

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otantamenetelmät

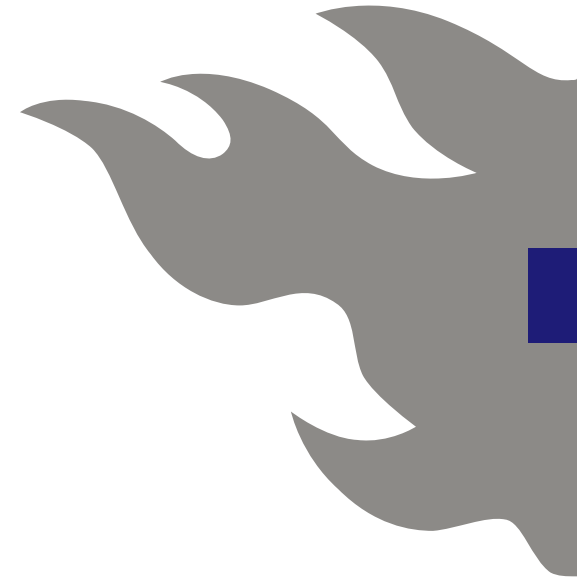
(78143)

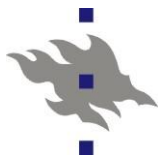
Syksy 2012

TEEMA 1

Risto Lehtonen

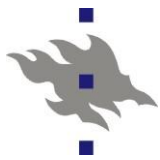
risto.lehtonen@helsinki.fi





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 mm. sisältyy valtiotieteellisen tiedekunnan [menetelmäkoriin](#)
- Soveltuvia jatkokursseja, kevät 2013
 - [Pienalue-estimointi](#) (periodi 3)
 - [Imputointimenetelmät](#) (periodi 3-4)



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

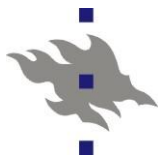


Sisältöalueita

- **Otannan perusmenetelmät**
 - yksinkertainen satunnaisotanta
 - systemaattinen otanta

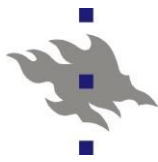
- **Lisäinformaation käyttö otanta-asetelmassa**
 - PPS-otanta (*Sampling with probability-proportional-to-size*)
 - ositettu otanta

- **Lisäinformaation käyttö estimointiasetelmassa**
 - suhde-estimointi
 - regressioestimointi
 - yleistetty regressioestimointi (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-makrot CLAN ja EBLUPGREG
 - SPSS-toiminto Complex Samples
Useita mahdollisuuksia
- Stata - svy-ohjelmat



Esimerkkejä

- Esimerkkejä otosperusteisista tiedonkeruista ja tutkimuksista
- OECD:n Pisa-tutkimussarja [MyPISA](#)
- [Terveys 2000](#), Terveys 2010
- [EU SILC](#) - Statistics on Income and Living Conditions



Tilastokeskuksen työvoimatutkimus

- Tilastokeskuksen työvoimatutkimus
- Laatuseloste
- EU-LFS (Labour Force Survey)

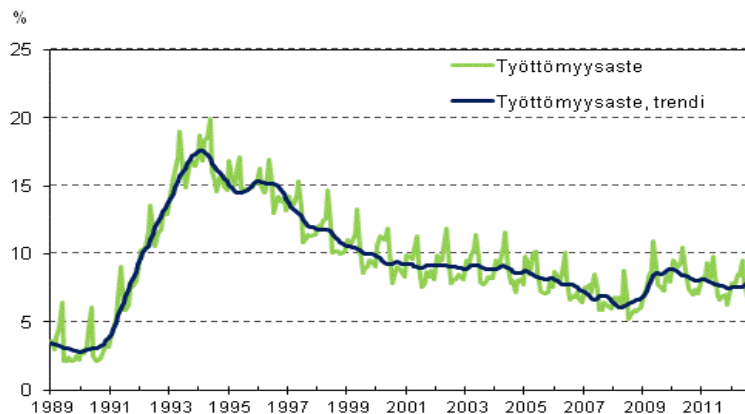
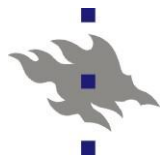


Table 1.1 Real survey data sets used in examples and case studies.

