HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otantamenetelmät
(78143)
Syksy 2014

TEEMA 1

Risto Lehtonen
risto.lehtonen@helsinki.fi

Kurssin soveltuvuus

- Kurssi soveltuu tilastotieteen aine- tai syventäviä opintoja suorittaville opiskelijoille ja tilastotieteen sivuaineopiskelijoille sekä myös yliopistoissa, korkeakouluissa ja tutkimuslaitoksissa toimiville jatko-opiskelijoille ja tutkijoille.
  - Kurssi on pakollinen tilastotieteen maisteriopintojen **yhteiskuntatilastotieteen** linjalla
- Kurssi sisältyy valtiotieteellisen tiedekunnan menetelmäkoriin
- Soveltuvia jatkokursseja, kevät 2015
  - Small area estimation (Risto Lehtonen) (periodi 3)
  - Imputation methods (Seppo Laaksonen) (periodi 3)

Tavoitteet

- Kurssilla annetaan yleiskuva **tilastollisista otantamenetelmistä** ja niihin liittyvästä **estimoinnista** sekä menetelmien käytöstä eri tieteenalojen empiirisessä tutkimuksessa
- Kurssin keskeinen teema on **lisäinformaation käyttö**
  1) otannassa ja
  2) estimoinnissa
- Kurssi on luonteeltaan soveltava
- Kurssisivu

Sisältöalueita

- **Otannan perusmenetelmät**
  - Yksinkertainen satunnaisotanta
  - Systemaattinen otanta
- Lisäinformaation käyttö **otanta-asetelmassa**
  - PPS-otanta (*Sampling with probability-proportional-to-size*)
  - Ositettu otanta
- Lisäinformaation käyttö **estimointiasetelmassa**
  - Suhde-estimointi
  - Regressioestimointi
  - Yleistetty regressioestimointi ja kalibrointi

Työkaluja

- Otoksien poiminta eri menetelmillä
  - SAS-proseduuri SURVEYSELECT
  - SPSS-toiminto
    Analyze -> Complex Samples -> Select a Sample
- Tunnuslukujen (piste-estimaatit, varianssit, keskivirheet, asetelmakertoimet) estimointi
  - SAS/SURVEY-proseduurit
    SURVEYMEANS, SURVEYREG
  - SAS-makro CLAN
  - SPSS-toiminto Complex Samples
  - Stata - svy-ohjelmat
  - Survey analysis in R (Thomas Lumley)

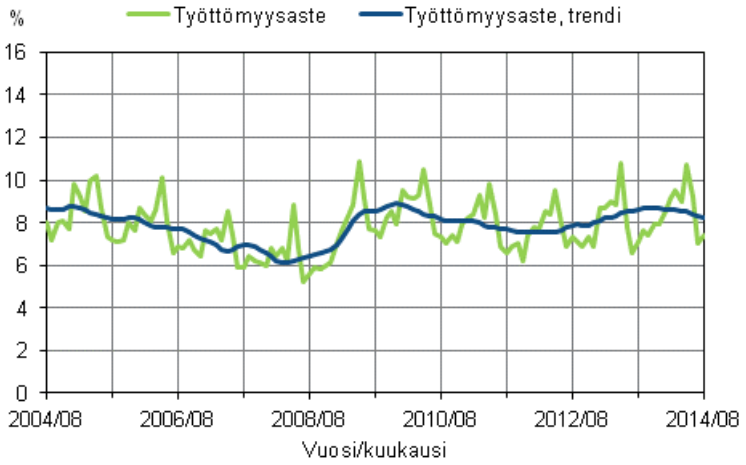
Esimerkkejä

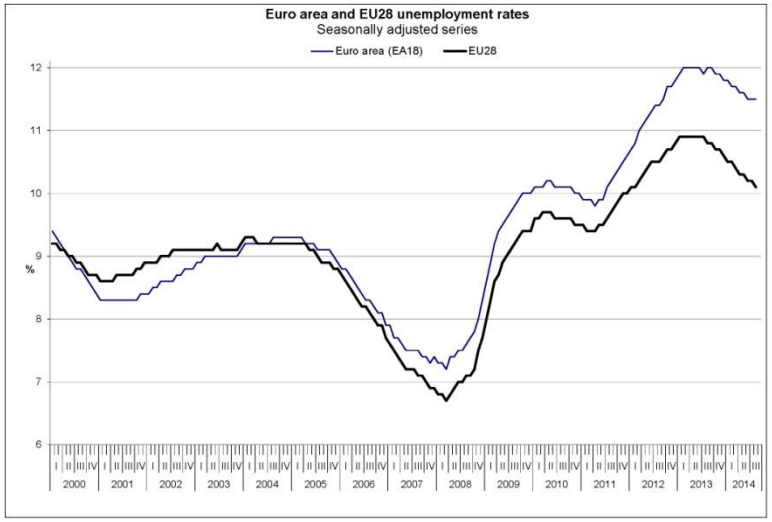
- Esimerkkejä otosperusteisista tiedonkeruista ja tutkimuksista
- OECD:n Pisa-tutkimussarja PISA
- Terveys 2000, kartta Terveys 2011
- European Social Survey ESS
  - Otanta-asetelmat ja data
  - ESS Sampling Guidelines

Tilastokeskuksen työvoimatutkimus

- Tilastokeskuksen työvoimatutkimus
- Laatuseloste

- EU-LFS (Labour Force Survey)





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://vliss.helsinki.fi//
- Laaksonen S. (2013). *Surveymetodiikka*. Bookboon.com.
- Pahkinen E. (2012). *Kyselytutkimusten otanta-menetelmät ja aineistonanalyysi*. Jyväskylä: JULPU.
- Pahkinen E. ja Lehtonen R. (1989). *Otanta-asetelmat ja tilastollinen analyysi*. Helsinki: Gaudeamus.
- Vehkalahti K. (1998). *Kyselytutkimuksen mittarit ja menetelmät.* Helsinki: Tammi.

