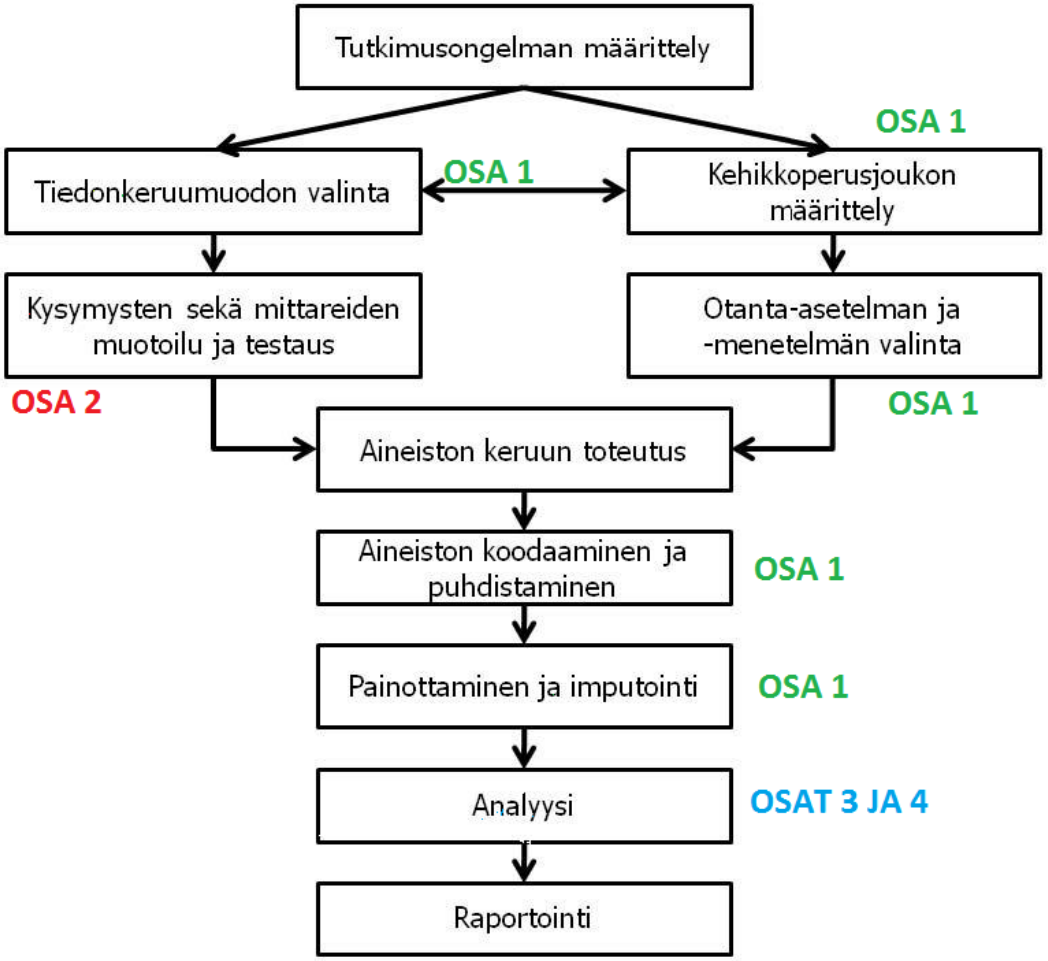
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Sosiaalitutkimuksen tilastolliset menetelmät Osa 1 - Diat 3 Otanta-asetelmat ja survey-aineiston käsittely

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

Versio 21.1.2015

Otosperusteinen survey: Otanta-asetelma tutkimusasetelman osana: Vastauskato



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
    - Muita syitä: Ks. ESS2010-report
      ESS 2010 Response rates (Sweden, France)
- **Eräkato** (*Item nonresponse*)
  - Osa henkilön haastattelutiedoista on saatu mutta osa muuttuja-arvoista puuttuu
- Molempia puuttuneisuuden lajeja esiintyy yleisesti käytännössä!

