HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi

(78405)
Kevät 2014
TEEMA 1

Risto Lehtonen
risto.lehtonen@helsinki.fi

Harjoitustyö

- Aineopinnot
  - Vapaaehtoinen mutta suositeltava (2 op)
- Syventävät opinnot
  - Pakollinen (2 op)
- Työn palautus maaliskuun 2014 loppuun mennessä

Teemoja ja näkökulmia

- **Hierarkkinen** (monitasoinen) otanta-aineisto
  - **Ryväsotantaan** perustuva aineisto
  - **Riippuvien havaintojen** tilanne
- Lineaariset ja logistiset mallit otanta-aineistoille
  - Asetelmaperusteinen analyysi
    - Kiinteiden tekijöiden mallit
  - Malliperusteinen analyysi
    - Sekamallit
- Esimerkkejä:
  - Harjoitusaineisto: OHC-aineisto
  - Case Studies: PISA, ESS
- Tilastollinen ohjelmisto
  - SAS, SPSS, R

Kirjallisuutta

- Lehtonen R. and Pahkinen E. (2004). Practical Methods for Design and Analysis of Complex Surveys. Second Edition. Chichester: John Wiley & Sons. E-kirja: dawsonera
  - Web extension: VLISS-virtual laboratory in survey sampling http://vliss.helsinki.fi/
- Chambers R.L. and Skinner C.J. (Eds.) (2004). Analysis of Survey Data. Chichester: Wiley.
- Goldstein H. (2011). Multilevel Statistical Models, 4th Edition. London: Arnold.
- Heeringa S.G., West B.T. and Berglund P.A. (2010). Applied Survey Data Analysis. Chapman and Hall/CRC.
- SAS/STAT 9.2 - Introduction to Survey Sampling and Analysis Procedures.
- Lumley T. (2010) . Complex Surveys: A Guide to Analysis Using R. Wiley.

TEEMA 1
JOHDANTO

Empiirinen kvantitatiivinen tutkimusprosessi - Otosperusteinen

Survey = Empiiris-kvantitatiivinen (yhteiskunta)tutkimus

- **Survey-projektin vaiheet:**

I Suunnittelu ja testaus

1. Tutkimusongelman muotoilu
2. Tutkimusasetelman laadinta
3. Otanta-asetelman laadinta
4. Tiedonkeruuvälineiden valmistus
5. Testaus laboratorio-oloissa ja pilotointi kentällä

II Tiedonkeruuoperaatiot

6. Otoksen poiminta
7. Tiedonkeruu
8. Tiedostonmuodostus
   - editointi, imputointi
   - katoanalyysi
   - painokertoimien muodostus

III Tilastollinen analyysi

9. Eksplorointi ja kuvailu
   - tunnuslukujen laskenta,
   - taulukointi
   - graafiset kuvailut
   - piste-estimointi
   - väliestimointi
10. Analyysi ja tulkinta
    - tilastollinen mallinnus

IV Raportointi ja jälkihoito

11. Julkaisut ja artikkelit
12. Opinnäytetyöt
13. Esitelmät
14. Sähköiset tuotteet
15. Dokumentointi ja arkistointi

Survey-asetelma

- Tutkimusasetelma (*Study design*)
  - Poikkileikkaustutkimus (*Cross-sectional*)
  - Pitkittäistutkimus (*Longitudinal*)
- Otanta-asetelma (*Sampling design*)
  - Ositus (*Stratification*)
  - Ryvästyminen (*Clustering*)
  - Painotus (*Weighting*)
- Analyysiasetelma (*Analysis design*)
  - Joukko analyysimenetelmiä joiden avulla voidaan reagoida tutkimusasetelman ja otanta-asetelman ominaisuuksiin

Kuvaileva ja analyyttinen survey

YHTEENVETO: KUVAILEVA JA ANALYYTTINEN SURVEY

	KUVAILEVA	ANALYYTTINEN
Tulosmuuttujat	Muutamia	Useita
Yleistystaso	Kiinteä perusjoukko	"Superpopulaatio"
Estimoitavat parametrit	Kuvailevia, esim. totaalit, keskiarvot	Analyyttisia, esim. regressiokertoimet
Estimaattori-tyypit	Lineaarisia, esim. totaalin HT-estimaattori	Epälineaarisia, esim. regressiokertoimen PNS-estimaattori
Varianssien estimointi	Analyyttisesti	Approksimatiivisesti
Ulkoisen lisäinfon käyttö analyysissa	Tärkeää	Vähemmän tärkeää
Malliavusteinen estimointi	Käytetään paljon	Ei juurikaan käytetä
Monimuuttuja-analyysi	Ei käytetä	Käytetään paljon
Tilastollinen testaus	Ei käytetä	Käytetään paljon
Painojen skaalaus	Perusjoukon taso (N)	Otostaso (n)
Tilastolliset ohjelmistot	SAS, GES, CLAN, SUDAAN	SAS, SPSS, SUDAAN, WesVar, Stata, MLwiN

