Examples of Standard Error Adjustment

Obtaining a Statistic Using Both SRS and Complex Survey Methods in the R Package 'survey'

This resource document will provide you with an example of the analysis of a variable in a complex sample survey dataset using the package '*survey*' by Thomas Lumley (2012, v.3.28-2) available for use in the R statistical programming language. 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 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 R

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 (the only two options available for export in the ECB). You can always double click the data file to open it in SAS or SPSS 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. Then export the file to comma-separated values format (.csv). When you save the file, make sure it goes into the working folder where your analysis will take place for this exercise.

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

Step 2- R startup steps

Open R. Change the default directory to your working directory (where the files needed for this exercise have been stored).

Call the first two packages (libraries) you will need using the following command lines:

```
library(Hmisc)
library(foreign)
```

Use the following import code to bring the data into R. Note that this changes the various missing codes to the R system-missing code:

```
eclskdata <-read.table("ECLSK_c1c2_panel_demo.csv", sep=",", header = TRUE,
na.strings=c("NA",".","-9","-8","-7","-1"))
```

Set the variable names to lower case for readability and then look at the variables and their numeric position in the data:

```
names(eclskdata)  <- tolower(names(eclskdata))
names(eclskdata)
```

Examine the data briefly to see that it looks as you expect. *C2R4MSCL*, *GENDER, etc.* are now *c2r4mscl*, *gender, etc.* For a large dataset such as this, the following shows just the first 50 cases:

```
head(eclskdata, n=50)
```

You notice that R recognizes the "." and -9 as missing values and calls them "NA" for the variables *c2r4mscl*, *gender,* etc.

Since you will be analyzing gain scores in Math, create the new variable *mathgain.*

```
eclskdata$mathgain=c2r4mscl-c1r4mscl
```

Add a count variable to the dataset for checking n at various stages

```
eclskdata$countn <- c(1)
```

Now make your data available to R for analyses:

```
attach(eclskdata)
```

At this point, you may opt to save your data back out to .csv in case you want to keep the changes you have just made to the dataset for later use.

```
write.csv(eclskdata,file="eclskdata_example.csv", row.names=FALSE)
```

R 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. Later, you will compare these results with those obtained from using the complex survey design. A summary of results from these three analyses are presented in Appendix A for your reference.

Remember that in R you will need to specify na.rm=TRUE to make some of these commands work correctly where there are missing values. Now run the following to find out how many cases there are, number of missing and some other summary statistics for gender and math gain scores:

```
length(eclskdata$countn)
length(eclskdata$gender)
summary(eclskdata$gender)
summary(eclskdata$mathgain)
```

Obtain the SRS mean and standard error for math gain score for all sampled children:

```
mean(eclskdata$mathgain, na.rm=TRUE)
all_stderr_mgain <-
sqrt(var(eclskdata$mathgain,na.rm=TRUE)/length(na.omit(eclskdata$mathgain)))
all_stderr_mgain
length(eclskdata$mathgain)
```

Make subset datasets for boys and girls.

```
subset(eclskdata, subset=(gender==1)) -> boys.math
attach(boys.math)
subset(eclskdata, subset=(gender==2)) -> girls.math
attach(girls.math)
```

Produce counts, missing data summaries, and mean gain scores of boys.

```
summary(boys.math$mathgain)
mean(boys.math$mathgain, na.rm=TRUE)
boys_stderr_mgain <-
sqrt(var(boys.math$mathgain,na.rm=TRUE)/length(na.omit(boys.math$mathgain)))
boys_stderr_mgain
length(boys.math$mathgain)
```

While the number of boys is reported as 10,950, only 9,000 have a math gain score due to missing values in either of round one or two of data collection. Next, produce the counts, missing data summaries, and mean gain scores of girls.

```
summary(girls.math$mathgain)
mean(girls.math$mathgain, na.rm=TRUE)
girls_stderr_mgain <-
sqrt(var(girls.math$mathgain,na.rm=TRUE)/length(na.omit(girls.math$mathgain)))
girls_stderr_mgain
length(girls.math$mathgain)
```

While the number of girls is reported as 10,446, only 8,702 have a math gain score, once the missing are removed from the calculation. We see from the report that the average math gain score of boys is 10.53 and the average math gain score for girls is 10.18 score points. 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 their accompanying standard errors). To find out if the difference in means is statistically significant under the assumption of a simple random sample, we will run a *t*-test.

```
t.test(mathgain ~ gender, data = eclskdata)
```

With the *p* value of 0.001, or probability of 0.001 that the results are different only by chance, you might conclude that there is a gender difference in math performance between girls and boys for the kindergarten year. **This method of estimating the average gain scores is misleading.** Even in the SRS analyses, when we have a main sampling weight, we must apply it using the code below.

