Otantamenetelmät (78143)
Syksy 2008

Risto Lehtonen
risto.lehtonen@helsinki.fi

Otantamenetelmät

Luennoija: Prof. Risto Lehtonen

Luennot
Tiistaisin klo 14–18
4.11.–2.12.2008 (yhteensä 20 tuntia)
Exactum C323

Harjoitukset
Torstaisin klo 12–15
13.11.–4.12.2008 (yhteensä 12 tuntia)
Mikroluokka C128

Loppukuulustelu
Tiistai 2.12.2008 klo 14–16 Exactum C323

Oppikirja
Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second Edition. Chichester: John Wiley & Sons.

Web-materiaali
VLISS-virtual laboratory in survey sampling, http://mathstat.helsinki.fi/VLISS/

Otantamenetelmät

Soveltuva jatkokurssi
Pienalue-estimointi, kevät 2009

Ilmoittautuminen
Kurssille ilmoittaudutaan WebOodissa

Suoritustapa
Luentoja, seminaari-istuntoja ja käytännön harjoituksia (yht. 32 t)

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

Harjoituksissa käytetään tilastollisia ohjelmistoja (pääasiassa SAS, SPSS)

Laajuus
6/8 op

Tavoitteet

Kurssilla annetaan yleiskuva tilastollisista otantamenetelmistä ja niiden käytöstä eri tieteenalojen empiirisessä tutkimuksessa.

Esiteltäviä menetelmiä ovat yksinkertainen satunnaisotanta, systemaattinen otanta ja ositettu satunnaisotanta sekä vaativampina menetelminä ryväsotanta, moniasteinen otanta ja PPS-menetelmä (todennäköisyys suhteellinen kokoon) ja menetelmiin liittyvä piste-estimointi ja väliestimointi.

Muita käsiteltäviä aiheita ovat otoskoon määrittelyn perusteet ja lisäinformaation käyttö otannassa ja estimoinnissa.

Esimerkkejä annetaan aidoista tilanteista, ml. ihmisiä, kotitalouksia ja yrityksiä koskevat otosperusteiset tiedonkeruut ja tutkimukset.

Lisäksi tarkastellaan otantaan ja estimointiin soveltuvia tilastollisia ohjelmistoja.

Kurssi soveltuu tilastotieteen aine- tai syventäviä opintoja suorittaville opiskelijoille sekä myös yliopistoissa, korkeakouluissa ja tutkimuslaitoksissa toimiville jatko-opiskelijoille ja tutkijoille.

Kirjallisuutta

- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys. Second Edition*. Chichester: John Wiley & Sons.
- Pahkinen E. ja Lehtonen R. (1989). *Otanta-asetelmat ja tilastollinen analyysi*. Helsinki: Gaudeamus.

- **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:

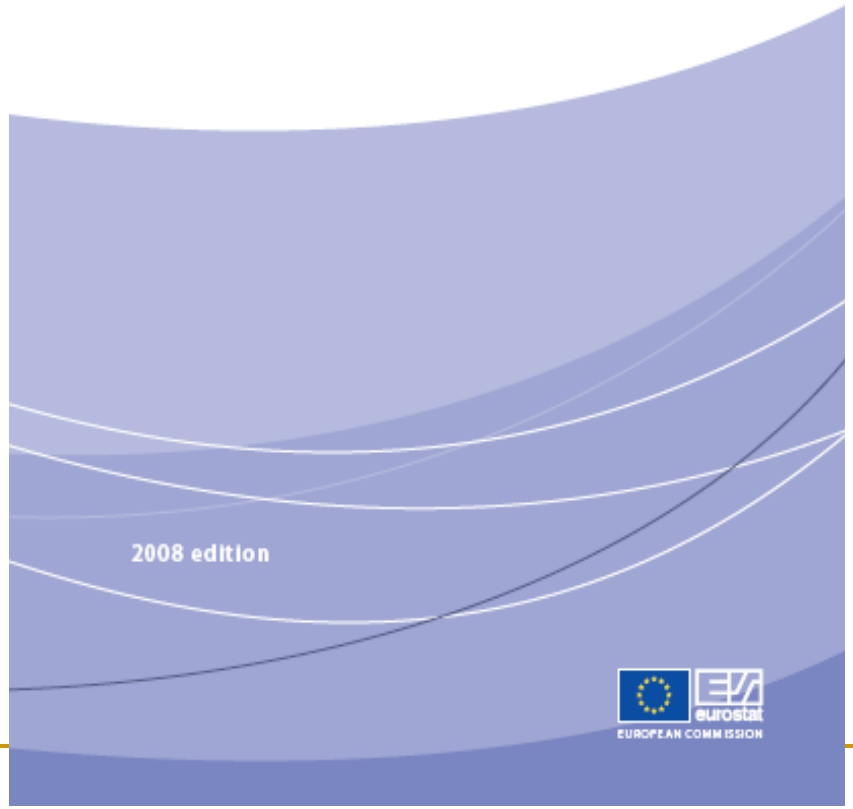
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF

