ühemõõtmelise analüüsiga, nt
    - tunnuse enda jaotus
    - tunnuse põhjal arvutatavad jaotusparameetrid ja näitajad
- Mitmemõõtmeline erind – märkimisväärselt erinev väärtus **kahe või rohkema tunnuse kombinatsioonis**
  - Näide:
    - 15-aastane individ ei ole erind
    - 2000 eurose kuusissetulekuga individ ei ole erind
    - Küll aga on erind 15-aastane individ, kes saab kuus 2000 eurot
  - Tuvastatavad mitmemõõtmelise analüüsiga, nt
    - mitmene jaotus
    - regressioonimudeli jääkide analüüs

# Erindid: liigid ja võimalused tuvastamiseks

- Näide: ESS 2014

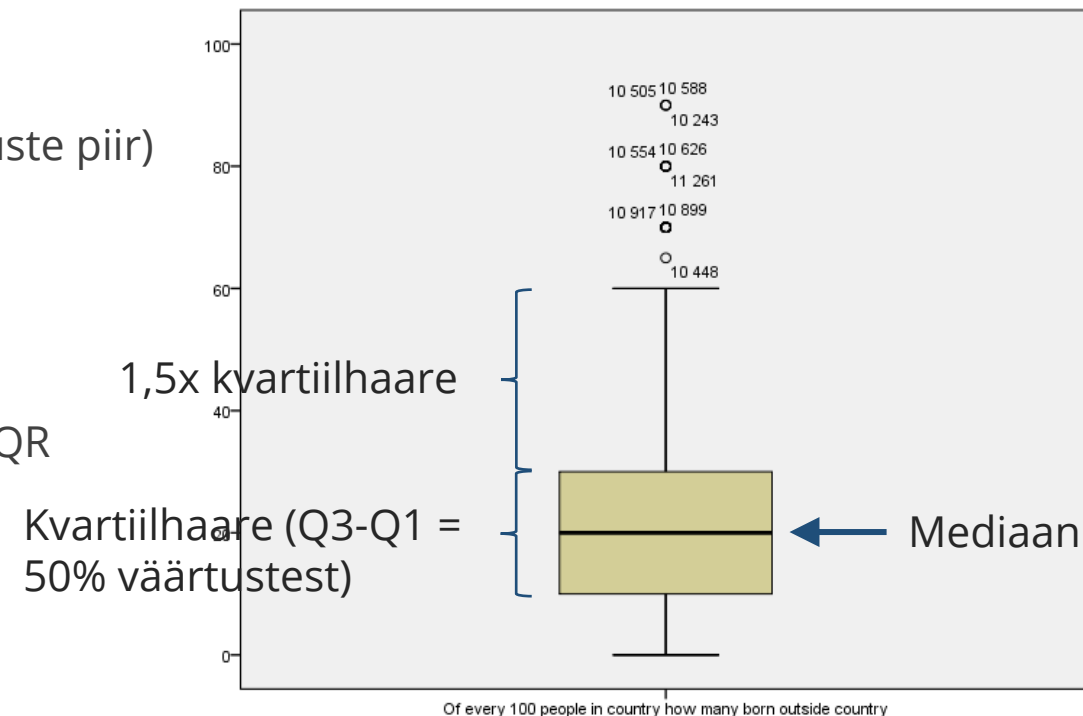
Kui mitu igast 100st Eestis elavast inimesest on Teie arvates sündinud väljaspool Eestit?

INTERVJUEERIJAL: Kui vastaja ütleb „ei oska öelda“; siis öelge: „Palun andke hinnanguline number“.

- Tunnusnoimbro: Of every 100 people in country how many born outside country

# Erindid: liigid ja võimalused tuvastamiseks

- Ühemõõtmelised erindid
  - Vihjed erindite olemasolule: suur asümmeetriakordaja, mediaani ja keskmise suur erinevus, miinimumi, maksimumi ja keskmise võrdlus
  - Tunnuse jaotus tabelis
  - Visuaalne jaotus: histogramm
  - Visuaalne jaotus: karpdiagramm
    - Põhineb variatsioonirea kvartiilidel:
    - Q1 = alumine kvartiil (alumise/esimeste 25% väärtuste piir)
    - Q2 = mediaan
    - Q3 = ülemine kvartiil
    - Q3-Q1 = kvartiilhaare (IQR, *interquartile range*)
    - Erindid:  $x_i < Q1 - 1,5 \cdot IQR \mid x_i > Q3 + 1,5 \cdot IQR$
    - Äärmuslikud erindid:  $x_i < Q1 - 3 \cdot IQR \mid x_i > Q3 + 3 \cdot IQR$



