HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Sosiaalitutkimuksen tilastolliset menetelmät Osa 3 – Diat 1 Otanta-asetelmat ja tilastollinen analyysi

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

Osan 3 kuvaus

- Jaksolla annetaan kuva tilastollisen otannan roolista yhteiskuntatieteiden sovellusalojen tutkimusprosessissa, erilaisista otantamenetelmistä ja niiden käytöstä sekä otanta-asetelmalla kerätyn aineiston tilastollisen analyysin perusteista.
- Menetelmiä valaistaan esimerkeillä, joissa käytetään aitoja tutkimusaineistoja.
  - European Social Survey -tutkimussarja
  - PISA-tutkimussarja
  - Kelan työterveyshuoltotutkimus
- Alan tilastollisia ohjelmistoja (SPSS, SAS, R) esitellään

Pääotsikot

Empiirinen kvantitatiivinen tutkimusprosessi

Tutkimusaineistojen lähteitä

Otoksen poiminta

Aineiston analysointi

Kirjallisuutta

- Lehtonen, Risto & Pahkinen, Erkki (2004). Practical Methods for Design and Analysis of Complex Surveys John Wiley & Sons.
  - Ladattavissa dawsoneran kautta
- Pahkinen, Erkki & Lehtonen, Risto (1989). Otanta-asetelmat ja tilastollinen analyysi, Gaudeamus.
- Laaksonen, Seppo (2011). Surveymetodiikka bookboon.com
- Tilastokeskus (2007). Laatua tilastoissa, 2. uudistettu painos, Tilastokeskus, Käsikirjoja 43.
- Eurostat (2008). Survey Sampling Reference Guidelines

Empiirinen kvantitatiivinen tutkimusprosessi - Otosperusteinen

Survey = Empiiris-kvantitatiivinen (yhteiskunta)tutkimus

- **Survey-projektin vaiheet:**

I Suunnittelu ja testaus

1. Tutkimusongelman muotoilu
2. Tutkimusasetelman laadinta
3. Otanta-asetelman laadinta
4. Tiedonkeruuvälineiden valmistus
5. Testaus laboratorio-oloissa ja pilotointi kentällä

II Tiedonkeruuoperaatiot

6. Otoksen poiminta
7. Tiedonkeruu
8. Tiedostonmuodostus
   - editointi, imputointi
   - katoanalyysi
   - painokertoimien muodostus

III Tilastollinen analyysi

9. Eksplorointi ja kuvailu
   - tunnuslukujen laskenta,
   - taulukointi
   - graafiset kuvailut
   - piste-estimointi
   - väliestimointi
10. Analyysi ja tulkinta
    - tilastollinen mallinnus

IV Raportointi ja jälkihoito

11. Julkaisut ja artikkelit
12. Opinnäytetyöt
13. Esitelmät
14. Sähköiset tuotteet
15. Dokumentointi ja arkistointi

Tärkeä huomio

- Osavaiheet:

3. Otanta-asetelman laadinta
6. Otoksen poiminta
8. Tiedostonmuodostus
9. Eksplorointi ja kuvailu
10. Analyysi ja tulkinta
15. Dokumentointi ja arkistointi

- Tutkimusprosessin aikaisemmat vaiheet 3, 6 ja 8 pitää osata ottaa huomioon aineiston analyysivaiheissa 9 ja 10

- Otanta-asetelmaan ja tiedostonmuodostukseen liittyvät asiat pitää osata dokumentoida aineiston myöhempää käyttöä varten

Tutkimusaineistojen lähteitä: Rekisteriaineistot

- Hallinnollinen rekisteri
  - Hallinnollisen prosessin oheistuote
  - Päivittyy jatkuvasti
  - Kela: Sosiaalivakuutuksen tietokannat
  - Verohallitus: Verotietokanta
  - Väestörekisterikeskus: Väestön keskusrekisteri
- Tilastorekisteri (Tilastokeskus)
  - Usean hallinnollisen rekisterin yhdistelmä
  - Rekisteriseloste: Tulonjakotilasto
  - StatFin – Tilastotietokannat
- Rekistereitä käytetään otantakehikkoina ja tutkimustiedon lähteinä

