Variance estimation of some EU-SILC based indicators at regional level

Jean Monet Lecture in Pisa – 2018

Ralf Münnich
Trier University, Faculty IV
Chair of Economic and Social Statistics

Pisa, 08st May 2018

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
	$\mathrm{E}\,\widehat{\tau}$	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?

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?

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?

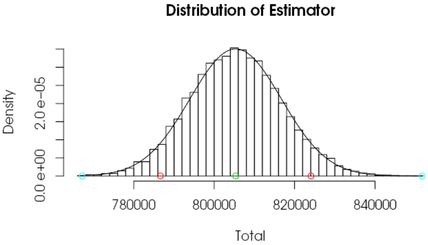
Why do we need variance estimation

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

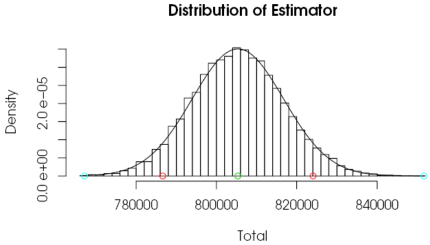
Do we need special measures for variance estimators or variance estimates?

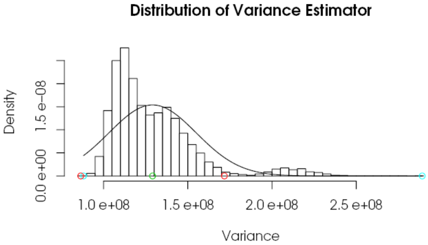
Example: Men in Hamburg



τ:	805258.00	N:	1669690				
$\hat{\tau}$:	805339.10	$V\hat{\tau}$:	1.29e+008	$E\left(\widehat{V}(\tau)\right)$:	1.29e+008	$V\left(\widehat{V}(\tau)\right)$:	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%):			

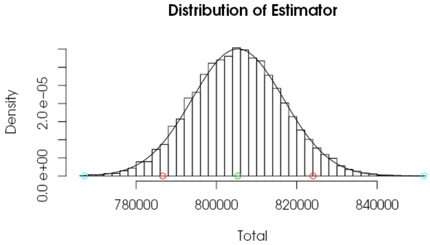
Example: Men in Hamburg

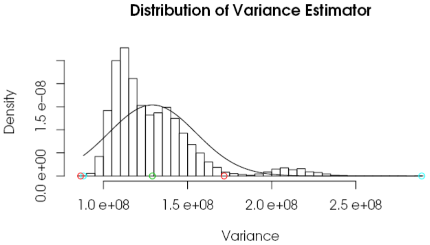




τ:	805258.00	N:	1669690				
$\hat{\tau}$:	805339.10	$V\hat{\tau}$:	1.29e+008	$E\left(\widehat{V}(\tau)\right)$:	1.29e+008	$V\left(\widehat{V}(\tau)\right)$:	6.72e+014
Bias Est:	81.10	MSE Est:	1.29e+008	Bias Var:	-3.78e+005	MSE Var:	6.72e+014
Skew Est:	0.0747	Curt Est:	3.0209	Skew Var:	1.8046	Curt Var: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

