

## Data validation

LPO 9951 | Fall 2018

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 need 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.

## 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 plans.dta

. // set up data for survey commands
. svyset psu [pw = bystwt], str(strat_id) singleunit(scaled)

      pweight: bystwt
      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 flpsepln

. // change values for vars in local that in (-4,-8,-9) to missing
. foreach myvar in `allvar' {
2.     replace `myvar' = . if `myvar' == -4
3.     replace `myvar' = . if `myvar' == -8
4.     replace `myvar' = . if `myvar' == -9
5. }
(648 real changes made, 648 to missing)
(276 real changes made, 276 to missing)
(0 real changes made)
(648 real changes made, 648 to missing)
(171 real changes made, 171 to missing)
(0 real changes made)
(648 real changes made, 648 to missing)
(276 real changes made, 276 to missing)
(0 real changes made)
(648 real changes made, 648 to missing)
(276 real changes made, 276 to missing)
(0 real changes made)
(1131 real changes made, 1131 to missing)
(781 real changes made, 781 to missing)
(46 real changes made, 46 to missing)
```

## 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-composite	Freq.	Percent	Cum.
-----+-----			
{don't know}	1,450	9.52	9.52
less than high school graduation	128	0.84	10.36
high school graduation or ged only	983	6.45	16.81
attend or complete 2-year college/schoo	879	5.77	22.58

attend college, 4-year degree incomplet		561	3.68	26.26
graduate from college		5,416	35.55	61.81
obtain master's degree or equivalent		3,153	20.69	82.50
obtain phd, md, or other advanced degree		2,666	17.50	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   =    16160
Number of PSUs   =    751      Population size =  3408319
                               Design df       =     390
```

```
_prop_1: bystexp = {don't know}
_prop_2: bystexp = less than high school graduation
_prop_3: bystexp = high school graduation or ged on
_prop_4: bystexp = attend or complete 2-year colleg
_prop_5: bystexp = attend college, 4-year degree in
_prop_6: bystexp = graduate from college
_prop_7: bystexp = obtain master's degree or equiva
_prop_8: bystexp = obtain phd, md, or other advance
```

-----				
		Linearized		
		Proportion	Std. Err.	[95% Conf. Interval]
-----+				
bystexp				
_prop_1		.0987875	.0030196	.0930076 .1048851
_prop_2		.0094831	.00098	.007738 .0116172
_prop_3		.0724693	.0030538	.0666899 .0787074
_prop_4		.0643949	.0028925	.0589365 .0703211
_prop_5		.0389852	.0018459	.0355139 .0427808
_prop_6		.3578959	.0046507	.3488048 .3670902
_prop_7		.1971035	.004424	.1885502 .2059464
_prop_8		.1608805	.0039873	.1531947 .1688749
-----				

## 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 `... using <file>` we save an

.rtf version of the same table. We can easily paste this table in a paper. If you are feeling bold, you could output the table in LaTeX format and incorporate into your LaTeX-formatted document.

```
. estimates store expect_tab

. // save as table using esttab
. esttab expect_tab using expect_tab.rtf, b(3) se(4) ///
>   varlabels(_prop_1 "Unsure" ///
>             _prop_2 "Less than HS" ///
>             _prop_3 "HS or GED" ///
>             _prop_4 "AA/AS" ///
>             _prop_5 "Some college" ///
>             _prop_6 "BA/BS" ///
>             _prop_7 "MA/MS" ///
>             _prop_8 "PhD or Prof") ///
>   replace
(output written to expect_tab.rtf)
```

```
.
. estpost svy: tabulate byrace bystexp, row percent
(running tabulate on estimation sample)
```

Number of strata	=	361	Number of obs	=	15236
Number of PSUs	=	751	Population size	=	3408318.6
			Design df	=	390

---

student~s								
race/ethn								
icity-com								
posite	{don~t k	less tha	high sch	attend o	attend c	graduate	obtain m	obtain p
amer, in	14.59	.4457	9.99	7.469	2.905	30.2	16.34	18.06
asian, h	10.01	1.104	3.363	3.207	4.165	33.54	21.2	23.42
black or	8.565	1.405	8.16	5.541	5.99	37.35	15.14	17.85
hispanic	12.79	1.26	10.5	6.849	6.581	34.14	15.47	12.42
hispanic	13.23	2.064	9.48	5.459	5.418	35.99	15.5	12.86
multirac	8.286	1.243	7.065	4.854	3.947	35.29	21.42	17.89
white, n	9.393	.6163	6.559	7.07	2.854	35.86	21.74	15.9
Total	9.879	.9483	7.247	6.439	3.899	35.79	19.71	16.09

---



---

| how far

	in school
	student
	thinks
student^s	will
race/ethn	get-compo
icity-com	site
posite	Total

-----+-----	
amer, in	100
asian, h	100
black or	100
hispanic	100
hispanic	100
multirac	100
white, n	100
Total	100
-----	

Key: row percentages

Pearson:

Uncorrected chi2(42) = 338.2312  
Design-based F(32.43, 12648.62)= 5.6934 P = 0.0000

saved vectors:

e(b) = row percentages  
e(se) = standard errors of row percentages  
e(lb) = lower 95% confidence bounds for row percentages  
e(ub) = upper 95% confidence bounds for row percentages  
e(deff) = deff for variances of row percentages  
e(deft) = deff for variances of row percentages  
e(cell) = cell percentages  
e(row) = row percentages  
e(col) = column percentages  
e(count) = weighted counts  
e(obs) = number of observations

row labels saved in macro e(labels)

column labels saved in macro e(eqlabels)

. estimates store expect\_tab2

```
.
. esttab expect_tab2 using expect_tab2.rtf, se nostar replace unstack ///
> varlabels(`e(labels)') eqlabels(`e(eqlabels)')
(output written to expect_tab2.rtf)
```

```

. // post clean table to output window
. esttab expect_tab, b(3) se(4) ///
>     varlabels(_prop_1 "Unsure" ///
>               _prop_2 "Less than HS" ///
>               _prop_3 "HS or GED" ///
>               _prop_4 "AA/AS" ///
>               _prop_5 "Some college" ///
>               _prop_6 "BA/BS" ///
>               _prop_7 "MA/MS" ///
>               _prop_8 "PhD or Prof")

```

```

-----
                        (1)
                Proportion
-----
bystexp
Unsure                0.099***
                      (0.0030)
Less than HS          0.009***
                      (0.0010)
HS or GED              0.072***
                      (0.0031)
AA/AS                  0.064***
                      (0.0029)
Some college          0.039***
                      (0.0018)
BA/BS                  0.358***
                      (0.0047)
MA/MS                  0.197***
                      (0.0044)
PhD or Prof           0.161***
                      (0.0040)
-----
N                      16160
-----

```

Standard errors in parentheses  
\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

.

*NB:* The /// at the end of each line in the `esttab` commands tells Stata to move to the next line but that the command isn't yet finished. Without this, the options would stretch far on one line: bad coding practice. I could have also changed the delimiter to ; like we did when reading in NCES datasets in the earlier lecture.

### **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 reveals that there were also correctly done. Now we need to check 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.

*Init: 25 August 2015; Updated: 17 October 2017*