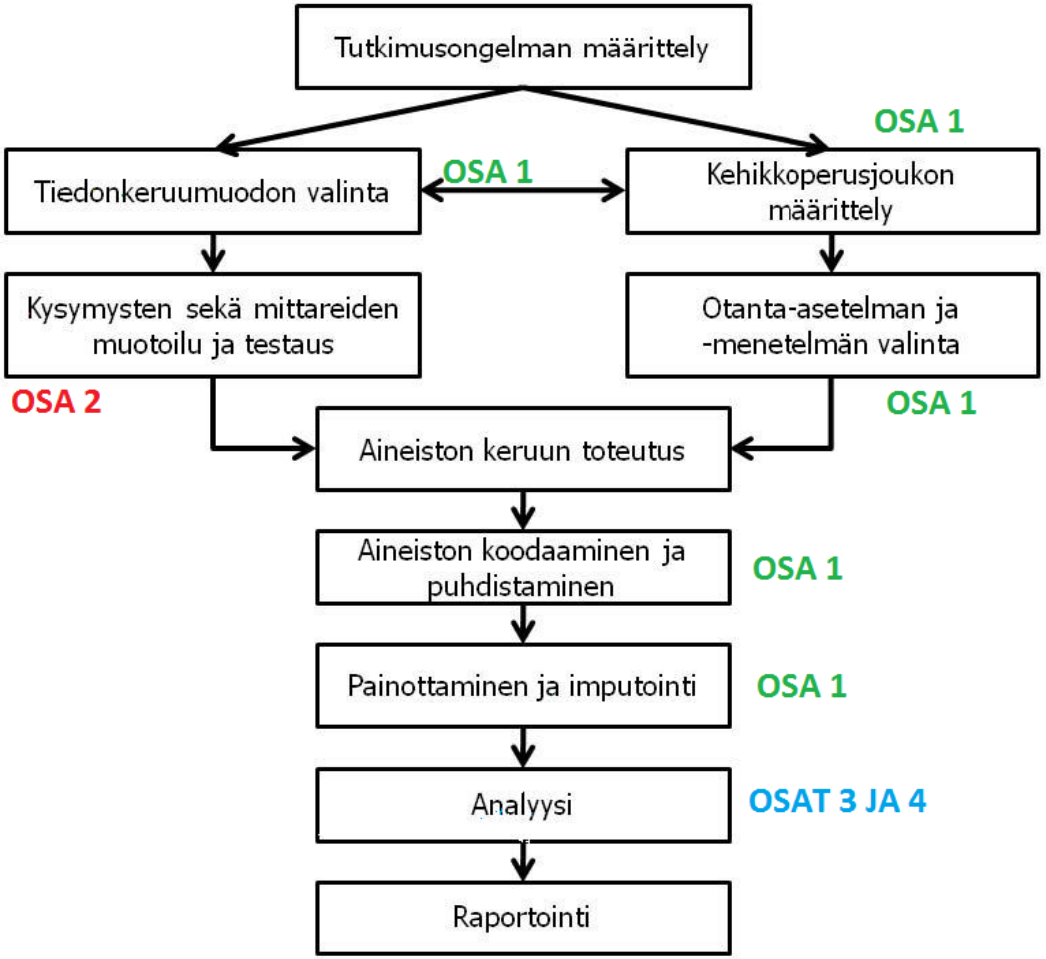
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

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

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

Versio 16.1.2015

Otosperusteinen survey: Otanta-asetelma tutkimusasetelman osana: Otanta ja painotus



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 (*Design weight*)