ISSN 1977-0375

eurostat
Methodologies and
Working papers

Survey sampling reference guidelines

Introduction to sample design and estimation techniques



OSA I
Survey-prosessi

Empiirinen kvantitatiivinen tutkimusprosessi
Survey-prosessi

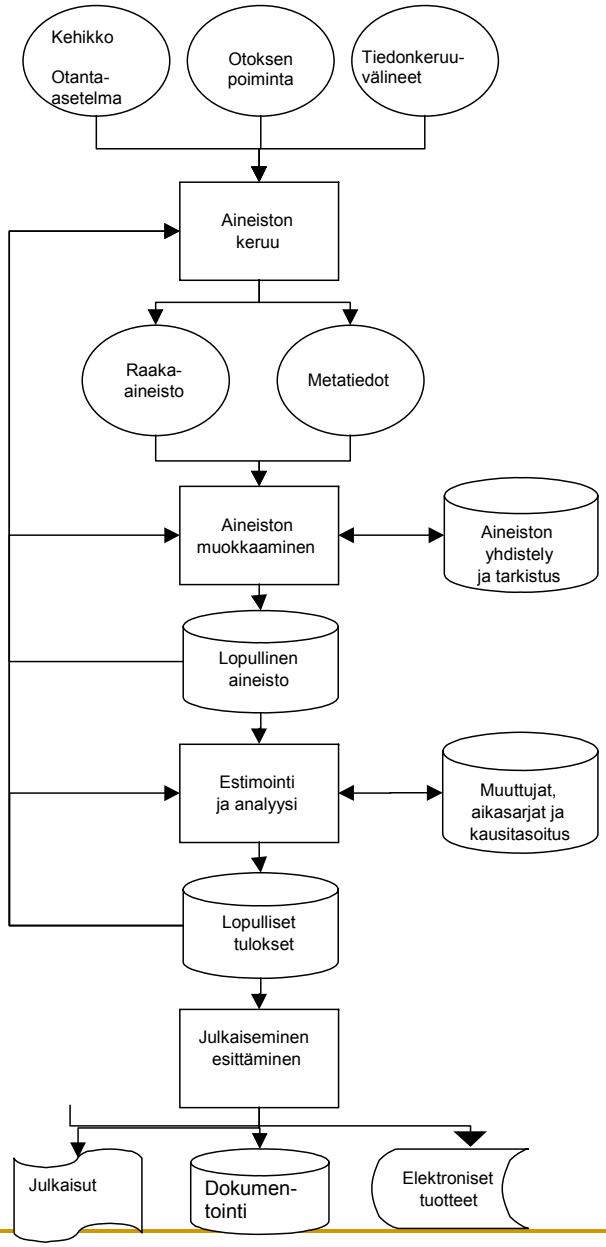
Survey = Empiiris-kvantitatiivinen (yhteiskunta)tutkimus

- **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.



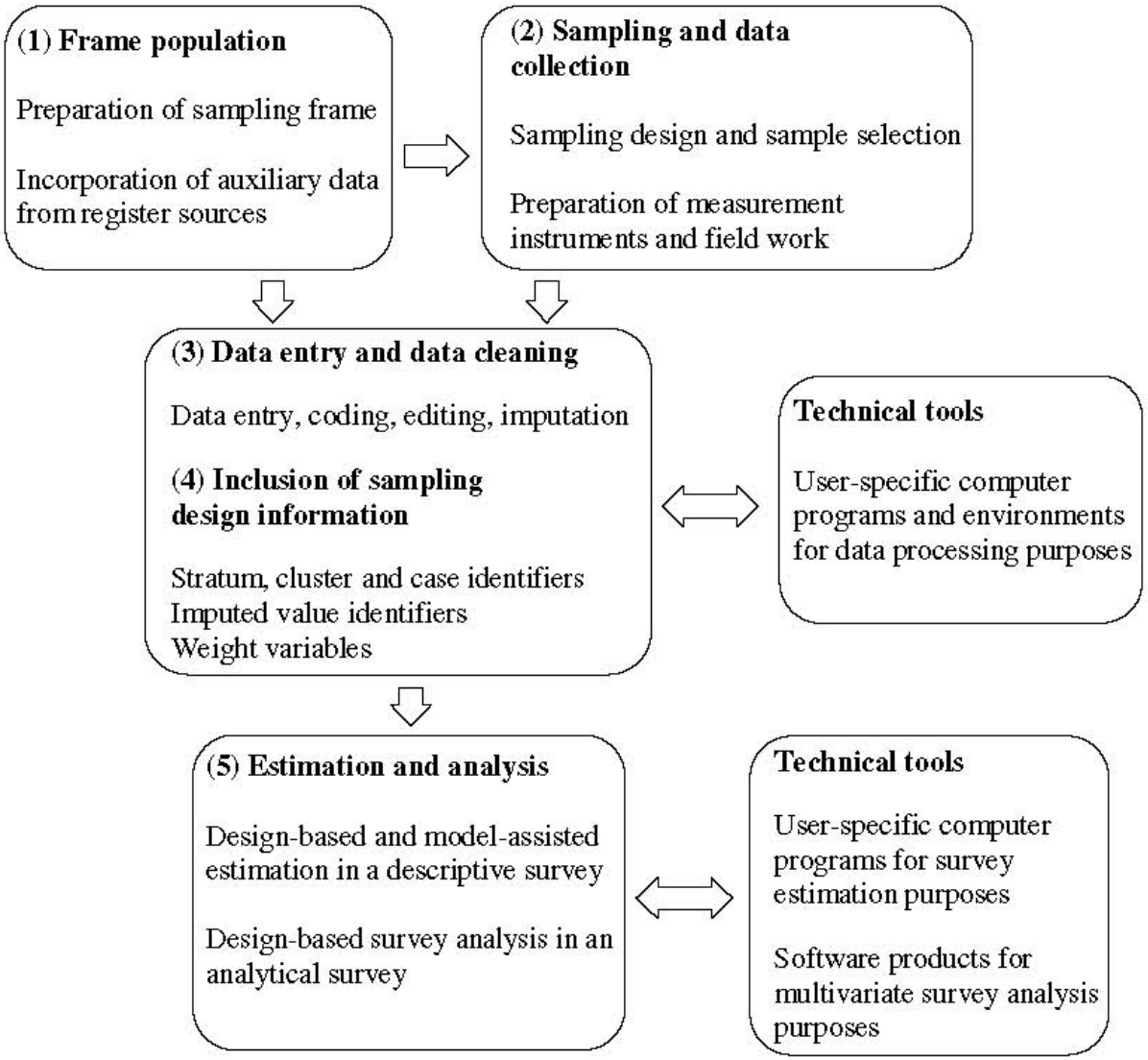


Figure 1.1 Flow chart for design-based estimation and analysis of complex survey data.

HY Otantamenetelmät syksy 2008 Risto Lehtonen

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) 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 **Kelan tutkimuksia ja selvityksiä** □ Terveysturvan väestötutkimukset □ Vanhempien kokemukset perhevapaiden käytöstä □ Kyselytutkimus Kelan etuuksista ja niiden toimeenpanosta □ Kela-barometri **Tilastokeskuksen tutkimuksia ja tilastoja** □ Työvoimatutkimus □ Kulutustutkimus **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 □ Kelan Mini-Suomi-terveystutkimus □ Tilastokeskuksen Tulonjakotutkimus □ EU:n European Community Household Panel ECHP □ EU SILC (Statistics on Income and Living Conditions)	

Kuvailevat ja analyyttiset otantatutkimukset

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

OSA II
Johdantoa:
Otanta ja estimointi, vastauskadon hallinta

Otanta-asetelmat (1)

