

Examples of Standard Error Adjustment

Obtaining a Statistic Using Both SRS and Complex Survey Methods in Stata

This resource document will provide you with an example of the analysis of a variable in a complex sample survey dataset using Stata. A subset of the public-use version of the Early Child Longitudinal Studies ECLS-K rounds one and two data from 1998 accompanies this example, as well as a Stata "do" (program) file. The stratified probability design of the ECLS-K requires that researchers use statistical software programs that can incorporate multiple weights provided with the data in order to obtain accurate descriptive or inferential statistics.

Research question

This dataset training exercise will answer the research question "Is there a difference in mathematics achievement gain from fall to spring of kindergarten between boys and girls?"

Step 1- Get the data ready for use in Stata

There are two ways for you to obtain the data for this exercise. You may access a training subset of the ECLS-K Public Use File prepared specifically for this exercise by clicking [here](#), or you may use the ECLS-K Public Use File (PUF) data that is available at

<http://nces.ed.gov/ecls/dataproducts.asp>.

If you use the training dataset, all of the variables needed for the analysis presented herein will be included in the file. If you choose to access the PUF, extract the following variables from the online data file (also referred to by NCES as an ECB or "electronic code book"):

CHILDID	CHILD IDENTIFICATION NUMBER
C1R4MSCL	C1 RC4 MATH IRT SCALE SCORE (fall)
C2R4MSCL	C2 RC4 MATH IRT SCALE SCORE (spring)
GENDER	GENDER
BYCW0	BASE YEAR CHILD WEIGHT FULL SAMPLE
BYCW1 through C1CW90	BASE YEAR CHILD WEIGHT REPLICATES 1 through 90
BYCWSTR	BASE YEAR CHILD STRATA VARIABLE
BYCWPSU	BASE YEAR CHILD PRIMARY SAMPLING UNIT

Export the data from this ECB to SAS or SPSS format. Then use software to convert the SAS or SPSS to Stata (the ECB does not currently export directly to Stata). The Stat/Transfer program does this. To successfully use the syntax file provided with this exercise, be sure to name your file, 'ECLSK_c1c2_panel_demo'. Finally, download the Stata program ("do") file prepared for this exercise by clicking [here](#).

Statistical Analysis of NCES Datasets Employing a Complex Sample Design > Examples > Slide 11 of 13

Step 2- Use Stata to calculate an estimate and accompanying standard error

Start Stata, and then open the Stata "do" file. The code in the do file is detailed below along with the accompanying Stata output, which is shown below in ten point font. Note that the Stata commands from the do file are preceded by periods. You will notice that in this example dataset all the missing codes found in the original ECB version of the data have been changed to ".".

First do a "describe" command, to see a description of the variables.

```
. describe
Contains data from C:\ECLS Data\ECLSK c1c2 panel demo 8-31-13.dta
  obs:          21,409
  vars:          101             11 Sep 2013 13:19
  size:         8,542,191
-----
variable name   storage  display  value  variable label
                type    format   label
-----
childid        str8    %8s      child identification number
gender         byte    %8.0g    sex
bycw0         float  %9.0g    c1c2 child panel weight full...
c1r4rscl      float  %9.0g    c1 rc4 reading irt scale score
c1r4mscl      float  %9.0g    c1 rc4 math irt scale score
c2r4rscl      float  %9.0g    c2 rc4 reading irt scale score
c2r4mscl      float  %9.0g    c2 rc4 math irt scale score
bycw1         float  %9.0g    c1c2 child panel weight replica
* Replicate weight variables bycw2 through bycw90 have been omitted from this listing for brevity.
```

Note that gender is a categorical variable where a value of "1" represents male and "2" female. Next do a "tab gender" command to see the distribution of cases by gender.

```
. tab gender

      gender |          Freq.      Percent      Cum.
-----+-----
      male |          10,950          51.18          51.18
      female |          10,446          48.82         100.00
-----+-----
      Total |          21,396         100.00
```

Next create the math gain score variable that will be used in the analysis.

```
. generate mathgain= c2r4mscl- c1r4mscl
(3706 missing values generated)
```

Statistical Analysis of NCES Datasets Employing a Complex Sample Design > Examples > Slide 11 of 13

Using the list command, you can show several cases to examine the data values and check that the data look as expected. In this example, the first 20 cases are shown.

```
. list childid gender clr4msc c2r4mscl mathgain in 1/20
```

	childid	gender	clr4mscl	c2r4mscl	mathgain
1.	0001001C	female	39.54	50.1	10.56
2.	0001002C	female	44.44	58.93	14.49
3.	0001003C	female	28.57	32.81	4.240002
4.	0001004C	male	26.62	34.37	7.749998
5.	0001005C	female	40.88	38.86	-2.02
6.	0001006C	male	23.57	35.25	11.68
7.	0001007C	male	19.65	23.6	3.950001
8.	0001008C	female	36.27	53.31	17.04
9.	0001009C	male	20.82	29.09	8.27
10.	0001010C	female	26.85	32.22	5.370001
11.	0001011C	female	47.36	59.89	12.53
12.	0001012C	male	62.32	78.65	16.33
13.	0001013C	male	22.27	27.44	5.17
14.	0001014C	male	32.57	.	.
15.	0002002C	male	15.97	21.97	5.999999
16.	0002003C	male	40.94	48.69	7.75
17.	0002004C	female	33.26	35.24	1.980003
18.	0002005C	female	20.32	28.91	8.59
19.	0002006C	female	20.23	31.35	11.12
20.	0002007C	female	23.7	31.04	7.34

Stata data analysis under the different assumptions

For comparison purposes, you will first run the analysis as if this data were SRS, that is, a simple random sample **with no weight** adjustments for sampling design or nonresponse. In this first run, you will not apply any weight. In the second run, you will repeat a standard analysis (assuming SRS) **with the main sampling weight**. Finally, we will use not only the main sampling weight, but also the 90 replicate weights necessary to properly account for the complex sample design to calculate accurate estimates and their accompanying standard errors.

