Pienalue-estimointi (78189)
Kevät 2009

Risto Lehtonen
risto.lehtonen@helsinki.fi

Pienalue-estimointi

Luennoija: Prof. Risto Lehtonen

Laajuus 6/8 op

Tyyppi
Aineopintojen erikoiskurssi
Syventävien opintojen erikoiskurssi

Luennot
Maanantaisin ja tiistaisin klo 16–18 Exactum C323
9.3.–16.4.2009 (yhteensä 20 tuntia)

Harjoitukset
Torstaisin klo 12–15 Mikroluokka C128
12.3.–16.4.2009 (yhteensä 12 tuntia)

Loppukuulustelu Tiistai 21.4.2009 klo 16–18 Exactum C323

Suoritustapa
Aineopinnot: Loppukuulustelu (6 op) tai loppukuulustelu ja (vapaaehtoinen) harjoitustyö (8 op)
Syventävät opinnot: Loppukuulustelu ja (pakollinen) harjoitustyö (8 op)

Tavoitteet

Kurssin tavoitteena on perehdyttää opiskelija perusjoukon osajoukkoja koskevan estimoinnin (small area estimation, SAE) lähestymistapoihin, teorioihin, laskentamenetelmiin ja sovelluksiin.

Kurssilla käsitellään asetelmaperusteisia malliavusteisia (design-based model assisted) menetelmiä, kuten yleistetyt regressioestimaattorit (GREG) ja kalibrointimenetelmät, sekä malliperusteisia (model-based) menetelmiä, kuten synteettiset ja EBLUP-estimaattorit.

Lisäksi tarkastellaan estimointiin soveltuvia tilastollisia ohjelmistoja.

Sovellukset ovat pääasiassa yhteiskuntatieteellisiltä ja terveystieteellisiltä aloilta.

Käytännön harjoituksissa käytetään laskentatyökaluina mm. SAS-ohjelmistoa ja erillisiä alan ohjelmia.

Kurssin menestyksekästä suorittamista edesauttaa, jos opiskelijalla on perustiedot otantamenetelmistä ja tilastollisista malleista.

Kurssi soveltuu tilastotieteen aine- tai syventäviä opintoja suorittaville opiskelijoille sekä myös yliopistoissa, korkeakouluissa ja tutkimuslaitoksissa toimiville jatko-opiskelijoille ja tutkijoille.

Hyödyllisiä taustatietoja

- Otantamenetelmät
  - Lehtonen R. and Djerf K. (2008). *Survey sampling reference guidelines.* Luxembourg: Eurostat Methodologies and Working papers
  - Saatavilla vapaasti osoitteessa: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF
- Tilastollisten mallien perusteita
  - Lineaariset mallit
  - Yleistetyt lineaariset mallit

SAE: Kirjallisuutta

- Rao J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons.
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys. Second Edition*. Chichester: John Wiley & Sons.

  http://books.google.fi/books?id=Xb-m5Xg74F4C

Web extension:

VLISS-Virtual Laboratory in Survey Sampling
http://mathstat.helsinki.fi/VLISS/

SAE: Web-materiaalia

- EURAREA Project
  http://www.statistics.gov.uk/eurarea/

- EWORSAE – the European Working Group on Small Area Estimation
  http://sae.wzr.pl/
- Pienalue-estimoinnin kongressit
  http://cio.umh.es/sae2009/

SAE: Kirjallisuutta

- Lehtonen R. and Djerf K. (eds.) (2001). *Lecture Notes on Estimation for Population Domains and Small Areas.* Statistics Finland: Reviews 2001/15.
- Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Vol. 29B. Sample Surveys: Inference and Analysis.* New York: Elsevier. (Forthcoming)

SAE: Laskentatyökaluja

- SAS
  - Procedures SURVEYMEANS
    - DOMAIN-lause
- EURAREA-projekti
  - SAS-makro Standard estimators
  - SAS-makro EBLUPGREG
- Ohjelma DOMEST
  - Ari Veijanen
- R-kielisiä ohjelmia

Ohjelma ja aikataulurunko		
Ma 9.3.2009 klo 16–18 C323	1. luento	Johdanto Eri lähestymistavat
Ti 10.3.	Ei luentoa (työmatka)	
To 12.3. klo 12–15 C128	1. demot	SAS-johdattelu
Ma 16.3. klo 16–18 C323	2. luento	Asetelmaperusteinen estimointi 1
Ti 17.3. klo 16–18 C323	3. luento	Asetelmaperusteinen estimointi 2
To 19.3. klo 12–15 C128	2. demot	HT-estimointi SAS Proc Surveymeans

Ma 23.3. klo 16–18 C323	4. luento	Malliavusteinen estimointi 1
Ti 24.3. klo 16–18 C323	5. luento	Malliavusteinen estimointi 2
To 26.3. klo 12–15 C128	3. demot	GREG-estimointi SAS Macro EBLUPGREG
Ma 30.3. klo 16–18 C323	6. luento	Malliperusteinen estimointi 1
Ti 31.3. klo 16–18 C323	7. luento	Malliperusteinen estimointi 2
To 2.4. klo 12–15 C128	4. demot	EBLUP-estimointi SAS Macro EBLUPGREG
Ma 6.4. klo 16–18 C323	8. luento	Esimerkkejä 1
Ti 7.4. klo 16–18 C323	9. luento	Esimerkkejä 2
To 16.4. klo 12–15 C128	10. luento 5. demot	Yhteenveto Ohjelman DOMEST esittely
Ti 21.4. klo16–18 C323	Lopputentti	

