HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otantamenetelmät
(78143)
Syksy 2013

TEEMA 1 JATKUU…

Risto Lehtonen
risto.lehtonen@helsinki.fi

Yksinkertainen satunnaisotanta

- Yksinkertainen satunnaisotanta
  - SRSWOR: *Simple random sampling without replacement* – palauttamatta-tyyppinen
  - SRSWR: *Simple random sampling with replacement* – palauttaen-tyyppinen
- SRSWOR ja SRSWR
  - Ks. Tekninen yhteenveto I
  - SAS-laskenta
- Ks. VLISS Training Key 28
  - Analysing an SRS sample

Bernoulli-otanta

- *Survey sampling reference guidelines* *s. 17*

- **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 *k*th 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

- **Appendix 1**. Example of sample selection using Bernoulli sampling

- We create a sampling frame consisting *N*=2000 elements
- We want to select (about) *n*=200 units to the sample.
- Sampling fraction PI = *n*/*N*=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

Bernoulli-otanta

- Bernoulli-otannassa tehdään N toisistaan riippumatonta Bernoulli-koetta
- Otoskoko n on binomisesti jakautunut satunnaismuuttuja
- Odotettu otoskoko $\pi = n/N$
- Toteutunut otoskoko voi poiketa odotusarvosta
  $E(n) = N\pi = 2000 \times 0.1 = 200$
- Varianssi:
  $Var(n) = N(1 - \pi)\pi = 2000 \times 0.9 \times 0.1 = 180$
- Keskihajonta = 13.4

Systemaattinen otanta SYS
Systematic sampling

- **Poimintamenettely**

  a) Määritä 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

- Vaihtoehtoja
  - Ks. Lehtonen-Pahkinen (2004) Sect. 2.4

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
- Estimointi
  - Ks. Lehtonen-Pahkinen (2004) Sect. 2.4

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
    - Ks: Lehtonen-Pahkinen (2004) Example 2.3

Systemaattinen otanta

- Ks. Tekninen yhteenveto I sivu 4
- Sisäkorrelaatio *Intra-class correlation*
- 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
- Tilastokeskuksen tulonjakotilasto Laatuseloste

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

- S**uhteellinen 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, jos ositteiden koot vaihtelevat
- **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
- Lehtonen-Djerf (2008)
- 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, kun kaksi ehtoa on voimassa:
- Kokoa mittaava muuttuja *z* korreloi voimakkaasti tulosmuuttujan *y* kanssa
  JA
- Tulosmuuttujan *y* ja kokomuuttujan (apumuuttuja) *z* suhde pysyy (likimain) vakiona yli koko perusjoukon
  - Vastaavan lineaarisen regressiomallin vakiotermi = 0

Malli on muotoa $$y_k = \beta_1 z_k + \varepsilon_k$$

PPS: Tekninen yhteenveto I ja VLISS

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

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

- Ks. Tekninen yhteenveto I Sivut 7-10
- Ks. Lehtonen-Pahkinen (2004) Section 2.5
- Ks. VLISS Training Key 54
  The effective use of auxiliary information in PPS sampling
  Ks. Jaettu moniste

Poisson-poiminta

- PPSWOR-otannan erikoistapaus
- Otoskoko *n* on satunnaismuuttuja
  - Vrt. Bernoulli-otanta SRSWOR- erikoistapauksena
- Otoskoon odotusarvo $\sum_{k=1}^{N} \pi_k$
- Varianssi $\sum_{k=1}^{N} \pi_k(1-\pi_k)$
- Poiminta:
  - Laske PI $= \pi_k = nZ_k/T_z$
  - Generoi tasajakaumasta Uniform(0,1) *N* riippumatonta satunnaismuuttujaa UNI
  - Jos UNI ≤ PI niin IND = 1, muulloin IND = 0
  - Alkiot joille IND=1 kuuluvat Poisson-otokseen

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
- SRSWOR-estimaatin arvo riippuu vahvasti siitä, ovatko suuret kunnat mukana otoksessa
  - Kyllä: Suuri estimaatti
  - Ei: Pieni estimaatti…
  - Katsotaan tarkemmin harjoituksissa
- Parempi estimointi:
- Ositettu otanta
  - Osite 1:Kaupungit
  - Osite 2: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)
  - Ositemuuttujana Stratum tai URB85
  - Tasakiintiöinti (Equal allocation)

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

Kokonaismäärän estimointi

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

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

Ositettu SRSWOR-otos ja estimointi

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

Obs	ID	stratum	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 stratum

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

Ositettu PPSWOR-otos

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

Obs	stratum	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

d) Ositettu systemaattinen PPS

Table 2.8 A systematic PPS sample ($n = 8$) from the *Province'91* population.

