Weight calibration and the survey bootstrap

Stas Kolenikov

Department of Statistics
University of Missouri-Columbia

March 7, 2011
Motivating questions

1. Why are the large scale samples always so complex?
2. Why do I need to use weights?
3. What is calibration, and why is it useful for complex survey data?
4. If I have calibrated my data, shouldn’t the standard errors go down?
5. How can I see that? What variance estimator should I use?
6. Simulation: interplay between calibration and various variance estimation methods.

Scattered additional ideas:
- Empirical likelihood
- Survey bootstrap
Introduction
Complex survey data
NAEP

Weights and calibration
Role of weights
Overview
Empirical likelihood

Variance estimation
Linearization
Calibration
Replication

Simulation
Setup
Parameter estimates
Variance estimates

Conclusion

References
Complex survey data

Most large scale data sets utilize features of complex survey sampling designs (Cochran 1977, Kish 1965, Korn & Graubard 1999):

- stratification: sampling independently within parts of population or from different frames
- clustering: sampling groups of observation units
- varying probabilities of selection: optimize performance of the sampling design
- multiple stages of selection: districts, schools, classes, students

Some features are aimed at improving precision, while others, at minimizing costs.
NAEP sample design

- A, B, C, . . . , Alpha, Beta, Gamma, Delta samples

<table>
<thead>
<tr>
<th>School Type</th>
<th>Grade</th>
<th>Assessment</th>
<th>Writing</th>
<th>Pilot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>4</td>
<td>State/National - Alpha</td>
<td>✪</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>National - Gamma</td>
<td>✪</td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>4</td>
<td>National - Delta</td>
<td>✪</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>4</td>
<td>State/National - Alpha</td>
<td>✪</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Stratification of public schools by jurisdiction, private schools by affiliation
- Systematic sampling of schools with probability proportional to size
- Sampling of students to attain (approximately) self-weighting sample
- Oversampling of (areas and schools with large % of) minorities
- Matrix sampling of items into booklets
Introduction
Complex survey data
NAEP

Weights and calibration
Role of weights
Overview
Empirical likelihood

Variance estimation
Linearization
Calibration
Replication

Simulation
Setup
Parameter estimates
Variance estimates

Conclusion

References
Setting

- Finite population $\mathcal{U}$ with $L$ strata and $N_h$ units in each
- Designs: sampling with replacement, $n_h$ units from stratum $h$
- Sample $s$ of size $n = \sum_h n_h = \text{total number of PSUs}$
- Estimates:
  - $t[y]$ for $T[Y]$
  - $\hat{\theta} = f(t_1, \ldots, t_k)$ for $\theta = f(T_1, \ldots, T_k)$
  - $\hat{\theta}$ from estimating equations $u(x, \hat{\theta}) = 0$
- Goal: estimation of design variance
Weighted estimation

- Expansion weight:
  \[ w_i = \frac{1}{\pi_i} = \frac{1}{\Pr[\text{unit } i \text{ in sample}]} \]

- Expansion estimator (Horvitz & Thompson 1952) of the total:
  \[ t[y] = \sum_{i \in s} w_i y_i \]

- Weighted mean:
  \[ \bar{y}_w = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i} = \frac{t[y]}{t[1]} \]

- Analytical use:
  \[ \hat{\beta}_{WLS} = (X'WX)^{-1} X'Wy, \quad W = \text{diag}(w_1, \ldots, w_n) \]

  Pseudo-maximum likelihood:
  \[ \sum_{i \in s} w_i l(\theta, x_i) \to \max \theta \]
Sampling weights

Pfeffermann (1993), Binder & Roberts (2009):

1. The weights can be used to test and protect against nonignorable sampling designs which could cause selection bias.
   - Without the weights, the parameters may be biased if the selection probabilities depend on characteristics of the units.

2. The weights can be used to protect against misspecification of the model holding in the population.
   - Estimation will be consistent for the census regression, even though a formal one.
Replicate weights

Besides the main sampling weights, microdata often come with (dozens or hundreds of) replicate weights.

