

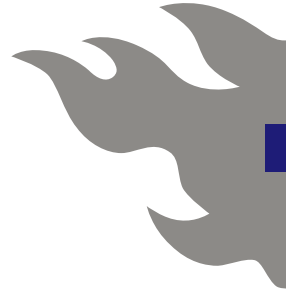


HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

# Otantamenetelmät (78143) Syksy 2009

## TEEMAT 3 & 4

Risto Lehtonen  
[risto.lehtonen@helsinki.fi](mailto: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 rypäeseen 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:  $deff(\hat{\rho}) = 1 + (\bar{m} - 1)\hat{\rho}_{\text{int}}$

missä  $deff(\hat{\rho})$  on asetelmakerroin  
(*design effect*)

## Esimerkkiaineisto: Työterveyshuoltotutkimus OHC

### ■ Otanta-asetelma

- Ositettu yksi- ja kaksiasteinen ryväotanta
- 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



# Ohjelma DOMEST

- Stand-alone Java program for estimation of totals and means for domains and small areas
- Developed by Dr. Ari Veijanen (Statistics Finland and University of Helsinki)
- Estimators
  - HT and Hájek estimator
  - GREG and SYN with linear model
  - GREG and SYN with linear mixed model
  - EBLUP with linear mixed model

Risto Lehtonen

15

The screenshot shows the 'Domain Estimation' window of the DOMEST software. The interface includes a menu bar (File, Help, Report, Interrupt), a toolbar, and several panels for configuration and results.

**Domain Estimation Panel:**

- Domain: C(domain)
- Select Model: Linear Mixed Model (selected)
- Select Estimator: EBLUP(y)
- Select Statistics: Domain totals, Variance and MSE (checked)

**Results Panel:**

Current table: Totals: EBLUP(y), MSE of EBLUP(y), 1 of EBLUP(y), ...

**Estimated Domain Totals of y in pj**

Unplanned domains defined by C(domain) in sample omacbs and in population sj

**Mixed Model 1. Linear mixed model**

$$y = 0.348 + 0.543x + u(C(domain)) + e$$

Variance(s): 0.167, var(p): 0.331  
Random intercepts (C(domain)): independent  
Fitted by REML. Algorithm converged.

**Methods**

- EBLUP(y) / Mixed Model 1. EBLUP estimator of the conditional expectation of domain total given random effects, sum of fitted values.
- MSE of EBLUP(y) / Mixed Model 1. Mean crossproduct prediction error

domain	Population Size	Sample Size	EBLUP(y)	VMSE of EBLUP(y)	$\sqrt{v_{y_i}}$	$\sqrt{v_{y_j}}$
1	60	6	1299.849	33.623	22.487	10.481
2	120	13	2520.045	51.247	35.206	12.709
3	84	14	1684.812	46.133	27.157	9.809
4	60	5	1696.648	42.756	30.181	13.521
5	66	7	1749.374	41.511	28.702	12.086
6	204	19	4691.956	77.450	54.277	20.228
7	48	6	840.822	23.327	14.998	7.403
8	47	6	1047.955	24.152	16.075	7.169
9	40	6	626.253	21.149	13.661	7.235
10	174	14	3631.264	71.704	50.289	17.764

Three decimals | Scientific notation | Report | Print | Export

Risto Lehtonen

16





## 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 2010
  - [Otanta-aineistojen analyysi](#)
  - [Imputointimenetelmät](#)
  - [Painotusmenetelmät surveyssä](#)
  - [Tilastolliset tietosuojamenetelmät](#)
- Pro gradu yhteiskuntatilastotieteen alalta
  - Tilastokeskus
  - Kelan tutkimusosasto
  - Muu valtionhallinnon tutkimuslaitos tai virasto