Oheismateriaalia

- 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

- Berger Y. et al. (2013). Handbook on precision requirements and variance estimation for ESS households surveys. Luxembourg: Eurostat Methodologies and Working papers.

- Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. In: C. R. Rao and D. Pfeffermann (eds.), *Handbook of statistics, vol. 29(B). Sample surveys: theory, methods and inference.* Elsevier.

Table 1.1 Real survey data sets used in examples and case studies.

Name of survey	Type of primary sampling unit PSU	Number of strata, clusters and elements in the survey data set: Strata	Clusters (PSU:s)	Elements
Census register data set				
(1) *Province'91* Population (data for one province)	Municipality	2	8 regional groups of municipalities	32 municipalities
Real survey data sets adjusted for pedagogical use				
(2) Mini-Finland Health Survey (data for males aged 30–64 years)	Municipality	24	48 municipalities	2699 persons
(3) Occupational Health Care Survey (data for establishments with 10 workers or more)	Industrial establishment	5	250 industrial establishments	7841 employees
Real survey data sets used in case studies				
(4) Passenger Transport Survey	Person	25	(Element-level sampling)	11711 persons
(5) Wages Survey	Business firm	25	744 firms	13987 employees
(6) Health Security Survey (data for one stratum)	Household	1	878 households	2071 persons
(7) PISA 2000 Survey (data for 7 countries)	School	7	1388 schools	32101 pupils

Ajatuksia virittävä harjoitus

- HS 26.10.2014 artikkeli, lainaus:
  - Yt-neuvottelut koskevat jo kaikkia aloja
  - Tänä vuonna yt-neuvotteluihin on kutsuttu ainakin 92 000 henkeä.
  - Työpaikkansa on menettänyt 0,39 prosenttia työllisistä.
  - Teemu Luukka HS
- Grafiikka (Petri Salmén) 26.10.2014
  - Kuviossa on kyseisen ajankohdan perusjoukkodata
  - Työttömäksi jääneet on luokiteltu yrityksen/laitoksen mukaan
- Miten poimisit tästä perusjoukosta otoksen haastattelutiedonkeruuta varten?
- Mitä asioita mielestäsi pitää ottaa huomioon otanta-asetelman laadinnassa?

Teemat

- Teema 1: Otannan perusmenetelmät ja lisäinformaation käyttö otanta- asetelmassa
- Teema 2: Lisäinformaation käyttö estimointiasetelmassa: Malliavusteinen estimointi
- Teema 3: Erityiskysymyksiä
- Teema 4: Tilastolliset ohjelmistot

Teema 1

OTANNAN PERUSMENETELMÄT JA LISÄINFORMAATION KÄYTTÖ OTANTA-ASETELMASSA

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

Tietoarkistot: Tärkeitä otosperusteisten valmisaineistojen varastoja

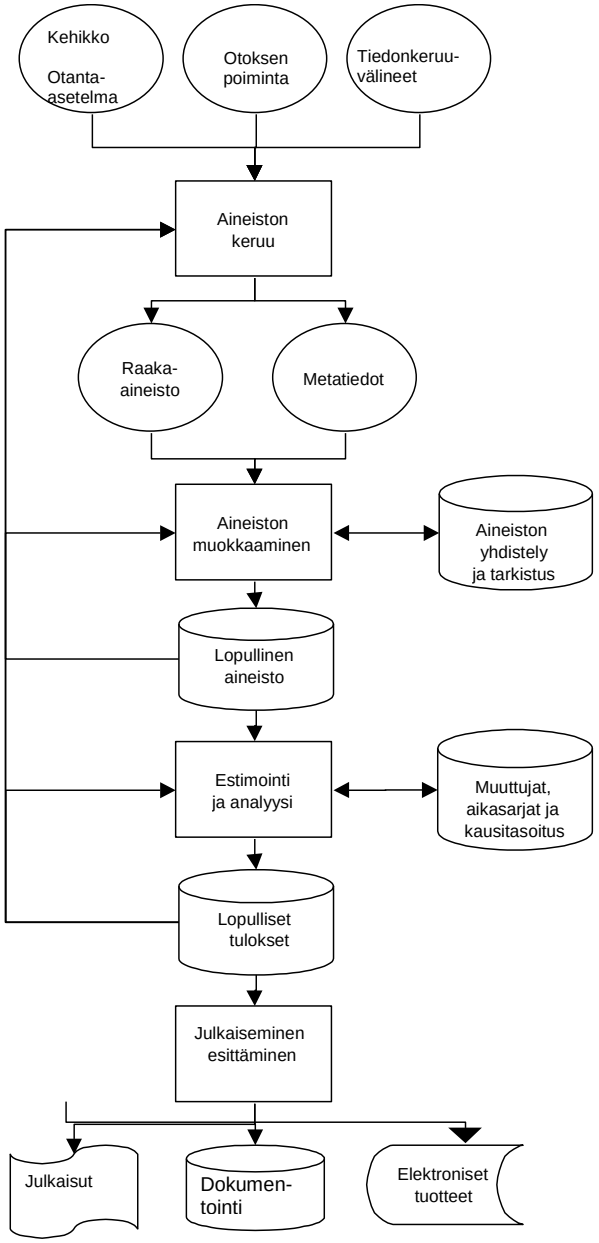
- Pääasiassa **otosperusteisia** kotimaisia ja kansainvälisiä aineistoja
  - Perustuvat suoraan tiedonkeruuseen
  - Kyselyaineistot, haastatteluaineistot
- Yhteiskuntatieteellinen tietoarkisto FSD
  - Tampereen yliopiston yhteydessä
- Council of European Social Science Data Archives CESSDA
- Esimerkki
  - European Social Survey ESS (2002-2012)
  - Riippumattomia poikkileikkausaineistoja