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) <i>Province'91</i> 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	32 101 pupils



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. (2011). 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.



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
- 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.



Teemat

- Teema 1: Otannan perusmenetelmät ja lisäinformaation käyttö otantaasetelmassa
- 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 - 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

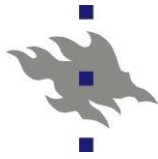
- 6. **Otoksen poiminta**
- 7. Tiedonkeruu
- 8. Tiedostonmuodostus

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



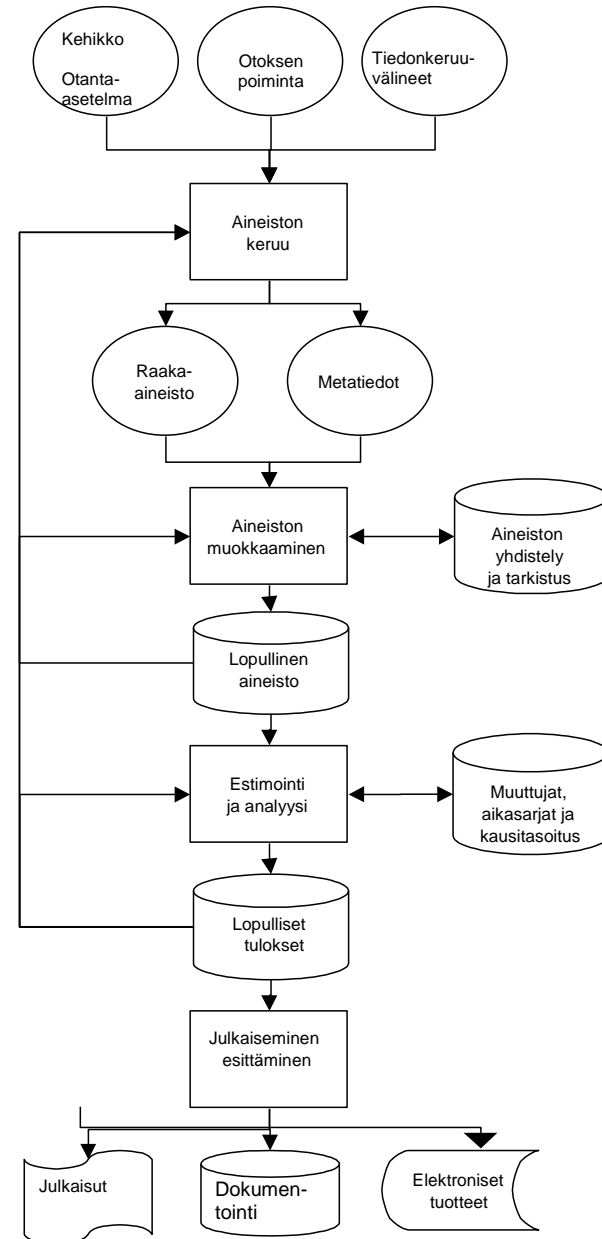
Survey-prosessi

■ 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.



