Variance Estimation in Complex Surveys Using Resampling Methods

eFRAME Summer School on measurement of well-being and social progress

Ralf Münnich

University of Trier
Faculty IV, Economics
Economic and Social Statistics Department

11. September 2013
### Unemployment in Germany (Saarland)

<table>
<thead>
<tr>
<th>Unemployed</th>
<th>14 – 24</th>
<th>25 – 44</th>
<th>45 – 64</th>
<th>65 +</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women τ</td>
<td>2.387</td>
<td><strong>7.248</strong></td>
<td>4.686</td>
<td><strong>128</strong></td>
<td>14.449</td>
</tr>
<tr>
<td>Men τ</td>
<td>4.172</td>
<td>9.504</td>
<td>10.588</td>
<td>0</td>
<td>24.264</td>
</tr>
<tr>
<td>E τ</td>
<td>4.172</td>
<td>9.505</td>
<td>10.598</td>
<td>0</td>
<td>24.275</td>
</tr>
<tr>
<td>∑ τ</td>
<td>6.559</td>
<td>16.752</td>
<td>15.274</td>
<td>128</td>
<td>38.713</td>
</tr>
<tr>
<td>E ∑ τ</td>
<td>6.558</td>
<td>16.743</td>
<td>15.282</td>
<td>128</td>
<td>38.711</td>
</tr>
</tbody>
</table>

- **True values** in Saarland
- Estimated values from the German Microzensus (1% sample survey)
- Base-population in Germany (82 Mio. inhabitants)
- Equivalent quality of table estimates?
- Does the sampling design influence modelling?
Surveys and quality measurement
Surveys and Samples

Definition of a *survey*

A survey, in general, is the methodology for gathering information via samples on persons, households, or other units. In a survey

- planning and development,
- pretest,
- survey design,
- implementation,
- data gathering, editing and processing, as well as
- analysis

play a vital role.
Introduction to Survey Statistics

- What would we like to achieve with drawing samples?
  - To avoid a census (complete inventory)
  - To gather information (*rapidly*)
  - To gain *early information* (e.g., polls)

- What do we have to consider while drawing samples?
  - Costs of a survey
  - Response burden (effort)
  - Quality of the *production process*

- ... anything else?
  - Selection of an appropriate sampling design
  - Selection of (a set) of estimator(s)

- How do we measure the outcome?
  - This shall be a difficult task!
How can we measure quality?

- Main emphasis is put on sample surveys
- Statistical items of interest:
  - Sampling distribution
    - Theoretical distribution
    - Approximate distribution
    - Simulated distribution
  - Focus on properties
    - Large sample properties
    - Small sample properties
- View of Official Statistics
Data quality: Eurostat definition

Relevance of the statistical concept:
   End-user, *user needs*, hierarchical structure and contents

Accuracy and reliability:
   ▶ Sampling errors: standard error, CI coverage
   ▶ Non-sampling errors: nonresponse, coverage error, measurement errors

Timeliness and punctuality: Time and duration from data acquisition until publication

Coherence and comparability: Preliminary and final statistics, annual and intermediate statistics (regions, domains, time)

Accessibility and clarity: Publication of data, analysis and method reports

Completeness (not part of the Code of Practice) Source:
   http://www.epp.eurostat.ec.europa.eu/→ ESS/Quality
Evaluation of samples and surveys

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?
Some first examples of surveys
Short overview to the German Microcensus

- 1% household and person survey each year (interviewers)
- Regional stratification by
  - Federal states
  - Regional strata (214 in Germany)
  - House size classes
  - Sampling units (1 of 100 selected in sample)
- Variables of interest (for the simulation study)
  - Unemployed (ILO definition; estimation variable)
  - Jobless (registered in Nuremberg)
  - Nationality (German, non-German), gender, regional stratum
  - Imputation: age class, non-response class (special combination of HH size, nationality, gender, age)
### BAW: unemployed men under 25
(non-German, longtime unemployed)

<table>
<thead>
<tr>
<th>Model</th>
<th>$E(\hat{\tau})$</th>
<th>$V(\hat{\tau})$</th>
<th>$E(\hat{V}(\tau))$</th>
<th>$V(\hat{V}(\tau))$</th>
<th>CV (%)</th>
<th>CI (90%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff</td>
<td>54724</td>
<td>2.13e+006</td>
<td>2.20e+006</td>
<td>4.12e+010</td>
<td>2.67</td>
<td>90.65</td>
</tr>
<tr>
<td>Ratio</td>
<td>54754</td>
<td>2.36e+006</td>
<td>2.44e+006</td>
<td>6.06e+010</td>
<td>2.81</td>
<td>90.52</td>
</tr>
<tr>
<td>RatioGGK</td>
<td>54840</td>
<td>2.44e+006</td>
<td>2.46e+006</td>
<td>6.31e+010</td>
<td>2.85</td>
<td>90.05</td>
</tr>
<tr>
<td>RatioRS</td>
<td>55657</td>
<td>3.21e+006</td>
<td>2.55e+006</td>
<td>8.23e+010</td>
<td>3.62</td>
<td>82.52</td>
</tr>
<tr>
<td>Regr</td>
<td>54720</td>
<td>2.05e+006</td>
<td>2.10e+006</td>
<td>3.41e+010</td>
<td>2.61</td>
<td>90.57</td>
</tr>
<tr>
<td>RegrGGK</td>
<td>54720</td>
<td>2.05e+006</td>
<td>2.08e+006</td>
<td>3.32e+010</td>
<td>2.62</td>
<td>90.37</td>
</tr>
<tr>
<td>RegrRS</td>
<td>54742</td>
<td>2.05e+006</td>
<td>1.98e+006</td>
<td>2.95e+010</td>
<td>2.61</td>
<td>89.46</td>
</tr>
<tr>
<td>Total</td>
<td>54693</td>
<td>6.71e+006</td>
<td>6.66e+006</td>
<td>1.42e+011</td>
<td>4.74</td>
<td>89.76</td>
</tr>
<tr>
<td>Regr</td>
<td>4787</td>
<td>2.52e+005</td>
<td>2.43e+005</td>
<td>3.50e+009</td>
<td>10.49</td>
<td>88.82</td>
</tr>
<tr>
<td>RegrGGK</td>
<td>4767</td>
<td>2.24e+005</td>
<td>2.08e+005</td>
<td>3.15e+009</td>
<td>9.93</td>
<td>87.70</td>
</tr>
<tr>
<td>RegrRS</td>
<td>4012</td>
<td>2.26e+005</td>
<td>1.34e+005</td>
<td>2.43e+009</td>
<td>22.53</td>
<td>36.35</td>
</tr>
<tr>
<td>Total</td>
<td>4783</td>
<td>6.38e+005</td>
<td>6.42e+005</td>
<td>1.31e+010</td>
<td>16.70</td>
<td>89.68</td>
</tr>
</tbody>
</table>

where $\tau_1 = 54734$ and $\tau_2 = 4781$. 