Vastauskato (puuttuneisuus) - 2

- **Vastauskato on ongelma analyysin kannalta**
- Osallistuminen (vastausalttius) ei välttämättä ole tutkittavien ilmiöiden kannalta satunnaista, vaan voi riippua erilaisista taustatekijöistä
- **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**
  - *Selection bias* = valikoitumisharha

Puuttuneisuuden tyypittelyä - 1

- MCAR – *Missing Completely at Random*
  - **Puuttuneisuus on täysin satunnaista**
  - **Vastausalttius** ja **tulosmuuttujat** ovat **toisistaan riippumattomia**
  - Harvoin voimassa - mutta usein oletetaan olevan!
    - Esim: Kulutustutkimuksessa suurituloisten vastausalttius on (likimain) sama kuin pienituloisten
    - MCAR-oletus ei välttämättä ole voimassa!
- Terminologiaa:

  Seppo Laaksonen (2013) Surveymetodiikka (Luku 9.3)
- Klassikko:

  Rubin D. (2004) Multiple Imputation for Nonresponse in Surveys. Wiley.

Puuttuneisuuden tyypittelyä - 2

- MAR – *Missing at Random*
  - **Puuttuneisuus on ehdollisesti satunnaista**
  - Vastausalttius ja tulosmuuttujat ovat **ehdollisesti riippumattomia** ehdolla taustamuuttujat
  - **Vastauskadon oikaisumenetelmien oletus**
    - Uudelleenpainotusmenetelmät (jälkiositus)
    - Mallinnusmenetelmät (logistinen malli)
    - Esim: Suurituloisilla (likimain) sama vastausalttius ja pienituloisilla samoin, mutta vastausalttiudet voivat poiketa näiden ryhmien välillä
    - Katokorjaus edellyttää tietoja tuloista **kaikilta otoshenkilöiltä** (sekä vastanneet että ei-vastanneet eli katohenkilöt)

ESIMERKKI: Katokorjaus MAR-oletuksen varassa (hypoteettinen kulutustutkimus)

- Otoskoko 500 henkilöä, vastanneita 260 henkilöä
  - Vastausosuus p = 260/500 = 0.52 eli 52 %
- Katokorjaus MAR-oletuksen varassa
  - Oletetaan että vastausalttius ja tulosmuuttujat (esim. kulutusmenot kulutusyksikköä kohti) ovat toisistaan riippumattomia **ehdolla tulot**
- Oletetaan, että tutkimuksessa tulotiedot ovat käytettävissä **kaikilta otoshenkilöltä**
- Muodostetaan tuloviidennekset
- Tuloviidenneksittäinen jakauma
- Laadittu painomuuttuja yhdistetään analyysitiedostoon
- Laaditaan uusi analyysipaino
  - Alkuperäisen (korjaamattoman) analyysipainon ja uuden tuloryhmä-kohtaisen painon tulo
  - Painojen uudelleenskaalaus niin, että uuden analyysipainomuuttujan keskiarvo yli analyysiaineiston = 1

Tuloviidennes	Otos	Vastanneet	Vastaus-osuus p	Paino- = 1/p
1 (pienituloisin)	100	50	0.50	2.00
2	100	60	0.60	1.67
3	100	70	0.70	1.43
4	100	50	0.50	2.00
5 (suurituloisin)	100	30	0.30	3.33
Kaikki	500	260	0.52	1.92

Puuttuneisuuden tyypittelyä - 3

- NMAR – *Not Missing at Random*
  - **Puuttuneisuus ei ole satunnaista**
  - **Pätee jos MCAR ja MAR eivät ole voimassa**
  - Tilanne hankalasti (jos lainkaan) hallittavissa menetelmällisesti!
    - NMAR-tilanteessa vastausalttius ja kiinnostuksen kohteena oleva tulosmuuttuja korreloivat keskenään ja vastausalttius riippuu tekijöistä, joita ei voida kontrolloida datassa olevien muuttujien avulla
  - Esim: Vastausalttiudet poikkeavat suurituloisten ja pienituloisten ryhmien välillä sekä näiden ryhmien sisällä
  - Aineistossa ei ole muuttujia, joiden avulla voidaan muodostaa vastausalttiuden suhteen homogeenisia osajoukkoja tuloryhmien sisällä

Vastauskadon vaikutusten oikaisu - 1

- **Yksikkökadon** vaikutusten oikaisu tilastollisilla menetelmillä (*Adjustment for unit nonresponse*)
  - Uudelleenpainotusmenetelmät (*Reweighting*)
    - **Jälkiositus** (*Post-stratification*) – **ESS-paino PSPWGHT**
    - RHG-menetelmä (*Response Homogeneity Groups)*
    - Muodostetaan vastausalttiuden suhteen sisäisesti homogeenisia osajoukkoja (vrt. edellinen esimerkki)
  - Vastauskadon tilastollinen mallinnus
    - Esim: Logistinen regressiomalli
    - Mallinnetaan vastausalttiutta (binäärinen vaste)
  - Menetelmillä muokataan analyysipainoja
  - **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:

