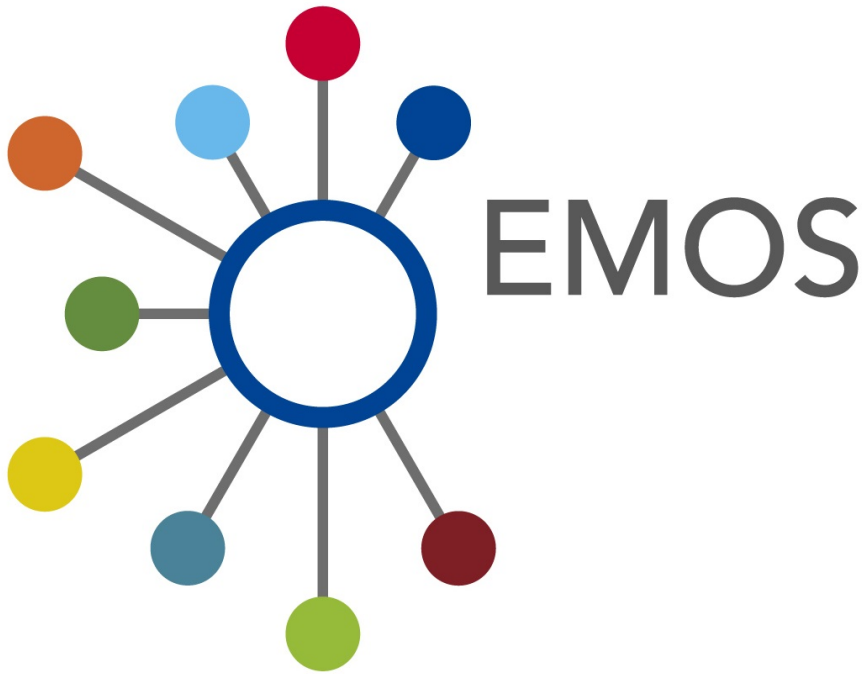
EMOS Webinar
Advanded Topics in Survey Sampling

Ralf Münnich
Economic and Social Statistics
Trier University



Basic definition (cf. Chapter 2)

Definition 2.1 (reconsidered)

Let $\mathcal{U}$ be a finite population with N elements. A sample $\mathcal{S}$ is generated with the help of a sample selection scheme that selects elements from the universe. The set of all possible samples with respect to the sample selection scheme is denoted by $\mathbb{S}$. The probability of drawing a pre-specified sample i is $P(\mathcal{S}_i)$.

The following holds:

$$\sum_{\mathcal{S} \in \mathbb{S}} P(\mathcal{S}) = 1 \qquad .$$

The tuple $(\mathbb{S}, P)$ is called sampling design. A deeper insight into theory and terminology is given in e.g. Cassel et al. (1977), Gabler (1990) as well as in Hedayat und Sinha (1991).

First order inclusion probabilities

The selection probability of element i in draw j is of special interest and denoted by $\psi_{i,j}$. These probabilities vary from draw to draw when sampling without replacement is chosen. We now define:

First order inclusion probabilities

The first order inclusion probability π_i is the probability, that element i is selected in the sample (independently from the draw!). Then:

$$\pi_i = \sum_{j=1}^{|\mathbb{S}|} \mathcal{I}(i \in \mathcal{S}_j) \cdot P(\mathcal{S}_j) \quad .$$

In general, we use the inverse inclusion probability $d_i = 1/\pi_i$ (IIP) and refer to the quantity d_i as design weight.

Properties of first order inclusion probabilities

The total probability over all samples is:

$$\sum_{i=1}^{|\mathbb{S}|} P(\mathcal{S}_i) = 1 \qquad .$$

For any estimator $\widehat{\pi}$ follows:

$$E(\widehat{\pi}) = \sum_{i=1}^{|\mathbb{S}|} \widehat{\pi}(\mathcal{S}_i) \cdot P(\mathcal{S}_i) \qquad \text{and} \qquad V(\widehat{\pi}) = \sum_{i=1}^{|\mathbb{S}|} (\widehat{\pi}(\mathcal{S}_i) - \pi)^2 \cdot P(\mathcal{S}_i)$$

We refer to design-unbiasedness, iff $E(\widehat{\pi}) = \pi$.

For fixed sample size n designs, we get

$$\sum_{i \in \mathcal{U}} \pi_i = n \qquad \text{and} \qquad \sum_{i \in \mathcal{S}} \frac{1}{\pi_i} = N \qquad .$$

SRS (revisited)

In SRS, the selection probability is $\psi_i = 1/N$ for each element. The inclusion probability for SRS in case of WR is drawn from the complementary event that element i is not included in the sample:

$$\pi_{i,\text{WR}} = 1 - \left(1 - \frac{1}{N}\right)^n \quad .$$

As inclusion probability for SRSWOR, we get

$$\pi_i = \binom{N-1}{n-1} \Big/ \binom{N}{n} = \frac{n}{N} \quad .$$

For $n = 1$, $\psi_i = \pi_i$. For large N, the inclusion probabilities for SRS are approximately $\pi_i \doteq n/N$ (cf. Särndal et al., 1992, p. 51).

Bernoulli Sampling and Poisson Sampling

Bernoulli Sampling (BERN)

For BERN, we get $\pi_i = \pi$ and $\pi_{ij} = \pi^2 \quad (i \neq j)$.
The sample size is random with $E(n) = N\pi$ and $V(n) = N\pi(1-\pi)$.
The selection is drawn sequentially with selection probability π.

Poisson Sampling (POIS)

POIS is a generalization of BERN with non-constant π_i. Due to the fact that POIS is WR, we get $\pi_{ij} = \pi_i \cdot \pi_j \quad (i \neq j)$.
The sample size is random with

$$E(n) = \sum_{i \in \mathcal{U}} \pi_k \qquad \text{and} \qquad V(n) = N \sum_{i \in \mathcal{U}} \pi_k (1 - \pi_k).$$

The selection is drawn sequentially with selection probabilities π_k.

Probability proportional to size sampling

One major problem of sampling with unequal probabilities is the design and especially the drawing mechanism (algorithm: cf. Section 1.4). One major drawback of these classes of designs is the calculation of the second order inclusion probabilities π_{ij} which may be cumbersome. These values will be needed for computations of variances estimates. However, some approximation formulae may be applied (cf. variance estimation methods).

One very efficient sampling design is the probability proportional to size method. In order to calculate these values, an auxiliary variable is needed that yields full information on universe-level. This, however, should be highly correlated to the variable of interest (estimation variable).

Account selection for audit sample: book values.
WR: pps WOR: πps

Probability proportional to size sampling (continued)

WR sampling: pps

The sample selection uses the sampling probability

$$\psi_i = \frac{x_i}{\sum_{j \in \mathcal{U}} x_j} \quad . \qquad \text{Hence, } \psi_j \propto x_j.$$

WOR sampling: πps

WOR sampling is done with the help of the inclusion probabilities

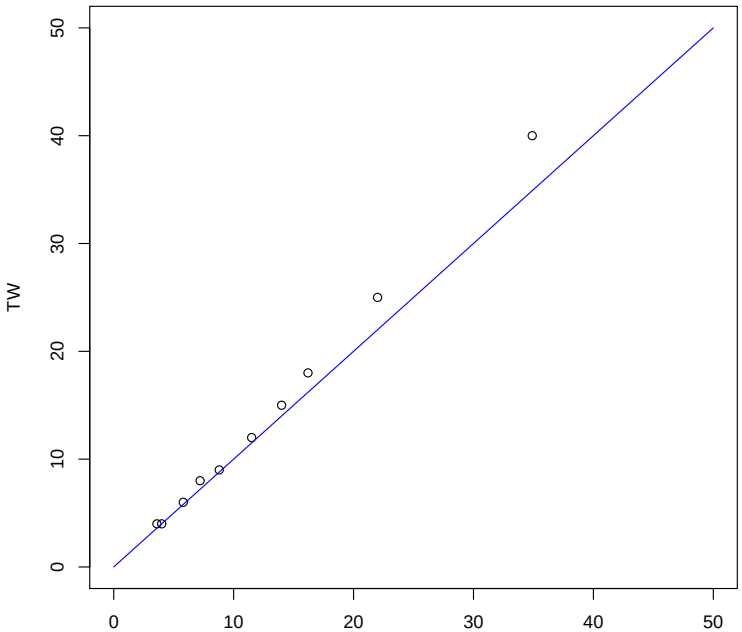
$$\pi_i = \frac{n \cdot x_i}{\sum_{j \in \mathcal{U}} x_j}$$

Problem: Determination of π_{ij} and the real sampling algorithm!
Further, $0 \leq \psi_i, \pi_i \leq 1$ (the latter has to be guaranteed!).

Example 4.1: Account selection for audit sample

In an inventory with $N = 10$ commodities, the following accounts (audit units) are given (cf. Lohr, 1999, pp. 202):

i	1	2	3	4	5	6	7	8	9	10	$\sum$
AV	4	4	6	8	9	12	15	18	25	40	141
BV	3.6	4.0	5.8	7.2	8.8	11.5	14.0	16.2	22.0	34.9	128



The book values (BV) are used to compute the probabilities proportional to size: 0.02813; 0.03125; 0.04531; 0.05625; 0.06875; 0.08984; 0.10938; 0.12656; 0.17188; 0.27266.

The Hansen-Hurwitz estimator (HH)

For unequal probability sampling (UPS) in case of WR, the Hansen-Hurwitz estimator

$$\widehat{\tau}_{\mathrm{HH}} = \frac{1}{n} \sum_{i \in \mathcal{S}} \frac{y_i}{\psi_i}$$

is unbiased for the total value τ_Y of the estimation variable Y.

$$V(\widehat{\tau}_{\mathrm{HH}}) = \frac{1}{n} \sum_{i \in \mathcal{S}} \psi_i \left(\frac{y_i}{\psi_i} - \tau \right)^2 \quad ,$$

with unbiased variance estimator

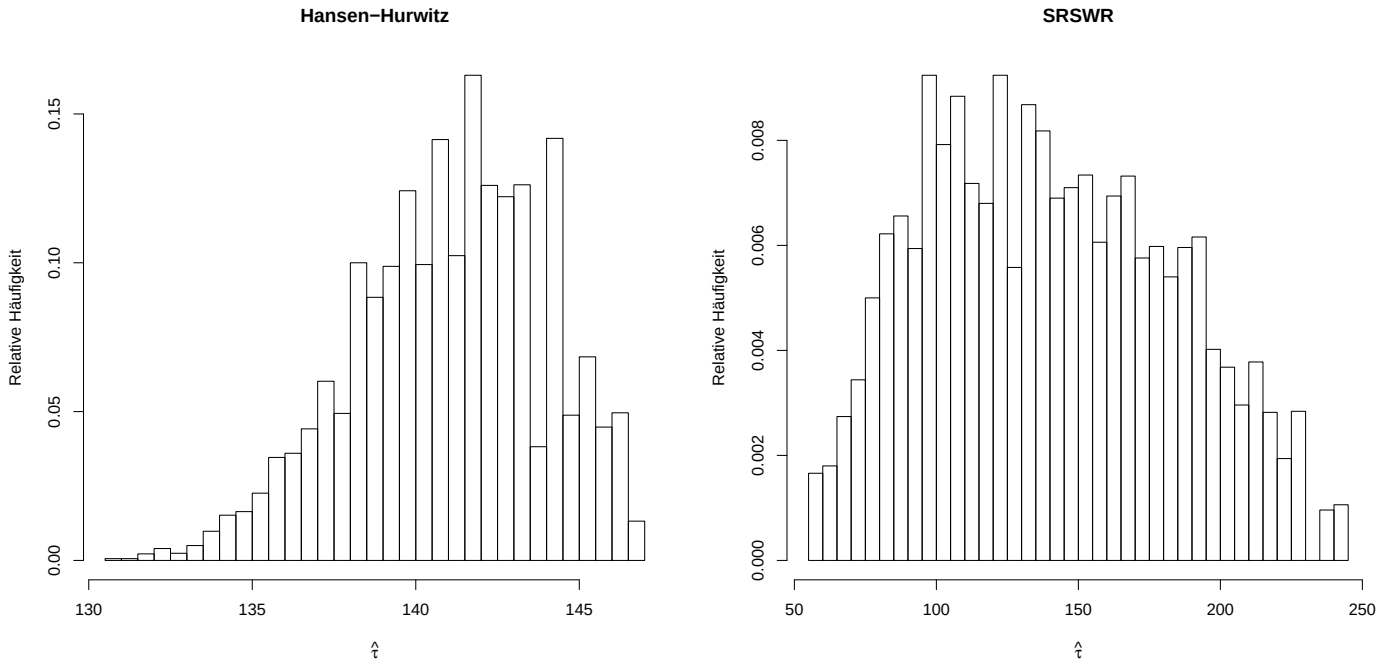
$$\widehat{V}(\widehat{\tau}_{\mathrm{HH}}) = \frac{1}{n \cdot (n-1)} \sum_{i \in \mathcal{S}} \psi_i \left(\frac{y_i}{\psi_i} - \widehat{\tau}_{\mathrm{HH}} \right)^2 \quad .$$

Example 4.2 (4.1 continued)

$n = 4$ elements 8, 10, 6 and 9 are drawn with pps. We get

$$\widehat{\tau}_{\text{HH}} = \frac{1}{4} \cdot \left(\frac{18}{0{,}12656} + \frac{40}{0{,}27266} + \frac{12}{0{,}08984} + \frac{25}{0{,}17188} \right) = 141{,}9867 \quad .$$

The true value is $\tau = 141$. SRS would yield $\widehat{\tau}_{\text{SRS}} = 237{,}5$. A simulation with $R = 10.000$ repetitions results in:

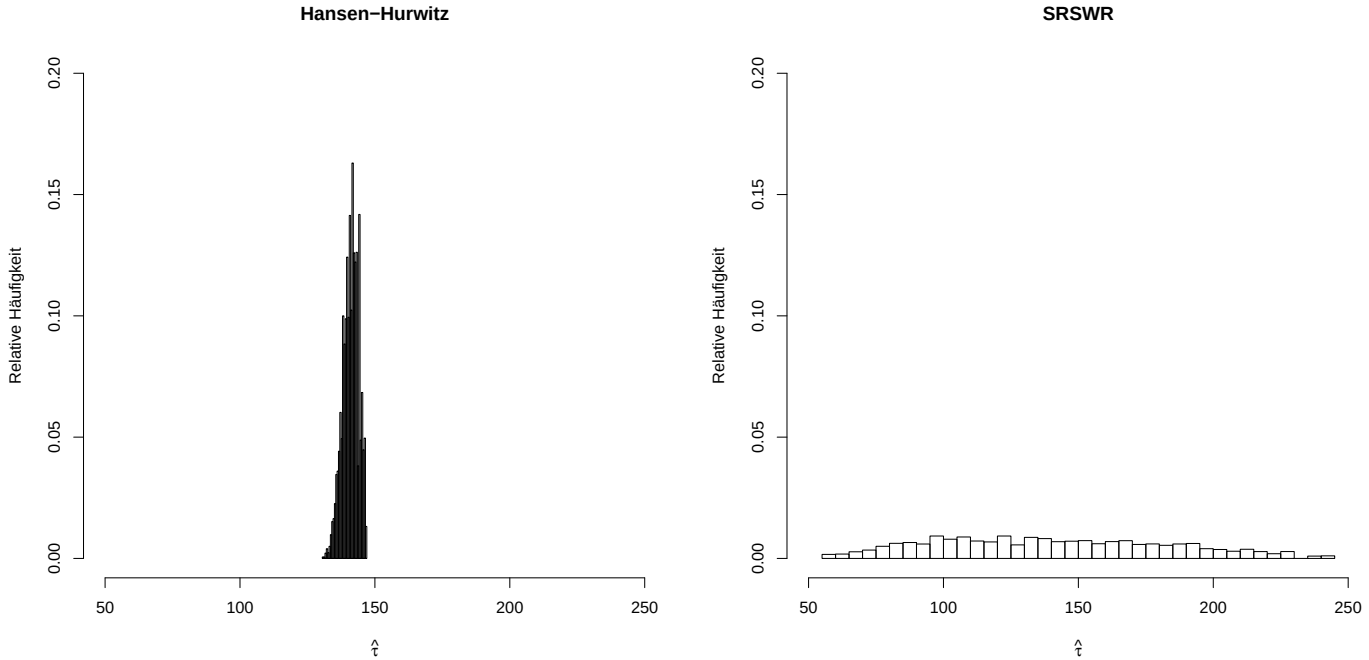


Example 4.2 (4.1 continued)

$n = 4$ elements 8, 10, 6 and 9 are drawn with pps. We get

$$\widehat{\tau}_{\text{HH}} = \frac{1}{4} \cdot \left(\frac{18}{0{,}12656} + \frac{40}{0{,}27266} + \frac{12}{0{,}08984} + \frac{25}{0{,}17188} \right) = 141{,}9867 \quad .$$

The true value is $\tau = 141$. SRS would yield $\widehat{\tau}_{\text{SRS}} = 237{,}5$. A simulation with $R = 10.000$ repetitions results in:



The Horvitz-Thompson estimator (HT)

For unequal probability sampling the HT estimator for the total τ_Y of a variable Y is (WOR)

$$\widehat{\tau}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} = \sum_{i \in \mathcal{S}} d_i \cdot y_i \quad ,$$

where $d_i = 1/\pi_i$ denotes the design weights (inverse inclusion probabilities).
Assuming a design with positive second order inclusion probabilities

$$V(\widehat{\tau}) = \sum_{i \in \mathcal{U}} \pi_i(1-\pi_i) \cdot \left(\frac{y_i}{\pi_i}\right)^2 + 2 \cdot \sum_{\substack{i,j \in \mathcal{U} \\ i<j}} \sum (\pi_{ij} - \pi_i \cdot \pi_j) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j}$$

is the variance of the HT estimator $\widehat{\tau}_{\mathrm{HT}}$.

Properties of the HT estimator

- HT estimators: A class of homogeneous, linear unbiased estimators
- The HT estimator is admissible;
  the HH is not necessarily admissible (improved HH estimator applying the Rao-Blackwell theorem)
- Negative variance estimates may result in some designs!
- The *design effect* should be considered.

Cassel, C.; Särndal, C.-E.; Wretman, J.H. (1977): Foundations of inference in survey sampling. Wiley.
Gabler, S. (1990): Minimax Solutions in Sampling from Finite Populations. Lecture Notes in Statistics, 64. Springer.
Hedayat, A.S.; Sinha, B.K. (1991); Design and Inference in Finite Population Sampling, Wiley.