Otanta-asetelma *sampling design*

- Niiden sääntöjen ja menetelmien kokonaisuus, jolla **otos** poimitaan määritellystä **perusjoukosta**
  - Tavoiteperusjoukko
  - Kohdeperusjoukko
  - Kehikkoperusjoukko
    - Ylipeitto
    - Alipeitto

Otanta-asetelma: Sisältymistodennäköisyys

- *N* alkion perusjoukko
- Jokaisella perusjoukon alkiolla *k* on tunnettu, nollaa suurempi todennäköisyys π_k tulla mukaan *n* alkion otokseen

$0 < \pi_k \leq 1$

perusjoukon alkiolle *k*,

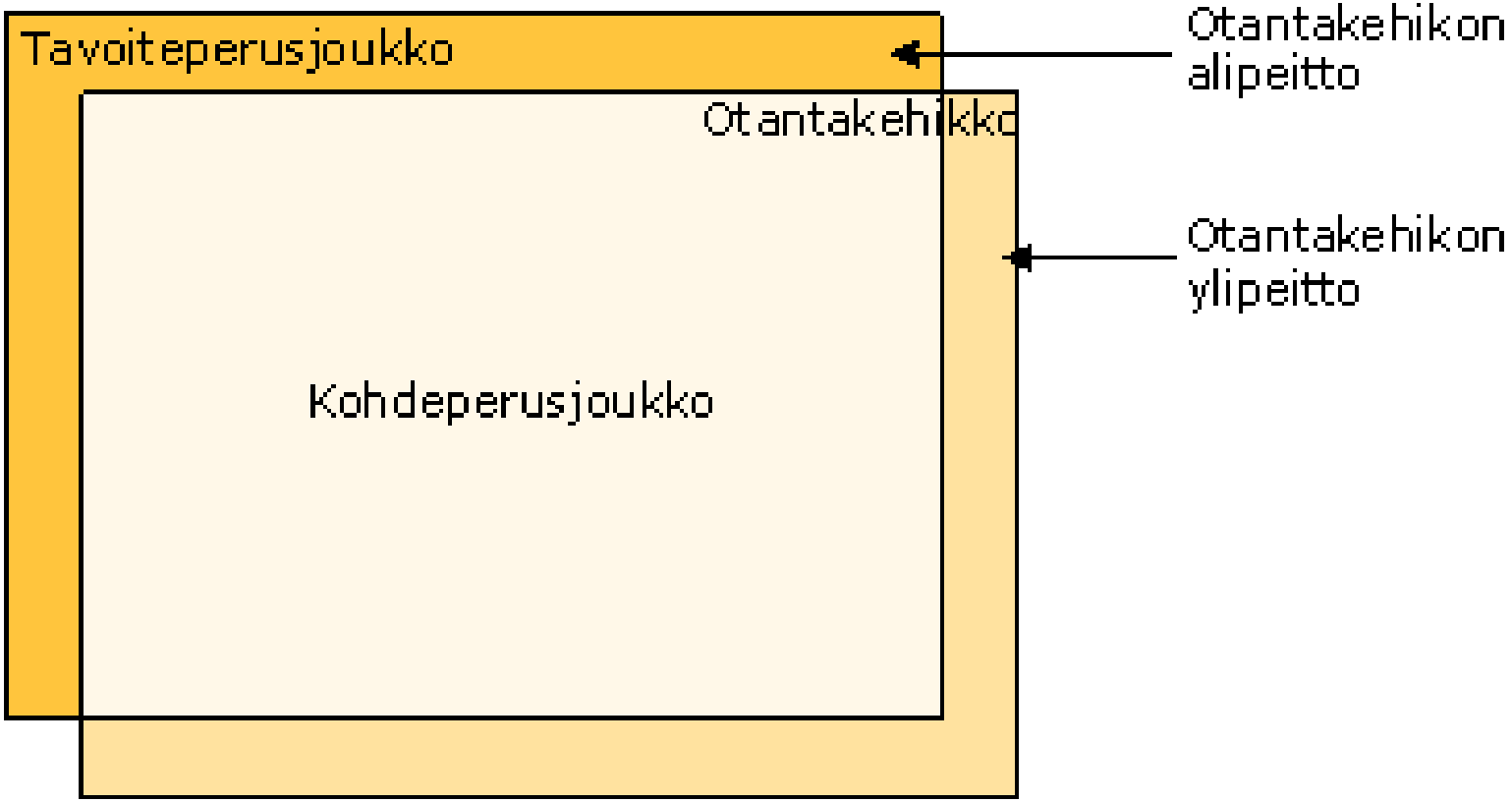
$k = 1,...,N$

missä *N* on perusjoukon

alkioiden lukumäärä

Otantakehikon alipeitto ja ylipeitto

Tilastokeskus: Laatua tilastoissa -käsikirja



Otos *Sample*

- Perusjoukon osajoukko
- Poimitaan jollain satunnaisotannan menetelmällä (*Random sampling, Probability sampling*)
- Poiminnassa käytetään sisältymis-todennäköisyyksiä (*Inclusion probability*)

- Miksi satunnaisotanta?
  - Otoksesta saatavat tulokset voidaan yleistää koskemaan koko kiinnostuksen kohteena olevaa perusjoukkoa tai hypoteettista mallia
  - Tilastollinen päättely
    - Piste-estimaatit
    - Keskivirheet
    - Luottamusvälit
    - Tilastollinen testaus

Huomioita sisältymistodennäköisyydestä

- Nollaa suurempi
- Voi olla = 1
  - Milloin?
- Voi olla yhtäsuuri kaikille alkioille
- Voi vaihdella
  - Alkioryhmittäin
    - Ositettu otanta
  - Alkioittain
    - PPS-otanta (otanta alkion kokoon suhteutetuin todennäköisyyksin)

- Sis.todennäköisyyttä käytetään painokertoimien muodostamisessa
- **Asetelmapaino** (*design weight*)
  - Totaalien estimointi
- **Analyysipaino** (*analysis weight*)
  - Muut analyysitilanteet
- **Uudelleenpainotus**