Rekisteriaineistot: Tärkeitä rekisteri-perusteisten tutkimusaineistojen lähteitä

- **Hallinnollinen rekisteri**
  - Hallinnollisen prosessin oheistuote
  - Jatkuva päivitys
    - Kela: Sosiaalivakuutuksen tietokannat
    - Verohallitus: Verotietokanta
    - Väestörekisterikeskus: Väestön keskusrekisteri
- **Tilastorekisteri** (Tilastokeskus)
  - Usean hallinnollisen rekisterin yhdistelmä
  - Rekisteriseloste: Tulonjakotilasto
  - StatFin – Tilastotietokannat
- Rekisteritutkimuksen tukikeskus ReTki
- Rekistereitä käytetään tutkimustiedon lähteinä ja tutkimusten otantakehikkoina

Survey-prosessi

- **Kaavio 1.** Survey-hankkeen operationaaliset vaiheet.
  - Muokattu lähteestä: Sundgren B. 1999. Information systems architecture for national and international statistical offices. Guidelines and recommendations. Geneva: United Nations, Statistical Standards and Studies 51. (Tilastokeskus, Laatukäsikirja)
- **Kaavio 2.**
  - Lehtonen R. and Pahkinen E. (2004). *Practical methods for Design and Analysis of Complex Surveys. Second Edition.* Chichester: John Wiley & Sons.



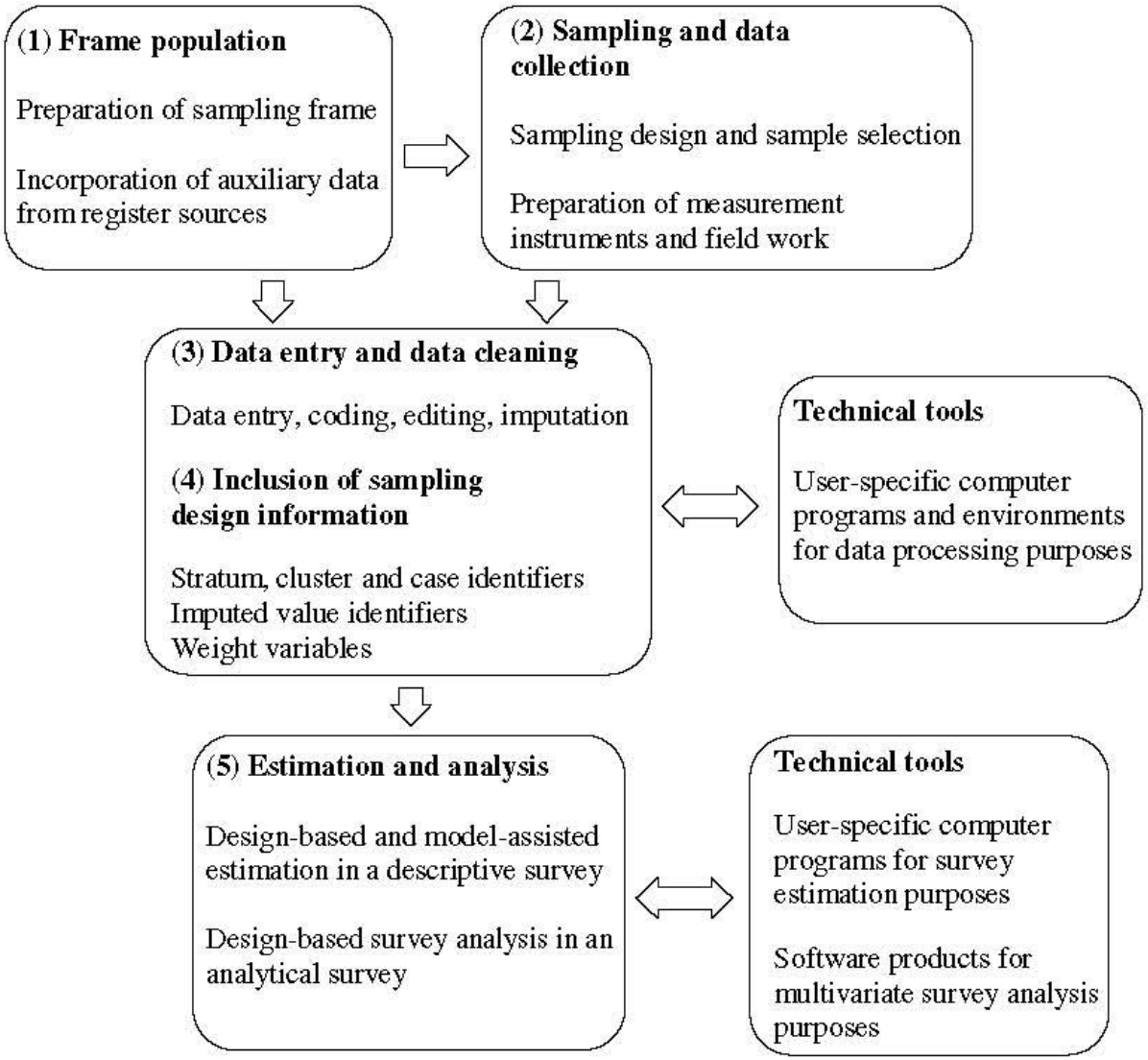


Figure 1.1 Flow chart for design-based estimation and analysis of complex survey data.

