HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi
(78136 , 78405)
Kevät 2010
TEEMA 1

Risto Lehtonen
risto.lehtonen@helsinki.fi

Otanta-aineistojen analyysi

Laajuus
6/8 op.

Tyyppi
78136 Otanta-aineistojen analyysi (aineopintojen valinnainen erikoiskurssi)
78405 Otanta-aineistojen analyysi (syventävien opintojen 1. erikoiskurssi)

Luentoajat
Luennot tiistaisin 26.1.–23.2.2010 klo 14–18 C323 (Exactum), yhteensä 20 tuntia.
Lisäksi harjoituksia mikroluokassa C128 torstaisin 28.1.–4.3.2010 klo 12–15, yhteensä 15 tuntia.
HUOM: To 25.2. ei ole demoja, viimeiset ovat 4.3.

Suoritustapa
Aineopinnot: Loppukuulustelu (6 op) tai loppukuulustelu ja (vapaaehtoinen) harjoitustyö (8 op)
Syventävät opinnot: Loppukuulustelu ja (pakollinen) harjoitustyö (8 op)

Loppukuulustelu
Tiistai 16.3.2010 klo 14–16 C323 (Exactum),

Harjoitustyö

- Aineopinnot
  - Vapaaehtoinen mutta suositeltava (2 op)
- Syventävät opinnot
  - Pakollinen (2 op)
- Työn palautus maaliskuun 2010 loppuun mennessä

Teemoja ja näkökulmia

- Hierarkkinen (monitasoinen) aineisto
- Ryväsotantaan perustuva aineisto
- Lineaariset mallit
- Logistiset mallit
- Sekamallit
- Esimerkkejä: OHC-aineisto
- Case Study: PISA
- Tilastollinen ohjelmisto: SAS, SPSS

Tavoitteet ja sisältö

- Kurssin tavoitteena on perehdyttää opiskelija otanta-aineistojen tilastolliseen analyysiin tilanteissa, joissa aineistossa on **hierarkkinen rakenne**
  - Esimerkiksi aineisto on kerätty jollakin mutkikkaalla tiedonkeruuasetelmalla
- Hierarkkisia rakenteita tuottaa esimerkiksi **moniasteinen otanta-asetelma**, johon sisältyy **ryvästyminen**

Tavoitteet ja sisältö

- Mutkikkaita otanta-asetelmia, jotka tuottavat aineistoon hierarkkisia rakenteita, käytetään laajasti eri tieteenalojen empiirisessä tutkimuksessa
- Hyviä esimerkkejä ovat
  - Kelan työterveyshuoltotutkimus (demoaineisto)
  - PISA-tutkimussarja
  - Terveys 2000 -tutkimus

Tavoitteet ja sisältö

- Kurssilla käsitellään tilastollisia menetelmiä, joilla otanta-asetelman ominaispiirteitä:
  - ositus
  - ryvästyminen
  - painokertoimet

  voidaan ottaa huomioon tilastollisen analyysin yhteydessä

Tilastollinen päättely

- Asetelmaperusteiset (*design-based*) menetelmät
- Malliperusteiset (*model-based*) menetelmät
  - Yleistetyt estimointiyhtälöt (GEE, *generalized estimating equations*)
  - Sekamallit (*mixed models*)
- Tilastolliset testit
  - Asetelmaperusteiset Waldin testisuureet
  - Rao-Scott-korjatut testisuureet
- Tilastollinen mallinnus
  - lineaariset mallit ja logistiset mallit

Tilastollinen ohjelmisto

- SAS, SPSS, Stata, R
- SAS-ohjelmisto
  - SURVEYFREQ,
  - SURVEYREG,
  - SURVEYLOGISTIC
  - GENMOD
  - MIXED
  - GLIMMIX

Kirjallisuutta

- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys. Second Edition*. Chichester: John Wiley & Sons.
- **Web extension:**
  VLISS-Virtual Laboratory in Survey Sampling
  http://mathstat.helsinki.fi/VLISS/

