HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Sosiaalitutkimuksen tilastolliset menetelmät Osa 3 - Diat 2 Otanta-asetelmat ja tilastollinen analyysi

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

Otanta-aineiston tilastollinen analyysi

- Lähtökohta: Analysoitavana otosperusteinen henkilöaineisto, jossa **on hierarkkinen (monitasoinen) rakenne**
- Tutkimusasetelmia/Otanta-asetelmia, jotka tuottavat aineistoon hierarkkisia rakenteita
  - **Moniasteinen ryväsotanta-asetelma**
  - **Pitkittäisasetelma/Paneeliasetelma**
- HUOM: Hierarkkinen rakenne tuottaa aineistoon *havaintojen keskinäistä korreloituneisuutta*
- HUOM: **Havaintojen** keskinäinen korrelaatio on eri asia kuin **muuttujien** välinen korrelaatio

Otanta-aineiston tilastollinen analyysi

- HUOM:
- Usein tilastollisten analyysimenetelmien opetuksessa oletetaan, että havainnot ovat toisistaan riippumattomia (korreloimattomia)
- Oletus vastaa sitä, että otos on poimittu alkiotason perusjoukosta yksinkertaisella satunnaisotannalla (SRS, *simple random sampling*)
- Mutkikkaammissa otantatilanteissa, kuten ryväsotannassa, tämä oletus ei päde

1. Alkiotasoinen otanta
Element sampling

- Kohdeperusjoukko: Alkiotasoinen
- Kehikkoperusjoukko: Alkiotasoinen
- Otantayksikkönä perusjoukon alkio
- Alkiotason otos poimitaan valitulla otantamenetelmällä suoraan kehikkoperusjoukosta
  - Esim. Henkilöotos väestörekisteristä
  - Yritysotos yritysrekisteristä

2. Yksi- ja kaksiasteinen ryväsotanta
One-stage / Two-stage cluster sampling

- **Yksiasteinen** ryväsotanta

1. aste: Rypäiden poiminta ryvästason perusjoukosta
Otantayksikkö: Perusjoukon alkioiden muodostama luonnollinen ryhmä eli **ryväs** (*cluster*)

- Alkiotason otokseen otetaan kaikki otosrypäiden alkiot

- **Kaksiasteinen** ryväsotanta

1. aste: Rypäiden poiminta ryvästason perusjoukosta
2. aste: Alkioiden poiminta otokseen tulleista rypäistä

- Alkiotason otokseen otetaan otosrypäistä poimitut otosalkiot

a. Poikkileikkausasetelma
Cross-sectional design

- Tiedonkeruu
- Ajallinen poikkileikkaus
- Tutkimusasetelmasta johtuva havaintoyksiköiden korreloituneisuus
  - Onko? -Ei ole
- Otanta-asetelmasta johtuva havaintojen korreloituneisuus
  - Onko?
- Riippuu otanta-asetelmasta!
  - Ryväsotanta: Kyllä
  - Alkiotasoinen otanta: Ei ole

b. Pitkittäisasetelma / Paneeliasetelma
Longitudinal / Panel design

- Paneelitutkimus, toistomittaus, seurantatutkimus, rotaatiopaneeli
  - Samoja yksikköjä koskeva ajassa toistuva tai jatkuva tiedonkeruu
- Tutkimusasetelmasta johtuva havaintojen korreloituneisuus – Onko?
  - Toistomittauksesta johtuva positiivinen autokorrelaatio
- Otanta-asetelmasta johtuva havaintojen korreloituneisuus – Onko?
  - Riippuu jälleen otanta-asetelmasta!