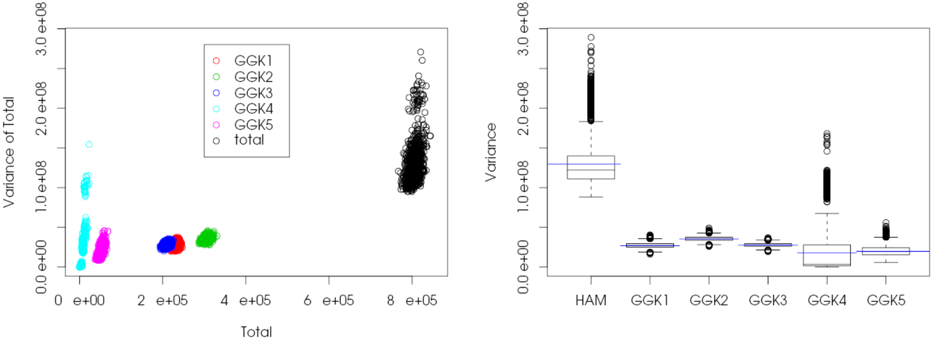
Example: Men in Hamburg





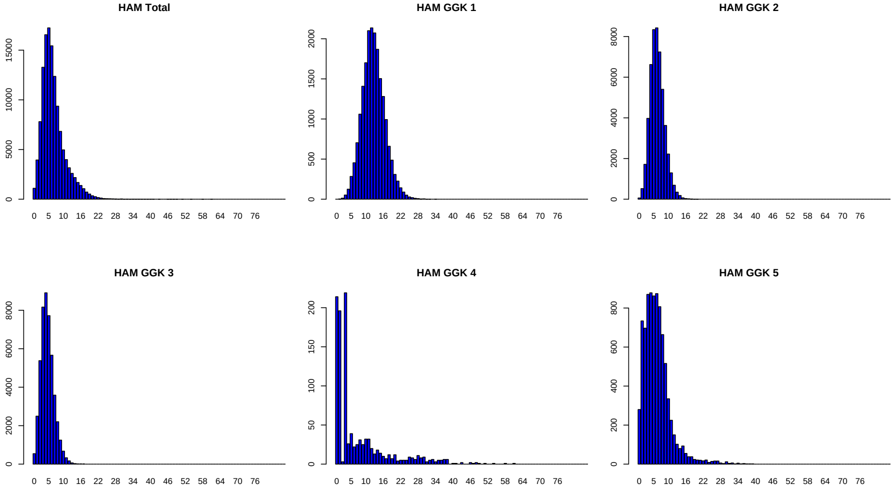
τ:	805258.00	N:	1669690				
$\hat{\tau}$:	805339.10	$V\hat{\tau}$:	1.29e+008	$E\left(\widehat{V}(\tau)\right)$:	1.29e+008	$V\left(\widehat{V}(\tau)\right)$:	6.72e+014
Bias Est:	81.10	MSE Est:	1.29e+008	Bias Var:	-3.78e+005	MSE Var:	6.72e+014
Skew Est:	0.0747	Curt Est:	3.0209	Skew Var:	1.8046	Curt Var: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

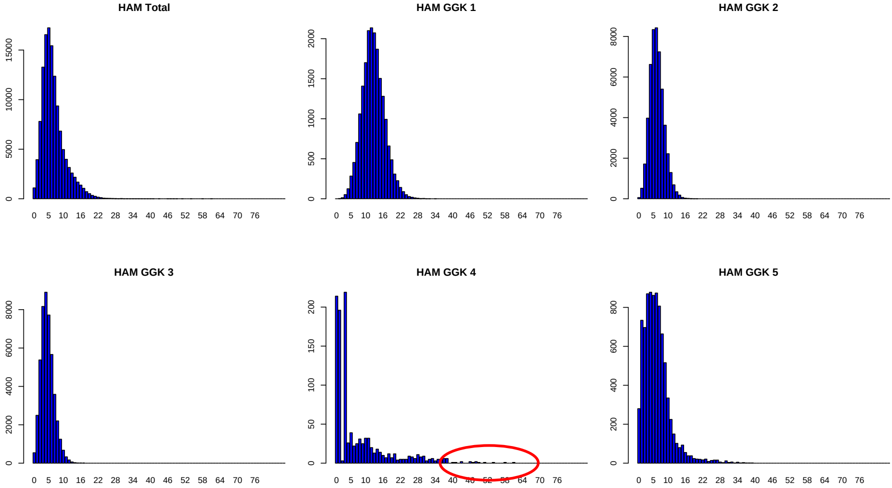


	GGK1	GGK2	GGK3	GGK4	GGK5	total
Persons	468293	651740	439745	9940	99970	1669690
Sampling units	173	446	414	10	75	1118

Distribution of Men in HAM (per SU)

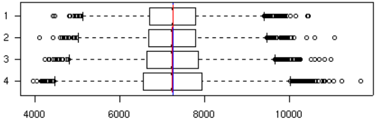


Distribution of Men in HAM (per SU)

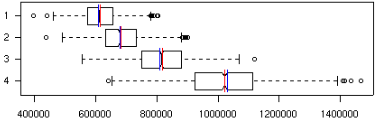


Unemployed women, 25 – 44

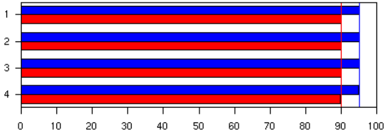
Raking estimator



variance estimator

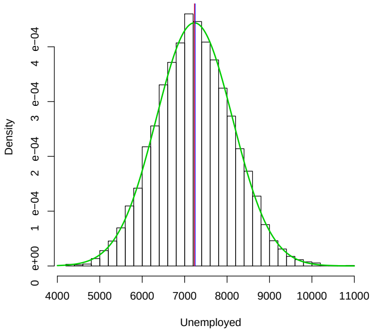


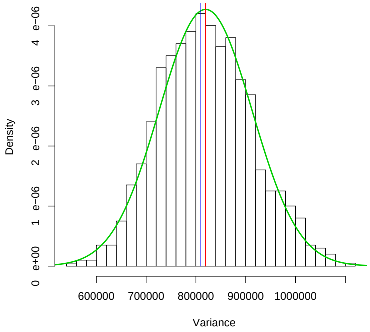
NR rates: 1: 5%, 2: 10%, 3: 25%, 4: 40%



95% 90%

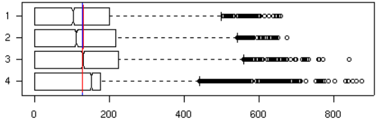
Unemployed women, 25 – 44, distribution of point and variance estimator (25% NR)