Sample design identifiers			Element	Size measure	Study variables	
STR	CLU	WGHT	LABEL	HOU85	UE91	LAB91
1	1	1.000	Jyväskylä	26 881	4123	33 786
2	10	1.004	Jyväsk.mlk.	9230	1623	13 727
2	4	1.893	Keuruu	4896	760	5919
2	7	2.173	Äänekoski	4264	767	5823
2	32	2.971	Viitasaari	3119	568	4011
2	26	4.762	Pihtipudas	1946	331	2543
2	18	6.335	Kuhmoinen	1463	187	1448
2	13	13.730	Kinnula	675	129	927

Table 2.9 Estimates under a PPSSYS design ($n = 8$); the *Province'91* population.

Statistic	Variables	Parameter	Estimate	s.e	c.v	deff
Total	UE91	15 098	15 077	521	0.03	0.0035

- Lehtonen-Pahkinen Example 2.6

Yhteenveto, *n* = 8

Otos	Totaali	s.e
SRSWOR	26440	13282
STR	13892	4029
PPSWOR	14580	635
PPS-SYS	15 077	521
Perusjoukossa	15098	0

VLISS – PPS sampling

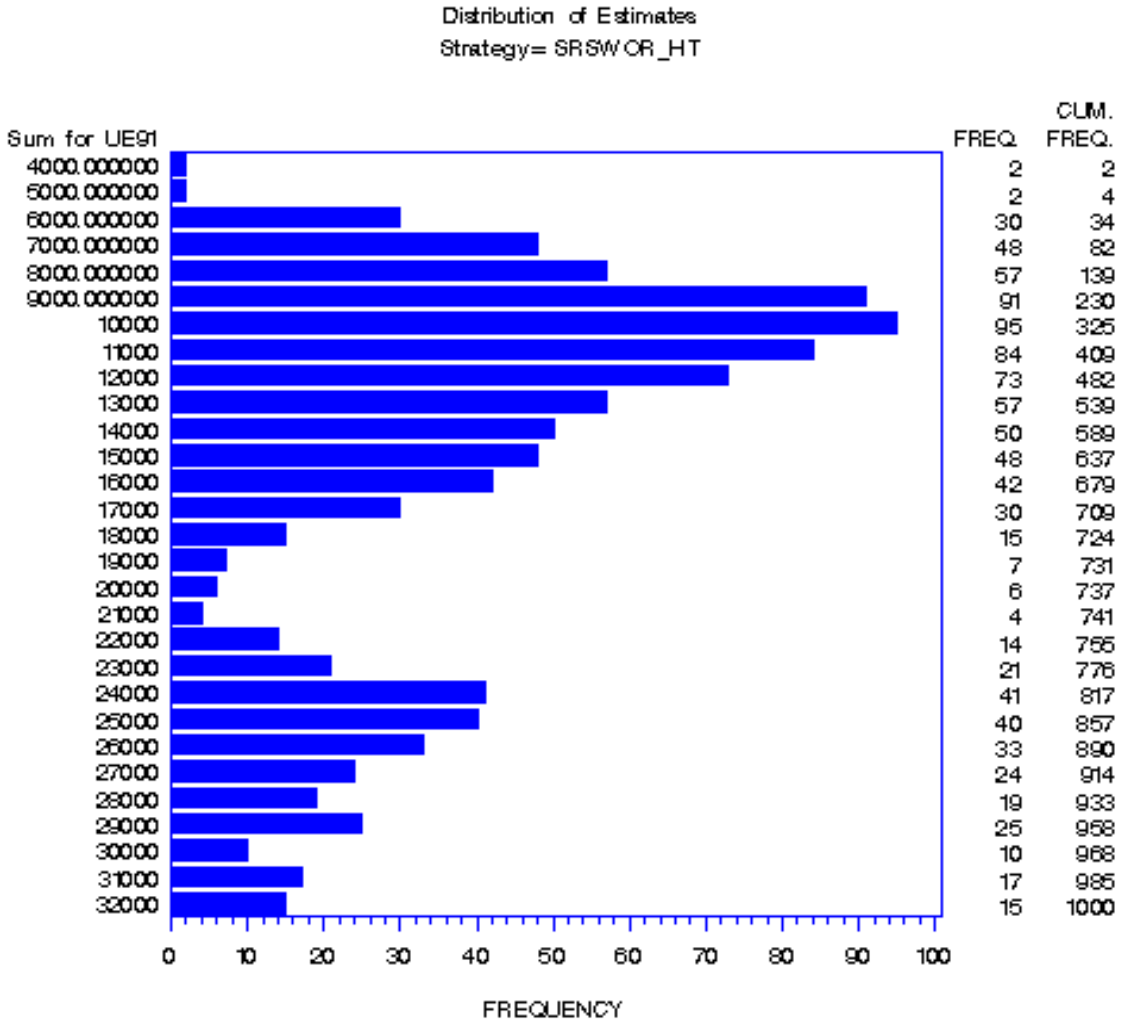
- VLISS Training Key 54
- Simulation experiment
  - 1,000 simulated samples of size $n = 8$
  - Estimation total of Y-variable UE
  - Measures of performance of HT estimator
    - Monte Carlo mean and standard error
    - Bias, ARB (absolute relative bias)
    - RMSE (Root mean squared error)
  - Size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