Trier, 11. September 2013 | Ralf Münich | 12 (122) | Variance Estimation in Complex Surveys
Some first examples of surveys

**BAW: unemployed, separate regression estimator**

![Distribution of Estimator](image1)

![Distribution of Variance Estimator](image2)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>54734.10</td>
</tr>
<tr>
<td>$N$</td>
<td>10112200</td>
</tr>
<tr>
<td>$\hat{\tau}$</td>
<td>54742.04</td>
</tr>
<tr>
<td>$\text{var} \hat{\tau}$</td>
<td>2.05e+06</td>
</tr>
<tr>
<td>$E(\text{var}(\tau))$</td>
<td>1.98e+06</td>
</tr>
<tr>
<td>$V(\hat{\tau})$</td>
<td>2.95e+10</td>
</tr>
<tr>
<td>Bias Est.</td>
<td>7.94</td>
</tr>
<tr>
<td>MSE Est.</td>
<td>2.05e+06</td>
</tr>
<tr>
<td>Bias Var.</td>
<td>-6.90e+04</td>
</tr>
<tr>
<td>MSE Var.</td>
<td>3.42e+10</td>
</tr>
<tr>
<td>Skew Est.</td>
<td>0.0597</td>
</tr>
<tr>
<td>Curt Est.</td>
<td>2.9789</td>
</tr>
<tr>
<td>Skew Var.</td>
<td>0.1206</td>
</tr>
<tr>
<td>Curt Var.</td>
<td>2.9455</td>
</tr>
<tr>
<td>CI (90%)</td>
<td>89.46 (5.1;5.4)</td>
</tr>
<tr>
<td>CI (95%)</td>
<td>94.47 (2.5;3.0)</td>
</tr>
</tbody>
</table>

Trier, 11. September 2013 | Ralf M"unnich | 13 (122) | Variance Estimation in Complex Surveys
Some first examples of surveys

BAW: non-German unemployed under 25
separate regression estimator

\( \tau \): 4780.79  \( N \): 10112200
\( \hat{\tau} \): 4011.98  \( V(\hat{\tau}) \): 2.26e+005  \( E(\hat{V}(\tau)) \): 1.34e+005  \( V(\hat{V}(\tau)) \): 2.43e+009
Bias Est: -768.81  MSE Est: 8.17e+005  Bias Var: -9.24e+004  MSE Var: 1.10e+10
Skew Est: -0.1401  Curt Est: 3.0031  Skew Var: 0.5365  Curt Var: 3.5587
CI (90%): 36.35 (0.1; 63.6)  CI (95%): 45.25 (0.0; 54.7)

Trier, 11. September 2013 | Ralf Münich | 14 (122) | Variance Estimation in Complex Surveys
Some first examples of surveys

Example: Men in Hamburg

\[ \tau: 805258.00 \quad N: 1669690 \]
\[ \hat{\tau}: 805339.10 \quad V \hat{\tau}: 1.29e+008 \quad E(\hat{V}(\tau)): 1.29e+008 \quad V(\hat{V}(\tau)): 6.72e+014 \]

Bias Est: 81.10 \quad MSE Est: 1.29e+008 \quad Bias Var: -3.78e+005 \quad MSE Var: 6.72e+014

Skew Est: 0.0747 \quad Curt Est: 3.0209 \quad Skew Var: 1.8046 \quad Curt Var: 6.9973

CI (90\%): 90.16 (4.1;5.7) \quad CI (95\%): 94.79 (2.0;3.2)
Some first examples of surveys

**Total Estimate Separated by House Size Class (GGK)**

<table>
<thead>
<tr>
<th></th>
<th>GGK1</th>
<th>GGK2</th>
<th>GGK3</th>
<th>GGK4</th>
<th>GGK5</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>468293</td>
<td>651740</td>
<td>439745</td>
<td>9940</td>
<td>99970</td>
<td>1669690</td>
</tr>
<tr>
<td>Sampling units</td>
<td>173</td>
<td>446</td>
<td>414</td>
<td>10</td>
<td>75</td>
<td>1118</td>
</tr>
</tbody>
</table>

Trier, 11. September 2013 | Ralf Münich | 16 (122) | Variance Estimation in Complex Surveys
Some first examples of surveys

Distribution of Men in HAM (per SU)

HAM Total

HAM GGK 1

HAM GGK 2

HAM GGK 3

HAM GGK 4

HAM GGK 5

Trier, 11. September 2013 | Ralf Münich | 17 (122) | Variance Estimation in Complex Surveys
Some first examples of surveys

Aim of sample surveys

Survey designer

- Gathering information on unknown population parameters
- Selection of sample design
  - Cost effectiveness of survey (burden etc.)
  - Frame effectiveness and practicability
  - Efficiency of estimates
    (e.g. stratification and optimal allocation)
- Needs of valid auxiliary information

Researcher

- ... is interested in estimating models
- How do sampling designs influence models?
- *Beware large Gelman factors!!!*
Variance estimation methods
Preliminary comments on variance estimation

So far we have seen:

1. Many designs
2. Many estimators of different kinds
3. Some practical issues (incl. non-response)

What exactly has variance estimation to fulfil?

▶ Deliver adequate quality measures
▶ Consider practical issues

But

▶ Can we apply general methods like for point estimation?
▶ Do we encounter contradictions in optimizing strategies between point and variance estimators?
Direct variance estimators
Direct variance estimation methods

In StrRS we take the variance

\[ V(\hat{\mu}_{\text{StrRSWOR}}) = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{\sigma_q^2}{n_q} \cdot \frac{N_q - n_q}{N_q - 1} \]

which can be estimated by

\[ \hat{V}(\hat{\mu}_{\text{StrRSWOR}}) = s_{\hat{\mu}}^2 = \sum_{q=1}^{L} \gamma_q^2 \cdot \frac{s_q^2}{n_q} \cdot \frac{N_q - n_q}{N_q} \] (M.o.Z.)

(unbiased estimate).
Variance estimation for the HT

The variance

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

of the HT estimate $\hat{\tau}$ can be estimated by

$$\hat{V}_{HT}(\hat{\tau}) = \sum_{i \in S} (1 - \pi_i) \cdot \left(\frac{y_i}{\pi_i}\right)^2 + 2 \cdot \sum_{i,j \in S} (1 - \frac{\pi_i \cdot \pi_j}{\pi_{ij}}) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j}.$$
Sen-Yates-Grundy

The Sen-Yates-Grundy variance estimator

\[
V_{SYG}(\hat{\tau}) = \sum_{i,j \in U} \sum_{i < j} (\pi_i \cdot \pi_j - \pi_{ij}) \cdot \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2
\]

has an unbiased variance estimate

\[
\hat{V}_{SYG}(\hat{\tau}) = \sum_{i,j \in S} \sum_{i < j} \frac{\pi_i \cdot \pi_j - \pi_{ij}}{\pi_{ij}} \cdot \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.
\]

Problem: \( \pi_{ij} \)

(calculation and memory usage)
An Experimental Study in Two Stage Sampling
Variance Estimator of a Total for Two Stage Sampling

- **Variance:**

\[
V(\hat{\tau}_{2St}) = L^2 \cdot \left( \frac{L - l}{L} \right) \cdot \frac{\sigma^2_e}{l} + \frac{L}{l} \sum_{q=1}^{L} \left( \frac{N_q - n_q}{N_q} \right) \cdot N_q^2 \cdot \frac{\sigma^2_q}{n_q}
\]

with \( \sigma^2_e = \frac{1}{L - 1} \sum_{q=1}^{L} \left( \tau_q - \frac{\tau}{L} \right)^2 \), \( \sigma^2_q = \frac{1}{N_q - 1} \sum_{i=1}^{N_q} (y_{qi} - \bar{y}_q)^2 \)


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

\[
E \left[ L^2 \cdot \left( \frac{L - l}{L} \right) \cdot \frac{s^2_e}{l} \right] = L^2 \cdot \left( 1 - \frac{l}{L} \right) \cdot \frac{\sigma^2_e}{l} + \frac{L}{l} \left( 1 - \frac{l}{L} \right) \sum_{q=1}^{L} V(\hat{\tau}_q)
\]
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)

Big study: http://ameli.surveystatistics.net
Münnich, Bruch, and Zins (2012): Presentation at NetSILC II
Experimental Study: Scenarios

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

Variance Estimates for the Total

Scenario 1

Scenario 2

Scenario 3

Relative Bias of the Variance Estimation
An Experimental Study in Two Stage Sampling