Statistical Analysis of NCES Datasets Employing a Complex Sample Design > Examples > Slide 11 of 13

First, the command below calculates mean math scores without weighting.

```
. mean mathgain, over(gender)

Mean estimation                Number of obs    =    17702

      male: gender = male
      female: gender = female

-----+-----
      Over |           Mean   Std. Err.   [95% Conf. Interval]
-----+-----
mathgain  |
      male |    10.52627   .0752745    10.37872    10.67381
      female |    10.17633   .0683613    10.04233    10.31032
-----+-----
```

The output above indicates that the average math score gain for boys is estimated as 10.53 with a standard error of 0.075. The average math score for girls is estimated as 10.18 with a standard error of 0.068. The answer to our main question about whether the difference of 0.35 in the gain scores of boys and girls depends on the accuracy of the mean gain scores and of these standard errors. If you run a t-test on these data, it will indicate that the difference is statistically significant at the $\alpha = 0.5$ level with a p value of 0.035.

However, the method shown above of estimating the average gain scores is misleading. Even in SRS analyses, when we have a main sampling weight, we must apply it.

Next, run the same analysis using the main sampling weight, bycw0.

```
. mean mathgain [pweight=bycw0], over(gender)

Mean estimation                Number of obs    =    17702

      male: gender = male
      female: gender = female

-----+-----
      Over |           Mean   Std. Err.   [95% Conf. Interval]
-----+-----
mathgain  |
      male |    10.36798   .0854712    10.20045    10.53551
      female |    10.15387   .0793504     9.998333    10.3094
-----+-----
```

The weighted analysis above shows that the average estimated mathematics score gain for boys has been reduced to 10.37 (with a slight increase in the standard error from 0.075 to 0.085) as compared to the unweighted estimate. The estimate for girls is reduced slightly to 10.15, and the standard error has increased from 0.068 to 0.079. Thus the gender difference is now 0.21, but a t-test run using the full sample weight gives a significance result similar to the unweighted version run previously.

Statistical Analysis of NCES Datasets Employing a Complex Sample Design > Examples > Slide 11 of 13

This method is also misleading because the complex sample design has not been taken into account to calculate the correct standard errors that accompany the properly weighted estimates.

Using “svyset” to account for the complex survey design.

In Stata, obtaining correct standard errors for complex survey designs uses the "svyset" command which identifies the main sampling weight, the replicate weights, and the primary method, which for ECLS-K data is "jackknife."

```
. svyset [pweight= bycw0], jkrweight(bycw1-bycw90) vce(jackknife)
      pweight: bycw0
      VCE: jackknife
      MSE: off
      jkrweight: bycw1 bycw2 bycw3 bycw4 bycw5 bycw6 bycw7 bycw8 bycw9 bycw10
* Replicate weight variables bycw11 through bycw90 have been omitted from this listing for brevity.
```

Now the correct standard errors can be obtained by using the prefix "svy" before the mean command.

```
. svy: mean mathgain, over(gender)
(running mean on estimation sample)
```

```
Jackknife replications (90)
Survey: Mean estimation
Number of strata =      1      Number of obs      =    20682
                                Population size      =   3804698
                                Replications          =      90
                                Design df             =      89

      male: gender = male
      female: gender = female
```

	Over	Mean	Jackknife Std. Err.	[95% Conf. Interval]	
mathgain					
male		10.36798	.1203234	10.1289	10.60706
female		10.15387	.1252753	9.904948	10.40279

Notice in the output above that the means are the same as for the weighted analysis, but now the standard errors are considerably higher, 0.120 for boys and 0.125 for girls.

Statistical Analysis of NCES Datasets Employing a Complex Sample Design > Examples > Slide 11 of 13

Now we can run a t-test for the difference in math gain scores between boys and girls using the regress command.

```
. svy: regress mathgain gender
(running regress on estimation sample)
```

```
Jackknife replications (90)
Survey: Linear regression
Number of strata = 1
Number of obs = 20683
Population size = 3804697.5
Replications = 90
Design df = 89
```

mathgain	Coef.	Jackknife Std. Err.	t	P> t	[95% Conf. Interval]	
gender	-.2141126	.1172426	-1.83	0.071	-.4470711	.0188459
_cons	10.58209	.2018742	52.42	0.000	10.18097	10.98321

We see that the standard error for the mean difference of 0.214 is 0.117, and the p value is 0.071, and not statistically significant at the $\alpha = 0.05$ level.

Step 3- Examine Results

We have now answered our research question, at least in the descriptive way. The gender difference in mean math gain score for children in ECLS-K from fall to spring is estimated at 0.21 (standard error of the contrast mean = 0.12) and does not represent a statistically significant difference in performance. We were only able to arrive at the correct answer by going away from the SRS analyses. With a complex survey data analysis program such as Stata, we can calculate appropriate standard errors that will give us more useful and accurate results when conducting significance testing or in creating confidence intervals in subsequent analysis steps.