Bootstrap Methods for Complex Sampling Designs in Finite Population

Erika Antal and Yves Tillé
University of Neuchâtel

Colloque Deville
Neuchâtel 2009
Objectives and Interest of the Method

- Method for bootstrapping from sample selected with complex design.
- The bootstrap sample has the same size (in expectation) as the original sample.
- No need of a reconstruction of pseudo-population (Gross, 1980; Chao and Lo, 1985).
- No need of rescaling the bootstrapped units (Mac Carthy and Snowden, 1985; Rao and Wu, 1988; Kuk, 1989; Rao et al., 1992; Sitter, 1992a,b; Booth et al., 1994; Holmberg, 1998; Preston and Henderson, 2007).
- In the bootstrap sample: use of the same inclusion probabilities.
- No need of correction of the variance.
- The variance of the bootstrapped estimators is directly an estimator of variance.
Multinomial Sampling or Simple Sampling with Replacement

- Well-known result (Bol’shev, 1965; Johnson et al., 1997)
  Let \((X_1, \ldots, X_N)\) a vector of independent random variables with Poisson distribution, \(S_k \sim \text{Pois} \left( \lambda \right)\) then

\[
S = (S_1, \ldots, S_N) = \left[ (X_1, \ldots, X_N) \left| \sum_{k=1}^{N} S_k = n \right. \right]
\]

has a multinomial distribution. Thus

\[
\Pr(S = s) = \binom{n}{s_1 \ldots s_k \ldots s_N}^{-1} \prod_{k \in U} \left( \frac{\pi_k}{n} \right)^{s_k}, \text{ for all } s \in \mathcal{R}_n,
\]

where

\[
\mathcal{R}_n = \left\{ s \in \mathbb{N}^N \left| \sum_{k=1}^{N} s_k = n \right. \right\}.
\]
In a multinomial sampling

- $S_k \sim Bin \left( n \frac{1}{N} \right)$
- $\mathbb{E}(S_k) = \frac{n}{N}$
- $\text{var}(S_k) = \frac{n}{N} \left( 1 - \frac{1}{N} \right)$
Consider now a vector of $N$ independent geometric random variables $X_k$:
\[
\Pr(X_k = x_k) = (1 - p)p^{x_k}, \ x_k = 0, 1, 2, 3, \ldots
\]
Then we say that
\[
S = (S_1, \ldots, S_N) = \begin{bmatrix}
(X_1, \ldots, X_N) \\
\sum_{k=1}^{N} S_k = n
\end{bmatrix}
\]
is a sampling with over-replacement.
We get
\[
\Pr(S_1 = x_1, \ldots S_N = x_N) = \frac{1}{\binom{N + n - 1}{n}}.
\]
Sampling with over-replacement

- $S_k$ has a negative hypergeometric distribution
  \[ \Pr(S_k = j) = \frac{\binom{N - 1 + n - j - 1}{n - j}}{\binom{N + n - 1}{n}}, j = 0, \ldots, n, \]
- $E(S_k) = \frac{n}{N}$
- $\text{var}(S_k) = \frac{n(N - 1)(N + n)}{N^2(N + 1)}$. 
Comparison of simple sampling designs

- Under simple random sampling without replacement

\[
\text{var}(\hat{Y}) = N^2 \frac{N-n}{nN} \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2,
\]

- Under simple random sampling with replacement (multinomial)

\[
\text{var}(\hat{Y}) = \frac{N(N-1)}{n} \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2,
\]

- Under simple random sampling with over-replacement

\[
\text{var}(\hat{Y}) = N(N-1) \frac{N+n}{n(N+1)} \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2,
\]
A one-one sampling design is only a tool for bootstrapping.