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

where

$w_{1i} = 1/(\pi_i \hat{\theta}_i)$ is the reciprocal of the product of the inclusion probability π_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 sopivasti
  - 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!
  - SPSS Learning Module: Missing data

ESS 2010 – Suomi: TRUST-muuttujat Eräkato – Item nonresponse

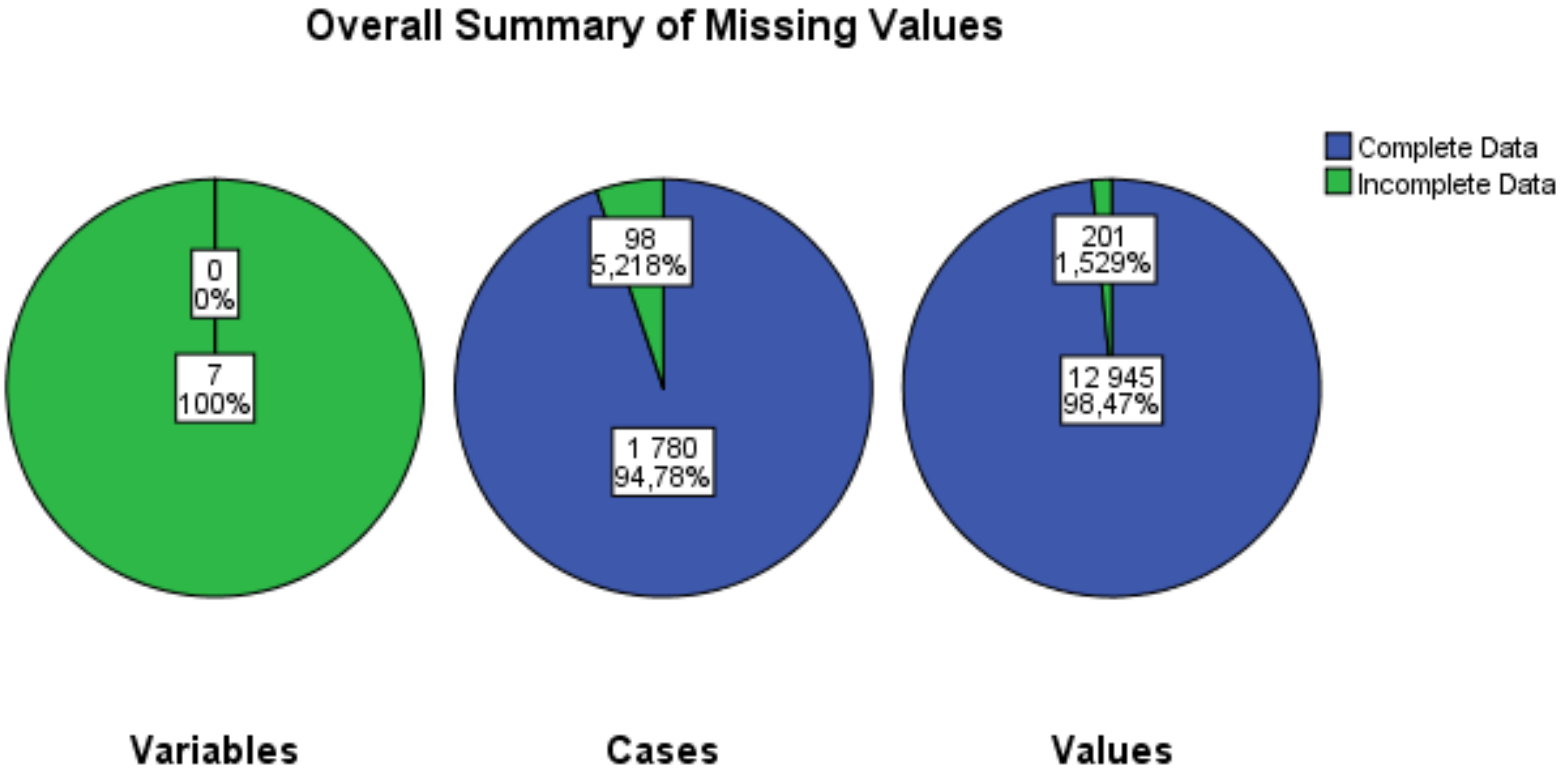
- **SPSS MVA – Missing Value Analysis**
- Aineistossa havaintoja kaikkiaan n = 1878
- TRUST-muuttujat: Puuttuvia arvoja 98 henkilöllä kaikkiaan 201 (ks. seuraavan sivun kuviot)