Variance Estimates for the ARPR

Scenario 1

Scenario 2

Scenario 3

Relative Bias of the Variance Estimation
Linearized variance estimators
Variance approximations

In order to avoid calculating these double sums with second order inclusion probabilities, several approximations are proposed

\[
\hat{V}(\hat{\tau}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{n_h}{n_h - 1} \cdot \sum_{j=1}^{n_h} \left(\frac{y_j}{\pi_j} - \frac{1}{n_h} \sum_{k=1}^{n_h} \frac{y_k}{\pi_k}\right)^2
\]


\[
\hat{V}(\hat{\tau}) = \frac{1}{1 - \sum_{i \in S} a_i^2} \cdot \sum_{i \in S} (1 - \pi_i) \cdot \left(\frac{y_i}{\pi_i} - \sum_{j \in S} a_j \cdot \frac{y_j}{\pi_j}\right)^2
\]

(n non random) with \(a_i = \frac{1 - \pi_i}{\sum_{j \in S} 1 - \pi_j}\).
General idea of linearization methods

Estimating $\hat{V}(\hat{\theta})$ may become cumbersome, in either design or type of estimator. Hence, it seems convenient to look for alternative values which use simple total estimation (eg via HT) where

$$\hat{V}(\hat{\theta}) \approx \hat{V}(\hat{\tau}_z^*) .$$

One approach may be Taylor’s method with

$$f(y) = f(y_0) + f'(y_0) \cdot (y - y_0) + R(x)$$

such that only simple terms result which vanishing remainder.
Variance estimation for calibration estimators

Variance estimation using a *residual* technique

\[ \hat{V}(\hat{\tau}) = \sum_{k \in S} \sum_{l \in S} (w_k \cdot w_l - w_{kl}) \cdot (g_k \cdot e_k) \cdot (g_l \cdot e_l) \]

where

\[ e_k = y_k - x'_k \cdot \hat{B} \]

(Asymptotics via superpopulation models).

Remarks:

- Some variants, e.g. using the adjusted weights.
- Variable selection may become important.
Basis for linearization methods

\( \tau \) is the parameter of interest:

\[
\tau = N \cdot \int y \, dF(y)
\]

Next, we get the estimator

\[
\hat{\tau} = \sum_{i \in S} w_i(S) \cdot y_i = \sum_{i=1}^{N} w_i(S) \cdot Y_i
\]

Estimating equations and influence functions

$\theta_0$ is defined implicitly via

$$\int u(y, \theta_0) \, dF(y) = 0.$$ 

The estimating equation for $\theta_0$ estimate $\hat{\theta}$ is the defined as

$$\int \hat{u}(y, \hat{\theta}_0) \, d\hat{F}(y) = 0,$$

where $\hat{u}(y, \hat{\theta})$ is the estimate of $u(y, \theta)$. Finally we get

$$\hat{\theta} - \theta \approx u^*(y)$$

approximately as:

$$u^*(y) = - \left( \frac{\partial \mathbb{E}(u(y, \theta))}{\partial \theta} \bigg|_{\theta=\theta_0} \right)^{-1} u(y, \theta_0).$$

The influence values $u^*(y)$ can directly used for variance estimation.
Quantiles from estimating equations

For estimating quantiles of a population we get

\[ u(y) = \mathbb{1}(y \leq \theta) - p \]

with

\[ u^*(y) = -\frac{1}{f(\theta)} \cdot (\mathbb{1}(y \leq \theta) - p) \]

Binder / Kovačević (1995), see also Deville (1999)
At-risk-of-poverty rate (1)

Starting with the poverty line $\alpha \cdot y_p$ we get as at-risk-of-poverty rate

$$\text{ARPR}_{\alpha;p} = F(\alpha \cdot y_p)$$

Next we estimate $\text{ARPR}_{\alpha;p}$ with

$$\text{\hat{ARPR}}_{\alpha,p} = \hat{F}(\alpha \cdot \hat{y}_p)$$

where

$$\hat{y}_p = \inf_{y \in \mathbb{R}} \left( \hat{F}(y) > p \right)$$

At-risk-of-poverty rate (2)

Using the values $\hat{\tau}_{u^*} = \sum_{i \in S} w_i \cdot u_i^*$ we get

$$\hat{V}(\text{ARPR}_\alpha, p) \approx \hat{V}(\hat{\tau}_{u^*})$$.

Applying this to the at-risk-of-poverty rate we obtain

$$u_i^* = \frac{1}{N} \cdot (1(y_i \leq \alpha y_p) - \text{ARPR}_\alpha, p)$$

and in the case of StrRS the variance estimator

$$\hat{V}(\text{ARPR}_\alpha, p) \approx \sum_{h=1}^{H} N_h^2 \cdot \frac{s_{u_i^*(h)}^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right).$$

Remark: ARPR$_{\alpha, p}$ is in general estimated from the sample (see Deville, 1999).
Revisiting the GINI estimate

\[ \hat{\text{GINI}}_y = \frac{1}{\hat{\tau}_Y} \cdot \sum_{i \in S} \frac{1}{\pi_i} \cdot \left( 2 \cdot \frac{1}{\hat{N}} \sum_{j \in S} \frac{1}{\pi_j} \cdot \delta(y_j \leq y_i) - 1 \right) \cdot y_i \]

we can derive the influence values

\[ u_i^* = \frac{1}{\hat{\tau}_Y} \cdot \left( 2 \cdot y_i \cdot \hat{F}(y_i) - (\hat{\text{GINI}}_y + 1) \cdot y_i \right) \]
Woodruff linearization

In analogy to the above mentioned methods we try to linearize the estimate \( \hat{\theta} = \sum_{i \in S} w_i \cdot z_i \) where

\[
    z_i = \sum_{j=1}^{k} \frac{\partial}{\partial \theta_i} h(\theta_1, \ldots, \theta_k) \cdot y_{j,i}.
\]

\( \theta = h(\theta_1, \ldots, \theta_k) \) is meant to be a function of \( k \) totals.

Assuming \( \hat{\theta} = \hat{\tau}_1/\hat{\tau}_2 \) we obtain

\[
    \hat{z}_i = \frac{\hat{\tau}_1}{\hat{\tau}_2} \cdot \left( \frac{y_{1,i}}{\hat{\tau}_1} - \frac{y_{2,i}}{\hat{\tau}_2} \right).
\]

The variance estimation is analogously to the estimating equation approach (see Andersson / Nordberg, 1994).
Quintile share ratio

Starting the quintiles of the groups R and P we get

\[
\hat{\mu}_R = \sum_i w_i \cdot (y_i - y_i \cdot 1(y_i \leq \hat{y}_{0.8})) / \sum_i w_i \cdot (1 - 0.8)
\]

\[
\hat{\mu}_P = \sum_i w_i \cdot y_i \cdot 1(y_i \leq \hat{y}_{0.2}) / \sum_i w_i \cdot 0.2
\]

Next, we define the QSR as a function of four totals:

\[
\widehat{QSR} = \frac{\hat{\mu}_R}{\hat{\mu}_P} = \frac{\hat{\tau}_1}{\hat{\tau}_2} / \frac{\hat{\tau}_3}{\hat{\tau}_4}
\]