  - Vastauskadon korjausta varten
    - Voidaan soveltaa sekä asetelmapainoon että analyysipainoon

Asetelmapaino (*design weight*)

Asetelmapaino: $w_k = 1/\pi_k$ otosalkiolle k,

$k = 1,\ldots,n$, missä π_k on alkion k sisältymis-

todennäköisyys ja n on otoskoko

Asetelmapainolle pätee $\sum_{k=1}^{n} w_k = N$,

missä N on perusjoukon alkioiden lkm

Asetelmapainoja tarvitaan kun estimoidaan

kokonaismääriä (esim. työttömien kokonaismäärä)

Analyysipaino (*analysis weight*)

Uudelleenskaalattu painokerroin, esim.

$$w_k^* = (n/N)w_k, \quad k = 1,...,n,$$

missä n on otoskoko ja N on perusjoukon koko

Analyysipainoille pätee $\sum_{k=1}^{n} w_k^* = n$ (otoskoko)

joten analyysipainojen keskiarvo = 1

Analyysipainoja käytetään yleensä tilastollisen analyysin yhteydessä

HUOM: SRS-otokselle analyysipaino = 1

Uudelleenpainotus (*Reweighting*)

- Asetelma- ja analyysipainojen lisäksi tarvitaan usein painojen muokkausta vastauskadon (*nonresponse*) vaikutusten oikaisemiseksi
  - Uudelleenpainotus
  - Estimoidaan ensin vastaustodennäköisyys (*response probability*)
    - Aineiston osajoukoissa tai
    - Alkioittain
  - Korjataan analyysipainoja estimoitujen vastaustodennäköisyyksien avulla
  - Esimerkiksi: PISA-tutkimus, Terveys 2000,…

Esimerkki: Health 2000 – Weighting procedures

Design weight $w_{hik} = 1/\pi_{hik}$ where π_{hik} denotes the inclusion probability of person *k* in cluster *i* of stratum *h* in the population.

WARNING: The sum of the sampling weights over the sample data set is equal to the size of the population *N*. That weight should not be used as a weight variable in the analysis!

