HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Pienalue-estimointi (78189)
Kevät 2013

Risto Lehtonen
Helsingin yliopisto

Pienalue-estimointi

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

- **Asetelmaperusteiset** malliavusteiset (*design-based model assisted*) menetelmät,
  - Yleistetyt regressioestimaattorit (GREG) ja kalibrointimenetelmät

- **Malliperusteiset** (*model-based*) menetelmät
  - Synteettiset ja EBLUP-tyyppiset estimaattorit.

Pienalue-estimointi

- Lisäksi tarkastellaan estimointiin soveltuvia tilastollisia ohjelmistoja.
- Sovellukset ovat pääasiassa yhteiskuntatieteellisiltä ja terveystieteellisiltä aloilta.
- **Käytännön harjoituksissa** käytetään laskentatyökaluina SAS-ohjelmiston proseduureja ja makroja sekä erikoisohjelmia kuten DOMEST.

- Kurssin suorittaminen ym. käytännön seikat
  - Kurssin kotisivu

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 ja teoriaa
  - Lineaariset mallit
  - Yleistetyt lineaariset mallit

Maailmanlaajuinen trendi

- Yhteiskunnassa on lisääntyvä tarve tuottaa luotettavia tietoja alueellisille ja muille populaation (perusjoukon) osajoukoille
  - *Estimation for domains*
  - *Small area estimation*
- EU:n tutkimuksen puiteohjelmien projekteja
  - EURAREA Project (2001-2004)
  - AMELI Project (2008-2012)
  - SAMPLE Project (2008-2012)
  - ESSnet SAE (2009-2012)

SAE Conferences

- EWORSAE
  European Working Group on Small Area Estimation http://sae.wzr.pl/

  SAE2005 (University of Jyväskylä)
  SAE2007 (University of Pisa)
  SAE2009 (University of M. Hernandez, Elche)
  SAE2011 (University of Trier)
  Forthcoming
  SAE2013 (Bangkok)

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://vliss.helsinki.fi/

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.

SAE: Laskentatyökaluja

- SAS 9.3
  - Procedure SURVEYMEANS
  - Procedure SURVEYREG
  - Procedure SURVEYLOGISTIC
  - Osajoukkoja koskeva estimointi: DOMAIN-lause
- EURAREA-projekti
  - SAS-makro: Standard estimators
  - SAS-makro: EBLUPGREG
- Ohjelma DOMEST
  - Ari Veijanen & Risto Lehtonen
- R-kielisiä ohjelmia

JOHDANTO LÄHESTYMISTAPOJA

Käsitteitä ja määritelmiä

- Perusjoukon osajoukko *Domain*
- Pienalue *Small area*
  - Lääni, maakunta, seutukunta, kunta…
  - Väestön demografiset ja sosioekonomiset osajoukot
  - Yritysten toimialakohtaiset osajoukot
- Estimoidaan otosaineiston perusteella:
  - Kokonaismääriä *Totals*
  - Keskiarvoja *Means*
  - Osuuksia *Proportions*
  - Mediaaneja…

  määritellyille perusjoukon osajoukoille (*domains, small areas*)

Erikoistapaus - SAE

- *Small area estimation,* SAE
  - Estimointi tilanteessa, jossa osajoukkojen otoskoko on pieni
  - Vaihtoehtoinen määritelmä (Partha Lahiri):

    Small area = Domain of interest, for which the sample size is not adequate to produce reliable **direct estimates**