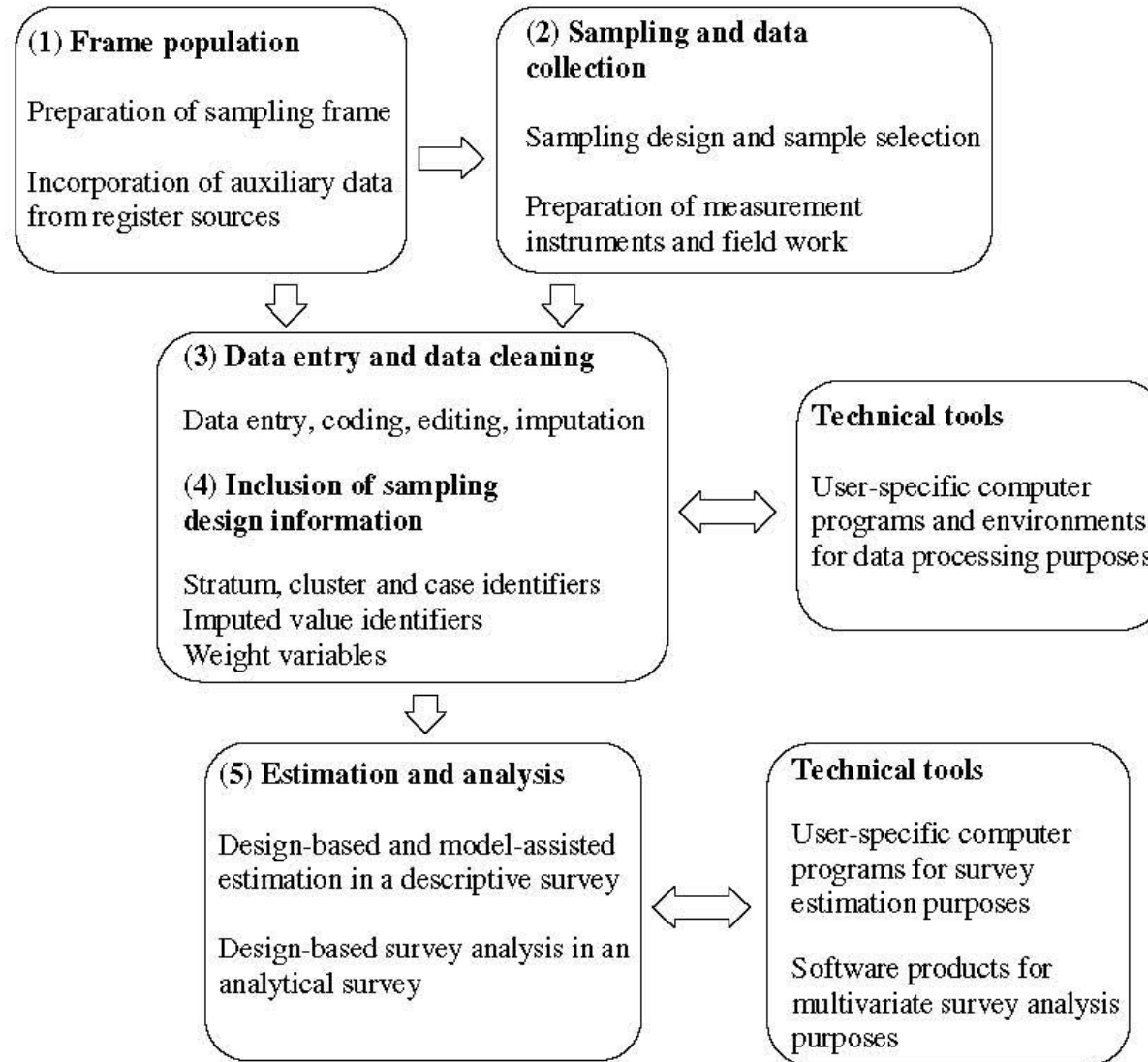


Figure 1.1 Flow chart for design-based estimation and analysis of complex survey data.

HY Otantamenetelmät syksy 2011 Risto Lehtonen

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

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



Tiedonkeruumenetelmiä

- Tyypillisiä tiedonkeruumenetelmiä
Mode of data collection
 - PAPI
 - *Paper-and-pencil*
 - Paperilomake

 - CAPI
 - *Computer assisted personal interview*
 - Tietokoneavusteinen käyntihaastattely

 - CATI
 - *Computer assisted telephone survey*
 - Tietokoneavusteinen puhelinhaastattelu