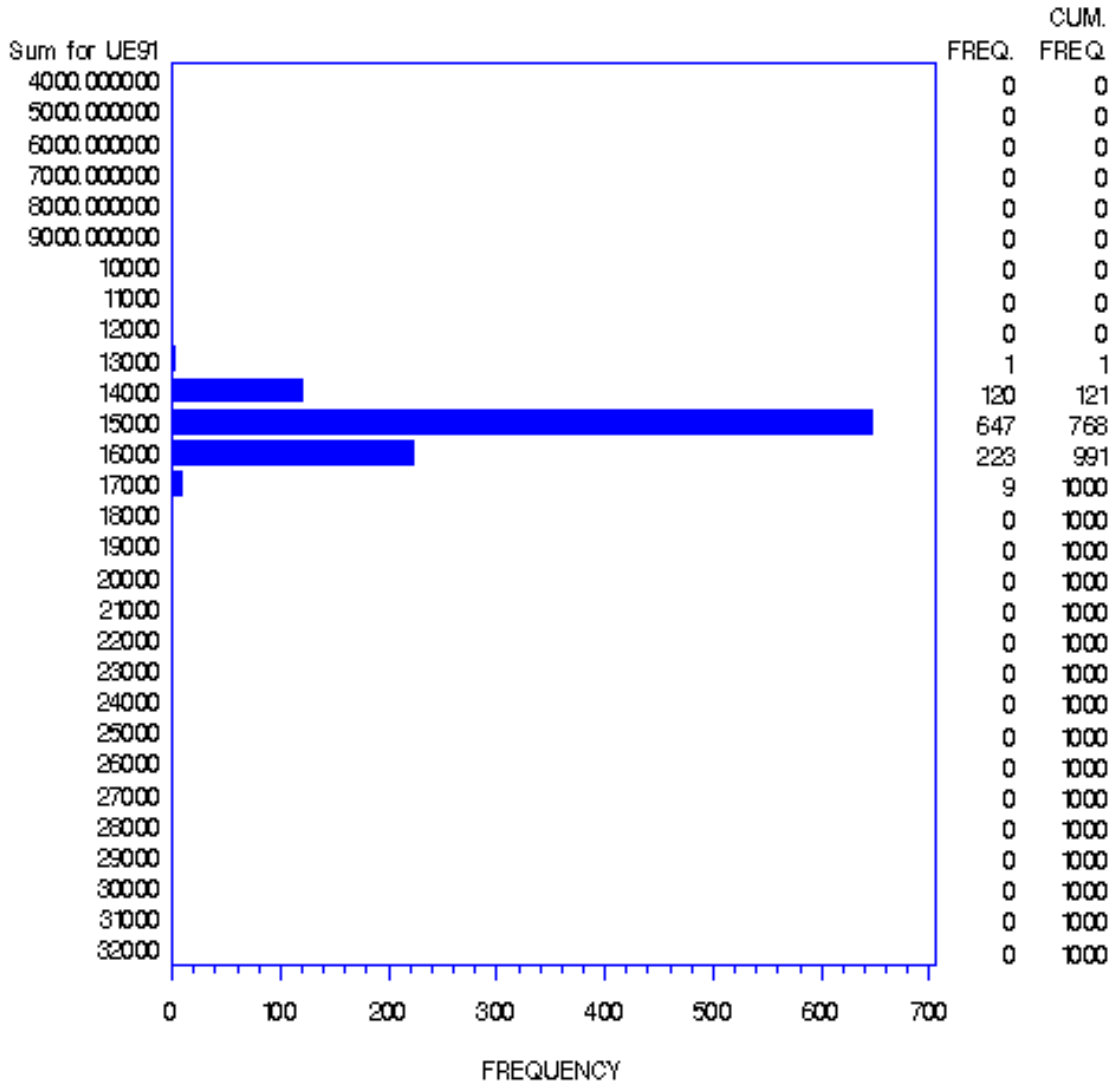
VLISS – PPS sampling

	Population	Mean of			Standard	Root
Strategy	Total	Estimates	Bias	ARB	error	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

