HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otantamenetelmät
(78143)
Syksy 2009

TEEMA 1

Risto Lehtonen
risto.lehtonen@helsinki.fi

Otantamenetelmät

Luennot
Tiistaisin klo 14–18
3.11.–1.12.2008 (yhteensä 20 tuntia), Exactum C323

Harjoitukset
Torstaisin klo 12–15
5.11.–3.12.2008 (yhteensä 15 tuntia), Mikroluokka C128
Harjoituksissa käytetään tilastollisia ohjelmistoja (pääasiassa SAS, SPSS)

Loppukuulustelu
Tiistai 8.12.2008 klo 14–16 Exactum C323

Suoritustapa
Luentoja ja käytännön harjoituksia (yht. 35 t)

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

Ilmoittautuminen
Kurssille ilmoittaudutaan WebOodissa

Harjoitustyö

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

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 mm. sisältyy valtiotieteellisen tiedekunnan ns. menetelmäkoriin
- Soveltuvia jatkokursseja
  - Otanta-aineistojen analyysi, kevät 2010
  - (Pienalue-estimointi, kevät 2011)

Tavoitteet

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

Tavoitteet

- Otannan perusmenetelmät
  - yksinkertainen satunnaisotanta
  - systemaattinen otanta
- Lisäinformaation käyttö otanta-asetelmassa
  - PPS-otanta (*sampling with probability-proportional-to-size method*)
  - ositettu otanta
- SAS-proseduuri SURVEYSELECT
- SAS-proseduuri SURVEYMEANS

Tavoitteet

- Lisäinformaation käyttö estimointiasetelmassa
  - suhde-estimointi
  - regressioestimointi
  - yleistetty regressioestimointi (kalibrointi)
- SAS-proseduuri SURVEYREG

Tavoitteet

- Esimerkkejä otosperusteisista tiedonkeruista ja tutkimuksista
  - Pisa
  - Terveys 2000
  - Tilastokeskuksen työvoimatutkimus
  - EU SILC
- Otantaan ja estimointiin soveltuvat tilastolliset ohjelmistot
  - SAS - SURVEY-proseduurit
  - SPSS - Complex samples -moduli
  - Stata - svy-ohjelmat

Kirjallisuutta

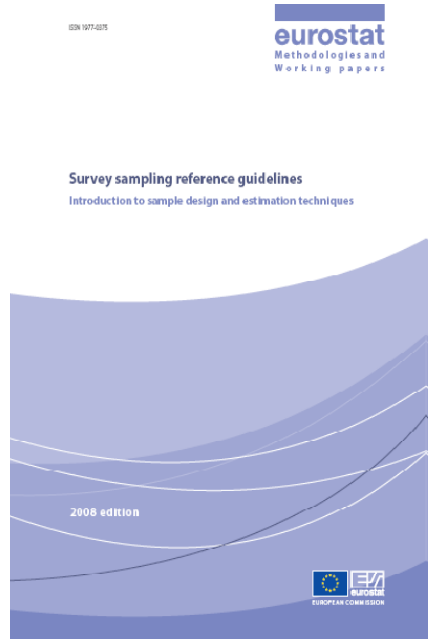
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys. Second Edition*. Chichester: John Wiley & Sons.
- Pahkinen E. ja Lehtonen R. (1989). *Otanta-asetelmat ja tilastollinen analyysi*. Helsinki: Gaudeamus.
- **Web extension:**
- VLISS-Virtual Laboratory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

- 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 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 - 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

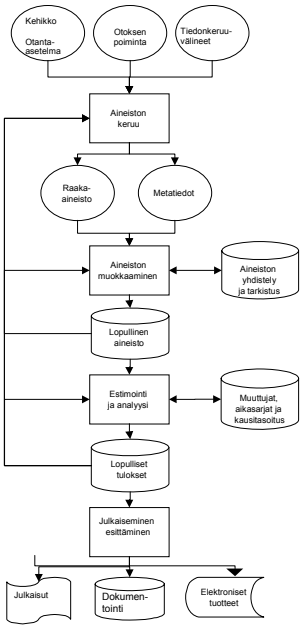
II Tiedonkeruuoperaatiot
6. Otoksen poiminta
7. Tiedonkeruu
8. Tiedostonmuodostus

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.