Example 4.3

In an inventory with $N = 100$ items, a total book value of $N \cdot \mu_X = 686.471$ € was observed. A sample with size $n = 25$ was taken by SRSWR. The sample yielded a mean inventory value of $\widehat{\mu}_Y = 4.620{,}76$ with corresponding mean book value of $\widehat{\mu}_X = 4.664{,}80$. The regression coefficient in the sample is calculated as $\widehat{B} = 1{,}1063$.
Estimation without using auxiliary information leads to

$$\widehat{\tau}_{Y\,\mathrm{SRS}} = N \cdot \widehat{\mu}_{Y\,\mathrm{SRS}} = 100 \cdot 4.620{,}76 = 462.076 \quad .$$

Applying regression estimation yields

$$\widehat{\tau}_{Y\,\mathrm{Reg}} = 100 \cdot \big(4.620{,}76 + 1{,}1063 \cdot \big(6.864{,}71 - 4.664{,}80\big)\big) = 705.451{,}7 \quad .$$

Example 4.4 (4.3 continued)

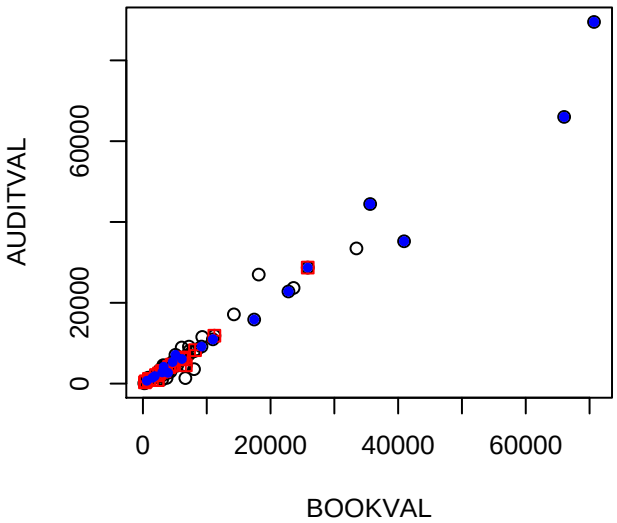
An alternative sampling design is πps. The first order inclusion probabilities are given as

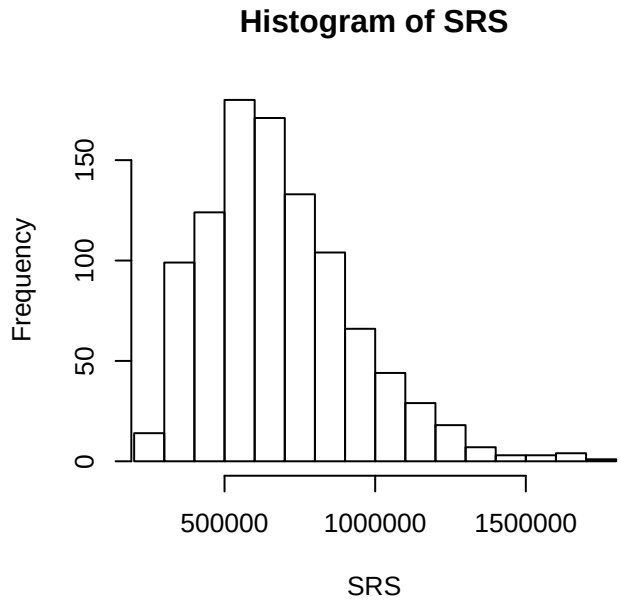
$$\pi_i = n \cdot y_i \Big/ \sum_{j=1}^{N} y_j \quad .$$

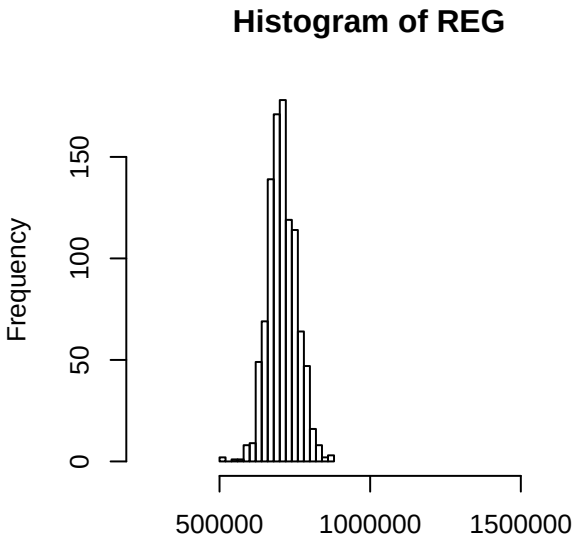
The Horvitz-Thompson estimate then is

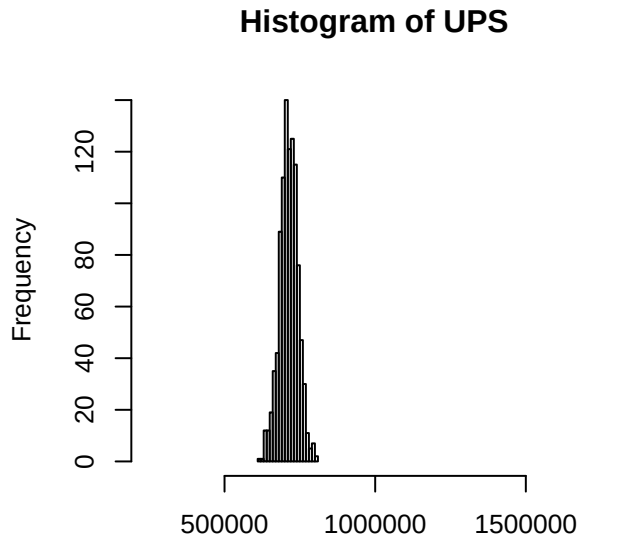
$$\widehat{\tau}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} = 729.908{,}2 \quad .$$

The following scatter plot includes the values in the universe, the sample according to Example 4.2 (red) and 4.3 (blue). Further, the simulated estimator distributions of the three estimators are shown as histograms.


AUDITVAL
0
20000
60000
0
20000
40000
60000
BOOKVAL


Histogram of SRS
Frequency
0
50
100
150
500000
1000000
1500000
SRS


Histogram of REG
Frequency
0
50
100
150
500000
1000000
1500000


Histogram of UPS
Frequency
0
20
40
60
80
120
500000
1000000
1500000

Basu's elephants (1971)

An essay on the logical foundations of survey sampling, Part One. In: Foundations of Statistical Inference (1971) edited by Godambe, V.P. and Sprott, D.A. (S. 203-242), Toronto.

A circus owner plans to ship his 50 elephants. Since the weighing of elephants is sophisticated, he aims to select $n = 1$, but which one. The director argues to take Samba whose weight was close to the average last time. The circus statistician is mad about this recommendation since the selection is non-random. He suggests:

$$\pi_{\text{Samba}} = \frac{99}{100} \quad \text{and} \quad \pi_i = \frac{1}{4.900} \quad \text{for all others.}$$

This yields:

If Samba is chosen: $\hat{\tau}_{\text{HT}} = 100/99 \cdot Y_{\text{Samba}}$

If Jumbo is chosen: $\hat{\tau}_{\text{HT}} = 4.900 \cdot Y_{\text{Jumbo}}$

... and the circus statistician looses his job!

Unequal probability designs

The problem of UPS sampling is:

- The selection procedure
- Determination of second order inclusion probabilities

We differ:

- Brewer sampling
- Midzuno sampling
- Maximum entropy sampling
- Tillé sampling
- Sampford sampling

Recommended reading: Yves Tillé (2006), Sampling Algorithms, Springer.

Motivation - Household surveys in Germany

European Statistics Code of Practice, Principle 14:

European Statistics are consistent internally, [...] it is possible to combine and make joint use of related data from different sources.

1. **Census**
   - Different levels of stratification (BL, KRS, SMP)
   - Benchmarks and auxiliary variables from different sources:
     - Known totals, e.g. from registers
     - Estimated totals from other suvery (with other methods, e.g. GREG, small area methods etc.)

   → Coherence und Consistency

   → One-Number-Census (one general vector of weights)

2. **Household surveys**
   - Coherence between different surveys in integrated housefold surveys (LFS, SILC,...)

Classical calibration

- **Known:** Population size N, design-weights d_i $(i = 1, \ldots, N)$, known totals $\boldsymbol{\tau}_{\mathbf{x}} \in \mathrm{R}^p$, sample $\mathcal{S}$ of size n.
- **T.b.d.:** Calibration weights $w_i := g_i d_i$ $(i = 1, \ldots, n)$
- **Goal:** Estimate $\hat{\tau}_Y^{\text{cal}} := \sum_{i \in \mathcal{S}} d_i g_i y_i$

Calibration problem (P)

$$\min_{g \in \mathrm{R}^n} \sum_{i \in \mathcal{S}} d_i D(g_i)$$

$$s.t. \sum_{i \in \mathcal{S}} d_i g_i \mathbf{x_i} = \boldsymbol{\tau}_{\mathbf{x}}$$

	$D(g_i)$
GREG	$\frac{1}{2} \cdot (g_i - 1)^2$
Raking Ratio	$g_i \log(g_i) - g_i + 1$
ML Raking	$g_i - 1 - \log(g_i)$

- Distance function $D(\cdot)$ determines the *penalty*
- In case of GREG distance, the optimal solution g^* of (P) is:

$$\hat{\tau}_Y^{\text{cal}} = \sum_{i \in \mathcal{S}} d_i g_i^* y_i = \hat{\tau}_y^{\text{GREG}} \quad \text{(cf. Deville, Särndal, 1992)}$$

Problems

- Negative weights $w_i = g_i d_i$ may occur for the GREG
- Extreme weights can happen (Gelman bound, see Gelman, 2007)

  $\rightarrow$ Box-constraints for g_i (see Münnich et al., 2012)

- No simultaneous calibration for different levels of stratification
- Large variation of even no solution in case of many benchmarks
- No *approximative* calibration on subgroups directly foreseen

  $\rightarrow$ Relaxation of selected benchmarks

Iterative solution

Semismooth Newton algorithm (SSN)

Generalized calibration

- Extension using box-constraints and relaxation (Münnich et al., 2012)

Generalized calibration problem (P*)

$$\min_{(\mathbf{g},\boldsymbol{\epsilon})\in R^{n+p^{rel}}} \sum_{i\in\mathcal{S}} d_i \frac{(g_i-1)^2}{2} + \sum_{k\in J} \delta_k \frac{(\epsilon_k-1)^2}{2}$$

$$s.t. \sum_{i\in\mathcal{S}} d_i g_i \mathbf{x}_i^{I} = \boldsymbol{\tau}_{\mathbf{X}^I}$$

$$\sum_{i\in\mathcal{S}} d_i g_i \mathbf{x}_i^{II} - \boldsymbol{\epsilon}\cdot\boldsymbol{\tau}_{\mathbf{X}^{II}} = \mathbf{0}$$

$$\mathbf{m} \leq \mathbf{g} \leq \mathbf{M}$$

$$\mathbf{L} \leq \boldsymbol{\epsilon} \leq \mathbf{U}$$

- No closed solution, iterative solution with sophisticated SSN algorithm
- Solution yields further information (e.g. violations within areas or outcomes of variables etc.)

Balanced Sampling - Framework

- Goal: estimate $\tau_y = \sum_{i \in \mathcal{U}} y_i$
- $\mathbf{x_i} \in \mathrm{R}^p$ vector of auxiliary variables available for each $i \in \mathcal{U}$
- Use auxiliary information as *balancing variables*
- $\mathcal{S}$: sample selected according to a probability sampling design $P(\mathcal{S})$
- $\mathbb{S}$: set of all possible samples
- π_i inclusion probability of unit i

Recommended reading: Tillé (2006)

Balanced Sampling - Strategy

Balancing strategy

A sample $s \in \mathcal{S}$ is said to be balanced if

$$\hat{\tau}_{\mathbf{x}} \equiv \sum_{i \in \mathcal{S}^*} \frac{\mathbf{x_i}}{\pi_i} = \sum_{i \in \mathcal{U}} \mathbf{x_i} \equiv \tau_{\mathbf{x}}$$

with inclusion probability π_i of unit i.
A design satisfying this equation for all possible $\mathcal{S}^* \in \mathcal{S}$ is called a *balanced sampling design*.

Algorithms:

- The Cube method: Deville and Tillé, (2004)
- Rejective sampling: Hajek (1981) and Fuller (2009)

Balanced Sampling - Why?

- Ensure consistency between survey estimates and the true population totals for the balancing variables
- May be interpreted as an *a priori calibration*
- Reduction of the set of possible samples $\mathcal{S}$: eliminate the undesirable samples
- Variance reduction: the variance of the resulting estimators may be reduced if there is a linear relationship between the characteristic of interest y_i and the balancing variables $\mathbf{x_i}$ $(i = 1, ..., N)$

Balanced Sampling - The Cube method

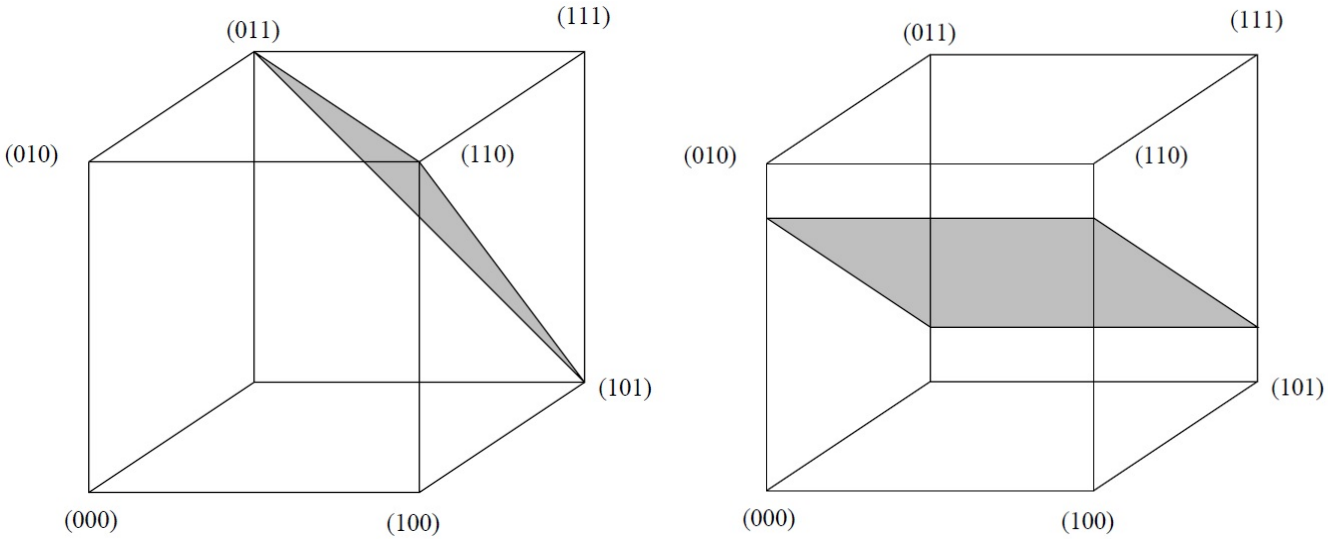
- Assign an inclusion probability π_i to every population unit prior to sampling
- Select a sample so that the balancing constraints are (approximately) satisfied

$$\hat{\tau}_{\mathbf{x}} \equiv \sum_{i \in s} \frac{\mathbf{x_i}}{\pi_i} = \sum_{i \in \mathcal{U}} \mathbf{x_i} \equiv \tau_{\mathbf{x}}$$

- The Cube method is based on a geometric representation of the set of the samples $\mathcal{S}$.
- A sample is completely characterized by a N-vector of 0 and 1
- Balancing equations define an linear subspace $Q \subset \mathrm{R}^N$

Balanced Sampling - The Cube method

- Reduction of the space of all samples $\mathcal{S}$ subject to given auxiliary information $\mathbf{x_i}$ (*balancing*)
- Graphical illustration (dimension $N = 3$):



- *Question:* Are there any (how many?) samples s in the reduced space $Q \subset \mathrm{R}^N$?

Balanced Sampling - The Cube method

- The Cube method consists of two distinct steps:
  - *A flight phase*: random walk starting from the vector of inclusion probabilities $\pi = (\pi_1, ..., \pi_N)^T$
  - *A landing phase*: complete the sample selection process, e.g., by successively relaxing the balancing constraints
- Inclusion probabilities remain *exactly respected* at each phase.
- Thus, the Horvitz-Thompson estimator is unbiased for τ_y
- **Issues:**
  - What, if no sample s can be found which satisfies exactly the balancing equation?
  - Is it possible to relax the balancing equation, e.g. the subspace Q, so that the balancing equation is *approximately* satisfied?

Balanced Sampling - Rejective sampling

Rejective sampling algorithms as an *alternative* to the Cube method:

- Relaxation of the balancing equation the avoid a small reduced space $Q \subset \mathcal{U}$
- Use rejective sampling with a balancing tolerance $\gamma > 0$ specified by the user
- Using rejective sampling, the discrepancy between $\hat{\tau}_{\mathbf{x}}^{\mathbf{p}}$ and the totals $\tau_{\mathbf{x}}$ is perfectly controlled through the balancing tolerance $\gamma > 0$
- Final π_i's are generally unknown
- See Chauvet, Haziza, and Lesage (2016)

Unemployment in Saarland

Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
Men	τ	4.172	9.504	10.588	0	24.264
$\sum$	τ	6.559	16.752	15.274	128	38.713

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

Unemployment in Saarland

Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
	$E\,\widehat{\tau}$	2.387	7.238	4.684	128	14.436
Men	τ	4.172	9.504	10.588	0	24.264
	$E\,\widehat{\tau}$	4.172	9.505	10.598	0	24.275
$\sum$	τ	6.559	16.752	15.274	128	38.713
	$E\,\widehat{\tau}$	6.558	16.743	15.282	128	38.711

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

Unemployment in Saarland

Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
	$\mathrm{E}\,\widehat{\tau}$	2.387	7.238	4.684	128	14.436
Men	τ	4.172	9.504	10.588	0	24.264
	$\mathrm{E}\,\widehat{\tau}$	4.172	9.505	10.598	0	24.275
$\sum$	τ	6.559	16.752	15.274	128	38.713
	$\mathrm{E}\,\widehat{\tau}$	6.558	16.743	15.282	128	38.711

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

Unemployment in Saarland

Unemployed		14 – 24	25 – 44	45 – 64	65 +	$\sum$
Women	τ	2.387	7.248	4.686	128	14.449
	$\mathrm{E}\,\widehat{\tau}$	2.387	7.238	4.684	128	14.436
Men	τ	4.172	9.504	10.588	0	24.264
	$\mathrm{E}\,\widehat{\tau}$	4.172	9.505	10.598	0	24.275
$\sum$	τ	6.559	16.752	15.274	128	38.713
	$\mathrm{E}\,\widehat{\tau}$	6.558	16.743	15.282	128	38.711

- *True* values in Saarland
- Estimates from the Microcensus
- Is the quality of the cell estimates identical?