# Erindid: liigid ja võimalused tuvastamiseks

- Ühemõõtmelised erindid

- Z-skoor

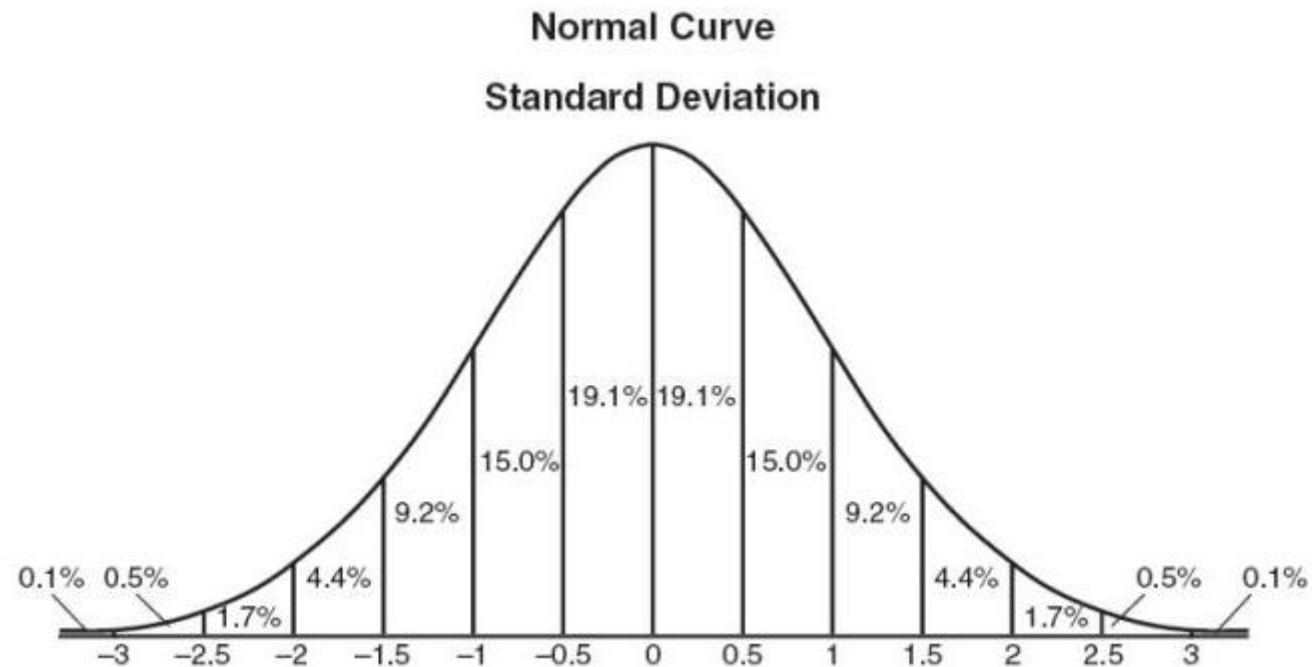
- Tunnus standardiseeritakse: tunnuse väärtuseid nihutatakse nii, et

- keskmise  $m_x$  saab väärtuse 0,
      - tunnuse väärtusi väljendatakse standardhälbe ( $s_x$ ) ühikutes

$$z = \frac{x - m_x}{s_x}$$

- Erindi lävendiks on erinevus keskmisest mõõdetuna standardhälvetes, nt
      - väärtus paikneb vähemalt/rohkem kui 2,5 või 3 või 3,5 standardhälbe kaugusel keskmisest

