





Sotsiaalse analüüsi meetodid:  
kvantitatiivne lähenemine

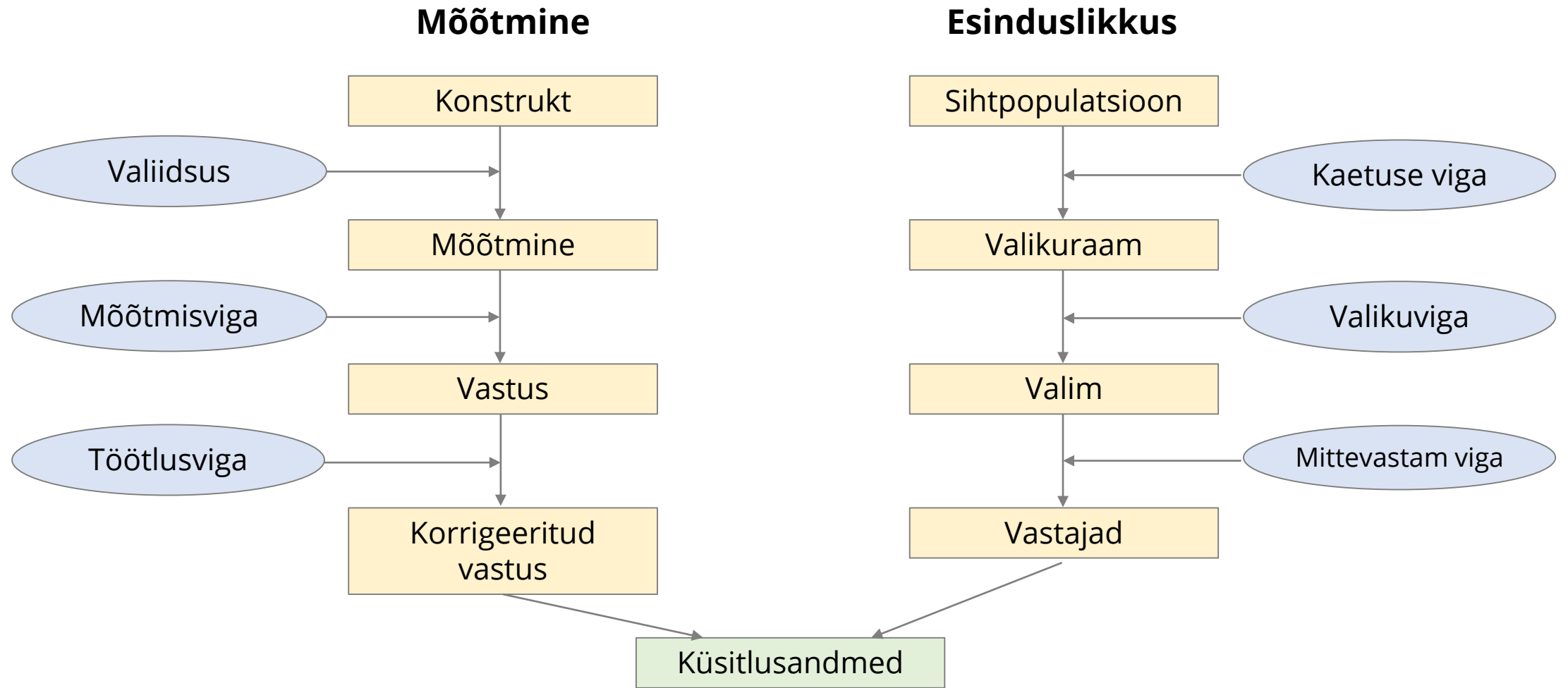
Andmete kvaliteedi hindamine,  
puuduvad väärtused, erindid

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, mida püütakse hinnata usaldusvahemike, olulisuse tõenäosuse abil
  - 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