Nettikysely

- web-CASI
 - *internet-based computer assisted self-interviewing*
 - Itsevalikoituva web-kysely
Self-selection web survey
 - *Computer Assisted Web Survey CAWI*
 - Internet-kysely, nettikysely
-
- tai jokin yllä listattujen yhdistelmä
Mixed-mode survey



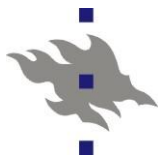
Nuoria työttömiä koskeva nettikysely mediassa 2009

- "Työttömyys hävettää entistä harvempia nuoria"
 - julkaistu ma 31.8.2009 klo 05:56
 - [YLE Uutiset](#)
- "Monet nuoret työttömät suhtautuvat työttömyyteensä myönteisesti. Uuden tutkimuksen mukaan yli 40 prosenttia työttömistä nuorista ei pidä työttömyyttä pahana asiana, jos toimeentulo on muuten turvattu."



Nuoria työttömiä koskeva nettikysely

- Ministry of Labour, työministeriö
 - Toukokuu 2009
 - $n = 716$ nuorta työtöntä (16-29 v.)
 - Web-lomake on ollut MOL:n sivustolla
- Nuori työtön on löytänyt lomakkeen työtä tai työvoimatoimenpiteitä koskevaa tietoa etsiessään
- "Oletko nuori aikuinen, joka on ollut joskus työttömänä tai olet par- aikaa työtön...- jos , niin vastaa..."



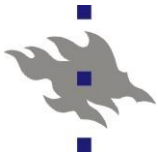
Nettikyselyn metodiikkaa

- Millaisia yleistyksiä nettikyselyn perusteella voidaan tehdä?
- Kriittinen näkökulma:
- Jelke Bethlehem (2008). How accurate are self-selection web surveys? Statistics Netherlands, [Discussion paper](#) (08014).

<http://www.cbs.nl/NR/rdonlyres/EEC0E15B-76B0-4698-9B26-8FA04D2B3270/0/200814x10pub.pdf>



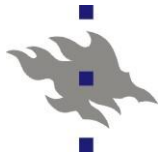
Summary: A web survey seems to be an attractive means of collecting survey data, because it provides simple, cheap and fast access to a large group of people. However, there are pitfalls. Due to methodological problems, the quality of the outcomes of web surveys may be seriously affected. This paper addresses one of these problems, and that is self-selection of respondents. Self-selection leads to a lack of representativity and thus to biased estimates. The effect of self-selection on the distributional characteristics of estimators is described in detail. It is shown that the bias of estimators in self-selection surveys can be much larger than in surveys based on traditional probability samples. A simulation study also shows what can go wrong in a self-selection web survey. It is explored whether some correction techniques (adjustment weighting and use of reference surveys) can improve the quality of the outcomes. It turns out that there is no guarantee for success.



Nettikyselyn metodiikkaa

- Menetelmällinen tarkastelu ja sovellus Kelan aineistoihin:
- Jani Miettinen (2011). Itsevalikoituminen verkkokyselytutkimuksen haasteena. Helsingin yliopisto, tilastotieteen pro gradu –tutkielma.

https://helda.helsinki.fi/bitstream/handle/10138/29260/itsevalikoituminen_verkkokyselytutkimuksen_haasteena.pdf?sequence=2



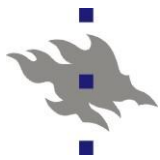
Kuvaileva ja analyyttinen survey

HY Otantamenetelmät syksy 2011

YHTEENVETO 2 KUVAILEVA JA ANALYYTTINEN SURVEY

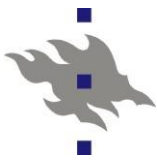
	KUVAILEVA	ANALYYTTINEN	
Tulosmuuttajat	Muutamia	Useita	
Yleistystaso	Kiinteä perusjoukko	"Superpopulaatio"	
Estimoitavat parametrit	Kuvailevia, esim. totaalit, keskiarvot	Analyttisiä, esim. regressiokertoimet	
Estimaattori-tyypit	Lineaarisia, esim. totaalin HT-estimaattori	Epälineaarisia, esim. regressiokertoimen PNS-estimaatto ri	
Varianssien estimointi	Analyttisesti	Approksimatiivisesti	
Ulkoisen lisäinfon käyttö analyysissä	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)	Otostasoa (n)	
Tilastolliset ohjelmistot	SAS, GES, CLAN, SPSS, SUDAAN	SAS, SPSS, SUDAAN, Stata, MLwiN	