Rekisteritutkimuksen tukikeskus ReTki

- Tavoitteena edistää kansallisten rekisterien tutkimuskäyttöä erityisesti terveys- ja sosiaalitieteissä

- **Perustehtävät**
  - tarjota tietoa rekistereistä ja niiden tutkimuskäytöstä
  - järjestää koulutusta rekisteritutkimuksesta
  - neuvoa rekisteriaineistojen tutkimuskäyttöön liittyvissä asioissa
  - ylläpitää rekisteriviranomaisten ja tutkimuslaitosten yhdyshenkilöiden verkostoa

Tietoarkistot: Tärkeitä valmisaineistojen varastoja

- Pääasiassa **otosperusteisia** aineistoja
  - Perustuvat suoraan tiedonkeruuseen
  - Kyselyaineistot, haastatteluaineistot
- Yhteiskuntatieteellinen tietoarkisto FSD
  - Tampereen yliopiston yhteydessä
- Council of European Social Sciende Data Archives CESSDA
- Esimerkki
  - European Social Survey ESS (2002-2010)
  - Riippumattomia poikkileikkausaineistoja/

ESS – European Social Survey

- ESS vaihe 5 (2010)
- Tiedonkeruu
  - Tyypillisesti käyntihaastattelu (CAPI)
  - Tiedonkeruulomake
- Keskimääräinen vastausprosentti 70 %
  - Vastauskato (unit nonresponse): 30 %
  - Vaihtelee maittain
- Tutkimusaineisto
  - noin 39 000 henkilöä
  - Noin 600 muuttujaa
- Datapankki: http://ess.nsd.uib.no/

ESS – Otanta-asetelmat

- Otanta-asetelmat vaihtelevat maittain
  - Maakohtaiset otoskoot likimain yhtäsuuria
  - Erisuuret sisältymistodennäköisyydet maittain
    - Painomuuttujien käyttö analyysissa
  - Yksinkertaiset / Monimutkaiset otanta-asetelmat
  - Suomi: Systemaattinen otanta
  - Belgia: Ositettu kaksiasteinen ryväsotanta
- ESS5 - 2010 DOCUMENTATION REPORT

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

ESS – Tilastollinen analyysi

- Otanta-asetelmaan reagointi on välttämätöntä tilastollisen analyysin yhteydessä
- Miksi?
- Pätevän tilastollisen päättelyn suorittamiseksi
- Tavoitteena on yleistää otosaineistosta saatavat tulokset koskemaan koko perusjoukkoa, josta otos on poimittu
  - Tilastollinen estimointi
  - Tilastollinen testaus
  - Tilastollinen mallinnus

Mitä tietoja aineistossa tulee olla pätevää analyysia varten?

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

Esimerkkl: Suomen ESS-data

- Käytännössä näitä kaikkia tietoja ei aina välttämättä tarvita
  - Riippuu asetelman yksinkertaisuudesta / monimutkaisuudesta

- Suomen ESS-aineisto: Yksinkertainen tilanne
  - Systemaattinen otanta
  - Sisältymistodennäköisyydet samoja kaikille
  - Ei ositusta
  - Ei ryvästymistä

  - Painomuuttuja = 1 kaikille

EsimerkkI: Belgian ESS-data

- Belgian ESS-aineisto: Mutkikkaampi tilanne
  - Ositettu kaksiasteinen ryväsotanta
  - Sisältymistodennäköisyydet vaihtelevat ositteittain
  - Provinssiperusteinen ositus
  - Alueelliset poimintarypäät
  - Painomuuttujat vaihtelevat ositteittain

ESS – Painotuksen tarve analyysissa

- 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 – Painomuuttujat

- Painomuuttujien avulla voidaan kompensoida
  - Erisuurien sisältymistodennäköisyyksien vaikutus
  - Vastauskadon vaikutus (ESS: ei ole tehty)

- ESS- painomuuttujat DWEIGHT ja PWEIGHT
  - Erilainen käyttö **maakohtaisessa** analyysissa ja **yhdistetyn** aineiston analyysissa
- Miltä data näyttää?

