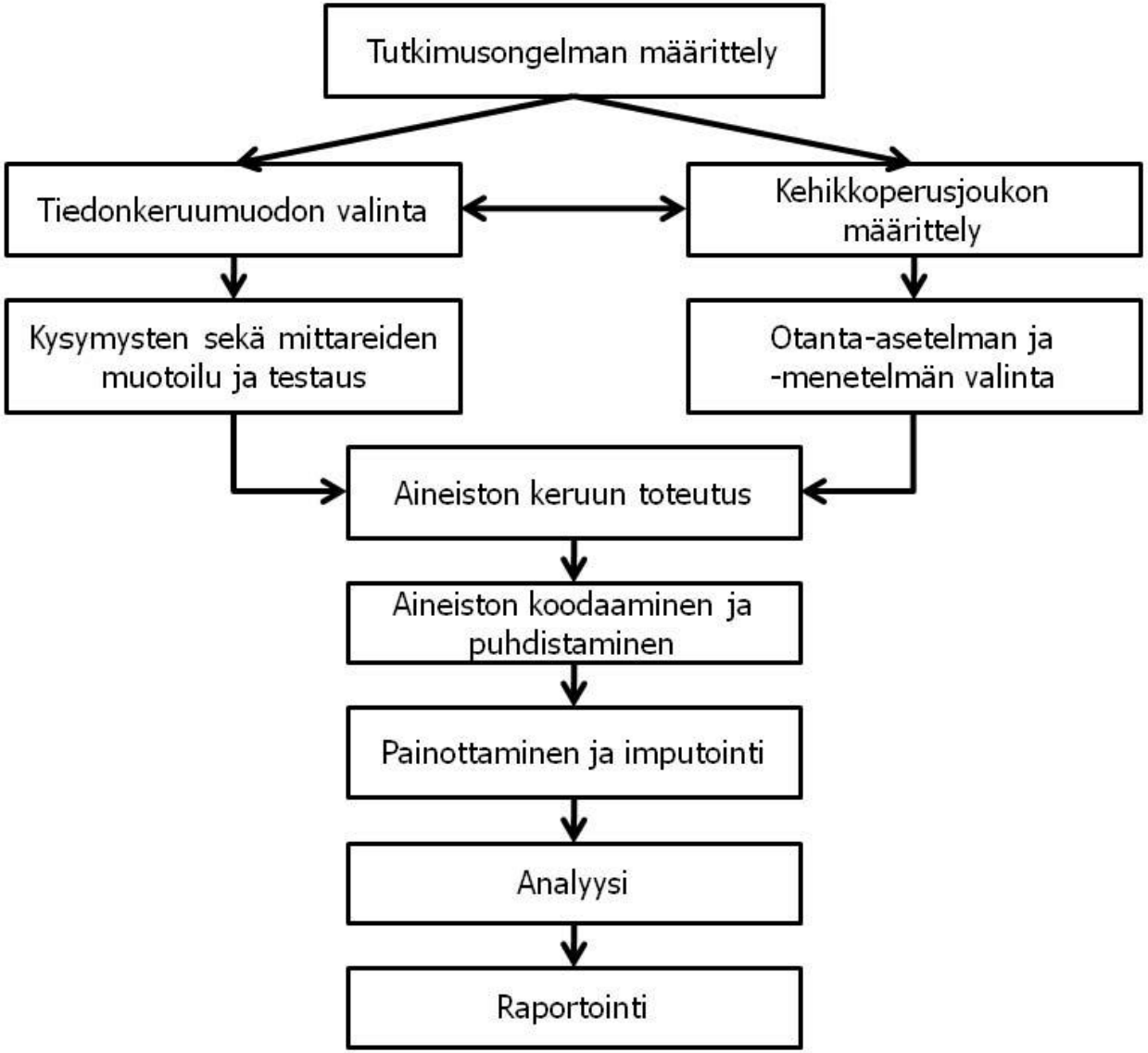
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Sosiaalitutkimuksen tilastolliset menetelmät Osa 1 – Diat 2 Otanta-asetelmat ja survey-aineiston käsittely

Risto Lehtonen, Helsingin yliopisto
risto.lehtonen@helsinki.fi

Otanta-asetelma tutkimusasetelman osana



Kuvio 2.2. Kyselytutkimuksen prosessi (Groves et al. 2009, s. 149).