ESS 2010 - Tavoiteperusjoukko

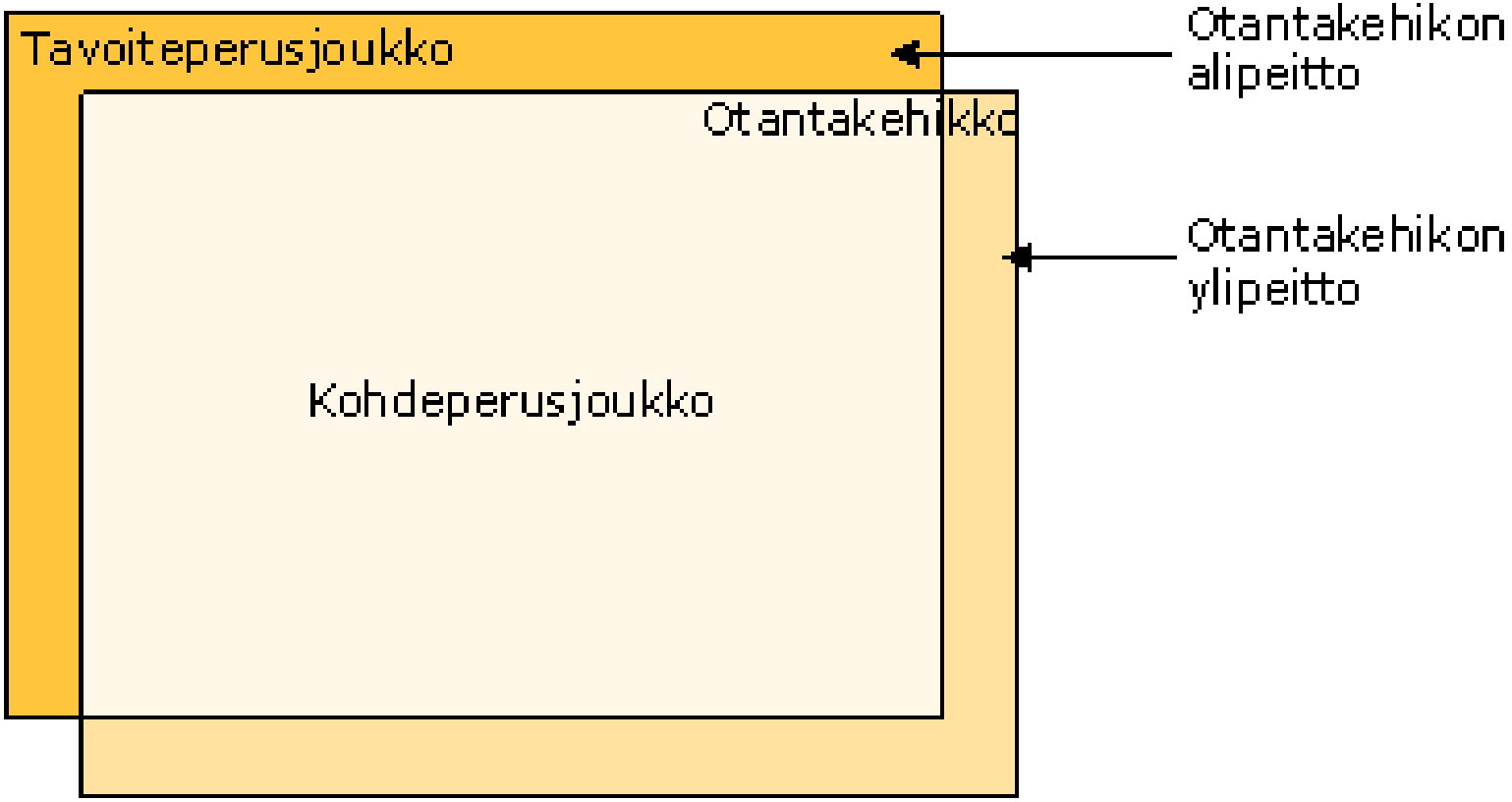
- **Universe: All persons aged 15 and over resident within private households, regardless of their nationality, citizenship, language or legal status**, in the following participating countries:

- European Union countries - Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech …jne.