Kirjallisuutta

- 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

TEEMA 1
JOHDANTO

Empiirinen kvantitatiivinen tutkimusprosessi - Survey-prosessi

Survey = Empiiris-kvantitatiivinen (yhteiskunta)tutkimus

- **Survey-hankkeen vaiheet:**
  - **I Suunnittelu ja testaus**
  - **II Tiedonkeruuoperaatiot**
  - **III Tilastollinen analyysi**
  - **IV Raportointi ja jälkihoito**
- **Vaiheet osavaiheineen:**
  - **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**
  - **III Tilastollinen analyysi**
  - **9. Eksplorointi ja kuvailu**
  - **10. Analyysi ja tulkinta**
  - **IV Raportointi ja jälkihoito**
  - **11. Julkaisut ja artikkelit**
  - **12. Opinnäytetyöt**
  - **13. Esitelmät**
  - **14. Sähköiset tuotteet**
  - **15. Dokumentointi ja arkistointi**

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	Fi 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

- 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,

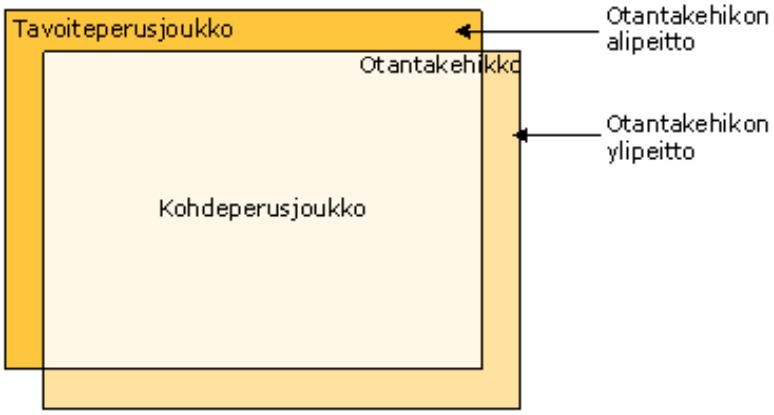
$k = 1,...,N$

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
- 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

Otanta-asetelmaan reagointi

- Aineiston hierarkkisuuteen ja asetelman muihin ominaisuuksiin reagointi tilastollisen analyysin yhteydessä
- Analysoitavassa aineistossa tulee otanta-asetelman mukaisesti olla…
  - **Ryväsindikaattorimuuttuja**
  - **Ositeindikaattorimuuttuja**
  - **Painomuuttujat**
    - Asetelmapaino
    - Analyysipaino

Asetelmapaino (*design weight*)

Asetelmapaino: $w_k = 1/\pi_k$ otosalkiolle k,
$k = 1,...,n$, missä π_k on alkion k sisältymis-
todennäköisyys ja n on otosalkioiden lkm

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 kadon (*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

Sampling 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.

Alkiotason otanta

(1) Alkiotason otanta (*element sampling*)

- Otantayksikkönä on perusjoukon alkio (esim. henkilö).
- Otos poimitaan valitulla otantamenetelmällä suoraan perusjoukon alkioiden muodostamasta kehikkoperusjoukosta
  - Väestörekisteri, toimipaikkarekisteri jne.

Ryväsotanta

(2) Ryväsotanta (*cluster sampling*)

- Otantayksikkönä on perusjoukon alkioiden muodostama luonnollinen ryhmä eli **ryväs** (*cluster*)
- Esim:
  - Toimipaikka (OHC Survey)
  - Kunta, terveyskeskuspiiri
    - Terveys 2000
  - Koulu, opetusryhmä
    - PISA
- **Esimerkkejä ryväsyksiköistä omalta toiminta-alueeltasi?**

Otanta-asetelma voi olla…

- Yksinkertainen
  - Systemaattinen otanta
    - Poiminta suoraan alkiotason kehikkoperusjoukosta
  - Ositettu systemaattinen otanta
    - Alkioiden ositus ja kiintiöinti
    - Systemaattinen otanta kustakin ositteesta
- Mutkikas (*Complex survey*)
  - Ositettu kaksiasteinen otanta
    - 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
- Kehikkoperusjoukon huono saatavuus voi edellyttää ryväsotantaa
  - Koulusaavutus-tutkimukset
  - Pisa
- Tutkimusasetelma voi edellyttää ryväsotantaa
  - Terveys 2000

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 osiin. Kustakin osiosta 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)