Havaintojen korreloituneisuuden lähteitä: Tutkimusasetelma ja otanta-asetelma

Otanta-asetelma	Tutkimusasetelma	
	a. Poikkileikkaus-asetelma	**b. Pitkittäisasetelma**
1. Alkiotason otanta	1a. Ei havaintojen korreloituneisuutta	1b. Positiivinen autokorrelaatio
2. Ryväsotanta	2a. Positiivinen rypäänsisäinen korrelaatio	2b. Ristikkäinen autokorrelaatio ja ryväskorrelaatio

Otanta-aineiston tilastollinen analyysi

- Käsitellään tilastollisia menetelmiä, joilla tutkimusasetelman / otanta-asetelman **hierarkkinen rakenne** ja siitä seuraava **havaintojen positiivinen korreloituneisuus** voidaan ottaa huomioon tilastollisen analyysin yhteydessä
- Miksi tämä on tärkeää?
- Tilastollisen päättelyn pätevyyden takia

Ryväsotanta (*Cluster sampling*)

- Otantayksikkönä on perusjoukon alkioiden muodostama luonnollinen ryhmä eli **ryväs** (*cluster*)
  - Toimipaikka (OHC Survey)
  - Kunta, terveyskeskuspiiri: Terveys 2000
  - Koulu, opetusryhmä: PISA

- **Esimerkkejä tyypillisistä ryväsyksiköistä omalta toiminta-alueeltasi?**

Esimerkkejä hierarkkisista ryväsotanta-aineistoista

- Kelan työterveyshuoltotutkimus (demoaineisto)
  - Rypäinä toimipaikat
  - Analysoidaan työntekijäaineistoa
- PISA-tutkimussarja
  - Rypäinä koulut/koululuokat
  - Analysoidaan oppilasaineistoa
- Terveys 2000 ja Terveys 2010 –tutkimukset
- Osin myös ESS (useissa maissa)
  - Alueelliset rypäät
  - Analysoidaan henkilöaineistoa

Esimerkki: OHC-aineisto

- **Kelan työterveyshuoltotutkimus**
  Occupational Health Care Survey
- Otanta-asetelma
  - Ositettu yksi- ja kaksiasteinen ryväsotanta
  - Toimipaikat rypäinä
  - Ositus rypään koon ja toimialan mukaan
    - Pienet toimipaikat: Yksiasteinen otanta
    - Suuret toimipaikat: Kaksiasteinen otanta
  - Henkilötasolla itsepainottuva (*self-weighting*) otos
    - Henkilötason analyysipainot = 1 kaikille

OHC Survey: Demonstraatioaineisto

- Rajaus pedagogista käyttöä varten
  - Toimipaikat, joissa vähintään 10 työntekijää
  - *H* = 5 ositetta (*strata*)
  - *m* = 250 toimipaikkaa (ryvästä, *clusters*)
  - *n* = 7841 henkilöä
  - 10 muuttujaa
  - Vaihteleva määrä otosrypäitä per osite
- Aineisto on saatavilla linkistä

VLISS-Virtual laboratory in survey sampling

OHC Survey: Aineiston muuttujat

Variables in Creation Order

#	Variable	Type	Len	Label
1	OSITE	Num	8	Stratum identifier
2	RYVAS	Num	8	Cluster identifier
3	ID	Num	8	Element identifier
4	SEX	Num	8	Gender
5	AGE	Num	8	Age in years
6	AGE2	Num	8	Age under/over 45
7	PHYS	Num	8	Physical health hazards of work
8	CHRON	Num	8	Chronic morbidity
9	PSYCH	Num	8	Psychic strain - 1st princomp
10	PSYCH2	Num	8	Psychic strain - dichotomy

Vaatimuksia analyysityökaluille OHC-data

- Aineiston hierarkkinen rakenne
  - Kaksiasteinen ositettu ryväsotanta
- **Rypäiden positiivinen sisäkorrelaatio**
  - Havainnot pareittain korreloituneita rypäiden (toimipaikkojen) sisällä
  - Otettava huomioon analyysissa
- Sisäkorrelaation tunnusluvut
  - **Asetelmakerroin** deff (*design effect)*
  - **Sisäkorrelaatio** *(intra-cluster correlation)*