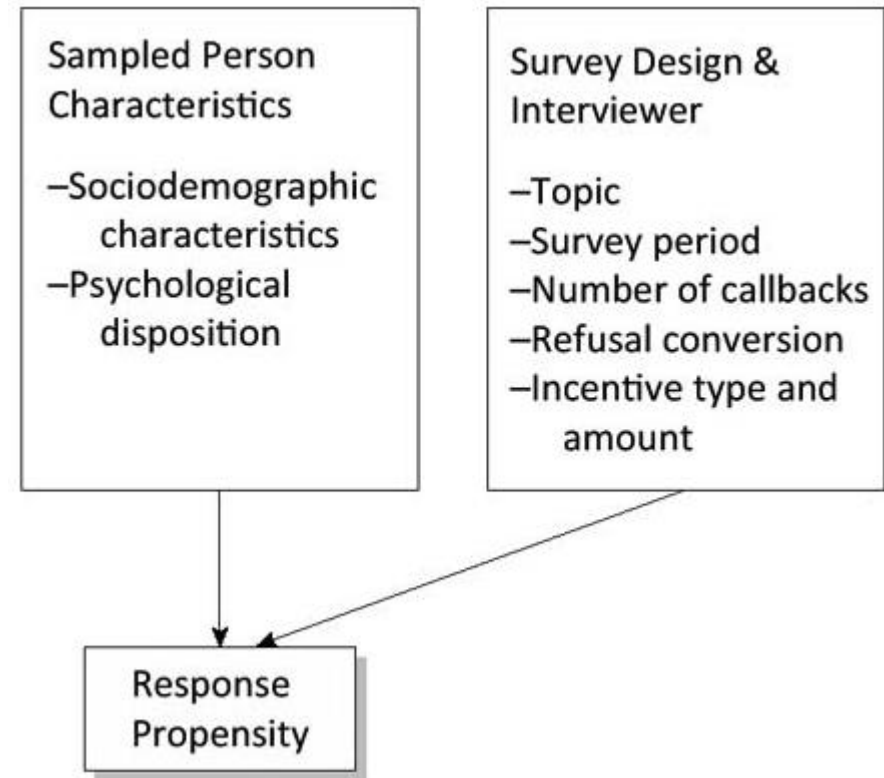
# Andmete esinduslikkus, tulemuste üldistatavus

- 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 ja esinduslikkusest
- 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
  - $\varphi$  – tõenäosus vastata
  - $x$  – mingi andmestikus olev tunnus
  - $y$  – vaadeldav tunnus (tunnus, milles esinevate andmelünkade juhuslikkusest oleme huvitatud)
- siis  $\varphi$  ei ole seotud  $x, y$  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 jätta 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
- See on paraku reeglina ebarealistlik

# Andmelünkade (mittevastamise) liigid

- *Missing at random (MAR)* ehk juhuslikud lüngad
- Andmelünkade esinemine ei sõltu uuritavast tunnusest
- Kui
  - $\varphi$  – tõenäosus vastata
  - $x$  – mingi andmestikus olev tunnus
  - $y$  – vaadeldav tunnus (tunnus, milles esinevate andmelünkade juhuslikkusest oleme huvitet),
- siis  $\varphi$  (tõenäosus vastata) on seotud  $x$ -ga, aga mitte  $y$ -ga
- Näide kehakaalu tunnuses esinevate lünkade kohta:
  - naised jätavad uuringus oma kaalu sagedamini ütlemata kui mehed (seos  $x$ -ga),
  - samas kergemate ja raskemate inimeste tõenäosuses kaal öelda või ütlemata jätta erinevust ei ole
    - st lünkade esinemine tunnuses  $y$  ei sõltu vaatlemata jäänud väärtustest selles tunnuses