Evaluation of samples and surveys (rpt.)

Practicability

Costs of a survey

Accuracy of results

- Standard errors
- Confidence interval coverage
- Disparity of sub-populations

Robustness of results

In order to adequately evaluate the estimates from samples, *appropriate* evaluation criteria have to be considered.

Why do we need variance estimation

Most *accuracy measures* are based on variances or variance estimates!

- Measures for point estimators
  - Bias, variance, MSE
  - CV, relative root MSE
  - Bias ratio, confidence interval coverage
  - Design effect, effective sample size
- Problems with measures:
  - *Theoretical* measures are problematic
  - Estimates from the sample (e.g. bias)
  - Availability in simulation study
  - Does large sample theory help much?
  - Small sample properties

Do we need special measures for variance estimators or variance estimates?

Why do we need variance estimation

Most *accuracy measures* are based on variances or variance estimates!

- Measures for point estimators
  - Bias, variance, MSE
  - CV, relative root MSE
  - Bias ratio, confidence interval coverage
  - Design effect, effective sample size
- Problems with measures:
  - *Theoretical* measures are problematic
  - Estimates from the sample (e.g. bias)
  - Availability in simulation study
  - Does large sample theory help much?
  - Small sample properties

Do we need special measures for variance estimators or variance estimates?

Why do we need variance estimation

Most *accuracy measures* are based on variances or variance estimates!

- Measures for point estimators
  - Bias, variance, MSE
  - CV, relative root MSE
  - Bias ratio, confidence interval coverage
  - Design effect, effective sample size
- Problems with measures:
  - *Theoretical* measures are problematic
  - Estimates from the sample (e.g. bias)
  - Availability in simulation study
  - Does large sample theory help much?
  - Small sample properties

Do we need special measures for variance estimators or variance estimates?

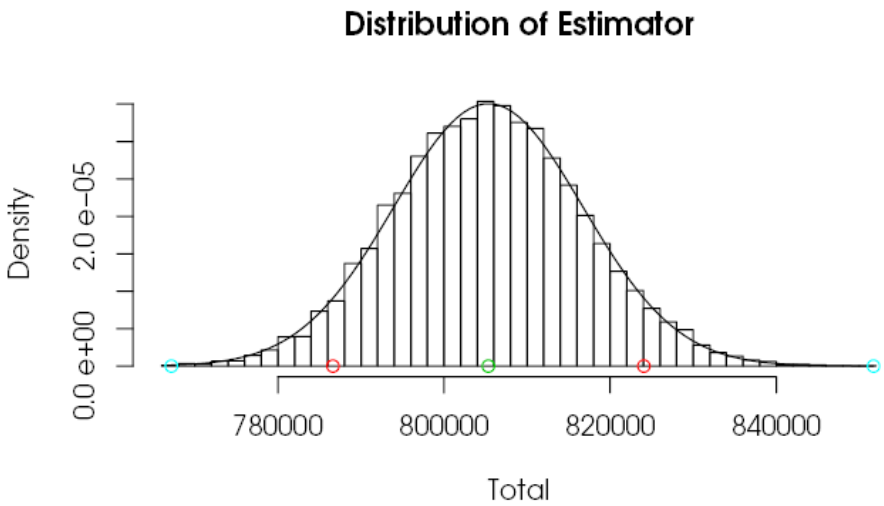
Why do we need variance estimation

Most *accuracy measures* are based on variances or variance estimates!

- Measures for point estimators
  - Bias, variance, MSE
  - CV, relative root MSE
  - Bias ratio, confidence interval coverage
  - Design effect, effective sample size
- Problems with measures:
  - *Theoretical* measures are problematic
  - Estimates from the sample (e.g. bias)
  - Availability in simulation study
  - Does large sample theory help much?
  - Small sample properties

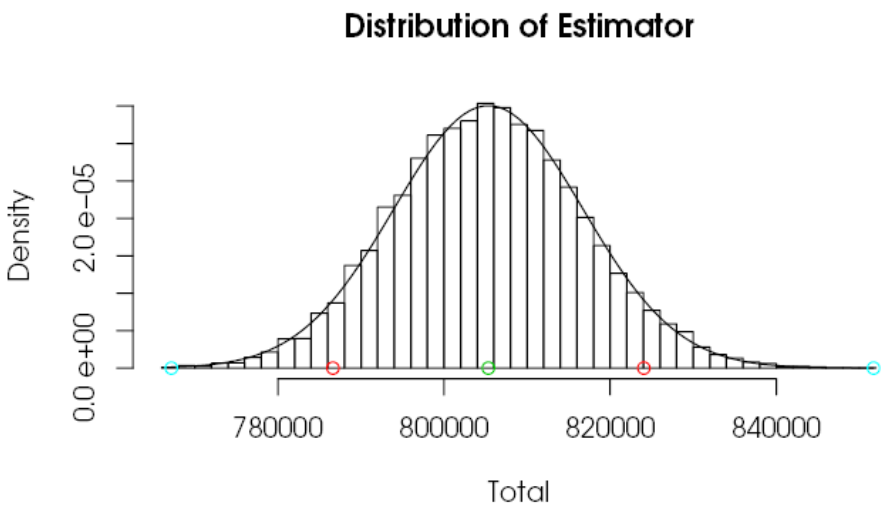
Do we need special measures for variance estimators or variance estimates?

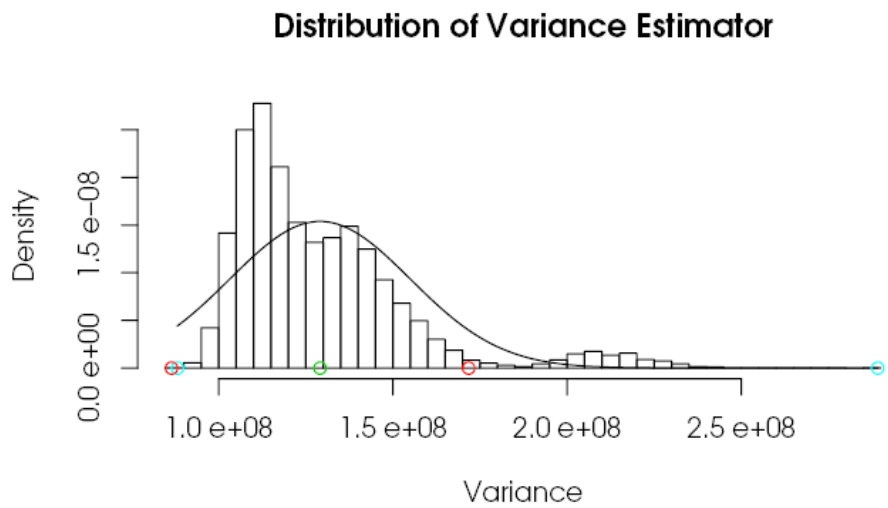
Example: Men in Hamburg



τ:	805258.00	N:	1669690				
$\hat{\tau}$:	805339.10	$V\hat{\tau}$:	1.29e+008	$E(\widehat{V}(\tau))$:	1.29e+008	$V(\widehat{V}(\tau))$:	6.72e+014
Bias Est:	81.10	MSE Est:	1.29e+008	Bias Var:	-3.78e+005	MSE Var:	6.72e+014
Skew Est:	0.0747	Curt Est:	3.0209	Skew Var:		Curt Var:	
CI (90%):				CI (95%):			

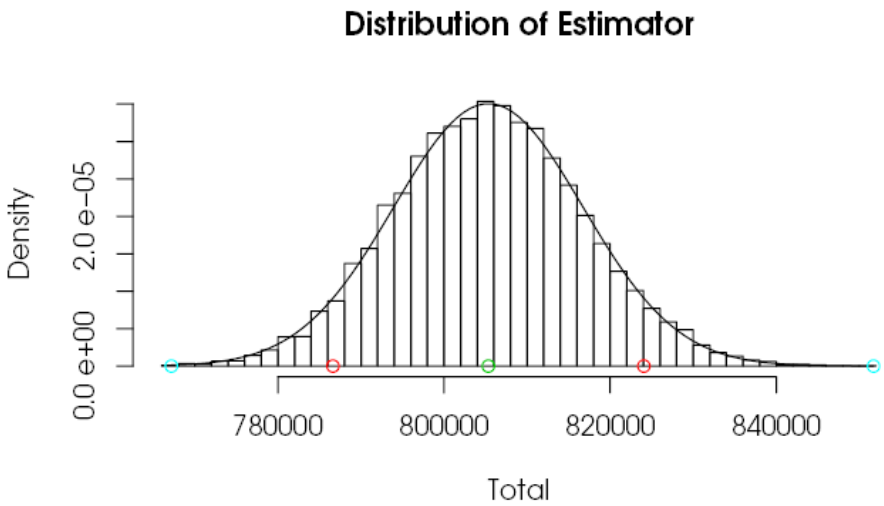
Example: Men in Hamburg

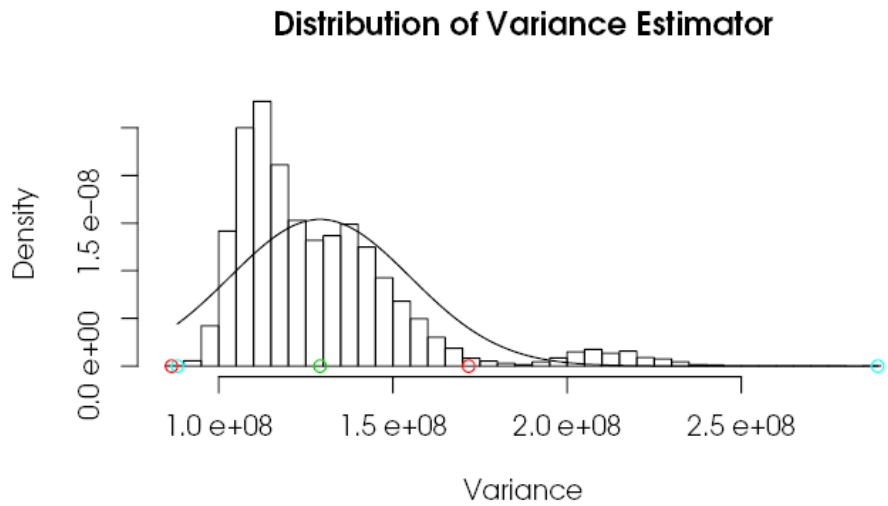




τ:	805258.00	N:	1669690				
$\hat{\tau}$:	805339.10	$V\hat{\tau}$:	1.29e+008	$E(\widehat{V}(\tau))$:	1.29e+008	$V(\widehat{V}(\tau))$:	6.72e+014
Bias Est:	81.10	MSE Est:	1.29e+008	Bias Var:	-3.78e+005	MSE Var:	6.72e+014
Skew Est:	0.0747	Curt Est:	3.0209	Skew Var:	1.8046	Curt Var:	6.9973
CI (90%):				CI (95%):			

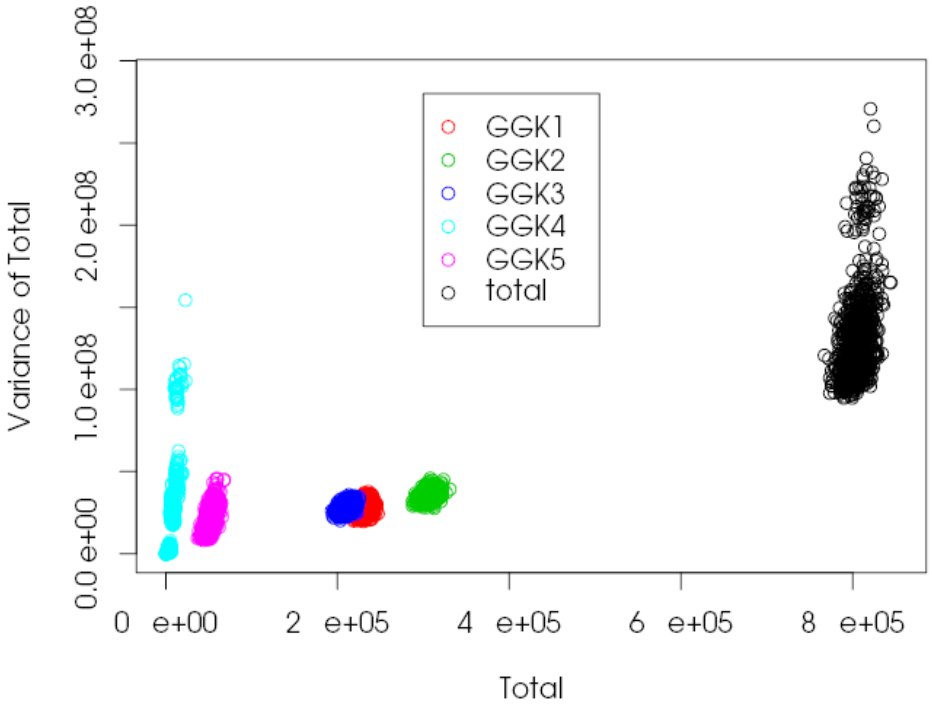
Example: Men in Hamburg

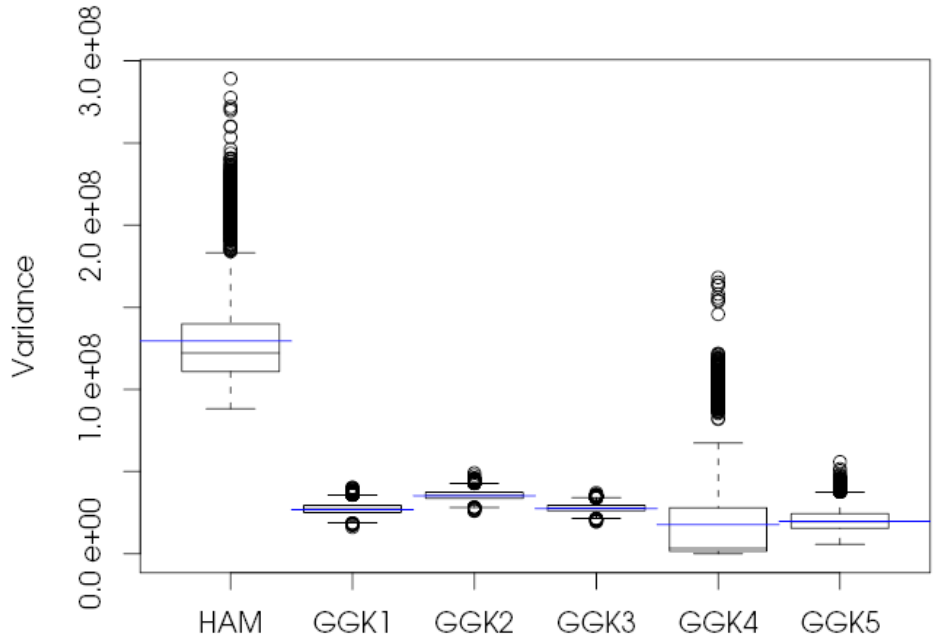




τ:	805258.00	N:	1669690				
$\hat{\tau}$:	805339.10	$V\hat{\tau}$:	1.29e+008	$E(\widehat{V}(\tau))$:	1.29e+008	$V(\widehat{V}(\tau))$:	6.72e+014
Bias Est:	81.10	MSE Est:	1.29e+008	Bias Var:	-3.78e+005	MSE Var:	6.72e+014
Skew Est:	0.0747	Curt Est:	3.0209	Skew Var:	1.8046	Curt Var:	6.9973
CI (90%):	90.16	(4.1;5.7)		CI (95%):	94.79	(2.0;3.2)	

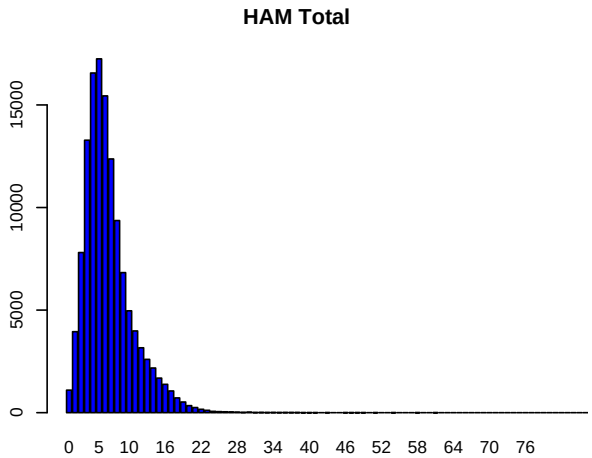
Total Estimate Separated by House Size Class (GGK)