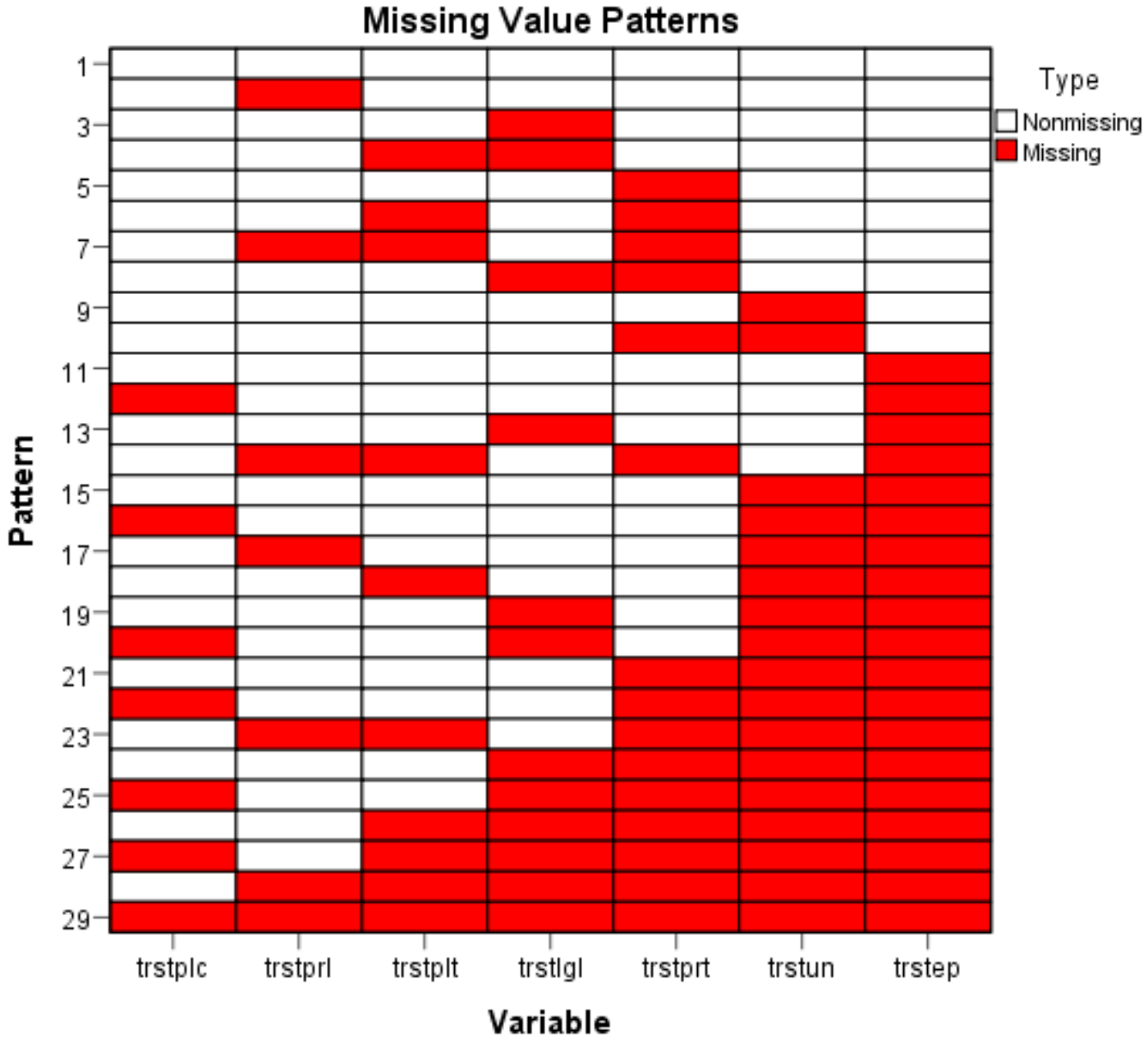
Univariate Statistics

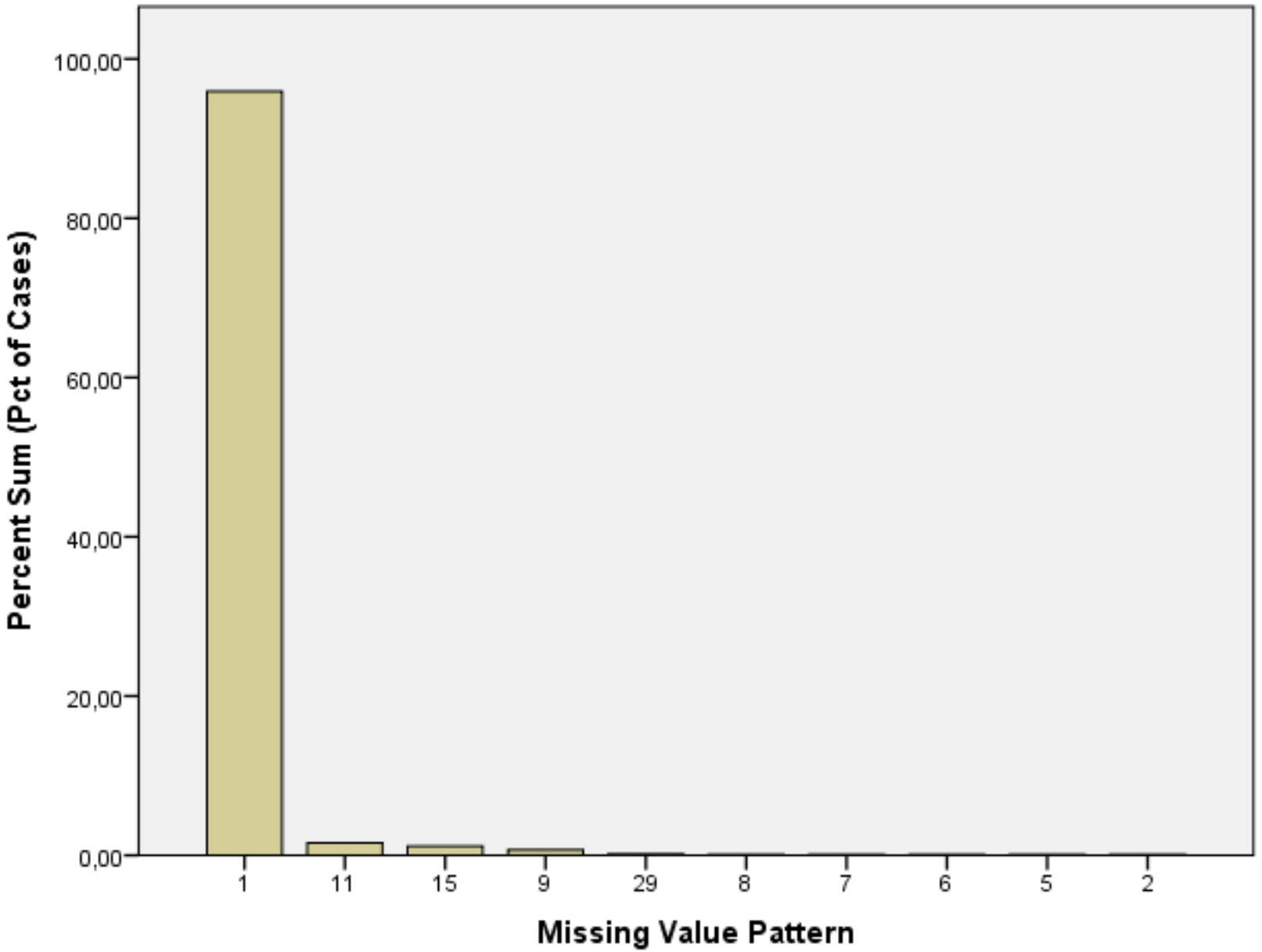
	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
trstprl	1866	5,38	2,254	12	,6	0	0
trstlgl	1862	6,91	2,001	16	,9	127	0
trstplc	1869	8,03	1,663	9	,5	79	0
trstplt	1863	4,43	2,181	15	,8	0	0
trstprt	1855	4,54	2,158	23	1,2	0	0
trstep	1806	5,09	2,181	72	3,8	0	0
trstun	1824	6,55	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