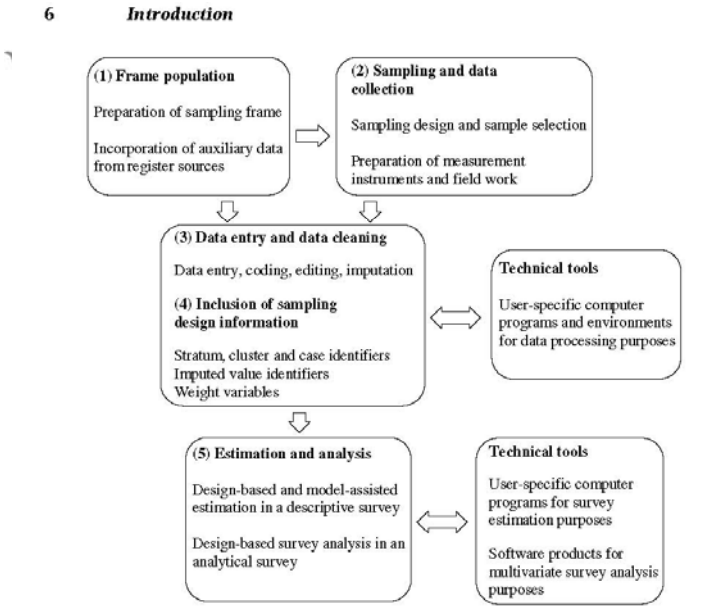


Figure 1.1 Flow chart for design-based estimation and analysis of complex survey data.

HY Otantamenetelmät syksy 2009 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)	

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*
- *Self-selection web survey*
- 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..."

Nettikysely… ja vaihtoehdot?

- Millaisia yleistyksiä nettikyselyn perusteella voidaan tehdä?
  - *How accurate are self-selection web surveys?* Jelke Bethlehem, Statistics Netherlands, Discussion paper (08014)

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

- Millaisia vaihtoehtoja nettikyselylle voisi olla?

Kuvaileva ja analyyttinen survey

HY Otantamenetelmät syksy 2009

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 2009

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

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
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,

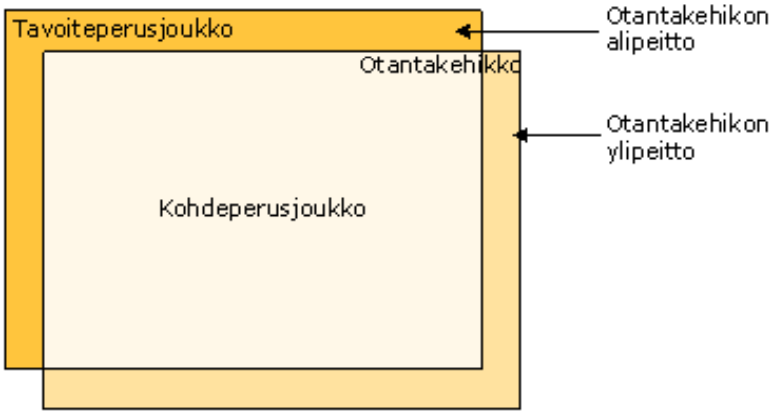
$k = 1,\ldots,N$

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

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
- 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	33 786	67 200	26 881
2	Jämsä	1	2	11.07	666	6016	12 907	4663
3	Jämsänkoski	1	2	13.83	528	3818	8118	3019
4	Keuruu	1	2	12.84	760	5919	12 707	4896
5	Saarijärvi	1	3	14.62	721	4930	10 774	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	13 727	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.

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

VLISS-Virtual Laboratory in Survey Sampling

- 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

- www.math.helsinki.fi/VLISS/

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: Terveys 2000 -tutkimus

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
  - **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-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 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 osittеisiin. 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

Otannan perusmenetelmät

- 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

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 $\bar{y}$

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)**

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

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

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

SRSWOR-otanta perusjoukosta Population91

UE91 totaalin T estimaatti, varianssiestimaatti ja 95 % luottamusväli:

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

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

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

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

Tekninen yhteenveto I

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

Bernoulli-otanta

- Katso *Survey sampling reference manual* (Lehtonen and Djerf 2008)
  - Sivu 17
  - Appendix 1
  - Vähän muokattu tätä esitystä varten

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 PI, IND and UNI, to the sampling frame data set. PI 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 kth population element is included in the sample if UNI ≤ PI, 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 HUOM