Analysis weight $w^*_{hik} = \frac{n}{N} \times \frac{1}{\pi_{hik}\hat{\theta}_{hik}}$ where $\hat{\theta}_{hik}$ denotes the

estimated response probability of sample person *k* in cluster *i* of stratum *h* .

NOTE: The sum of analysis weights over the sample data set is equal to the size *n* of the sample data set. Can be used in the analysis.

Otanta-asetelma voi olla…

- Yksinkertainen
  - Systemaattinen otanta
    - Poiminta suoraan alkiotason kehikkoperusjoukosta
  - Ositettu systemaattinen otanta
    - Alkiotason perusjoukon ositus ja otoksen kiintiöinti ositteisiin
    - Systemaattinen otanta kustakin ositteesta
- Mutkikas (*Complex survey*)
  - Ositettu kaksiasteinen otanta
  - Esimerkiksi:
    - Rypäiden poiminta ryvästason perusjoukosta PPS-otannalla
    - Alkioiden poiminta otosrypäistä systemaattisella otannalla

Ryväsotannan motivaatio

- Tiedonkeruumenetelmän kannalta voi olla edullista käyttää ryväsotantaa
  - Käyntihaastattelut
  - Rypäänä kotitalous
  - Kliiniset menetelmät
  - Rypäänä terveyskeskus
  - Koulututkimukset
  - Rypäänä koulu
  - Pisa
- Alkiotason kehikkoperusjoukon puuttuminen voi edellyttää ryväsotantaa
  - Kotitaloustutkimukset useissa Euroopan maissa
- Tutkimusasetelma voi edellyttää ryväsotantaa
  - Työterveyshuoltotutkimus (OHC)
  - Terveys 2010 -tutkimus

Otanta-aineiston analyysi

- Otanta-asetelmaan reagointi on välttämätöntä tilastollisen analyysin yhteydessä
- Miksi?
- Pätevän tilastollisen päättelyn suorittamiseksi
- Tavoitteena on yleistää otosaineistosta saatavat tulokset koskemaan koko perusjoukkoa, josta otos on poimittu
  - Tilastollinen estimointi
  - Tilastollinen testaus
  - Tilastollinen mallinnus

Mitä tietoja aineistossa tulee olla pätevää analyysia varten?

- Otanta-asetelman mukaiset muuttujat
  1. **Asetelmapaino**
     Sisältymistodennäköisyyden käänteisluku
  2. **Analyysipaino**
     Skaalattu asetelmapaino, tarvittaessa **katokorjattu**
     Keskiarvo yli aineiston = 1
  3. **Ositeindikaattori**
     Osoittaa, mihin ositteeseen havaintoyksikkö kuuluu
  4. **Ryväsindikaattori**
     Osoittaa, mihin poimintarypääseen havainto kuuluu
- Tarvittaessa myös indikaattorimuuttuja, joka kertoo onko muuttujan tieto **imputoitu** vai ei

ESS – Painotuksen tarve analyysissa

- **Käytännössä kaikkia tietoja 1-4 ei aina välttämättä tarvita tai ei voida käyttää**
  - Tarve vaihtelee maittain ja riippuu asetelman yksinkertaisuudesta / monimutkaisuudesta
  - Käyttömahdollisuus riippuu siitä, mitä tietoja (muuttujia) analyysitiedostossa on!
    - HUOM: Kansallisten tietoarkistojen ja Norjan tietoarkiston sisältöerot (Esim. Suomi)
- ESS-dokumentti:
  Weighting European Social Survey Data
- Kysymyksiä ja vastauksia, esim:
- *Do tables run on the ESS website need to be weighted? - Almost certainly yes.*

Esimerkki: Suomen ja Belgian ESS-datat

- Suomen ESS-aineisto (Norjan tietoarkistossa):
- Yksinkertainen tilanne
  - **Systemaattinen otanta**
  - Sisältymistodennäköisyydet samoja kaikille
  - Ei ositusta eikä ryvästymistä
  - Painomuuttuja (asetelmapaino) = 1 kaikille
- Belgian ESS-aineisto: Mutkikkaampi tilanne
  - **Ositettu kaksiasteinen ryväsotanta**
  - Sisältymistn vaihtelee ositteittain
  - Provinssiperusteinen ositus
  - Alueelliset poimintarypäät
  - Analyysipainot ovat samoja ositteiden sisällä mutta vaihtelevat ositteiden välillä