Risto Lehtonen 2011



Lisätieto – auxiliary information

- Yleensä perusjoukon alkioista on käytettävissä **otoksen ulkopuolista lisätietoa** apumuuttujien muodossa
- Lisätietoa saadaan eri rekisterilähteistä
 - tilastorekisterit, hallinnolliset rekisterit, viralliset tilastot
- Apumuuttujat yhdistetään otosaineistoon **identifikaatiomuuttujien** avulla
 - henkilötunnus, yritystunnus, kuntanumero jne.



Lisätieto – auxiliary information

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



Lisätiedon kaksi käyttötapaa

A. Lisätiedon käyttö otanta-asetelmassa

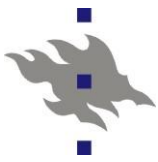
- Tavoitteena tehokkaan otanta-asetelman konstruointi
 - mahdollisimman pienet keskivirheet
 - Ositettu otanta (*Stratified sampling* STR)
 - PPS-otanta
 - poiminta otosyksikön kokoon suhteutetuin todennäköisyyksin; *Probability Proportional to Size*
- SAS Procedure [SURVEYSELECT](#)
- SPSS Complex Samples module: [CSSELECT](#)



Lisätiedon kaksi käyttötapaa

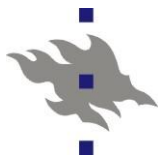
B. Lisätiedon käyttö estimointiasetelmassa

- Tavoitteena estimoinnin tehostaminen poimitulle otokselle
 - keskivirheiden pienentäminen käytetyn otanta-asetelman puitteissa
 - Regressioestimointi
 - Suhde-estimointi
 - Kalibrointimenetelmät
 - Jälkiosittaminen...
- SAS Procedure SURVEYMEANS
- SAS Procedure SURVEYREG



Strategia

- **Otanta-asetelman ja estimointiasetelman yhdistelmä**
- **Lisäinformaation sisällyttäminen**
 - Otanta-asetelmaan
 - Estimointiasetelmaan
 - Otanta-asetelmaan ja estimointiasetelmaan
- **Tilastollisen mallin käyttö estimointiasetelmassa**



Esimerkkejä strategioista

- Yksinkertainen satunnaisotanta SRS
 - Ei lisäinfoa, ei tilastollista mallia
- PPS-otanta
 - Lisäinformaatio otanta-asetelmassa
 - Perusjoukon alkion kokotieto
 - Ei 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 vakiotermejä)

Regressioestimointi

SRS*reg

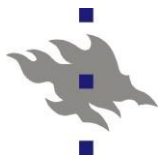
Jatkuva

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, nolaa suurempi todennäköisyys π_k tulla mukaan n alkion otokseen