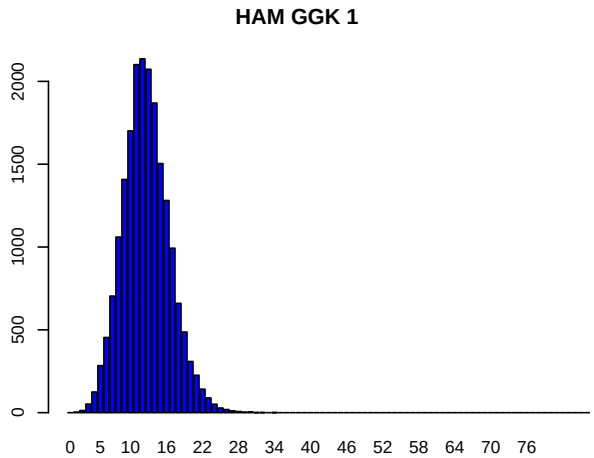


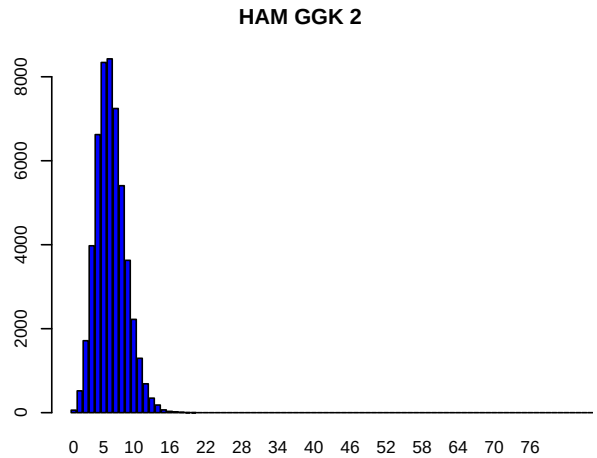


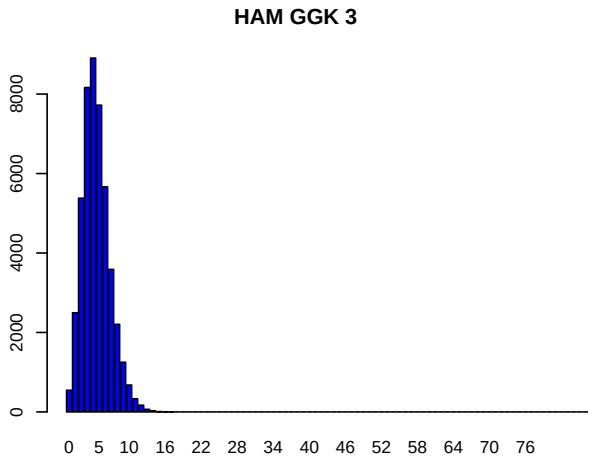
	GGK1	GGK2	GGK3	GGK4	GGK5	total
Persons	468293	651740	439745	9940	99970	1669690
Sampling units	173	446	414	10	75	1118

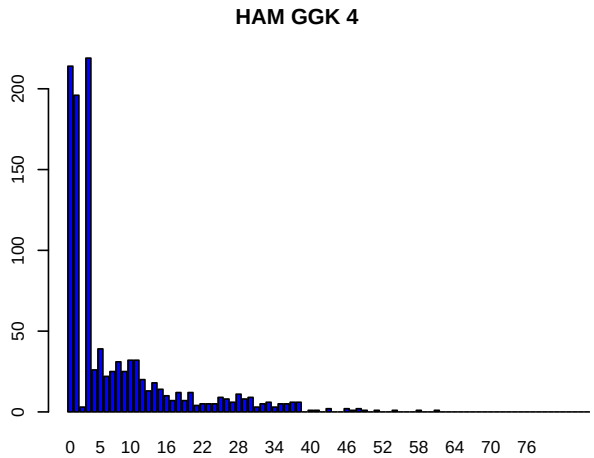
Distribution of Men in HAM (per SU)

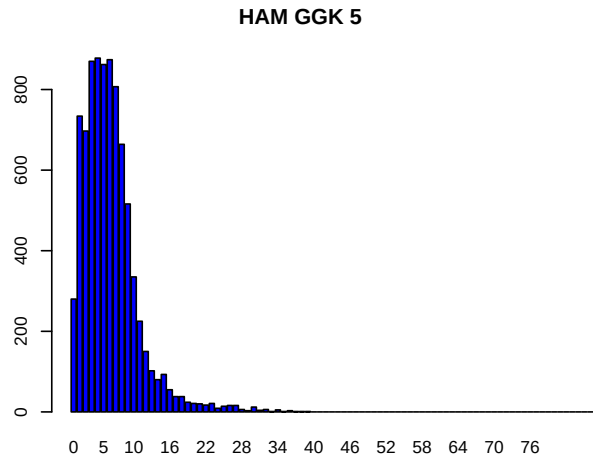




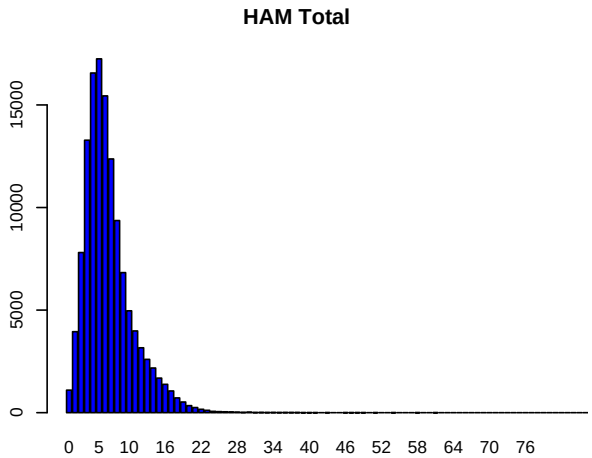


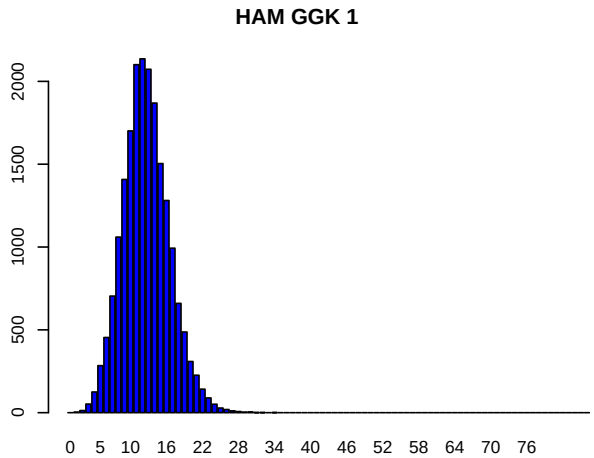


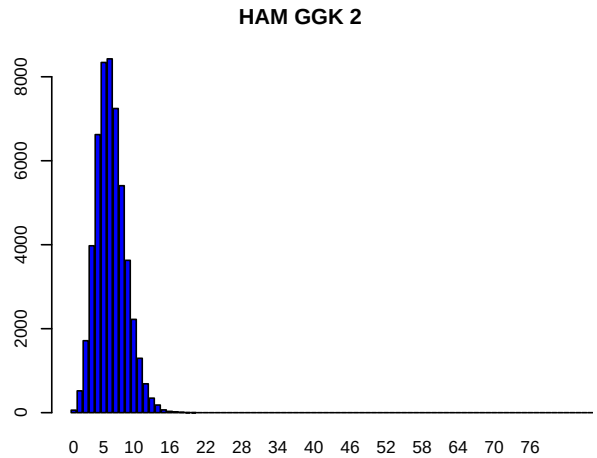


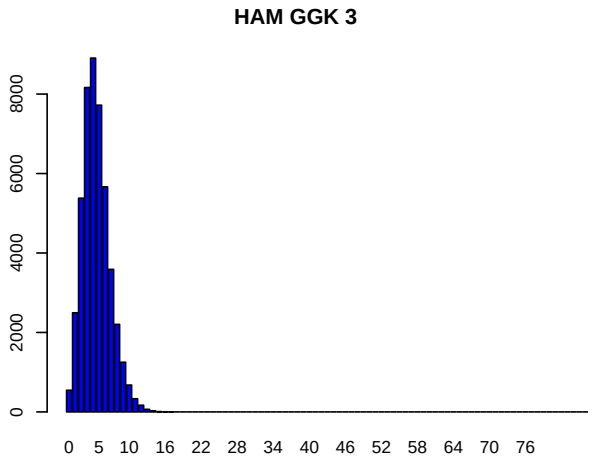


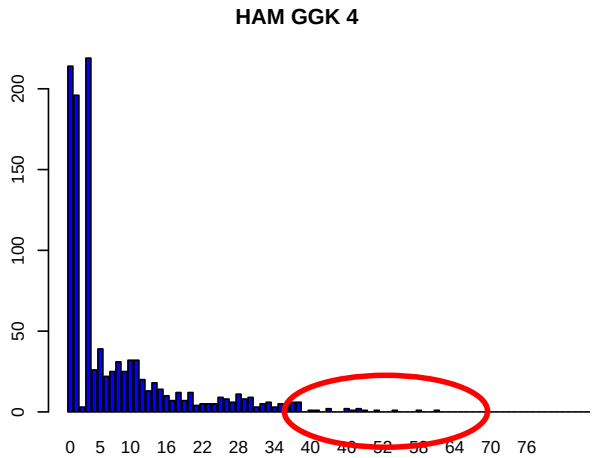
Distribution of Men in HAM (per SU)

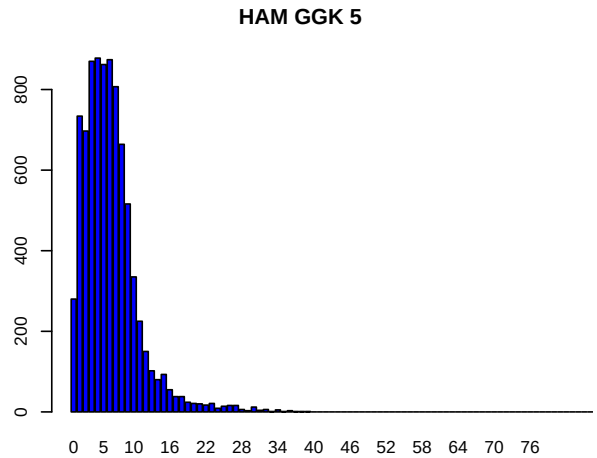






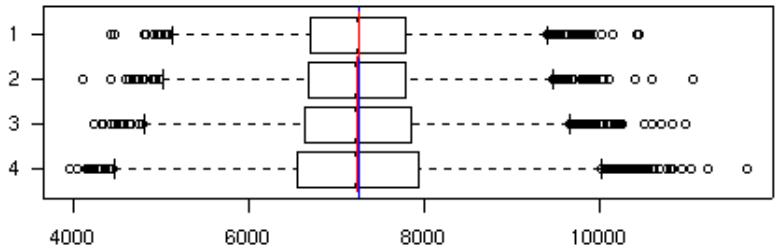




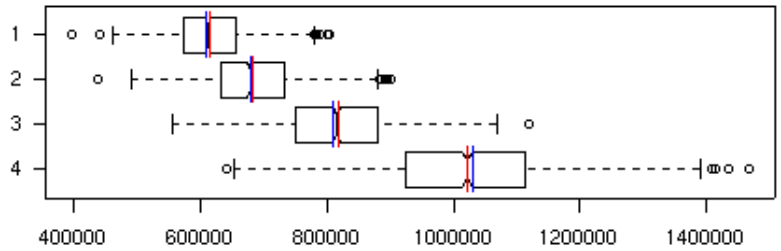


Unemployed women, 25 – 44

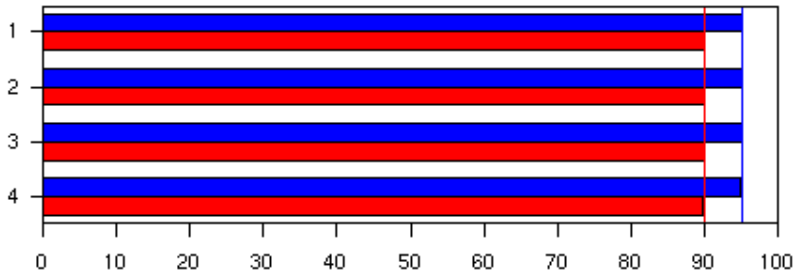
Raking estimator



variance estimator

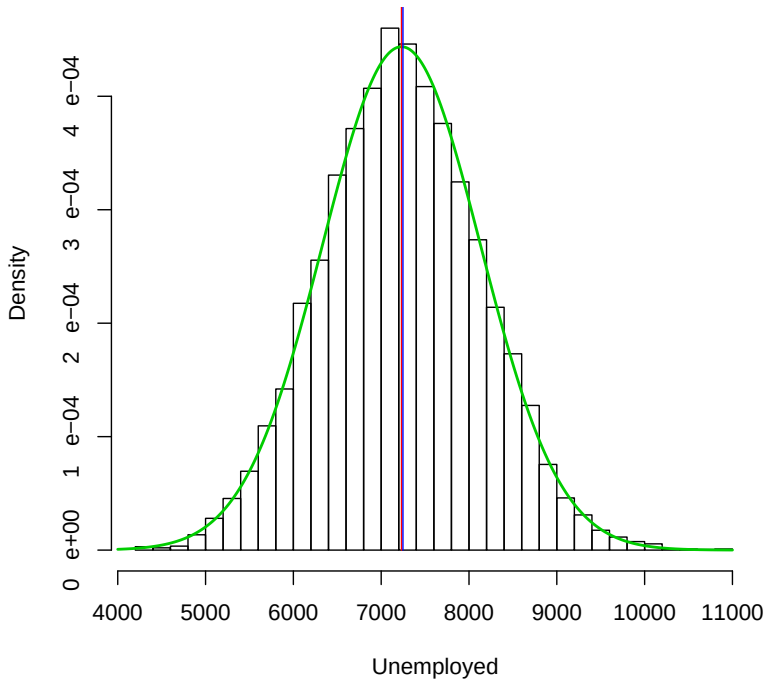


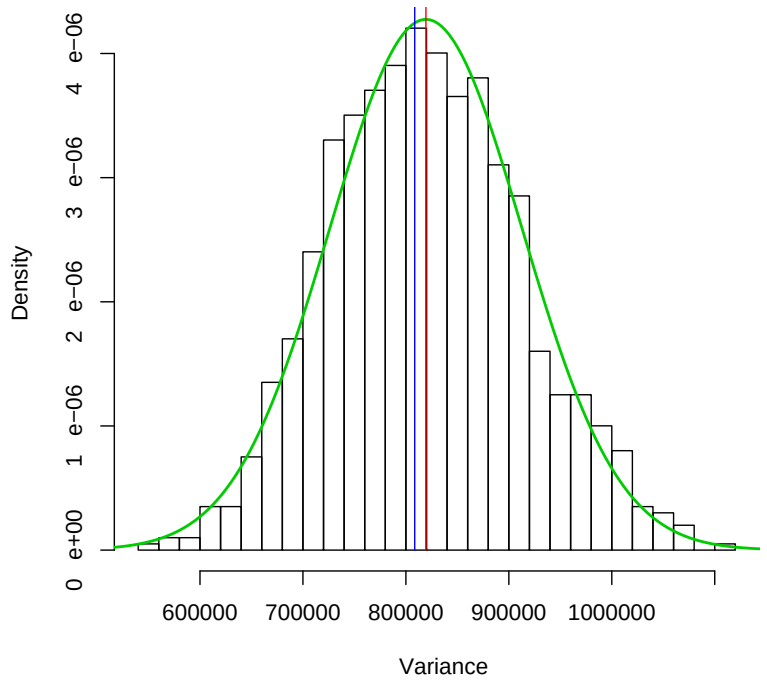
NR rates: 1: 5%, 2: 10%, 3: 25%, 4: 40%



95% 90%

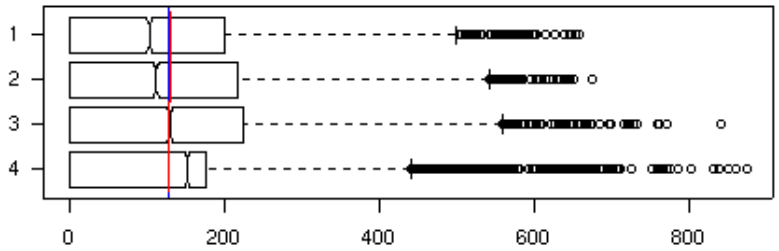
Unemployed women, 25 – 44, distribution of point and variance estimator (25% NR)



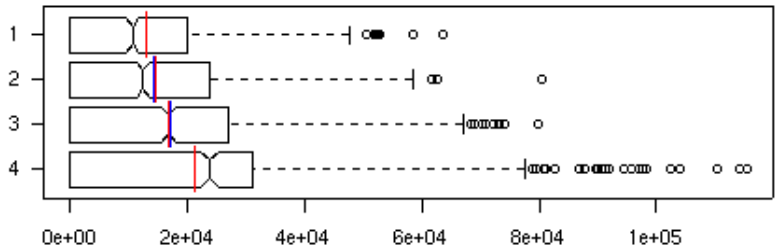


Unemployed women, 65 +

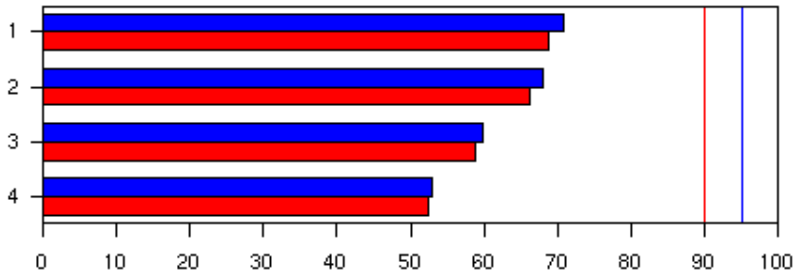
Raking estimator



variance estimator

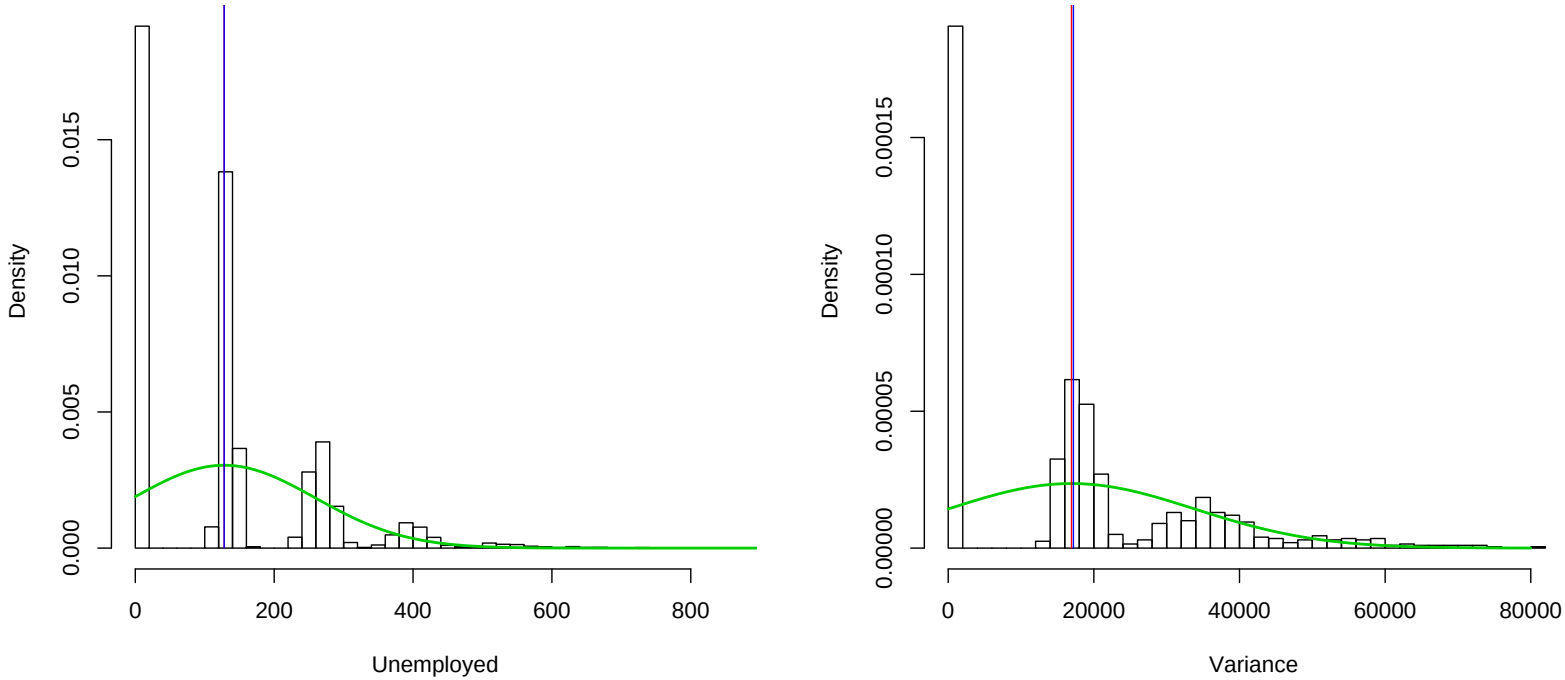


NR rates: 1: 5%, 2: 10%, 3: 25%, 4: 40%



95% 90%

Unemployed women, 65 +, distribution of point and variance estimator (25% NR)



Direct variance estimator for Two Stage Sampling

- *Direct variance estimator*:

$$\hat{\mathrm{V}}\left(\hat{\tau}_{2St}\right) = L^2 \cdot \left(\frac{L-l}{L}\right) \cdot \frac{s_e^2}{l} + \frac{L}{l} \sum_{q=1}^{l} \left(\frac{N_q - n_q}{N_q}\right) \cdot N_q^2 \cdot \frac{s_q^2}{n_q}$$

$$\text{with} \quad s_e^2 = \frac{1}{l-1} \sum_{q=1}^{l} \left(\hat{\tau}_q - \frac{\hat{\tau}}{L}\right)^2, s_q^2 = \frac{1}{n_q - 1} \cdot \sum_{i=1}^{n_q} \left(y_{qi} - \overline{y}_q\right)^2$$

cf. Lohr (1999), p. 147.