Tiivistelmä: Otantamenetelmät I

Otantamenetelmä	Poimintatapa
SRS *Simple random sampling* Yksinkertainen satunnaisotanta	Otos poimitaan perusjoukosta satunnaislukujen avulla
SYS *Systematic sampling* Systemaattinen otanta	Otos poimitaan tasavälisesti listasta tai rekisterinä olevasta tietokannasta
STR *Stratified sampling* Ositettu otanta	Perusjoukon alkiot jaetaan ensin homogeenisiin ositteisiin. Kustakin ositteesta poimitaan SRS tai SYS otos

Tiivistelmä: Otantamenetelmät II

Otantamenetelmä	Poimintatapa
CLU *Cluster sampling* Ryväsotanta	Perusjoukon alkiot muodostavat luonnollisia osajoukkoja eli rypäitä
- Yksiasteinen *one-stage*	1)Rypäiden perusjoukosta poimitaan otosrypäät 2) Kaikki otosrypäiden alkiot tulevat alkiotason otokseen
- Kaksiasteinen *two-stage*	1) Rypäiden perusjoukosta poimitaan otosrypäät 2) Otosrypäiden alkioista poimitaan alkiotason otokset SRS:llä tai SYS:llä
PPS *Selection with Probabilities Proportional to Size*	Sisältymistodennäköisyys on suhteessa alkion kokoon

Tiivistelmä: Otantamenetelmät III

	SRS	SYS	STR	CLU	STR-CLU	PPS
Sisältymis-todennäköi-syys(*)	Vakio *n/N*	Vakio *n/N*	Voi vaih-della(**)	Voi vaih-della	Voi vaih-della	Voi vaih-della
Lisä-informaatio	Ei tarvita	Ei tarvita (***)	Osite-indi-kaattori	Ryväs-indi-kaattori	Osite- ja ryväs-indik.	Koko-tieto

(*) Sisältymistodennäköisyys = todennäköisyys sille, että *N* alkion perusjoukkoon kuuluva alkio sisältyy otokseen, jonka koko on *n* alkiota

(**) Sisältymistodennäköisyys voi vaihdella alkioryhmittäin (ositettu otanta) tai alkioittain (PPS-otanta)

(***) SYS: Voidaan käyttää (implisiittinen osittaminen lajittelemalla perusjoukko ennen poimintaa)

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)
  - *iid*-oletus (*independent identically distributed)*
- Oletus vastaa sitä, että otos on poimittu alkiotason perusjoukosta yksinkertaisella satunnaisotannalla
  - SRS, *simple random sampling*
- Mutkikkaammissa otanta-asetelmissa riippumattomuusoletus ei päde
  - Ryväsotanta

1. Alkiotasoinen otanta
Element sampling

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

  - Esim: Suomen ESS-aineisto 2010

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

- Esim: Belgian ESS-aineisto 2010

a. Poikkileikkausasetelma
Cross-sectional design

- Ajallinen poikkileikkaus
- Tutkimusasetelmasta johtuva havaintoyksiköiden korreloituneisuus
  - Onko?
- Otanta-asetelmasta johtuva havaintojen korreloituneisuus
  - Onko?
- Riippuu otanta-asetelmasta!
  - Ryväsotanta: Pohditaan luennolla
  - Alkiotasoinen otanta: Pohditaan luennolla

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

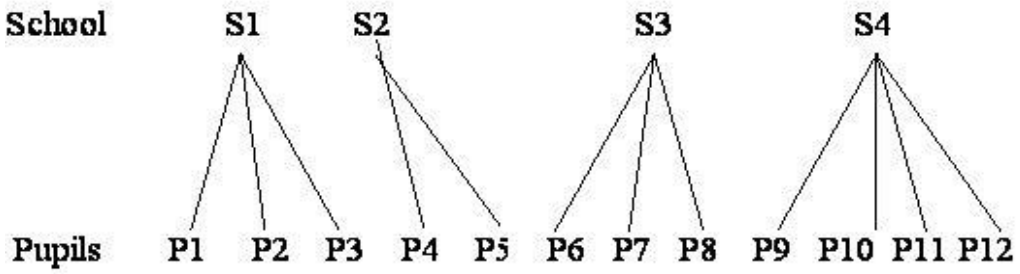
Otanta-asetelma	Tutkimusasetelma	
	a. Poikkileikkaus-asetelma	**b. Pitkittäisasetelma**
1. Alkiotason otanta	1a. Ei havaintojen korreloituneisuutta	1b. Positiivinen autokorrelaatio
2. Ryväsotanta	2a. Positiivinen rypäänsisäinen korrelaatio	2b. Ristikkäinen autokorrelaatio ja ryväskorrelaatio