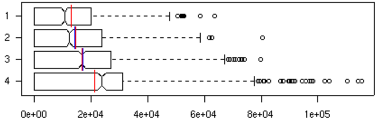


Unemployed women, 65 +

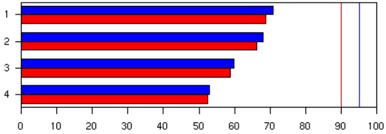
Raking estimator



variance estimator

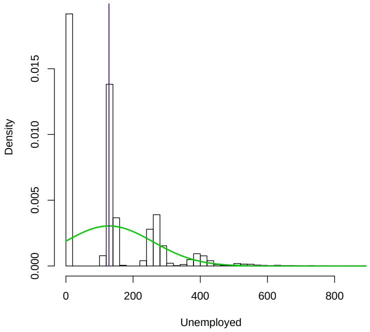


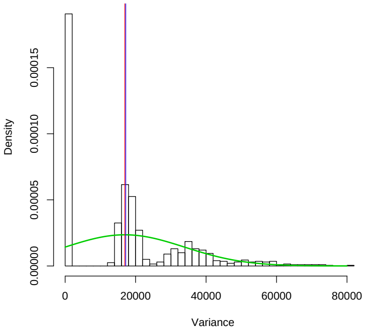
NR rates: 1: 5%, 2: 10%, 3: 25%, 4: 40%



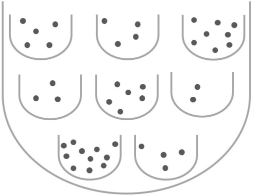
95% 90%

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



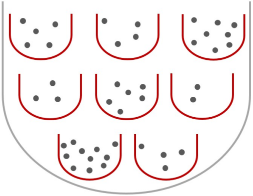


Framework of Two Stage Samples



Framework of Two Stage Samples

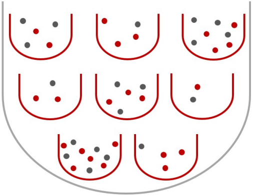
Stratified Sampling



	stage I	stage II
stratified sampling	100%	
single stage cluster sampling		
two stage cluster sampling		

Framework of Two Stage Samples

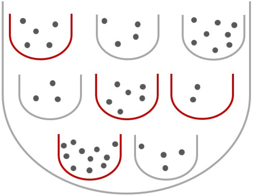
Stratified Sampling



	stage I	stage II
stratified sampling	100%	some
single stage cluster sampling		
two stage cluster sampling		

Framework of Two Stage Samples

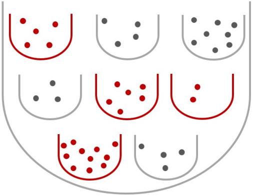
Single Stage Cluster Sampling



	stage I	stage II
stratified sampling	100%	some
single stage cluster sampling	some	
two stage cluster sampling		

Framework of Two Stage Samples

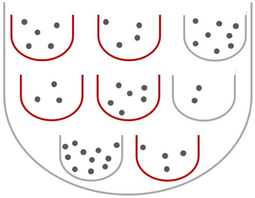
Single Stage Cluster Sampling



	stage I	stage II
stratified sampling	100%	some
single stage cluster sampling	some	100%
two stage cluster sampling		

Framework of Two Stage Samples

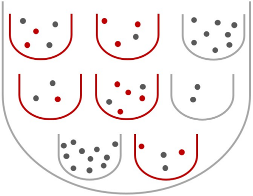
Two Stage Cluster Sampling



	stage I	stage II
stratified sampling	100%	some
single stage cluster sampling	some	100%
two stage cluster sampling	some	

Framework of Two Stage Samples

Two Stage Cluster Sampling



	stage I	stage II
stratified sampling	100%	some
single stage cluster sampling	some	100%
two stage cluster sampling	some	some

Direct variance estimator for Two Stage Sampling

- *Direct variance estimator*:

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

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

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):

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

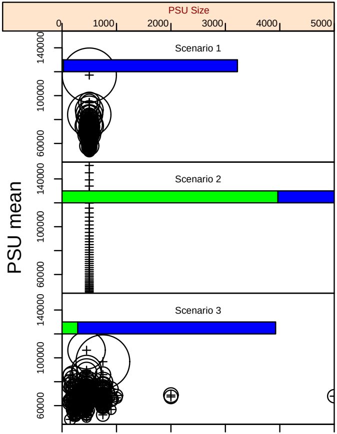
Experimental Study: Sampling Design

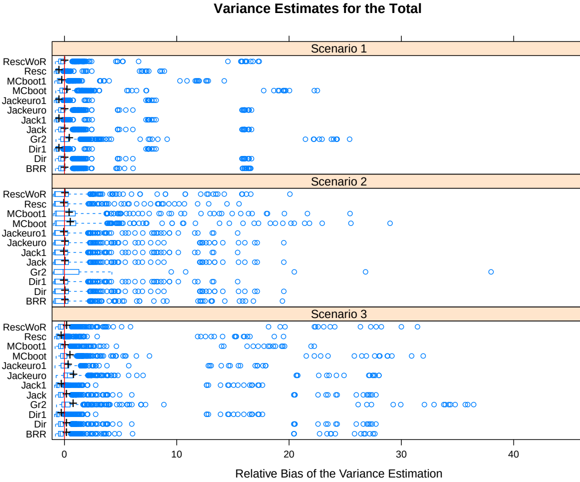
- Two stage sampling with stratification at the first stage, 25 strata
- 1. Stage: Drawing 4 PSU in each stratum (contains 8 PSU on average, altogether 200 PSU)
- 2. Stage: Proportional allocation of the sample size (1,000 ultimate sampling units, USU) to the PSU (contains 500 USU on average, altogether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

$$\pi_{ij} = \sum_{\mathcal{S} \in \mathbb{S}} P(\mathcal{S}) \cdot \mathbb{1}(i \in \mathcal{S}) \cdot \mathbb{1}(j \in \mathcal{S}) \qquad ,$$