ESS – Esimerkki, painojen vaikutus

- 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

- Vaikutus keskiarvoihin
  - Painottamattomat ja painotetut keskiarvot voivat poiketa paljon
  - 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, esim. 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

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

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

ESS – Yhteiskuntatieteellinen tutkimus

- Esimerkki (full text)

ACTA SOCIOLOGICA 2010

Unemployment and Subjective Well-being

An Empirical Test of Deprivation Theory, Incentive Paradigm and Financial Strain Approach

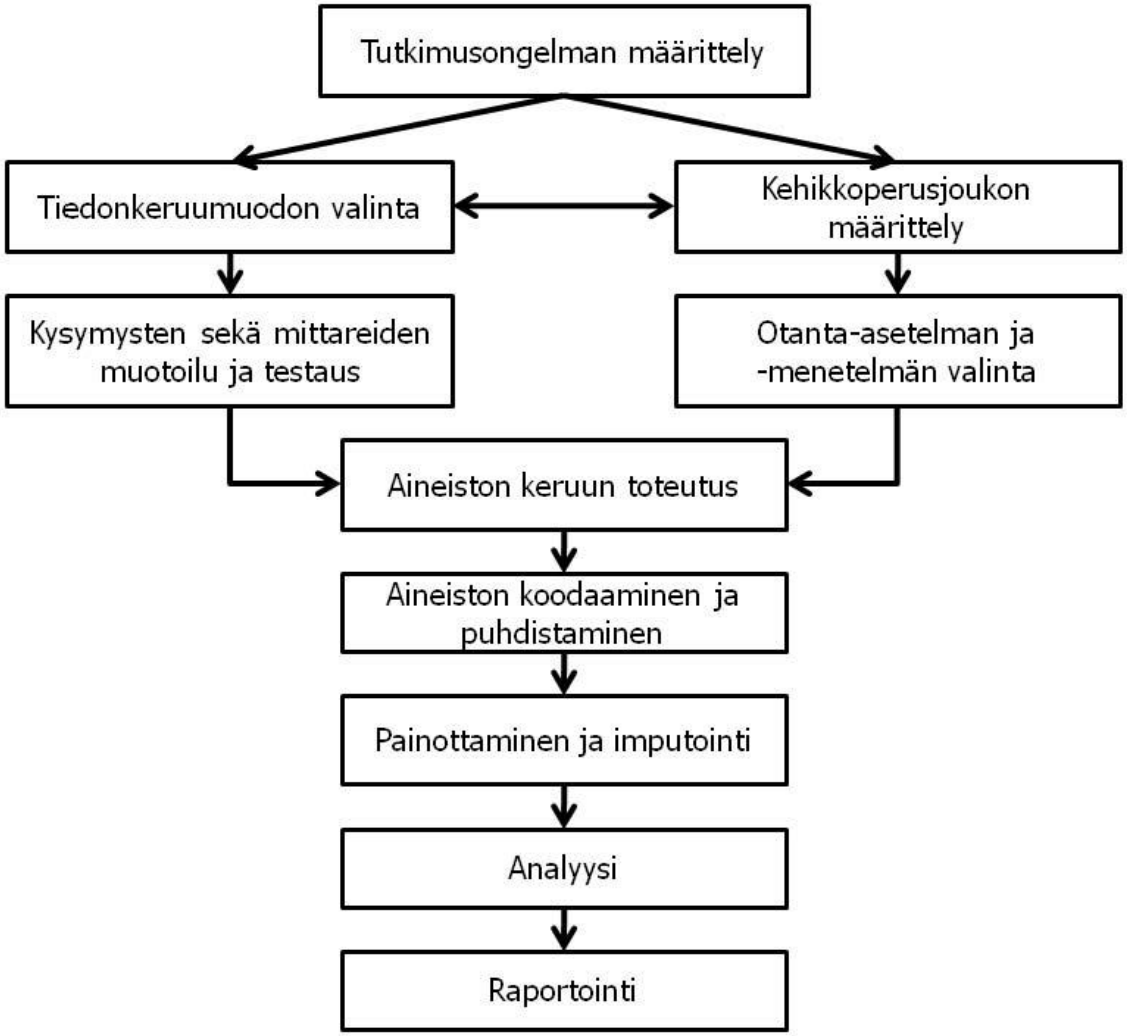
Heikki Ervasti

Department of Social Research, University of Turku, Turku, Finland

Takis Venetoklis

Government Institute for Economic Research, Helsinki and Department of Social Research, University of Turku, Finland

Itse kerätyt otosaineistot (Jani Miettinen, pro gradu)



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

Otosperusteisen tutkimuksen työvaiheet listattuna - 1

- Otanta-asetelman laadinta
  - Perusjoukon määrittely
  - Kehikkoperusjoukon muodostaminen
  - Otoskoon kiinnittäminen
  - Otantamenetelmien valinta
- Otoksen poiminta kehikkoperusjoukosta
  - Sisältymistodennäköisyyksien laskenta
  - Otanta
  - Otostiedoston muodostus
- Tiedonkeruuoperaatiot
  - Haastattelut, kyselyt ym.

Työvaiheet listattuna - 2

- Tiedonkeruuoperaatiot
  - Haastattelut, kyselyt ym.

- Aineiston muodostus
  - Asetelmapainojen muodostus
  - Katoanalyysi