Clustered data structure

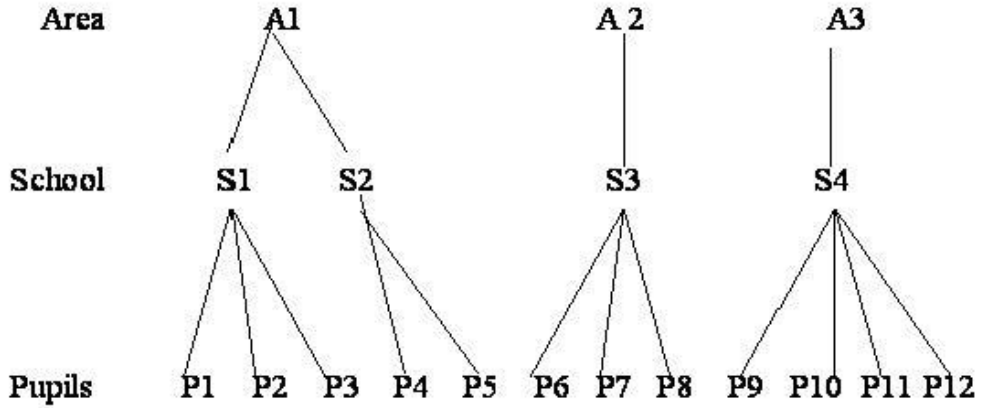
- Stratified multi-stage sampling design
- Hierarchically structured data
  Clustered data, Multilevel data
- **Cluster = a grouping containing *lower level* elements in the population or sample**
- Examples: clustered or multilevel structures
  - Schools – Students
  - Establishments – Staff members
  - Health centers – Patients
  - Neighborhoods – Households – Household members
  - Persons – measurement occasion for a person

Two-level and three-level nested structures

- Two-level nested structure with schools as clusters



- Three-level nested structure clustered by area and school



http://www.bristol.ac.uk/cmm/learning/multilevel-models/data-structures.html

Esimerkkejä hierarkkisesti rakentuneista ryväsotanta-aineistoista

Tutkimus-aineisto	Tutkimus-asetelma	Otanta-asetelma	Ryväs-rakenne	Havainto-yksikkö
Terveys 2000 ja 2010	Poikki-leikkaus	2-asteinen ositettu ryväsotanta	Terveys-keskus-piiri	Henkilö
PISA	Poikki-leikkaus	1-asteinen ositettu ryväsotanta	Koulu tai opetus-ryhmä	Oppilas
SILC	Rotaatio-paneeli	1-asteinen ositettu ryväsotanta	Koti-talous	Koti-talouden jäsen
ESS	Toistettujen poikki-leikkausten sarja	1- ja 2-asteinen ositettu ryväsotanta	Alue	Henkilö

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
- oletetaan, että analyysipainot=1 kaikille

$\hat{p}$ on estimoitu osuustunnusluku

$v_{clu}(\hat{p})$ on ryväsotanta-asetelman mukainen otosvarianssi

$v_{srs}(\hat{p})$ on yksinkertaiseen satunnaisotantaan perustuva otosvarianssi (tässä binominen varianssilauseke)

Design effect *deff*: extended formulas

Design effect, *deff* (Kish 1965) measures the magnitude of the clustering effect to variance (standard error) estimate for $\hat{\theta}$

Estimated overall deff (1):

$$deff(\hat{\theta}) = \frac{\hat{v}_{des}(\hat{\theta})}{\hat{v}_{srs}(\hat{\theta}^*)}$$

where
$\hat{\theta}$ is weighted estimate and $\hat{\theta}^*$ is the corresponding unweighted estimate, both based on the same net sample size *n*

$\hat{v}_{des}(\hat{\theta})$ is based on the actual complex sampling design
$\hat{v}_{srs}(\hat{\theta}^*)$ is the SRS-based variance estimate

Deff (2):

$$deff(\hat{\theta}) = \frac{\hat{v}_{des}(\hat{\theta})}{\hat{v}_{srs}(\hat{\theta})}$$

Mitä asetelmakertoimesta voi päätellä?