In order to apply linearized variance estimation, we have to derive the linearized variables.
QSR: linearized variables

\[ u_{1i} = y_i - ((y_i - y_{0,8}) \cdot 1(y_i \leq \hat{y}_{0,8}) + 0,8 \cdot y_{0,8}) \]
\[ u_{2i} = 0,2 \]
\[ u_{3i} = (y_i - y_{0,2}) \cdot 1(y_i \leq \hat{y}_{0,2}) + 0,2 \cdot y_{0,2} \]
\[ u_{4i} = 0,2 \]

\[ u_{5i} = (u_{1i} - \frac{\hat{\tau}_1}{\hat{\tau}_2} \cdot u_{2i}) \cdot \frac{1}{\hat{N} \cdot 0,2} = \hat{\mu}_R \]
\[ u_{6i} = (u_{3i} - \frac{\hat{\tau}_3}{\hat{\tau}_4} \cdot u_{4i}) \cdot \frac{1}{\hat{N} \cdot 0,2} = \hat{\mu}_P \]

where \( \hat{\tau}_2 = \hat{\tau}_4 = \hat{N} \cdot 0,2 \). Finally, we get

\[ z_i = (u_{5i} - \frac{QSR}{\hat{\tau}_3} \cdot u_{6i}) \cdot \frac{\hat{\tau}_4}{\hat{\tau}_3} \]
Study on Poverty Measurement
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
Recent Developments (European view)

- Start: Nygård and Sandström (1985) in JOS
- 90ies: different approaches applied to non-linear statistics
  Deville (1999)
- FP6: (composite) indicators with imputation
- FP7: measures of poverty and social exclusion (Laeken)
  - Lisbon goals
  - Year of combating poverty
  - AMELI and SAMPLE
  - ESSnet on SILC
- In parallel: World bank method, Molina/Rao, Tillé
- Beyond GDP (FP8)
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)
Application to Poverty and Inequality Indicators

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

$$V(\hat{\theta}) \approx V\left(\sum_{i \in s} w_i \cdot z_i\right)$$

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).

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARPR:</td>
<td>Deville (1999)</td>
</tr>
<tr>
<td>GINI:</td>
<td>Kovačević and Binder (1997)</td>
</tr>
<tr>
<td>QSR:</td>
<td>Hulliger and Münnich (2007)</td>
</tr>
<tr>
<td>RMPG:</td>
<td>Osier (2009)</td>
</tr>
</tbody>
</table>
## The AMELI Sample Designs

<table>
<thead>
<tr>
<th>ID</th>
<th>PSU</th>
<th>Strata</th>
<th>$\pi_{il}$</th>
<th>SM</th>
<th>Alloc.</th>
<th>SSU</th>
<th>St.</th>
<th>$\pi_{ill}$</th>
<th>SM</th>
<th>Alloc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>HID</td>
<td>–</td>
<td>srs</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1.4a</td>
<td>HID</td>
<td>NUTS2</td>
<td>srs</td>
<td>1</td>
<td>prop</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1.5a</td>
<td>HID</td>
<td>NUTS2*DOU</td>
<td>pps</td>
<td>HHG</td>
<td>prop</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2.6</td>
<td>CIT</td>
<td>NUTS2*DOU</td>
<td>srs</td>
<td>1</td>
<td>prop</td>
<td>HID</td>
<td>–</td>
<td>pps</td>
<td>HHG</td>
<td>–</td>
</tr>
<tr>
<td>2.7</td>
<td>CIT</td>
<td>NUTS2*DOU</td>
<td>srs</td>
<td>1</td>
<td>prop</td>
<td>HID</td>
<td>–</td>
<td>srs</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**Notes:** SM: measure of size; Alloc: allocation; prop.: proportional; srs: simple random sampling without replacement; pps: sampling proportional to size (Midzuno); $p_k(\cdot)$: sampling design at the $k$th stage; $\pi_{il}$ and $\pi_{ill}$: sample inclusion probability at the first and second stage; **Variables:** HID: household identifier; HHG: household size; CIT: municipality identifier (LAU1); DOU: degree of urbanization.
### Sampling Fractions in the Study

<table>
<thead>
<tr>
<th>ID(s)</th>
<th>variable sample size</th>
<th>fixed $n = 6000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$f_1$</td>
<td>$f_{II}$</td>
</tr>
<tr>
<td>1.2 / 1.4a</td>
<td>1%; 5%</td>
<td>–</td>
</tr>
<tr>
<td>1.5a</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>2.7 / 2.6</td>
<td>5%; 1.25%</td>
<td>20%; 80%</td>
</tr>
<tr>
<td>2.7 / 2.6</td>
<td>25%; 6.25%</td>
<td>20%; 80%</td>
</tr>
</tbody>
</table>

Here: we use a fixed sample size of $n = 6,000$ households.
Characteristics of the AMELI universe
### Study on Poverty Measurement

<table>
<thead>
<tr>
<th></th>
<th>ARPR</th>
<th>RMPG</th>
<th>QSR</th>
<th>GINI</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>rbr boot approx naive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rbr boot approx naive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rbr boot approx naive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Variance Estimation in Complex Surveys

<table>
<thead>
<tr>
<th></th>
<th>ARPR</th>
<th>RMPG</th>
<th>QSR</th>
<th>GINI</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Trier, 11. September 2013 | Ralf Münich | 52 (122) | Variance Estimation in Complex Surveys
<table>
<thead>
<tr>
<th>2.7</th>
<th>ARPR</th>
<th>RMPG</th>
<th>QSR</th>
<th>GINI</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
</tr>
<tr>
<td>2.6</td>
<td>ARPR</td>
<td>RMPG</td>
<td>QSR</td>
<td>GINI</td>
<td>MEAN</td>
</tr>
<tr>
<td></td>
<td><img src="image6.png" alt="Graph" /></td>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
<td><img src="image9.png" alt="Graph" /></td>
<td><img src="image10.png" alt="Graph" /></td>
</tr>
<tr>
<td>1.5a</td>
<td>ARPR</td>
<td>RMPG</td>
<td>QSR</td>
<td>GINI</td>
<td>MEAN</td>
</tr>
<tr>
<td></td>
<td><img src="image11.png" alt="Graph" /></td>
<td><img src="image12.png" alt="Graph" /></td>
<td><img src="image13.png" alt="Graph" /></td>
<td><img src="image14.png" alt="Graph" /></td>
<td><img src="image15.png" alt="Graph" /></td>
</tr>
<tr>
<td>1.4a</td>
<td>ARPR</td>
<td>RMPG</td>
<td>QSR</td>
<td>GINI</td>
<td>MEAN</td>
</tr>
<tr>
<td></td>
<td><img src="image16.png" alt="Graph" /></td>
<td><img src="image17.png" alt="Graph" /></td>
<td><img src="image18.png" alt="Graph" /></td>
<td><img src="image19.png" alt="Graph" /></td>
<td><img src="image20.png" alt="Graph" /></td>
</tr>
</tbody>
</table>
## Coverage Rates (in %) of Indicator Estimates

<table>
<thead>
<tr>
<th>Direct/appr.</th>
<th>1.2</th>
<th>1.4a</th>
<th>1.5a</th>
<th>2.6</th>
<th>2.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARPR</td>
<td>95.070</td>
<td>94.700</td>
<td>94.950</td>
<td>89.340</td>
<td>90.640</td>
</tr>
<tr>
<td>RMPG</td>
<td>94.640</td>
<td>94.790</td>
<td>94.550</td>
<td>92.930</td>
<td>92.650</td>
</tr>
<tr>
<td>QSR</td>
<td>94.620</td>
<td>95.260</td>
<td>94.850</td>
<td>83.880</td>
<td>83.690</td>
</tr>
<tr>
<td>GINI</td>
<td>94.440</td>
<td>95.090</td>
<td>95.140</td>
<td>84.230</td>
<td>85.550</td>
</tr>
<tr>
<td>MEAN</td>
<td>94.850</td>
<td>95.070</td>
<td>95.320</td>
<td>78.720</td>
<td>79.960</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bootstrap</th>
<th>1.2</th>
<th>1.4a</th>
<th>1.5a</th>
<th>2.6</th>
<th>2.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARPR</td>
<td>95.100</td>
<td>94.910</td>
<td>94.810</td>
<td>87.850</td>
<td>93.070</td>
</tr>
<tr>
<td>RMPG</td>
<td>94.410</td>
<td>94.750</td>
<td>94.600</td>
<td>92.390</td>
<td>94.940</td>
</tr>
<tr>
<td>QSR</td>
<td>94.280</td>
<td>95.180</td>
<td>94.220</td>
<td>82.210</td>
<td>88.260</td>
</tr>
<tr>
<td>GINI</td>
<td>94.240</td>
<td>94.770</td>
<td>94.660</td>
<td>81.890</td>
<td>90.070</td>
</tr>
<tr>
<td>MEAN</td>
<td>94.620</td>
<td>95.260</td>
<td>95.090</td>
<td>77.630</td>
<td>90.340</td>
</tr>
</tbody>
</table>
Other Variance Estimation Work in AMELI