Hierarkkinen data

- Kurssilla käsitellään tilastollista analyysia **hierarkkisessa** (**monitasoisessa**) aineistossa
- Kun aineistossa on havaintoyksiköitä ryhmittelevä luonnollinen rakenne eikä havaintoyksiköiden riippumattomuusoletus ole voimassa, on tämä otettava huomioon tilastollisessa mallintamisessa
  - Muuten vaarana on, että analyysitulokset sekä niistä tehtävät sisällölliset tulkinnat ja johtopäätökset ovat virheellisiä

Terminologiaa

Hierarkkisesti rakentunut aineisto

Havaintojen korreloituneisuus

- *Correlated data, Dependent data*
- *Clustered data, Cluster correlated data*

Autokorrelaatio – Aikadimensio

Spatiaalinen korrelaatio – Tiladimensio

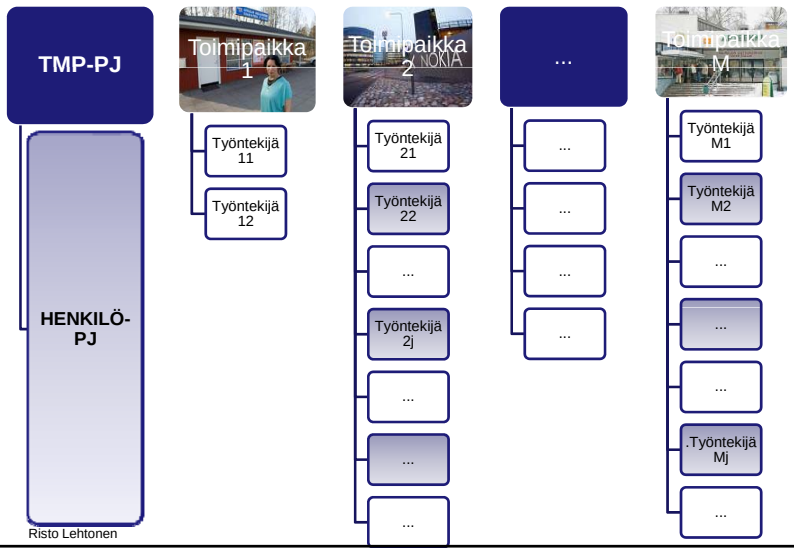
Sisäkorrelaatio – Ryhmän sisäinen korrelaatio

Hierarkkinen malli *Hierarchical model*

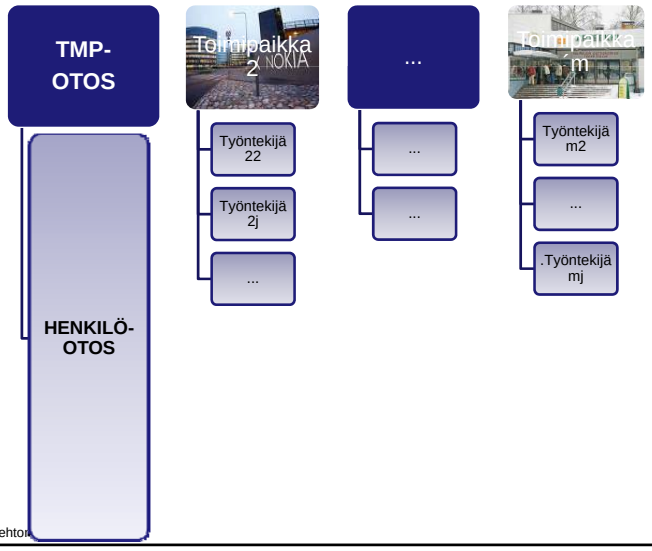
Monitasomalli *Multilevel model*

Sekamalli *Mixed model*

OHC Survey
Kaksitasoinen hierarkkinen rakenne
Kaksiasteinen ryväsotanta: Perusjoukot



OHC Survey
Kaksitasoinen hierarkkinen rakenne
Kaksiasteinen ryväsotanta: Poimittu otos



Havaintojen korreloituneisuuden lähteitä: Tutkimusasetelma ja otanta-asetelma

Otanta-asetelma	Tutkimusasetelma	
	a. Poikkileikkaus-asetelma	b. Pitkittäisasetelma
1. Alkiotason otanta		
2. Ryväsotanta		

