HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otantamenetelmät
(78143)
Syksy 2010

TEEMAT 3 & 4

Risto Lehtonen
risto.lehtonen@helsinki.fi

Teema 3

ERITYISKYSYMYKSIÄ

Otannan erityiskysymyksiä

- Ryväsotanta
  - Survey sampling reference guidelines (2008 edition) Introduction to sample design and estimation techniques Section 3.2.5 Cluster sampling
- Otoskoon määrääminen
  - Survey sampling reference guidelines (2008 edition) Introduction to sample design and estimation techniques Section 3.3 Sample size determination
- Vastauskadon hallinnan perusteita
  - Lehtonen-Pahkinen (2004) Chapter 3
  - VLISS Training Key 114, 117, 123

Sisäkorrelaatio ryväsotannassa

- Samaan rypääseen kuulumisella on taipumus samankaltaistaa alkioita tutkittavien ilmiöiden suhteen
- Koulututkimukset
  - Rypäänä opetusryhmä
  - Oppimistulokset, esim. PISA
- Työolotutkimukset
  - Rypäänä työpaikka
  - Työolot, esim. Kelan työterveyshuoltotutkimus
  - OHC data, VLISS

Sisäkorrelaatio

Sisäkorrelaatio $\hat{\rho}_{\text{int}}$ *Intra - cluster correlation*

Likimääräinen kaava $\hat{\rho}_{\text{int}} = (deff - 1)/(\bar{m} - 1)$

missä $\bar{m}$ on keskimääräinen ryväskoko

Rypäät ovat usein positiivisesti sisäkorreloituneita

eli $\hat{\rho}_{\text{int}} > 0$ kun $deff > 1$

Lisäksi on voimassa: $deff(\hat{p}) = 1 + (\bar{m} - 1)\hat{\rho}_{\text{int}}$

missä $deff(\hat{p})$ on asetelmakerroin

(*design effect*)

Esimerkkiaineisto: Työterveyshuoltotutkimus OHC

- **Otanta-asetelma**
  - Ositettu yksi- ja kaksiasteinen ryväsotanta
  - Toimipaikat rypäinä
  - Ositus rypään koon ja toimialan mukaan
    - Pienet toimipaikat: Yksiasteinen otanta
    - Suuret toimipaikat: Kaksiasteinen otanta
- Henkilötasolla likimain **itsepainottuva** (*self-weighting*) otos
- **Demonstraatioaineisto SAS-data OHC**
- Rajaus:
  - Toimipaikat, joissa vähintään 10 työntekijää
  - *H* = 5 ositetta (*strata*)
  - *m* = 250 toimipaikkaa (ryvästä, *clusters*)
  - *n* = 7841 henkilöä
- Vaihteleva määrä otosrypäitä per osite

Deff-estimaatit OHC (Lehtonen&Pahkinen 2004)

Table 5.8

Averages of design-effect estimates of proportion estimates of selected groups of binary response variables in the OHC Survey data set (number of variables in parentheses).

Study variable	Mean deff
Physical working conditions (12)	6.5
Psycho-social working conditions (11)	3.3
Psychosomatic symptoms (8)	2.0
Psychic symptoms (9)	1.8

Teema 4

OHJELMISTO

Ohjelmisto

- SAS-ohjelmisto
  - SAS-proseduurit
  - SAS-makrot
- SPSS module Complex Samples
- Stata:n svy-ohjelmat
- Erikoisalue: Pienalue-estimointi
  - SAS-makro EBLUPGREG
  - Ohjelma Domest
  - R-kieliset ohjelmat
- Software

SAS-ohjelmisto, kuvailu

- Survey-proseduurit, joilla otanta-asetelma (ositus, ryvästyminen, painotus) voidaan ottaa huomioon estimoinnissa
- SURVEYMEANS
  - Keskiarvot ja kokonaismäärät
- SURVEYREG
  - Regressioestimointi
- SURVEYFREQ
  - Ristiintaulukointi ja perustestit

SAS-ohjelmisto, analyysi

- SURVEYFREQ
  - Monipuolinen valikoima tilastollisia testejä
- SURVEYREG
  - Lineaariset mallit
- SURVEYLOGISTIC
  - Logistiset mallit

SAS-proseduuri SURVEYMEANS

- Asetelmaperusteinen estimointi
- Kokonaismäärien, keskiarvojen ja osuuksien estimointi koko aineistossa ja osajoukoissa
- Osajoukkoestimointi (domains)
  - BY statement
  - Ositekohtainen estimointi (*planned domains*)
  - DOMAIN statement
  - Estimointi muuntyyppisille osajoukoille (*unplanned domains*)

SAS-makrot, pienalue-estimointi

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

- Asetelmaperusteinen estimointi
  - GREG with linear fixed-effects models
- Malliperusteinen estimointi
  - Standard estimators (EBLUP)
  - Estimators with spatial or temporal effects
  - Estimators for cross-classifications
- Proper MSE estimation