- **Otanta-asetelma** (*sampling design*)
  - Niiden sääntöjen ja menetelmien kokonaisuus, jolla **otos** poimitaan määritellystä **perusjoukosta**
    - Tavoiteperusjoukko
    - Kohdeperusjoukko
    - Kehikkoperusjoukko
      - Ylipeitto
      - Alipeitto

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

$0 < \pi_k \leq 1$

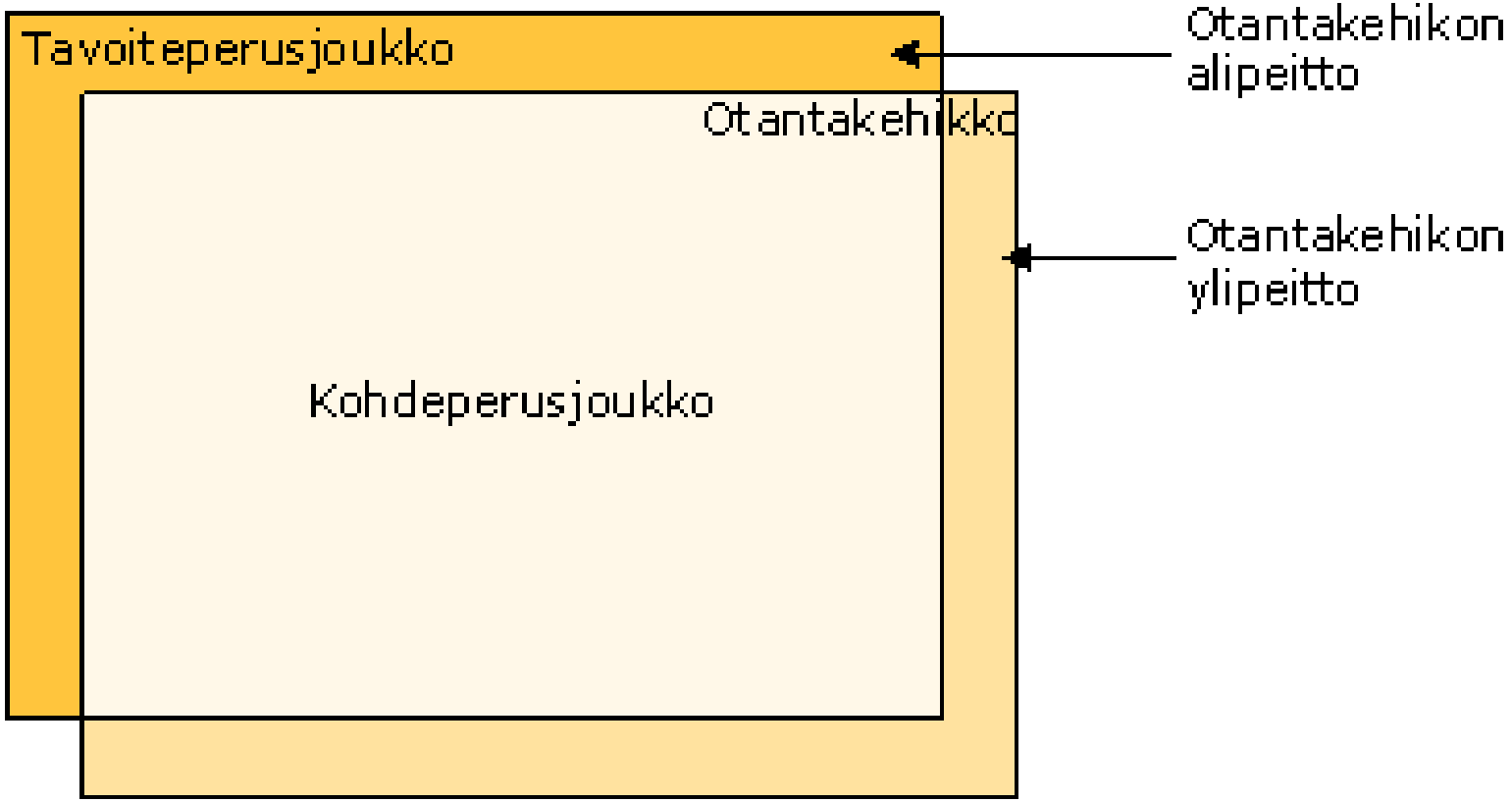
perusjoukon alkiolle k,

$k = 1,\ldots,N$

missä N on perusjoukon alkioiden lukumäärä

Otantakehikon alipeitto ja ylipeitto

Tilastokeskus: Laatua tilastoissa -käsikirja



VLISS-Virtual Laboratory in Survey Sampling

www.math.helsinki.fi/VLISS/

Province'91 perusjoukko (*population*) (entinen K-S lääni)
Tilastoyksikkönä (alkiona) kunta
N = 32 kuntaa
Tulosmuuttuja 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

- **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	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	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.

Otanta-asetelmat (2)

- Otos (*Sample*)
  - Perusjoukon osajoukko
  - Poimitaan jollain todennäköisyysotannan (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
    - Tilastolliset mallit

Otanta-asetelmat (3)

- Huomioita **sisältymis-todennäköisyydestä**
  - Nollaa suurempi
  - Voi olla = 1
    - Milloin?
  - Voi olla yhtäsuuri kaikille alkioille
  - Voi vaihdella
    - Alkioryhmittäin
      - Ositettu otanta
    - Alkioittain
      - PPS-otanta (otanta alkion kokoon suhteutetuin todennäköisyyksin
- Sisältymistodennäköisyyttä käytetään painokertoimien muodostamisessa
  - **Asetelmapaino** (design weight)
    - Totaalien estimointi
  - **Analyysipaino** (analysis weight
    - Muut analyysitilanteet
  - **Uudelleenpainotus**
    - Vastauskadon korjausta varten
      - Voidaan soveltaa sekä asetelmapainoon että analyysipainoon

Otanta-asetelmat (4)

- Huomioita asetelmapainosta

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

$k = 1,\ldots,n,$ 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

Esimerkki: Yksinkertainen satunnaisotanta SRS

SRS-otanta, $n = 8$ otosalkiota

Perusjoukossa $N = 32$ kuntaa

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

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

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

TRAINING KEY 28 Analysing an SRS sample

Otanta-asetelmat (5)

- 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

Otanta-asetelmat (6)

- 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: 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.