Havaintojen korreloituneisuuden lähteitä: Tutkimusasetelma ja otanta-asetelma

Otanta-asetelma	Tutkimusasetelma	
	a. Poikkileikkaus-asetelma	b. Pitkittäisasetelma
1. Alkiotason otanta	1a. Ei havaintojen korreloituneisuutta	
2. Ryväsotanta		

Havaintojen korreloituneisuuden lähteitä: Tutkimusasetelma ja otanta-asetelma

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

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	

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

1. Alkiotasoinen otanta
Element sampling

- Kohdeperusjoukko
  - Alkiotasoinen
- Kehikkoperusjoukko
  - Alkiotasoinen
- Otantayksikkönä perusjoukon alkio
- Otos poimitaan valitulla otantamenetelmällä suoraan kehikkoperusjoukosta
  - Esimerkiksi: Ositettu systemaattinen otanta
- Analyysiyksikkö
  - Alkiotasoinen
  - Esimerkiksi henkilö

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 otosrypäistä
  - Alkiotason otokseen otetaan otosrypäistä poimitut otosalkiot
- **Ryvästyyppejä**
  - Kotitalous, koulu tai opetusryhmä, toimipaikka, terveyskeskus

a. Poikkileikkausasetelma
Cross-sectional design

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

b. Pitkittäisasetelma / Paneeliasetelma
Longitudinal / Panel design

- Paneelitutkimus, toistomittaus, rotaatiopaneeli
  - Samoja yksikköjä koskeva ajassa toistuva tiedonkeruu
- Tutkimusasetelmasta johtuva havaintojen korreloituneisuus
  - Toistomittauksesta johtuva positiivinen autokorrelaatio
- otanta-asetelmasta johtuva havaintojen korreloituneisuus
  - Riippuu otanta-asetelmasta!
  - Alkiotasoinen otanta - ei sisäkorrelaatiota
  - Ryväsotanta - rypäiden sisäkorrelaatio

Esimerkkejä hierarkkisesti rakentuneista ryväsotanta-aineistoista

Tutkimus-aineisto	Tutkimus-asetelma	Otanta-asetelma	Ryväs-rakenne	Havainto-yksikkö
Terveys 2000	Poikki-leikkaus	2-asteinen ositettu ryväsotanta	Terveys-keskus-piiri	Henkilö
PISA	Poikki-leikkaus	1-asteinen ositettu ryväsotanta	Koulu tai opetus-ryhmä	Oppilas
ECHP	Paneeli	1-asteinen ositettu ryväsotanta	Koti-talous	Koti-talouden jäsen
OHC Survey	Poikki-leikkaus	2-asteinen ositettu ryväsotanta	Toimi-paikka	Työn-tekijä

Kelan työterveyshuoltotutkimus Occupational Health Care Survey OHC

- Tutkimusasetelma: Poikkileikkaustutkimus
- 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 otos
- Havaintojen riippuvuus
  - Rypäiden positiivinen sisäkorrelaatio

OHC-data

- Demonstraatioaineisto: SAS-data OHC
- Rajaus
  - Toimipaikat, joissa vähintään 10 työntekijää
  - *H* = 5 ositetta (*strata*)
  - *m* = 250 toimipaikkaa (ryvästä, *clusters*)
  - *n* = 7841 henkilöä
  - 10 muuttujaa
  - Vaihteleva määrä otosrypäitä per osite
- VLISS-Virtual laboratory in survey sampling
- OHC data

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 kuvailu- ja analyysityökaluille - OHC-data

- Aineiston hierarkkinen rakenne
  - Yksi- ja kaksiasteinen ositettu ryväsotanta
- **Rypäiden positiivinen sisäkorrelaatio**
  - Havainnot pareittain korreloituneita rypäiden sisällä
- Tutkimus- ja otanta-asetelma otettava huomioon analyysin yhteydessä
- Korreloituneisuuden 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 (SRS)
  - Otanta-asetelma on optimoitu tutkittavaa ilmiötä varten
  - Otanta-asetelmassa ja/tai estimointiasetelmassa on käytetty tehokkaasti lisäinformaatiota
    - PPS-otanta
    - Malliavusteinen estimointi

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	Number of observations in data set: 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.

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
- Yhteenvetotaulukko