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 **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

Basic Approach (2)

- Variance of an estimator Θ is embarrassingly easy to calculate, once you have the replicate weights:

$$V(\Theta) = F \sum_{r=1}^{R} f_r \left(\Theta - \Theta_r\right)^2$$

$$V(\Theta) = F \sum_{r=1}^{R} f_r \left(\Theta - \Theta_r\right)\left(\Theta - \Theta_r\right)'$$

across R replicates

Θ is the full-sample estimate

Θ_r is the estimate of Θ 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

- **survwgt** (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)

Stata Results

```
. use http://www.stata-press.com/data/r8/nhanes2d.dta, clear

. survwgt create brr , strata(strata) psu(psu) weight(finalwgt)

Obtaining hadamard matrix file...
Generating replicate weights................................

Created weights and set svr values:
   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>

.
```

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/oprobit, 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

Stata Results

```
. svymean bpsystol, by(sex)

Survey mean estimation

pweight:   finalwgt                                 Number of obs    =     10351
Strata:    strata                                   Number of strata =        31
PSU:       psu                                      Number of PSUs   =        62
                                                    Population size  = 1.172e+08

------------------------------------------------------------------------------
Mean     Subpop. |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
-----------------+------------------------------------------------------------
bpsystol         |
            Male |   129.9253    .6432933    128.6132    131.2373    5.482304
          Female |   124.2027    .7051858    122.7644    125.6409    5.162487
------------------------------------------------------------------------------

. svrmean bpsystol, by(sex)

Survey mean estimation, replication (brr) variance method

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

------------------------------------------------------------------------------
Mean     Subpop. |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
-----------------+------------------------------------------------------------
bpsystol         |
            Male |   129.9253    .6442178    128.6114    131.2391    5.498073
          Female |   124.2027    .7094868    122.7557    125.6497    5.225652
------------------------------------------------------------------------------
```

svrtab

[SVY] p. 78

```
. 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               Number of obs      =     10349
Replicate weights:    brr_1...               Population size    = 1.171e+08
Number of replicates: 32                     Degrees of freedom =        31
k (Fay's method):     0.000

----------------------------------------------
1=white, |
2=black, |        diabetes, 1=yes, 0=no
3=other  |          0             1      Total
---------+------------------------------------
   White |   0.968046      0.031954   1.000000
         |   (0.0020)      (0.0020)
         | [0.96,0.97]   [0.03,0.04]
         |
   Black |   0.940965      0.059035   1.000000
         |   (0.0062)      (0.0062)
         | [0.93,0.95]   [0.05,0.07]
         |
   Other |   0.979664      0.020336   1.000000
         |   (0.0102)      (0.0102)
         | [0.94,0.99]   [0.01,0.06]
         |
   Total |   0.965754      0.034246   1.000000
         |   (0.0018)      (0.0018)
         | [0.96,0.97]   [0.03,0.04]
----------------------------------------------
  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                Number of obs      =     10351
Replicate weights:      brr_1...                Population size    = 1.172e+08
Number of replicates:   32                      Degrees of freedom =        31
k (Fay's method):       0.000                   F(    6,      26)  =     80.78
                                                Prob > F           =    0.0000

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |  -.0325996   .0058292    -5.59   0.000    -.0444884   -.0207109
      weight |    .049074    .003289    14.92   0.000      .042366    .0557821
         age |   .1541151   .0211487     7.29   0.000     .1109819    .1972482
        age2 |  -.0010746   .0002063    -5.21   0.000    -.0014953   -.0006539
      female |   -.356497   .0870778    -4.09   0.000    -.5340933   -.1789007
       black |   .3429301   .1484334     2.31   0.028     .0401982    .6456619
       _cons |   -4.89574   1.132855    -4.32   0.000    -7.206214   -2.585267
------------------------------------------------------------------------------

. lincom female+black

 ( 1)  female + black = 0

------------------------------------------------------------------------------
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0135669   .1727493    -0.08   0.938    -.3658915    .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

------------------------------------------------------------------------------
             |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
stats        |
       stat1 |   168.4599   .1469785   1146.15    0.000     168.1601    168.7597
       stat2 |   9.699111   .0754005    128.63    0.000     9.545331    9.852891
------------------------------------------------------------------------------
```

svrest (2)

```
. 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.