YHTEENVETO 1. Aineisto-optiot tiedonkeruun tavan ja kattavuuden mukaan.

TIEDONKERUUTAPA	KATTAVUUS PERUSJOUKON SUHTEEN	
	A. OSITTAINEN KATTAVUUS: OTOSTUTKIMUS	**B. TÄYSI KATTAVUUS: KOKONAISTUTKIMUS**
1. SUORA TIEDONKERUU **Tietolähde** **Haastattelututkimus** Tietokoneavusteinen käyntihaastattelu *Computer Assisted Personal Interview* CAPI Tietokoneavusteinen puhelinhaastattelu *Computer Assisted Telephone Interview* CATI Tietokoneavusteinen kysely *Computer Assisted Self-interview* CASI *Computer Assisted Web Survey* CAWI Tiedonkeruu kynä-ja-paperi -menetelmällä *Paper-and-Pencil Interview* PAPI **Postikysely** **Internet-kysely, Web-kysely, eSurvey**	**Optio 1a. Suoraan tiedonkeruuseen perustuva otostutkimus** Perinteinen otostutkimuksen tyyppi **Tilastokeskuksen tutkimuksia ja tilastoja** □ Työvoimatutkimus □ Kulutustutkimus **Kelan tutkimuksia ja selvityksiä** □ Terveysturvan väestötutkimukset **Monikansallisia tutkimuksia** □ European Social Survey ESS □ PISA	**Optio 1b. Suoraan tiedonkeruuseen perustuva kokonaistutkimus** Perinteinen kokonaistutkimuksen tyyppi □ Tilastokeskuksen väestölaskennat (vuoteen 1985 saakka)
2. EPÄSUORA TIEDONKERUU **Tietolähde:** **Rekisteri** Kattaa kohdeperusjoukon Päivitetään säännöllisesti **Hallinnollinen rekisteri** Hallinnollisen proseduurin oheistuote **Tilastorekisteri** Usean hallinnollisen rekisterin yhdistelmä	**Optio 2a. Hallinnolliseen rekisteriaineistoon perustuva otostutkimus** Puhtaana muotona harvinainen □ Poikkeuksena Tilastokeskuksesta saatavat tilastorekistereiden otosaineistot	**Optio 2b. Hallinnolliseen rekisteriin tai tilastorekisteriin perustuva kokonaistutkimus** Tämä surveyn tyyppi on yleistymässä Aineistolähteet □ Rekisteriperusteiset väestölaskennat □ Sosiaalivakuutuksen rekisterit □ Väestörekisteri □ Yritysrekisteri □ Verotusrekisterit □ Kelan lääketutkimukset
3. TIEDONKERUUTAPOJEN YHDISTELMÄ **Tietolähde:** Suoran ja epäsuoran tiedonkeruun yhdistelmä	**Optio 3. Otostutkimus, joka perustuu suoran tiedonkeruun ja rekisteriaineiston yhdistelyyn** Tämä surveyn tyyppi on yleistymässä □ KTL:n Terveys 2000 ja Terveys 2010 □ Kelan Mini-Suomi-terveystutkimus □ Tilastokeskuksen Tulonjakotutkimus □ EU:n European Community Household Panel ECHP □ EU SILC (Statistics on Income and Living Conditions)	

Kuvaileva ja analyyttinen survey

HY Otantamenetelmät syksy 2011

YHTEENVETO 2 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, SPSS, SUDAAN	SAS, SPSS, SUDAAN, Stata, MLwiN

Risto Lehtonen 2011

Alkiotasoinen lisätieto
Unit-level auxiliary information

- Kohteena olevista yksiköistä on usein käytettävissä **otoksen ulkopuolista lisätietoa** (*auxiliary information*) apumuuttujien muodossa
- Alkiotasoista lisätietoa saadaan rekisteri-perusteisista tietokannoista ja muista lähteistä
  - tilastorekisterit, hallinnolliset rekisterit, viralliset tilastot
- Apumuuttujat yhdistetään otosaineistoon **identifikaatiomuuttujien** avulla
  - henkilötunnus, yritystunnus, kuntanumero jne.

Aggregoitu lisätieto
Aggregate level auxiliary information

- Apumuuttujatieto voi olla saatavilla myös **perusjoukon kokonaismäärätietoina**
  - Viralliset tilastot ja vastaavat
- Jotta lisätiedon käytöstä on tehokkuushyötyä estimoinnissa, **tulee apumuuttujien korreloida tutkittavien tulosmuuttujien kanssa**
  - Hyötykriteeri: Estimoinnin tehostuminen eli estimaattorin varianssin ja keskivirheen pieneneminen

Lisätiedon kaksi käyttötapaa

A. Lisätiedon käyttö otanta-asetelmassa

- Tavoitteena **tehokkaan otanta-asetelman** konstruointi
  - mahdollisimman pienet keskivirheet
  - Ositettu otanta (*Stratified sampling* STR)
  - PPS-otanta
    - poiminta otosyksikön kokoon suhteutetuin todennäköisyyksin; *Probability Proportional to Size*