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 jaetaan kahteen pääryhmään:
  - **eräkatoon** (*item non-response*) ja
  - **yksikkökatoon** (*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 kokonaistutkimuksiin 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
  - Uudelleenpainotus-menetelmät
    - 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-asetelmat (7)

- **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 ositteisiin
      - Ositekohtaiset otoskoot
    - Alkioiden poimintamenetelmän valinta kullakin otanta-asteella ja ositteessa
      - Yksinkertainen satunnaisotanta SRS
      - Systemaattinen otanta
      - PPS-otanta

Otanta-asetelmat (8)

- **(1) Alkiotasoinen 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-asetelmat (9)

- Otanta-asetelma voi olla…
  - Yksinkertainen
    - Systemaattinen otanta
      - Poiminta suoraan alkiotason kehikkoperusjoukosta (rekisteristä, listasta…)
    - 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

Otanta-asetelmat (10)

- **Ryväsotannan motivaatio**
  - Tiedonkeruumenetelmän kannalta voi olla edullista käyttää ryväsotantaa
    - Käyntihaastattelut
      - Rypäänä kotitalous
    - Kliiniset menetelmät
      - Rypäänä terveyskeskus
  - Kehikkoperusjoukkojen huono saatavuus voi edellyttää ryväsotantaa
    - Koulusaavutus-tutkimukset
      - Pisa
  - **Tutkimusasetelma voi edellyttää ryväsotantaa**
    - **Terveys 2000**
  - **Ryväsotanta oman toiminta-alueesi näkökulmasta?**

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)

Tekninen yhteenveto I

- Peruskäsitteet
- Ks. Tekninen yhteenveto I
  - sivut 1-2

Yksinkertainen satunnaisotanta SRS (1)
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

Yksinkertainen satunnaisotanta (2)

Perusjoukko: *N* alkiota

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

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

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

Perusjoukon alkion *k* sisältymistodennäköisyys π_k

Otos: *n* alkiota

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

Otosalkion *k* asetelmapaino w_k

Kokonaismäärän estimaattori $\hat{t}$

Keskiarvon estimaattori $\bar{y}$

Yksinkertainen satunnaisotanta (3)

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 (4)

Totaalin varianssiestimaattori ja **keskivirhe (SRSWOR)**

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

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

Keskiarvon varianssiestimaattori ja **keskivirhe**

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

Tekninen yhteenveto I

- SRSWOR ja SRSWR
- Ks. Tekninen yhteenveto I
  - sivu 3
- Ks. VLISS
  - Training Key 28
    - Analysing an SRS sample

Bernoulli-otanta

- Katso *Survey sampling reference manual* (Lehtonen and Djerf 2008)
  - Sivu 17
  - Appendix 1

Bernoulli-otanta

- **Example.** *Bernoulli sampling* provides an example of an SRS-WOR type sampling scheme. In this method, the sample size is not fixed in advance but is a random variate whose expectation is n, the desired sample size. This property leads to a variation in the sample size with the expected value $N\pi$ and variance $N(1 - \pi)\pi$, where π stands for the inclusion probability. The randomness in the sample size is relatively unimportant in large samples.

Bernoulli-otanta

- Let us briefly introduce the technique. To carry out Bernoulli sampling, we need to carry out the following steps:
- Step 1. Fix the value of the inclusion probability π, where $0 < \pi < 1$, so that the expected sample size will be $N\pi$, the product of the population size and the inclusion probability. If the desired sample size is n, then $\pi = n/N$.

Bernoulli-otanta

- Step 2. Append three variables, let say PROB, IND and UNI, to the sampling frame data set. PROB is set equal to the chosen value of π, and IND is set to zero, for all N population elements. For UNI, a value from a uniform distribution over the range (0, 1) is drawn independently for each population element, starting from the first element. A pseudo random number generator can be used in generating the random numbers.

Bernoulli-otanta

- Step 3. The decision rule for inclusion of a population element in the sample is the following. The *k*th population element is included in the sample if UNI < π, and correspondingly, we set IND = 1 for the selected element (otherwise, the value of IND remains zero).

Bernoulli-otanta

- Step 4. Treat all population elements sequentially by using Step 3.
- When Steps 1 to 4 are completed, the sum of IND over the sampling frame appears to be close (or, equal) to the desired sample size *n*. The elements having IND = 1 constitute the Bernoulli sample. The procedure can be easily programmed for example with Excel, SAS or SPSS.
  - Appendix 1. contains a short example of Bernoulli sampling.

Bernoulli-otanta

- **Appendix 1**. Example of sample selection using Bernoulli sampling
- We create a sampling frame consisting 2000 elements and want to select about 200 units to the sample.
- Sampling fraction UNI = 200/2000 = 0.1.
- All elements in the frame are assigned a pseudo random number from Uniform distribution, PI.
- Those elements with PI ≤ 0.1 are selected and selection indicator IND is given value 1.
- If the unit was not selected, IND is set 0.

Bernoulli-otanta – SAS-koodi

```
Data Bernoulli;
UNI=200/2000;
Do I=1 to 2000;
  PI=Ranuni(0);
  If PI<=UNI then IND=1;
  Else IND=0;
  Output;
End;
Proc Print;
Sum IND;
Run;
```

I	UNI	PI	IND
1	0.1	0.83976	0
2	0.1	0.50375	0
3	0.1	0.08013	1
4	0.1	0.87756	0
5	0.1	0.13501	0
6	0.1	0.41416	0
7	0.1	0.10639	0
8	0.1	0.28283	0
9	0.1	0.16496	0
10	0.1	0.88332	0
...			
1991	0.1	0.67351	0
1992	0.1	0.11558	0
1993	0.1	0.78235	0
1994	0.1	0.66004	0
1995	0.1	0.08314	1
1996	0.1	0.19041	0
1997	0.1	0.77828	0
1998	0.1	0.07666	1
1999	0.1	0.53644	0
2000	0.1	0.35678	0
Sum			201

Bernoulli-otanta

- In Bernoulli sampling the sample size is a random quantity and this example shows that we received one unit too much.
  - The simplest way to obtain a fixed sample size would be to sort the frame by the random number and select exactly 200 cases (from the beginning, end or just at any point as long as the random numbers are used for selection).

