Flexible variance estimation in complex sample surveys: rescaled bootstrap in multistage, pps surveys

NTTS 2013
Malgorzata Osiewicz & Sébastien Pérez-Duarte

Usual disclaimer applies
Outline of talk

1. **Why bootstrap?**

2. **Application of rescaled bootstrap to**
   - a. Stratified simple random sampling
   - b. Multistage stratified SRS
   - c. PPS with replacement
   - d. Rao-Hartley-Cochran PPS
   - e. Multistage mix of RHC PPS and stratified SRS

3. **Simulation exercise**
User requirements on variance estimation

• **Variance estimation is a clear **requisite** **from users**
  – need to reveal uncertainty in estimates
  – obtain valid inferences, measure quality
  – emphasis on solutions applicable to multinational level, e.g. a composite “euro area” or “European union”

• **Variance estimation should be as straightforward as possible, and suitable for means, medians, regressions**
  – err on pragmatism on the user side
Sample designs of the HFCS

- HFCS = Household Finance and Consumption Survey

- Variety of sample designs
  - one stage – stratified SRS (BE, CY, FI, LU, MT)
  - two stages – e.g. selection of municipalities (AT, ES, FR, IT, GR, PT, SI)
  - three stages (DE)

- PSU selection with or without PPS
- Sampling unit = addresses, (fiscal) households, or persons
The rescaled bootstrap

- Within the bootstrap family, the **rescaled bootstrap** is a good contender [Rao & Wu (1988), Rao, Wu & Yue (1992)]

- **Advantages:**
  - suitable for arbitrary (i.e. **harmonisable**) number of replicate weights
  - asymptotically sound
  - appropriate for large-fraction sampling
  - straightforward to implement and use
  - can be combined with calibration
The rescaled bootstrap (2)

- Covers both smooth and non-smooth statistics
- Applicable to stratified designs, easily combinable across countries
- Method:
  - sample with replacement $m_h$ units from each stratum $(n_h)$

$$w^*_hi = \left(1 - \lambda_h + \lambda_h \frac{n_h}{m_h} r^*_hi\right) w_{hi} \text{ with } \lambda_h = \left(\frac{m_h(1-f_h)}{n_h-1}\right)^{1/2}$$

- in HFCS, $m_h = n_h - 1$, at least 1000 replicate weights, calibrated
Extension to multistage samples

- Preston (2009) in the case of **without** replacement bootstrap
- Osiewicz & Perez-Duarte (2012) convert to **with** replacement bootstrap:

\[
W_{hijk}^* = \left(1 + \Delta_{hi} + \Delta_{hij} + \Delta_{hik}\right)W_{hijk}
\]

First stage

\[
\Delta_{hi} = -\lambda_h + \lambda_h \frac{n_h}{m_h} r_{hi}^*
\]

Second stage

\[
\Delta_{hij} = -\lambda_{hi} \sqrt{\frac{n_h}{m_h}} r_{hi}^* + \lambda_{hi} \sqrt{\frac{n_h}{m_h}} \frac{n_{hi}}{m_{hi}} r_{hij}^*
\]

Third stage

\[
\Delta_{hijk} = -\lambda_{hij} \sqrt{\frac{n_h}{m_h}} \sqrt{\frac{n_{hi}}{m_{hi}}} r_{hij}^* + \lambda_{hij} \sqrt{\frac{n_h}{m_h}} \sqrt{\frac{n_{hi}}{m_{hi}}} \frac{n_{hij}}{m_{hij}} r_{hijk}^*
\]

\[
\lambda_h = \sqrt{\frac{m_h (1 - f_h)}{n_h - 1}}
\]

\[
\lambda_{hi} = \sqrt{\frac{m_{hi} f_h (1 - f_{hi})}{n_{hi} - 1}}
\]

\[
\lambda_{hij} = \sqrt{\frac{m_{hij} f_h f_{hi} (1 - f_{hij})}{n_{hij} - 1}}
\]
PPS sampling – wish list

• No obvious PPS procedure that
  1. Is simple to apply
  2. Has exact first-order inclusion probabilities
  3. Has positive second-order inclusion probabilities
  4. Second-order inclusion probabilities are easy to compute
  5. Sen-Yates-Grundy variance estimator is positive
Generalized bootstrap

