Sampling II

LPO 9951 | Fall 2020

PURPOSE

In the last lecture, we discussed a number of ways to properly estimate the means and variances of complex survey designs. In this lecture, we'll discuss how to use Stata's internal `svy` commands and various variance estimation methods to mor e easily and correctly estimate what we want.

```
. capture log close                        // closes any logs, should they be open

. set linesize 90

. log using "sampling_part2.log", replace    // open new log
------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s1-06-sampling/sampling_part2.log
  log type:  text
 opened on:   6 Oct 2021, 10:59:45
```

Complex survey designs: Cluster sampling and stratification

In the NCES surveys you'll be using this semester, the designers combined a design that includes multistage cluster sampling with stratification. In ECLS, for example, the designers designated counties as *PSUs*. They next stratified the sample by creat ing strata that combined census region with msa status, percent minority, and per capita income. They then randomly selected schools within each *PSU* (schools were the *SSUs*) and then randomly selected kindergarteners within each school (students were the *TSUs*). They then created two strata for each school with Asian and Pacific Islander students in one stratum and all other students in the other. Students were randomly sampled within this second stratum. The target number of children per school w as 24.

Weights in complex survey designs such as the one employed with ECLS are calculated via the same that we discussed in the last lecture. Nothing changes except for the layers of complexity. The good news, however, is that we a researchers don't have to c ompute the weights ourselves. Instead, we can use information provided by the survey makers.

The *PSUs* that are provided by NCES are what is known as "analysis *PSUs*". They aren't the identifier for the actual school or student. Instead, they are allocated within strata (many times 2 *PSU* per strata). Strata themselves may be analysis strata , that is, not the same strata that were used to run the survey. Oftentimes, this is done in service of further protecting the anonymity of participants. As far your analyses go, the end result is the same, but sometimes this can be a source of confusio n.

Variance estimation in complex survey designs

There are four common options for estimating variance in complex survey designs:

1. Taylor series linearized estimates
2. Balanced repeated replication (BRR) estimates
3. Jackknife estimates
4. Bootstrap estimates

Remember that these are all estimates: you cannot directly compute the variance of quantities of interest from complex surveys. Instead, you must use one of these techniques, with trade-offs for each. We'll be using a couple of datasets for this lesson:

- *nhanes*, which is a health survey conducted using a complex survey design that comes with a variety of weights
- *nmihs_bs*, which is a survey of births that comes with bootstrap replicate weights

Let's start with the *nhanes* dataset from which we'd like to get average height weight and age for the US population. First, let's get the naive estimate:

```
. clear all                                   // clear memory


. set more off                                // turn off annoying "__more__" feature

. webuse nhanes2f, clear

. preserve

. keep sampl stratid psuid

. save nhanes2f_s, replace
file nhanes2f_s.dta saved

. restore

. mean age height weight

Mean estimation                             Number of obs = 10,337

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
         age |    47.5637   .1693381      47.23177    47.89564
      height |   167.6512   .0950124       167.465    167.8375
      weight |   71.90088   .1510277      71.60484    72.19692
```

```
--------------------------------------------------------------
```

We can also take a look at the sampling design, particularly the designation of strata and *PSUs*:

```
. tab stratid psuid

   Stratum |
identifier | Primary sampling unit
    , 1-32 |     PSU 1      PSU 2 |     Total
-----------+----------------------+----------
         1 |       215        165 |       380 
         2 |       118         67 |       185 
         3 |       199        149 |       348 
         4 |       231        229 |       460 
         5 |       147        105 |       252 
         6 |       167        131 |       298 
         7 |       270        206 |       476 
         8 |       179        158 |       337 
         9 |       143        100 |       243 
        10 |       143        119 |       262 
        11 |       120        155 |       275 
        12 |       170        144 |       314 
        13 |       154        188 |       342 
        14 |       205        200 |       405 
        15 |       189        191 |       380 
        16 |       177        159 |       336 
        17 |       180        213 |       393 
        18 |       144        215 |       359 
        20 |       158        125 |       283 
        21 |       102        111 |       213 
        22 |       173        128 |       301 
        23 |       182        158 |       340 
        24 |       202        232 |       434 
        25 |       139        115 |       254 
        26 |       132        129 |       261 
        27 |       144        139 |       283 
        28 |       135        163 |       298 
        29 |       287        215 |       502 
        30 |       166        199 |       365 
        31 |       143        165 |       308 
        32 |       239        211 |       450 
-----------+----------------------+----------
     Total |     5,353      4,984 |    10,337 
```

It's important to remember that these are *analysis* PSUs and strata, not the exact ones that were used in the survey design itself. Essentially the original strata are reassigned names that allow for deidintification, and then psus are assigned within the strata.