Survey-prosessin työvaiheet - 1

- **Otanta-asetelman laadinta ja dokumentointi**
  1. Otanta-asteiden kiinnittäminen
     - Yksiasteinen, kaksiasteinen, moniasteinen
  2. Kullekin otanta-asteelle määritellään:
     - Perusjoukot: Tavoiteperusjoukko, kohdeperusjoukko, kehikkoperusjoukko
     - Otantamenetelmä ja otoskoko
  3. Otanta-asetelman dokumentointi käyttäjiä varten
- **Otoksen poiminta kehikkoperusjoukoista**
  1. Otoksen poiminta valitulla menetelmällä
  2. Otostiedoston muodostus (SPSS, SAS, Excel)

Otanta-asetelma *sampling design*

- Otanta-asetelma on niiden sääntöjen ja menetelmien kokonaisuus, jolla **otos** poimitaan määritellystä **perusjoukosta**
- Otos = Perusjoukon osajoukko
- Otos poimitaan jollain **satunnaisotannan** eli todennäköisyysotannan menetelmällä (*Random sampling, Probability sampling*)
- Otannassa käytetään perusjoukon alkioiden **sisältymistodennäköisyyksiä**
- **Asetelmapaino** = 1 / sisältymistodennäköisyys

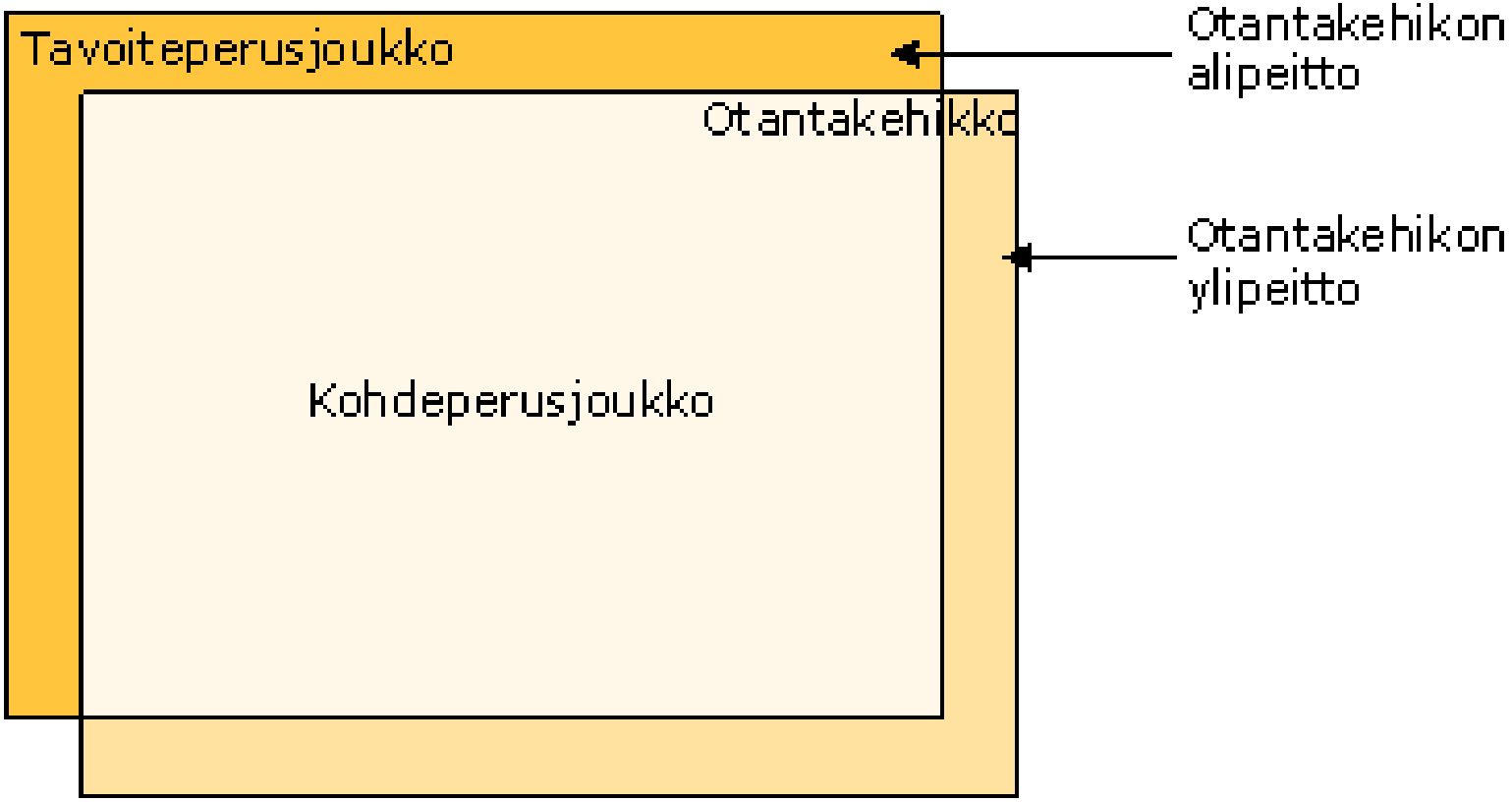
ESS - Tavoiteperusjoukko

- **All persons aged 15 and over resident within private households, regardless of their nationality, citizenship, language or legal status**, in the following participating countries: European Union countries - Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech …jne.

