Data validation

LPO 9951 | Fall 2020

Data validation refers to the process of ensuring that the characteristics of your data match the known characteristics of the population as measured by other analysts. If you have large discrepancies between your estimates and the estimates compiled by others, this is a clear "red flag" that something has gone wrong. Usually this is a problem that can be solved by going back to cleaning the data, but sometimes your sample may diverge in important ways from the samples collected by others. You will ne ed to state why this is the case in your write-up of the data.

Data validation can be done in several ways:

- You can compare the estimates from your dataset with the estimates from another analysis of the same dataset. This is what we will do with the datasets used in this class.
- Sometimes you will be the first one to analyze your dataset. In this case, you need to look for others who have collected similar samples and compare with them.
- Sometimes you won't have any other samples to work with. In this case, you'll need to see if there are population data that might be useful. Many people use the Census as a "check" on the data they have collected.
- Last, you need to use common sense. If you have data on private elite institutions of higher education, and you calculate an average tuition of $2,000, you can rest assured that you have not found a hidden bargain but rather a flaw in your data.

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

. set linesize 90

. log using "validation.log", replace    // open new log
--------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s1-08-validation/validation.log
  log type:  text
 opened on:  21 Oct 2020, 11:20:03

. clear all                              // clear memory


. global ddir "../../data/"
```

Calculating estimates and comparing them with known results

Today, we'll use the `plans` dataset. We're going to compare our results with several tables published by NCES. Let's start with educational expectations of high school sophomores. We start by survey setting the data:

```
. use ${ddir}plans.dta

. svyset psu [pw = bystuwt], str(strat_id) singleunit(scaled)

      pweight: bystuwt
          VCE: linearized
  Single unit: scaled
     Strata 1: strat_id
         SU 1: psu
        FPC 1: <zero>
```

Account for missing data

The next step is to account for missing data properly:

```
. local allvar bystexp bysex byrace byses1 f1psepln

. mvdecode `allvar', mv(-9/-2)
     bystexp: 924 missing values generated
       bysex: 819 missing values generated
      byrace: 924 missing values generated
      byses1: 924 missing values generated
    f1psepln: 1958 missing values generated

. recode bystexp (-1=8 )
(bystexp: 1450 changes made)

. label define expect 1 "Less than HS" ///                                      2 '


. label values bystexp expect

. label define race 1 "American Indian/AK Native" ///
                                   2 "Asian/PI" ///
                                   3 "African American/Black" ///
                                   4 "Hispanic No Race Specified" ///
                                   5 "Hispanic, Race Specified" ///
                                   6 "Multiracial, non Hispanic" ///
                                   7 "White"
```

```
. label values byrace race
```

Get estimates

Next, we tabulate expectations for college and compare it to a known estimate.

```
. tab bystexp

  how far in |
      school |
     student |
 thinks will |
get-composit |
           e |      Freq.     Percent        Cum.
-------------+-----------------------------------
Less than HS |        128        0.84        0.84
      HS/GED |        983        6.45        7.29
        2 Yr |        879        5.77       13.06
    Attend 4 |        561        3.68       16.74
   BA Degree |      5,416       35.55       52.29
    Master's |      3,153       20.69       72.99
         PhD |      2,666       17.50       90.48
 Don't Know' |      1,450        9.52      100.00
-------------+-----------------------------------
       Total |     15,236      100.00

. svy: proportion bystexp
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =     361        Number of obs   =     16,160
Number of PSUs   =     751        Population size =  3,408,319
                                  Design df       =        390

      _prop_1: bystexp = Less than HS
      _prop_2: bystexp = HS/GED
      _prop_3: bystexp = 2 Yr
      _prop_4: bystexp = Attend 4
      _prop_5: bystexp = BA Degree
      _prop_6: bystexp = Master's
      _prop_8: bystexp = Don't Know'
```

```
--------------------------------------------------------------
             |             Linearized           Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
bystexp      |
     _prop_1 |   .0094831     .00098       .007738    .0116172
     _prop_2 |   .0724693   .0030538      .0666899    .0787074
     _prop_3 |   .0643949   .0028925      .0589365    .0703211
     _prop_4 |   .0389852   .0018459      .0355139    .0427808
     _prop_5 |   .3578959   .0046507      .3488048    .3670902
     _prop_6 |   .1971035    .004424      .1885502    .2059464
         PhD |   .1608805   .0039873      .1531947    .1688749
     _prop_8 |   .0987875   .0030196      .0930076    .1048851
--------------------------------------------------------------
```

Once you create estimates from a command like `proportion` you can save them for later, using the `estimates store` command. These can be replayed using `replay` and can be brough back into memory using `restore`