Systemaattinen otanta SYS (1)
Systematic sampling

- **Poimintamenettely**
  - a) Määrää poimintaväli
    $q = N/n$
  - b) Valitse satunnaisesti ensimmäinen otokseen poimittava alkio väliltä [1,q]
  - c) Poimi ensimmäisestä poimitusta lähtien joka q:s alkio.
    Saadaan n alkion otos
- SYS-poiminta on teknisesti helppo toteuttaa esim. numeroidusta kehikkoperusjoukosta manuaalisesti tai koneellisesti atk-rekisteristä
- SYS-otantaa käytetään usein päämenettelynä poimittaessa alkiotason otoksia atk-rekistereistä
  - SAS Procedure SURVEYSELECT

Systemaattinen otanta SYS (2)

- **SYS-otannan erikoistapauksia:**
  - Satunnaisjärjestyksessä oleva perusjoukko
    - estimointi palautuu SRSWOR-tilanteeseen
  - Implisiittisesti ositettu perusjoukko
    - perusjoukon alkiot on lajiteltu tiettyjen kriteerien mukaan ennen poimintaa
    - estimointi palautuu ositetun otannan tilanteeseen
    - käytännössä usein sovellettu menetelmä

Systemaattinen otanta (3)

Sisältymistodennäköisyys $\pi_k = n/N$ on vakio

Kokonaismäärän eli totaalin T estimaattori

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

Keskiarvon $\bar{Y}$ estimaattori

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

Tekninen yhteenveto I

- SYS-otanta
- Sisäkorrelaatio (intra-class correlation)
- Ks. Tekninen yhteenveto I
  - sivu 3
  - Ote Lehtonen-Pahkinen (2004)
- Ks. VLISS
  - Training Key 45
    - Intra-class correlation

Lisätiedon käyttö otannan ja estimoinnin yhteydessä (1)

- 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.
  - 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 käyttö… (2)

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

Lisätiedon käyttö… (3)

- **B. Lisätiedon käyttö estimointiasetelmassa**
  - Tavoitteena estimoinnin tehostaminen poimitulle otokselle
    - keskivirheiden pienentäminen käytetyn otanta-asetelman puitteissa
  - Regressioestimointi
  - Suhde-estimointi
  - Kalibrointimenetelmät
  - Jälkiosittaminen…
- SAS Proc SURVEYMEANS
- Sas Macro CLAN

Lisätiedon käyttö… (4)

- Strategia
  - Otanta-asetelman ja estimointiasetelman yhdistelmä
- Ks. Lehtonen-Pahkinen (2004)
  - Strategian määritelmä
  - Laajennettu asetelmakertoimen määritelmä
    - Strategian deff
  - Table 3.12

Strategia (1)

- Lehtonen - Pahkinen (2004) pp. 88-89:
- The concept of *estimation strategy* will be used referring to a combination of the sampling design and the appropriate estimator.
  - The model-assisted strategies to be discussed are shown in Table 3.12.
  - In the design-based reference strategies, no auxiliary information is used.
  - **Tehtävä:** Miten liittäisit taulukkoon 3.12 strategiat, joissa lisäinformaatio sisällytetään otanta-asetelmaan?

Strategia (2)

Table 3.12 Estimation strategies for population total

Strategy	*Auxiliary information*		*Assisting model*
	Design-based strategies		
SRSWOR	Not used		None
SRSWR	Not used		None
	Model-assisted strategies		
Poststratification	SRS*pos	Discrete	ANOVA
Ratio estimation	SRS*rat	Continuous	Regression (no intercept)
Regression estimation	SRS*reg	Continuous	Regression

Ositettu otanta STR *Stratified sampling* (1)

- **Tavoite**
  - Tehokas otanta-asetelma muodostamalla perusjoukon alkioista ennen otoksen poimintaa tutkittavan ilmiön kannalta sisäisesti homogeenisia, toisensa poissulkevia ositteita (*stratum; strata*)
  - Ositteet ovat toisistaan riippumattomia osaperusjoukkoja
  - Kullekin ositteelle voidaan tarvittaessa kiinnittää oma otanta-asetelma
    - Joustavuusperiaate

Ositettu otanta STR *Stratified sampling* (2)

- **Työvaiheet:**
  - **(1) Osituskriteerien valinta**
    - alueelliset ositteet, esim. lääni
    - demografiset ositteet, ikäryhmän ja sukupuolen mukaan (henkilöotokset)
    - toimialan mukaiset ositteet (yritysotannat)
  - **(2) Kehikkoperusjoukon osittaminen ositteisiin**
    - kunkin kehikkoperusjoukon alkion kiinnittäminen yhteen (ja vain yhteen) ositteeseen
  - **(3) Otoksen kiintiöinti ositteisiin**
    - määritellään kustakin ositteesta poimittavien alkioiden lukumäärä niin, että kokonaisotoskoko on *n*
  - **(4) Otoksen poiminta kustakin ositteesta**
    - kustakin ositteesta poimitaan valitulla otantamenetelmällä (esim. SYS) alkiotason otos valitun kiintiöintimenetelmän mukaisesti
  - SAS Procedure SURVEYSELECT

Ositettu otanta STR *Stratified sampling* (3)

Otoksen kiintiöinti ositteisiin

- **Suhteellinen kiintiöinti**
  - kustakin ositteesta poimitaan alkioita otokseen ositteen suhteellista osuutta koko perusjoukossa vastaava määrä
  - sisältymistodennäköisyys on vakio *n/N*
- **Tasakiintiöinti**
  - kustakin ositteesta poimitaan yhtä monta otosalkiota
  - sisältymistodennäköisyydet vaihtelevat ositteittain

- **Optimaalinen kiintiöinti**
  - suurista ositteista ja ositteista, joissa on suuri variaatio, poimitaan suhteessa enemmän alkioita kuin pienistä ositteista ja ositteista, joissa on pieni variaatio
  - sisältymistodennäköisyydet vaihtelevat ositteittain

- HUOM: Ositusmuuttujien arvot tulee olla tiedossa kaikille perusjoukon alkioille ennen otoksen poimintaa

Tekninen yhteenveto I

- Ositettu otanta STR
- Ks. Tekninen yhteenveto I
  - Sivut 5-6
- Ks. Ote Lehtonen-Pahkinen (2004)
  - Stratified sampling
- Ks. VLISS
  - Training Key 63
    - Design effect and allocation under stratified sampling

PPS-otanta (1)

- PPS: *Selection with probabilities proportional to size*
  - Poiminta otosyksiköiden koon mukaisin todennäköisyyksin
  - Perusjoukon alkion otokseen sisältymis-todennäköisyys riippuu alkion kokoa mittaavan muuttujan *z* arvosta
  - **Kokoa mittaavan muuttujan arvo tulee olla tiedossa kaikilta perusjoukon alkioilta ennen poimintaa**
  - Käytetään usein yritysotannoissa
  - Muita esimerkkejä
    - PISA-tutkimukset