  ….

  Non-European Union countries -
  Israel, Norway, Switzerland, Russian Federation, Ukraine.

Otantakehikon alipeitto ja ylipeitto
Tilastokeskus: Laatua tilastoissa -käsikirja



Survey-prosessin työvaiheet - 2

- **Aineiston keruun toteutus**
  - Haastattelut, CAPI, CATI, PAPI,…
- **Tutkimusaineiston muodostus** (SPSS, SAS)
  1. Asetelmapainojen muodostus
  2. Vastauskadon analyysi
  3. Aineiston tarkistaminen ja editointi
  4. Puuttuvien tietojen imputointi (tarvittaessa)
  5. Uudelleenpainotus ja analyysipainojen muodostus
  6. Asetelmaindikaattoreiden liittäminen aineistoon
     - Ositeindikaattorit, ryväsindikaattorit
  7. Lisätietojen liittäminen aineistoon (rekisteritiedot)
- **Tilastollinen analyysi ja raportointi**

Miksi satunnaisotanta?

- Otoksesta saatavat tulokset voidaan **yleistää tilastollisen päättelyn keinoin** koskemaan koko kiinnostuksen kohteena olevaa perusjoukkoa (tai hypoteettista mallia)
- Tilastollinen päättely
  - Piste-estimaatit, esim. keskiarvot
  - Keskivirheet ja luottamusvälit
  - Tilastolliset testit
  - Tilastollinen mallinnus

Alkiotasoiset otanta-asetelmat
Element sampling

- Otantayksikkönä on perusjoukon alkio (*element, unit*)
  Esim. henkilö, yritys, maatila,…
- Otos poimitaan valitulla otantamenetelmällä suoraan perusjoukon alkioiden muodostamasta kehikkoperusjoukosta
  - Henkilöotos väestörekisteristä
  - Yritysotos yritysrekisteritä
  - Maatilaotos maatilarekisteristä

Ryväsotannan asetelmat
Cluster sampling

- Otantayksikkönä on perusjoukon alkioiden muodostama luonnollinen ryhmä eli ryväs (*cluster*)
- Esim: PISA
  Ryväsyksikkönä koulu (opetusryhmä)
- **Ryvästason otanta**
  Poimitaan kouluotos koulujen perusjoukosta
- **Alkiotason otanta**
  Poimitaan oppilasotokset (esim. opetusryhmä, luokka) kouluotokseen tulleista kouluista

Otanta-asetelma voi olla…

- Yksinkertainen
  - **Systemaattinen otanta**
    - Poiminta suoraan alkiotason kehikko-perusjoukosta
  - **Ositettu systemaattinen otanta**
    - Alkioiden ositus ja ositteiden kiintiöinti
    - Systemaattinen otanta kustakin ositteesta
- Mutkikas
  - **Ositettu kaksiasteinen ryväsotanta**

    1. aste: Rypäiden poiminta ryvästason perusjoukosta esim. PPS-otannalla

    2. aste: Alkioiden poiminta otosrypäistä systemaattisella otannalla

Otanta-asetelma: ESS 2010 - Suomi

- **Sampling procedure: Finland**
- Sampling frame:
  - Population database (total register).
  - Foreign citizens are included if they have residency status.
- Sampling Design
  - **Single stage equal probability systematic sample** (no clustering).
  - Implicit stratification by region, sex and age.

Tiivistelmä: Otantamenetelmät I

Otantamenetelmä	Poimintatapa
SRS *Simple random sampling* Yksinkertainen satunnaisotanta	Otos poimitaan perusjoukosta satunnaislukujen avulla
SYS *Systematic sampling* Systemaattinen otanta	Otos poimitaan tasavälisesti listasta tai rekisterinä olevasta tietokannasta
STR *Stratified sampling* Ositettu otanta	Perusjoukon alkiot jaetaan ensin homogeenisiin ositteisiin. Kustakin ositteesta poimitaan SRS tai SYS otos

Tiivistelmä: Otantamenetelmät II

Otantamenetelmä	Poimintatapa
CLU *Cluster sampling* Ryväsotanta	Perusjoukon alkiot muodostavat luonnollisia osajoukkoja eli rypäitä
- Yksiasteinen *one-stage*	1)Rypäiden perusjoukosta poimitaan otosrypäät 2) Kaikki otosrypäiden alkiot tulevat alkiotason otokseen
- Kaksiasteinen *two-stage*	1) Rypäiden perusjoukosta poimitaan otosrypäät 2) Otosrypäiden alkioista poimitaan alkiotason otokset SRS:llä tai SYS:llä
PPS *Selection with Probabilities Proportional to Size*	Sisältymistodennäköisyys on suhteessa alkion kokoon

Mitä tietoja aineistossa tulee olla pätevää analyysia varten?