We can use the weights supplied with *nhanes* to get accurate estimates of the means, but the variance estimates will be off:

```
. mean age height weight [pw = finalwgt]

Mean estimation                          Number of obs = 10,337

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
         age |   42.23732   .1617236      41.92031    42.55433
      height |   168.4625   .1139787      168.2391     168.686
      weight |   71.90869   .1802768      71.55532    72.26207
--------------------------------------------------------------
```

`svyset` and `svy: <command>`

To aid in the analysis of complex survey data, Stata has incorporated the `svyset` command and the `svy:` prefix, with its suite of commands. With `svyset`, you can set the *PSU* (and *SSU* and *TSU* if applicable), the wei ghts, and the type of variance estimation along with the variance weights (if applicable). Once set, most Stata estimation commands such as `mean` can be combined with `svy:` in order to produce correct estimates.

Variance estimators

Taylor series linearized estimates

Taylor series linearized estimates are based on the general strategy of Taylor series estimation, which is used to linearize a non-linear function in order to describe the function in question. In this case, a Taylor series is used to approximate the fu nction, and the variance of the result is the estimate of the variance.

The basic intuition behind a linearized estimate is that the variance in a complex survey will be a nonlinear function of the set of variances calculated within each stratum. We can calculate these, then use the first derivative of the function that wou ld calculate the actual variance as a first order approximation of the actual variance. This works well enough in practice. To do this, you absolutely must have multiple *PSUs* in each stratum so you can calculate variance within each stratum.

This is the most common method and is used as the default by Stata. You must, however, have within-stratum variance among *PSUs* for this to work, which means that you must have at least two *PSUs* per stratum. This lonely PSU problem is common and diff icult to deal with. We'll return the lonely PSU later.

To set up a dataset to use linearized estimates in Stata, we use the `svyset` command:

```
. svyset psuid [pweight = finalwgt], strata(stratid)

Sampling weights: finalwgt
             VCE: linearized
     Single unit: missing
        Strata 1: stratid
 Sampling unit 1: psuid
           FPC 1: <zero>
```

Now that we've set the data, every time we want estimates that reflect the sampling design, we use the `svy: <command>` format:

```
. svy: mean age height weight
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 31           Number of obs   =      10,337
Number of PSUs   = 62           Population size = 117,023,659
                                Design df       =          31

--------------------------------------------------------------
             |             Linearized
             |       Mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
         age |   42.23732   .3034412      41.61844    42.85619
      height |   168.4625   .1471709      168.1624    168.7627
      weight |   71.90869   .1672315      71.56762    72.24976
--------------------------------------------------------------
```

As you can see, the parameter estimates (means) are exactly the same as using the weighted sample, but the standard errors are quite different: nearly twice as large for age, but actually smaller for weight.

Balanced repeated replication (BRR) estimates

In a balanced repeated replication (BRR) design, the quantity of interests is estimated repeatedly by using half the sample at a time. In a survey which is designed with BRR in mind, each sampling stratum contains two *PSUs*. BRR proceeds by estimating the quantity of interest from one of the *PSUs* within each stratum. For H strata, $2H$ replications are done, and the variance of the quantity of interest across these strata forms the basis for the estimate.

BRR weights are usually supplied with a survey. These weights result in appropriate half samples being formed across strata. BRR weights should generally be used when the sample was designed with them in mind, and not elsewhere. This can be a serious co mplication when survey data are subset.

To get variance estimates using BRR in Stata, you either need to have a set of replicate weights set up or you need to create a set of balanced replicates yourself. If the data has BRR weights estimates can be obtained as follows:

```
. webuse nhanes2brr, clear

. svyset [pw=finalwgt], brrweight(brr*) vce(brr)

Sampling weights: finalwgt
             VCE: brr
             MSE: off
     BRR weights: brr_1 .. brr_32
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: <observations>
           FPC 1: <zero>

. svy: mean age height weight
(running mean on estimation sample)

BRR replications (32)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
................................

Survey: Mean estimation          Number of obs   =      10,351
                                 Population size = 117,157,513
                                 Replications    =          32
                                 Design df       =          31

--------------------------------------------------------------
             |                BRR
             |       Mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
         age |   42.25264   .3013406      41.63805    42.86723
      height |   168.4599     .14663      168.1608    168.7589
      weight |   71.90064   .1656452       71.5628    72.23847
--------------------------------------------------------------
```

The `brrweight` option specified which variables constitute the brr weights, while the `vce` option says that variance should be calculated using the balanced repeated replication approach.