- Necessary for variance estimation
- It suffices for the end user to run the model with all sets of weights and compute the variance estimator as the empirical variance of the results with different weights.
Accuracy of estimates

Stratification

Population

Sample

Weighting

\( \bar{y} \)
Accuracy of estimates

Stratification

Weighting

Population

Sample

Calibration

\( y \)

Sample data

Complex survey data
NAEP

Weights and calibration
Role of weights
Overview
Empirical likelihood

Variance estimation
Linearization
Calibration
Replication

Simulation
Setup
Parameter estimates
Variance estimates

Conclusion

Other expertise

References
Calibration


- Auxiliary information: variables $x_k$
- Either $x_k$’s are known for every unit $k$ in the population, or...
- ... the population totals, or independent control totals $T[x] = \sum_{k \in U} x_k$ are known
- Human population samples: demography (from census data or CPS)
- Calibrate the design weights $d_i \mapsto w_i$ so that calibration equations (benchmark constraints) hold:

$$\sum_{i \in s} w_i x_i = \sum_{k \in U} x_k \quad (1)$$

- Typical procedures: raking, generalized raking (Deville, Sarndal & Sautory 1993)
Extensions

- Instrumental variables calibration (Kott 2006)
- Calibration using information at different levels (Estevao & Särndal 2006, Isaki, Tsay & Fuller 2004)
- Restricted weight ranges (Chen, Sitter & Wu 2002, Théberge 2000)
Linear calibration

• A popular functional form:

\[ w_i = d_i(1 + \lambda' x_i) \]  \hspace{1cm} (2)

where \( \lambda \) is the vector of Lagrange multipliers

• Solution:

\[ \lambda = (\sum_{i \in s} d_i x_i x_i')^{-1} \left( \sum_{k \in U} x_k - \sum_{i \in s} w_i x_i \right) \]

• The resulting expansion estimator of the total is identical to GREG estimator (Särndal, Swensson & Wretman 1992)

• Disadvantage: weights may become negative.
Empirical likelihood

- Owen (1988) proposed to construct confidence intervals for the population mean using profile empirical likelihood function:

\[
R(\mu) = \left\{ \prod_{i=1}^{n} nw_i \mid w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i y_i = \mu \right\}, \tag{3}
\]

- Maximum empirical likelihood estimator:

\[
\hat{\mu} = \text{arg max } R(\mu)
\]

- Asymptotic distributional results:

\[
\hat{\mu} \text{ is asy. normal;} \quad -2 \log R(\hat{\mu}) \xrightarrow{d} \chi^2
\]

- Extensions from i.i.d. to regression models (Owen 1991), models based on estimating equations (Qin & Lawless 1994), . . .
Pseudo-EL for survey data

  \[ \tilde{l}(w) = \sum_{i \in s} d_i \ln w_i \]  

- Calibration use:
  \[ \tilde{l}(w) \rightarrow \max_w \text{ s.t. } N \sum_{i \in s} w_i x_i = \sum_{k \in U} x_k, w_i \geq 0, \sum_{i \in s} w_i = 1 \]  

- Solution:
  \[ w_i = \frac{Nd_i}{\sum_{j \in s} d_j 1 + \lambda' (x_i - \bar{X})} \]
  \[ 0 = \sum_{i \in s} \frac{d_i}{\sum_{j \in s} d_j 1 + \lambda' (x_i - \bar{X})} \frac{x_i - \bar{X}}{\sum_{j \in s} d_j 1 + \lambda' (x_i - \bar{X})} \]
Pseudo-EL for survey data

- Convex hull condition: the population value $\bar{X}$ is in the convex hull of the sample $x_i$
- Non-negative weights
- Convex optimization problem, solution always exists, and algorithms find it reliably (Wu 2004b)
- Chen, Sitter & Wu (2002) discuss how to utilize EL methods to construct bounded weights (may have to relax the calibrating constraints)
- Combining information from multiple surveys (Wu 2004a)
Variance estimation goals


Reporting and analytic purposes:

• a survey analyst needs standard errors to include in the report

• an applied researcher needs standard errors to construct confidence intervals and test their substantive models

Design purposes:

• a sample designer needs to know population variances to come up with efficient designs, strata allocations, small area estimators
**Horvitz-Thompson estimator**

- Expansion estimator

\[
    t[y] = \sum_{i \in s} d_i y_i = \sum_{i \in s} y_i / \pi_i
\]

- Variance (with respect to sampling, i.e., design-variance):

\[
    \nabla \{ t[y] \} = \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l}
\]

- Estimator of variance:

\[
    v \{ t[y] \} = \sum_{i,j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j}
\]
Linearization variance estimator

By taking Taylor series expansion near the population value $\theta$, one can obtain linearization estimators for:

- Function of moments:
  \[
  \nu_L[\hat{\theta}] \approx \hat{\text{MSE}}[\hat{\theta}] \approx \nu \left[ \sum_k \frac{\partial f}{\partial t_k} t_k \right]
  \]

- Example: ratio estimator of the total $t_r = t[y] T[x] / t[x]$, variance estimator
  \[
  \nu_L[t_r] = \frac{T[x]^2}{t[x]^2} \nu[e_i], \quad e_i = y_i - rx_i, \quad T[x] \neq 0
  \]

Remind notation: slide 2
Linearization variance estimator

By taking Taylor series expansion near the population value $\theta$, one can obtain linearization estimators for:

- Estimator defined by estimating equations:

$$nL[\hat{\theta}] \approx \text{MSE}[\hat{\theta}] \approx (\nabla u)^{-1} v[u(x, \hat{\theta})](\nabla u)^{-1}^T$$

- Examples: regression (Fuller 1975)

$$\sum_{i \in s} w_i u(x_i, y_i, \theta) \equiv \sum_{i \in s} w_i x_i (y_i - x_i^T \theta) = 0$$

- Closely related to White (1980) heteroskedasticity-robust estimator

- Other examples: GLM (Binder 1983), pseudo-MLE (Skinner 1989)

Remind notation: slide 2
Calibration estimators


- Calibration variables \( \mathbf{x} \)
- Parameter of interest: total of \( y \)
- Variance of the calibrated estimator:

\[
\mathbb{V}[t_{\text{cal}}] = \sum_{ij \in \mathcal{U}} (\pi_{ij} - \pi_i \pi_j)(d_i E_i)(d_j E_j),
\]

where \( E_i = y_i - \mathbf{x}_i' \mathbf{B} \) are residuals from the census regression (c.f. \( y_i - \bar{Y} \) for H-T estimator, \( y_i - r x_i \) for ratio)

- Variance estimator:

\[
\nu[t_{\text{cal}}] = \sum_{ij \in \mathcal{s}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} (w_i e_i)(w_j e_j)
\]

where \( e_i = y_i - \mathbf{x}_i' \hat{\mathbf{\beta}} \) are residuals from the sample (weighted) LS regression
Replication methods

For a given estimation procedure \((X_1, \ldots, X_n) \mapsto \hat{\theta}\):

1. To create data for replicate \(r\), reshuffle PSUs, omitting some and/or repeating others, according to a certain replication scheme
2. Using the original estimation procedure and the replicate data, obtain parameter estimate \(\hat{\theta}(r)\)
3. Repeat Steps 1–2 for \(r = 1, \ldots, R\)
4. Estimate variance/MSE as
   \[
   \nu_m[\hat{\theta}] = \frac{A}{R} \sum_{r=1}^{R} (\hat{\theta}(r) - \bar{\theta})^2 \tag{7}
   \]
   where \(A\) is a scaling constant, \(\bar{\theta} = \sum_r \hat{\theta}(r) / R\) for variance estimation and \(\bar{\theta} = \hat{\theta}\) for MSE estimation

Alternative implementation: use replicate weights \(w_{hij}^{(r)}\) instead of actually resampling the data.
The jackknife