PPS-otanta (2)

- PPS-otanta on erittäin tehokas menetelmä, jos kokoa mittaava muuttuja korreloi voimakkaasti tulosmuuttujan kanssa
- PPS-otoksien poiminta
  - SAS Procedure SURVEYSELECT

Sisältymistodennäköisyys

$$\pi_k = n \times z_k \,/\, \sum_{k=1}^{N} z_k$$

missä z_k on kokomuuttujan arvo perusjoukon alkiolle *k*

Tekninen yhteenveto I

- PPS-otanta
- Ks. Tekninen yhteenveto I
  - Sivut 7-10
- Ks. Ote Lehtonen-Pahkinen (2004)
  - PPS sampling
- Ks. VLISS
  - Training Key 54
    - The effective use of auxiliary information in PPS sampling

Otoksien poiminta käytännössä

- **SAS Procedure SURVEYSELECT**
  - SRS – yksinkertainen satunnaisotanta
    - SRSWOR -palauttamatta
    - SRSWR - palauttaen
  - SYS- systemaattinen otanta
  - STR – ositettu otanta
  - PPS-otanta
    - PPSWOR
    - PPSWR
    - PPSSYS
    - Ositettu PPS

ESIMERKKI (1): Otoksen poiminta Province-perusjoukosta

- **SAS Procedure SURVEYSELECT**
  - Ositettu SRSWOR, *n*=8
    - 2 ositetta (kaupungit/muut kunnat)
    - Tasakiintiöinti (Equal allocation)

```
proc surveyselect
   data=province91
   out=otos
   sampsize=(4,4)
   seed=9876543
   method=srs stats;
   strata kumu;
run;
```

ESIM. (2) Kokonaismäärän estimointi

- Työttömien kokonaismäärän UE91 estimointi äsken poimitusta ositetusta SRSWOR-otoksesta:

```
proc surveymeans data=otos N=32 sum;
  strata kumu;
  weight SamplingWeight;
  var UE91;
  run;
```

HUOM: SURVEYSELECT tuottaa painomuuttujan SamplingWeight arvot otostiedostoon kaikissa otantamenetelmissä

ESIM. (3): Poimittu SRSWOR-otos

(1) SRSWOR-otos / n=8 kuntaa

Obs	ID	LABEL	UE91	SamplingWeight
1	1	Jyvaskyla	4123	4
2	4	Keuruu	760	4
3	5	Saarijarvi	721	4
4	15	Konginkangas	142	4
5	18	Kuhmoinen	187	4
6	26	Pihtipudas	331	4
7	30	Toivakka	127	4
8	31	Uurainen	219	4
Sum				32

ESIM. (4): Totaaliestimaatti ja keskivirhe

Statistics

Variable	Sum	Std Dev
UE91	26440	13282

- Sum = Estimoitu totaali
- Std Dev = totaaliestimaatin keskivirheen estimaatti

ESIM. (5): Totaaliestimaatti ja keskivirhe

Horvitz-Thompson-estimaattori

SRSWOR-otokselle

$$\hat{t} = \sum_{k=1}^{n} w_k y_k = 26440$$

missä $w_k = 1/\pi = 4$ on asetelmapaino

$s.e(\hat{t}) = 13282$ mikä on kovin suuri!

HUOM: Perusjoukossa $t = 15098$

ESIM. (6): Huomioita SRSWOR-otoksesta

- Tulosmuuttujan UE91 jakauma vino
- Muutama suuri arvo
  - Jyväskylä
  - Jkl mlk
- Estimaatin arvo riippuu vahvasti siitä, ovatko suuret kunnat mukana otoksessa
  - Kyllä: Suuri estimaatti
  - Ei: Pieni estimaatti…
  - Katsotaan tarkemmin PC-demoissa
- Parempi estimointi
  - Ositettu otanta
    - Kaupungit
    - Muut kunnat
  - PPS-otanta
    - Käytetään otannassa kokomuuttujaa
    - Tässä HOU85
      - Väestölaskennasta saatu kotitalouksien lukumäärä kussakin perusjoukon kunnassa

ESIM. (7): Ositettu SRSWOR-otos

(2a) Oma ositettu SRSWOR-otos / n=8 kuntaa

Obs	ID	kumu	LABEL	UE91	Sampling Weight
1	1	1	Jyvaskyla	4123	1.75
2	2	1	Jamsa	666	1.75
3	3	1	Jamsankoski	528	1.75
4	5	1	Saarijarvi	721	1.75
5	11	2	Kannonkoski	153	6.25
6	18	2	Kuhmoinen	187	6.25
7	19	2	Kyyjarvi	94	6.25
8	27	2	Pylkonmaki	98	6.25

Statistics

Variable	Sum	Std Dev
UE91	13892	4029.562414

ESIM. (8): Ositettu PPSWOR-otos

(3a) Oma PPSWOR-otos / n=8 kuntaa

Obs	osite	ID	LABEL	UE91	HOU85	Sampling Weight
1	1	1	Jyvaskyla	4123	26881	1.00000
2	2	9	Joutsa	194	1823	5.08361
3	2	3	Jamsankoski	528	3019	3.06970
4	2	5	Saarijarvi	721	3730	2.48457
5	2	7	Aanekoski	767	4264	2.17341
6	2	2	Jamsa	666	4663	1.98744
7	2	4	Keuruu	760	4896	1.89286
8	2	10	Jyvaskmlk	1623	9230	1.00406

Statistics

Variable	Sum	Std Dev
UE91	14580	635.260481

ESIM. (9): Yhteenveto estimointituloksista, $n = 8$

Otos	Totaali	s.e	
SRSWOR	26440	13282	(huonoin)
STR	13892	4029	
PPSWOR	14580	635	(paras)
Perusjoukossa	15098	0	

VLISS – PPS sampling (1)

- VLISS Training Key 54
- Simulation experiment
  - Measures of performance
    - Monte Carlo mean and standard error
    - Bias, ARB (absolute relative bias)
    - RMSE (Root mean squared error)
  - Size variables

    1) HOU85 (number of households in a municipality)