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:

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

As unbiased estimator can be applied:

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

Examples approximations

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

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

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

- 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)

Main Idea & Example (cf. Lohr, 2010)

Non-linear statistic $f(\theta)$,
e.g. $f(\theta) = \theta \cdot (1 - \theta)$

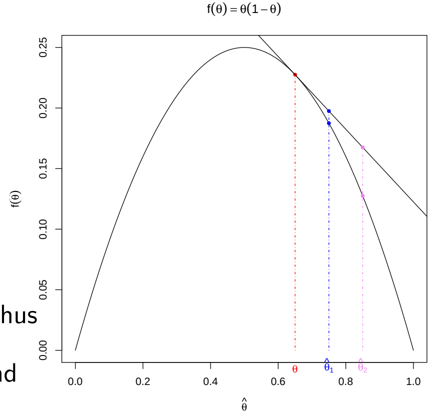
If $\widehat{\theta}$ is close to θ, then $\widehat{f(\theta)} = f\left(\widehat{\theta}\right)$ will be close to the tangent line with slope $f'(\theta) = 1 - 2 \cdot \theta$

Linearization using first derivative:

$$\widehat{f(\theta)} \approx f(\theta) + f'(p)\left(\widehat{\theta} - \theta\right) \text{, thus}$$

$$V\left(\widehat{f(\theta)}\right) \approx \left(f'(\theta)\right)^2 \cdot V\left(\left(\widehat{\theta} - \theta\right)\right) \text{ and}$$

$$\widehat{V}\left(\widehat{f(\theta)}\right) = \left(f'\left(\widehat{\theta}\right)\right)^2 \cdot \widehat{V}\left(\widehat{\theta}\right) \quad .$$



Taylor Linearization

Let $A \subseteq \mathbb{R}^p$ be an open set with $\boldsymbol{\tau}, \widehat{\boldsymbol{\tau}} \in A$ and let f be twice continuously differentiable on A. Then, with Taylor's theorem

$$
\begin{aligned}
\widehat{\Theta} &= f(\widehat{\boldsymbol{\tau}}) \\
&= f(\boldsymbol{\tau}) + (\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau})^t Df(\boldsymbol{\tau}) + R_n \\
&= \Theta + \left(\sum_{k=1}^{p} (\widehat{\tau}_k - \tau_k) c_k \right) + R_n ,
\end{aligned}
$$

with some remainder term R_n ($n \in \mathbb{N}$ sample size) and $c_k = \frac{\partial f}{\partial \tau_k}(\boldsymbol{\tau})$.

It can be shown that R_n is negligible for n "large enough" and this yields

$$\begin{aligned}\widehat{\Theta} &\approx \Theta + \sum_{k=1}^{p}(\widehat{\tau}_k - \tau_k)c_k \\ &= C + \sum_{k=1}^{p} c_k\widehat{\tau}_k .\end{aligned}$$

So $C + \sum_{k=1}^{p} c_k\widehat{\tau}_k$ is a linear approximation for $f(\widehat{\boldsymbol{\tau}})$ and this yields

$$V(\widehat{\Theta}) \approx V(C + \sum_{k=1}^{p} c_k\widehat{\tau}_k) = V(\sum_{k=1}^{p} c_k\widehat{\tau}_k).$$

Example: Ratio Estimator - I

Consider the ratio of two totals

$$R := f(\tau_1, \tau_2) = \frac{\tau_1}{\tau_2}.$$

A ratio estimator is given through

$$\widehat{R} = f(\widehat{\tau}_1, \widehat{\tau}_2) = \frac{\widehat{\tau}_1}{\widehat{\tau}_2}.$$

Example: Ratio Estimator - II

Using Taylor yields:

$$
\begin{aligned}
\widehat{R} \;=\; f(\widehat{\tau}_1, \widehat{\tau}_2) \;&\approx\; R + \sum_{k=1}^{2} (\widehat{\tau}_k - \tau_k) \frac{\partial f(\boldsymbol{\tau})}{\partial \tau_k} \\
&=\; R + (\widehat{\tau}_1 - \tau_1) \cdot \frac{\partial f(\tau_1, \tau_2)}{\partial \tau_1} + (\widehat{\tau}_2 - \tau_2) \cdot \frac{\partial f(\tau_1, \tau_2)}{\partial \tau_2} \\
&=\; R + (\widehat{\tau}_1 - \tau_1) \cdot \frac{1}{\tau_2} + (\widehat{\tau}_2 - \tau_2) \cdot \frac{-\tau_1}{(\tau_2)^2} \\
&=\; R + \frac{1}{\tau_2} \cdot \widehat{\tau}_1 - R - \frac{\tau_1}{(\tau_2)^2} \cdot \widehat{\tau}_2 + R \\
&=\; R + \frac{1}{\tau_2} \cdot \widehat{\tau}_1 - R \cdot \frac{1}{\tau_2} \widehat{\tau}_2
\end{aligned}
$$