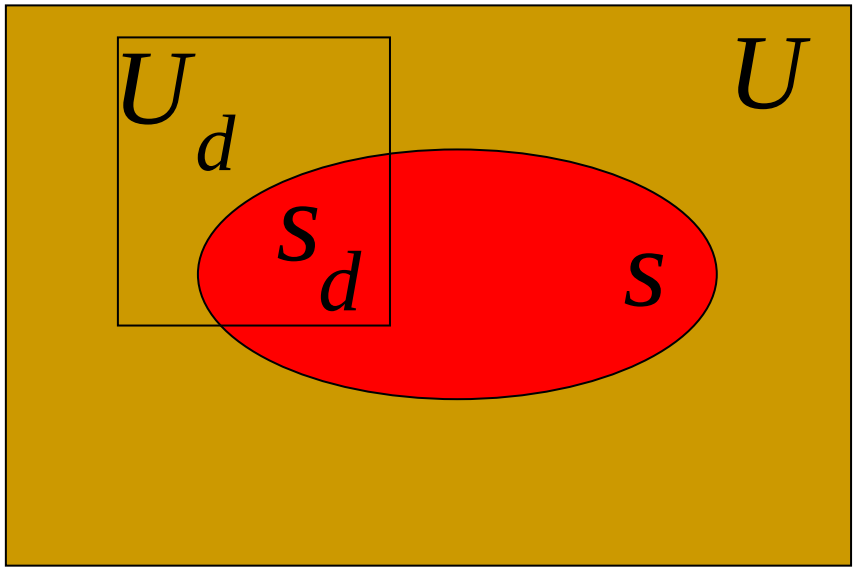
JOHDANTO
Eri lähestymistavat

Käsitteitä ja määritelmiä

- Perusjoukon osajoukko
- *Domain*

- Pienalue
- *Small area*
  - Lääni, maakunta, seutukunta, kunta…
  - Väestön demografiset ja sosioekonomiset osajoukot
  - Yritysten toimialakohtaiset osajoukot

- *”Small”?*
  - **Osajoukon otoskoko on pieni**



U Perusjoukko

s Otos

U_d Perusjoukon osajoukko d

s_d Otos osajoukossa d

$s_d = s \cap U_d$

Estimointitehtävä

Osajoukot

Estimoinnin kohteena olevia osajoukkoja on yleensä runsaasti

U_d, s_d, $d = 1, \ldots, D$

D suuri

Estimoitavat parametrit

Osajoukkojen kokonaismäärät (totaalit) T_d ja keskiarvot $\bar{Y}_d$

$$T_d = \sum_{k \in U_d} Y_k, \; \bar{Y}_d = \sum_{k \in U_d} Y_k / N_d, \; d = 1, \ldots, D$$

Esimerkki

ILO-työttömien lukumäärä maakunnittain sukupuolen ja ikäryhmän mukaan

Tilastokeskuksen työvoimatutkimuksen perusteella

Käytettävissä olevat tulot keskimäärin henkilöä kohti kunnittain Tilastokeskuksen

EU-SILC-tutkimuksen mukaan

Estimointitehtävä: Estimaattorin tyyppi

- Perusjoukon osajoukkoja koskevien parametrien estimointi
- *Estimation for domains*
  - **Asetelmaperusteiset menetelmät**
  - *Design-based methods*
- Pienalue-estimointi
- *Small area estimation*
  - **Malliperusteiset menetelmät**
  - *Model-based methods*

Suora (direct) ja epäsuora (indirect) estimaattori

(Lehtonen and Veijanen 2009)

A *direct* estimator uses values of the variable of interest only from the time period of interest and only from units in the domain of interest (Federal Committee on Statistical Methodology, 1993).

A Horvitz-Thompson type estimator

$$\hat{t}_d = \sum_{k \in s_d} y_k / \pi_k$$

provides a simple example of direct estimator.

In model-assisted estimation, direct estimators are constructed by using models fitted separately in each domain; an example is a model

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta}_d + \varepsilon_k, \; k \in U_d,$$

with domain specific auxiliary *x*-data and a vector of regression coefficients $\boldsymbol{\beta}_d$, $d = 1, \ldots, D$.

A direct domain estimator can still incorporate auxiliary data outside the domain of interest. This is relevant if accurate population data about the auxiliary *x*-variables are only available at a higher aggregate level.

An *indirect* domain estimator uses values of the variable of interest from a domain and/or time period other than the domain and time period of interest (Federal Committee on Statistical Methodology, 1993).

For example, if a linear model

$$Y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k\,,\ k \in U\ ,$$

with a common vector $\boldsymbol{\beta}$ is used as an assisting model, the resulting domain estimator will be indirect.