Tyypillinen estimointitehtävä

Määritellään ja identifioidaan osajoukot

Osajoukkojen U_d lkm D on yleensä suuri

Spesifioidaan tulosmuuttujan y parametrit

Osajoukkototaalit $t_d = \sum_{k \in U_d} y_k$

Keskiarvot $\bar{Y}_d = t_d / N_d, \; d = 1, ..., D$

missä N_d on osajoukon koko pj:ssa

Esimerkkejä

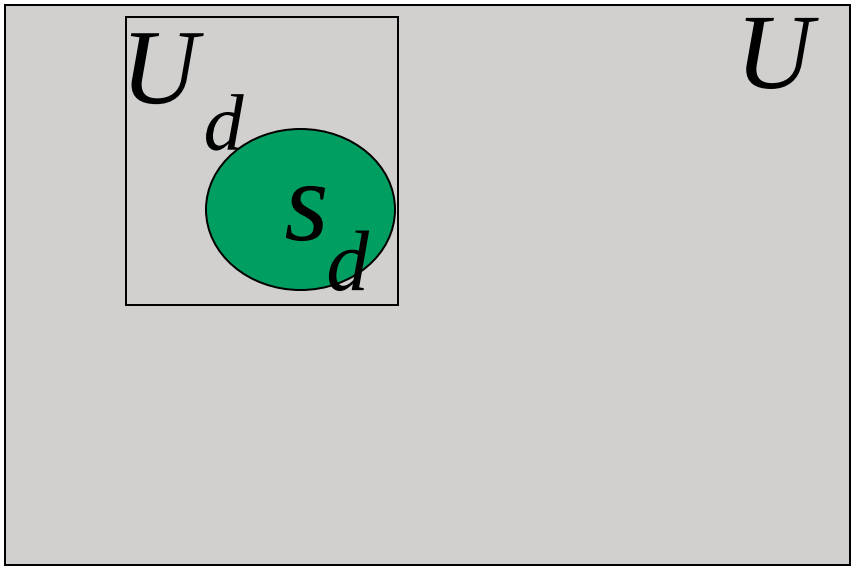
- Työttömien kokonaismäärän estimointi alueittain sukupuolen ja ikäryhmän mukaan muodostetuissa osajoukoissa
  - Tilastokeskuksen työvoimatutkimuksen aineisto
- Kotitalouksien käytettävissä olevien tulojen mediaanin estimointi kunnittain
  - EU:n SILC-tutkimusaineisto
- Alueellisten köyhyysasteiden estimointi
  - EU:n SILC-tutkimusaineisto

Tärkeitä kysymyksiä ja käsitteitä

- Osajoukkojen tyyppi?
  - Suunnitellut / Ei-suunnitellut osajoukot
    Planned domains / Unplanned domains
- Aineistolähteet?
  - Otosaineisto *Sample data*
  - Lisäinformaatio *Auxiliary data*
- Estimaattorin tyyppi?
  - Suora / Epäsuora *Direct / Indirect*
  - Asetelmaperusteinen / Malliperusteinen
    Design based / Model-based
- Mallin tyyppi?
  - Lineaarinen / Epälineaarinen *Linear / Non-linear*
  - Kiinteät vaikutukset / Sekamallit (*Mixed models*)

Osajoukon tyyppi

- Suunnitellut osajoukot *Planned domains*
  - Usein tärkeimmät osajoukkotyypit pyritään määrittelemään otanta-asetelmassa **ositteiksi** (*strata*)
  - Osajoukkojen otoskoot on kiinnitetty
  - Otoskoot hallitaan kiintiöintimenetelmillä (*allocation*)
  - Liian pienet otoskoot voidaan välttää
- Ei-suunnitellut osajoukot *Unplanned domains*
  - Osajoukkojen otoskoot ovat satunnaismuuttujia
  - Voi tulla osajoukkoja joiden otoskoko pieni
  - Käytännössä yleinen tilanne (Miksi?)