It's helpful to take a look at how BRR weights are related to PSUs and strata

```
. merge 1:1 sampl using nhanes2f_s

    Result                      Number of obs
    -----------------------------------------
    Not matched                            14
        from master                        14  (_merge==1)
        from using                          0  (_merge==2)

    Matched                            10,337  (_merge==3)
    -----------------------------------------

. order sampl finalwgt psu stratid brr*

. gsort stratid psuid
```

Jackknife estimates

The Jackknife is a general strategy for variance estimation, so named by Tukey because of its general usefulness. The strategy for creating a jackknifed estimate is to delete every observation save one, then estimate the quantity of interest. This is re peated for every single observation in the dataset. The variance of every estimate computed provides an estimate of the variance for the quantity of interest.

In a complex sample, this is done by *PSUs*, deleting each *PSU* one at a time and re-weighting the observations within the stratum, then calculating the parameter of interest. The variance of these parameters estimates is the within-stratum variance es timate. The within stratum variances calculated this way are then averaged across strata to give the final variance estimate.

The jackknife is best used when Taylor series estimation cannot be done, for instance in the case of lonely *PSUs*.

```
. webuse nhanes2jknife, clear
```

In Stata, the command is:

```
. svyset [pweight = finalwgt], jkrweight(jkw_*) vce(jackknife)

   Sampling weights: finalwgt
                VCE: jackknife
                MSE: off
  Jackknife weights: jkw_1 .. jkw_62
        Single unit: missing
           Strata 1: <one>
    Sampling unit 1: <observations>
              FPC 1: <zero>
```

Now we can compare the naive estimates with the `svyset` estimates:

```
. mean age weight height

Mean estimation                         Number of obs = 10,351

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
         age |   47.57965   .1692044      47.24798    47.91133
      weight |   71.89752   .1509381      71.60165    72.19339
      height |   167.6509   .0949079      167.4648    167.8369
--------------------------------------------------------------

. svy: mean age weight height
(running mean on estimation sample)

Jackknife replications (62)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
............

Survey: Mean estimation

Number of strata = 31            Number of obs   =      10,351
                                 Population size = 117,157,513
                                 Replications    =          62
                                 Design df       =          31

--------------------------------------------------------------
             |             Jackknife
             |       Mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
         age |   42.25264   .3026765      41.63533    42.86995
      weight |   71.90064   .1654453      71.56321    72.23806
      height |   168.4599   .1466141      168.1609    168.7589
--------------------------------------------------------------

. merge 1:1 sampl using nhanes2f_s

    Result                      Number of obs
    -----------------------------------------
    Not matched                            14
        from master                        14  (_merge==1)
        from using                          0  (_merge==2)

    Matched                            10,337  (_merge==3)
    -----------------------------------------
```

```
. order sampl finalwgt psu stratid jkw_*

. gsort stratid psuid

. browse sampl finalwgt psuid stratid jkw_*
```

Bootstrap estimates

The bootstrap is a more general method than the jackknife. Bootstrapping involves repeatedly resampling within the sample itself and generating estimates of the quantity of interest. The variance of these replications (usually many, many replications) p rovides an estimate of the total variance. In NCES surveys, within stratum bootstrapping can be used, with the sum of the variances obtained used as an estimate of the population variance. Bootstrapping is an accurate, but computationally intense method of variance estimation.

As with the jackknife, bootstrapping must be accomplished by deleting each *PSU* within the stratum one at a time, re-weighting, calculating the estimate, than calculating the bootstrap variance estimate from the compiled samples.

```
. webuse nmihs_bs, clear

. svyset idnum [pweight = finwgt], vce(bootstrap) bsrweight(bsrw*)

 Sampling weights: finwgt
              VCE: bootstrap
              MSE: off
Bootstrap weights: bsrw1 .. bsrw1000
      Single unit: missing
         Strata 1: <one>
  Sampling unit 1: idnum
            FPC 1: <zero>

. gen birthwgtlbs = birthwgt * 0.0022046
(7 missing values generated)

. mean birthwgtlbs

Mean estimation                          Number of obs = 9,946

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
 birthwgtlbs |   6.272294   .0217405      6.229678     6.31491
--------------------------------------------------------------
```

```
. svy: mean birthwgtlbs
(running mean on estimation sample)

Bootstrap replications (1,000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
.................................................. 1,000

Survey: Mean estimation            Number of obs   =      9,946
                                   Population size = 3,895,562
                                   Replications    =      1,000

--------------------------------------------------------------
             |   Observed   Bootstrap         Normal-based
             |       mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
 birthwgtlbs |    7.39743   .0143754      7.369255    7.425606
--------------------------------------------------------------

. gsort finwgt

. browse idnum finwgt bsrw*
```

