

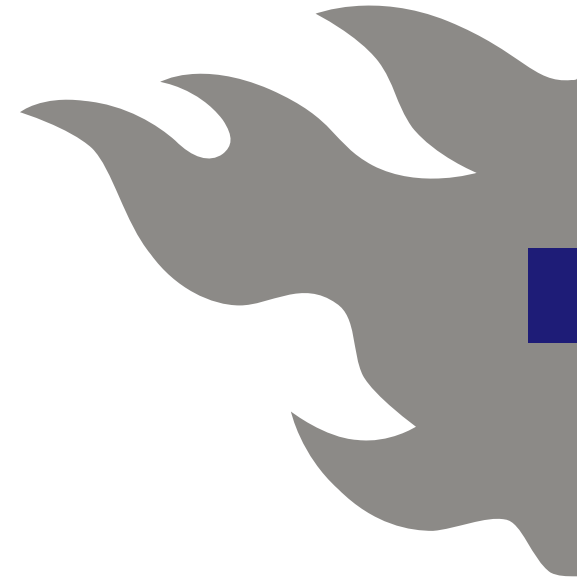
HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

# Sosiaalitutkimuksen tilastolliset menetelmät

## Osa 1 - Diat 3

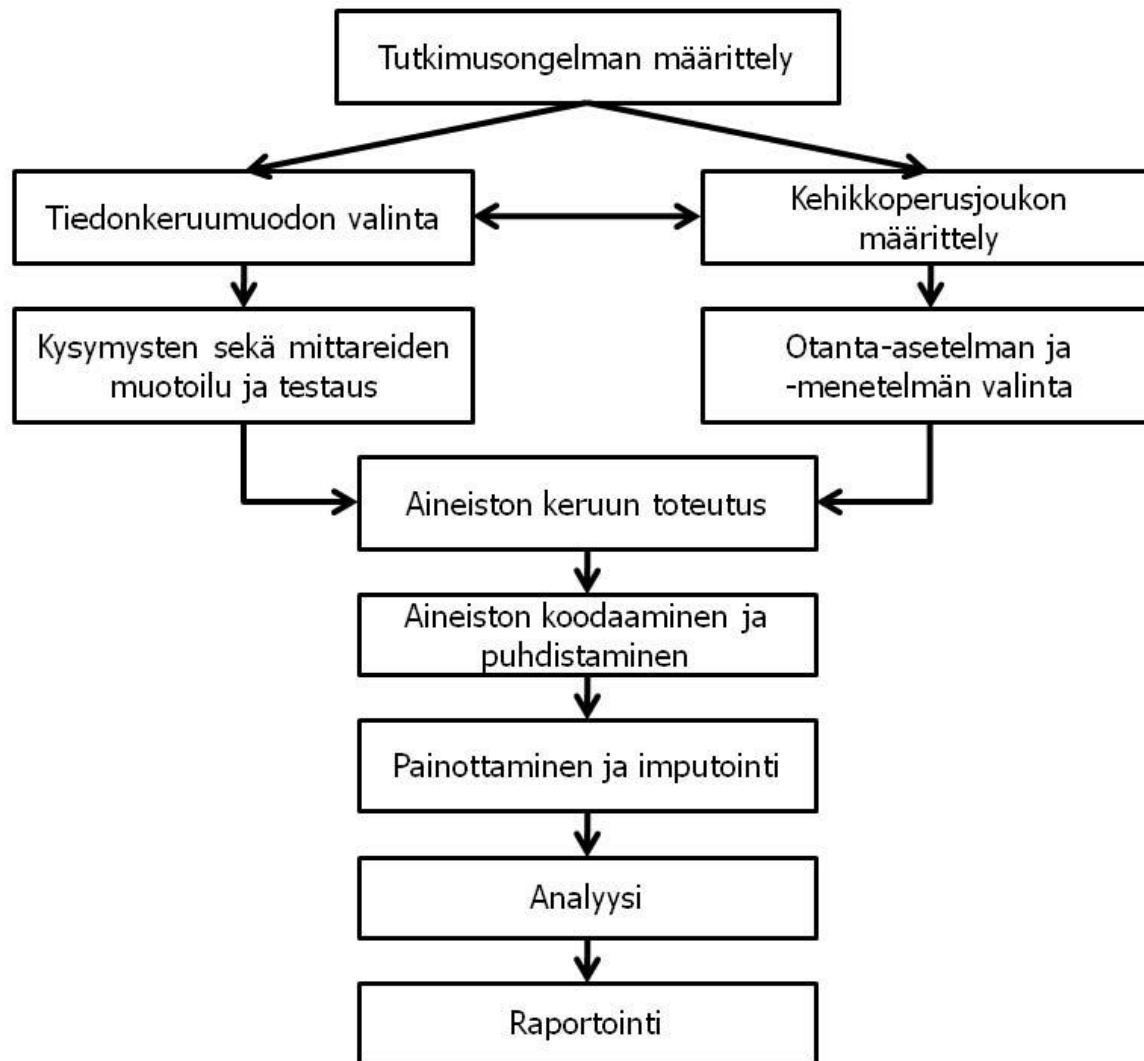
### Otanta-asetelmat ja survey-aineiston käsittely