- Replicates: omit only one PSU from the entire sample
- Replicate weights: if unit \( k \) from stratum \( g \) is omitted,

\[
w_{hij}^{(gk)} = \begin{cases} 
0, & h = g, i = k \\
\frac{ng}{ng-1} w_{hij}, & h = g, i \neq k \\
w_{hij}, & h \neq g 
\end{cases}
\]

- Number of replicates: \( R = n \)
- Variance estimators:

\[
v_{J1} = \sum_{h} \frac{n_h - 1}{n_h} \sum_{i} (\hat{\theta}(hi) - \hat{\theta}^h)^2
\]
\[
v_{J2} = \sum_{h} \frac{n_h - 1}{n_h} \sum_{i} (\hat{\theta}(hi) - \hat{\theta})^2
\]
Calibration bootstrap

Stas Kolenikov
U of Missouri

Introduction
Complex survey data
NAEP

Weights and calibration
Role of weights
Overview
Empirical likelihood

Variance estimation
Linearization
Calibration
Replication

Simulation
Setup
Parameter estimates
Variance estimates

Conclusion
Other expertise
References

Balanced repeated replication

(BRR)

- Design restriction: \( n_h = 2 \) PSUs/stratum
- Replicates (half-samples): omit one of the two PSUs from each stratum
- Replicate weights:

\[
 w_{hij}^{(r)} = \begin{cases} 
 2w_{hij}, & \text{PSU } hi \text{ is retained} \\
 0, & \text{PSU } hi \text{ is omitted} 
\end{cases}
\]

- (2nd order) balance conditions:
  - each PSU is used \( R/2 \) times
  - each pair of PSUs is used \( R/4 \) times
- Number of replicates:
  \( L \leq R \leq 2^L \)
- McCarthy (1969):
  \( L \leq R = 4m \leq L + 3 \) using Hadamard matrices
- Scaling factor in (7):
  \( A = 1 \)
Rescaling bootstrap (RBS)

Rao & Wu (1988): for parameter $\theta = f(\bar{x})$,

1. Sample with replacement $m_h$ out of $n_h$ units in stratum $h$.
2. Compute pseudo-values

$$\tilde{x}_h^{(r)} = \bar{x}_h + m_h^{1/2} (n_h - 1)^{-1/2} (\bar{x}_h^{(r)} - \bar{x}_h),$$

$$\tilde{x}^{(r)} = \sum_h W_h \tilde{x}_h^{(r)}, \quad \tilde{\theta}^{(r)} = f(\tilde{x}^{(r)})$$

(8)

3. Repeat steps 1–2 for $r = 1, \ldots, R$.
4. Compute $\nu_{RBS}[\hat{\theta}]$ using (7) with $A = 1$. 
Scaling of weights

Rao, Wu & Yue (1992): weights can be scaled instead of values.

- \( m_{hi}^{(r)} \) = # times the \( i \)-th unit from stratum \( h \) is used in the \( r \)-th replicate
- Replicate weight is

\[
W_{hik}^{(r)} = \left[ 1 - \left( \frac{m_h}{n_h - 1} \right)^{1/2} + \left( \frac{m_h}{n_h - 1} \right)^{1/2} \frac{n_h}{m_h} m_{hi}^{(r)} \right] W_{hik}
\]

- Equivalent to RBS for functions of moments
- Applicable to \( \hat{\theta} \) obtained from estimating equations
Bootstrap scheme options

Choice of $m_h$:

- $m_h \leq n_h - 1$ to ensure non-negative replicate weights
- $m_h = n_h - 1$: no need for internal scaling
- $m_h = n_h - 3$: matching third moments (Rao & Wu 1988)
- Simulation evidence (Kovar, Rao & Wu 1988): for $n_h = 5$, the choice $m_h = n_h - 1$ leads to more stable estimators with better coverage than $m_h = n_h - 3$

Choice of $R$:

- No theoretical foundations
- Popular choices: $R = 100, 200$ or $500$
- $R \geq$ design degrees of freedom $= n - L$
Mean bootstrap

Yung (1997), Yeo, Mantel & Liu (1999)

- Confidentiality protection: units with a weight of 0 belong to the same PSU, risk of identification.
- Replace the number of bootstrap draws $m^{(*r)}_h$ by

$$\bar{m}^{(*r)}_h = \frac{1}{K} \sum_{k=(r-1)K+1}^{rK} m^{(*k)}_h$$

- Take $K$ large enough so that $\Pr^*[\bar{m}^{(*r)}_h = 0]$ is small.
- Proceed to compute the bootstrap weights (9).
- Compute $v_{MBOOT}$ using (7) with scaling factor $A = K$.
- Number of resulting weight variables $= R/K$.