- SAS Procedure SURVEYSELECT
- SPSS Complex Samples module: CSSELECT

Lisätiedon kaksi käyttötapaa

B. Lisätiedon käyttö estimointiasetelmassa

- Tavoitteena **estimoinnin tehostaminen poimitulle otokselle**
  - keskivirheiden pienentäminen käytetyn otanta-asetelman puitteissa
  - Regressioestimointi
  - Suhde-estimointi
  - Kalibrointimenetelmät
  - Jälkiosittaminen…
- SAS Procedure SURVEYMEANS / SURVEYREG
- Vastaavat SPSS-toiminnot
- Vastaavat R-toiminnot

Strategia

- **Otanta-asetelman ja estimointiasetelman yhdistelmä**
- Lisäinformaation sisällyttäminen
  - Otanta-asetelmaan
  - Estimointiasetelmaan
  - Otanta-asetelmaan ja estimointiasetelmaan
- Tilastollisten mallien käyttö estimointiasetelmassa

Esimerkkejä strategioista

- Yksinkertainen satunnaisotanta SRS
  - Ei lisäinfoa, ei tilastollista mallia
- PPS-otanta
  - Lisäinformaatio otanta-asetelmassa
    - Perusjoukon alkion kokotieto
  - Ei (eksplisiittistä) tilastollista mallia
- SRS ja regressioestimointi
  - Ei lisäinfoa otannassa
  - Lisäinfo estimoinnissa: Jatkuva apumuuttuja
  - Tilastollinen malli: Regressiomalli

Estimointistrategioita perusjoukon kokonaismäärälle

Strategia		*Lisäinformaatio*	*Avustava malli*
Asetelmaperusteinen strategia			
SRSWOR		Ei käytetä	Ei ole
SRSWR		Ei käytetä	Ei ole
PPSWOR		Kokotieto	Ei ole
Malliavusteinen strategia			
Jälkiositus	SRS*pos	Diskreetti	ANOVA
Suhde-estimointi	SRS*rat	Jatkuva	Regressiomalli (ei vakiotermiä)
Regressioestimointi	SRS*reg	Jatkuva	Regressiomalli

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*,

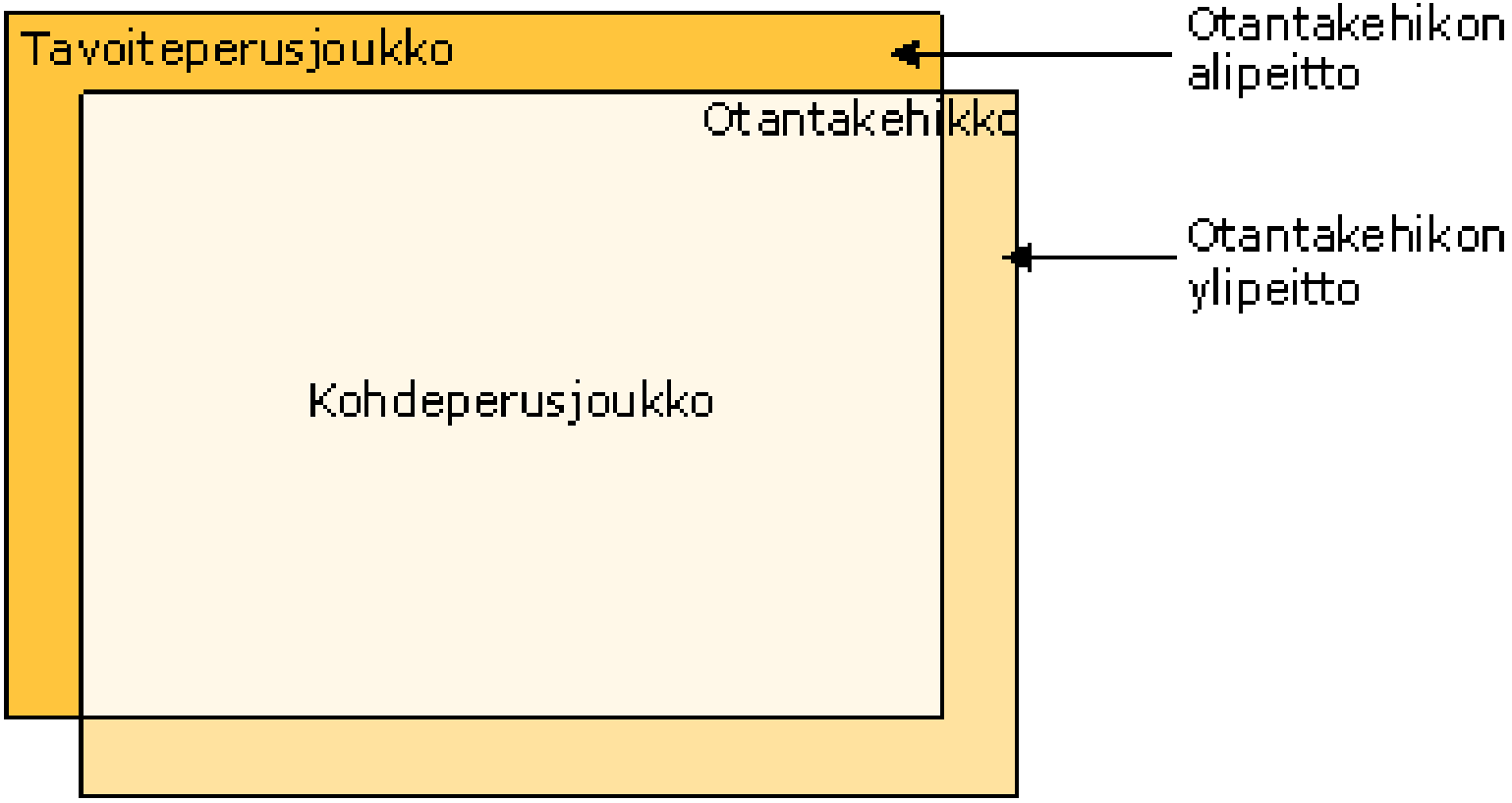
$k = 1,...,N$

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
  - 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