- Beaumont and Patak (2012) application to Poisson sampling

- Applicable to unequal probability sample, generic Horvitz Thompson estimator

- Calculation of joint inclusion probabilities still an issue
With replacement PPS sampling

- Variance has a simple expression
  
  \[
  V(\hat{Y}) = \frac{1}{n} \sum_{k=1}^{N} p_k \left( \frac{y_k}{p_k} - \bar{Y} \right)^2 = \frac{1}{n} \sum_{k=1}^{N} p_k \left( \frac{y_k^2}{p_k} - \bar{Y}^2 \right)
  \]

- Variance estimator has simple expression as well
  
  \[
  \hat{V}(\hat{Y}) = \frac{1}{n} \frac{1}{n - 1} \sum_{k=1}^{N} \left( \frac{y_k}{p_k} - \bar{Y} \right)^2
  \]

- Replication: draw with replacement, equal probabilities, size \( m \)

- Replicate weights: \( w_k^* = \left( 1 - \lambda + \lambda \frac{n}{m} r_k^* \right) w_k \), with

  \[
  \lambda = \left( \frac{m}{n-1} \right)^{1/2}
  \]. If \( m = n - 1 \), \( w_k^* = \frac{n}{n-1} r_k^* w_k \)
In this case the variance estimator has a very simple expression:

\[ \hat{V}(\hat{t}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{\hat{t}_i}{p_i} - \hat{t} \right)^2 \]

Only first stage enters the calculation

Bootstrap: resample with-replacement as before in the first stage, but do not resample in the second stage
Rao-Hartley-Cochran PPS

- $p_i$ “size” probability of unit $i$
- **RHC sample:**
  1. Randomly group the $H$ units in $h$ groups of $H_k$
  2. Sample one unit from each group, with pps
- The estimator is $\hat{Y} = \sum_{k=1}^{h} \frac{y_k}{p_k/P_k} = \sum_{k=1}^{h} P_k \frac{y_k}{p_k}$ with $P_k = \sum_{i \in \text{group } k} p_i$
- Variance is $V(\hat{Y}) = \frac{h(\sum_{k=1}^{h} H_k^2 - H)}{H(H-1)} \left( \sum_{i=1}^{H} \frac{y_i^2}{hp_i} - \frac{Y^2}{h} \right)$
- Variance estimator:
  $$\hat{V}(\hat{Y}) = \frac{\sum_{k=1}^{h} H_k^2 - H}{H^2 - \sum_{k=1}^{h} H_k^2} \left( \sum_{k=1}^{h} P_k \left( \frac{y_k}{p_k} - \hat{Y} \right)^2 \right)$$
Resampling the Rao-Hartley-Cochran PPS

- Rao-Wu (1988) – Resample \( m \) units with replacement with probabilities proportional to \( P_k \)

- Replicate weights: \( w_k^* = \left( 1 - \mu m^{1/2} + \mu \frac{m^{-1/2}}{P_k} r_k^* \right) w_k \)

  with \( \mu^2 = (\sum_{k=1}^h H_k^2 - H)/(H^2 - \sum_{k=1}^h H_k^2) \)

- Compare with standard RSB: \( w_k^* = \left( 1 - \lambda + \lambda \frac{h}{m} r_k^* \right) w_k \)

  with \( \lambda = \left( \frac{m(1-f)}{h-1} \right)^{1/2} \)
Two stage sample with RHC on 1\textsuperscript{st} stage

- General \textbf{applicability}, e.g. pps selection of municipalities, SRS of households
- Variance estimator (Rao-Hartley-Cochran):

\[ \hat{V}(\hat{Y}) = \mu^2 \sum_{k=1}^{h} P_k \left( \frac{\hat{Y}_k}{p_k} - \hat{Y} \right)^2 + \sum_{k=1}^{h} P_k \frac{N_k^2}{p_k} \left( \frac{1}{n_k} - \frac{1}{N_k} \right) s_k^2 \]
Two stage RHC bootstrap

- **RSB resample:**
  - $m$ PSUs according to with-replacement pps, probability $P_k$ (first stage)
  - In each PSU resampled (and independently), with-replacement SRS of SSUs (second stage)

