Examples of Standard Error Adjustment

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

This resource document will provide you with an example of the analysis of a variable in a complex sample survey dataset using SAS. 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 an SAS syntax 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 SAS

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"):

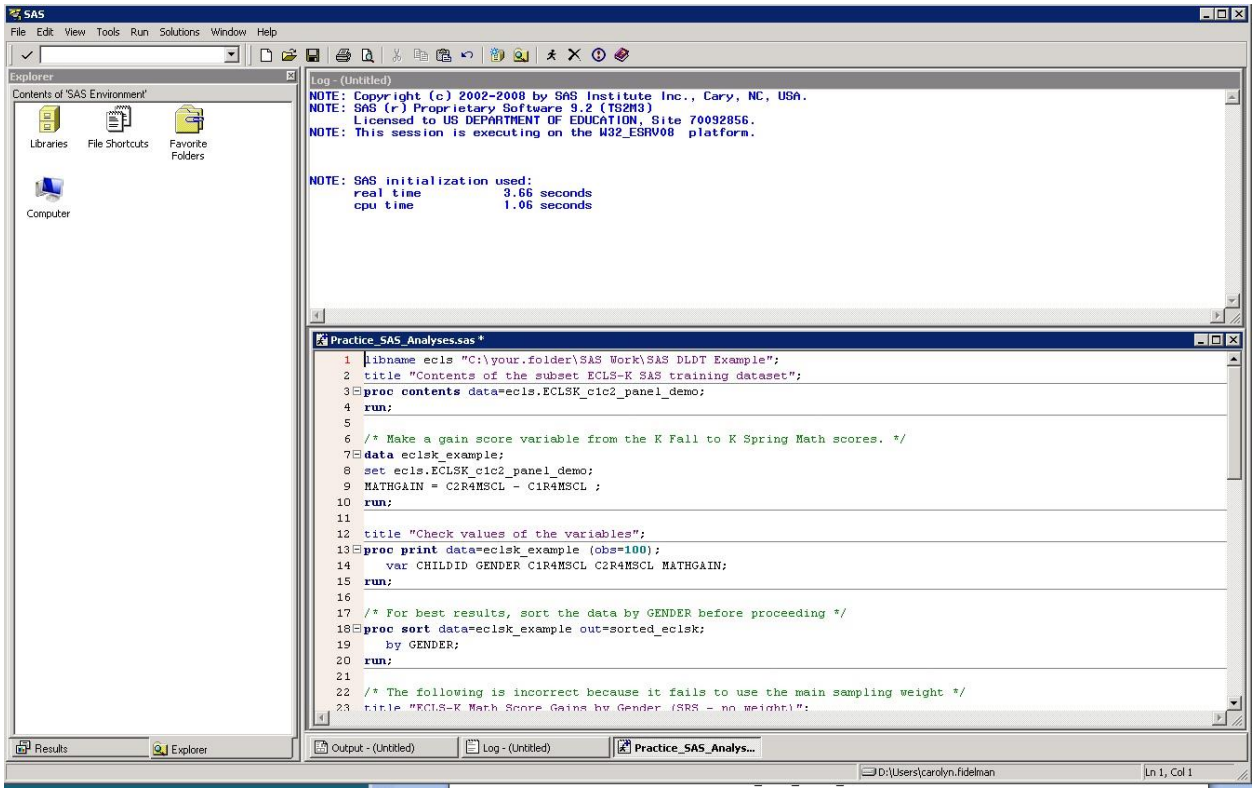
CHILDID	CHILD IDENTIFICATION NUMBER
C1R4MSCL	C1 RC4 MATH IRT SCALE SCORE (fall)
C2R4MSCL	C2 RC4 MATH IRT SCALE SCORE (spring)
GENDER	GENDER
BYCW0	BASE YEAR CHILD WEIGHT FULL SAMPLE
BYCW1 through C1CW90	BASE YEAR CHILD WEIGHT REPLICATES 1 through 90
BYCWSTR	BASE YEAR CHILD STRATA VARIABLE
BYCWPSU	BASE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

Export the data from this ECB to SAS format. You can always double click the data file to open it in SAS and to examine its contents, making sure all the variables needed for this exercise are available for use. Delete any extraneous variables that may show up when you extract the data. To successfully use the syntax file provided with this exercise, be sure to name your file, 'ECLSK_c1c2_panel_demo'.

Finally, download the SAS syntax file that has been prepared for this exercise by clicking here.

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

Start SAS, then open the training SAS syntax file 'Practice_SAS_Analyses.sas' from File/Open Program. The syntax you will use in this exercise will open in the editor window. Below is a screen shot of the typical SAS programming environment with the SAS syntax file "Practice_SAS_Analyses.sas" open in the editor window. All the code that is in that syntax file is also commented below.