Planned domains

U Population

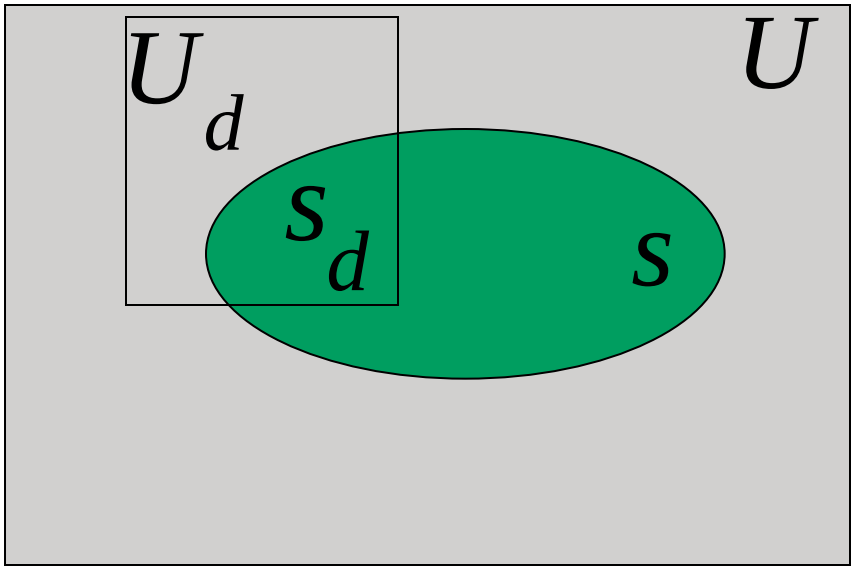
U_d Population domain d

Domains = Strata

$s_d \subset U_d$ Sample in domain d

Sample size n_d in domain d is fixed

$d = 1,...,D$



Unplanned domains

U Population

s Sample

U_d Population domain d

$s_d = s \cap U_d$ Sample in domain d

Sample size n_d in domain d is random

$d = 1,...,D$

Suora ja epäsuora estimaattori

- Suora estimaattori *Direct estimator*
  - *Direct* domain estimator uses values of the variable of interest *y* only from the time period of interest and only from units in the domain of interest (Federal Committee on Statistical Methodology, 1993)
  - **Suunniteltujen** osajoukkorakenteiden tilanne
- Epäsuora estimaattori *Indirect estimator*
  - *Indirect* domain estimator uses values of the variable of interest *y* from a domain and/or time period other than the domain and time period of interest
  - **Ei-suunniteltujen** osajoukkorakenteiden tilanne

Esimerkki: Suora HT-estimaattori

Asetelmaperusteinen **Horvitz-Thompson** (HT) estimaattori on muotoa

$$\hat{t}_{dHT} = \sum_{k \in s_d} y_k / \pi_k \ , \ d = 1, \ldots, D$$

missä π_k on sisältymistodennäköisyys alkiolle *k* ja s_d on otos osajoukosta U_d

HT käyttää y-arvoja vain osajoukosta s_d

$\hat{t}_{dHT}$ on **suora** estimaattori

Esimerkki: Suora GREG-estimaattori

Suora asetelmaperusteinen malliavusteinen GREG-estimaattori

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} (y_k - \hat{y}_k) / \pi_k$$

käyttää lineaarisia malleja jotka on spesifioitu erikseen kullekin osajoukolle:

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta}_d + \varepsilon_k, \; k \in U_d, \; d = 1, \ldots, D$$

missä $\boldsymbol{\beta}_d$ on osajoukkokohtainen, ja

$\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_d$ ovat sovitteita, laskettu jokaiselle $k \in U_d$

Esimerkki: Epäsuora GREG-estimaattori

Epäsuora asetelmaperusteinen GREG-estimaattori

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} (y_k - \hat{y}_k) / \pi_k$$

käyttää koko aineistolle sovitettua lineaarista mallia

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

jossa vektori $\boldsymbol{\beta}$ on yhteinen kaikille osajoukoille

ja $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ lasketaan kaikille $k \in U$

Esimerkki: Epäsuora SYN-estimaattori

Epäsuora malliperusteinen synteettinen SYN-estimaattori

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k \ ,$$

käyttää koko aineistolle sovitettua lineaarista mallia

$$Y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k \ , \ d = 1, ..., D$$

jossa $\boldsymbol{\beta}$ on yhteinen kaikille osajoukoille

ja $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ lasketaan kaikille $k \in U$

“Voiman lainaaminen”
“*Borrow strength*”

- Epäsuorat estimaattorit pyrkivät “lainaamaan voimaa”
  - Muista osajoukoista (spatiaalinen dimensio)
  - Saman osajoukon aikaisemmista mittauksista (temporaalinen dimensio)
- Tyypillistä erityisesti pienten osajoukkojen tilanteissa (pieni otoskoko)
- “*Borrowing strength*” käytetään usein malliperusteisissa SAE-tilanteissa

Estimointitehtävä

- **A. Suuri osajoukko – Large domain**
- Osajoukko jossa on mahdollista tuottaa riittävällä tarkkuudella **asetelmaperusteinen suora** (*direct*) estimaatti
- **Tämän kurssin alue: *Estimation for domains and small areas* eli tilanteet A ja B**

- **B. 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*

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 GLMM*

Asetelmaperusteiset menetelmät

- **Suorat estimaattorit**
  - Horvitz-Thompson (HT) –estimaattorit
  - Hájek-tyyppiset estimaattorit
- **Malliavusteiset estimaattorit**
  - Suoria tai epäsuoria estimaattoreita
  - Yleistetyt regressioestimaattorit (*generalized regression estimators*) GREG
  - Kalibrointiestimaattorit
  - *Model calibration MC / Model free calibration*
    - Särndal, Swensson and Wretman (1992)
    - Lehtonen, Särndal and Veijanen (2003, 2005)
    - Lehtonen and Pahkinen (2004), Chapter 6
    - Lehtonen and Veijanen (2009, 2012)