Lonely *PSUs*

The most common problem that students have with complex surveys is what is known as "lonely *PSUs*." When you subset the data, you may very well end up with a sample that does not have mutliple *PSUs* per stratum. There are several options for what do i n this case:

- Eliminate the offending data by dropping strata with singleton *PSUs*. This is a terrible idea.
- Reassign the *PSU* to a neighboring stratum. This is okay, but you must have a reason why you're doing this.
- Assign a variance to the stratum with a singleton *PSU*. This could be the average of the variance across the other strata. This process is also known as "scaling" and generally is okat, but you should take a look at how different this stratum is fr om the others before proceeding.

The svyset command includes three possible options for dealing with loney *PSUs*. Based on the above, I recommend you use the `singleunit(scaled)` command, but with caution and full knowledge of the implications for your estimates.

Design Effects

Design effects are pretty old-school and shouldn't be used. That said, you will see these used in some older articles. These were used because most statistical programming languages weren't able to compute variance estimates from complex surveys up until about 2010. As a patchwork solution, the survey provider would calculate standard errors for some commonly used estimates from some common variables and look at how much bigger they were than naive estimates. The ratio between these would be averaged and called a design effect. For instance, if standard errors from a Taylor series linearized estimate were on average 1.3 times as big as naive standard errors then the design effect was 1.3. Do not use this approach, for hopefully obvious reasons.