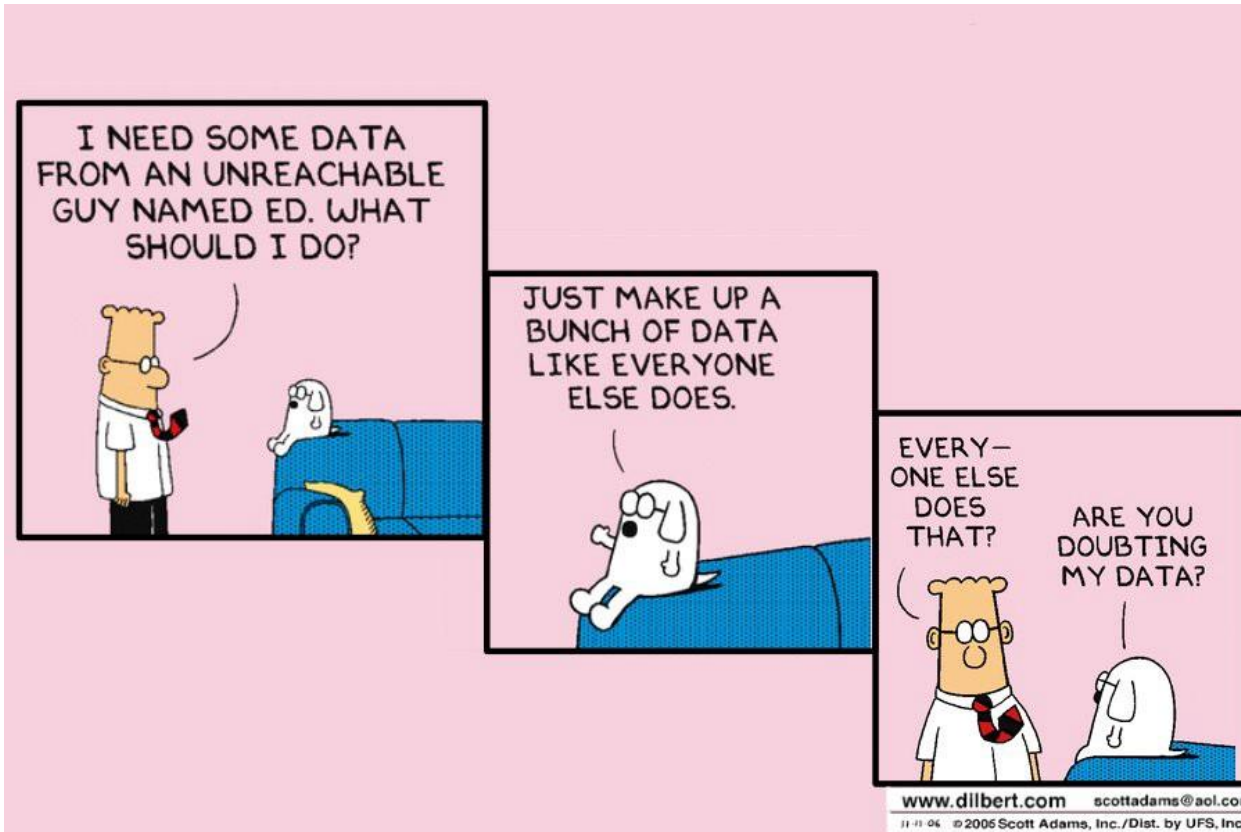
# Andmelünkade (mittevastamise) liigid

- Eelnevas MAR näites on lünkade esinemine seotud sooga
- Kuidas saab siis öelda, et lünklikkus on juhuslik?
- Tõesti mõnevõrra eksitav termin – justkui võiks lünklikkust ignoreerida
- Parem oleks *Missing Conditionally at Random*, aga akronüüm ajaks asja päris segaseks...
- Silmas peetakse seda, et
  - Lünklikkus ei sõltu tunnuse enda mõõtmata väärtustest, vaid muudest tunnustest =>
  - Kui võtame arvesse lünklikkust tekitavad tegurid (need muud tunnused), saab esinduslikkuse kadu vältida
  - Kuidas, sellest räägime natuke hiljem

# Andmelünkade (mittevastamise) liigid

- *Missing not at random (MNAR)* ehk mittejuhuslikud lüngad
- $\varphi$  (tõenäosus vastata) on seotud  $y$ -ga ja seda ei ole võimalik täielikult seletada  $x$  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
- MNAR on keeruline või isegi võimatu tuvastada, põhimõtteliselt ainult muu kvaliteetse uuringu või kordusuuringu abil (Valliant et al 2013)

# Mida andmelünkadega teha?



# Täielik mittevastamine => kaalumine

- Täielikust mittevastamisest ehk kutsele mittevastamisest tulenevat esinduslikkuse kadu saab vähendada andmete kaalumise (nt järelkihistamiskaaludega)
- Vastajate seas alaesindatud gruppidele antakse analüüsis suurem kaal
- Üleesindatud gruppidele antakse analüüsis väiksem kaal
- NB! Andmete kaalumine võimaldab esinduslikkuse kadu vähendada, kui kehtib MCAR või MAR
- Kui kehtib MNAR, 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