- The estimator is unbiased, but the first and second term do not estimate the variance at the respective stage (cf. Särndal et al. 1992, p. 139 f., Lohr 1999, p. 210):

$$\mathrm{E}\left[L^2 \cdot \left(\frac{L-l}{L}\right) \cdot \frac{s_e^2}{l}\right] = L^2 \cdot \left(1 - \frac{l}{L}\right) \cdot \frac{\sigma_e^2}{l} + \frac{L}{l}\left(1 - \frac{l}{L}\right) \sum_{q=1}^{L} \mathrm{V}\left(\hat{\tau}_q\right)$$

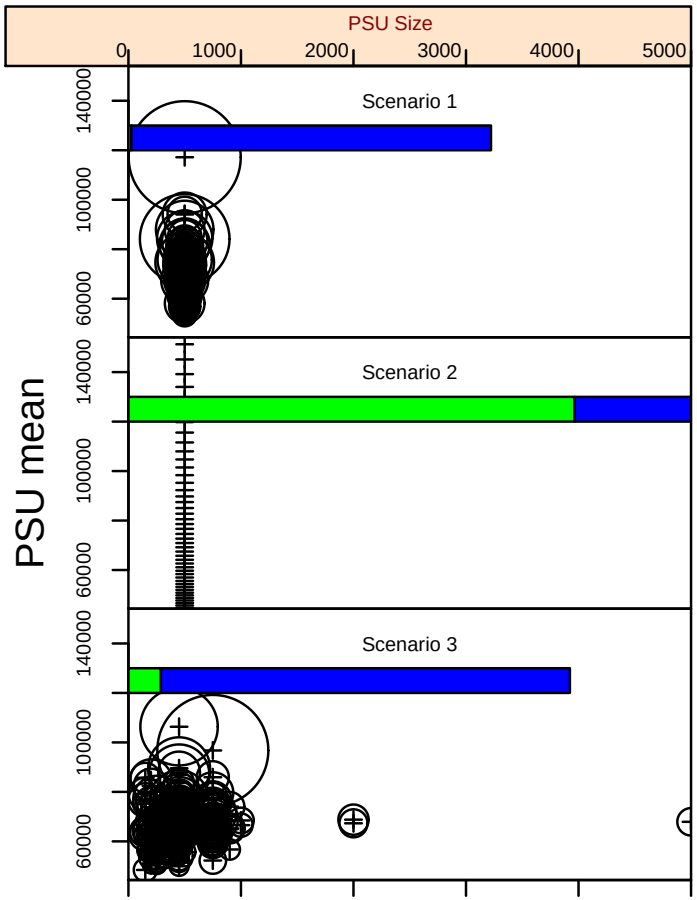
Experimental Study: Sampling Design

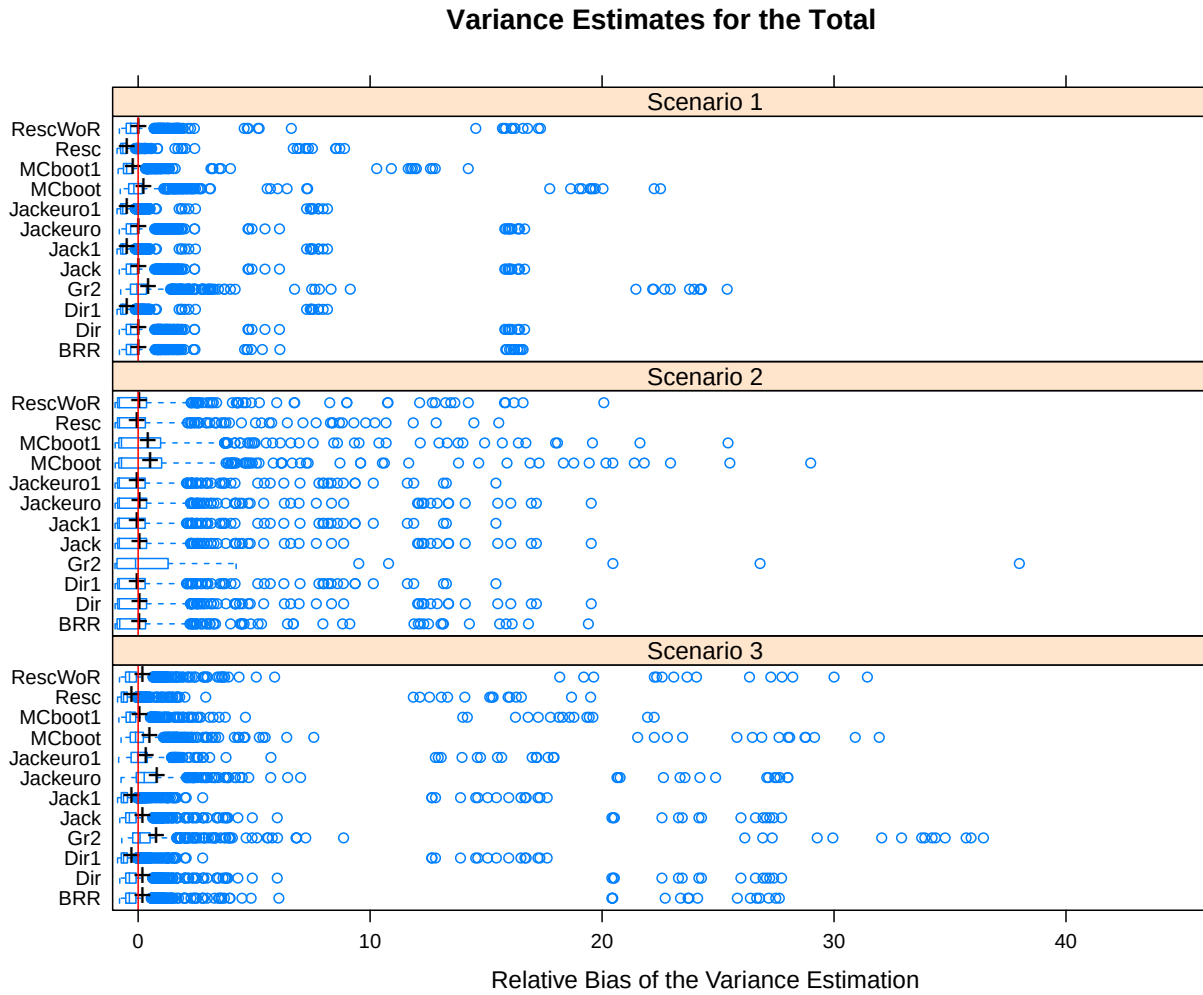
- Two stage sampling with stratification at the first stage, 25 strata
- 1. Stage: Drawing 4 PSU in each stratum (contains 8 PSU on average, altogether 200 PSU)
- 2. Stage: Proportional allocation of the sample size (1,000 USU) to the PSU (contains 500 USU on average, altogether 100,000 USU)

Experimental Study: Scenarios

- *Scenario 1* : Units within PSU are heterogeneous with respect to the variable of interest $Y \sim LN(10, 1.5^2)$, PSU are of equal size
- *Scenario 2* : Units within PSU are homogeneous with respect to the variable of interest, PSU are of equal size
- *Scenario 3* : Units within PSU are heterogeneous with respect to the variable of interest $Y \sim LN(10, 1.5^2)$, PSU are of unequal size

Variance Estimates for the Total





Second order inclusion probabilities

In case of unequal probability sampling designs, we also need the second order inclusion probabilities for variance estimation:

Second order inclusion probability

The probability that both elements i and j are drawn in the sample is denoted by

$$\pi_{ij} = \sum_{\mathcal{S} \in \mathbb{S}_{i,j}} \mathcal{I}(i,j \in \mathcal{S}) \cdot P(\mathcal{S}) \quad ,$$

and is called second order inclusion probability.

From this definition, we can conclude that $\pi_{ii} = \pi_i$ holds.

Sen-Yates-Grundy variance estimator

Alternatively, for designs with fixed sample sizes, we can use the Sen-Yates-Grundy variance estimator:

$$V_{\text{SYG}}(\widehat{\tau}) = -\frac{1}{2}\sum_{\substack{i,j\in\mathcal{U}\\ i\neq j}}\sum (\pi_{ij} - \pi_i \cdot \pi_j) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

$$= \sum_{\substack{i,j\in\mathcal{U}\\ i<j}}\sum (\pi_i \cdot \pi_j - \pi_{ij}) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

As unbiased estimator can be applied:

$$\widehat{V}_{\text{SYG}}(\widehat{\tau}) = \sum_{\substack{i,j\in\mathcal{S}\\ i<j}}\sum \frac{\pi_i \cdot \pi_j - \pi_{ij}}{\pi_{ij}} \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Examples approximations

- In presence of a sampling design with maximum entropy the following general approximation of the variance results:

$$V_{approx}(\hat{\tau}) = \sum_{i \in \mathcal{U}} \frac{b_i}{\pi_i^2} \cdot (y_i - y_i^*)^2$$

$$y_i^* = \pi_i \cdot \frac{\sum_{j \in \mathcal{U}} b_i \cdot y_i / \pi_j}{\sum_{j \in \mathcal{U}} b_j}$$

- Hájek approximation:

$$b_i^{Hajek} = \frac{\pi_i \cdot (1 - \pi_i) \cdot N}{N - 1}$$

Cf. Matei and Tillé (2005) or Hulliger et. al (2011)

Linearization of nonlinear statistics

Suppose we have d different study variables and we want to estimate parameter θ of the finite population $\mathcal{U}$ of size N, which has the following form

$$\theta = f(\boldsymbol{\tau}) \ , \tag{1}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k, \ldots, \tau_d)$ and $\tau_k = \sum_{i \in \mathcal{U}} y_{ki}$, with y_{ki} as the observation of k-th study variable of the i-th element in $\mathcal{U}$.
Then we substitute the unknown parameter vector $\boldsymbol{\tau}$ in (1) by its estimate $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_k, \ldots, \hat{\tau}_d)$, which yields

$$\hat{\theta} = f(\hat{\boldsymbol{\tau}}) \ ,$$

with $\hat{\tau}_k = \sum_{i \in s} y_{ki} d_i$ as the estimated total of the k-th study variable and w_i is the design weight of the i-th element in $\mathcal{S}$.
Further, it is assumed that $\hat{\tau}_k$ is a consistent estimator of τ_k.

In case the function f is continuously differentiable up to order two at each point in the open set $\mathbb{S}$ containing $\boldsymbol{\tau}$ and $\hat{\boldsymbol{\tau}}$, we can use a Taylor expansion

$$\hat{\theta} - \theta = \sum_{k=1}^{d} \left[\frac{\partial f(p_1, \ldots, p_d)}{\partial \tau_k} \right]_{\mathbf{p}=\boldsymbol{\tau}} (\hat{\tau}_k - \tau_k) + R(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) \,, \qquad (2)$$

where

$$R(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) = \frac{1}{2!} \sum_{k=1}^{d} \sum_{l=1}^{d} \left[\frac{\partial^2 f(p_1, \ldots, p_d)}{\partial p_k \partial p_l} \right]_{\mathbf{p}=\ddot{\boldsymbol{\tau}}} (\hat{\tau}_k - \tau_k)(\hat{\tau}_l - \tau_l)$$

and $\ddot{\boldsymbol{\tau}}$ is in the interior of line segment $L(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}})$ joining $\boldsymbol{\tau}$ and $\hat{\boldsymbol{\tau}}$. For the remainder term R we have $R = O_p(r_n^2)$, where $r_n \to 0$ as $n \to \infty$. Further, we have $\hat{\theta} - \theta = O_p(r_n)$. Thus, in most applications it is common practice to regard R as negligible in (2) for sample sizes large enough.

This justifies the use of the following approximation:

$$\hat{\theta} - \theta \approx \sum_{k=1}^{d} \left[\frac{\partial f(p_1, \ldots, p_d)}{\partial \tau_k} \right]_{\mathbf{p}=\boldsymbol{\tau}} (\hat{\tau}_k - \tau_k) \ . \tag{3}$$

Note, that in expression (3) only the linear part of the Taylor series is kept. Now, we can use (3) to derive an approximation of the mean square error (MSE) of $\hat{\theta}$ which is given by

$$\begin{aligned} \text{MSE}\left(\hat{\theta}\right) &\approx V\left(\sum_{k=1}^{d} \left[\frac{\partial f(p_1, \ldots, p_d)}{\partial \tau_k} \right]_{\mathbf{p}=\boldsymbol{\tau}} \hat{\tau}_k \right) \\ &= \sum_{k=1}^{d} a_k^2 V(\hat{\tau}_k) + 2 \sum_{k=1}^{d} \sum_{\substack{l=1 \\ k<l}}^{d} a_k a_l Cov(\hat{\tau}_k, \hat{\tau}_l) \ , \end{aligned} \tag{4}$$

where $a_k = [\frac{\partial f(p_1, \ldots, p_d)}{\partial \tau_k}]_{\mathbf{p}=\boldsymbol{\tau}}$.

Because $\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$, where $\text{Bias}(\hat{\theta}) = \hat{\theta} - \theta$, we can approximate the variance of $\hat{\theta}$ by $\text{MSE}(\hat{\theta})$ since $V(\hat{\theta})$ is of higher order than $\text{Bias}(\hat{\theta})^2$ for unbiased or at least consistent estimators. Thus, we can use (4) as an approximation of the design variance of $\hat{\theta}$.

The methodology holds for *smooth* functions. For non-differentiable functions (e.g. quantiles), influence functions could be applied (cf. Deville, 1999). Finally, estimating equations may also lead to linearized variables (cf. Binder and Patak, 1994). Finally, *only* the linearized variables have to be derived in order to apply the linear methodology for non-linear statistics (e.g. poverty indicators).

To estimate (4), we could simply substitute the variances and covariances with their corresponding estimates. This, however, might become unpractical if d, the number of estimated totals in $\hat{\boldsymbol{\tau}}$, becomes large. To evade this problem, Woodruff (1971) suggested the following:

$$\begin{aligned}\text{MSE}\left(\hat{\theta}\right) &\approx V(\sum_{k=1}^{d} a_k \hat{\tau}_k) \approx V(\sum_{k=1}^{d} a_k \sum_{i=1}^{n} w_i y_{ki}) \\ &\approx V(\sum_{i=1}^{n} w_i \sum_{k=1}^{d} a_k y_{ki}) \approx V(\sum_{i=1}^{n} w_i z_i) \quad ,\end{aligned}$$

where

$$z_i = \sum_{k=1}^{d} a_k y_{ki} \qquad \text{with} \qquad \text{MSE}\left(\hat{\theta}\right) \approx V\left(\sum_{i \in s} w_i z_i\right) \quad . \qquad (5)$$

CLAN: Function of Totals

Andersson and Nordberg introduced easy to computer SAS macros in order to produce linearized values for functions of totals:

Let $\theta = \tau_1 \circ \tau_2$ a function of totals from $\circ \in \{+, -, \cdot, /\}$. Then

Operator	z transformation
$+$	$z_k = y_{1k} + y_{2k}$
$-$	$z_k = y_{1k} - y_{2k}$
$\cdot$	$z_k = \theta \cdot (y_{1k}/t_1 + y_{2k}/t_2)$
$/$	$z_k = \theta \cdot (y_{1k}/t_1 - y_{2k}/t_2)$

The proof follows from applying Woodruff's method. Now, any functions using the above operators of totals can be recursively developed, which can be integrated in software (cf. Andersson and Nordberg, 1994).

Evidence-based Policy Decision Based on Indicators

- Indicators are seen as *true* values
- In general, indicators are simply survey variables
- No modelling is used to
  - Improve quality and accuracy of indicators
  - Disaggregate values towards domains and areas
- Reading naively point estimator tables may lead to misinterpretations
  - Change (Münnich and Zins, 2011)
  - Benchmarking (change in European policy)

- How accurate are estimates for indicators (ARPR, RMPG, GINI, and QSR)?
- This leads to applying the adequate variance estimation methods

Linearization and Resampling Methods

The statistics in question (the Laeken indicators) are highly non-linear.