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

- Katso ESS5 - 2010 DOCUMENTATION REPORT - THE ESS DATA ARCHIVE, Edition 3.2

Otantakehikon alipeitto ja ylipeitto
Tilastokeskus: Laatua tilastoissa -käsikirja



Survey-prosessin työvaiheet - 2

- **Aineiston keruun toteutus**
  - Haastattelut/Kyselyt, 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**

ESIMERKKI survey-prosessin työvaiheiden dokumentaatiosta – ESS – Suomen data

Finland

33 Data collector
37 Mode of data collection
37.1 Main questionnaire
37.2 Supplementary questionnaire
38 Type of research instrument
39 Field work period(s)
40 Geographic unit
42 Sampling procedure
43 Fieldwork procedures
43.1 Interviewer selection
43.2 Briefing of interviewers
43.3 Employment status of interviewers
43.4 Payments of interviewers
43.5 Advance information
43.6 Call schedules
43.7 Respondent incentives
43.8 Strategies for refusal conversion
43.9 Pretest
44 Control operation
44.1 Interviews
44.2 Refusals
44.3 Non-contacts
45 Cleaning operations
45.1 Consistency checks and verifications performed before deposit to the data archive
45.2 Checking and control of main questionnaire CAPI program(s)
45.3 Verification of optical scanning or keying of main questionnaire
47 Response rates
47.1 Break down of response and non response, main questionnaire
47.2 Supplementary questionnaires
48 Estimates of Sampling error
49 Weighting
50 Other study-related materials
50.6 Relationship status/legal marital status
50.6.a Legal civil union
50.6.b Cohabiting
50.6.c Legal separation
50.7 Occupation coding

- Lähde: ESS5 - 2010 DOCUMENTATION REPORT - THE ESS DATA ARCHIVE, Edition 3.2

Miksi todennäköisyysotanta?

- Otoksesta saatavat tulokset voidaan **yleistää tilastollisen päättelyn keinoin** koskemaan koko kiinnostuksen kohteena olevaa perusjoukkoa (tai hypoteettista mallia)
  - **Tilastollinen päättely**
    Kiinnostuksen kohteena olevien tunnuslukujen (keskiarvo, osuus, mediaani, regressiokerroin…)
    - Piste-estimaatit
    - Keskivirheet
    - Luottamusvälit
- Hypoteesien tilastollinen testaus
- Tilastollinen mallinnus

Alkiotasoiset otanta-asetelmat
Element sampling

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

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

Esim: Otanta-asetelma: ESS 2010 – Suomi

OTE raportista ESS5 - 2010 DOCUMENTATION REPORT - THE ESS DATA ARCHIVE - Edition 3.2

- **Sampling procedure: Finland**
- Sampling frame:
  - Population database (total register).
- Sampling Design:
  - **Single stage equal probability systematic sample** (no clustering).
  - Implicit stratification by region, sex and age.

- **Poimintaproseduuri: Suomi**
- Otantakehikko:
  - Väestötietokanta (kokonais-rekisteri).
- Otanta-asetelma:
  - **Yksiasteinen systemaattinen otanta yhtäsuurin sisältymis-todennäköisyyksin** (ei ryvästymistä).
  - Implisiittinen ositus alueen, sukupuolen ja iän mukaan.

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 tietolähteestä ja siitä, mitä tietoja (muuttujia) analyysitiedostossa on!
  HUOM: Norjan tietoarkiston NSD ja Suomen tietoarkiston FSD – Suomi sisältöerot (muuttujanimet, painomuuttujat)

- ESS EduNet: Weighting the ESS data

- 2014 versio: Weighting European Social Survey Data
- Aiempi versio: 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

NSD:n kansainväliset tietovarannot

- **Analyysipainot** DWEIGHT ja PSPWGHT
  - Keskiarvo yli aineiston = 1
  - Painomuuttujan avulla kompensoidaan:
    - DWEIGHT: Erisuurien sisältymistodennäköisyyksien vaikutus (mutta ei katokorjausta)
    - PSPWGHT: Erisuurien sisältymistodennäköisyyksien vaikutus ja katokorjaus (jälkiositus, *poststratification)*
  - ESS-suositus:
    - DWEIGHT tai PSPWEIGHT pitää käyttää aina maakohtaisissa tarkasteluissa ja maiden välisissä vertailuissa
    - POIKKEUS: Jos PSPWEIGHT = 1 datan kaikille henkilöille, niin vastaavaa painomuuttujaa ei tarvitse käyttää analyysissa

ESS – Painomuuttujien käyttö – 2

NSD:n kansainväliset tietovarannot

- **Väestöosuuspaino** PWEIGHT
  - Vaihtelee maakohtaisesti
  - Kompensoidaan maiden väestöjen kokoerot
- **Yhdistetty painomuuttuja**