Risto Lehtonen, Helsingin yliopisto  
[risto.lehtonen@helsinki.fi](mailto:risto.lehtonen@helsinki.fi)

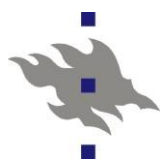




# Survey-tutkimusprosessi - Revisited



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



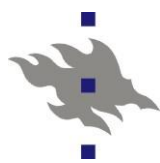
# Vastauskato (puuttuneisuus) - 1

## ■ Yksikkökato (*Unit nonresponse*)

- Otoshenkilöltä ei ole saatu mitään haastattelutietoja
  - Otoshenkilöä ei ole tavoitettu
  - Otoshenkilö on tavoitettu mutta kieltäytynyt osallistumasta haastatteluun

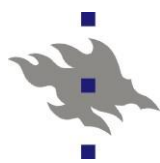
## ■ Eräkato (*Item nonresponse*)

- Osa henkilön haastattelutiedoista saatu mutta osa muuttuja-arvoista puuttuu
- Molempia puuttuneisuuden lajeja esiintyy yleisesti käytännössä!



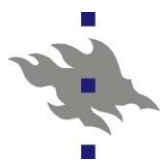
## Vastauskato (puuttuneisuus) - 2

- **Vastauskato on ongelma analyysin kannalta**
- Osallistuminen ei välttämättä ole tutkittavien ilmiöiden kannalta satunnaista
- Vastaajien valikoituminen: Vastausalttius (tuntematon parametri) yleensä vaihtelee henkilöiden ja henkilöryhmien välillä
- Ongelmia tulee erityisesti, jos vastausalttius ja tutkittavat ilmiöt ovat korreloituneita
- Tämä (voi) aiheuttaa analyysituloksiin **harhaa** (*Bias*)!



# Puuttuneisuuden tyypittelyä - 1

- MCAR – *Missing Completely at Random*
  - Osohenkilön **vastausalttius** ja **tulosmuuttujat** ovat **toisistaan riippumattomia**
  - Harvoin voimassa - mutta usein oletetaan olevan!
  - Esim: Kulutustutkimuksessa suurituloisten vastausalttius (likimain) sama kuin pienituloisten
    - Ei välttämättä ole voimassa!
- Terminologiaa:  
Seppo Laaksonen (2013) [Surveyymetodiikka](#) (Luku 9.3)



## Puuttuneisuuden tyypittelyä - 2

- MAR – *Missing at Random*
  - Vastausalttius ja tulosmuuttujat ovat **ehdollisesti riippumattomia** ehdolla taustamuuttujat
  - **Vastauskadon oikaisumenetelmien oletus**
  - Esim: Suurituloisilla (likimain) sama vastausalttius ja pienituloisilla samoin, mutta vastausalttiudet voivat poiketa näiden ryhmien välillä
- NMAR – *Not Missing at Random*
  - **Puuttuneisuus ei ole satunnaista**
  - Tilanne hankalasti hallittavissa menetelmällisesti!