Huomioita asetelmapainosta

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

$k = 1,...,n,$ missä n on otoskoko

Useissa otanta-asetelmissa painolle pätee

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

Asetelmapainoja tarvitaan kun estimoidaan kokonaismääriä (esim. työttömien kokonaismäärä)

HUOM: Muissa tilanteissa kannattaa käyttää analyysipainoa

Analyysipaino

- Analyysipainon (*analysis weight*) laadinta

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

tehdään uudelleenskaalattu painokerroin

$$w_k^* = (n/N)w_k$$

missä n on otoskoko ja N on perusjoukon koko

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

joten analyysipainojen keskiarvo = 1

HUOM: SRS-otokselle analyysipaino = 1

- **Taulukko**
- Province91-perusjoukko
- *N* = 32 kuntaa
- Tulosmuuttuja
  - UE91
- Apumuuttujat
  - STR osite
    - Kuntamuoto
  - HOU85
    - Kotitalouksien lkm

- Lähde: Lehtonen R. and Pahkinen E. (2004). Practical Methods for Design and Analysis of Complex Surveys. Second Edition. Wiley.

Table 2.1 The *Province'91* population. Percentage unemployment (%UE) and totals of unemployed persons (UE91), labour force (LAB91), population in 1991 (POP91) and number of households (HOU85) by municipality in the province of Central Finland in 1985.

ID	LABEL	STR	CLU	%UE	UE91	LAB91	POP91	HOU85
	Urban			**12.67**	**8022**	**63 314**	**129 460**	**49 842**
1	Jyväskylä	1	1	12.20	4123	33786	67 200	26 881
2	Jämsä	1	2	11.07	666	6016	12907	4663
3	Jämsänkoski	1	2	13.83	528	3818	8118	3019
4	Keuruu	1	2	12.84	760	5919	12707	4896
5	Saarijärvi	1	3	14.62	721	4930	10774	3730
6	Suolahti	1	5	15.12	457	3022	6159	2389
7	Äänekoski	1	3	13.17	767	5823	11 595	4264
	Rural			**12.63**	**7076**	**56 011**	**125 124**	**41 911**
8	Hankasalmi	2	5	15.07	391	2594	6080	2179
9	Joutsa	2	6	9.38	194	2069	4594	1823
10	Jyväskylän mlk.	2	7	11.82	1623	13727	29349	9230
11	Kannonkoski	2	4	18.64	153	821	1919	726
12	Karstula	2	4	13.53	341	2521	5594	1868
13	Kinnula	2	8	13.92	129	927	2324	675
14	Kivijärvi	2	8	15.63	128	819	1972	634
15	Kongink angas	2	3	21.04	142	675	1636	556
16	Konnevesi	2	5	12.91	201	1557	3453	1215
17	Korpilahti	2	1	11.15	239	2144	5181	1793
18	Kuhmoinen	2	2	12.91	187	1448	3357	1463
19	Kyyjärvi	2	4	11.31	94	831	1977	672
20	Laukaa	2	5	12.11	874	7218	16 042	4952
21	Leivonmäki	2	6	10.65	61	573	1370	545
22	Luhanka	2	6	10.34	54	522	1153	435
23	Multia	2	7	11.24	119	1059	2375	925
24	Muurame	2	1	9.79	296	3024	6830	1853
25	Petäjävesi	2	7	15.08	262	1737	3800	1352
26	Pihtipudas	2	8	13.02	331	2543	5654	1946
27	Pylkönmäki	2	4	17.98	98	545	1266	473
28	Sumiainen	2	3	12.80	79	617	1426	485
29	Säynätsalo	2	1	10.28	166	1615	3628	1226
30	Toivakka	2	6	11.72	127	1084	2499	834
31	Uurainen	2	7	16.47	219	1330	3004	932
32	Viitasaari	2	8	14.16	568	4011	8641	3119
	Whole province			**12.65**	**15 098**	**119 325**	**254 584**	**91 753**

Sources: Statistics Finland: Population Census 1985. Statistics Finland (1992): Statistical Yearbook of Finland, Volume 87. Ministry of Labour of Finland (1991): Employment Service Statistics, November 30, 1991.

Esimerkki: Yksinkertainen satunnaisotanta SRS *Simple random sampling*

SRS-otanta, $n = 8$ otosalkiota

Perusjoukossa $N = 32$ kuntaa

Sisältymistn $\pi_k = \pi = 8/32 = 0.25$

Asetelmapaino $w_k = 1/\pi_k = 1/0.25 = 4$

$$\sum_{k=1}^{8} w_k = N = 32$$

VLISS-Virtual Laboratory in Survey Sampling

- Province'91 perusjoukko (*population*)

(entinen Keski-Suomen lääni)

  - Tilastoyksikkönä (alkiona) kunta
  - *N* = 32 kuntaa
- Tulosmuuttuja UE91: Työttömien lukumäärä läänissä
- **VLISS-toteutus**