- Resampling methods
  Kovačević and Yung (1997)
  - Balanced repeated replication
  - Jackknife
  - Bootstrap
- Linearization methods
  - Taylor's method
  - Woodruff linearization, Woodruff (1971) or Andersson and Nordberg (1994)
  - Estimating equations, Kovačević and Binder (1997)
  - Influence functions, Deville (1999)
  - Demnati and Rao (2004)

Application to Poverty and Inequality Indicators

Using the linearized values for the statistics ARPR, GINI, and QSR to approximate their variances.

Calibrated weights w_i: z_i are residuals of the regression of the linearized values on the auxiliary variables used in the calibration (cf. Deville, 1999).

Indicator $\mathcal{I}$	Source
ARPR:	Deville (1999)
GINI:	Kovačević and Binder (1997)
QSR:	Hulliger and Münnich (2007)
RMPG:	Osier (2009)

For CI estimation, empirical likelihood methods may be preferable (cf. Berger, De La Riva Torres, 2015).

Resampling methods

- Idea: Draw repeatedly (sub-)samples from the sample in order to build the sampling distribution of the statistic of interest
- Estimate the variance as variability of the estimates from the resamples
- Methods of interest
  - Random groups
  - Balanced repeated replication (balanced half samples)
  - Jackknife techniques
  - Bootstrap techniques
- Some remarks:
  - If it works, one doesn't need second order statistics for the estimate
  - May be computationally exceptional
  - What does influence the quality of these estimates

Random groups

- Mahalanobis (1939)
- Aim: Estimate variance of statistic θ
- Random partition of sample into R groups (independently)
- $\widehat{\theta}_{(r)}$ denotes the estimate of θ on r-th subsample
- Random group points estimate:

$$\widehat{\theta}_{\text{RG}} = \frac{1}{R} \cdot \sum_{r=1}^{R} \widehat{\theta}_{(r)}$$

- Random group variance estimate:

$$\widehat{V}(\widehat{\theta}_{\text{RG}}) = \frac{1}{R} \cdot \frac{1}{R-1} \cdot \sum_{r=1}^{R} \left(\widehat{\theta}_{(r)} - \widehat{\theta}_{\text{RG}}\right)^2$$

- Random selection versus random partition!

Balanced repeated replication

- Originally we have two observations per stratum
- Random partioning of observations into two groups
- $\widehat{\theta}_r$ is the estimate of the r-th selection using the H *half samples*
- Instead of recalling all possible $R \ll 2^H$ replications, we use a balanced selection via Hadamard matrices
- We obtain:

$$\widehat{\theta}_{\text{BRR}} = \frac{1}{R} \cdot \sum_{r=1}^{R} \widehat{\tau}_r \text{ and } \widehat{V}_{\text{BRR}}\left(\widehat{\theta}\right) = \frac{1}{R} \sum_{r=1}^{R} \left(\widehat{\theta}_r - \widehat{\theta}\right)^2 \quad .$$

- May lead to highly variable variance estimates, especially when H is small (cf. Davison and Sardy, 2004). Repetition of random grouping may be useful (cf. Rao and Shao, 1996)
- Use special weighting techniques for improvements

Delete-1-Jackknife

- Resampling by omitting (deleting) one element in each resample
- $\widehat{\theta}_{\_i}$ is used in n resamples
- Originally designed for bias estimation

Bootstrap

- Resampling by subsamples of size n
- Number of resamples B is arbitrary
- WR *only*

The Jackknife

Originally, the Jackknife method was introduced for estimating the bias of a statistic (Quenouille, 1949).
Let $\widehat{\theta}(Y_1, \ldots, Y_n)$ be the statistic of interest for estimating the parameter θ. Then,

$$\widehat{\theta}_{-i} = \widehat{\theta}(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$$

is the corresponding statistic omitting the observation Y_i which is therefore based on $n-1$ observations. Finally, the delete-1-Jackknife (d1JK) bias for θ is

$$\widehat{\mathrm{B}}_{\mathrm{d1JK}}\left(\widehat{\theta}\right) = (n-1) \cdot \left(\frac{1}{n} \sum_{i \in \mathcal{S}} \widehat{\theta}_{-i} - \widehat{\theta} \right)$$

(cf. Shao und Tu, 1995).

The jackknife (continued)

From the bias follows immediately the Jackknife point estimate

$$\begin{aligned}\widehat{\theta}_{\mathsf{d1JK}} &= \widehat{\theta} - \widehat{\mathsf{B}}_{\mathsf{d1JK}}\left(\widehat{\theta}\right) \\ &= n \cdot \widehat{\theta} - \frac{n-1}{n} \sum_{i \in \mathcal{S}} \widehat{\theta}_{-i}\end{aligned}$$

which is a delete-1-Jackknife bias corrected estimate. This estimator is under mild smoothness conditions of order n^{-2}.

Jackknife variance estimation

Tukey (1958) defined the so-called jackknife pseudo values $\widehat{\theta}_i^* := n \cdot \widehat{\theta} - (n-1) \cdot \widehat{\theta}_{-i}$ which yield under the assumption of stochastic independency and approximately equal variance of the $\widehat{\theta}_i^*$. Finally

$$\begin{aligned}\widehat{\mathrm{V}}_{\mathrm{d1JK}}\left(\widehat{\theta}\right) &= \frac{1}{n(n-1)} \cdot \sum_{i \in \mathcal{S}} \left(\widehat{\theta}_i^* - \overline{\widehat{\theta}}^*\right)^2 \\ &= \frac{n-1}{n} \sum_{i \in \mathcal{S}} \left(\widehat{\theta}_{-i} - \frac{1}{n} \sum_{i \in \mathcal{S}} \widehat{\theta}_{-j}\right)^2 \quad .\end{aligned}$$

Problem: What is $\widehat{\theta}_i^*$ and $\widehat{\mathrm{V}}_{\mathrm{d1JK}}\left(\widehat{\theta}\right)$ for $\widehat{\theta} = \overline{Y}$?

Advantages and disadvantages of the jackknife

- Very good for *smooth* statistics
- Biased for the estimation of the median
- Needs special weights in stratified random sampling (missing independency of jackknife resamples)

$$\widehat{V}_{\text{d1JK,strat}}(\widehat{\theta}) = \sum_{h=1}^{H} \frac{(1-f_h)\cdot(n_h-1)}{n_h} \cdot \sum_{i=1}^{n_h} \left(\widehat{\theta}_{h,-i} - \overline{\widehat{\theta}}_h\right)^2$$

  where $-i$ indicates the unit i that is left out.
- Specialized procedures are needed for (really) complex designs (cf. Rao, Berger, and others)
- Huge effort in case of large samples sizes (n):
  - Grouped jackknife (m groups; cf. Kott and R-package EVER)
  - Delete-d-jackknife (m replicates with d sample observations eliminated simultaneously; $m \ll \binom{n}{d}$)

Bootstrap resampling

- Theoretical bootstrap
- Monte-Carlo bootstrap:
  Random selection of size n (SRS) yields

$$\widehat{V}_{\text{Boot,MC}} = \frac{1}{B-1}\sum_{i=1}^{B}\left(\widehat{\theta}^{*}_{n,i} - \frac{1}{B}\sum_{j=1}^{B}\widehat{\theta}^{*}_{n,j}\right)^2 \quad .$$

- Special adaptions are needed in complex surveys
- Insufficient estimates in WOR sampling and higher sample fractions

Monte-Carlo Bootstrap

Efron (1982):

1. Estimate $\widehat{F}$ as the empirical distribution function (non-parametric maximum likelihood estimation);
2. Draw bootstrap samples from $\widehat{F}$, that is

$$X_1^*, \ldots, X_n^* \overset{\text{IID}}{\sim} \widehat{F}$$

   of size n;
3. Compute the bootstrap estimate $\widehat{\tau}_{n,i}^* = \widehat{\tau}\left(X_1^*, \ldots, X_n^*\right)$;
4. Repeat 1. to 3. B times (B arbitrarily large) and compute finally the variance

$$\widehat{\mathrm{V}}_{\text{Boot,MC}} = \frac{1}{B-1} \sum_{i=1}^{B} \left(\widehat{\tau}_{n,i}^* - \frac{1}{B} \sum_{j=1}^{B} \widehat{\tau}_{n,j}^* \right)^2 .$$

Properties of the Monte-Carlo Bootstrap

The bootstrap variance estimates converge by the law of large numbers to the *true* (theoretical) bootstrap variance estimate (cf. Shao and Tu, 1995, S. 11)

$$\widehat{\mathrm{V}}_{\mathrm{Boot,MC}} \xrightarrow{\text{a.s.}} V_{\mathrm{Boot}} \quad .$$

Analogously, one can derive the bootstrap bias of the estimator by

$$\widehat{\mathrm{B}}_{\mathrm{Boot,MC}} = \frac{1}{B}\sum_{i=1}^{B} \widehat{\tau}^{*}_{n,i} - \widehat{\tau} \quad .$$

Bootstrap confidence intervals

- Via variance estimation

$$\left[\widehat{\tau} - \sqrt{\widehat{\mathrm{V}}_{\mathrm{Boot,MC}}\left(\widehat{\tau}\right)} \cdot z_{1-\alpha/2}; \widehat{\tau} - \sqrt{\widehat{\mathrm{V}}_{\mathrm{Boot,MC}}\left(\widehat{\tau}\right)} \cdot z_{\alpha/2}\right]$$

- Via bootstrap resamples:

$$z_1^* = \frac{\widehat{\tau}_1^* - \widehat{\tau}}{\sqrt{\widehat{\mathrm{V}}_{\mathrm{Boot,\ MC}}\left(\widehat{\tau}_1^*\right)}} \quad , \ldots , \quad z_B^* = \frac{\widehat{\tau}_B^* - \widehat{\tau}}{\sqrt{\widehat{\mathrm{V}}_{\mathrm{Boot,\ MC}}\left(\widehat{\tau}_B^*\right)}}$$

From this empirical distribution, one can calculate the $\alpha/2$- and $(1-\alpha/2)$ quantiles $z^*_{\alpha/2}$ and $z^*_{1-\alpha/2}$ respectively by

$$\left[\widehat{\tau} - \sqrt{\widehat{\mathrm{V}}_{\mathrm{Boot,MC}}\left(\widehat{\tau}\right)} \cdot z^*_{B(1-\alpha/2)}; \widehat{\tau} - \sqrt{\widehat{\mathrm{V}}_{\mathrm{Boot,MC}}\left(\widehat{\tau}\right)} \cdot z_{B\alpha/2}\right]$$

This is referred to as the *studentized* bootstrap confidence interval.

Rescaling bootstrap

- *Rescaling bootstrap*: In case of multistage sampling only the first stage is considered. l^* (must be chosen) instead of l PSU are drawn with replacement (see Rao, Wu and Yue, 1992, Rust, 1996) The weights are adjusted by:

$$w_{qi}^* = \left[\left(1 - \left(\frac{l^*}{l-1}\right)^{1/2}\right) + \left(\frac{l^*}{l-1}\right)^{1/2} \cdot \left(\frac{l}{l^*}\right) \cdot r_q\right] \cdot w_{qi}.$$

- *Rescaling bootstrap without replacement*: From the l units of the sample, $l^* = \lfloor l/2 \rfloor$ units are drawn without replacement (see Chipperfield and Preston, 2007). In case of single stage sampling, the weights are adjusted by:

$$w_i^* = \left(1 - \lambda + \lambda \cdot \frac{n}{n^*} \cdot \delta_i\right) \cdot w_i, \text{with } \lambda = \sqrt{n^* \cdot \frac{(1-f)}{(n-n^*)}},$$

where δ_i is 1 when element i is chosen and 0 otherwise. For multistage designs (cf. Preston, 2009) the weights are adjusted at each stage by adding the term $-\lambda_G \cdot (\prod_{g=1}^{G-1} \sqrt{(n_g/n_g^*)} \cdot \delta_g) + \lambda_G \cdot (\prod_{g=1}^{G-1} \sqrt{(n_g/n_g^*)} \cdot \delta_g) \cdot (n_G/n_G^*) \cdot \delta_g$ at each stage G with $\lambda_G = \sqrt{n_G^* (\prod_{g=1}^{G-1} f_g) \cdot \frac{(1-f_G)}{(n_G - n_G^*)}}$.

Example: Rescaling bootstrap WOR for a three stage design

First stage:

$$w^*_{hq} = \left(1 - \lambda_h + \lambda_h \cdot \frac{l_h}{l^*_h} \cdot \delta_{hq}\right) \cdot w_{hq},$$

where $\lambda_h = \sqrt{l^*_h \cdot \frac{(1 - f_h)}{(l_h - l^*_h)}}$

and δ_{hq} is 1 when PSU q in stratum h is drawn and 0 else

Second stage:

$$w^*_{hqk} = (1 - \lambda_h + \lambda_h \frac{l_h}{l^*_h}\delta_{hq} - \lambda_{hq}\sqrt{\frac{l_h}{l^*_h}}\delta_{hq} + \lambda_{hq}\sqrt{\frac{l_h}{l^*_h}}\delta_{hq}\frac{m_{hq}}{m^*_{hq}}\delta_{hqk})w_{hqk}\frac{w_{hq}}{w^*_{hq}},$$

where $\lambda_{hq} = \sqrt{m^*_{hq} \cdot f_h \cdot \frac{(1 - f_{hq})}{(m_{hq} - m^*_{hq})}}$

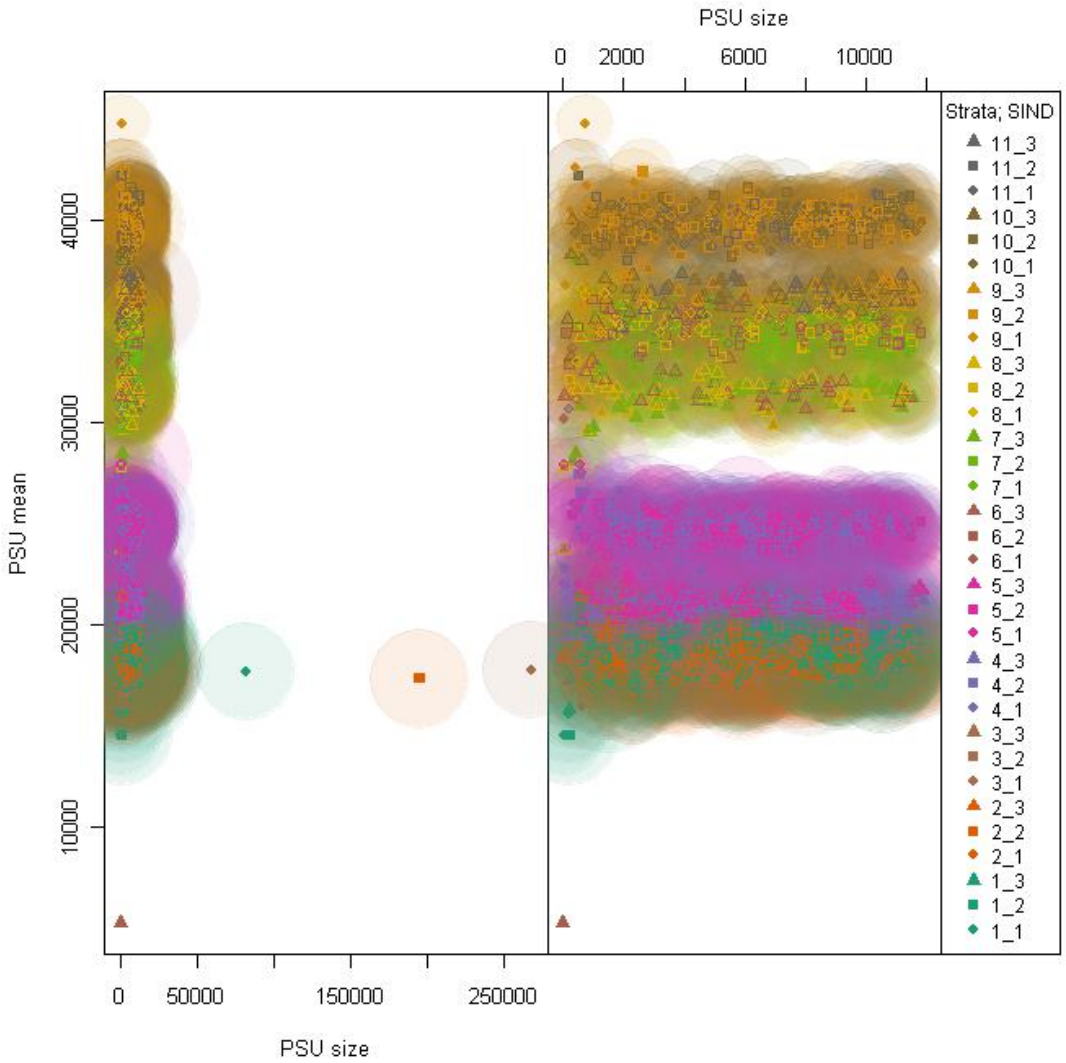
and δ_{hqk} is 1 when SSU k in PSU q in stratum h is drawn and 0 else