```
. estimates store expect_tab

. estimates replay expect_tab

-------------------------------------------------------------------------------------
Model expect_tab
-------------------------------------------------------------------------------------

Survey: Proportion estimation

Number of strata =     361        Number of obs   =     16,160
Number of PSUs   =     751        Population size =  3,408,319
                                  Design df       =        390

      _prop_1: bystexp = Less than HS
      _prop_2: bystexp = HS/GED
      _prop_3: bystexp = 2 Yr
      _prop_4: bystexp = Attend 4
      _prop_5: bystexp = BA Degree
      _prop_6: bystexp = Master's
      _prop_8: bystexp = Don't Know'

--------------------------------------------------------------
             |             Linearized           Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
bystexp      |
     _prop_1 |   .0094831     .00098       .007738    .0116172
```

```
     _prop_2 |   .0724693   .0030538      .0666899    .0787074
     _prop_3 |   .0643949   .0028925      .0589365    .0703211
     _prop_4 |   .0389852   .0018459      .0355139    .0427808
     _prop_5 |   .3578959   .0046507      .3488048    .3670902
     _prop_6 |   .1971035    .004424      .1885502    .2059464
         PhD |   .1608805   .0039873      .1531947    .1688749
     _prop_8 |   .0987875   .0030196      .0930076    .1048851
--------------------------------------------------------------

. estimates restore expect_tab
(results expect_tab are active now)
```

Estimates can be stored using a simplified approach, using `eststo` and then the name of the estimates to be stored.

```
. eststo expect_tab: svy: tabulate bystexp
(running tabulate on estimation sample)

Number of strata   =       361                  Number of obs     =     15,236
Number of PSUs     =       751                  Population size   =  3,408,319
                                                Design df         =        390

----------------------
how far   |
in school |
student   |
thinks    |
will      |
get-compo |
site      | proportion
----------+-----------
 Less tha |      .0095
   HS/GED |      .0725
     2 Yr |      .0644
 Attend 4 |       .039
 BA Degre |      .3579
 Master's |      .1971
      PhD |      .1609
 Don't Kn |      .0988
          |
    Total |          1
----------------------
  Key:  proportion  =  cell proportion
```

Nicer tables

We get output in the console, but let's use the `eststo` and `esttab` commands to store our estimates and produce nicer tables. Using `esttab` alone, we'll get a nicely formatted table in the console. By adding `... us ing <file>` we save an `.rtf` version of the same table. We can easily paste this table in a paper.

```
. estpost svy: tabulate bystexp
(running tabulate on estimation sample)

Number of strata   =       361                  Number of obs     =     15,236
Number of PSUs     =       751                  Population size   =  3,408,319
                                                Design df         =        390

----------------------
how far   |
in school |
student   |
thinks    |
will      |
get-compo |
site      | proportion
----------+-----------
 Less tha |      .0095
   HS/GED |      .0725
     2 Yr |      .0644
 Attend 4 |       .039
 BA Degre |      .3579
 Master's |      .1971
      PhD |      .1609
 Don't Kn |      .0988
          |
    Total |          1
----------------------
  Key:  proportion  =  cell proportion

saved vectors:
             e(b) =  cell proportions
            e(se) =  standard errors of cell proportions
            e(lb) =  lower 95% confidence bounds for cell proportions
            e(ub) =  upper 95% confidence bounds for cell proportions
          e(deff) =  deff for variances of cell proportions
          e(deft) =  deft for variances of cell proportions
          e(cell) =  cell proportions
         e(count) =  weighted counts
           e(obs) =  number of observations
```

```
. eststo expect_tab

. esttab expect_tab using expect_tab.rtf, ///                 b(3) /// /* 3 decimal po

(output written to `"expect_tab.rtf"')
```

Validate with published data

Now that we have a clean table to look at, is this the same as Table 2 on page 22 of the report? Yes. Checking the standard errors on page B-3 reveals that these were also correctly done. Now we need to chec k this for all of the other variables in our dataset.

Not-so-quick Exercise

> I want you to replicate Table 34 on page 128 of NCES 2005-338. We'll split this up, but I want the class to come up with a single table that has exactly the same results as the NCES document.

```
. log close
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s1-08-validation/validation.log
  log type:  text
 closed on:  21 Oct 2020, 11:20:03
----------------------------------------------------------------------------------

. exit
```