We expect the allowance from the Commission for publication very soon:

**WP3 / D3.1** Variance Estimation for Complex Surveys
**WP3 / D3.2** Variance Estimation for Indicators of Poverty and Social Exclusion
Variance estimation for change (development)

**WP7 / D7.1** Report on the Simulation Results (with long appendix)

http://ameli.surveystatistics.net
Resampling methods
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 computational exceptional
  - What does influence the quality of these estimates
Random groups

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

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

- Random group variance estimate:

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

- Random selection versus random partition!
Resampling methods
Balanced repeated replication

Balanced repeated replication

- Originally we have two observations per stratum
- Random partitioning of observations into two groups
- $\hat{\theta}_r$ 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:

$$\hat{\theta}_{\text{BRR}} = \frac{1}{R} \cdot \sum_{r=1}^{R} \hat{\tau}_r$$

$$\hat{V}_{\text{BRR}}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}_r - \hat{\theta})^2$$

- Use special weighting techniques for improvements
Delete-1-Jackknife
- Resampling by omitting (deleting) one element in each resample
- \( \hat{\theta}_i \) 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 $\hat{\theta}(Y_1, \ldots, Y_n)$ be the statistic of interest for estimating the parameter $\theta$. Then,

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

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

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

(cf. Shao und Tu, 1995).
The jackknife (continued)

From the bias follows immediately Hieraus Jackknife point estimate

\[
\hat{\theta}_{d1JK} = \hat{\theta} - \hat{B}_{d1JK}(\hat{\theta}) \\
= n \cdot \hat{\theta} - \frac{n - 1}{n} \sum_{i \in S} \hat{\theta}_i
\]

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
\[ \hat{\theta}^*_i := n \cdot \hat{\theta} - (n - 1) \cdot \hat{\theta}_-; \]
which yield under the assumption of stochastic independency and approximately equal variance of the \( \hat{\theta}^*_i \) finally

\[
\hat{V}_{d1JK}(\hat{\theta}) = \frac{1}{n(n - 1)} \cdot \sum_{i \in S} (\hat{\theta}^*_i - \bar{\theta}^*)^2
\]

\[
= \frac{n - 1}{n} \sum_{i \in S} \left( \hat{\theta}_- - \frac{1}{n} \sum_{i \in S} \hat{\theta}_- \right)^2.
\]

Problem: what is \( \hat{\theta}^*_i \) and \( \hat{V}_{d1JK}(\hat{\theta}) \) for \( \hat{\theta} = \bar{Y} \)?
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)

\[
\hat{V}_{d1JK, strat}(\hat{\theta}) = \sum_{h=1}^{h} \frac{(1 - f_h) \cdot (n_h - 1)}{n_h} \cdot \sum_{i=1}^{n_h} (\hat{\theta}_{h,-i} - \bar{\hat{\theta}}_h)^2
\]

where \(-i\) indicates the unit \(i\) 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; \(m \ll \binom{n}{d}\))
Estimation of the median \((n \text{ pair})\)

\[ \hat{V}_{d1JK}(\hat{\theta}) = \frac{n-1}{n} \cdot \sum_{i \in S} (\hat{\theta}_{-i} - \hat{\theta})^2 \]

\[ = \frac{n-1}{n} \cdot \left( \sum_{i=1}^{m} (Y_{[m+1]} - \frac{1}{2}(Y_{[m]} + Y_{[m+1]})^2) \right) \]

\[ + \sum_{i=m+1}^{n} (Y_{[m]} - \frac{1}{2}(Y_{[m]} + Y_{[m+1]})^2) \]

\[ = \frac{n-1}{4} \cdot (Y_{[m]} - Y_{[m+1]})^2. \]

Finally,

\[ n \cdot \hat{V}_{d1JK}(\hat{\theta}) \xrightarrow{\text{d}} \sigma^2 \cdot \left( \chi^2_2 / 2 \right)^2, \]

where \(\sigma^2\) is the asymptotic variance of \(\sqrt{n} \cdot (\hat{\theta} - \theta)\). The inconsistency follows from the missing smoothness of quantile estimation and the use of the median.
Bootstrap resampling

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

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

- Special adaptations are needed in complex surveys
- Insufficient estimates in WOR sampling and higher sample fractions
Bootstrap resampling (continued)

\[
\hat{V}_{\text{Boot}}(\hat{\theta}) = \int \cdots \int (\hat{\tau}_n(x_1, \ldots, x_n) - \int \cdots \int \hat{\tau}_n(y_1, \ldots, y_n) \\
\hat{f}(y_1) \cdot \ldots \cdot \hat{f}(y_n) \ dy_1 \ldots dy_n) \cdot \hat{f}(x_1) \cdot \ldots \cdot \hat{f}(x_n) \ dx_1 \ldots dx_n
\]

\[
= V_\ast(\hat{\tau}_n(X_1^\ast, \ldots, X_n^\ast) | x_1, \ldots, x_n)
\]

is the classical variance of the random vector conditional to the sample with distribution \( \hat{F} \) (cf. Shao and Tu, 1995). \( \{X_1^\ast, \ldots, X_n^\ast\} \) is an iid sample from \( \hat{F} \) which is called bootstrap sample.

The variance \( V_\ast(\cdot | X_1, \ldots, X_n) \) is the conditional bootstrap variance given the data \( X_1, \ldots, X_n \). For \( \hat{F} = F \) one yields

\[
V(\hat{\tau}) = V_\ast(\hat{\tau}_n(X_1^\ast, \ldots, X_n^\ast) | X_1, \ldots, X_n)
\]
Bootstrap variance estimation

In case of $\hat{F} \neq F$ the equivalence of the variances do not hold anymore. However, the bootstrap variance $\hat{V}_{\text{Boot}}(\hat{\theta})$ can be used as variance estimator of the variance for the statistic $\hat{\tau}$ (see Shao, 2003, S. 381).

As an example for a theoretic bootstrap variance estimation one can apply $\hat{\tau}^* = \overline{X}^*$ for the estimator $\hat{\tau} = \overline{X}$. Finally, one gets the bootstrap variance estimator

$$V(\overline{X}^*) = \frac{n-1}{n^2} \cdot s^2_{\overline{X}} = \frac{n-1}{n} V(\overline{X}) .$$

(siehe Rao und Wu, 184, S. 106 f.). This equals the classical variance of the estimator $\overline{X}$ except the finite population correction.
Monte-Carlo-Bootstrap

Efron (1982):

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

$$X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}$$

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

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

\[ \hat{V}_{\text{Boot,MC}} \xrightarrow{\text{a.s.}} V_{\text{Boot}}. \]

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

\[ \hat{B}_{\text{Boot,MC}} = \frac{1}{B} \sum_{i=1}^{B} \hat{\tau}^*_n,i - \hat{\tau}. \]
Bootstrap confidence intervals

- **Via variance estimation**

\[
\left[ \hat{\tau} - \sqrt{\hat{V}_{\text{Boot,MC}}(\hat{\tau})} \cdot z_{1-\alpha/2}, \hat{\tau} - \sqrt{\hat{V}_{\text{Boot,MC}}(\hat{\tau})} \cdot z_{\alpha/2} \right]
\]