  - Editointi ja imputointi
  - Lopullisten painomuuttujien muodostus
  - Asetelmaindikaattoreiden liittäminen
    - Ositus, ryvästyminen

- Analyysi ja raportointi

Esimerkki: Nettikyselyt (Jani Miettinen, pro gradu)

Taulukko 3.1. Verkkokyselytutkimuksen asetelmat (Couper 2000).

Otosperusteiset menetelmät	**Ei-satunnaiset menetelmät**
(1) Verkkosivujen käyttäjäkyselyt	(6) Itsevalikoituneet verkkokyselyt
(2) Listauksista kerätyt otokset	(7) Vapaaehtoiset paneelitutkimukset
(3) Vaihtoehto vastata verkon välityksellä	(8) Viihdegallupit
(4) Paneeli etukäteen värvätyistä Internetin käyttäjistä	
(5) Paneeli etukäteen värvätyistä populaation edustajista	

Nettikysely

- Itsevalikoituva web-kysely (verkkokysely)
  Self-selection web survey

- "Työttömyys hävettää entistä harvempia nuoria”
  - julkaistu ma 31.8.2009 klo 05:56
  - YLE Uutiset

- ”Monet nuoret työttömät suhtautuvat työttömyyteensä myönteisesti. Uuden tutkimuksen mukaan yli 40 prosenttia työttömistä nuorista ei pidä työttömyyttä pahana asiana, jos toimeentulo on muuten turvattu.”

Nuoria työttömiä koskeva nettikysely

- Ministry of Labour (MOL), työministeriö
  - Toukokuu 2009
  - *n* = 716 nuorta työtöntä (16-29 v.)
  - Web-lomake on ollut MOL:n sivustolla

- Nuori työtön on löytänyt lomakkeen työtä tai työvoimatoimenpiteitä koskevaa tietoa etsiessään

- "Oletko nuori aikuinen, joka on ollut joskus työttömänä tai olet par- aikaa työtön...- jos , niin vastaa...”

Nettikysely… ja vaihtoehdot?

- Millaisia yleistyksiä itsevalikoituvan nettikyselyn perusteella voidaan tehdä?