- Selection with replacement of $n$ units in a sample of size $n$.
- $E(S_k) = 1$
- $\text{var}(S_k) = 1$
- Can be implemented by selecting a part of the observation by multinomial sampling and another part by sampling with over-replacement.
Fundamental problem of the bootstrap in survey sampling

- $\hat{Y}$: an estimator of the total $Y$.
- $\widehat{\text{var}}(\hat{Y})$: an estimator of the variance of $\hat{Y}$.
- $\hat{Y}^*$: an estimator of the total $Y$ obtained by resampling.
- $\text{var}(\hat{Y}^*|S)$: conditional variance of $\hat{Y}^*$ given the sample $S$.
- A ‘good’ bootstrap method must be such that
  \[ \mathbb{E}(\hat{Y}^*|S) = \hat{Y} \quad \text{and} \quad \text{var}(\hat{Y}^*|S) = \widehat{\text{var}}(\hat{Y}) \]
- PROBLEM: in a complex design an estimator of the variance $\widehat{\text{var}}(\hat{Y})$ can be very different from the variance $\text{var}(\hat{Y}^*|S)$.
- Thus the resampling design MUST be different from the sampling design, except in very special cases.
Fundamental problem of the bootstrap in survey sampling

- $S$ the sample $S_k$ number of times the unit is selected in the sample
- $\mathbb{E}(S_k) = \pi_k$ and $\text{cov}(S_k, S_\ell) = \Delta_{k\ell}$
- $S^*$ bootstrap sample $S^*_k$ number of times the unit is resampled
- $\text{var}(\hat{Y}^* | S)$ conditional variance of $\hat{Y}^*$ given the sample $S$.
- $\mathbb{E}(S^*_k | S) = 1$ (to have $\mathbb{E}(\hat{Y}^* | S) = \hat{Y}$)
- $\text{cov}(S^*_k, S^*_\ell | S) = \Sigma_{k\ell}$.
- In order to have $\text{var}(\hat{Y}^* | S) = \text{var}(\hat{Y})$, we must have

\[
\Sigma_{k\ell} = \frac{\Delta_{k\ell}}{\pi_{k\ell}}.
\]
Simple Example: Poisson Sampling

- $S_k$ are independent Bernoulli random variables
- $E(S_k) = \pi_k$, $\text{var}(S_k) = \pi_k(1 - \pi_k)$, and $\text{cov}(S_k, S_\ell) = 0$.
- $S^*$ must be such that

$$E(S^*_k|S) = 1, \quad \text{cov}(S^*_k, S^*_\ell|S) = 0,$$

$$\text{var}(S^*_k|S) = \frac{\text{var}(S_k)}{\pi_k} = 1 - \pi_k.$$
Select a sample $S_{kA}^*$ with a Poisson sampling (without replacement) from $S$ with inclusion probabilities $\pi_k$.

From the set of units of $S$ such that $S_{kA}^* = 0$, select a sample $S_{kB}^*$ according to a Poisson random variable (with replacement) with a parameter 1 i.e.

$$\Pr(S_{kB}^* = x|S) = \begin{cases} \frac{e^x}{x!}, x = 0, 1, 2, 3, \ldots & \text{if } S_{kA}^* = 0 \\ 0 & \text{if } S_{kA}^* = 1. \end{cases}$$

The resampling design is $S_k^* = S_{kA}^* + S_{kB}^*$. 
Sampling design: sampling with replacement.

Resampling with a one-one design.

In this case,

\[
\text{var}(\hat{Y}^*|S) = \frac{N^2}{n(n-1)} \sum_{k \in S} (y_k - \bar{Y})^2.
\]

Note that, if the resampling design is a simple random sampling with replacement, then the variance is slightly underestimated:

\[
\text{var}(\hat{Y}^*|S) = \frac{N^2}{n^2} \sum_{k \in S} (y_k - \bar{Y})^2.
\]
Unequal Probability Sampling With Replacement

- Sampling design: unequal probability sampling with replacement.
- Resampling with a one-one design.

\[
\text{var}(\hat{Y}^*|S) = \frac{n}{n-1} \sum_{k \in S} \left( \frac{y_k}{\pi_k} - \frac{\hat{Y}}{n} \right)^2 ,
\]
Sampling design: sampling without replacement.

Resampling with a composite design:

- A sample $S_A$ of $m = n^2/N$ units is selected in $S$ with simple random sampling without replacement.
- In the non-selected units, select a sample $S_B$ with a one-one design.
- The bootstrap sample is $S^* = S_A + S_B$.

\[
\text{var}(\hat{Y}^*|S) = N^2 \frac{N - n}{n(N - 1)} \frac{1}{n - 1} \sum_{k \in S} (y_k - \hat{Y})^2.
\]
 Sampling design: sampling without replacement.