Example: Ratio Estimator - III

Therefore

$$\begin{aligned} V(\widehat{R}) &\approx V\left(\frac{1}{\tau_2}\widehat{\tau}_1 - R \cdot \frac{1}{\tau_2}\widehat{\tau}_2\right) \\ &= \left(\frac{1}{\tau_2}\right)^2 \left[V(\widehat{\tau}_1) + (-R)^2\, V(\widehat{\tau}_2) + 2(-R)\,Cov(\widehat{\tau}_1, \widehat{\tau}_2)\right] \\ &= \frac{1}{(\tau_2)^2}\left[V(\widehat{\tau}_1) + R^2 V(\widehat{\tau}_2) - 2R\,Cov(\widehat{\tau}_1, \widehat{\tau}_2)\right]. \end{aligned}$$

CLAN: Function of Totals

Andersson and Nordberg introduced easy to compute 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.

$$\mathrm{V}(\hat{\theta}) \approx \mathrm{V}\Big(\sum_{i \in \mathcal{S}} w_i \cdot z_i\Big)$$

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)

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:

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

- Random group variance estimate:

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

- 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

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

(cf. Shao und Tu, 1995).

The jackknife (continued)

From the bias follows immediately the Jackknife point estimate

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

which is a delete-1-Jackknife bias corrected estimate. This estimator is under milde 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

$$\widehat{\mathrm{V}}_{\mathrm{d1JK}}\big(\widehat{\theta}\big) = \frac{1}{n(n-1)} \cdot \sum_{i \in \mathcal{S}} \Big(\widehat{\theta}_i^* - \overline{\widehat{\theta}}^*\Big)^2$$

$$= \frac{n-1}{n} \sum_{i \in \mathcal{S}} \left(\widehat{\theta}_{-i} - \frac{1}{n} \sum_{i \in \mathcal{S}} \widehat{\theta}_{-j}\right)^2 \qquad .$$

Problem: What is $\widehat{\theta}_i^*$ and $\widehat{\mathrm{V}}_{\mathrm{d1JK}}\big(\widehat{\theta}\big)$ 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)

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

  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

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

- 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
$$X_1^*, \ldots, X_n^* \quad \overset{\text{IID}}{\sim} \quad \widehat{F}$$
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
$$\widehat{\mathrm{V}}_{\text{Boot,MC}} = \frac{1}{B-1} \sum_{i=1}^{B} \left(\widehat{\tau}_{n,i}^* - \frac{1}{B} \sum_{j=1}^{B} \widehat{\tau}_{n,j}^*\right)^2 \qquad .$$

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)

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

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

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

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. I^* (must be chosen) instead of I PSU are drawn with replacement (cf. Rao, Wu and Yue, 1992, Rust, 1996) The weights are adjusted by:

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

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

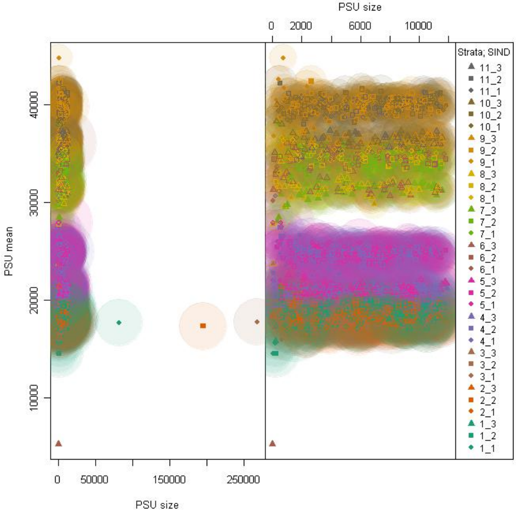
$$w_i^* = \left(1 - \lambda + \lambda \cdot \frac{n}{n^*} \cdot \delta_i\right) \cdot w_i, \text{with } \lambda = \sqrt{n^* \cdot \frac{(1-f)}{(n-n^*)}},$$