Missing Value Patterns
Type
Nonmissing
Missing
Pattern
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
trstplc
trstprl
trstplt
trstlgl
trstprt
trstun
trstep
Variable



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

Surveyn aineistonmuodostus analyysia varten (revisited)

- **Tutkimusaineiston luonnin työvaiheita**

1. Asetelmapainojen muodostus
2. Vastauskadon analyysi
3. Aineiston tarkistaminen ja editointi
4. Puuttuvien tietojen imputointi
   - Eräkadon oikaisu
5. Uudelleenpainotus ja analyysipainojen muodostus
   - Yksikkökadon oikaisu
6. Asetelmaindikaattoreiden liittäminen aineistoon
   - Ositeindikaattorit, ryväsindikaattorit (tarvittaessa)
7. Lisätietojen liittäminen aineistoon
   - Rekisteritietoja ja muita saatavilla olevia tietoja esim. virallisista tilastolähteistä

ESS 2010: Avoimet kv. aineistot (NSD)

- **Kaikkia vaiheita 1-7 ei välttämättä ole aina toteutettu**
  - Riippuu osin maakohtaisista ratkaisuista ja tiedostoversiosta
- **ESS 2010, Norjan tietovarasto NSD:**
  - Alkuperäiset painot DWEIGHT ja PWEIGHT
  - Uusi painomuuttuja PSPWGHT (jälkiositus, *post-stratification*) ESS5 – 2010 -raportti vers. 3.2:

    ESS5 - 2010 Documentation Report Edition 3.2
    ESS5 edition 3.2 (26.11.14) - Applies to datafile ESS5 edition 3.2
    Changes from edition 3.1: Weighting. Information regarding post-stratification weights updated
    Documentation of ESS: Post-Stratification Weights
    (25th April 2014)

ESS 2010: FSD:n Suomen aineisto

- **Yksikkökadon oikaisu** (adjustointi)
  - Datassa kaksi painomuuttujaa
    - korjaavat saadun aineiston jakaumia populaation väestöjakaumia vastaaviksi
  - Paino_1 (**Korotuspaino**): Summa = 15-vuotta täyttäneiden suomalaisten lukumäärä (4,4 milj.)
    HUOM: Vastaavaa painoa EI ole NSD:n Suomi-datassa!
  - Paino_2 (**Analyysipaino**):
    - Pohjana väestön ikä-sukupuoli-jakauma, äidinkieli, muunnettu suuraluejako ja kuntaryhmitys
    - Vastauskadon (*unit nonresponse*) aiheuttaman harhan oikaisu
    - Kalibrointimenetelmä (Deville and Särndal 1992)
    - Analyysipainojen skaalaus: keskiarvo = 1 summa = aineiston henkilöiden lkm (vastaa NSD:n PSPWGHT-painoa)
  - Ks: FSD2682 European Social Survey 2010: Suomen aineisto, Koodikirja

[PAINO_1] Korotuspaino

Kysymysteksti

Korotuspaino

Kuvailevat tunnusluvut

tunnusluku	arvo
kelvollisten havaintojen lkm	1878
minimi	1648.62
maksimi	3516.09
keskiarvo	2344.12
keskihajonta	453.83

[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 – Kaksi esimerkkiä Revisited

- Esimerkki1

http://www.ess.fi/uutiset/kotimaa/2013/11/05/tutkimus-suomi-on-euroopan-neljanneksi-onnellisin-maa

- Esimerkki2

- Millaisia menetelmällisiä / sisällöllisiä kysymyksiä heräsi näistä ESS-esimerkeistä?
  - Tutkimusasetelmat?
  - Tiedonkeruuasetelmat?
  - Aineistot?
  - Tutkimuskysymykset?
  - Analyysit?
  - Päätelmät / Yleistykset?
  - Tulosten vertailtavuus?
    - **ESS-tulosten vertailtavuus yli maiden?**

ETELÄ-SUOMEN SANOMAT

Tutkimus: Suomi on Euroopan neljänneksi onnellisin maa

KOTIMAA 5.11.2013, klo 17:21.

Suomi on eurooppalaisen yhteiskuntatutkimuksen mukaan Euroopan neljänneksi onnellisin maa. Suomalaisten onnellisuus on hieman noussut kahdessa vuodessa.