- Then replicate weights are:

  $$w_{ki}^* = \left( 1 - \lambda + \frac{\lambda}{mP_k} r_{ki}^* - \frac{\lambda_k}{\sqrt{mP_k}} r_k^* + \frac{\lambda_k}{\sqrt{mP_k} m_k} n_k r_{ki}^* \right) w_{ki}$$

  with $\lambda^2 = m \mu^2$ and $\lambda_k = \sqrt{m_k(1 - f_k) p_k / (n_k - 1)}/P_k$
Simulation – one stage PPS

- NUTS3 statistics on demography and area
- 1462 regions
- Size variable = total population
  - Test with-replacement pps vs. RHC
    - Sample sizes 50, 100, 200
      - 500 replicate weights
    - 5000 draws for bootstrap variance
      - 500,000 draws for empirical variance
RHC sampling compared to WIR pps

Decrease in the variance due to RHC

<table>
<thead>
<tr>
<th>Area</th>
<th>Birth rate</th>
<th>Death rate</th>
<th>Female population</th>
<th>Births</th>
<th>Total population</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>200</td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>50</td>
</tr>
</tbody>
</table>

- Decrease in variance due to RHC is shown across various measures, such as birth rate, death rate, female population, births, total population, and deaths. The graph illustrates the percentage decrease in variance from 50 to 200, indicating the effectiveness of RHC sampling compared to WIR pps.
Simulation results

Variance of the total

![Graph showing variance of total for different categories such as Area, Birth rate, Death rate, Female population, Births, and Deaths, with the relative bootstrap variance on the x-axis and the categories on the y-axis. The graph compares RHC and With replacement PPS.]
Simulation results

Variance of the median

Area
Birth rate
Death rate
Female population
Births
Deaths

Relative bootstrap variance

Deaths
Births
Female population
Death rate
Birth rate
Area
200
100
50
200
100
50
200
100
50
200
100
50
200
100
50
200
100
50

RHC With replacement PPS
Simulation results

Variance of the 10\textsuperscript{th} percentile

- Area
- Birth rate
- Death rate
- Female population
- Births
- Deaths

Relative bootstrap variance

RHC With replacement PPS
Simulation results – RMSE of the total

With replacement pps  RHC

Area | Birth rate | Death rate | Female population | Male population | Births | Deaths | Area | Birth rate | Death rate | Female population | Male population | Births | Deaths
---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
50 | 50% | 0% | 200% | 100% | 150% | 200% | 250% | 300% | 350% | 0% | 50% | 100% | 150% | 200% | 250% | 300% | 350%
200 | 200% | 150% | 100% | 50% | 0% | 50% | 100% | 150% | 200% | 250% | 300% | 350% | 0% | 50% | 100% | 150% | 200% | 250% | 300% | 350%
Conclusions

• **General framework for replicate weights in multistage surveys combining PPS (Rao-Hartley-Cochran type) and stratified simple random sampling**

• **Extensions:**
  - Multistage simulation (work in progress)
  - Rao & Wu (1988) resampling from (i, i’) for general PPS sampling
Thank you for your time!
Background slides
What is the HFCS?

• **Endeavour of the** Eurosystem of Central Banks + some NSIs
• **Collects** micro-level data on household finances and consumption
  – Covering a broad range of issues including real and financial assets, liabilities, consumption and saving, as well as income, employment, pension entitlements, etc.
• **Conducted in all 17 euro area countries**
  – Data will be representative at the country level, and comparable across countries
• **First wave currently close to finished** (around April 2013)
• **Anonymized micro data will be made available to the research community, pre-registration possible**
General characteristics

- **Decentralized effort.** Each NCB/NSI finances and conducts its own wealth survey
- **Survey frequency:** every 2 / 3 years
- **Sample size:** ca. 62,000 completed interviews
- **Panel component:** BE, DE, ES, IT, NL
- **Wealthy HHs oversampled in several countries using different methods**
General characteristics (cont’d)

• **ECB coordinates / ensures common methodology:**
  – Probability sampling
  – Multiple imputation
  – Re-weighting / calibration to external sources
  – Variance estimation/replicate weights

• **Ex-ante commitment on output harmonisation**
  (common definitions / output variables / blueprint questionnaire) ➔ no precedent for wealth surveys!

**HFCS unique case:** joint data collection on euro area household income, consumption and wealth via a single survey tool