Resampling with a composite design:
- A sample $S_A$ of units is selected in $S$ with unequal probability sampling and inclusion probabilities $\phi_k$ without replacement.
- In the non-selected units, select a sample $S_B$ with a one-one design.
- The bootstrap sample is $S^* = S_A + S_B$. 
The choice

\[
1 - \phi_k = \min \left[ \frac{n}{n-1} (1 - \pi_k) \left( 1 - \frac{1 - \pi_k}{\sum_{j \in U} S_j (1 - \pi_j)} \right), 1 \right].
\]

will be very close to the estimator of variance proposed by Hájek (1981):

\[
\text{var}(\hat{Y}^* | S) \approx \hat{\text{var}}(\hat{Y}) = \sum_{k \in S} c_k \left( \frac{y_k}{\pi_k} - \frac{\sum_{k \in S} c_k y_k / \pi_k}{\sum_{k \in S} c_k} \right)^2,
\]

\[
c_k = \frac{n}{n-1} (1 - \pi_k).
\]
Other choices for $c_k$:

- The more simple choice is $\phi_k = \pi_k$. In this case, $\text{var}(S_k^*|S) = 1 - \pi_k$. This choice will reconstruct a matrix that is very close to an estimator of variance proposed by Deville and Tillé (2005).

- A third choice can be $\phi_k$ such that

$$1 - \phi_k = \min \left[ -\sum_{j \in U \setminus k} S_j \frac{\Delta_{kj}}{\pi_{kj}}, 1 \right],$$

which has the interest of reconstructing exactly the diagonal of $\Delta_{k\ell}/\pi_{k\ell}$. 


Two-stage Designs

- Possibility to generalize to two-stage and two-phases designs.
- All the primary units must not be selected.
Conclusion

- In order to reconstruct the ‘good’ estimator of variance, the resampling design must be an *ad hoc* procedure that reproduces the first and the second moment.
- The interest is that the rescaling is not necessary.
- The bootstrap sample can be treated as the original sample (calibration, etc.)
Population, functions and sampling designs

- Generated population of 284 units.
- Variance of the estimator of four functions:
  - Total
  - Ratio of two totals Y and X
  - Median
  - Index of Gini
- Three different sampling designs:
  - Poisson sampling
  - Simple Random Sampling Without Replacement (srswor)
  - Unequal Probability Sampling Without Replacement (UPwor)
Goal: \( \text{var}(\hat{Y}^* | S) = \text{var}(\hat{Y}) \)

Comparison of the results of:
- the bootstrap with replacement (generally used bootstrap)
- the bootstrap without replacement (e.g. creating artificial population)
- the new method proposed here.
Other function

Other statistics:

- \( \text{var}(\hat{\theta}) \approx \text{var}_{\text{sim}}(\hat{\theta}) \)
- **Goal:** \( \text{var}(\hat{\theta}^*|S) = \text{var}_{\text{sim}}(\hat{\theta}) \)
- Comparison of the results of
  - the bootstrap with replacement (generally used bootstrap)
  - the new method proposed here.
Measures of comparison

Measures used for the comparison:

- **Ratio (R)** = \[ \frac{\text{var}(\theta^*|S)}{\text{var}_{\text{sim}}(\theta)} \]

- **Relative Bias (RB)** = \[ \frac{\text{var}(\theta^*|S) - \text{var}_{\text{sim}}(\theta)}{\text{var}_{\text{sim}}(\theta)} = \frac{B}{\text{var}_{\text{sim}}(\theta)} \]

- **Relative Root Mean Squared Error (RRMSE)** = \[ \sqrt{\left[ B^2 + \text{var}[\text{var}(\theta^*|S)] \right]} \]

- **Bias Ratio (BR)** = \[ \frac{B}{\sqrt{\text{var}[\text{var}(\theta^*|S)]}} \]
### Simulation results: Poisson Sampling