Chapter 2. Basic sampling techniques

2.1 Basic definitions 2.2 The Province ´91 population

2.3. Simple Random Sampling and design effect

TRAINING KEY 28 Analysing an SRS sample

http://vliss.helsinki.fi/

Uudelleenpainotus *Reweighting*

- Asetelma- ja analyysipainojen konstruoinnin lisäksi usein tarvitaan painojen muokkausta kadon (*nonresponse*) vaikutusten oikaisemiseksi
  - Uudelleenpainotus
  - Estimoidaan ensin vastaustodennäköisyys (*response probability*)
    - Aineiston osajoukoissa tai
    - Alkioittain
  - Korjataan analyysipainoja estimoitujen vastaustodennäköisyyksien avulla

Kolme esimerkkiä painotuksesta

- **PISA-tutkimussarja**
- Tyypillinen otanta-asema: Ositettu kaksiasteinen ryväsotanta
  - Poimintarypäänä koulu
  - Koulupopulaation alueellinen ositus ennen poimintaa
  - 1. aste: Otoskoulujen poiminta
  - 2. aste Oppilasryhmien poiminta otoskouluista
  - Oppilastason asetelmapainon muodostus: uudelleenpainotus
- **Terveys 2000 -tutkimus**
- Otanta-asema: Ositettu kaksiasteinen ryväsotanta
  - Poimintarypäänä terveyskeskuspiiri
  - Ryväspopulaation alueellinen ositus
  - 1. aste: Otos terveyskeskuspiirien populaatiosta
  - 2. aste: Henkilöotos otokseen tulleista terveyskeskuspiireistä
  - Henkilötason asetelma- ja analyysipainojen muodostus: uudelleenpainotus
- European Social Survey ESS : Weighting ESS sampling

ESIMERKKI: PISA
Weighting procedure (adjusted design weight)

Weight w_{ik} for student k in school i:

$$w_{ik} = w_{1i} \times w_{2ik} \times f_i, \quad i = 1,...,m \text{ and } k = 1,...,n_i,$$

where

$w_{1i} = 1/(\pi_i \hat{\theta}_i)$ is the reciprocal of the product of the inclusion probability π_i and the estimated participation probability $\hat{\theta}_i$ of school i;

$w_{2ik} = 1/(\pi_{k|i} \hat{\theta}_{k|i})$ is the reciprocal of the product of the conditional inclusion probability $\pi_{k|i}$ and estimated conditional response probability $\hat{\theta}_{k|i}$ of student k from within the selected school i;

f_i is an adjustment factor for school i to compensate any country-specific refinements in the survey design, and

m is the number of sample schools in a given country and n_i is the number of sample students in school i.

Esimerkki: Terveys 2000 -tutkimus Painotusmenetelmä

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. This weight is appropriate in statistical analysis.

Vastauskato (1)

- **Vastauskatoa** *(non-response)* esiintyy usein vapaaehtoisuuteen perustuvissa kysely- ja haastattelututkimuksissa.

- Vastauskato jaetaan kahteen pääryhmään:
  - **Eräkatoon** (*item non-response*) ja
  - **Yksikkökatoon** (*unit non-response*).

- **Eräkadolla** tarkoitetaan sellaista vastausta, jossa tutkimusyksiköltä on saatu vain osa tiedoista

- **Yksikkökadon** tapauksessa kaikki tutkimusyksikköä koskevat tiedot puuttuvat tai joudutaan hylkäämään

- Useimmiten vastaajat ja katoon jääneet yksiköt poikkeavat toisistaan tutkittavien ilmiöiden suhteen

- Tämä voi aiheuttaa tutkimustuloksiin harhaa

Vastauskato (2)

- **Yksikkökato**
  Unit nonresponse
  - Uudelleenpainotus-menetelmät
    - RHG-menetelmä
      Response homogeneity groups
  - Mallinnusmenetelmät
    - Logistinen katomalli
  - **Katoanalyysi ja katoon reagointi ovat empiirisen tutkimuksen tärkeitä työvaiheita**

- **Eräkato**
  Item nonresponse
  - Imputointimenetelmät
    - Hot deck
    - Lähimmän naapurin menetelmä
      Nearest neighbour method
    - Moni-imputointi
      Multiple imputation
  - **Imputointimenetelmien käyttö on yleistymässä eri tieteenaloilla ja sovelluksissa**

Otanta-asetelman laadintavaiheet

A. Perusjoukkojen määrittely

- Alkiotason perusjoukko
- Ryvästason perusjoukko

B. Otanta-asteiden määrittely

- Alkiotason otanta
- Ryväsotanta
  - Yksiasteinen otanta
  - Kaksiasteinen otanta
  - Moniasteinen otanta

C. Otantamenetelmien kiinnittäminen eri otanta-asteille

- Osittaminen
- Otoksen kiintiöinti
- Alkioiden poimintamenetelmän valinta kullakin otanta-asteella ja ositteessa

Alkiotason otanta ja ryväsotanta

(1) Alkiotason otanta (*element sampling*)

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

(2) Ryväsotanta (*cluster sampling*)

- Otantayksikkönä on perusjoukon alkioiden muodostama luonnollinen ryhmä eli **ryväs** (*cluster*)
- Esim:
  - 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 ryväsotanta**
  - 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 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)

Otannan perusmenetelmät

- Peruskäsitteet: **Tekninen yhteenveto I**
- Yksinkertainen satunnaisotanta
  SRS - *Simple random sampling*
- Bernoulli-otanta BER
- Systemaattinen otanta
  SYS - *Systematic sampling*
- Ei käytetä lisäinfoa
- Sisältymistn on sama kaikille pj:n alkioille