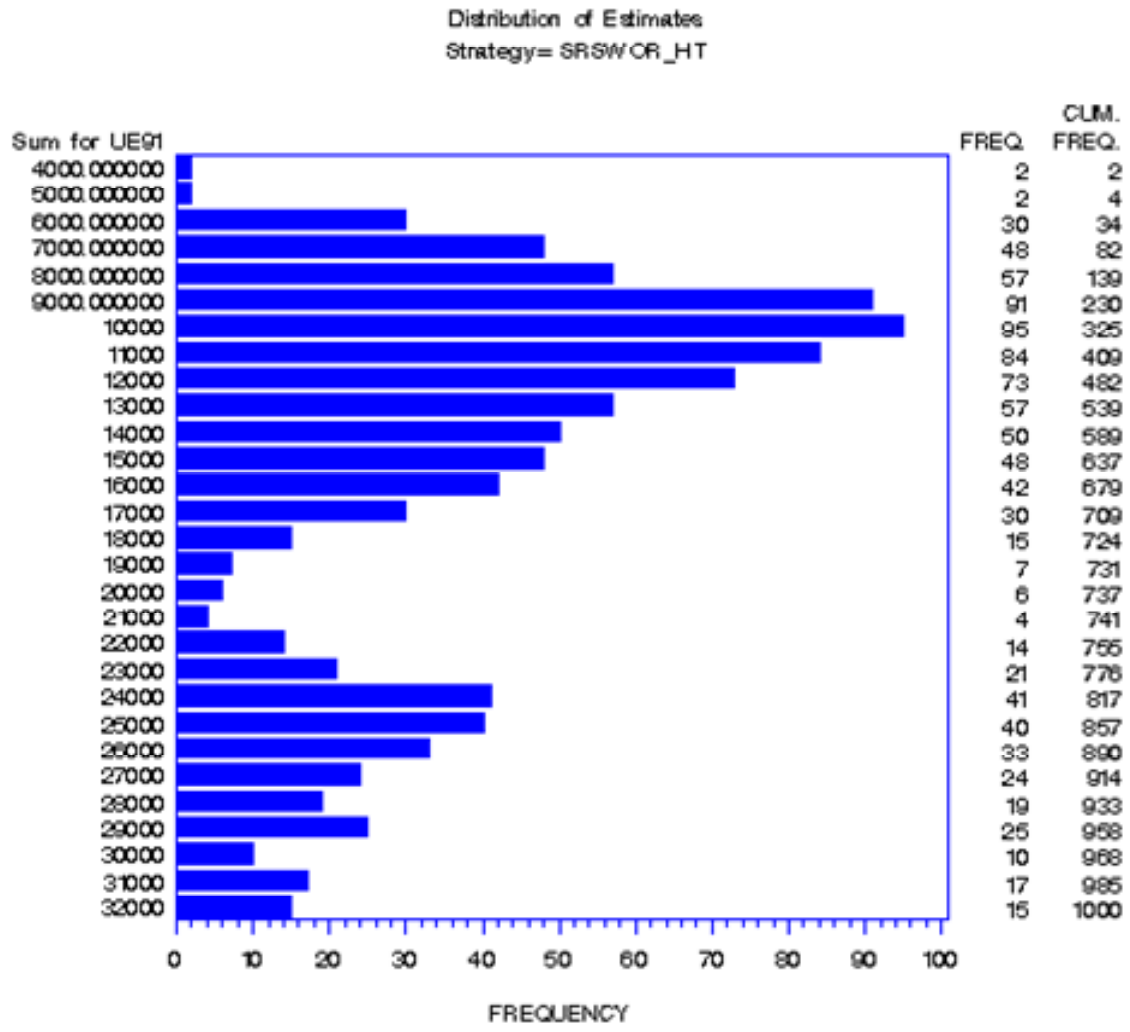
    2) Z (artificially created variable for pedagogical purposes, N(500,150))

    3) X (artificially created variable for pedagogical purposes, UE91 + 3000)

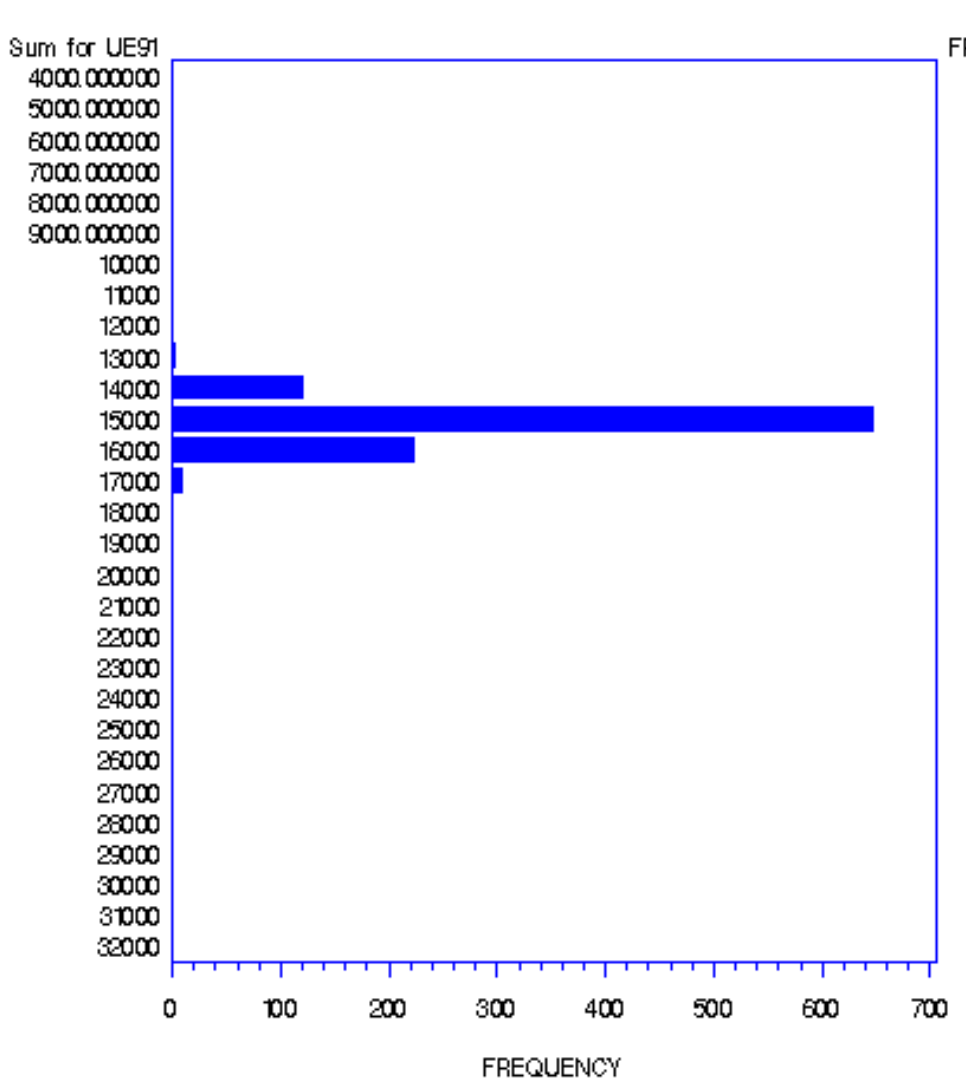
VLISS – PPS sampling (2)