The typical SAS start up steps are listed below, such as setting the library name and listing the contents of the file:

```
libname ecls "C:\your.folder\SAS Work\SAS DLDT Example";
title "Contents of the subset ECLS-K SAS training dataset";
proc contents data=ecls.ECLSK_c1c2_panel_demo;
run;
```

Next, referencing the training dataset, 'ecls.ECLSK_c1c2_panel_demo', you will create a temporary SAS dataset called 'eclsk_example' for use in the practice analyses. The temporary dataset will be a place where you can create a new gain score variable for

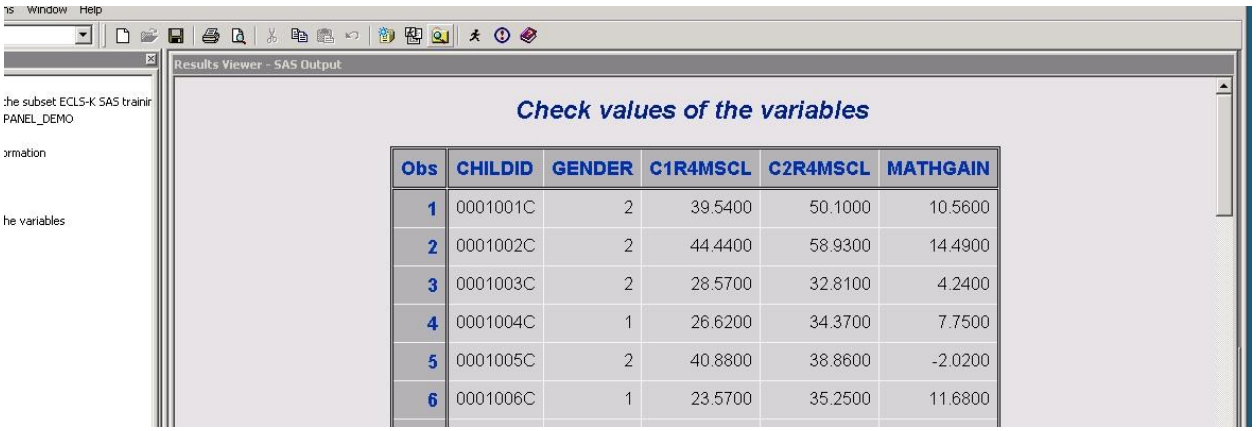
use in these analyses. You will notice that in this example dataset all the missing codes found in the original ECB version of the data have been changed to “.”.

```
/* Make a gain score variable from the K Fall to K Spring Math scores. */
data eclsk_example;
set ecls.ECLSK_c1c2_panel_demo;
MATHGAIN = C2R4MSCL - C1R4MSCL ;
run;
```

Use PROC PRINT to examine the data briefly to see that it looks as you expect. For a large dataset such as this, the following shows just the first 100 cases:

```
title "Check values of the variables";
proc print data=eclsk_example (obs=100);
   var CHILDID GENDER C1R4MSCL C2R4MSCL MATHGAIN;
run;
```

From this screen, you may examine the first few occurrences of MATHGAIN and verify that they are calculated in the way you expected.



Check values of the variables

Obs	CHILDID	GENDER	C1R4MSCL	C2R4MSCL	MATHGAIN
1	0001001C	2	39.5400	50.1000	10.5600
2	0001002C	2	44.4400	58.9300	14.4900
3	0001003C	2	28.5700	32.8100	4.2400
4	0001004C	1	26.6200	34.3700	7.7500
5	0001005C	2	40.8800	38.8600	-2.0200
6	0001006C	1	23.5700	35.2500	11.6800

Next, sort the data by the GENDER variable which is the classifier we will be are using to group children. SAS requires this kind of sorting.

```
/* For best results, sort the data by GENDER before proceeding */
proc sort data=eclsk_example out=sorted_eclsk;
   by GENDER;
run;
```

SAS data analysis under 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.

First, we will calculate simple descriptive statistics, the average math score gain of all children and then again, by gender.

```
/* The following is incorrect because it fails to use the main sampling
weight */
title "ECLS-K Math Score Gains by Gender (SRS - no weight)";
proc means data=sorted_eclsk  n nmiss min max mean median var stddev stderr;
     var MATHGAIN;
run;
proc means data=sorted_eclsk  n nmiss min max mean median var stddev stderr;
     by GENDER;
     var MATHGAIN;
run;
```

The output (pictured below) from the procedure 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 there is a statistically significant difference 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 with a p value of 0.035.

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

Results Viewer - SAS Output

ECLS-K Math Score Gains by Gender (SRS - no weight)

The MEANS Procedure

GENDER=.

Analysis Variable : MATHGAIN								
N	N Miss	Minimum	Maximum	Mean	Median	Variance	Std Dev	Std Error
1	12	8.6500015	8.6500015	8.6500015	8.6500015	.	.	.

GENDER=1

Analysis Variable : MATHGAIN								
N	N Miss	Minimum	Maximum	Mean	Median	Variance	Std Dev	Std Error
9000	1950	-48.1000023	75.3500042	10.5262678	9.5749998	50.9963039	7.1411696	0.0752745

GENDER=2

Analysis Variable : MATHGAIN								
N	N Miss	Minimum	Maximum	Mean	Median	Variance	Std Dev	Std Error
8702	1744	-19.3499985	46.1399994	10.1763261	9.4349995	40.6668051	6.3770530	0.0683613

Output - (Untitled) | Log - (Untitled) | Practice_SAS_Analyses.s... | Results Viewer - SAS Out...