Malliperusteiset menetelmät

- **Synteettiset estimaattorit SYN**
- **EBLUP- ja EBP-estimaattorit**
  - *EBLUP: Empirical Best Linear Unbiased Predictor*
    EBP: Empirical Best Predictor
  - Rao (2003)
  - EURAREA-projekti, Domest-ohjelma
- Bayes-menetelmät
  - Empirical Bayes, Hierarchical Bayes
- “Poverty mapping”
  - World Bank, Peter Lanjouw, Chris Elbers,…
  - PovMap Software

HUOM: Tilastollisen mallin rooli

- **Asetelmaperusteinen GREG, MC**
  - Malleja käytetään avustavina työkaluina
  - GREG-estimaattorit ja MC-estimaattorit ovat **malliavusteisia** (*model-assisted*)
- **Malliperusteiset SYN, EBLUP, EBP**
  - Nojautuvat tilastolliseen malliin
  - SYN-estimaattorit ovat **malliperusteisia** (*model-based / model-dependent*)
- HUOM: *All models are wrong but some are useful!*

Estimaattoreiden ominaisuuksia

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

Estimaattoreiden tilastollisten ominaisuuksien vertailu

Osajoukkojen totaalien t_d
estimaattoreiden $\hat{t}_d$ vertailu:

Harha:

Bias $Bias(\hat{t}_d) = E(\hat{t}_d) - t_d$

Varianssi:

Precision $Var(\hat{t}_d) = E(\hat{t}_d - E(\hat{t}_d))^2$

Keskineliövirhe:

Accuracy

$$MSE(\hat{t}_d) = E(\hat{t}_d - t_d)^2 = Var(\hat{t}_d) + Bias^2(\hat{t}_d)$$

Design-based properties of estimators

	Design-based methods HT, GREG, MC	Model-based methods SYN, EBLUP, EBP
Bias	**Design unbiased** (approximately) by the construction principle	**Design biased** Bias does not necessarily approach zero with increasing sample size
Precision (Variance)	**Large variance for small domains** Variance decreases with increasing sample size	**Small variance for small domains** Variance decreases with increasing 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 CI can be constructed	Valid design-based CI not necessarily obtained

Estimaattoreiden työnjako

- **Asetelmaperusteisia** estimaattoreita (HT, GREG, MC) käytetään tyypillisesti **suurille osajoukoille** (suuri otoskoko, pieni varianssi).

- **Malliperusteisia** estimaattoreita (SYN, EBLUP, EBP) käytetään yleensä **pienille osajoukoille**, (pieni otoskoko, pieni varianssi) joissa asetelmaperusteiset estimaattorit toimivat huonosti (suuri varianssi).

Natural application areas of estimation approaches by domain sample size

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

Applicability

0 Not at all

\+ Low ++ Medium +++ High

Selected literature - Design-based

- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second Edition. Chichester: John Wiley & Sons. Chapter 6.
- 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. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33, 99–119.

Selected literature - Design-based

- Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology* 24, 51–55.
- Lehtonen R., Särndal C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33–44.
- Lehtonen R., Särndal C.-E. and Veijanen A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition,* 7, 649–673.
- Lehtonen, R. and Veijanen, A. (2012). Small area poverty estimation by model calibration*. Journal of the Indian Society of Agricultural Statistics* 66, 125-133.

Selected literature - Model-based

- Rao J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons.

- Longford N. (2005). *Missing Data and Small-area Estimation: Modern Analytical Equipment for the Survey Statistician*. New York: Springer.

- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *JASA* 74, 269–277.

Selected literature - Model-based

- Battese, G.E., Harter, R.M., and Fuller, W.A. (1988), An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *JASA* 80, 28–36.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science* 9, 55–93.
- Jiang J. and Lahiri P. (2006). Mixed model prediction and small area estimation. *TEST* 15, 1–96.

Additional SAE materials

- EURAREA Project

Downloads

- AMELI Project

Downloads

- BIAS Project

Downloads

- Case Study
  Lehtonen R., Myrskylä M., Särndal C.-E. and Veijanen A. (2007)
  The role of models in model-assisted and model-dependent estimation for domains and small areas (Posteri)

- Pienalue-estimointi, tekninen tarkastelu
  OSA 1 (erillinen materiaali kurssisivulla)