Warning: no formal theory have been developed so far.
Pros and cons of resampling estimators

+ Analysts and data users only need software that does weighted estimation
+ No need to program specific estimators for each model
+ No need to release strata and sampling unit identifiers in public data sets (jackknife?)
  ‒ Computationally intensive, especially when \( n \sim 10^3 – 10^4 \)
  ‒ Post-stratification and non-response adjustments need to be performed on every set of weights
  ‒ Bulky data files with many weight variables
Goal of this study

Give the calibrated estimators to people!

• Variance estimation for calibrated estimators requires knowledge of what the calibrating variable are, and what their population totals are

• Such information is unlikely to be available to analysis and public micro data users

• Resampling estimators must play a role
Population

- 1/10 of IPUMS 2000 Census 5% sample (Ruggles, Alexander, Genadek, Goeken, Schroeder & Sobek 2010)
- Alaska, Hawaii, DC excluded
- 1396105 individuals
- 2052 PUMAs
- From 298 to 2080 individuals in each PUMA
- Demographic variables: age, gender, race, ethnicity, education, language spoken
- Economic variables: income, employment status

PUMA = Public Use Microdata Area: 100,000+ residents; whole central city, MSA, PMSA, non-metro places
Sample

• 41 strata $\approx$ state (4 New England states conglomerated; 5 Mountains states conglomerated)
• PSU = PUMA
• In population, $11 \leq N_h \leq 233$ PSU/stratum
• In sample, $2 \leq n_h \leq 6$ PSU/stratum (more in larger strata)
• Resulting sample size: $n = 120$ PSUs
• 1st stage of selection: $\approx$ PPS (Rao-Hartley-Cochran method)
• 2nd stage of selection: 10–16 persons per PSU (fewer in rural areas, more in urban areas)
• 2000 Monte Carlo replications
• Sample size: 1215–1319 individuals
Statistics

Statistics collected:

- English proficiency (6 categories)
- Marital status (6 categories)
- Logistic regression: high school graduate = (age + age^2) × sex + black + white + hispanic
- Gini coefficient of income inequality (Lorenz curve)

Calibration variables (if used):

- Age groups (0 to 10, 10 to 20, ...)
- Race (white, black, Asian)
- Gender
- Hispanic (0/1)
- Strata indicators
- Linear calibration (2), EL calibration (5)
Variance estimation methods

Linearization $+$

Deville & Sarndal (1992) estimator for calibrated data $+$

\[
\left\{ \text{Jackknife, Bootstrap } R = 100, \text{ Bootstrap } R = 990, \right. \\
\left. \text{Mean bootstrap } R = 693 \times 20 \right\} \\
\times \left\{ \text{No calibration, Calibration of the main weight only,} \\
\text{Calibration of the main and replicate weights} \right\}
\]
### Weights

<table>
<thead>
<tr>
<th></th>
<th>Calibration method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Mean CV</td>
<td>0.539</td>
</tr>
<tr>
<td>Median of $\min{w_i}$</td>
<td>205</td>
</tr>
<tr>
<td>Median of $\max{w_i}$</td>
<td>3268</td>
</tr>
<tr>
<td>Median of $\frac{\max{w_i}}{\min{w_i}}$</td>
<td>16.2</td>
</tr>
<tr>
<td>$\min \frac{w_i}{d_i}$</td>
<td>0.639</td>
</tr>
<tr>
<td>$\max \frac{w_i}{d_i}$</td>
<td>1.589</td>
</tr>
</tbody>
</table>