$$0 < \pi_k \leq 1$$

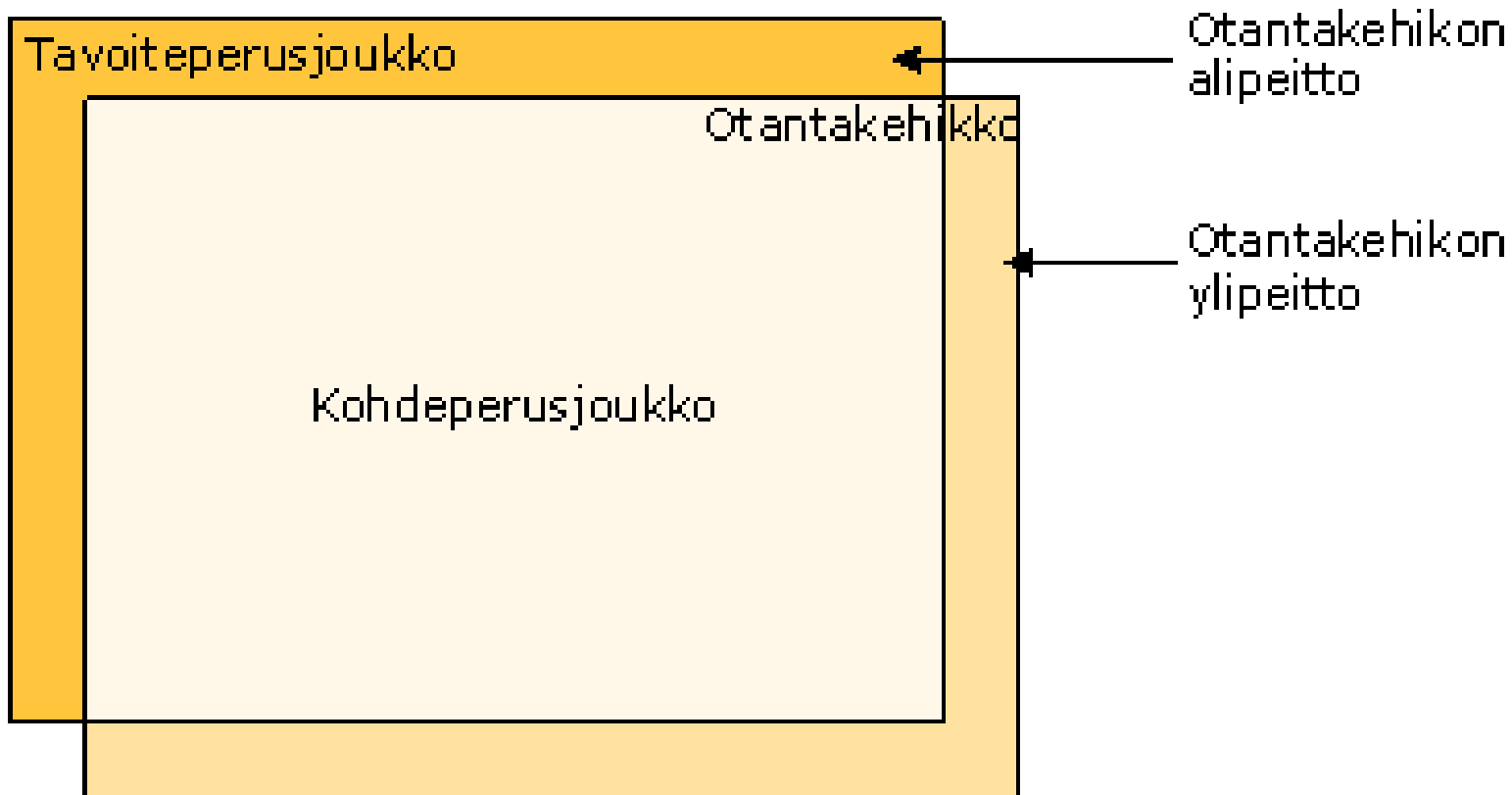
perusjoukon alkiolle k ,

$$k = 1, \dots, N$$

missä N on perusjoukon alkioiden lukumäärä

Otantakehikon alipeitto ja ylipeitto

Tilastokeskus: Laatusa tilastoissa -käsikirja



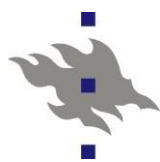


Otos Sample

- Perusjoukon osajoukko
- Poimitaan jollain satunnaisotannan menetelmällä
(*Random sampling, Probability sampling*)
- Poiminnassa käytetään sisältymistodennäköisyyksiä
(*Inclusion probability*)
- Miksi satunnaisotanta?
 - Otoksesta saatavat tulokset voidaan yleistää koskemaan koko kiinnostuksen kohteena olevaa perusjoukkoa tai hypoteettista mallia
 - Tilastollinen päättely
 - Piste-estimaatit
 - Kesquivirheet
 - 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 ,
 $k = 1, \dots, n$, missä n on otoskoko