- **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 PI = 200/2000 = 0.1.
- All elements in the frame are assigned a pseudo random number from Uniform distribution, UNI.
- Those elements with UNI ≤ 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;
PI=200/2000;
do i=1 to 2000;
   UNI=Ranuni(0);
   if UNI le PI then IND=1;
   else IND=0;
   output;
end;
proc print data=Bernoulli;
sum IND;
run;
```

SAS-toteutus

I	PI	UNI	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
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

Systemaattinen otanta SYS

- 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
  - Useita mahdollisia tapoja käytännön toteutuksessa Syntax

Systemaattinen otanta SYS

- **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 (STR) tilanteeseen
  - käytännössä usein sovellettu menetelmä

Systemaattinen otanta

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

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

Systemaattinen otanta

Asetelmavarianssi

$$V_{sys}(\hat{t}) = \sum_{j=1}^{q} (\hat{t}_j - T)^2 / q = V_{SRS}(\hat{t})(1 + (n-1)\rho_{\text{int}}) = N \times SSB,$$

missä $\hat{t}_j$ on j:nnen systemaattisen otoksen kokonaismäärän estimaattori

$q=N/n$ on poimintaväli

$$\rho_{\text{int}} = 1 - \frac{n}{n-1} \times \frac{SSW}{SST}$$

on sisäkorrelaatiokerroin, missä käytetään ANOVA-neliösummahajoitelmaa $SST = SSW + SSB$.

Systemaattinen otanta

Asetelmakerroin (parametri)

$$DEFF_{sys}(\hat{t}) = \frac{V_{sys}(\hat{t})}{V_{srs}(\hat{t})} = 1 + (n-1)\rho_{\text{int}}$$

Systemaattinen otanta on yksinkertaiseen satunnaisotantaan verrattuna:
- tehokkaampi, jos $-1/(n-1) < \rho_{\text{int}} < 0$,
- yhtä tehokas, jos $\rho_{\text{int}} = 0$,
- tehottomampi, jos $0 < \rho_{\text{int}} < 1$

Systemaattinen otanta

- **Varianssiestimaattori**
  - Asetelmavarianssille ei ole analyyttista estimaattoria - Miksi?
  - Asetelmavarianssin estimointi vaatii approksimaatioita
  - Estimointi kuten SRS, jos oletetaan, että kyseessä on **satunnaisjärjestyksessä** oleva perusjoukko (jolloin sisäkorrelaatio = 0)
  - Estimointi kuten STR (ositettu otanta, suhteellinen kiintiöinti), jos oletetaan **implisiittinen** ositus
    - Perusjoukon alkioiden lajittelu ennen SYS-poimintaa

Systemaattinen otanta

- Ks. Tekninen yhteenveto I sivu 4
- VLISS Training Key 45
  - In this exercise the intra-class correlation is negative (-0.08) which means that given the current sorting order of the population, systematic sampling will be more efficient than simple random sampling without replacement (SRSWOR). Note that the population was pre-sorted first by the variable URB85 (urbanicity) and within each URB85 class, by the variable Municipality for this exercise.

Lisäinfon käyttö otanta-asetelmassa

- Ositettu otanta
  STR - *Stratified sampling*
- PPS-otanta
  PPS: *Selection with probabilities proportional to size*

Ositettu otanta STR
Stratified sampling

- **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 Työvaiheet

(1) Osituskriteerien valinta
- alueelliset ositteet demografiset ositteet
- toimialan mukaiset ositteet (yritysotannat)

(2) Kehikkoperusjoukon osittaminen
- kunkin kehikkoperusjoukon alkion kiinnittäminen yhteen (ja vain yhteen) ositteeseen

(3) Otoksen kiintiöinti
- määritellään kustakin ositteesta poimittavien alkioiden lukumäärä niin, että kokonaisotoskoko on n

(4) Otoksen poiminta kustakin ositteesta
- kustakin ositteesta poimitaan valitulla otantamenetelmällä alkiotason otos valitun kiintiöintimenetelmän mukaisesti

Ositettu otanta STR Kiintiöinti

- **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
  - sisältymistn vaihtelee ositteittain
- **Optimaalinen kiintiöinti**
  - suurista ositteista ja ositteista, joissa on suuri variaatio, poimitaan suhteessa enemmän alkioita kuin pienistä ositteista ja ositteista, joissa on pieni variaatio
  - sisältymistn vaihtelee ositteittain
- HUOM: Ositusmuuttujien arvot tulee olla tiedossa kaikille perusjoukon alkioille ennen otoksen poimintaa

STR: Tekninen yhteenveto I ja VLISS

- Ositettu otanta STR
- Ks. Tekninen yhteenveto I
  - Sivut 5-6
- Asetelmakerroin *deff*
  - Lehtonen-Pahkinen (2004) s. 62-63
- Ks. VLISS
  - Training Key 63
    - Design effect and allocation under stratified sampling

PPS-otanta

- 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 pj:n alkioilta ennen poimintaa**
  - Käytetään usein yritysotannoissa
  - Muita esimerkkejä: Kouluotokset, alueotokset
    - PISA-tutkimukset, Terveys 2000
- PPS-otoksien poiminta
  - SAS Procedure SURVEYSELECT **Syntax**

PPS sampling

Sampling with probability proportional to size (PPS) is a method where auxiliary information has a key role. An auxiliary variable is assumed to be available as a measure of the size of a population element. Varying inclusion probabilities for population elements can be assigned using the size variable. Efficiency improves relative to SRS if the relationship between the study variable and the size variable is strong. PPS is often used in business surveys and in general, for situations where the sampling units vary with a size measure.

PPS-otanta

- PPS-otanta on erittäin tehokas menetelmä, jos kokoa mittaava muuttuja korreloi voimakkaasti tulosmuuttujan kanssa JA jos y-muuttujan ja z-muuttujan suhde pysyy (likimain) vakiona yli koko perusjoukon (lin. mallin vakiotermi = 0)

Sisältymistodennäköisyys

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

missä z_k on kokomuuttujan arvo perusjoukon alkiolle *k*

PPS: Tekninen yhteenveto I ja VLISS

- Kokonaismäärän perusestimaattori: Horvitz-Thompson (HT) -estimaattori

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

- Ks. Tekninen yhteenveto I
  - Sivut 7-10
- 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: Otoksien poiminta Province-perusjoukosta

a) SRSWOR-otos

- **SAS Procedure SURVEYSELECT**
  - SRSWOR, *n*=8