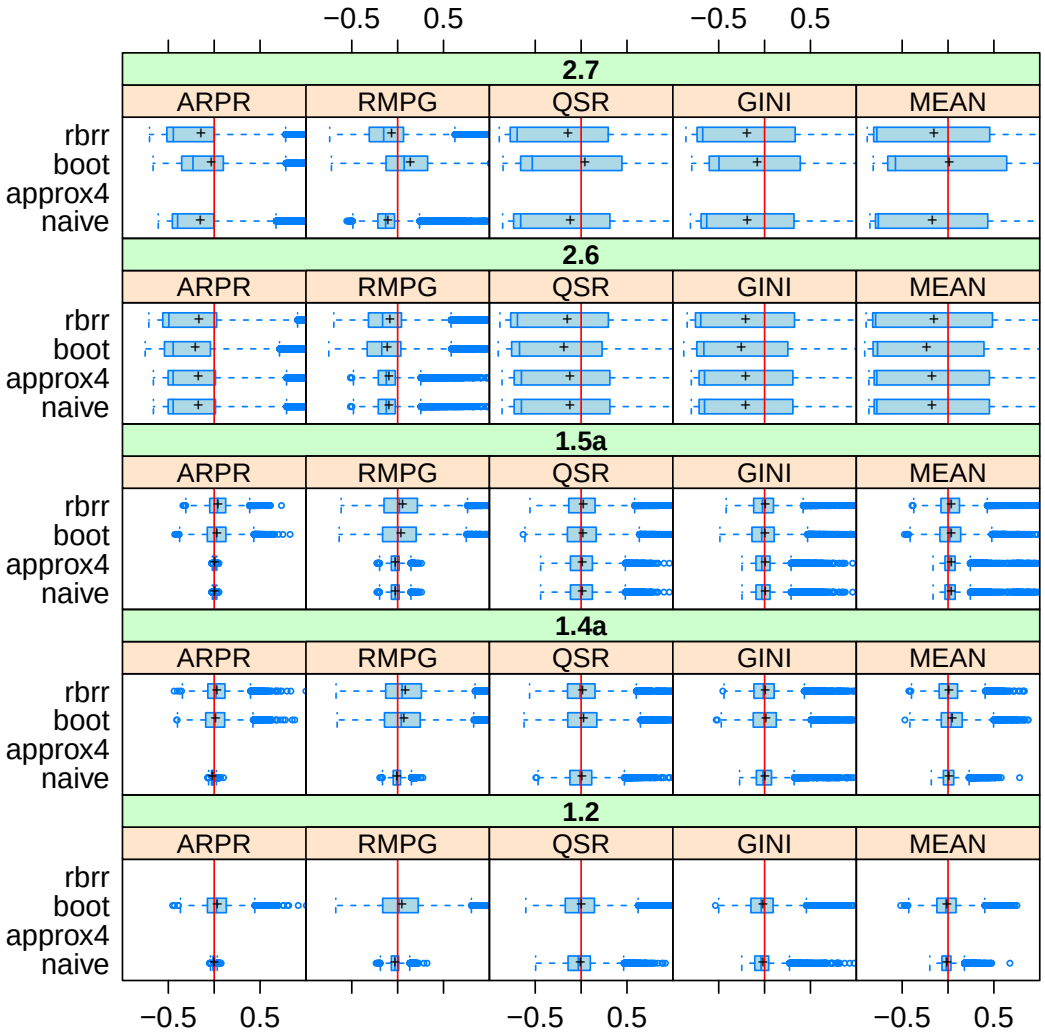
cf. Preston (2009)

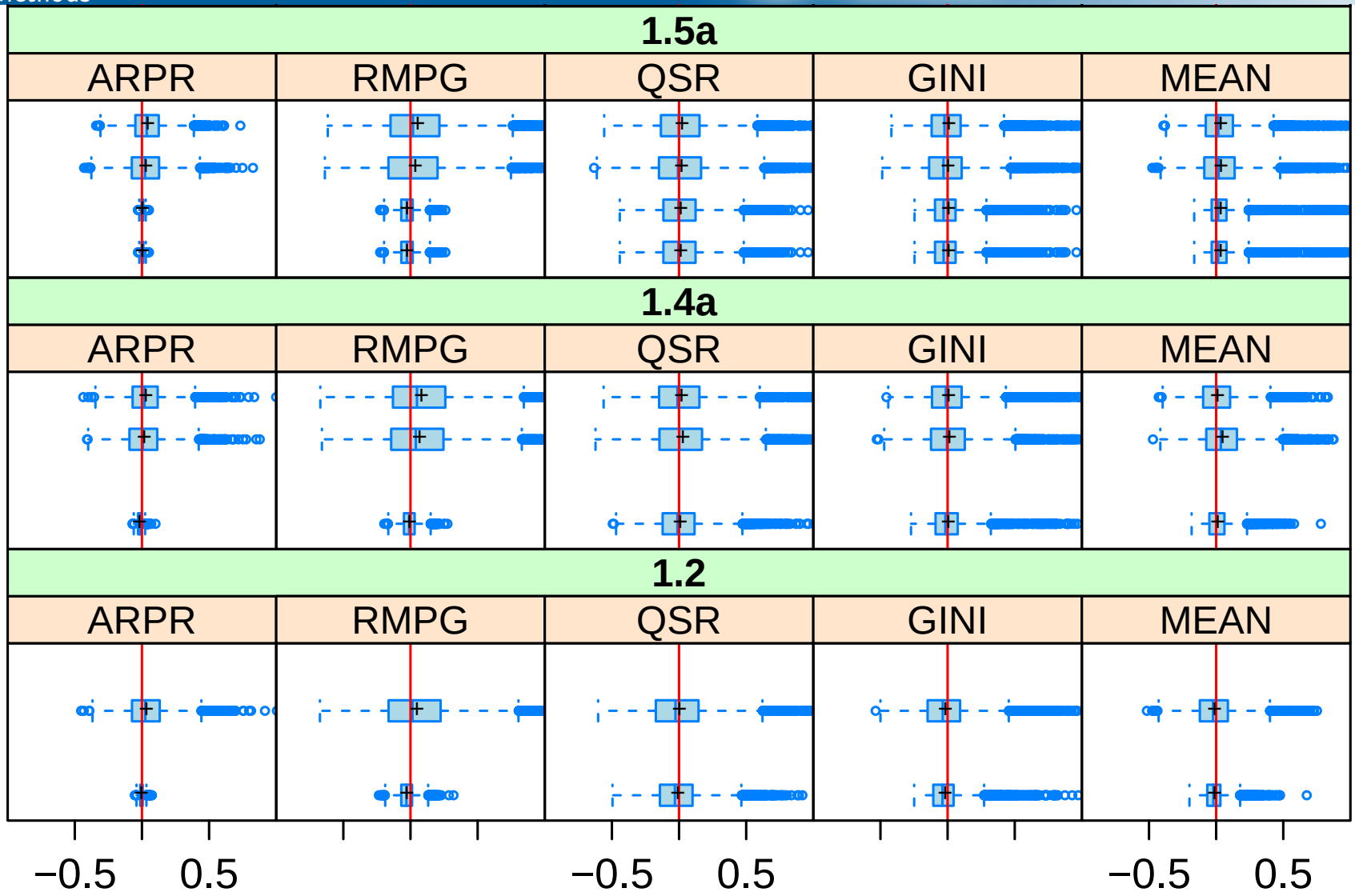
Comparison (cf. Bruch et al., 2011)

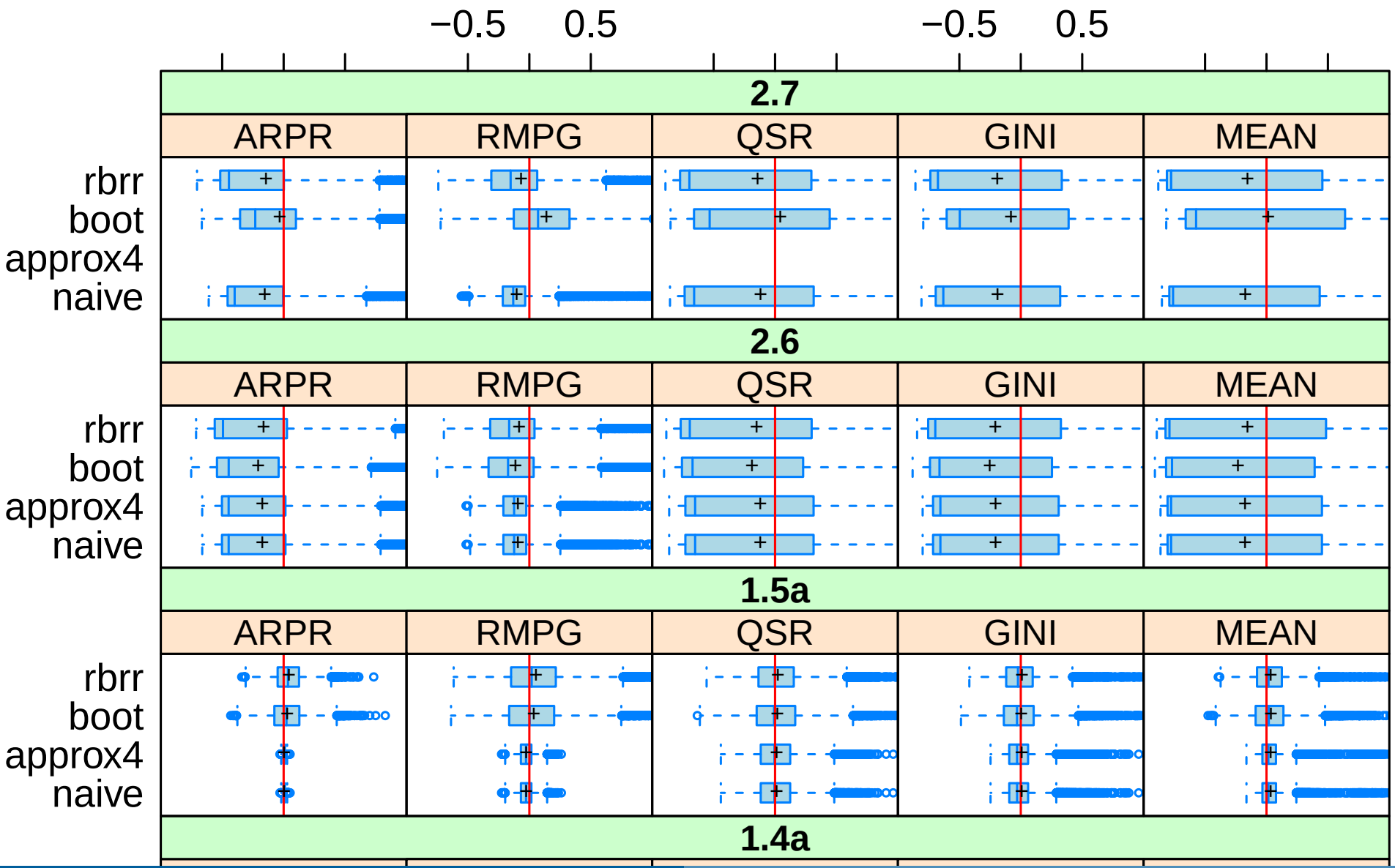
Method	BRR (Basic Model)	BRR (Group)	Delete-1 Jackknife	Delete-d Jackknife	Delete-a-Group Jackknife	Monte Carlo Bootstrap	Rescaling Bootstrap	Rescaling Bootstrap WoR
Statistic	Smooth and non-smooth	Smooth and non-smooth	Only for smooth statistics	Smooth and non-smooth	Smooth and non-smooth	Smooth and non-smooth	Smooth and non-smooth	Smooth and non-smooth
Stratification	Only when 2 elements per stratum	Required	Appropriate	Appropriate	Appropriate	Appropriate	Appropriate	Appropriate
Unequal Probability Sampling	Wolter (2007, p. 113)	Not considered	Berger (2007)	Not considered	Not considered	The ordinary Monte Carlo Bootstrap may lead to biased variance estimates	Not considered	Not considered
Sampling WR/WoR	WR	WoR	WR/WoR	WR/WoR	WR/WoR	WR	WR	WoR
FPC	Not considered	Considered	Possible	Possible	Possible	Not considered	Not considered	Considered