Onnellisuuden keskiarvolla mitattuna Tanska on paras, Islanti toinen ja Norja kolmas. Suomen kanssa samalle tasolle yltää Sveitsi. Sen sijaan ruotsalaisten onnellisuuskeskiarvo on laskenut ja jää nyt kuudenneksi.

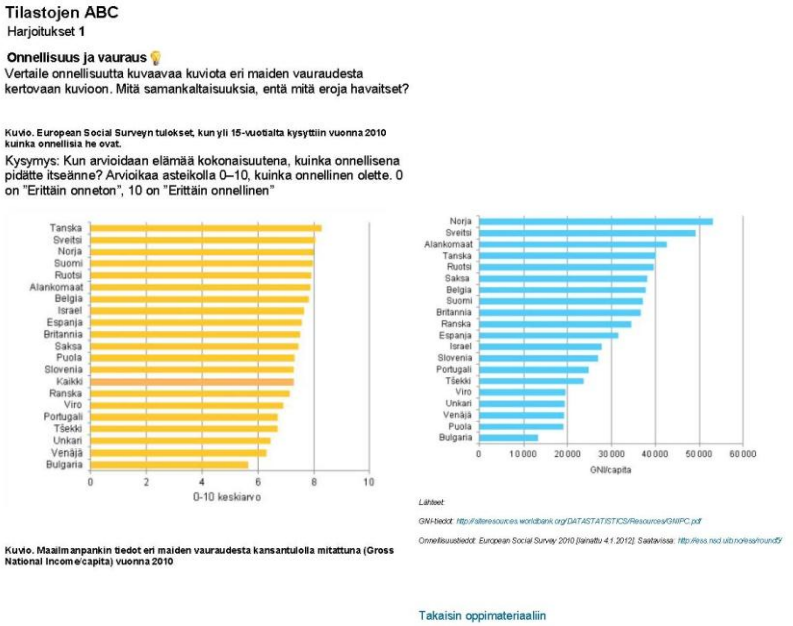
European Social Survey (ESS) -tutkimusta varten tiedusteltiin yli 15-vuotiaiden onnellisuudesta 29 maassa viime ja tänä vuonna.

Tutkimuksen otantaa valvovaan ryhmään kuului tilastotieteen professori Seppo Laaksonen Helsingin yliopistosta. Hän kertoo, että vastaajilta kysyttiin onnellisuudesta asteikolla 1–10, jossa 10 on hyvin onnellinen.

– Tässä vaiheessa uusimmat tulokset ovat käytettävissä suuresta osasta ESS-tutkimusmaita, mutta esimerkiksi Liettua ja Italia eivät vielä ole mukana. Loput tulokset saadaan ensi vuonna, Laaksonen kertoo.

Onnellisuustilaston peränpitäjäksi jää Bulgaria, jossa onnellisuus on vähentynyt entisestään. Onnellisuuden keskiarvo on noussut selvimmin Saksassa.

STT



Tilastojen ABC

Harjoitukset 1

Onnellisuus ja vauraus

Vertaile onnellisuutta kuvaavaa kuviota eri maiden vauraudesta kertovaan kuvioon. Mitä samankaltaisuuksia, entä mitä eroja havaitset?

Kuvio. European Social Surveyn tulokset, kun yli 15-vuotialta kysyttiin vuonna 2010 kuinka onnellisia he ovat.

Kysymys: Kun arvioidaan elämää kokonaisuutena, kuinka onnellisena pidätte itseänne? Arvioikaa asteikolla 0–10, kuinka onnellinen olette. 0 on "Erittäin onneton", 10 on "Erittäin onnellinen"

Lähteet:

GNI-tiedot: http://siteresources.worldbank.org/DATASTATISTICS/Resources/GNIPC.pdf

Onnellisuustiedot: European Social Survey 2010 [lainattu 4.1.2012]. Saatavissa: http://ess.nsd.uib.no/ess/round5/

Kuvio. Maailmanpankin tiedot eri maiden vauraudesta kansantulolla mitattuna (Gross National Income/capita) vuonna 2010

Takaisin oppimateriaaliin

Kirjallisuutta

- Laaksonen, Seppo (2013) Surveymetodiikka.
  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
- Rubin D. (2004) Multiple Imputation for Nonresponse in Surveys. Wiley.
- 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