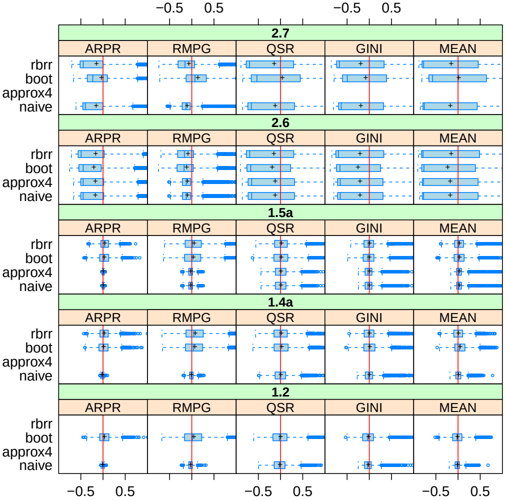
where δ_i is 1 when element i is chosen and 0 otherwise. For multistage designs (cf. Preston, 2009) the weights are adjusted at each stage by adding the term $-\lambda_G \cdot (\prod_{g=1}^{G-1} \sqrt{(n_g/n_g^*)} \cdot \delta_g) + \lambda_G \cdot (\prod_{g=1}^{G-1} \sqrt{(n_g/n_g^*)} \cdot \delta_g)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

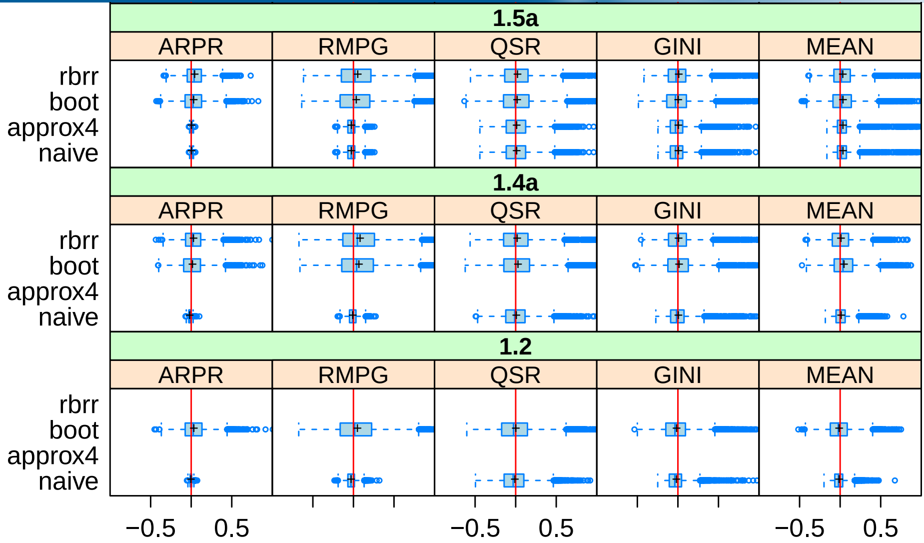
Comparison (cf. Bruch et al., 2011)

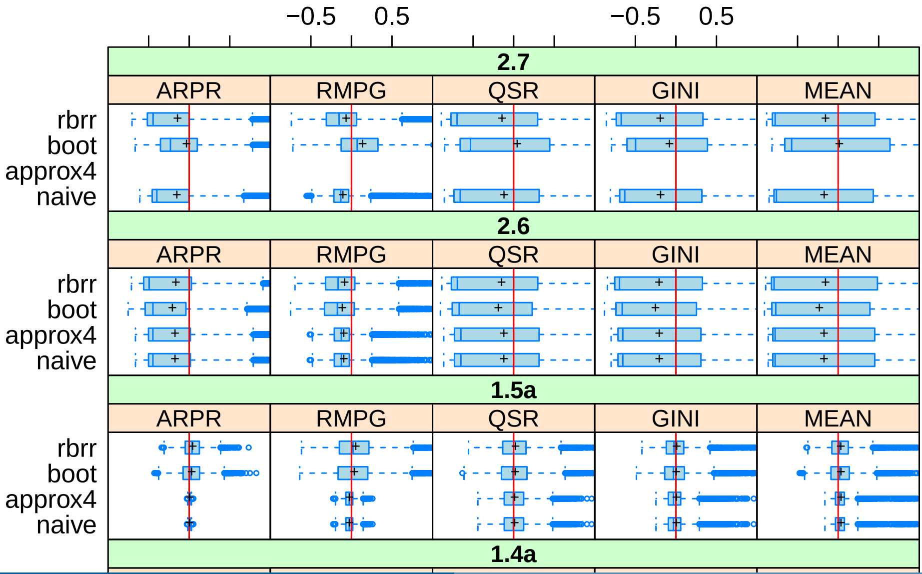
Method	BRR (Basic Model)	BRR (Group)	Delete-1 Jackknife	Delete-d Jackknife	Delete-a-Group Jackknife	Monte Carlo Bootstrap	Rescaling Bootstrap	Rescaling Bootstrap WoR
Statistic	Smooth and non-smooth	Smooth and non-smooth	Only for smooth statistics	Smooth and non-smooth	Smooth and non-smooth	Smooth and 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