Conclusion: calibration increases the spread of weights
Parameter estimates bias

Average relative bias of parameter estimates:

<table>
<thead>
<tr>
<th></th>
<th>Calibration method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Language</td>
<td>0.03%</td>
</tr>
<tr>
<td>Marital status</td>
<td>-0.14%</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>7.61%</td>
</tr>
<tr>
<td>Gini index</td>
<td>-0.40%</td>
</tr>
</tbody>
</table>

Conclusion: ??
## Parameter estimates efficiency

Average ratio of MSEs of parameter estimates:

<table>
<thead>
<tr>
<th>Parameter estimates</th>
<th>Calibration method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Language</td>
<td>1</td>
</tr>
<tr>
<td>Marital status</td>
<td>1</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>1</td>
</tr>
<tr>
<td>Gini index</td>
<td>1</td>
</tr>
</tbody>
</table>

Smaller than 1: calibrated is better.

**Conclusion:**

- substantial efficiency gains in simple statistics
- minute loss of efficiency in “complicated” statistics
Relative bias of variance estimates

Variance estimation for the descriptive statistics:

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>Calibration</th>
<th>Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>main</td>
<td>main+rep</td>
<td>main</td>
</tr>
<tr>
<td>Linearization</td>
<td>-2.5%</td>
<td>55.4%</td>
<td>55.0%</td>
</tr>
<tr>
<td>D &amp; S 1992</td>
<td>-10.5%</td>
<td>-10.5%</td>
<td></td>
</tr>
<tr>
<td>Jackknife</td>
<td>-1.8%</td>
<td>56.6%</td>
<td>35.2%</td>
</tr>
<tr>
<td>BS (R = 100)</td>
<td>-3.7%</td>
<td>53.5%</td>
<td>-1.8%</td>
</tr>
<tr>
<td>BS (R = 990)</td>
<td>-2.7%</td>
<td>55.4%</td>
<td>-1.1%</td>
</tr>
<tr>
<td>MBS 693 × 20</td>
<td>-2.7%</td>
<td>55.1%</td>
<td>-10.4%</td>
</tr>
</tbody>
</table>

Conclusion: calibration must be accounted for in variance estimation! Without such account, the design-consistent methods report the design variance of the uncalibrated point estimator.
Relative bias of variance estimates

<table>
<thead>
<tr>
<th></th>
<th>Calibration</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>EL</td>
<td>Linear</td>
<td></td>
</tr>
<tr>
<td></td>
<td>main</td>
<td>main+rep</td>
<td>main</td>
<td>main+rep</td>
</tr>
<tr>
<td>Logistic regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linearization</td>
<td>-11.2%</td>
<td>-11.8%</td>
<td>-11.1%</td>
<td></td>
</tr>
<tr>
<td>Jackknife</td>
<td>3.9%</td>
<td>3.9%</td>
<td>40.6%</td>
<td>4.6%</td>
</tr>
<tr>
<td>BS ($R = 100$)</td>
<td>13.0%</td>
<td>14.6%</td>
<td>15.5%</td>
<td>15.2%</td>
</tr>
<tr>
<td>BS ($R = 990$)</td>
<td>17.0%</td>
<td>16.7%</td>
<td>18.2%</td>
<td>17.4%</td>
</tr>
<tr>
<td>MBS 693 × 20</td>
<td>-10.4%</td>
<td>-11.0%</td>
<td>-16.4%</td>
<td>-10.3%</td>
</tr>
<tr>
<td>Gini index</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linearization</td>
<td>42.4%</td>
<td>38.6%</td>
<td></td>
<td>37.8%</td>
</tr>
<tr>
<td>Jackknife</td>
<td>-14.1%</td>
<td>-16.4%</td>
<td>14.6%</td>
<td>-16.9%</td>
</tr>
<tr>
<td>BS ($R = 100$)</td>
<td>-19.0%</td>
<td>-20.8%</td>
<td>-20.5%</td>
<td>-22.0%</td>
</tr>
<tr>
<td>BS ($R = 990$)</td>
<td>-18.4%</td>
<td>-20.5%</td>
<td>-19.7%</td>
<td>-21.1%</td>
</tr>
<tr>
<td>MBS 693 × 20</td>
<td>-18.2%</td>
<td>-20.2%</td>
<td>-25.3%</td>
<td>-20.9%</td>
</tr>
</tbody>
</table>

Conclusion: nothing works for Gini.
## CI coverage

Tail probabilities of the nominal 90% CI.