Yksinkertainen satunnaisotanta SRS
Simple random sampling

- Kunkin perusjoukon alkion otokseen sisältymis-todennäköisyys on vakio *n*/*N* missä *n* on otoskoko ja *N* on perusjoukon alkioiden lukumäärä
- Tekninen toteutus esimerkiksi satunnaislukujen avulla
- Erikoistapaukset:
  - SRSWOR
    - SRS **palauttamatta** (*without replacement*)
  - SRSWR
    - SRS **palauttaen** (*with replacement*)
- SAS Procedure SURVEYSELECT Syntax

Yksinkertainen satunnaisotanta SRS

MERKINTÖJÄ

Perusjoukko U, jossa on N alkiota

Perusjoukon tuntemattomat arvot $Y_1, Y_2, ..., Y_k, ..., Y_N$

Parametrit $T = \sum_{k \in U} Y_k = \sum_{k=1}^{N} Y_k$ kokonaismäärä

$\bar{Y} = \sum_{k=1}^{N} Y_k / N$ keskiarvo

Perusjoukon alkion k sisältymistodennäköisyys $\pi_k,\ k = 1, ..., N$

Otos s, jossa on n alkiota

Otoksesta mitatut arvot $y_1, y_2, ..., y_k, ..., y_n$

Otosalkion k asetelmapaino $w_k = 1/\pi_k,\ k = 1, ..., n$

Kokonaismäärän estimaattori $\hat{t} = \sum_{k \in s} w_k y_k = \sum_{k=1}^{n} w_k y_k$

Keskiarvon estimaattori $\bar{y} = \hat{t} / N$

Yksinkertainen satunnaisotanta SRS

Sisältymistodennäköisyys $\pi_k = n/N$ on vakio

Kokonaismäärän T **estimaattori**

$$\hat{t}_{SRS} = N\bar{y} = N\sum_{k=1}^{n} y_k / n = \frac{N}{n}\sum_{k=1}^{n} y_k ,$$

missä $\bar{y}$ on otoskeskiarvo ja N perusjoukon koko

Horvitz-Thompson (HT) estimaattori (1952)

$$\hat{t}_{HT} = \sum_{k\in s} w_k y_k = \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

missä $w_k = 1/\pi_k$ on **asetelmapaino** (*design weight*)

SRS-otanta: $w_k = N/n$

HT-estimaattorin harhattomuus

Totaali: $T = \sum_{k \in U} Y_k$

Estimaattori: $\hat{t}_{HT} = \sum_{k \in s} w_k y_k = \sum_{k \in s} \frac{y_k}{\pi_k}$

Olkoon otosindikaattori $I_k = 1$ kun $k \in s$, 0 muulloin

$$\hat{t}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} I_k \frac{y_k}{\pi_k}$$

$$E(\hat{t}_{HT}) = E(\sum_{k \in s} \frac{y_k}{\pi_k}) = \sum_{k \in U} E(I_k \frac{y_k}{\pi_k})$$

$$= \sum_{k \in U} \frac{y_k}{\pi_k} E(I_k) = \sum_{k \in U} \frac{y_k \pi_k}{\pi_k} = \sum_{k \in U} y_k = T$$

Yksinkertainen satunnaisotanta SRS

SRSWOR: Totaalin varianssiestimaattori ja **keskivirhe**

$$\hat{v}_{SRS}(\hat{t}) = N^2(1-\frac{n}{N})(\frac{1}{n})\sum_{k=1}^{n}(y_k-\bar{y})^2/(n-1) = N^2(1-\frac{n}{N})(\frac{1}{n})\hat{s}^2$$

$\text{s.e}_{SRS}(\hat{t}) = \sqrt{\hat{v}_{SRS}(\hat{t})}$ on estimoitu keskivirhe (*standard error* s.e)

missä $\bar{y} = \sum_{k=1}^{n} y_k / n$ on otoskeskiarvo

$\hat{s}^2 = \sum_{k=1}^{n}(y_k-\bar{y})^2/(n-1)$ on otosvarianssi

$(1-\frac{n}{N})$ on äärellisyyskorjaus (fpc, *finite population correction*)

SRSWOR: Keskiarvon varianssiestimaattori ja **keskivirhe**

$$\hat{v}_{SRS}(\bar{y}) = (1-\frac{n}{N})(\frac{1}{n})\hat{s}^2 \quad \text{ja} \quad \text{s.e}_{SRS}(\bar{y}) = \sqrt{\hat{v}_{SRS}(\bar{y})}$$

Perusjoukon *Population91* parametrit

UE91 totaali:

$$T = \sum_{k=1}^{N} Y_k = 15098$$

UE91 keskiarvo:

$$\bar{Y} = \sum_{k=1}^{N} Y_k / N = 15098/32 = 472$$

SRSWOR-otanta perusjoukosta (VLISS)

UE91 totaalin T estimaatti, varianssiestimaatti ja 95 % lv:

$$\hat{t} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \frac{N}{n} \sum_{k=1}^{n} y_k = \frac{32}{8} \sum_{k=1}^{8} y_k = 26440$$

$$\hat{v}_{SRS}(\hat{t}) = N^2 (1 - \frac{n}{N})(\frac{1}{n}) \sum_{k=1}^{n} (y_k - \bar{y})^2 / (n-1)$$

$$= 32^2 (1 - \frac{8}{32})(\frac{1}{8}) \hat{s}^2 = 13282^2$$

95 % LV: -4967 -- 57847 Käyttökelvoton tulos!

Pyritään tehokkaampaan estimointiin tuomalla lisäinformaatiota otanta-asetelmaan tai estimointiasetelmaan