## Vastauskadon vaikutusten oikaisu - 1

- **Yksikkökadon** vaikutusten oikaisu tilastollisilla menetelmillä (*Adjustment for unit nonresponse*)
  - Uudelleenpainotusmenetelmät (*Reweighting*)
    - RHG-menetelmä (*Response Homogeneity Groups*)
    - Muodostetaan vastausalttiuden suhteen sisäisesti homogeenisia osajoukkoja
  - Vastauskadon tilastollinen mallinnus
    - Esim: Logistinen regressiomalli
    - Mallinnetaan vastausalttiutta (binäärinen vaste)
  - Menetelmillä muokataan painokertoimia
- **Jotta menetelmiä voidaan käyttää, tarvitaan tietoja sekä vastanneista että ei-vastanneista!**



## ■ Esim: PISA - Weighting procedure (design weight)

■ Weight  $w_{ik}$  for student  $k$  in school  $i$ :

$$w_{ik} = w_{1i} \times w_{2ik} \times f_i, \quad i = 1, \dots, m \text{ and } k = 1, \dots, n_i,$$

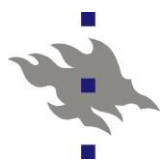
where

$w_{1i} = 1/(\pi_i \hat{\theta}_i)$  is the reciprocal of the product of the inclusion probability  $\pi_i$  and the estimated participation probability  $\hat{\theta}_i$  of school  $i$ ;

$w_{2ik} = 1/(\pi_{k|i} \hat{\theta}_{k|i})$  is the reciprocal of the product of the conditional inclusion probability  $\pi_{k|i}$  and estimated conditional response probability  $\hat{\theta}_{k|i}$  of student  $k$  from within the selected school  $i$ ;

$f_i$  is an adjustment factor for school  $i$  to compensate any country-specific refinements in the survey design, and  $m$  is the number of sample schools in a given country and  $n_i$  is the number of sample students in school  $i$ .





## Vastauskadon vaikutusten oikaisu - 2

- **Eräkadon** vaikutusten oikaisu tilastollisilla imputointimenetelmillä (*Imputation*)
- **”Yksinkertainen” imputointi** (*Single imputation*)
  - Puuttuva tieto korvataan **yhdellä uudella arvolla**
  - ”Hot deck” – puuttuva muuttuja-arvo ”lainataan” samankaltaiselta toiselta vastaajalta
  - Lähimmän naapurin menetelmä – arvo ”lainataan” ”naapurilta” – saadaan esim. lajittelemalla data
  - Regressiopohjaiset menetelmät, ym.
- **HUOM: Puuttuvan muuttuja-arvon korvaaminen saatujen havaintojen keskiarvolla on HUONO imputointimenetelmä, ei voida suositella!**



## Vastauskadon vaikutusten oikaisu - 3

### ■ **Moni-imputointi** (*Multiple imputation*)

- Puuttuva tieto korvataan **usealla uudella arvolla**
- Työvaiheet
  - Määrittele aineistolle imputointimalli
  - Generoi mallin avulla lukuisia puuttuvan tiedon korvaavia arvoja, esim. 10 arvoa
  - Saadaan 10 uutta aineistoa
  - Analysoi yhdistetty aineisto
  - SAS-proseduurit MI ja MIANALYZE
  - SPSS – Multiple Imputation ja SPSS-analyysit
- Imputointi vaatii harkintaa ja menetelmäosaamista!

# ESS 2010 – Suomi: TRUST-muuttujat

## Eräkato – Item nonresponse

- **SPSS MVA – Missing Value Analysis**
- Aineistossa havaintoja kaikkiaan n=1878
- TRUST-muuttujat: Puuttuvia arvoja 98 henkilöllä kaikkiaan 201 (ks. seuraavan sivun kuviot)

### Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes <sup>a</sup>	
				Count	Percent	Low	High
trstp1	1866	5,38	2,254	12	,6	0	0
trst1g1	1862	6,91	2,001	16	,9	127	0
trstp1c	1869	8,03	1,663	9	,5	79	0
trstp1t	1863	4,43	2,181	15	,8	0	0
trstp1r	1855	4,54	2,158	23	1,2	0	0
trstep	1806	5,09	2,181	72	3,8	0	0
trstun	1824	6,55	1,890	54	2,9	21	0

a. Number of cases outside the range (Q1 - 1.5\*IQR, Q3 + 1.5\*IQR).