Asetelmakerroin *Deff*

Asetelmakerroin (*Design effect*, *deff*) mittaa otanta-asetelman ryvästymisen vaikutusta estimaattorin varianssiin

Esimerkiksi **osuustunnusluvun** (suhteellisen osuuden) $\hat{p}$ estimoitu asetelmakerroin on:

$$deff(\hat{p}) = \frac{v_{clu}(\hat{p})}{v_{srs}(\hat{p})} = \frac{v_{clu}(\hat{p})}{\hat{p}(1-\hat{p})/n}$$

missä
$\hat{p}$ on estimoitu osuustunnusluku

v_{clu} on ryväsotanta-asetelman mukainen otosvarianssi

v_{srs} on yksinkertaiseen satunnaisotantaan perustuva otosvarianssi (tässä binominen varianssilauseke)

Mitä asetelmakertoimesta voi päätellä?

- *deff* < 1
  - Käytetty otanta-asetelma on **tehokkaampi** kuin yksinkertainen satunnaisotanta (SRS)
  - Otanta-asetelma on optimoitu tutkittavaa ilmiötä varten
    - Esim. Tilastokeskuksen työvoimatutkimus
  - Otanta-asetelmassa ja/tai estimointiasetelmassa on käytetty tehokkaasti lisäinformaatiota
    - PPS-otanta
    - Malliavusteinen estimointi
    - Käsitellään otantamenetelmien kurssilla s/2012

Mitä asetelmakertoimesta voi päätellä?

- *deff* = 1
  - Käytetty otanta-asetelma on **yhtä tehokas** kuin SRS
- *deff* > 1
  - Käytetty otanta-asetelma on **tehottomampi** kuin SRS
  - Tyypillistä **ryväsotanta-aineistoille**
  - Esim. OHC-aineisto, PISA, Terveys2000…

- HUOM: Otanta-asetelma on sitä tehokkaampi mitä pienempi on estimaattorin (tunnusluvun) varianssiestimaatti ja keskivirhe
  - Keskivirhe = estimaattorin varianssin neliöjuuri

OHC-aineisto: *Deff*-estimaatit (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

Rypäiden positiivisen sisäkorrelaation vaikutukset analyysin kannalta

- Vastaavankokoiseen alkiotasoiseen otanta-aineistoon verrattuna ryväsotanta-aineistossa:
  - Tehokas otoskoko pienenee
  - Tunnuslukujen keskivirheet kasvavat
  - Luottamusvälit (virhemarginaalit) suurenevat
  - Testisuureiden tilastollinen merkitsevyys heikkenee

Asetelmakerroin, sisäkorrelaatio ja tehokas otoskoko

Asetelmakerroin ja sisäkorrelaatio

$$\hat{\rho}_{\text{int}} = \frac{deff(\hat{p}) - 1}{\bar{n} - 1}$$

Tehokas otoskoko *(effective sample size)*:

$$n_{eff} = \frac{n}{deff(\hat{p})} = \frac{n}{1 + (\bar{n} - 1)\hat{\rho}_{\text{int}}}$$

missä
n on alkiotason otoskoko
$\bar{n}$ on rypäiden keskimääräinen otoskoko

Tehokas otoskoko ja sisäkorrelaatio OHC-aineistossa

- **Fysikaaliset työolot, voimakas positiivinen sisäkorrelaatio**

- Asetelmakerroin *deff* = 6.5
- Sisäkorrelaatio *rho* = 0.181

- Otoskoko *n* = 7841 henkilöä
- Tehokas otoskoko

  n(eff) = 7841/6.5 = 1206 henkilöä

Tehokas otoskoko ja sisäkorrelaatio OHC-aineisto

- **Psyykkiset oireet, heikko positiivinen sisäkorrelaatio**

- Asetelmakerroin *deff* = 1.8
- Sisäkorrelaatio *rho* = 0.026

- Otoskoko *n* = 7841 henkilöä
- Tehokas otoskoko

  n(eff) = 7841/1.8 = 4356 henkilöä

OHC Survey – Tilastollinen analyysi

- **Asetelmaperusteinen** (*Design-based*) tilastollinen analyysi
- (Lehtonen – Pahkinen 2004)
- **Ohjelmistot**
- SAS: SURVEY-proseduurit
  - SURVEYMEANS, SURVEYFREQ
  - SURVEYREG, SURVEYLOGISTIC
- SPSS
  - Complex Samples –moduli
- Stata
  - SVY-proseduurit

