Replication methods for analysis of complex survey data in Stata

North American Stata User’s Group Meeting
Boston
August 2004

Nicholas Winter
Cornell University
nw53@cornell.edu
Complex Survey Designs

- Stratification
- Clustering
- Unequal Probabilities of Selection

→ Traditional calculations give wrong point estimates; easily fixed through weighting

→ Traditional i.i.d. statistical calculations give wrong variance estimates
Approaches

• Linearization
  – Taylor Series Expansion for each statistic
  – This is the world of Stata’s \texttt{svy}
  – Must be programmed separately for every estimator
  – Requires information on stratum and PSU (ie, cluster) membership for each sample element

• Replication Methods
  – Take multiple “pseudo” samples (or replicates) from the dataset
  – Variance calculated from the variance of the estimator across the replicates
  – Once programmed, can plug in any estimator
  – More and more public use datasets now include replicate weights
Replication Methods

• Balance Repeated Replication (BRR)
  – 2-PSU per stratum designs
  – Each replicate consists of half the PSUs (1 per stratum)
  – $2^l$ possible samples; can use appropriately selected subset

• Survey Jackknife
  – Drop one PSU from each replicate
  – Designs
    • 2-PSU per stratum (JK2)
    • 2+ PSU per stratum (JKn)
    • unstratified (JK1)
Basic Approach

- Given sampling weights and sample design information, calculate \( R \) sets of replicate weights
- Weights set to zero for excluded PSU(s)
- Other weights adjusted accordingly
Variance of an estimator $\Theta$ is embarrassingly easy to calculate, once you have the replicate weights:

\[
V(\Theta) = F \sum_{r=1}^{R} f_r (\Theta - \Theta_r)^2
\]

\[
V(\Theta) = F \sum_{r=1}^{R} f_r (\Theta - \Theta_r) (\Theta - \Theta_r)'
\]

across R replicates

$\Theta$ is the full-sample estimate

$\Theta_r$ is the estimate of $\Theta$ in the r’th replicate

F is a technique-specific scaling factor

$f_r$ is a replicate-specific scaling factor (JKn only)
Advantages of Replication

- Easily extended to new techniques
  - No new programming for new estimators
- PSU and Stratum membership information may not be available
  - Privacy concerns are making replication more common on publicly-released datasets
- Easy to incorporate post-stratification or raking, and non-response adjustments into variance estimation
  - Simply apply the post-stratification (raking, NR adjustment) to each set of replicate weights in turn
Disadvantages

• Not implemented in Stata
Generating the weights

- **survwght** (available on SSC)
  - Given sampling weight, strata, and PSU information
  - Calculates BRR, JK1, JK2, JKn replicate weights
  - Also does post-stratification, raking, non-response adjustments
Creating Replicate Weights

- Using National Health & Nutrition Examination Study (NHANES)
Doing Analysis

• **svr** package (also available from SSC)

• Counterparts to official Stata’s **svy** commands:
  – svrmean, svrtotal, svrratio
  – svrtab
  – svrmodel (for regression-style models: regression, logit/probit, ologit/ooprobit, poisson, etc. etc.)

• And some extras
  – svrcorr calculates variances for correlation coefficients
  – svrest turns any command that accepts weights into a replication-based survey estimator (analogy to –simul- or –jknife-)
svrset

```
. svrset clear
. svrset set meth brr
. svrset set pw finalwgt
. svrset set rw brr_*
. svrset set dof 31
. svrset set fay 0
. svrset list
  meth  brr
    pw  finalwgt
    rw   brr_1  brr_2  brr_3  brr_4  brr_5  brr_6  brr_7  brr_8  brr_9  brr_10  brr_11
         brr_12  brr_13  brr_14  brr_15  brr_16  brr_17  brr_18  brr_19  brr_20  brr_21
         brr_22  brr_23  brr_24  brr_25  brr_26  brr_27  brr_28  brr_29  brr_30  brr_31
         brr_32
  dof  31
  fay  0
psun <not set>
```
svymean vs. svrmean

---

### svymean

```
svymean bpsystol, by(sex)
```

**Survey mean estimation**

- `pweight`: finalwgt
- `strata`: strata
- `psu`: psu

<table>
<thead>
<tr>
<th>Mean</th>
<th>Subpop.</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>bpsystol</td>
<td>Male</td>
<td>129.9253</td>
<td>0.6432933</td>
<td>128.6132</td>
<td>131.2373</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>124.2027</td>
<td>0.7051858</td>
<td>122.7644</td>
<td>125.6409</td>
</tr>
</tbody>
</table>

---

### svrmean

```
svrmean bpsystol, by(sex)
```

**Survey mean estimation, replication (brr) variance method**

- `Analysis weight`: finalwgt
- `Replicate weights`: brr_1...
- Number of replicates: 32
- `k (Fay's method)`: 0.000

<table>
<thead>
<tr>
<th>Mean</th>
<th>Subpop.</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>bpsystol</td>
<td>Male</td>
<td>129.9253</td>
<td>0.6442178</td>
<td>128.6114</td>
<td>131.2391</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>124.2027</td>
<td>0.7094868</td>
<td>122.7557</td>
<td>125.6497</td>
</tr>
</tbody>
</table>
svrtab race diabetes, row se ci fse(\%5.4f) fci(\%3.2f) fcell(\%7.6f)