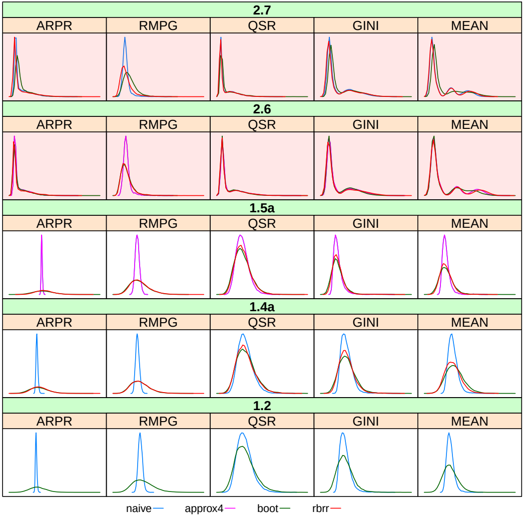
Characteristics of the AMELI universe

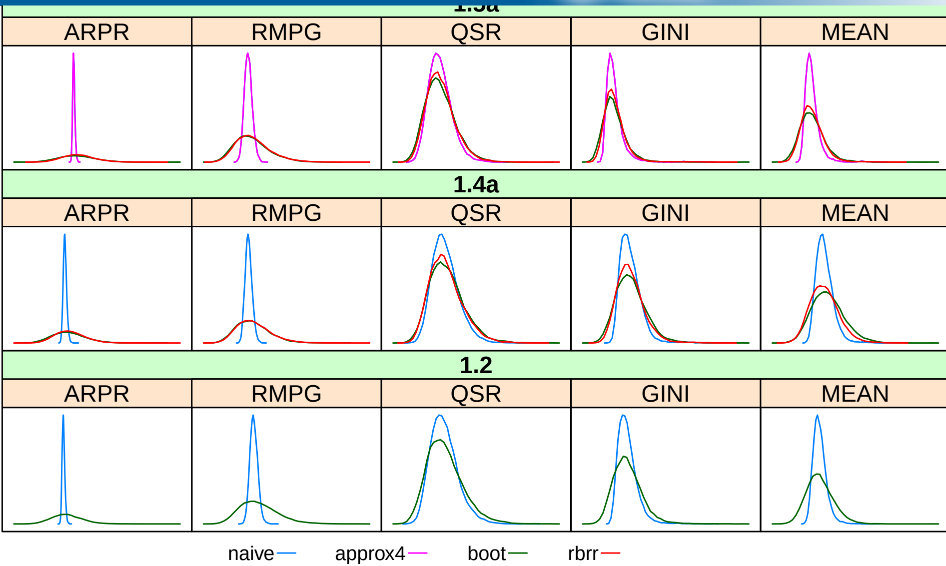


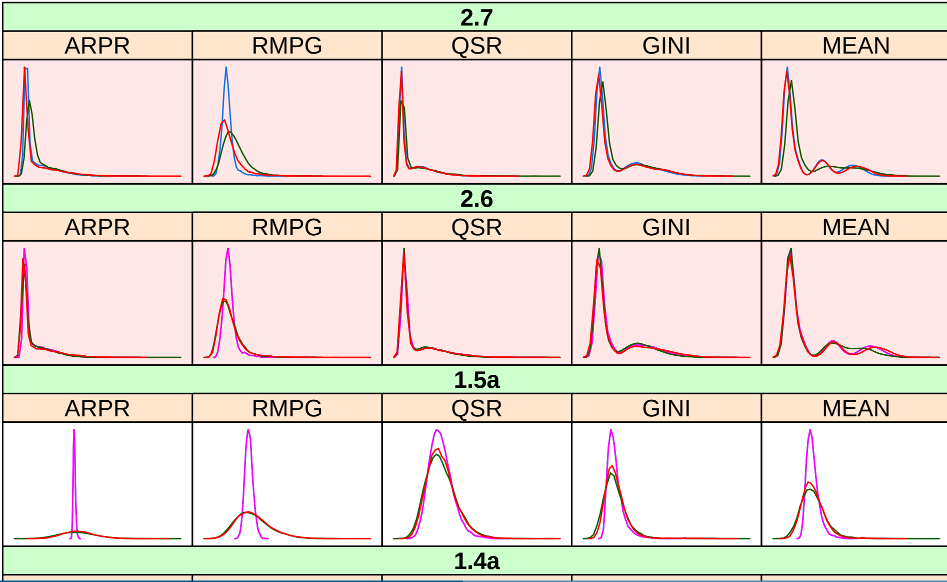

2.7
2.6
1.5a
1.4a
1.2
ARPR
RMPG
QSR
GINI
MEAN
rbrr
boot
approx4
naive
−0.5
0.5


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
1.5a
1.4a


2.7
ARPR
RMPG
QSR
GINI
MEAN
2.6
1.5a
1.4a
1.2
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

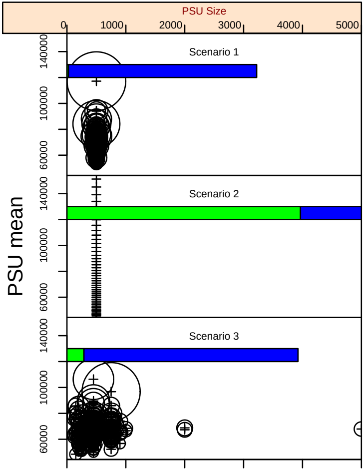
Direct/appr.	1.2	1.4a	1.5a	2.6	2.7
ARPR	95.070	94.700	94.950	89.340	90.640
RMPG	94.640	94.790	94.550	92.930	92.650
QSR	94.620	95.260	94.850	83.880	83.690
GINI	94.440	95.090	95.140	84.230	85.550
MEAN	94.850	95.070	95.320	78.720	79.960
Bootstrap	**1.2**	**1.4a**	**1.5a**	**2.6**	**2.7**
ARPR	95.100	94.910	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