In general, indirect estimators are attempting to "borrow strength" from other domains and/or in a temporal dimension. The concept of "borrowing strength" is often used in model-based small area estimation (e.g. Rao, 2003).

Indirect model-assisted estimators for domains are discussed in the literature (e.g. Estevao and Särndal, 1999, Lehtonen, Särndal and Veijanen, 2003, 2005, and Hidiroglou and Patak, 2004). Estevao and Särndal (2004) have argued in favour of direct estimators in the context of design-based estimation for domains.

Estimointitehtävä

- **Suuri osajoukko – Large domain**

- Osajoukko jossa on mahdollista tuottaa riittävällä tarkkuudella **asetelmaperusteinen suora** (*direct*) estimaatti

- **Pieni osajoukko – Small domain**
- Pieni osajoukko = Osajoukko jossa **ei ole** mahdollista tuottaa riittävällä tarkkuudella **asetelmaperusteinen suora** (*direct*) estimaatti

- Tarvitaan **malliperusteisia epäsuoria** (*indirect*) estimaattoreita
  - ”Voiman lainaaminen”
  - *Borrowing strength*

- Tämän kurssin alue: *Estimation for domains and small areas*

Voiman lainaaminen - Lisäinformaatio

- Kaikissa tarkasteltavissa menetelmissä on olennaista:
  - Perusjoukkoa koskevan lisäinformaation hyvä saatavuus
    - *Auxiliary data, auxiliary information*
    - Rekistereistä saatavat lisätiedot, apumuuttujat
  - Lisäinformaation tuonti estimointiproseduuriin tilanteeseen soveltuvien tilastollisten mallien avulla
    - Lineaariset mallit, logistiset mallit, sekamallit
    - Yleistetyt lineaariset sekamallit (Generalized linear mixed models)

Asetelmaperusteiset menetelmät

- **Asetelmaperusteiset suorat estimaattorit**
  - Horvitz-Thompson (HT) –estimaattori
  - Hájek-tyyppinen estimaattori
- **Asetelmaperusteiset malliavusteiset estimaattorit**
  - Suoria tai epäsuoria estimaattoreita
  - Yleistetyt regressioestimaattorit (generalized regression estimators) GREG
    - Särndal, Swensson and Wretman (1992)
    - Lehtonen, Särndal and Veijanen (2003, 2005)
    - Lehtonen and Pahkinen (2004), luku 6
    - Lehtonen and Veijanen (2009)

Malliperusteiset menetelmät

- **Synteettiset estimaattorit SYN**
- **EBLUP-estimaattorit**
- Empirical Best Linear Unbiased Predictor EBLUP
  - Rao (2003)
  - EURAREA-projekti
- **Statistics in Transition –lehden erikoisnumerot**
  - Joulukuu 2005
  - Maaliskuu 2006

Estimaattoreiden ominaisuuksia

- **Table 1**
- Asetelmaperusteiset estimaattorit HT, GREG
  - Likimain harhattomia
  - Varianssi voi kasvaa suureksi, jos osajoukon otoskoko on pieni
    - Varianssi pienenee otoskoon kasvaessa
- Malliperusteiset estimaattorit SYN, EBLUP
  - Harhaisia määritelmän mukaan (“All models are wrong”).
    - Harha ei pienene otoskoon kasvaessa!
  - Varianssi voi olla pieni myös pienissä osajoukoissa
  - MSE voi olla suuri jos harha on dominoiva

Table 1. Design-based properties of model-assisted and model-dependent estimators for domains and small areas

	Design-based model-assisted methods GREG and calibration estimators	**Model-dependent methods** Synthetic and EBLUP estimators
Bias	Design unbiased (approximately) by the construction principle	Design biased Bias does not necessarily approach zero with increasing domain sample size
Precision (Variance)	Variance may be large for small domains Variance tends to decrease with increasing domain sample size	Variance can be small even for small domains Variance tends to decrease with increasing domain sample size
Accuracy (Mean Squared Error, MSE)	MSE = Variance (or nearly so)	MSE = Variance + squared Bias Accuracy can be poor if the bias is substantial
Confidence Intervals	Valid design-based intervals can be constructed	Valid design-based intervals not necessarily obtained

Estimaattoreiden työnjako

- Asetelmaperusteisia estimaattoreita (HT, GREG) käytetään tyypillisesti suurille osajoukoille (suuri otoskoko, pieni varianssi).
- Malliperusteisia estimaattoreita (SYN, EBLUP) käytetään pienille osajoukoille, (pieni otoskoko, pieni varianssi) joissa asetelmaperusteiset estimaattorit toimivat huonosti (suuri varianssi).

Table 2. Application areas of estimation approaches by domain sample size

	DOMAIN SAMPLE SIZE		
ESTIMATION APPROACH	**Minor**	**Medium**	**Major**
Model-based			
Synthetic SYN	++	+	**0**
EBLUP	+++	++	++
Design-based			
Horvitz-Thompson HT	**0**	+	++
Model-assisted GREG	+	++	+++

Applicability

0 **Not at all**
+ **Low**
++ **Medium**
+++ **High**