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>Calibration EL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>main</td>
</tr>
<tr>
<td></td>
<td></td>
<td>main+rep</td>
</tr>
<tr>
<td>Descriptive statistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most methods</td>
<td>4.0–4.3% + 7.8–7.5%</td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linearization</td>
<td>6.2% + 6.4%</td>
<td>6.3% + 6.6%</td>
</tr>
<tr>
<td></td>
<td>6.2% + 6.4%</td>
<td>6.3% + 6.6%</td>
</tr>
<tr>
<td>Jackknife</td>
<td>5.0% + 5.1%</td>
<td>5.0% + 5.2%</td>
</tr>
<tr>
<td></td>
<td>5.0% + 5.1%</td>
<td>5.0% + 5.2%</td>
</tr>
<tr>
<td>BS ($R = 100$)</td>
<td>4.5% + 4.5%</td>
<td>4.6% + 4.6%</td>
</tr>
<tr>
<td></td>
<td>4.5% + 4.5%</td>
<td>4.6% + 4.6%</td>
</tr>
<tr>
<td>BS ($R = 990$)</td>
<td>4.3% + 4.2%</td>
<td>4.3% + 4.4%</td>
</tr>
<tr>
<td></td>
<td>4.3% + 4.2%</td>
<td>4.3% + 4.4%</td>
</tr>
<tr>
<td>Mean BS $693 \times 20$</td>
<td>6.0% + 6.3%</td>
<td>6.2% + 6.4%</td>
</tr>
<tr>
<td></td>
<td>6.0% + 6.3%</td>
<td>6.2% + 6.4%</td>
</tr>
<tr>
<td>Gini index</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most methods</td>
<td>3.5% + 10.0%</td>
<td></td>
</tr>
</tbody>
</table>

Monte Carlo margin of error: $3\sigma = 1.5\%$

Conclusion: notable asymmetry for fractions and Gini.
Conclusions I

Limited simulation evidence suggests:

- Calibration improves precision of the descriptive statistics, not so much for the model parameter estimates
- The linearization and the mean bootstrap are the most stable methods, followed by the jackknife and by other bootstrap methods (not shown)
- Calibration using strata indicators does not work well with the jackknife. With calibration on demographics only (not shown), the jackknife demonstrated performance comparable to linearization and the mean bootstrap
Conclusions II

- Performance of the mean bootstrap was virtually the same as that of linearization and jackknife, where applicable, or Deville & Sarndal (1992) estimator for calibrated weights: it had similar (downward) bias, stability, and coverage.

- End users of complex survey data sets with calibrated main weights will not be able to construct adequate replicate weights.

- There are analyses that neither variance estimation method can handle adequately.
What I covered was...

1. Introduction
   Complex survey data
   NAEP

2. Weights and calibration
   Role of weights
   Overview
   Empirical likelihood

3. Variance estimation
   Linearization
   Calibration
   Replication

4. Simulation
   Setup
   Parameter estimates
   Variance estimates

5. Conclusion

6. References
Calibration bootstrap

Stas Kolenikov
U of Missouri

Introduction
Complex survey data
NAEP

Weights and calibration
Role of weights
Overview
Empirical likelihood

Variance estimation
Linearization
Calibration
Replication

Simulation
Setup
Parameter estimates
Variance estimates

Conclusion

Other expertise

References

My expertise

- Survey statistics
- Latent variable (structural equation) modeling
- Econometrics
- Multilevel modeling
- Statistical programming and computing
References


References III


References IV


References


References VI