D:\Users\carolyn.fidelman

As the SAS comment code indicates, 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 as in the code below:

```
/* The following is better because it uses the main sampling weight */
title "ECLS-K Math Gains by Gender (SRS-weighted)";
proc means data=sorted_eclsk  n nmiss min max mean median var stddev stderr;
        weight BYCW0;
     var MATHGAIN;
run;
proc means data=sorted_eclsk  n nmiss min max mean median var stddev stderr;
     by GENDER;
        weight BYCW0;
     var MATHGAIN;
run;
```

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

Results Viewer - SAS Output

ECLS-K Math Gains by Gender (SRS-weighted)

The MEANS Procedure

GENDER=.

Analysis Variable : MATHGAIN								
N	N Miss	Minimum	Maximum	Mean	Median	Variance	Std Dev	Std Error
1	12	8.6500015	8.6500015	8.6500015	8.6500015	.	.	.

GENDER=1

Analysis Variable : MATHGAIN								
N	N Miss	Minimum	Maximum	Mean	Median	Variance	Std Dev	Std Error
9000	1828	-48.1000023	75.3500042	10.3679803	9.4500008	10954.40	104.6632798	0.0747628

GENDER=2

Analysis Variable : MATHGAIN								
N	N Miss	Minimum	Maximum	Mean	Median	Variance	Std Dev	Std Error
8702	1649	-19.3499985	46.1399994	10.1538677	9.4099979	8685.08	93.1937582	0.0686126

Output - (Untitled) | Log - (Untitled) | Practice_SAS_Analyses.s... | Results Viewer - SAS Out...

D:\Users\carolyn.fidelman

The output pictured above shows that the average mathematics score gain for boys appears to be 10.37 with little change in the standard error from the unweighted estimate. For girls it has changed to 10.15, but also nearly same standard error as before. A *t*-test run using the full sample weight gives a significance result similar to the unweighted version run previously.

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 commands from SUDAAN

In order to obtain the average gain in mathematics scale score for all children by gender while accounting for the complex survey design, we will run the following code.

```
/* SUDAAN CODE */
title "Population Estimate for Math Gains using SUDAAN";
PROC DESCRIPT data=sorted_eclsk filetype=sas design=jackknife;
   WEIGHT BYCW0;
   jackwgts BYCW1--BYCW90 / adjjack=.999999999;
   TABLES GENDER ;
   SUBGROUP GENDER ;
   LEVELS    2 ;
      var MATHGAIN;
run;
```

Notice that we specified "jackknife" for the design and included the 90 replicate weights in this code. SUDAAN will only recognize one kind of jackknife design and that is JK1. Also, we specified the main sampling weight as we did in the SRS analyses previously.

Notice that in order to see the output from a SUDAAN program you must choose the Output tab at the bottom of the SAS window or choose View/Output from the file menu at the top of the screen. It does not offer an html version of output as regular SAS does.

Once you run the *t*-test of the difference in means using the SUDAAN complex sample analysis program, you see that the *p* value is really above 0.07, well over the $p<0.05$ usually needed to indicate a statistically significant difference in two statistics.

```
/* Run a t test of the difference in means by gender */
 PROC DESCRIPT data=sorted_eclsk filetype=sas design=jackknife;
   WEIGHT BYCW0;
   jackwgts BYCW1--BYCW90 / adjjack=.999999999;
   var MATHGAIN;
      CONTRAST GENDER = (-1 1);
      PAIRWISE GENDER ;
      DIFFVAR GENDER = ( 1 2 ) ;
   SUBGROUP GENDER ;
   LEVELS    2 ;
      PRINT MEAN="Math" SEMEAN T_MEAN P_MEAN;
run;
```

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 SUDAAN, 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. The table below shows the results for all the analyses performed in this exercise.

Comparison of results for SAS (SRS) and SUDAAN (complex survey methods)

SRS no weight

ALL

N	N Miss	Mean	Median	Std Error
17703	3706	10.35	9.51	0.051

GENDER=1

N	N Miss	Mean	Median	Std Error
9000	1950	**10.53**	9.58	0.075

GENDER=2

N	N Miss	Mean	Median	Std Error
8702	1744	**10.18**	9.44	0.068

SRS weighted

ALL

N	N Miss	Mean	Median	Std Error
17703	3489	10.26	9.43	0.051

GENDER=1

N	N Miss	Mean	Median	Std Error
9000	1828	10.37	9.45	0.075

GENDER=2

N	N Miss	Mean	Median	Std Error
8702	1649	10.15	9.41	0.069

SUDAAN

ALL

N	Pop N	Mean	Median	Std Error
17702	3,804,698	10.26	9.42	0.11

GENDER=1

N	Pop N	Mean	Median	Std Error
9000	1,959,828	10.37	9.45	0.12

GENDER=2

N	Pop N	Mean	Median	Std Error
8702	1,844,869	10.15	9.40	0.13