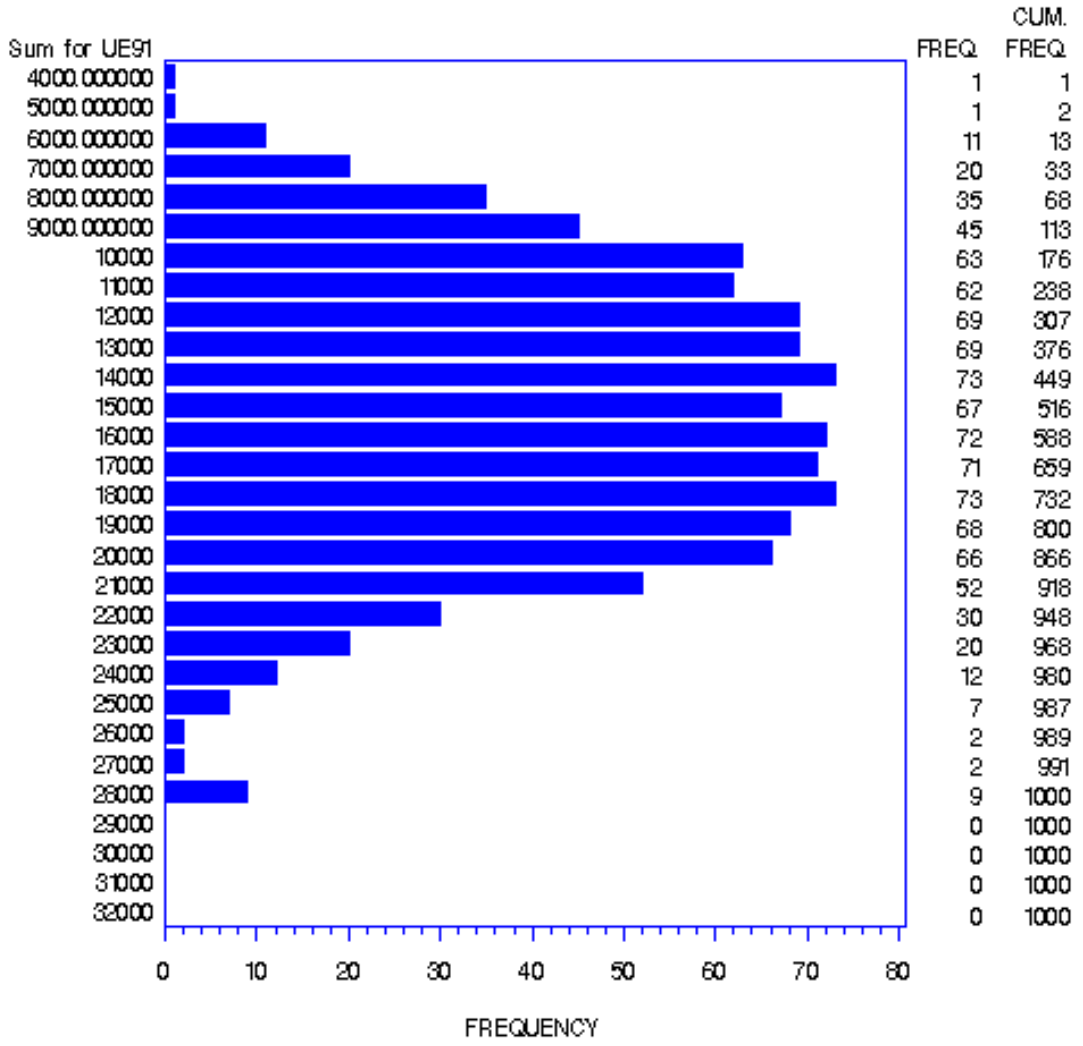
Distribution of SRSWOR_HT estimator



Distribution of PPSWOR estimator (aux.var. HOU85)