- Otanta-asetelman mukaiset muuttujat
  1. **Asetelmapaino**
     Sisältymistodennäköisyyden käänteisluku
  2. **Analyysipaino**
     Skaalattu asetelmapaino, tarvittaessa katokorjattu
     Keskiarvo yli aineiston = 1
  3. **Ositeindikaattori**
     Osoittaa, mihin ositteeseen havaintoyksikkö kuuluu
  4. **Ryväsindikaattori**
     Osoittaa, mihin poimintarypääseen havainto kuuluu
- Tarvittaessa myös indikaattorimuuttuja, joka kertoo onko muuttujan tieto **imputoitu** vai ei

ESS – Painotuksen tarve analyysissa

- **Käytännössä kaikkia tietoja 1-4 ei aina välttämättä tarvita tai ei voida käyttää**
- Tarve vaihtelee maittain ja riippuu asetelman yksinkertaisuudesta / monimutkaisuudesta
- Käyttömahdollisuus riippuu siitä, mitä tietoja (muuttujia) analyysitiedostossa on!
- ESS-dokumentti
  Weighting European Social Survey Data
- Kysymyksiä ja vastauksia, esim:
- *Do tables run on the ESS website need to be weighted? - Almost certainly yes.*

ESS – Painomuuttujien käyttö - 1

- **Analyysipaino** DWEIGHT
- Keskiarvo yli aineiston = 1
- Kompensoidaan
  - Erisuurien sisältymistodennäköisyyksien vaikutus
  - Vastauskadon vaikutus (ESS: ei ole tehty)
- ESS-suositus:
- DWEIGHT pitää käyttää aina maakohtaisissa vertailuissa
- Jos DWEIGHT = 1 datan kaikille henkilöille, niin painomuuttujaa ei tarvitse käyttää (esim. Suomen aineisto osana kv. aineistoa)

ESS – Painomuuttujien käyttö - 2

- **Väestöosuuspaino** PWEIGHT
  - Vaihtelee maakohtaisesti
  - Kompensoidaan maiden väestöjen kokoerot
- **Yhdistetty painomuuttuja**
  DWEIGHT*PWEIGHT
- ESS-suositus: Yhdistettyä painomuuttujaa pitää käyttää usean maan aineistoista yhdistetyn datan analyysissa
  - Esim: EU-maiden yhdistelmäaineiston analyysi
- ESS-suosituksia

ESS 2010 – Otanta-asetelmat ja aineistot

- **Otanta-asetelmat vaihtelevat maittain**
  - Maakohtaiset otoskoot likimain yhtäsuuria
  - Yksinkertaiset / Monimutkaiset otanta-asetelmat
  - Suomi: Systemaattinen otanta
  - Ranska: Kolmiasteinen ryväsotanta
- **Vastauskadon määrä vaihtelee maittain**
  - Suomi: Otoskoko 3200 henkilöä
    Saatu aineisto: 1878 henkilön haastattelutiedot
  - Osallistumisprosentti: 59.5 %
  - Ranska: 47 %
- ESS5 - 2010 DOCUMENTATION REPORT

ESIM. 1: Suomen ESS-data 2010

- Suomen ESS-aineisto: Yksinkertainen tilanne
  - Systemaattinen otanta
  - Sisältymistodennäköisyydet samoja kaikille
  - Implisiittinen ositus (pj:n lajittelu ennen otantaa)
  - **Maakohtainen analyysi:**
    Analyysipaino DWEIGHT = 1 kaikille
  - **Painomuuttujaa ei tarvita Suomen aineiston analyysissa, kun Suomi on osa kv. dataa**
  - **Yhdistelmäaineistojen analyysi**:
    Suomen väestöosuuspaino (vakio kaikille)
    Population size weight PWEIGHT = 0.24
    Painomuuttuja tarvitaan!