Vaihtoehto: Malliperusteinen (*model-based*) analyysi

- Rypäiden sisäkorreloituneisuuteen reagoidaan mallintamalla

- Tilastolliset sekamallit (*Mixed models*)
- Monitasomallit (*Multilevel models*)
- Hierarkkiset mallit (*Hierarchical models*)
- (Kaikki nämä termit viittaavat samaan laajaan yleistettyjen lineaaristen sekamallien perheeseen)

- Ohjelmistot: SAS, SPSS, R, ym.
- Laaja kirjallisuus

ESIMERKKI: Logistinen kovarianssianalyysi Lehtonen&Pahkinen (2004) Example 8.2

- Asetelmaperusteinen logistinen ANCOVA
- Binäärinen tulosmuuttuja:

  PSYCH2 Psyykkinen rasittuneisuus

  0: Lievä

  1: Vakava
- Luokiteltu selittäjä
  - Sukupuoli SEX (M/F)
- Jatkuva selittäjä
  - Ikä AGE (vuosina)
- Binääriset selittäjät
  - Työn fysikaaliset haitat: PHYS (0/1)
  - Pitkäaikaissairastavuus: CHRON (0/1)

Tilastollinen malli

- Logistinen ANCOVA-malli

  logit(P) = INTERCEPT + SEX + AGE + PHYS + CHRON + SEX*AGE + SEX*PHYS + SEX*CHRON

missä logit(P) = log(P/(1-P)) on “vedonlyöntisuhteen” (Odds Ratio) logaritmi
OR: “ristitulosuhde”, “suhteellinen riski”

Mallinnetaan todennäköisyyttä:

P = Prob(Psych2 = 1 | X)

Todennäköisyys kuulua vakavamman psyykkisen rasittuneisuuden luokkaan

Lopullinen (redusoitu) malli

- Eksploratiivinen analyysi tuotti lopulliseksi malliksi:

logit(P) = INTERCEPT + SEX + AGE
+ PHYS + CHRON + SEX*AGE

missä SEX, AGE, PHYS ja CHRON
ovat mallin **päävaikutustermejä**

SEX*AGE on **yhdysvaikutustermi**

SAS Procedure SURVEYLOGISTIC

```
proc surveylogistic data=ohc;
strata osite;
cluster ryvas;
class sex / param=ref;
model psych2(event=last)
      = sex age phys chron
      sex*age
      / link=logit rsquare;
run;
```

Lehtonen & Pahkinen (2004) Table 8.8

Table 8.8 Design-based logistic ANCOVA on overall psychic strain with the PML method.

Model term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value	Odds ratio	95% confidence interval for OR	
							Lower	Upper
Intercept	0.1964	1.56	0.1572	1.25	0.2127	1.22	0.89	1.66
Sex								
Males	−0.9926	1.43	0.2033	−4.88	0.0000	0.37	0.25	0.55
Females*	0	n.a.	0	n.a.	n.a.	1	1	1
Age	−0.0046	1.55	0.0041	−1.12	0.2624	1.00	0.99	1.00
Physical health hazards	0.2765	1.39	0.0596	4.64	0.0000	1.32	1.17	1.48
Chronic morbidity	0.5641	1.17	0.0575	9.82	0.0000	1.76	1.57	1.97
Sex, Age								
Males	0.0131	1.41	0.0051	2.56	0.0111	1.01	1.00	1.02
Females*	0	n.a.	0	n.a.	n.a.	1	1	1

* Reference class; parameter value set to zero.
n.a. not available.