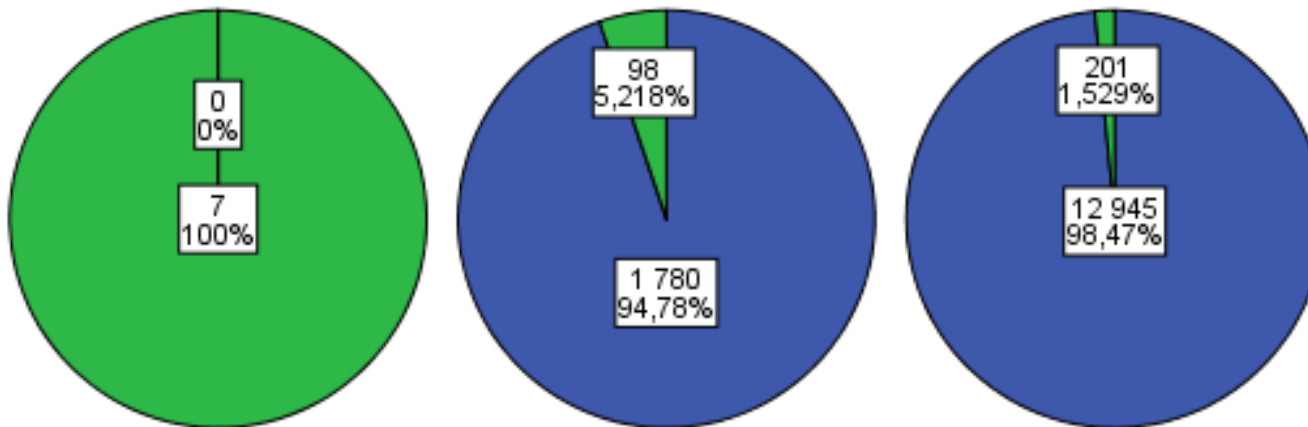


# ESS 2010 – Suomi: TRUST-muuttujat

## SPSS Multiple Imputation – Analyze Patterns

### Overall Summary of Missing Values

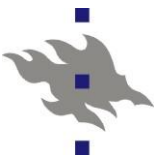
■ Complete Data  
■ Incomplete Data



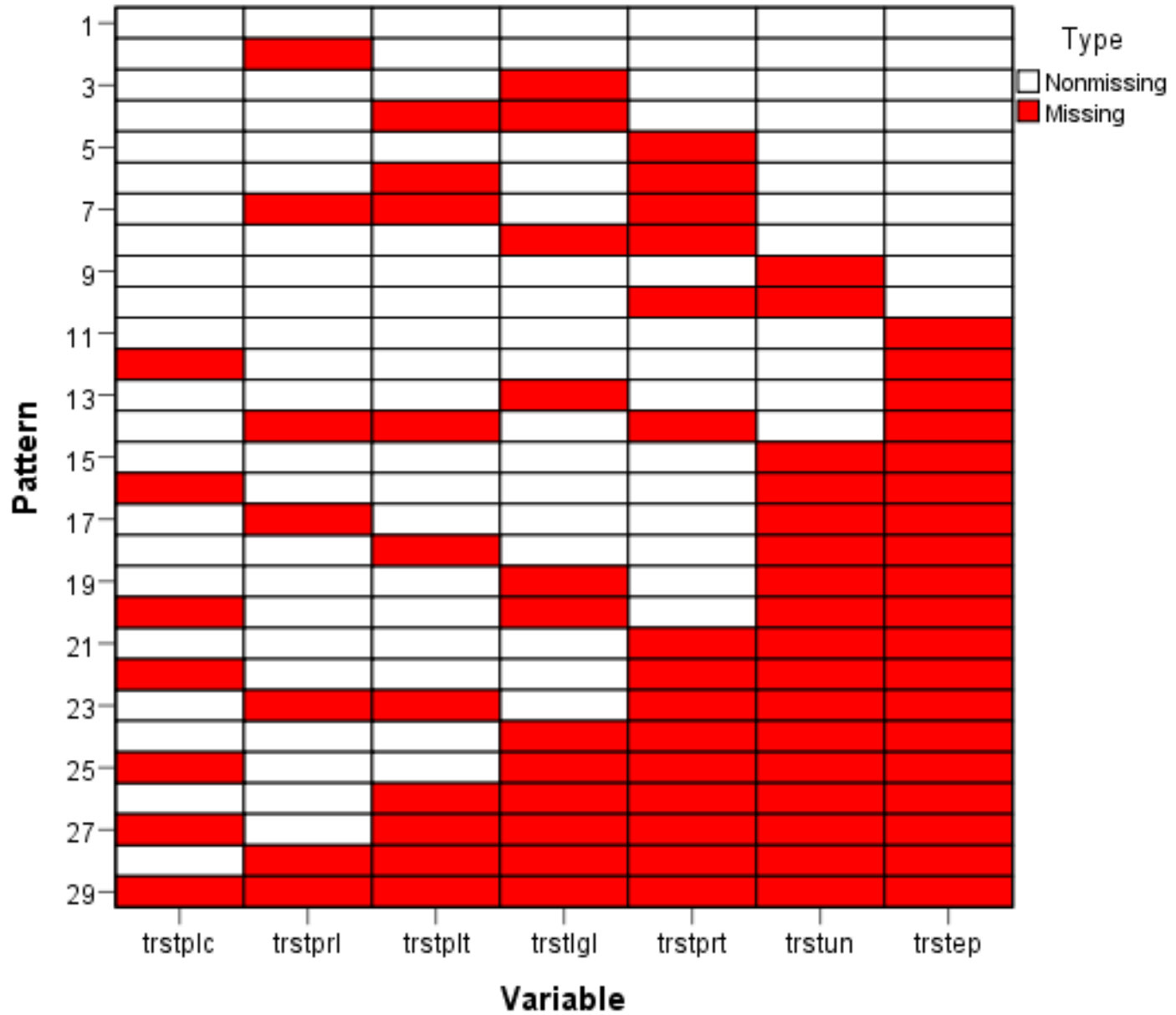
**Variables**

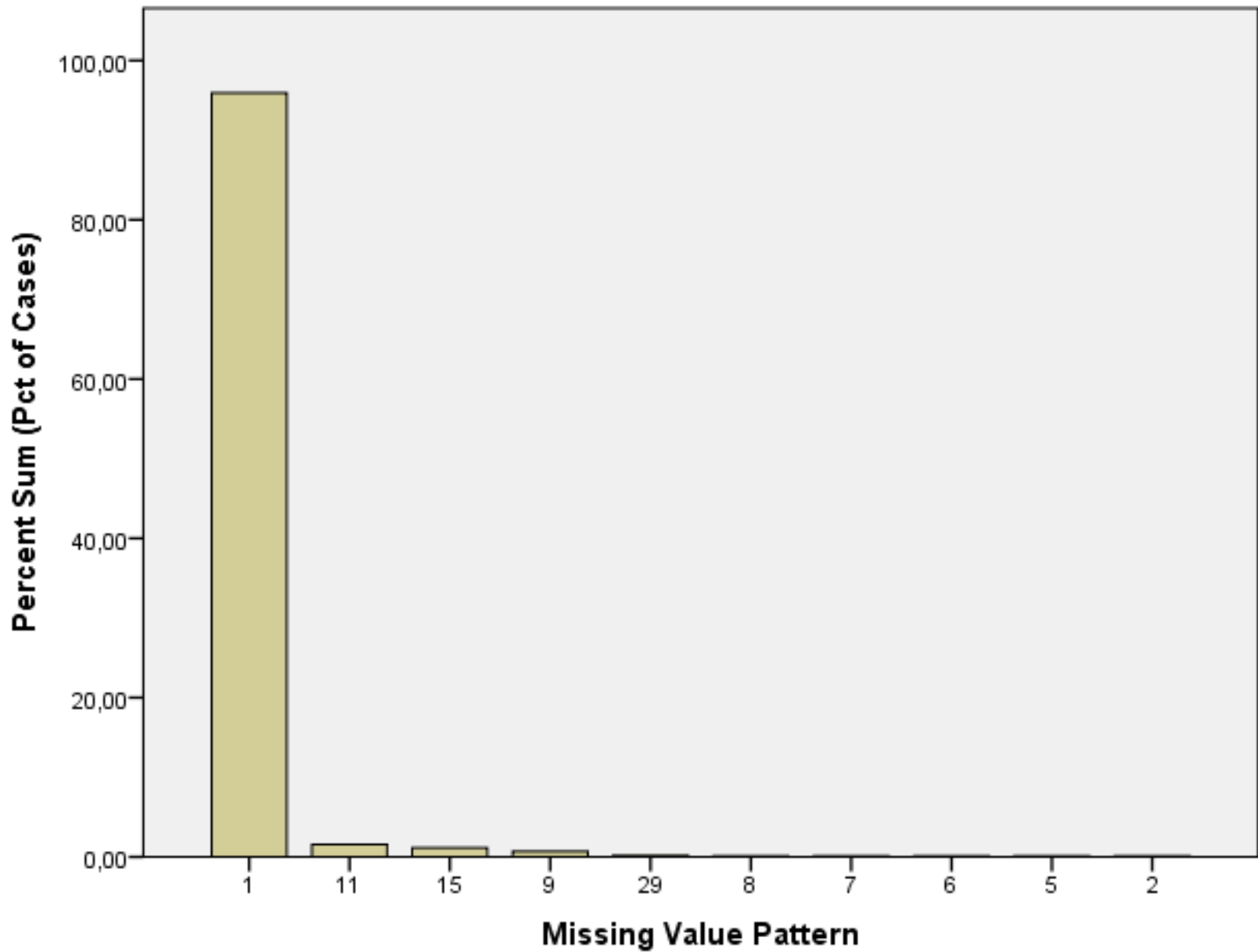
**Cases**

**Values**



## Missing Value Patterns





The 10 most frequently occurring patterns are shown in the chart.



# Surveyn aineistonmuodostus – Revisited!

## ■ Tutkimusaineiston luonnin työvaiheita

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, ryväsendikaattorit
7. Lisätietojen liittäminen aineistoon
  - Rekisteritietoja ja muita saatavilla olevia tietoja esim. virallisista tilastolähteistä



## ESS 2010: Avoimet kv. aineistot (NSD)

- **Kaikkia vaiheita 1-7 ei aina ole toteutettu**