- **via bootstrap resamples:**

\[
z_1^* = \frac{\hat{\tau}_1^* - \hat{\tau}}{\sqrt{\hat{V}_{\text{Boot,MC}}(\hat{\tau}_1^*)}}, \ldots, \quad z_B^* = \frac{\hat{\tau}_B^* - \hat{\tau}}{\sqrt{\hat{V}_{\text{Boot,MC}}(\hat{\tau}_B^*)}}
\]

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

\[
\left[ \hat{\tau} - \sqrt{\hat{V}_{\text{Boot,MC}}(\hat{\tau})} \cdot z_{B(1-\alpha/2)}^*, \hat{\tau} - \sqrt{\hat{V}_{\text{Boot,MC}}(\hat{\tau})} \cdot z_{B\alpha/2}^* \right]
\]

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

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:

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

where $\delta_i$ 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 $\lambda_G = \sqrt{n_G \cdot (\prod_{g=1}^{G-1} f_g) \cdot \frac{(1 - f_G)}{(n_G - n_G^*)}}$. 
Resampling methods
The bootstrap

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

and \( \delta_{hq} \) is 1 when PSU \( q \) in stratum \( h \) is drawn and 0 else

Second stage:

\[ w_{hqk}^* = \left( 1 - \lambda_h + \lambda_h \cdot \frac{l_h}{l_h^*} \cdot \delta_{hq} - \lambda_{hq} \cdot \sqrt{\frac{l_h}{l_h^*}} \cdot \delta_{hq} + \lambda_{hq} \cdot \sqrt{\frac{l_h}{l_h^*}} \cdot \delta_{hq} \cdot \frac{m_{hq}}{m_{hq}^*} \cdot \delta_{hqk} \right) \cdot w_{hqk} \cdot \frac{w_{hq}}{w} \]

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

and \( \delta_{hqk} \) is 1 when SSU \( k \) in PSU \( q \) in stratum \( h \) is drawn and 0 else

cf. Preston (2009)
Example: Rescaling bootstrap WoR for a three stage design

Third stage:

\[ w_{hqki}^* = (1 - \lambda_h + \lambda_h \cdot \frac{l_h}{l_h^*} \cdot \delta_{hq} - \lambda_h \cdot \sqrt{\frac{l_h}{l_h^*} \cdot \delta_{hq}} + \lambda_h \cdot \sqrt{\frac{l_h}{l_h^*} \cdot \delta_{hq}} \cdot \frac{m_{hq}}{m_{hq}^*} \cdot \delta_{hqk} \\
- \lambda_{hqk} \cdot \sqrt{\frac{l_h}{l_h^*} \cdot \delta_{hq}} \cdot \sqrt{\frac{m_{hq}}{m_{hq}^*} \cdot \delta_{hqk}} \\
+ \lambda_{hqk} \cdot \sqrt{\frac{l_h}{l_h^*} \cdot \delta_{hq}} \cdot \sqrt{\frac{m_{hq}}{m_{hq}^*} \cdot \delta_{hqk} \cdot \frac{n_{hqk}}{n_{hqk}^*} \cdot \delta_{hqki}} \cdot w_{hqki} \cdot \frac{w_{hq}}{w_{hq}^*} \cdot \frac{w_{hqk}}{w_{hqk}^*}, \]

where \( \lambda_{hqk} = \sqrt{n_{hqk}^* \cdot f_h \cdot f_{hq} \cdot \frac{(1 - f_{hqk})}{(n_{hqk} - n_{hqk}^*)}} \)

and \( \delta_{hqk} \) is 1 when USU \( i \) in SSU \( k \) in PSU \( q \) in stratum \( h \) is drawn and 0 else.

cf. Preston (2009)
Bootstrap without replacement

- WR plus specialized correction terms (difficult to compute)
- Mirror-match bootstrap
- Bernoulli Bootstrap
- Repeated half-sample bootstrap

cf. especially for multistage sampling Saigo (2010)

πPS bootstrap: Perez-Duarte (2012) ... under development
Resampling methods

The bootstrap

Replication weights

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

\[
\begin{align*}
 w_{hi} \cdot \left[ 1 + \left\{ \frac{(n_h - m_h) \cdot (1 - f_h)}{m_h} \right\}^{1/2} \right], & \quad \delta_{rh} = 1, \\
 w_{hi} \cdot \left[ 1 - \left\{ \frac{m_h \cdot (1 - f_h)}{n_h - m_h} \right\}^{1/2} \right], & \quad \delta_{rh} = -1,
\end{align*}
\]

where \( \delta_{rh} \) 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
Applications to poverty measurement (II)
EU-SILC

We use the EU-SILC with two different continuous income distributions

- Mikrocensus 2001
- Sub-population Hessen:

\[ n_h = (6, 48, 110, 14, 8, 44, 107, 3, 4, 51, 23, 21, 149, 14, 10, 82, 31, 5) \]

Sample fractions: 0.76% – 10.5%
Results: GINI coefficient

<table>
<thead>
<tr>
<th>Estimator</th>
<th>mean P.Est</th>
<th>var P.Est</th>
<th>mean V.Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value:</td>
<td>0.4378971</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GINI.linearized</td>
<td>0.43569128</td>
<td>0.00102439</td>
<td>0.00120201</td>
</tr>
<tr>
<td>GINI.brr</td>
<td>0.43500500</td>
<td>0.00095699</td>
<td>0.00092486</td>
</tr>
<tr>
<td>GINI.d1JK</td>
<td>0.43568935</td>
<td>0.00102426</td>
<td>0.00073022</td>
</tr>
<tr>
<td>GINI.boot99</td>
<td>0.43377631</td>
<td>0.00096958</td>
<td>0.00076771</td>
</tr>
<tr>
<td>GINI.boot499</td>
<td>0.43371329</td>
<td>0.00095573</td>
<td>0.00076769</td>
</tr>
<tr>
<td>true value:</td>
<td>0.4181957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GINI.linearized</td>
<td>0.41643888</td>
<td>0.00092309</td>
<td>0.00119264</td>
</tr>
<tr>
<td>GINI.brr</td>
<td>0.41621606</td>
<td>0.00083668</td>
<td>0.00084811</td>
</tr>
<tr>
<td>GINI.d1JK</td>
<td>0.41643780</td>
<td>0.00092296</td>
<td>0.00105660</td>
</tr>
<tr>
<td>GINI.boot99</td>
<td>0.41449881</td>
<td>0.00085095</td>
<td>0.00072890</td>
</tr>
<tr>
<td>GINI.boot499</td>
<td>0.41459596</td>
<td>0.00085365</td>
<td>0.00072457</td>
</tr>
</tbody>
</table>
Varianzschätzung: GINI
## Results: ARPR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>mean P.Est</th>
<th>var P.Est</th>
<th>mean V.Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value</td>
<td>0.1965685</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARPR.linearized</td>
<td>0.19048904</td>
<td>0.00023127</td>
<td>0.00021083</td>
</tr>
<tr>
<td>ARPR.brr</td>
<td>0.18999639</td>
<td>0.00021324</td>
<td>0.00023983</td>
</tr>
<tr>
<td>ARPR.d1JK</td>
<td>0.19050662</td>
<td>0.00022993</td>
<td>0.00166968</td>
</tr>
<tr>
<td>ARPR.boot99</td>
<td>0.19009913</td>
<td>0.00020982</td>
<td>0.00022526</td>
</tr>
<tr>
<td>ARPR.boot499</td>
<td>0.19005892</td>
<td>0.00020835</td>
<td>0.00022639</td>
</tr>
<tr>
<td>true value</td>
<td>0.1962637</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARPR.linearized</td>
<td>0.19068219</td>
<td>0.00026417</td>
<td>0.00021226</td>
</tr>
<tr>
<td>ARPR.brr</td>
<td>0.19055935</td>
<td>0.00023797</td>
<td>0.00027438</td>
</tr>
<tr>
<td>ARPR.d1JK</td>
<td>0.19069923</td>
<td>0.00026351</td>
<td>0.00160690</td>
</tr>
<tr>
<td>ARPR.boot99</td>
<td>0.19040288</td>
<td>0.00023023</td>
<td>0.00025073</td>
</tr>
<tr>
<td>ARPR.boot499</td>
<td>0.19047729</td>
<td>0.00022906</td>
<td>0.00025310</td>
</tr>
</tbody>
</table>
Varianzschätzung: ARPR
Results: QSR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>mean P.Est</th>
<th>var P.Est</th>
<th>mean V.Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value:</td>
<td>8.62098</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QSR.linearized</td>
<td>8.6308557</td>
<td>1.20583745</td>
<td>1.67286982</td>
</tr>
<tr>
<td>QSR.brr</td>
<td>8.71025303</td>
<td>1.24331111</td>
<td>1.19411971</td>
</tr>
<tr>
<td>QSR.d1JK</td>
<td>8.63537274</td>
<td>1.20642163</td>
<td>2.09821625</td>
</tr>
<tr>
<td>QSR.boot99</td>
<td>8.62079757</td>
<td>1.21137806</td>
<td>1.08477999</td>
</tr>
<tr>
<td>QSR.boot499</td>
<td>8.62777705</td>
<td>1.20474372</td>
<td>1.07938536</td>
</tr>
</tbody>
</table>