  DWEIGHT*PWEIGHT tai PSPWEIGHT*PWEIGHT
- ESS-suositus: Yhdistettyä painomuuttujaa tulee käyttää usean maan aineistoista yhdistetyn datan analyysissa
  - Esim: EU-maiden yhdistelmäaineiston analyysi (esim. Suomen ja Ruotsin yhdistelmäaineisto)
- Painotusdokumentit (myös kotisivulla):

Painomuuttujien kuvaus NSD:n aineistoissa
Weighting European Social Survey Data (2014 edition)

ESIMERKKI: Analyysipaino Suomen datassa

- Analyysipaino DWEIGHT =1 Norjan tietoarkiston (NSD) Suomen ESS 2010 –aineiston versiossa, jota käytetään tämän osuuden esimerkeissä
- Analyysipaino FSD:n koodikirjan mukaan

[PAINO_2] Analyysipaino

Kysymysteksti

Analyysipaino

Kuvailevat tunnusluvut

tunnusluku	arvo
kelvollisten havaintojen lkm	1878
minimi	0.70
maksimi	1.50
keskiarvo	1.00
keskihajonta	0.19

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
(Norjan tietovaraston aineisto)

- Suomen ESS-aineisto : Yksinkertainen tilanne
  - Systemaattinen otanta
  - Sisältymistodennäköisyydet samoja kaikille
  - Implisiittinen ositus (pj:n lajittelu ennen otantaa)
  - **Maakohtainen analyysi:**
    Analyysipaino DWEIGHT = 1 kaikille
    Painomuuttujaa DWEIGHT ei tarvita Suomen aineiston maakohtaisessa analyysissa
  - **Yhdistelmäaineistojen analyysi**:
    Suomen väestöosuuspaino (vakio kaikille)
    Population size weight PWEIGHT = 0.24
    Painomuuttuja PWEIGHT tarvitaan yhdistelmäaineistojen analyysissa

ESIM. 2: Ranskan ESS-data 2010
(Norjan tietovaraston aineisto)

- Ranskan ESS-aineisto: Mutkikkaampi tilanne
  - Kolmiasteinen ryväsotanta
    Three stage cluster sampling
  - **Maakohtainen analyysi:** DWEIGHT tarvitaan
    Henkilötason sisältymistodennäköisyydet vaihtelevat
    Analyysipainot vaihtelevat
    DWEIGHT: Mean=1, Min=0.098, Max=4.0
    Painomuuttuja DWEIGHT tarvitaan maakohtaisessa analyysissa
  - **Yhdistelmäaineistojen analyysi:**
    Väestöosuuspaino PWEIGHT = 3.1 (vakio kaikille)
    Yhdistetty paino DWEIGHT*PWEIGHT tarvitaan yhdistelmäaineistojen analyysissa

ESIM. 3: ESS – Yhdistelmäaineisto

(Norjan tietovaraston aineisto)

- **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

ESIM. 4: ESS 2010 - Painotuksen vaikutukset: Suomi ja Ranska (Norjan tietovaraston aineistot)

- **Suomen ESS-aineiston analyysi yksinään**
  - Ei tarvita DWEIGHT eikä PWEIGHT
  - Ei vaikutusta keskiarvoihin eikä keskivirheisiin
  - Miksi? Koska analyysipaino DWEIGHT = 1 kaikille henkilöille ja väestöosuuspainoa PWEIGHT ei tarvita
- **Suomen ja Ranskan ESS-vertailuanalyysi**
  - Tarvitaan DWEIGHT
  - Vaikutus Ranskan estimaatteihin (keskivirhe)
  - Ei vaikutusta Suomen estimaatteihin
- **Suomen ja Ranskan yhdistetyn ESS-aineiston analyysi**
  - Tarvitaan DWEIGHT*PWEIGHT

Maiden vertailu: Suomi ja Ranska

(Norjan tietovaraston aineisto)

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

(Norjan tietovaraston aineisto)

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

Laskentaesimerkki: 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

Laskentaesimerkki: 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

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