Using variance estimation from different surveys

```
. use ../../data/plans.dta, clear

. svyset psu [pw=f1pnlwt],strata(strat_id)

Sampling weights: f1pnlwt
             VCE: linearized
     Single unit: missing
        Strata 1: strat_id
 Sampling unit 1: psu
           FPC 1: <zero>

. mean bynels2m

Mean estimation                        Number of obs = 16,160

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
```

```
-------------+------------------------------------------------
    bynels2m |   44.44327   .1187556      44.21049    44.67604
--------------------------------------------------------------

. svy: mean bynels2m
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 361            Number of obs   =     16,160
Number of PSUs   = 751            Population size = 3,388,462
                                  Design df       =        390

--------------------------------------------------------------
             |             Linearized
             |       Mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
    bynels2m |   44.74391   .2618191      44.22915    45.25866
--------------------------------------------------------------

. use ../../data/hsls_belong.dta, clear

. renvars *, lower

. svyset [pw=w1parent], brr(w1parent???) vce(brr)

Sampling weights: w1parent
             VCE: brr
             MSE: off
     BRR weights: w1parent001 .. w1parent200
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: <observations>
           FPC 1: <zero>

. prop x3hscompstat

Proportion estimation                        Number of obs = 808

      _prop_1: x3hscompstat = High school diploma
      _prop_2: x3hscompstat = GED, certificate of attendance,
      _prop_3: x3hscompstat = Dropped out
      _prop_4: x3hscompstat = Still enrolled
      _prop_5: x3hscompstat = Status unknown

--------------------------------------------------------------
```

```
             |                                     Logit
             | Proportion   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
x3hscompstat |
     _prop_1 |   .7376238   .0154765      .7061348    .7668526
     _prop_2 |   .0433168   .0071615       .031245    .0597649
     _prop_3 |   .0482673   .0075401      .0354439    .0654156
     _prop_4 |   .0680693   .0088606      .0526049    .0876591
     _prop_5 |   .1027228   .0106805      .0835711    .1256616
--------------------------------------------------------------

. svy: prop x3hscompstat
(running proportion on estimation sample)

BRR replications (200)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200

Survey: Proportion estimation     Number of obs   =        808
                                  Population size = 218,060.05
                                  Replications    =        200
                                  Design df       =        199

      _prop_1: x3hscompstat = High school diploma
      _prop_2: x3hscompstat = GED, certificate of attendance,
      _prop_3: x3hscompstat = Dropped out
      _prop_4: x3hscompstat = Still enrolled
      _prop_5: x3hscompstat = Status unknown

--------------------------------------------------------------
             |                 BRR                 Logit
             | Proportion   std. err.     [95% conf. interval]
-------------+------------------------------------------------
x3hscompstat |
     _prop_1 |   .7019053   .0249524      .6504961    .7486743
     _prop_2 |   .0385162   .0101322      .0228223    .0642919
     _prop_3 |   .0535511   .0134007      .0325002    .0870105
     _prop_4 |   .0705174   .0125091      .0494954    .0995331
     _prop_5 |   .1355101   .0183957      .1031456    .1760363
--------------------------------------------------------------

. use ../../data/nhes_example.dta, clear
```

```
. replace dpcolor=. if inlist(dpcolor, -8 ,-7, -6 ,-5 ,-4 ,-3 ,-2, -1)
(1,847 real changes made, 1,847 to missing)

. svyset epsu [pw=fewt] ,strat(estratum) singleunit(scaled)

Sampling weights: fewt
             VCE: linearized
     Single unit: scaled
        Strata 1: estratum
 Sampling unit 1: epsu
           FPC 1: <zero>

. prop dpcolor

Proportion estimation                   Number of obs = 3,997

      _prop_1: dpcolor = 1 No
      _prop_2: dpcolor = 2 Yes, some of them
      _prop_3: dpcolor = 3 Yes, all of them

--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
dpcolor      |
     _prop_1 |   .0542907   .0035841      .0476777    .0617615
     _prop_2 |   .1788842   .0060621      .1673063    .1910794
     _prop_3 |   .7668251   .0066884      .7534565    .7796808
--------------------------------------------------------------

. svy: prop dpcolor
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =     3          Number of obs   =      3,997
Number of PSUs   = 3,997          Population size = 13,693,230
                                  Design df       =      3,994

      _prop_1: dpcolor = 1 No
      _prop_2: dpcolor = 2 Yes, some of them
      _prop_3: dpcolor = 3 Yes, all of them

--------------------------------------------------------------
             |             Linearized            Logit
             | Proportion   std. err.     [95% conf. interval]
```

```
-------------+----------------------------------------------------
dpcolor      |
     _prop_1 |   .0639356   .0056819      .0536596    .0760214
     _prop_2 |   .2238697   .0100379      .2048063    .2441626
     _prop_3 |   .7121947   .0106956      .6907792     .732701
------------------------------------------------------------------

. rename fewt finalwgt


. svyset epsu [pw=finalwgt] , vce(brr) brrweight(fewt*)

Sampling weights: finalwgt
             VCE: brr
             MSE: off
     BRR weights: fewt1 .. fewt80
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: epsu
           FPC 1: <zero>

. prop dpcolor

Proportion estimation                    Number of obs = 3,997

      _prop_1: dpcolor = 1 No
      _prop_2: dpcolor = 2 Yes, some of them
      _prop_3: dpcolor = 3 Yes, all of them

------------------------------------------------------------------
             |                                      Logit
             | Proportion   Std. err.     [95% conf. interval]
-------------+----------------------------------------------------
dpcolor      |
     _prop_1 |   .0542907   .0035841      .0476777    .0617615
     _prop_2 |   .1788842   .0060621      .1673063    .1910794
     _prop_3 |   .7668251   .0066884      .7534565    .7796808
------------------------------------------------------------------

. svy: prop dpcolor
(running proportion on estimation sample)

BRR replications (80)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..............................
```

```
Survey: Proportion estimation     Number of obs   =      3,997
                                  Population size = 13,693,230
                                  Replications    =         80
                                  Design df       =         79

      _prop_1: dpcolor = 1 No
      _prop_2: dpcolor = 2 Yes, some of them
      _prop_3: dpcolor = 3 Yes, all of them

--------------------------------------------------------------
             |                BRR              Logit
             | Proportion   std. err.     [95% conf. interval]
-------------+------------------------------------------------
dpcolor      |
     _prop_1 |   .0639356   .0005379      .0628732    .0650147
     _prop_2 |   .2238697   .0009822      .2219208    .2258307
     _prop_3 |   .7121947   .0010494      .7101015    .7142789
--------------------------------------------------------------

. log close
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s1-06-sampling/sampling_part2.log
  log type:  text
 closed on:   6 Oct 2021, 11:00:28
-------------------------------------------------------------------------------------

. exit
```