<table>
<thead>
<tr>
<th></th>
<th>Ratio</th>
<th>Relative Bias</th>
<th>RRMSE</th>
<th>Bias Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>0.9844</td>
<td>−0.0156</td>
<td>0.1643</td>
<td>−0.0892</td>
</tr>
<tr>
<td>Bootstrap WOR</td>
<td>1.0246</td>
<td>0.0246</td>
<td>0.0813</td>
<td>0.3174</td>
</tr>
<tr>
<td>New Method</td>
<td>0.9997</td>
<td>−0.0002</td>
<td>0.1447</td>
<td>−0.0014</td>
</tr>
<tr>
<td><strong>RATIO</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>0.9955</td>
<td>−0.0045</td>
<td>0.2370</td>
<td>−0.0188</td>
</tr>
<tr>
<td>New Method</td>
<td>0.9875</td>
<td>−0.0125</td>
<td>0.2376</td>
<td>−0.0529</td>
</tr>
<tr>
<td><strong>MEDIAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>1.0490</td>
<td>0.0490</td>
<td>0.5355</td>
<td>0.0918</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0502</td>
<td>0.0502</td>
<td>0.5619</td>
<td>0.0897</td>
</tr>
<tr>
<td><strong>GINI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>0.9625</td>
<td>−0.0375</td>
<td>0.6116</td>
<td>−0.0614</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0027</td>
<td>0.0027</td>
<td>0.6528</td>
<td>0.0042</td>
</tr>
</tbody>
</table>
Simulation results: Simple Random Sampling Without Replacement

<table>
<thead>
<tr>
<th></th>
<th>Ratio</th>
<th>Relative Bias</th>
<th>RRMSE</th>
<th>Bias Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>1.0101</td>
<td>0.0101</td>
<td>0.1662</td>
<td>0.0611</td>
</tr>
<tr>
<td>Bootstrap WOR</td>
<td>0.9897</td>
<td>−0.0103</td>
<td>0.1633</td>
<td>−0.0631</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0067</td>
<td>0.0067</td>
<td>0.1656</td>
<td>0.0407</td>
</tr>
<tr>
<td><strong>RATIO</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>0.9865</td>
<td>−0.0135</td>
<td>0.1939</td>
<td>−0.0698</td>
</tr>
<tr>
<td>New Method</td>
<td>0.9831</td>
<td>−0.0169</td>
<td>0.1936</td>
<td>−0.0874</td>
</tr>
<tr>
<td><strong>MEDIAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>1.0427</td>
<td>0.0427</td>
<td>0.4745</td>
<td>0.0903</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0703</td>
<td>0.0703</td>
<td>0.5188</td>
<td>0.1367</td>
</tr>
<tr>
<td><strong>GINI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>1.0088</td>
<td>0.0088</td>
<td>0.4152</td>
<td>0.0335</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0044</td>
<td>0.0044</td>
<td>0.2619</td>
<td>0.0169</td>
</tr>
</tbody>
</table>
Simulation results: Unequal Probability Sampling Without Replacement

<table>
<thead>
<tr>
<th></th>
<th>Ratio</th>
<th>Relative Bias</th>
<th>RRMSE</th>
<th>Bias Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>1.0383</td>
<td>0.0383</td>
<td>0.2412</td>
<td>0.0992</td>
</tr>
<tr>
<td>Bootstrap WOR</td>
<td>0.9440</td>
<td>−0.0560</td>
<td>0.1936</td>
<td>−0.0027</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0004</td>
<td>0.0004</td>
<td>0.1985</td>
<td>0.0019</td>
</tr>
<tr>
<td><strong>RATIO</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>1.0301</td>
<td>0.0301</td>
<td>0.2034</td>
<td>0.1499</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0296</td>
<td>0.0296</td>
<td>0.2090</td>
<td>0.1430</td>
</tr>
<tr>
<td><strong>MEDIAN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>0.9975</td>
<td>−0.0025</td>
<td>0.4293</td>
<td>−0.0059</td>
</tr>
<tr>
<td>New Method</td>
<td>1.0248</td>
<td>0.0248</td>
<td>0.4673</td>
<td>0.0536</td>
</tr>
<tr>
<td><strong>GINI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bootstrap WR</td>
<td>0.9307</td>
<td>−0.0693</td>
<td>0.8546</td>
<td>−0.0814</td>
</tr>
<tr>
<td>New Method</td>
<td>0.9615</td>
<td>−0.0385</td>
<td>0.9496</td>
<td>−0.0406</td>
</tr>
</tbody>
</table>
Future work

Three possible directions for examining the performance of this new method:

- other sampling designs (two stages sampling design, balanced sampling, stratified sampling design, etc.)
- other $\theta$ (variance, QSR, etc.)
- comparison with other existing resampling methods (BBR, other types of bootstrap methods, etc.)
Bibliography