```
proc surveyselect
   data=province91
   out=otos1
   sampsize=8
   seed=9876543
   method=srs stats;
run;
```

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

SURVEYMEANS, SRSWOR
Totaaliestimaatti ja keskivirhe Std Dev

Statistics

Variable	Sum	Std Dev
UE91	26440	13282

- Sum = Estimoitu totaali
- Std Dev = totaaliestimaatin keskivirheen estimaatti (s.e = standard error)
- HUOM: SURVEYSELECT tuottaa painomuuttujan SamplingWeight arvot otostiedostoon kaikissa otantamenetelmissä

Totaaliestimaatti ja s.e

Horvitz-Thompson-estimaatti 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 $t = 15098$

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

b) Ositettu SRSWOR-otos

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

```
proc surveyselect
   data=province91
   out=otos2
   sampsize=(4,4)
   seed=9876543
   method=srs stats;
   strata kumu;
run;
```


Kokonaismäärän estimointi

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

```
data kunta; input kumu _total_;
datalines;
1 7
2 25
;
proc surveymeans data=otos2 total=kunta sum;
   strata kumu;
   weight SamplingWeight;
   var UE91;
   run;
```

Ositettu SRSWOR-otos ja estimointi

(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

c) Ositettu PPS-otos

- **SAS Procedure SURVEYSELECT**
  - Ositettu PPSWOR, *n*=8
  - 2 ositetta (Jyväskylä /muut kunnat), muuttuja osite