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

Asetelmapainoja tarvitaan kun estimoidaan kokonaismääriä (esim. työttömien kokonaismäärä)

HUOM: Muissa tilanteissa kannattaa käyttää analyysipainoa

- **Taulukko**
- **Province91 -**
- perusjoukko
- **$N = 32$ kuntaa**
- **Tulosmuuttuja**
 - **UE91**
- **Apumuuttajat**
 - **STR osite**
 - Kuntamuoto
 - **HOU85**
 - Kotitalouksien
 - Ikä
- **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	11 395	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 mk.	2	7	11.82	1623	13727	29 349	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	Konginkangas	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	Pihlajavesi	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	Vitasaari	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

SRS-otanta, $n = 8$ otosalkiota

Perusjoukossa $N = 32$ kuntaa

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

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

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

TRAINING KEY 28 [Analysing an SRS sample](#)



VLISS-Virtual Laboratory in Survey Sampling

- Province'91 perusjoukko (*population*)
(entinen K-S lääni)
 - Tilastoyksikkönä (alkiona) kunta
 - $N = 32$ kuntaa
- Tulosuuttuja UE91: Työttömien lkm 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/>



Analyysipaino

■ Analyysipainon (*analysis weight*) laadinta

Tehdään uudelleenskaalattu painokerroin

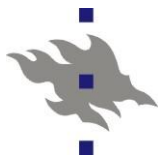
$$w_k^* = (n/N)w_k$$

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



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



Esimerkki: Health 2000 – Weighting procedures

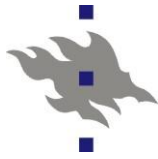
Sampling weight $w_{hik} = 1/\pi_{hik}$ where π_{hik} denotes the inclusion probability of person k in cluster i of stratum h in the population.

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

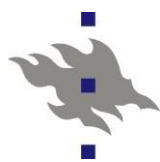
estimated response probability of sample person k in cluster i of stratum h .

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.



Vastauskadon hallinta (1)

- Tiedonkeruun eri vaiheissa havaintojen määrä usein pienenee erilaisista syistä.
- Perinteisesti **vastauskatoa** (*non-response*) esiintyy vapaaehtoisuuteen perustuvissa kysely- ja haastattelututkimuksissa.
- Vastauskato rekisteriaineistoissa?
- Vastauskato nettikyselyissä?
- Vastauskato jaetaan kahteen pääryhmään:
 - **eräkatoon** (*item non-response*) ja
 - **yksikkökattoon** (*unit non-response*).
- **Eräkadolla** tarkoitetaan sellaista vastausta, jossa tutkimusyksikkö antaa vain osan tiedoista hyväksyttävästi tai antaa sellaisen vastauksen, joka myöhemmissä aineiston tarkistuksissa joudutaan hylkäämään.



Vastauskadon hallinta (2)

- **Yksikkökadon** tapauksessa kaikki havaintoyksikköä koskevat tutkimustiedot puuttuvat tai joudutaan hylkäämään.
- Kadolla on vaikutusta tutkimuksen tuloksiin.
 - Tutkimusjoukon pienenemisellä tavoitteeseen eli perusjoukkoon nähden on useimmiten harmillisia vaikutuksia.
- Mikäli vastanneet ja kato ovat sekä tausta- että tutkimusmuuttujien suhteen samoin jakautuneita, otosvarianssi suurenee kadon vaikutuksesta.
- Myös kokonaistutkimukseen syntyy otosvarianssia tällä tavoin.
- Useimmiten vastaajat ja katoon jääneet yksiköt poikkeavat toisistaan, mikä aiheuttaa tutkimuksen tuloksiin virhettä, pahimmassa tapauksessa harhaa.