ESIM. 2: Ranskan ESS-data 2010

- Ranskan ESS-aineisto: Mutkikkaampi tilanne
  - Kolmiasteinen ryväsotanta
    Three stage cluster sampling
  - Henkilötason sisältymistodennäköisyydet vaihtelevat
  - Analyysipainot vaihtelevat
    DWEIGHT: Mean=1, Min=0.098, Max=4.0

    Ranskan väestöosuuspaino:
    PWEIGHT = 3.1 (vakio kaikille)
- **Painomuuttujat tarvitaan analyyseissa!**

ESIM. 3: ESS - Yhdistelmäaineisto

- **Nuorten kokema onnellisuus ja koettu terveydentila**
- **Yhdistetty data** 2002-2008, **kaikki maat**, *n* = 24822
- Painomuuttuja DWEIGHT * PWEIGHT
- Huomataan, että painojen käyttö vaikuttaa tässä vähän keskiarvoihin mutta kasvattaa keskivirheitä

Koettu terveydentila	Nuorten lukumäärä otoksessa	Onnellisuuden keskiarvo (skaala 0-10)		Keskiarvon keskivirhe	
		Ei painoja	Painotettu	Ei painoja	Painotettu
Huono	3763	6.8	6.8	0.034	0.047
Keskinkertainen	12072	7.6	7.5	0.014	0.021
Hyvä	8971	8.1	8.0	0.016	0.025

ESS – Painotuksen vaikutukset yhdistelmäaineistossa

- **Vaikutus keskiarvoihin**
  - Painottamattomat ja painotetut keskiarvot voivat poiketa, jos painot vaihtelevat
  - ESS-esimerkki: Ei merkittävää vaikutusta
- **Vaikutus keskivirheisiin**
  - Painotus yleensä kasvattaa keskivirheitä
  - ESS-esimerkki: Painotetut keskivirheet huomattavasti suurempia kuin painottamattomat
- **Seurauksia**
  - Luottamusvälit suurenevat
  - Tilastollisten testien merkitsevyydet heikkenevät

Luottamusvälit

- Painojen käytön seurauksia keskiarvolle $\bar{y}$
  - **Luottamusvälit suurenevat kun keskivirhe *s.e* kasvaa**
  - **E**sim. 95 % luottamusväli:

$$\bar{y} \pm 1.96 \times s.e(\bar{y})$$

(*s.e* on keskivirhe – *standard error of mean*)

Painottamaton:	7.67 – 7.71
Painotettu:	7.55 – 7.61

Regressiokertoimen t-testit

- **Tilastollisten testien merkitsevyydet heikkenevät, kun keskivirhe *s.e* kasvaa**

- Regressiomalli $y = \beta_0 + \beta_1 x_1 + \varepsilon$

- Regressiokertoimen β_1 nollasta poikkeamisen t-testisuure

$$t(\beta_1) = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$$

ESIM. 4: ESS 2010 - Painotuksen vaikutukset, Suomi ja Ranska

- **Suomen ESS-aineiston analyysi**
- Ei vaikutusta keskiarvoihin eikä keskivirheisiin
- Miksi? Koska analyysipaino DWEIGHT = 1 kaikille henkilöille ja väestöosuuspainoa PWEIGHT ei tarvita (HUOM: Suomen data osana kv. dataa!)
- **Suomen ja Ranskan ESS-vertailuanalyysi**
  - Tarvitaan DWEIGHT
  - Vaikutus Ranskan estimaatteihin (keskivirhe)
- **Suomen ja Ranskan yhdistetyn ESS-aineiston analyysi**
  - Tarvitaan DWEIGHT*PWEIGHT

Maiden vertailu: Suomi ja Ranska

Domain Analysis: Country Unweighted analysis					
Country	**Variable**	**Label**	**N**	**Mean**	**Std Error of Mean**
Finland	**trstplc**	Trust in the police	1869	8.031568	0.038451
France	**trstplc**	Trust in the police	1726	5.630939	0.055850

Domain Analysis: Country Weighted analysis Painomuuttuja = DWEIGHT					
Country	**Variable**	**Label**	**N**	**Mean**	**Std Error of Mean**
Finland	**trstplc**	Trust in the police	1869	8.031568	0.038451
France	**trstplc**	Trust in the police	1726	5.638801	0.064014

Yhdistelmäaineisto: Suomi ja Ranska

Yhdistelmäaineiston analyysi
Unweighted analysis (painottamaton)

Variable	Label	N	Mean	Std Error of Mean
trstplc	Trust in the police	3595	6.878999	0.038973

Yhdistelmäaineiston analyysi
Weighted analysis (painotettu)
Painomuuttuja = DWEIGHT*PWEIGHT

Variable	Label	N	Mean	Std Error of Mean
trstplc	Trust in the police	3595	5.824814	0.059411