Strategy	Population Total	Mean of Estimates	Bias	ARB	Standard error	Root MSE
SRSWOR_HT	15098	15360.5	262.5	1.74	7325.4	7330.1
PPS_HOU85	15098	15138.2	40.2	0.27	559.2	560.7
PPS_x	15098	15276.2	178.2	1.18	4609.6	4613.1
PPS_z	15098	15025.3	-72.7	0.48	7564.0	7564.4

Distribution of SRSWOR_HT estimator

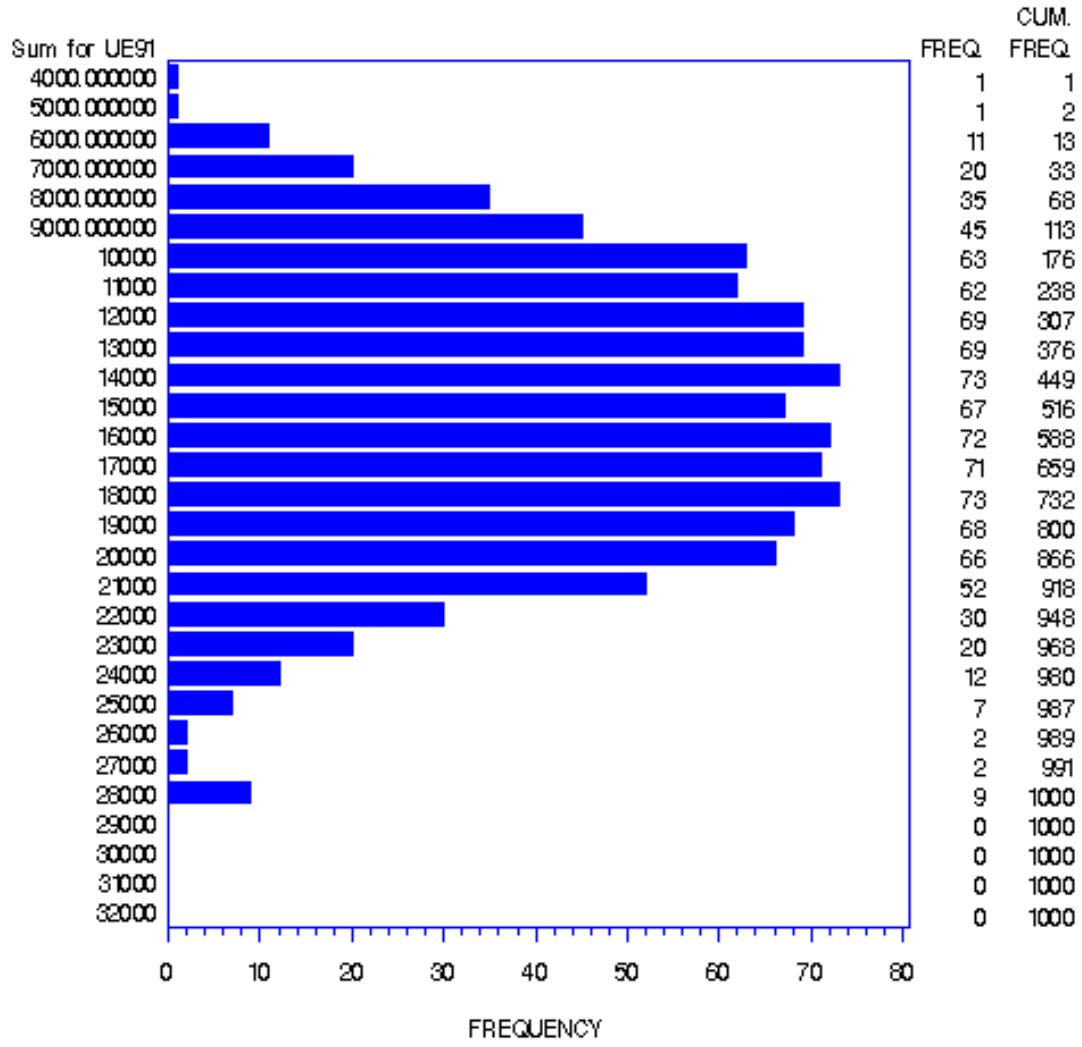


Distribution of PPSWOR estimator (aux.var. HOU85)



	FREQ.	CUM. FREQ.
4000.000000	0	0
5000.000000	0	0
6000.000000	0	0
7000.000000	0	0
8000.000000	0	0
9000.000000	0	0
10000	0	0
11000	0	0
12000	0	0
13000	1	1
14000	120	121
15000	647	768
16000	223	991
17000	9	1000
18000	0	1000
19000	0	1000
20000	0	1000
21000	0	1000
22000	0	1000
23000	0	1000
24000	0	1000
25000	0	1000
26000	0	1000
27000	0	1000
28000	0	1000
29000	0	1000
30000	0	1000
31000	0	1000
32000	0	1000

Distribution of PPSWOR estimator (aux.var. X)



	FREQ	CUM. FREQ
4000.000000	1	1
5000.000000	1	2
6000.000000	11	13
7000.000000	20	33
8000.000000	35	68
9000.000000	45	113
10000	63	176
11000	62	238
12000	69	307
13000	69	376
14000	73	449
15000	67	516
16000	72	588
17000	71	659
18000	73	732
19000	68	800
20000	66	866
21000	52	918
22000	30	948
23000	20	968
24000	12	980
25000	7	987
26000	2	989
27000	2	991
28000	9	1000
29000	0	1000
30000	0	1000
31000	0	1000
32000	0	1000

Distribution of PPSWOR estimator (aux.var. Z)