  - Riippuu osin maakohthaisten otanta-asetelmien eroista ja mahdollisesti tietosuoja-asioista
- ESS 2010, Norjan tietovarasto NSD:
- Analyysipainot DWEIGHT ja PWEIGHT ovat käytettävissä
  - Aineistojen muita ominaisuuksia
    - Yksikkökatoa ei ole oikaistu (adjustoitu)
      - Eräkatoa ei ole imputoitu
      - Asetelmaindikaattoreita on vaikea saada käyttöön – tilanne kuitenkin parantunut
      - Rekisteriperusteisia lisätietoja ei ole mukana





# ESS 2010: FSD:n **Suomen** aineisto

## ■ **Yksikkökadon oikaisu** (adjustointi)

- Aineistossa on kaksi painomuuttujaa, jotka korjaavat saadun aineiston jakaumia populaation väestösuhteita vastaaviksi
- **DWEIGHT**: Pohjana väestön ikä-sukupuoli-jakauma, äidinkieli, muunnettu suuraluejako ja kuntaryhmitys
  - Kalibrointimenetelmä (Deville and Särndal 1992)
  - Analyysipainojen skaalaus: keskiarvo = 1  
summa = aineiston henkilöiden lkm
- **PWEIGHT** (korotuspaino): Summa = 15-vuotta täyttäneiden suomalaisten lukumäärä (4,4 milj.)



# Kirjallisuutta

- Laaksonen, Seppo (2013) [Surveyymetodiikka](#).  
**Hyvä suomenkielinen perusteos, kattaa myös ESS-tutkimussarjan arviointia ja analyysia!**
- Lehtonen, Risto & Pahkinen, Erkki (2004) [Practical Methods for Design and Analysis of Complex Surveys](#) John Wiley & Sons.  
Ladattavissa [dawsoneran](#) kautta  
Web-laajennus: [VLISS](#)-Virtual laboratory in survey sampling
- De Vaus D. (2013) [Surveys in Social Research](#) □6th Ed. Routledge.
- Tilastokeskus (2007). [Laatua tilastoissa](#).  
2. uudistettu painos, Tilastokeskus, Käsikirjoja 43.  
Ladattavissa myös kurssin kotisivulta.