| true value:     | 8.041066   |           |            |
| QSR.linearized  | 8.07585630 | 0.99426342| 1.10653990 |
| QSR.brr         | 8.14088144 | 1.01576321| 1.02049213 |
| QSR.d1JK        | 8.07488981 | 0.99280725| 2.01360904 |
| QSR.boot99      | 8.06864356 | 1.00089130| 0.93755550 |
| QSR.boot499     | 8.06901036 | 0.99308870| 0.92878758 |
Applications to poverty measurement (II)

Varianzschätzung: QSR
Variance estimation in the presence of nonresponse
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)

→ In multivariate analysis often 30% to 40% of the data are lost with case deletion assuming MCAR!
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

→ We regard multiple imputation as most flexible for multipurpose complex surveys
Calibration for non-response

- Using the calibration approach directly yields biased estimates
- Update of calibration weights
  - Correction of response probability
  - Response homogeneity classes

before calibrating the first order inclusion probabilities

\[ \pi_k \rightarrow \pi_k;NR \]
Example: Unit non-response in volunteer panels

Microcensus (MC)
- 1% sample of German population
- 4 rotational quarters

Access Panel (DSP)
- Recruitment out of the latest MC quarter
- Unit non-response due to self-selection (recruitment rates vary enormously)

D-SILC
- Stratified sample from access panel
- Population estimates using weights
Variance estimation in the presence of nonresponse
Weighting methods

Two possibilities to compensate for unit non-response

Grossing up using response propensities

Grossing up directly

\[ \hat{w}_i, \text{D-SILC} = \hat{\pi}_{i,MC} \times g_{i,Pop} \times g_{i,MC} \times \hat{\pi}_{i,AP} \times \hat{\pi}_{i,RS} \]

\[ \hat{w}_i, \text{D-SILC} = \hat{\pi}_{i,MC} \times g_{i,Pop} \times g_{i,MC} \times \hat{\pi}_{i,RS} \]
Point estimates

Grossing up directly
Grossing up using response propensities
Särndal’s two-phase approach

\[ \hat{\nu}(\hat{\tau}) = \nu_{\text{SAM}}(\hat{\tau}) + \nu_{\text{NR}}(\hat{\tau}) \]

where

\[ \nu_{\text{SAM}}(\hat{\tau}) = \sum_{k=1}^{r} \sum_{l=1}^{r} (d_k d_l - d_{kl}) (g_k v_{sk} e_k) (g_l v_{sl} e_l) \]

\[ - \sum_{k=1}^{r} d_k (d_k - 1) v_{sk} (v_{sk} - 1) (g_k e_k)^2 \]

\[ \nu_{\text{NR}}(\hat{\tau}) = \sum_{k=1}^{r} d_k^2 v_{sk} (v_{sk} - 1) e_k^2 \]

with \( v_{sk}, g_k, \) and \( e_k \) according to Lundstrøm and Särndal (2002), sections 6.3 and 6.4.
Variance estimation in the presence of nonresponse
Weighting methods

Unemployed women, 25 – 44

Raking estimator

variance estimator

NR rates: 1: 5%, 2: 10%, 3: 25%, 4: 40%
Variance estimation in the presence of nonresponse
Weighting methods

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%

$\mathcal{E}_1 \quad \mathcal{E}_2 \quad \mathcal{E}_3$

95% 90%
Unemployed women, 65 +, distribution of point and variance estimator (25% NR)
Resampling under non-response

- General idea:
  Imputation in each resample

- Balanced Repeated Replication
  Shao, Chen, und Chen (1998)

- Jackknife
  Rao and Shao (1992); Shao and Tu (1995)
  Newer developments with specialised routines by Berger and Rao

- Bootstrap
  Shao und Sitter (1996)

Computational burden may be very high!
Cf. http://rpm.dacseis.de
Variance estimation in the presence of nonresponse
Resampling methods

Unemployed women, aged 14 - 24

Imputation, direct variance estimation
without jobless

Imputation, direct variance estimation
with jobless

1: 5%, 2: 10%, 3: 25%, 4: 40% non-response
Unemployed women, aged 14 - 24

Imputation, bootstrap variance estimation (Shao-Sitter)
without jobless

Imputation, bootstrap variance estimation (Shao-Sitter)
with jobless

1: 5%, 2: 10%, 3: 25%, 4: 40% non-response
Unemployed women, aged 14 - 24

Raking, residual variance estimation without jobless

Raking, residual variance estimation with jobless

1: 5%, 2: 10%, 3: 25%, 4: 40% non-response
Unemployed women, aged 14 - 24

Calibration, LS residual variance estimation without jobless

Calibration, LS residual variance estimation with jobless

1: 5%, 2: 10%, 3: 25%, 4: 40% non-response
The Multiple Imputation Principle (1)

Estimate 1

Estimate 2

Estimate 3

MI estimate

MI inference

Variance estimation in the presence of nonresponse
Multiple imputation

Trier, 11. September 2013 | Ralf Männich | 104 (122) | Variance Estimation in Complex Surveys
The Multiple Imputation Principle (2)

\[ \hat{\theta}, \hat{\text{var}}(\hat{\theta}) \]

Complete data

Missing values

Incomplete data

\[ \hat{\theta}_1, \hat{\text{var}}(\hat{\theta}_1) \]

Imputed data set 1

\[ \hat{\theta}_2, \hat{\text{var}}(\hat{\theta}_2) \]

Imputed data set 2

\[ \vdots \]

\[ \vdots \]

\[ \vdots \]

\[ \hat{\theta}_m, \hat{\text{var}}(\hat{\theta}_m) \]

Imputed data set \( m \)

\[ \hat{\theta}_{\text{MI}}, \hat{\text{var}}_{\text{MI}}(\hat{\theta}) \]
Variance estimation under multiple imputation

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

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

with

- within imputation variance $W = \frac{1}{m} \sum_{j=1}^{m} \hat{\text{var}}(\hat{\theta}^{(j)})$
- between imputation variance $B = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\theta}^{(j)} - \hat{\theta}_{MI})^2$

- Problem: the imputation has to be proper in Rubin’s sense.
Univariate Multiple Imputation Models for Complex Data

Simple case with 3 variables $A$, $B$ and $C$ each with missing data (Rubin 2003, applied in the NMES):

- “Begin by arbitrarily filling in all missing $B$ and $C$.
- Fit a model of $A|B, C$ using those units where $A$ is observed and impute the missing $A$ values.
- Toss the imputed $B$ values and fit a model of $B|A, C$ using those units where $B$ is observed and impute the missing $B$ values.
- Toss the imputed $C$ values and fit a model of $C|A, B$ using those units where $C$ is observed and impute the missing $C$ values.
- Iterate...”
MI Algorithm for Missing Continuous Data

- Assume the underlying data model of a linear regression

\[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon = X \beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \]

- Assume that \( y \) has \( n_{mis} \) missing data, variables \( X \) are fully observed or already imputed. \( y_{obs} \) and \( X_{obs} \) refer to the jointly observed part, \( X_{mis} \) to the missing part \( y_{mis} \).