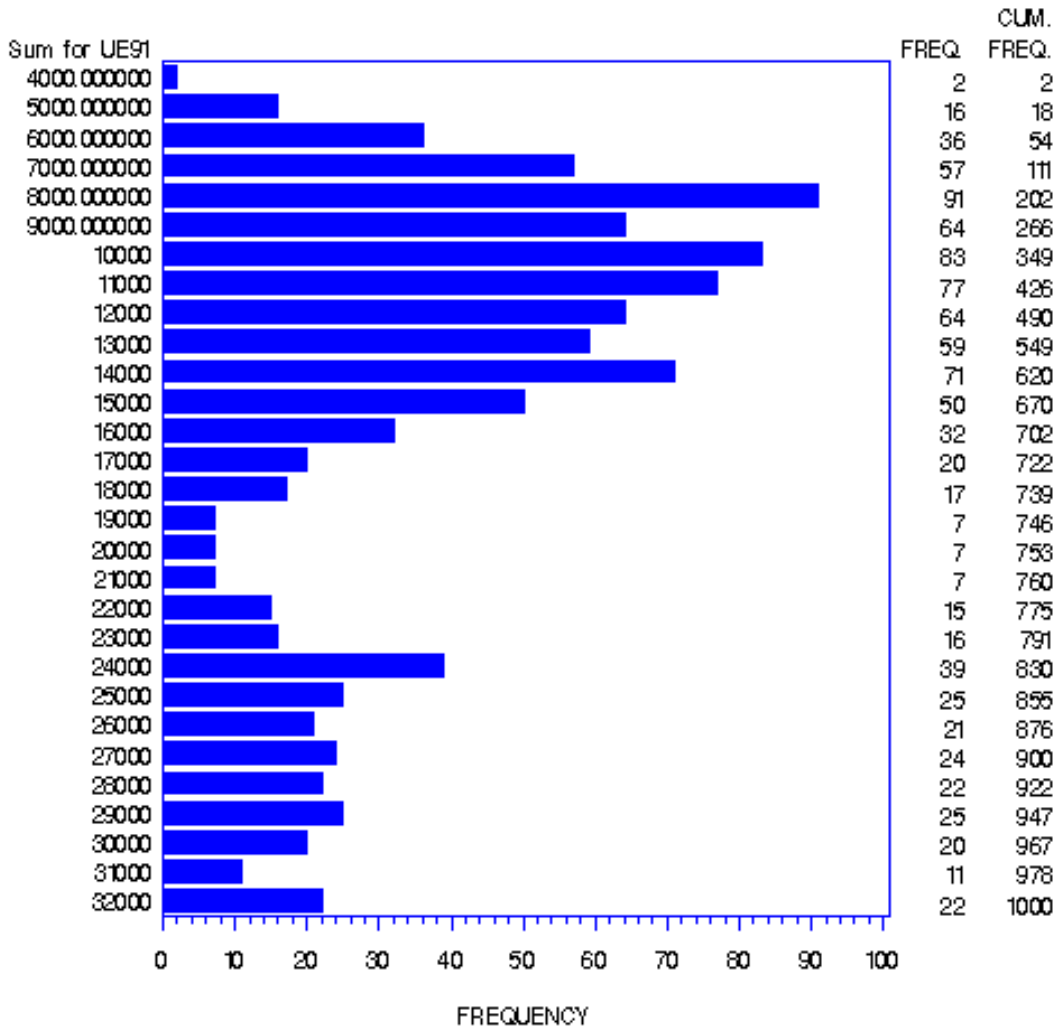


Distribution of PPSWOR estimator (aux.var. X)



	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