Suhteellinen riski
Odds Ratio OR

- Sukupuoli-ikävakioitu suhteellinen riski *Odds Ratio*, OR (asetelmaperusteinen 95% luottamusväli):

  OR(PHYS) = 1.32 (1.17, 1.48)
  OR(CHRON) = 1.76 (1.57, 1.97)

- Henkilöillä, joilla on pitkäaikainen sairaus, on 1.76 kertainen riski kuulua vakavamman psyykksien rasittuneisuuden luokkaan verrattuna henkilöihin joilla sairautta ei ole

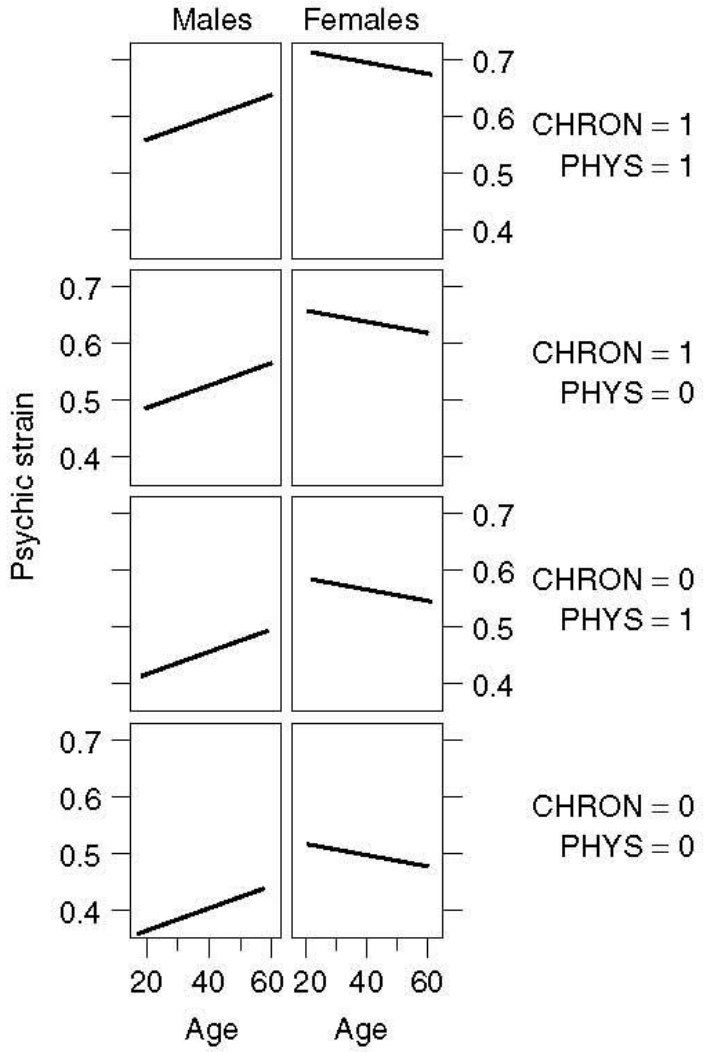

Figure 8.2 Fitted proportions of falling into the high psychic strain group for the final logistic ANCOVA model.

Kirjallisuutta: Analyysivaihe

- Lehtonen, Risto & Pahkinen, Erkki (2004). Practical Methods for Design and Analysis of Complex Surveys John Wiley & Sons.
  - Ladattavissa dawsoneran kautta
  - Chapter 7: Analysis of one-way and two-way tables
  - Chapter 8: Multivariate survey analysis
  - Chapter 9: More detailed case studies

- Tilastokeskus (2007). Laatua tilastoissa. 2. uudistettu painos, Tilastokeskus, Käsikirjoja 43.
  - Ladattavissa kurssin kotisivulta
  - Luku 2.12 *Tilastollinen estimointi ja analyysi*