Cross-tabulation with replication-based (brr) standard errors

Analysis weight: finalwgt
Replicate weights: brr_1...
Number of replicates: 32
k (Fay's method): 0.000

<table>
<thead>
<tr>
<th></th>
<th>diabetes, 1=yes, 0=no</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.968046</td>
<td>0.031954</td>
<td>1.000000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0020)</td>
<td>(0.0020)</td>
<td>[0.96, 0.97]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.93, 0.99]</td>
<td>[0.03, 0.04]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.940965</td>
<td>0.059035</td>
<td>1.000000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0062)</td>
<td>(0.0062)</td>
<td>[0.93, 0.95]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.93, 0.95]</td>
<td>[0.05, 0.07]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.979664</td>
<td>0.020336</td>
<td>1.000000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0102)</td>
<td>(0.0102)</td>
<td>[0.94, 0.99]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.94, 0.99]</td>
<td>[0.01, 0.06]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.965754</td>
<td>0.034246</td>
<td>1.000000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0018)</td>
<td>(0.0018)</td>
<td>[0.96, 0.97]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.93, 0.99]</td>
<td>[0.03, 0.04]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Key: row proportions
(standard errors of row proportions)
[95% confidence intervals for row proportions]

Pearson:
Uncorrected chi2(2) = 21.3483
Design-based F(1.52, 47.02) = 14.9413 P = 0.0000
svrmodel

[SVY] pp. 32-33

```
.svrmodel highbp height weight age age2 female black, cmd(logit)

Logit estimates with replicate-based (brr) standard errors

Analysis weight: finalwgt
Replicate weights: brr_1...
Number of replicates: 32
k (Fay's method): 0.000
Number of obs = 10351
Population size = 1.172e+08
Degrees of freedom = 31
F( 6, 26) = 80.78
Prob > F = 0.0000

|    | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----|--------|-----------|-------|------|----------------------|
| highbp |        |           |       |      |                      |
| height | -0.0325996 | 0.0058292 | -5.59 | 0.000 | -0.0444884 to -0.0207109 |
| weight | 0.049074 | 0.003289 | 14.92 | 0.000 | 0.042366 to 0.0557821    |
| age    | 0.1541151 | 0.0211487 | 7.29  | 0.000 | 0.1109819 to 0.1972482    |
| age2   | -0.0010746 | 0.002063 | -5.21 | 0.000 | -0.0014953 to -0.0006539   |
| female | -0.356497 | 0.0870778 | -4.09 | 0.000 | -0.5340933 to -0.1789007   |
| black  | 0.3429301 | 0.1484334 | 2.31  | 0.028 | 0.0401982 to 0.6456619     |
| _cons  | -4.89574 | 1.132855 | -4.32 | 0.000 | -7.206214 to -2.585267     |

.lincom female+black
( 1)  female + black = 0

|    | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----|--------|-----------|-------|------|----------------------|
| highbp |        |           |       |      |                      |
| (1)  | -0.0135669 | 0.1727493 | -0.08 | 0.938 | -0.3658915 to 0.3387577   |
```
Postestimation

• svr routines work with the usual post-estimation commands
  – test (nee svytest)
  – lincom
  – etc.
• Some ugly programming here, but it works . . .
My personal favorite: svrest

```
. svrest "sum height" "r(mean) r(sd)"

Estimates with replication (brr) based standard errors

Command:         sum height
Analysis weight:  finalwgt
Replicate weights:  brr_1...
Number of replicates:  32                      Degrees of freedom =  31

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>stats</td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>t</td>
<td>P&gt;</td>
<td>t</td>
</tr>
<tr>
<td></td>
<td>168.4599</td>
<td>.1469785</td>
<td>1146.15</td>
<td>.000</td>
<td>168.1601</td>
</tr>
<tr>
<td></td>
<td>9.699111</td>
<td>.0754005</td>
<td>128.63</td>
<td>.000</td>
<td>9.545331</td>
</tr>
</tbody>
</table>
```
. svrest "regress weight height sex" "e(r2)", matrices(e(b)) liststats

Estimates with replication (brr) based standard errors

Command:          regress weight height sex
Analysis weight:  finalwgt
Replicate weights: brr_1...
Number of replicates: 32               Degrees of freedom = 31

|        | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|----------------------|
| **mat1** |        |           |       |       |                      |
| height | .6600245 | .028589  | 23.09 | 0.000 | .6017168 - .7183323  |
| sex    | -3.901195 | .60011  | -6.50 | 0.000 | -5.125127 - .2.677263 |
| _cons  | -33.35467 | 5.58121 | -5.98 | 0.000 | -44.73762 - .21.97171 |
| **stats** |        |           |       |       |                      |
| stat1  | .2617663 | .0094207 | 27.79 | 0.000 | .2425527 - .2809799  |

Key: height  height
       sex    sex
       _cons _cons
       stat1 e(r2)
A Note on Development

• **svrmodel** is fairly simple, really

• Means, totals, ratios, and tabulation
  – Stata’s **svy** commands are implemented as ado files
  – Internal **_svy** calculates variances of means, totals & ratios
  – Called by svymean, svytotal, svyratio and svytab
  – svrcalc.ado is reverse-engineered to return results in the form that **_svy** does, as expected by **svymean, svytotal, svytab**, etc.