- Miten nettikyselyn yleistettävyyttä voidaan parantaa tilastotieteen menetelmillä?
  - Jani Miettinen, HY tilastotieteen gradu (2011)

- *How accurate are self-selection web surveys?* Jelke Bethlehem, Statistics Netherlands, Discussion paper (08014)

- Millaisia vaihtoehtoja nettikyselylle voisi olla?

Aineistotyyppien yhdistelmät

HY Otantamenetelmät syksy 2011 Risto Lehtonen
YHTEENVETO 1. Aineisto-optiot tiedonkeruun tavan ja kattavuuden mukaan.

TIEDONKERUUTAPA	KATTAVUUS PERUSJOUKON SUHTEEN	
	A. OSITTAINEN KATTAVUUS: OTOSTUTKIMUS	**B. TÄYSI KATTAVUUS: KOKONAISTUTKIMUS**
1. SUORA TIEDONKERUU **Tietolähde** **Haastattelututkimus** Tietokoneavusteinen käyntihaastattelu *Computer Assisted Personal Interview* CAPI Tietokoneavusteinen puhelinhaastattelu *Computer Assisted Telephone Interview* CATI Tietokoneavusteinen kysely *Computer Assisted Self-interview* CASI *Computer Assisted Web Survey* CAWI Tiedonkeruu kynä-ja-paperi -menetelmällä *Paper-and-Pencil Interview* PAPI **Postikysely** **Internet-kysely, Web-kysely, eSurvey**	**Optio 1a. Suoraan tiedonkeruuseen perustuva otostutkimus** Perinteinen otostutkimuksen tyyppi **Tilastokeskuksen tutkimuksia ja tilastoja** □ Työvoimatutkimus □ Kulutustutkimus **Kelan tutkimuksia ja selvityksiä** □ Terveysturvan väestötutkimukset **Monikansallisia tutkimuksia** □ European Social Survey ESS □ PISA	**Optio 1b. Suoraan tiedonkeruuseen perustuva kokonaistutkimus** Perinteinen kokonaistutkimuksen tyyppi □ Tilastokeskuksen väestölaskennat (vuoteen 1985 saakka)
2. EPÄSUORA TIEDONKERUU **Tietolähde:** **Rekisteri** Kattaa kohdeperusjoukon Päivitetään säännöllisesti **Hallinnollinen rekisteri** Hallinnollisen proseduurin oheistuote **Tilastorekisteri** Usean hallinnollisen rekisterin yhdistelmä	**Optio 2a. Hallinnolliseen rekisteriaineistoon perustuva otostutkimus** Puhtaana muotona harvinainen □ Poikkeuksena Tilastokeskuksesta saatavat tilastorekistereiden otosaineistot	**Optio 2b. Hallinnolliseen rekisteriin tai tilastorekisteriin perustuva kokonaistutkimus** Tämä surveyn tyyppi on yleistymässä Aineistolähteet □ Rekisteriperusteiset väestölaskennat □ Sosiaalivakuutuksen rekisterit □ Väestörekisteri □ Yritysrekisteri □ Verotusrekisterit □ Kelan lääketutkimukset
3. TIEDONKERUUTAPOJEN YHDISTELMÄ **Tietolähde:** Suoran ja epäsuoran tiedonkeruun yhdistelmä	**Optio 3. Otostutkimus, joka perustuu suoran tiedonkeruun ja rekisteriaineiston yhdistelyyn** Tämä surveyn tyyppi on yleistymässä □ KTL:n Terveys 2000 ja Terveys 2010 □ Kelan Mini-Suomi-terveystutkimus □ Tilastokeskuksen Tulonjakotutkimus □ EU:n European Community Household Panel ECHP □ EU SILC (Statistics on Income and Living Conditions)	

Jatkotarkastelut...

- Otanta- ja survey-metodiikka

- Syksy 2012

I periodi
Survey-metodiikka

II periodi
Otantamenetelmät