Vastauskadon hallinta (3)

■ Yksikkökato

Unit nonresponse

- Uudelleenpainotusmenetelmät
 - RHG-menetelmä
Response homogeneity groups
- Mallinnusmenetelmät
 - Logistinen katomalli
 - Terveys 2000
- **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
- Ryväsootanta
 - 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

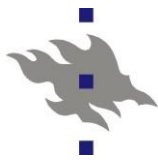
(*element sampling*)

- Otantayksikkönä on perusjoukon alkiio (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 ryväyksiköistä 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

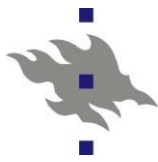
- Rypäiden poiminta ryvästason perusjoukosta PPS-otannalla
- Alkioiden poiminta otosrypäistä systemaattisella otannalla



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

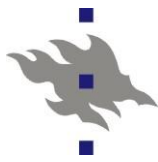
Otos poimitaan tasavälisesti listasta tai rekisterinä olevasta tietokannasta

STR

Stratified sampling

Ositettu otanta

Perusjoukon alkiot jaetaan ensin homogeenisiin 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

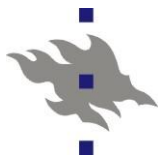
1) Rypäiden perusjoukosta poimitaan otosrypäät
2) Otosrypäiden alkiosta poimitaan alkiotason otokset SRS:llä tai SYS:llä

PPS

Selection with Probabilities

Proportional to Size

Sisällymistodennäköisyys on suhteessa alkion kokoon



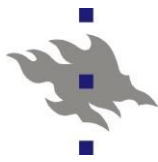
Tiivistelmä: Otantamenetelmät III

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

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

(**) Sisältymistodennäköisyys voi vaihdella alkioryhmittäin (ositettu otanta) tai alkiottain (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 sisällymistodennäköisyys on vakio n/N missä n on otoskoko ja N on perusjoukon alkioden lukumäärä
- Tekninen toteutus esimerkiksi satunnaislukujen avulla
- Erikoistapaukset:
 - SRSWOR
 - SRS **palauttamatta** (*without replacement*)
 - SRSWR
 - SRS **palauttaen** (*with replacement*)
- SAS Procedure SURVEYSELECT **Syntax**



Yksinkertainen satunnaisotanta

Perusjoukko: N alkiota

Perusjoukon tuntemattomat arvot $Y_1, Y_2, \dots, Y_k, \dots, Y_N$

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

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

Perusjoukon alkion k sisällymistodennäköisyys π_k

Otos: n alkiota

Otoksesta mitatut arvot $y_1, y_2, \dots, y_k, \dots, y_n$

Otosalkion k asetelmapaino w_k

Kokonaismäärän estimaattori \hat{t}

Keskiarvon estimaattori \bar{y}



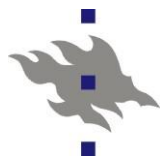
Perusjoukon *Population91* parametrit

UE91 totaali:

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

UE91 keskiarvo:

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



Yksinkertainen satunnaisotanta

Sisältymistodennäköisyys $\pi_k = n / N$ on vakio

Kokonaismäärän eli totaalin $T = \sum_{k=1}^N Y_k$ estimaattori

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

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

$$\hat{t} = \sum_{k=1}^n w_k y_k = \sum_{k=1}^n \frac{1}{\pi_k} y_k = \frac{N}{n} \sum_{k=1}^n y_k \quad \text{ja} \quad \bar{y} = \hat{t} / N$$

missä $w_k = N/n$ on asetelmapaino



Yksinkertainen satunnaisotanta

Totaalin varianssiestimaattori ja keskivirhe (SRSWOR)

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

$\text{s.e}_{SRS}(\hat{t}) = \sqrt{\hat{v}_{SRS}(\hat{t})}$ on 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

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

Keskiarvon varianssiestimaattori ja keskivirhe

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



SRSWOR-otanta perusjoukosta Population91

UE91 totaalin T estimaatti, varianssiestimaatti
ja 95 % luottamusväli:

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

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

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

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