EURAREA Project

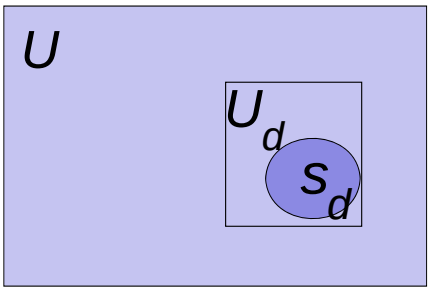
- The EURAREA "Standard" Estimators and performance criteria*
Office for National Statistics, UK
- Area-level Composite Estimator with Time-Varying Area Effect*
Office for National Statistics, UK
- **EBLUPGREG: Unit-level Composite Estimator with Spatial or Temporal Effects* Statistics Finland**
- Unit-level Composite Estimator with Spatial Effects*
ISTAT, Italy
- Small Area Estimation with Sampling Weights*
INE, Spain / UMH, Spain
- Cross Classifications with Two-way and Three-way tables*
ISTAT, Italy

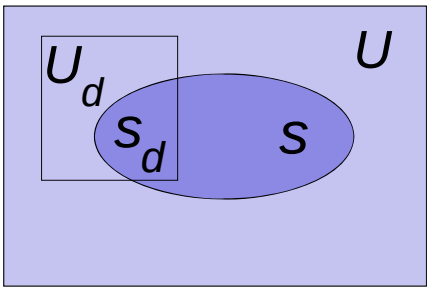
Estimation for domains and SAE

- *Domain estimation*
  - The estimation of population quantities for the desired population subgroups called domains (large or small)
    - Totals , Means, Proportions
    - Medians, Quantiles, Percentiles
    - More complex indicators…
- *Small area estimation*, SAE
  - Estimation for domains whose **sample size** is small or very small (even zero)

Domain estimation design

- Types of domains of interest
  - Planned domains / Unplanned domains
- Type of domain estimator
  - Direct / Indirect
  - Design-based / Model-based
- Availability of auxiliary (population) data
  - Unit-level / Aggregated (area-level)
- Type of model
  - Linear / Non-linear
  - Fixed-effects model / Mixed model
  - Generalized linear mixed models (GLMM)
- Accuracy measures
  - Variance estimators / MSE estimators





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**

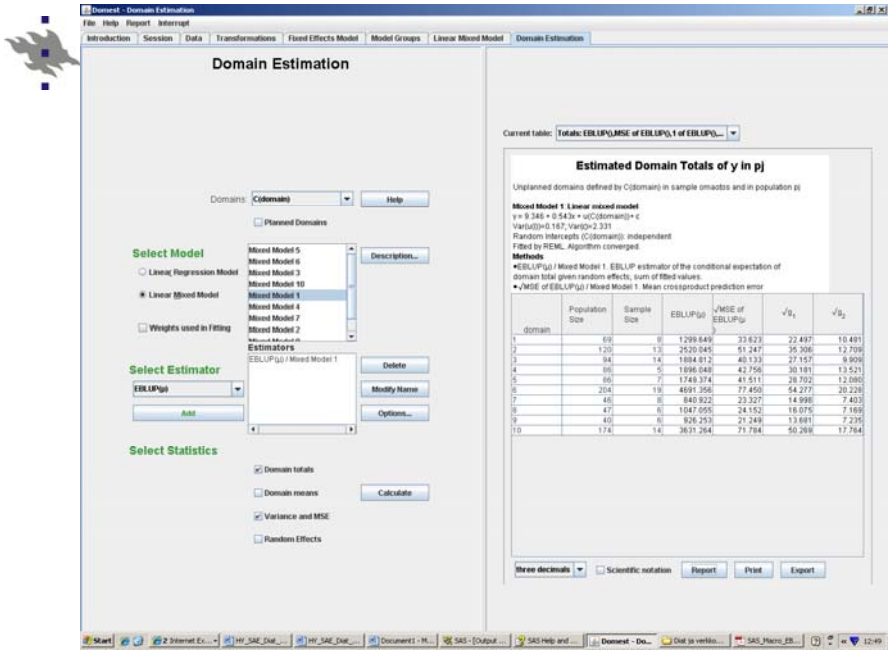
$d = 1,...,D$

Domain type and estimator type

Domain type	Estimator type	
	Direct	Indirect
Planned	Typical set-up	More rarely
Unplanned	More rarely	Typical set-up

Ohjelma DOMEST

- Stand-alone Java program for estimation of totals and means for domains and small areas developed by Dr. Ari Veijanen and Prof. Risto Lehtonen
- Design-based estimators
  - HT and Hájek estimator
  - GREG with linear model
  - GREG with linear mixed model
- Model-based estimators
  - EBLUP (Empirical best linear unbiased predictor) with linear mixed model



R-kieliset ohjelmat

- R-kielisiä ohjelmia on tekeillä eri EU-projekteissa (FP7)
  - SAMPLE project (EU FP7)
  - AMELI project (EU FP7)
- Demoissa käytetään SAS-ohjelmiston proseduureja
  - SURVEYSELECT
  - SURVEYMEANS
  - SURVEYREG

Kuinka jatkaa eteenpäin?

- Yhteiskuntatilastotiede, kevät 2011
  - Pienalue-estimointi
  - Tilastolliset tietosuojamenetelmät
  - Imputointimenetelmät
- Pro gradu yhteiskuntatilastotieteen alalta
  - Tilastokeskus
  - Kelan tutkimusosasto
  - Muu valtionhallinnon tutkimuslaitos tai virasto