# Täielik mittevastamine => kaalumine

- 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, mis on kaalude arvutamiseks aluseks
  - Tavaliste järelkihistamiskaalude puhul tähendab see umbes 3-5 tunnust
- Kaalumine parandab esinduslikkust teiste tunnuste suhtes:
  - niivõrd, kuivõrd andmelünkade esinemine teistes tunnustes on lineaarselt seotud järelkihistamiskaalude aluseks olevate tunnustega
- Tähendabki seda, et lünkade esinemine peab olema MAR
  - ehk lünklikkus ei ole küll täiesti juhuslik, st on seotud mingite tunnustega, aga võttes lünklikkust nendes tunnustes arvesse, saame lünklikkuse mõju esinduslikkusele elimineerida
  - selle mõju täielik elimineerimine eeldab, et lünklikkus muudes tunnustes on kas täiesti juhuslik või täielikult seletatav kaalumise aluseks olevate tunnustega

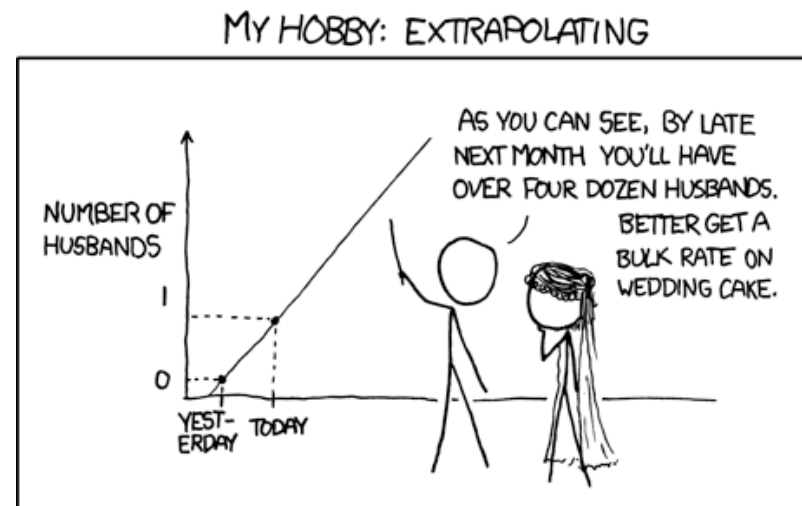
# Täielik mittevastamine => kaalumine

- Kas see on realistlik kõigi andmestikus olevate tunnuste suhtes?
- Ilmselt mitte, aga laias laastus sõltub küsitlusjärgsete (st kaalumata) andmete kvaliteedist (esinduslikkusest)
- Mida suurem on mittevastamise nihe enne andmete kaalumist, seda tõenäolisemalt püsib (või suureneb) esinduslikkuse kadu pärast kaalumist
- Kaalumine võib vähendada andmete esinduslikkust – kuidas on see võimalik?
- Seega, andmete kaalumine ei ole imerohi
- Siiski on sellest pigem kasu kui kahju, KUI andmekogumine on olnud metodoloogiliselt kvaliteetne
- MOTT: oluline on
  - teada, kuidas andmed on kogutud
  - tunda andmekogumise meetodeid
  - osata hinnata andmekvaliteeti



# Mida teha küsimusele mittevastamisest tuleneva esinduslikkuse kaoga?

- Kaalumise mõeldud vähendada kutsele mittevastamisest tulenevad esinduslikkuse kadu
- Küsimusele mittevastamisest tulenev esinduslikkuse kadu ikkagi probleem
- Seega, mida teha 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 MNAR, ei saa mitmest imputeerimist kasutada



# Mida teha küsimusele mittevastamisest tuleneva esinduslikkuse kaoga?

- 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 andmelünkadega indiviidide osakaal on väga väike, on tõenäolisem, et andmelüngad on (täiesti) juhuslikud või et nende mõju tulemustele on väike
- Mis on „väga väike“ osakaal, on subjektiivne – 2-3%, vb 5%
- Tegelikuses sõltub jällegi andmete üleüldisest kvaliteedist/esinduslikkusest
- Kas esinevad andmelüngad on MAR või MCAR?
  - Võrrelda oluliste tunnuste jaotuseid andmelünkadega ja valiidsete väärtustega indiviidide seas
  - sõltumatute kogumite  $t$ -test, Little'i test,  $\chi^2$ -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