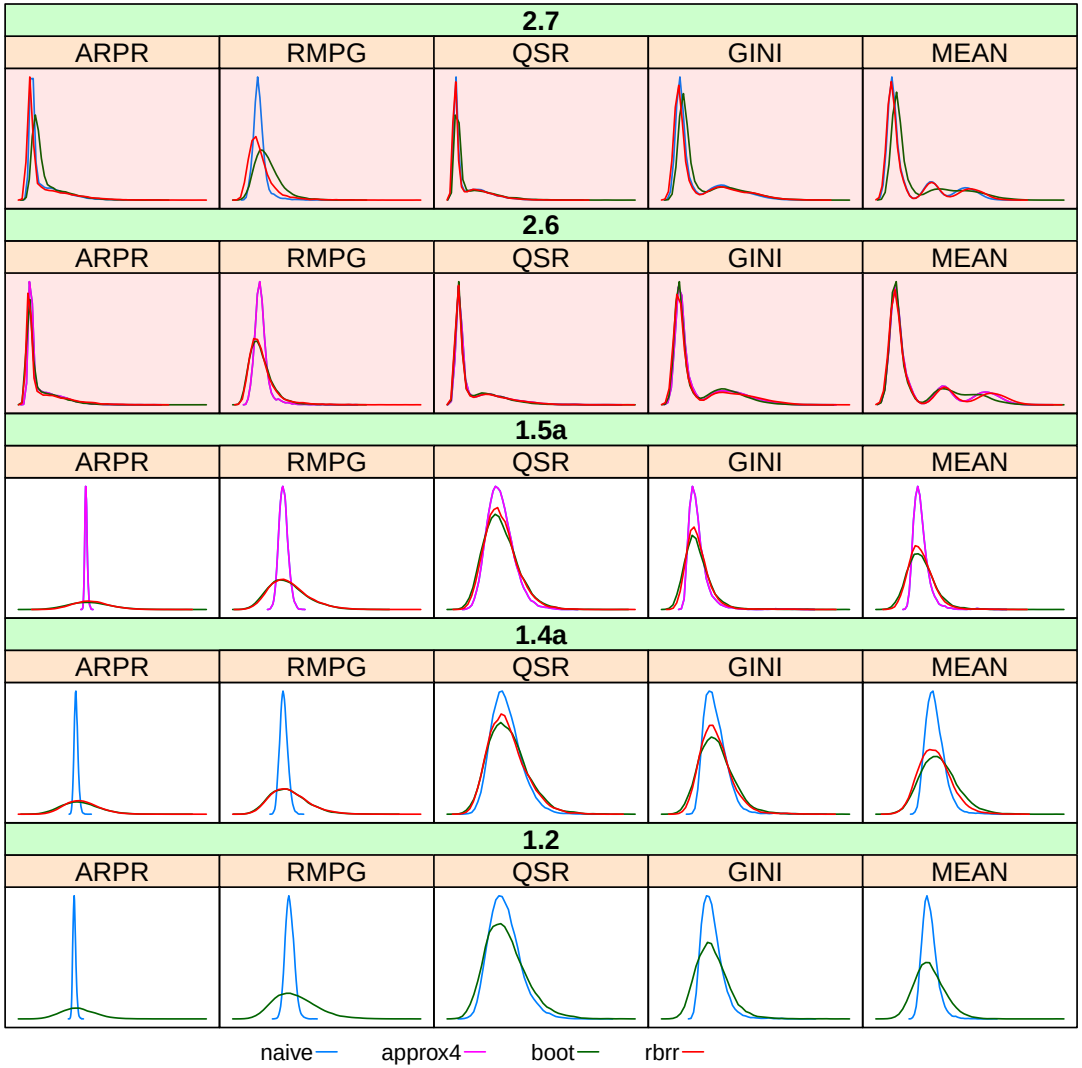
Characteristics of the AMELI universe

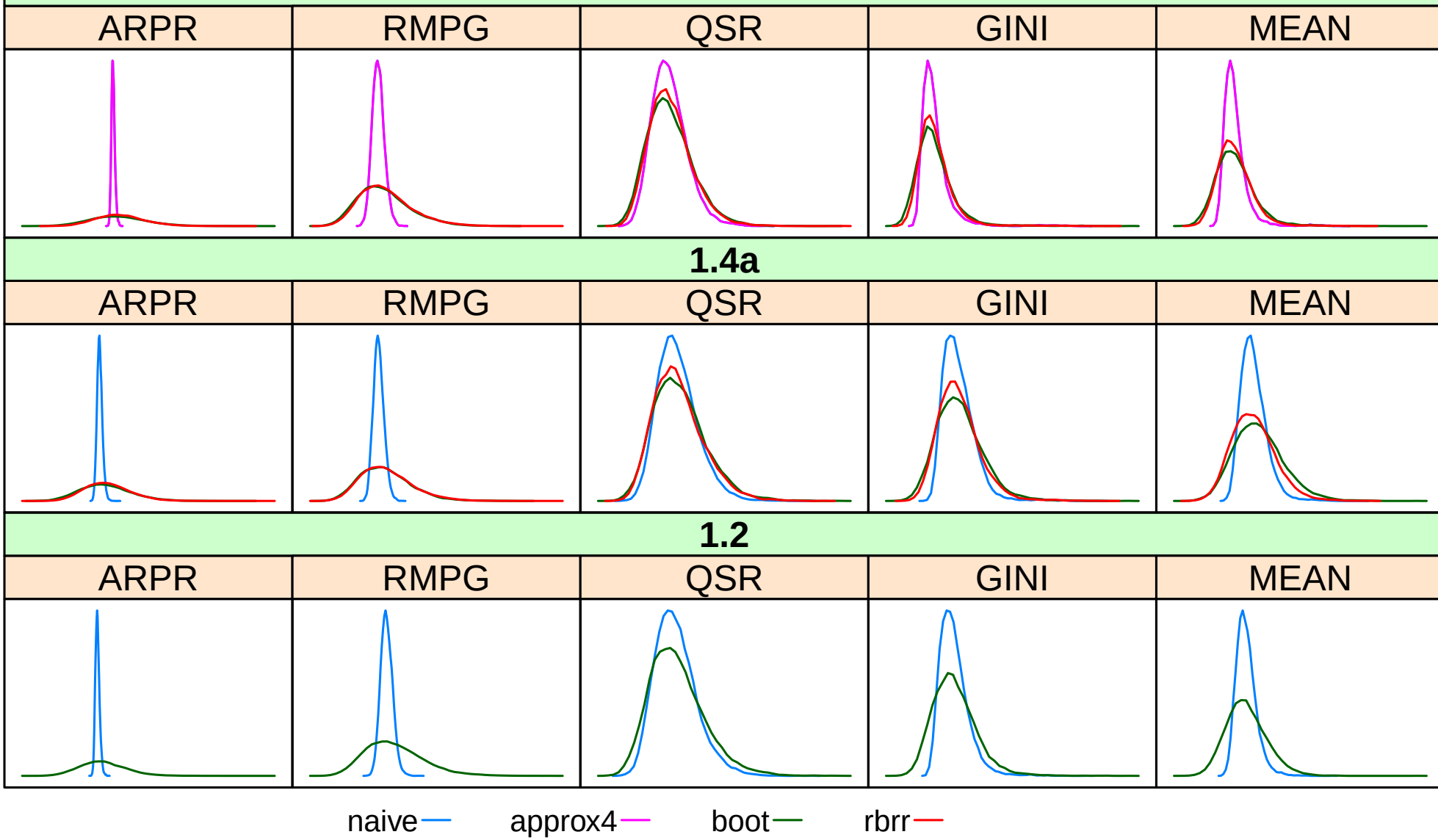


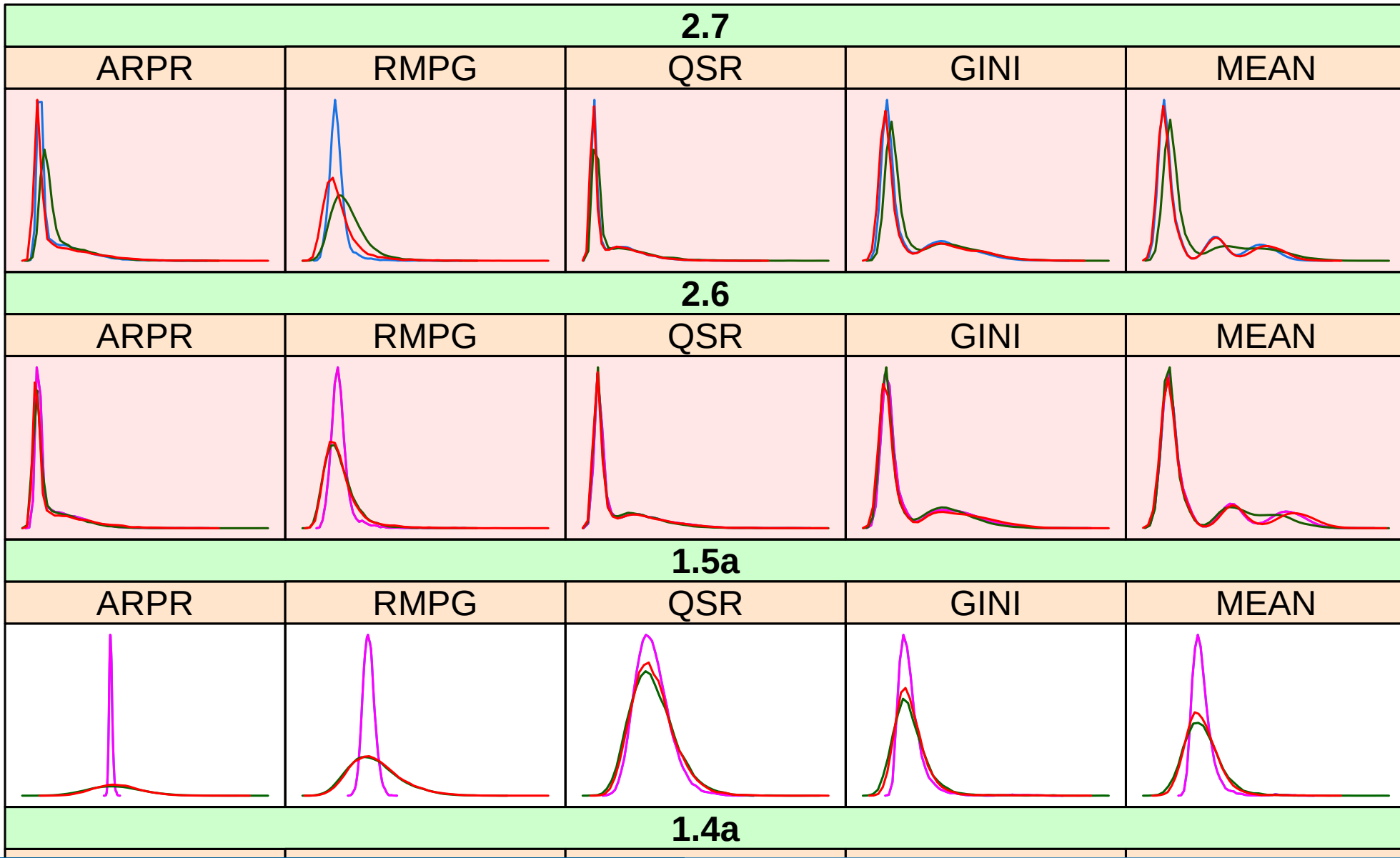

−0.5
0.5
2.7
2.6
1.5a
1.4a
1.2
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive


1.5a
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive
1.4a
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive
1.2
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive
−0.5
0.5
−0.5
0.5
−0.5
0.5


−0.5
0.5
−0.5
0.5
2.7
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive
2.6
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive
1.5a
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive
1.4a


2.7
ARPR
RMPG
QSR
GINI
MEAN
2.6
ARPR
RMPG
QSR
GINI
MEAN
1.5a
ARPR
RMPG
QSR
GINI
MEAN
1.4a
ARPR
RMPG
QSR
GINI
MEAN
1.2
ARPR
RMPG
QSR
GINI
MEAN
naive
approx4
boot
rbrr


1.3a
ARPR
RMPG
QSR
GINI
MEAN
1.4a
ARPR
RMPG
QSR
GINI
MEAN
1.2
ARPR
RMPG
QSR
GINI
MEAN
naive
approx4
boot
rbrr


2.7
ARPR
RMPG
QSR
GINI
MEAN
2.6
ARPR
RMPG
QSR
GINI
MEAN
1.5a
ARPR
RMPG
QSR
GINI
MEAN
1.4a

Coverage Rates (in %) of Indicator Estimates

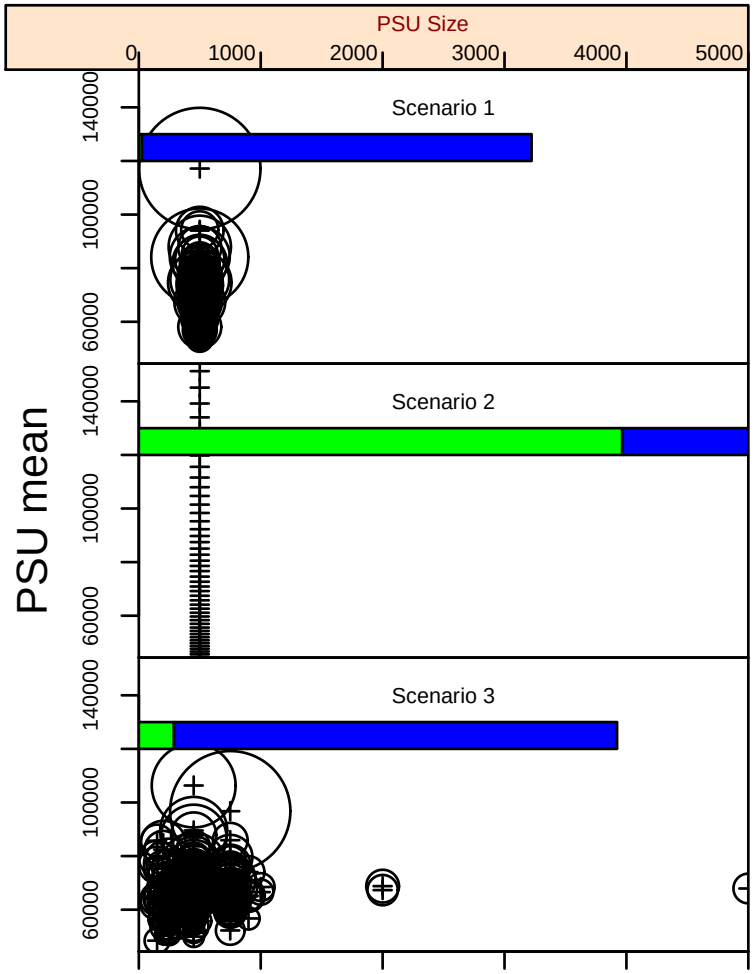
Direct/appr.	1.2	1.4a	1.5a	2.6	2.7
ARPR	95.070	94.700	94.950	89.340	90.640
RMPG	94.640	94.790	94.550	92.930	92.650
QSR	94.620	95.260	94.850	83.880	83.690
GINI	94.440	95.090	95.140	84.230	85.550
MEAN	94.850	95.070	95.320	78.720	79.960
Bootstrap	1.2	1.4a	1.5a	2.6	2.7
ARPR	95.100	94.910	94.810	87.850	93.070
RMPG	94.410	94.750	94.600	92.390	94.940
QSR	94.280	95.180	94.220	82.210	88.260
GINI	94.240	94.770	94.660	81.890	90.070
MEAN	94.620	95.260	95.090	77.630	90.340

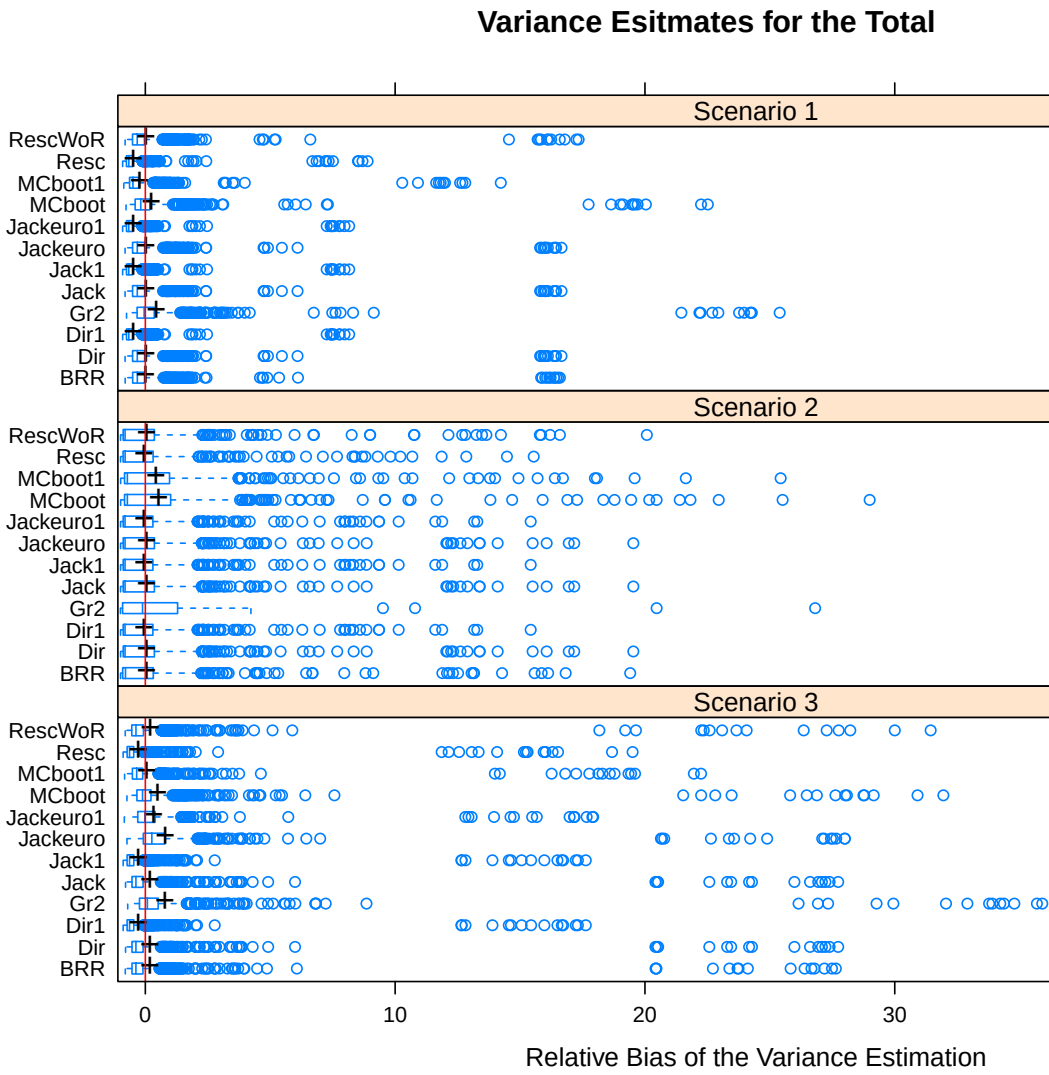
Experimental Study: Sampling Design

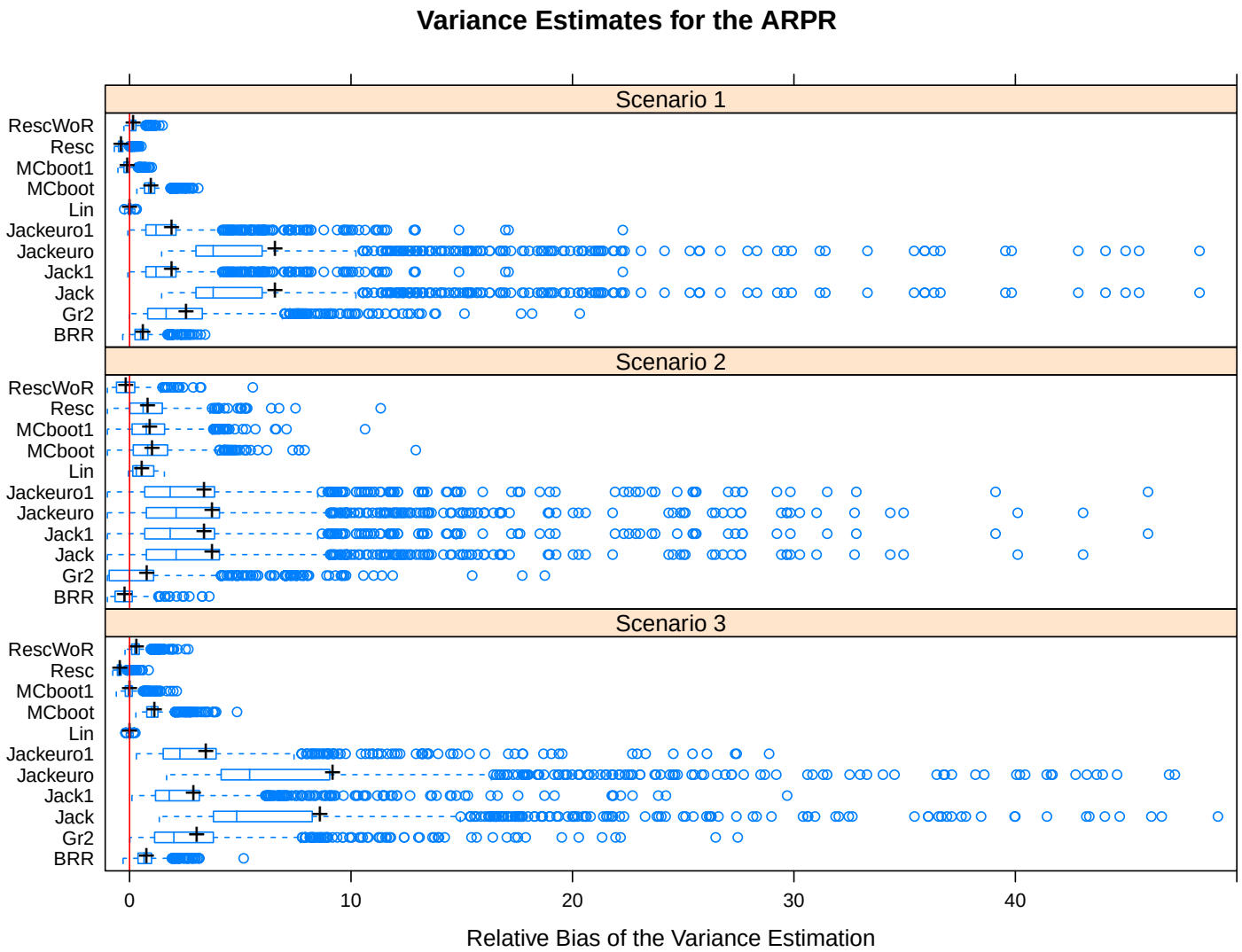
- Two stage sampling with stratification at the first stage, 25 strata
- 1. Stage: Drawing 4 PSU in each stratum (contains 8 PSU in average, altogether 200 PSU)
- 2. Stage: Proportional allocation of the sample size (1,000 USU) to the PSU (contains 500 USU in average, altogether 100,000 USU)

Experimental Study: Scenarios

- *Scenario 1* : Units within PSU are heterogeneous with respect to the variable of interest $Y \sim LN(10, 1.5^2)$, PSU are of equal size
- *Scenario 2* : Units within PSU are homogeneous with respect to the variable of interest, PSU are of equal size
- *Scenario 3* : Units within PSU are heterogeneous with respect to the variable of interest $Y \sim LN(10, 1.5^2)$, PSU are of unequal size


PSU Size
0
1000
2000
3000
4000
5000
Scenario 1
Scenario 2
Scenario 3
PSU mean
60000
100000
140000


Variance Esitmates for the Total
Scenario 1
Scenario 2
Scenario 3
RescWoR
Resc
MCboot1
MCboot
Jackeuro1
Jackeuro
Jack1
Jack
Gr2
Dir1
Dir
BRR
0
10
20
30
Relative Bias of the Variance Estimation


Variance Estimates for the ARPR
Scenario 1
Scenario 2
Scenario 3
RescWoR
Resc
MCboot1
MCboot
Lin
Jackeuro1
Jackeuro
Jack1
Jack
Gr2
BRR
0
10
20
30
40
Relative Bias of the Variance Estimation

Replication weights

- Doing resampling methods by adjusting the weights
- Advantage: Partial anonymization
  only the design weights are required (may not be fully true)
- BRR: Adjusting weights by

$$w_{h,i}^{(r)} := \begin{cases} w_{hi} \cdot \left[1 + \left\{ \frac{(n_h - m_h) \cdot (1 - f_h)}{m_h} \right\}^{1/2} \right], & \delta_{rh} = 1, \\ w_{hi} \cdot \left[1 - \left\{ \frac{m_h \cdot (1 - f_h)}{n_h - m_h} \right\}^{1/2} \right], & \delta_{rh} = -1, \end{cases}$$

  where δ_{rh} indicates if the first or second group in stratum h in replication r is chosen and $m_h = \lfloor n_h/2 \rfloor$ (cf. Davison and Sardy, 2004)
- Delete-1-Jackknife: The weights of the deleted unit are 0, all others are computed by $\frac{n_h}{n_h - 1} \cdot w_{hi}$
- Monte-Carlo Bootstrap: Computing weights by $w_{hi} \cdot c_{hi}$ where c_{hi} indicates how often unit i in stratum h is drawn with replacement

Some further issues on variance estimation

Design effect Determines the loss or gain of a complex sampling design in contrast to SRS (ratio of variances).

Variance estimation under imputation

- Calibration methods for compensating non-response
- Variance estimation under single imputation
- Multiple imputation

Variance functionals Provides a *functional form* of variances for estimates (eg totals in tables). This helps avoiding the computation of many sophisticated variances estimates in practice on the expense of precision.