# Z-skoor: võimalikud erindi lävendid



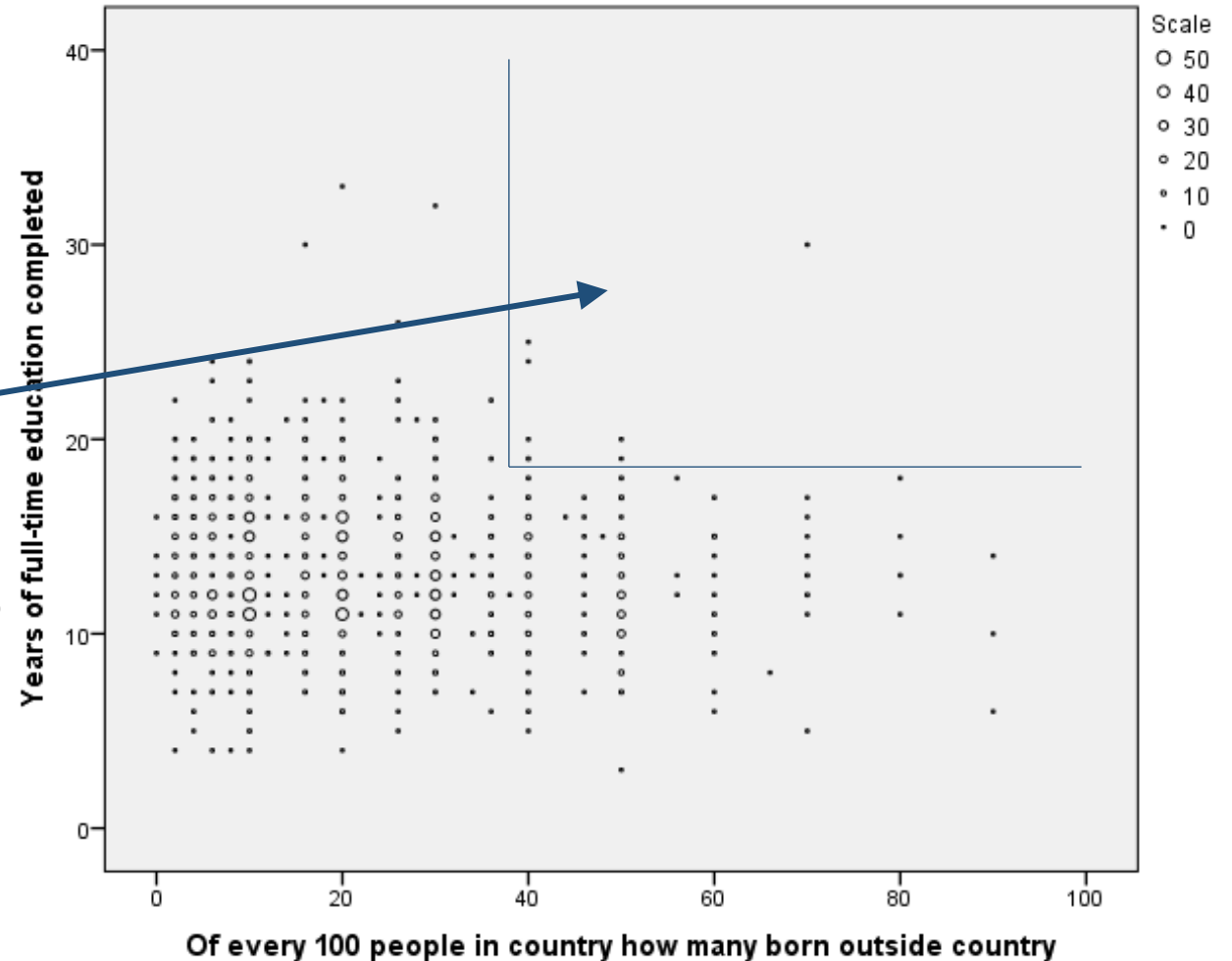
# Erindid: liigid ja võimalused tuvastamiseks

- Z-skoor
  - Standardiseeritud väärtused sõltuvad tunnuse jaotusest
  - Sh erinditest =>
  - Erindite olemasolu kahandab z-skooride „erandlikkust“ =>
  - Teatud määral sõltub see erindite tuvastamise meetod erindite olemasolust!



# Erindid: liigid ja võimalused tuvastamiseks

- Mitmemõõtmelised erindid
  - Visuaalne mitmemõõtmeline jaotus (nt hajuvusdiagramm)
- Põhjused?
  - Iseäralikud individid?
  - Kehva andmekvaliteediga individid?



# Erindid: liigid ja võimalused tuvastamiseks

- Mitmemõõtmelised erindid
  - Suured regressioonijäägid regressioonimudelis
  - Erindite algpõhjuste leidmiseks teha indikaatortunnus erindite ja normaalväärtuste eristamiseks => võrrelda muude tunnuste jaotuseid või keskmisi

# Erindid: põhjused ja käsitusviisid

- 1) Andmesisestusviga
  - teatud ulatuses võimalik järelkontrollida või ennetamiseks seadistada kriteeriumid
  - kui on selge, et tegu sisestusveaga...
    - ja võimalik tuvastada täpne väärtus => sisestada täpne väärtus
    - pole võimalik tuvastada täpset väärtust => vastus võimalik kustutada
- 2) Andmelünga kood jäetud lüngana defineerimata
  - kontrollimisel reeglina lihtsasti tuvastatav
  - defineerida andmelüngana
- 3) Ülekaetuse viga
  - respondendi vastused kustutatakse
- 4) Respondent on sihtpopulatsiooni esindaja, kellel ongi tunnuses ebatavaliselt erandlik väärtus –
  - tunnuse teisendamine
    - puhtalt paari erindi pärast pole mõttekas, kui muidu tunnus enam-vähem normaaljaotuse lähedane
  - võimalik respondent alles jätta, aga muuta erindi väärtust (*Winsorising*)
    - väärtusel väiksem mõju analüüsitulemustele
    - subjektiivne