```
proc surveyselect
  data=province91
  out=otos3
  sampsize=(1,7)
  seed=9876543
  method=pps stats;
  strata osite;
  size HOU85; * PPS-kokomuuttuja;
run;
```

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

Yhteenveto, *n* = 8

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

VLISS – PPS sampling

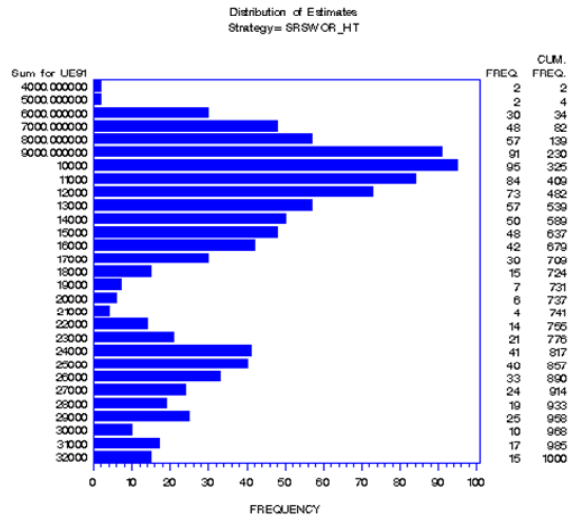
- VLISS Training Key 54
- Simulation experiment
  - 1,000 simulated samples of size *n* = 8
  - Estimation total of Y-variable UE
  - Measures of performance of HT estimator
    - Monte Carlo mean and standard error
    - Bias, ARB (absolute relative bias)
    - RMSE (Root mean squared error)
  - Size variables in PPS

    1) HOU85 (number of households in a municipality)
    2) X (artificially created variable for pedagogical purposes, UE91 + 3000), Y and X dependent but nonzero intercept
    3) Z (artificially created variable for pedagogical purposes, N(500,150)) , Y and Z independent

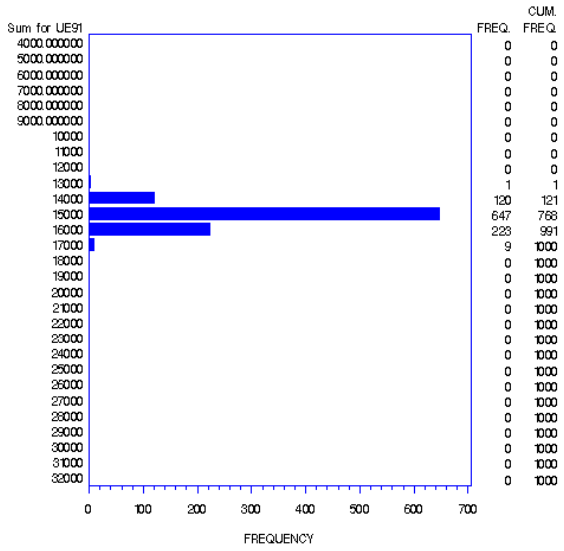
VLISS – PPS sampling

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

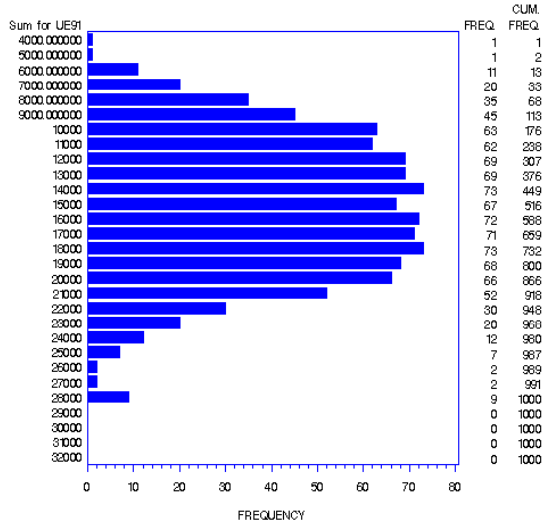
Distribution of SRSWOR_HT estimator



Distribution of PPSWOR estimator (aux.var. HOU85)

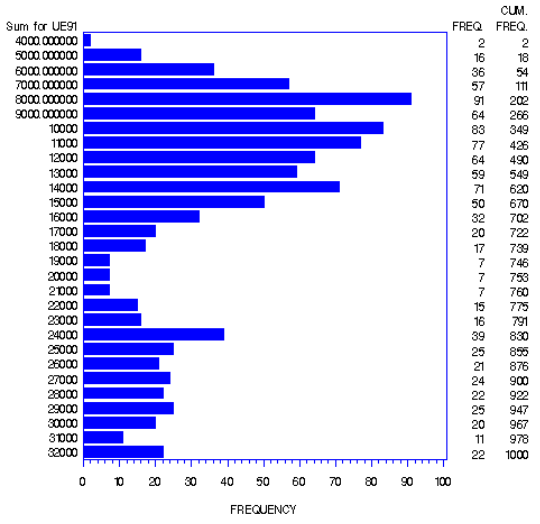


Distribution of PPSWOR estimator (aux.var. X)



	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)



	FREQ.	CUM. FREQ.
4000.000000	2	2
5000.000000	16	18
6000.000000	36	54
7000.000000	57	111
8000.000000	91	202
9000.000000	64	266
10000	83	349
11000	77	426
12000	64	490
13000	59	549
14000	71	620
15000	50	670
16000	32	702
17000	20	722
18000	17	739
19000	7	746
20000	7	753
21000	7	760
22000	15	775
23000	16	791
24000	39	830
25000	25	855
26000	21	876
27000	24	900
28000	22	922
29000	25	947
30000	20	967
31000	11	978
32000	22	1000