- Let \( \hat{\beta} \) and \( \hat{\sigma}^2 = (y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})/(n_{obs} - p) \) be the least squared estimates from the observed data.

- Multiple imputation procedure for \( j = 1, 2, \ldots m \):
  1. Draw \( \sigma^2|X \sim (y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})\chi_{n_{obs} - p}^{-2} \)
  2. Draw a vector of \( p \) variables from
     \[ \beta|\sigma^2, X \sim N(\hat{\beta}, \sigma^2(X_{obs}'X_{obs})^{-1}) \]
  3. Draw \( Y_{mis}|\beta, \sigma^2, X \sim N(X_{mis}\beta, \sigma^2) \) independently for every missing value \( i = 1, 2, \ldots, n_{mis} \).
MI Algorithm for Missing Binary Data

- Assume the underlying data model of a logistic regression

\[ \ln \left( \frac{p}{1 - p} \right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = X \beta, \quad p = P(Y = 1 | X). \]

- Assume that \( y \) has \( n_{mis} \) missing data, variables \( X \) are fully observed or already imputed.

- Let \( \hat{\beta} \) be the iterative least squares estimates from the observed data (or any approximate ML estimate) and \( \hat{V}(\hat{\beta}) \) its estimated covariance matrix.

- Apply the large sample normal approximation for \( j = 1, 2, \ldots, m \):
  1. Draw a vector of \( p \) variables from \( \beta | X \sim N(\hat{\beta}, \hat{V}(\hat{\beta})) \)
  2. For every \( i \in mis \) calculate \( p_i = 1/(1 + \exp(-X_i' \beta)) \)
  3. Draw \( n_{mis} \) independent uniform \((0,1)\) random numbers \( u_i \) for \( i = 1, 2, \ldots, n_{mis} \) and if \( u_i > p_i \) impute \( Y_i = 0 \), otherwise impute \( Y_i = 1 \).
Predictive Mean Matching

- Large sample approximation of the Logit model sometimes does not work well
- Either use stabilizing prior
- or predictive mean matching (Rubin 1986, Little 1988)
- Use the simple linear regression multiple imputation routine
- Create $Y_{imp}$ for all units $i = 1, 2, \ldots, n$
- For every unit $i$ with missing values, impute the observed value $Y_{obs,j}$ from a nearest neighbor unit $j$ in $Y_{imp,j}$ such that

$$|Y_{imp,i} - Y_{imp,j}| \rightarrow \text{min}$$

- Quite robust!
Unemployed in Saarland

Point Estimator

Variance Estimator

Raking, HT (MI logit, PMM), gGREG (MI logit, PMM), $m = 30, 5, 15$
Unemployed Women in Saarland (Age Group II)

Point Estimator

Variance Estimator

Raking, HT (MI logit, PMM), gGREG (MI logit, PMM), $m = 30, 5, 15$
Unemployed in Saarland (Subgroup)

Point Estimator

Variance Estimator

Raking, HT (MI logit, PMM), gGREG (MI logit, PMM), $m = 30, 5, 15$
Non-Bayesian multiple imputation

Bjørnstad (2004) has introduced non-Bayesian multiple imputation using classical hotdeck imputation methods in Rubin’s framework. An extension with discussion will be published in JOS soon (Bjørnstad, 2008):

\[
V(\hat{\theta}_{MI}) = \frac{1}{m} \sum_{j=1}^{m} \hat{V}(\hat{\theta}^*_j) + \left( k + \frac{1}{m} \right) \cdot \frac{\sum_{j=1}^{m} (\hat{\theta}^*_j - \hat{\theta}^*_{MI})^2}{m - 1}.
\]

Problem: variance inflation constant \( k \). Preliminary choice:

\[
k = \frac{1}{n - n_r} \cdot \frac{1}{1 - \frac{n_r}{n}}.
\]
Optimal constant $k$

- sub0: Jobless in the federal state Saarland
- sub1: Sub-population from regional stratum
- sub5: Sub population of larger buildings

Table: Optimal constants $k_{\text{opt}}$ for the HT (left) and GREG (right) estimator under multiple imputation with SI LFS2 and SI LFS 3 as well as under MI PRIMA (preliminary choice: 4/3)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>HT</th>
<th>GREG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sub0</td>
<td>sub1</td>
</tr>
<tr>
<td>Population</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imputation method</td>
<td>$k_{\text{opt},0}$</td>
<td>$k_{\text{opt},1}$</td>
</tr>
<tr>
<td>SI LFS2</td>
<td>1.6683</td>
<td>2.6946</td>
</tr>
<tr>
<td>SI LFS3</td>
<td>1.6908</td>
<td>2.6950</td>
</tr>
<tr>
<td>MI PRIMA</td>
<td>1.4426</td>
<td>2.2696</td>
</tr>
</tbody>
</table>
Summary and outlook
Variable selection methods

In almost all methods that use auxiliary information the problem of adequate variable selection arises.

- Do classical variable selection methods help survey statistics?
- Is the rule *the more the better* correct here?

- It seems to be the case that answers are different for point and variance estimation.

Work ongoing:
Knobelspies and Münnich (2008)
Nowadays: model-based estimation (small area)
Summary (1)

- Law of large numbers yields perfect results
- Central limit theorem sometimes hardly applicable
  - Outliers
  - Rare observations
  - Survey design
- Well posed problems:
  → many good solutions
- Problems with models in some situations
  - Logit models for imputation
  - Many variables for multiple imputation
  - Categorical variables
- Other quality aspects
  - Coherence
  - Disparity of regions
- http://www.dacseis.de
Summary (2)

- Variance estimation for nonlinear estimates: Linearization and BRR
  Improvements for jackknife and bootstrap
- Linearization
  - Good for GINI
  - QSR problematic
- Designs
  - Easy adaptations for linearization methods
  - $\pi_{ij}$: approximations from Deville and others?
  - Very complicated for resampling methods
- Consideration of item non-response
Outlook

- Improvements of resampling methods with regards to complex surveys
- Deeper investigation of the accuracy of linearization methods
- Simultaneous estimation of groups of estimates of variance estimates
  - Presentation of Partha Lahiri at ISI 2007
    - Variance stabilization as small-area problem
  - Census 2010/11 round on NUTS 5 level estimates
Addon for longitudinal estimates

- Linearization methods were best in many studies.
- ARPR is less sensitive towards skewed distributions but more tends to be biased.
- GINI and R8020 are relatively non-robust against very skewed distributions.

Alternative approach:
- Parametric estimation of income distributions (Graf, 2009)
- Model-based non-parametric estimation

Next steps:
- Calibrated version of the estimators (Demnati and Rao, 2004)
- Panel attrition (models from Nordberg, 2000) vs. MI
- Inclusion of longer periods (2 to 4 years) including at-persistence-of poverty rate

http://ameli.surveystatistics.net