- *deff* < 1
  - Käytetty otanta-asetelma on **tehokkaampi** kuin (SRS)
  - Otanta-asetelma on optimoitu tutkittavaa ilmiötä varten
  - Otanta-asetelmassa ja/tai estimointiasetelmassa on käytetty tehokkaasti lisäinformaatiota
    - PPS-otanta
    - Malliavusteinen estimointi (GREG)
  - Esimerkiksi Tilastokeskuksen kuukausittainen työvoimatutkimus

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 varianssiestimaatti (ja keskivirhe)

OHC-data: *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 SAS data OHC

- **Fysikaaliset työolot**
- 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 SAS data OHC

- **Psyykkiset oireet**
- Asetelmakerroin *deff* = 1.8
- Sisäkorrelaatio *rho* = 0.026
- Otoskoko *n* = 7841 henkilöä
- Tehokas otoskoko
  n(eff) = 7841/1.8 = 4356 henkilöä

PISA - Deff ja Eff

Table 2. Descriptive statistics for combined reading literacy score in the PISA 2000 Survey by country (in alphabetical order).

Country	Mean	Standard error	Design effect	Effective sample size of students	Number of observations in data set	
					Students	Schools
Brazil	402.9	3.82	8.33	476	3961	290
Finland	550.7	2.15	2.79	1600	4465	147
Germany	497.4	5.68	13.47	305	4108	183
Hungary	485.7	6.02	20.00	231	4613	184
Republic of Korea	526.6	3.66	12.99	351	4564	144
United Kingdom	531.4	4.08	14.08	564	7935	328
United States	517.0	5.16	6.93	354	2455	112
All	500.0			3881	32101	1388
Data source: OECD PISA database, 2001.						

OHC Survey –Asetelmaperusteinen analyysi

- **Asetelmaperusteisessa** (*Design-based*) analyysissa rypäiden sisäkorrelaatiorakenteet otetaan **häiriötekijöinä** (*nuisance effect*), joiden vaikutus ”puhdistetaan pois” analyysin yhteydessä
  - Lehtonen and Pahkinen (2004), luvut 5, 7 ja 8
- Asetelmaperusteisilla menetelmillä reagoidaan otanta-asetelman ominaisuuksiin:
  - Ositus (*stratification*)
  - Ryvästyminen (*clustering*)
  - Painokertoimet (*weighting*, analyysipainot)

OHC Survey –Asetelmaperusteinen analyysi - Ohjelmistot

- SAS: SURVEY-proseduurit
  - SURVEYMEANS – keskiarvot, kokonaismäärät
  - SURVEYFREQ – taulukointi, testit
  - SURVEYREG – lineaariset mallit
  - SURVEYLOGISTIC – logistiset mallit
  - SURVEYPHREG - elinaikamallit
- SPSS:
  - Complex Samples -moduli
- Stata: SVY-proseduurit
- R-kieliset funktiot ja ohjelmapaketit
  - Lumley T. (2010). Complex Surveys: A Guide to Analysis Using R. Wiley.

OHC Survey – Malliperusteinen analyysi

- **Malliperusteisessa** (*model-based*) analyysissa rypäiden sisäkorreloituneisuuteen reagoidaan **mallintamalla** korrelaatiorakenteita
  - Tilastolliset sekamallit (*Mixed models*)
  - Monitasomallit (*Multilevel models*)
  - Hierarkkiset mallit (*Hierarchical models*)
  - Kaikki nämä termit viittaavat samaan yleistettyjen lineaaristen sekamallien perheeseen
  - *Generalized Linear Mixed Models*, GLMM
- Ohjelmistot: SAS, SPSS, R, ym.
- Laaja kirjallisuus, esim:
  - Demidenko E. (2004) Mixed Models. Wiley.

Mallit ja ohjelmat: SAS-ohjelmisto

- Yleistetyt lineaariset mallit
- Esim:
  - Lineaarinen kiinteiden tekijöiden regressioanalyysi, ANOVA ja ANCOVA
  - Logistinen kiinteiden tekijöiden regressioanalyysi, ANOVA ja ANCOVA
- Yleistetyt lineaariset sekamallit (GLMM)
- Esim:
  - Lineaariset sekamallit
  - Logistiset sekamallit
- SAS – Mallit ja proseduurit