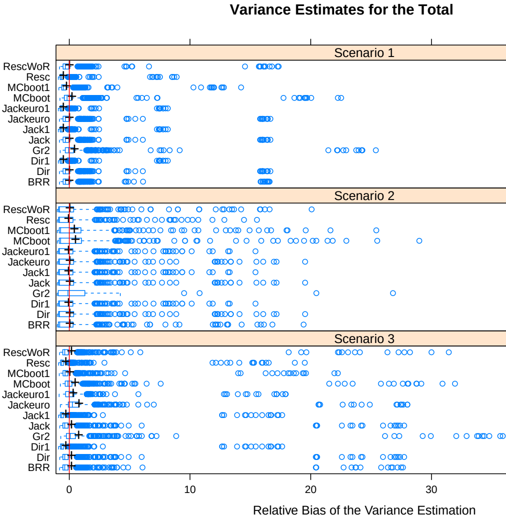
Experimental Study: Sampling Design

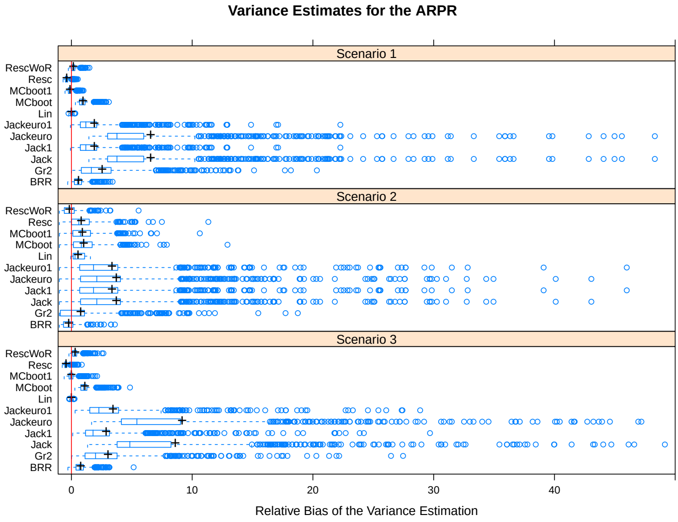
- Two stage sampling with stratification at the first stage, 25 strata
- 1. Stage: Drawing 4 PSU in each stratum (contains 8 PSU in average, altogether 200 PSU)
- 2. Stage: Proportional allocation of the sample size (1,000 USU) to the PSU (contains 500 USU in average, altogether 100,000 USU)

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
140000
100000
60000


Variance Estimates 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

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

  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 $\dfrac{n_h}{n_h - 1} \cdot w_{hi}$
- 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

Missing Data - *Everybody has them, nobody wants them*

Missingness may be either

- **MCAR** (missing completely at random),
- **MAR** (missing at random), or
- **MNAR** (missing not at random)

Rubin and Little (1987, 2002)

Methods to handle missing data

- Procedures based on the **available cases** only, i.e., only those cases that are completely recorded for the variables of interest
- **Weighting procedures** such as Horvitz-Thompson type estimators or raking estimators that adjust for nonresponse
- **Single imputation** and correction of the variance estimates to account for imputation uncertainty
- **Multiple imputation** (MI) according to Rubin (1978, 1987) and standard complete-case analysis
- **Model-based corrections** of parameter estimates such as the expectation-maximization (EM) algorithm

Variance estimation under multiple imputation

- Multiple imputation (Rubin, 1987): $\widehat{\theta}^{(j)}$ and $\widehat{V}(\widehat{\theta}^{(j)})$
- Multiple imputation point estimate $\widehat{\theta}_{MI} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\theta}^{(j)}$
- Multiple imputation variance Estimate

$$T = W + (1 + \frac{1}{m})B$$

with

within imputation variance $W = \frac{1}{m}\sum_{j=1}^{m}\widehat{V}(\widehat{\theta}^{(j)})$ and

between imputation variance $B = \frac{1}{m-1}\sum_{j=1}^{m}(\widehat{\theta}^{(j)} - \widehat{\theta}_{MI})^2$

- Problem: the imputation has to be proper in Rubin's sense.

What else to consider?

- Design Effects:
  - Are defined as the ratio of an estimator under a complex design over the corresponding estimator under a simple random sampling without replacement.
  - The enumerator was developed throughout the lecture.
  - The denominator is more difficult to estimate since the data are from a complex design.
  - The ESS uses design effects to determine effective sample sizes that allow for comparative surveys.
- Variance functionals.
- Variance estimation for change (cf. AMELI reports).