Using commands from the package 'weights'

Basic descriptive statistics and *t*-tests can be run using a weighted SRS approach through the R package *weights*.

```
library(weights)
```

In this run, we will repeat the analysis (assuming SRS) with the main sampling weight.

```
all_mgain_m <- weighted.mean(eclskdata$mathgain, eclskdata$bycw0, na.rm=TRUE)
all_mgain_m
```

Before obtaining the weighted standard error of the mean, you must first implement the function "weighted.var.se" :

```
weighted.var.se <- function(x, w, na.rm=TRUE)
{
  if (na.rm) { w <- w[i <- !is.na(x)]; x <- x[i] }
  n = length(w)
  xWbar = weighted.mean(x,w,na.rm=na.rm)
  wbar = mean(w)
  out = n/((n-1)*sum(w)^2)*(sum((w*x-wbar*xWbar)^2)-2*xWbar*sum((w-wbar)*(w*x-
wbar*xWbar))+xWbar^2*sum((w-wbar)^2))
  return(out)
}
```

This function computes the variance of a weighted mean following Cochran's 1977 definition. After implementing the function, you can reuse it as shown below at any time.

```
mgain_se <- weighted.var.se(eclskdata$mathgain, eclskdata$bycw0)
mgain_se
```

Using the separate dataset for boys that was defined earlier, you can answer the question "What are the weighted mean gain scores for boys?"

```
boysresult_m <- weighted.mean(boys.math$mathgain, boys.math$bycw0, na.rm=TRUE)
boysresult_m

boysresult_se <- weighted.var.se(boys.math$mathgain, boys.math$bycw0)
boysresult_se

describe(~boys.math$mathgain, weights=boys.math$bycw0)
```

Now obtain the weighted mean gain scores for girls.

```
girlsresult_m <- weighted.mean(girls.math$mathgain, girls.math$bycw0, na.rm=TRUE)
girlsresult_m

girlsresult_se <- weighted.var.se(girls.math$mathgain, girls.math$bycw0)
girlsresult_se

describe(~girls.math$mathgain, weights=girls.math$bycw0)
```

Finally, we will run a *t*-test to find out if the difference in boys' and girls' math gain scores is statistically different.

```
wtd.t.test(boys.math$mathgain, girls.math$mathgain, weight=boys.math$bycw0,
 weighty=girls.math$bycw0)
```

Your result will indicate that the difference is statistically significant. **However, this method is also misleading because the complex sample design has not been taken into account.** Now 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.

Using commands from the package 'survey'

First load the package 'survey'.

```
library(survey)
```

This R package by Thomas Lumley, professor of Professor of Biostatistics, Department of Statistics, University of Auckland (previously at the University of Washington), is described in detail at http://faculty.washington.edu/tlumley/survey/. Presentation slides on it are at http://faculty.washington.edu/tlumley/survey/survey-wss.pdf.

One problem that you must deal with in using this dataset is that a certain number of the records have a missing value in the weights. Survey will not work where these cases are present, therefore you must subset them out.

```
eclskdata_nomis <- subset(eclskdata, bycw0>=0)
summary(eclskdata_nomis$bycw0)
length(eclskdata_nomis$bycw0)
attach(eclskdata_nomis)
```

Using the new subset eclskdata_nomis you may proceed to the next step.

Specify the survey design for the ECLS data

Before specifying the survey design you will need to create a scaling variable to use in the rscales specification. The code below just gets the number of records and then makes the scaling variable that we will call *ecls_rscales*.

```
wgt_n <- length(eclskdata_nomis$childid)
wgt_n
eclsk_rscales  <-wgt_n/(wgt_n-1)
eclsk_rscales
```

You will also need to find out in what column range the replicates reside. When you got the names in one of the steps above you might have noticed the column numbers associated with the replicates. You can also find this out using the *grep* command:

```
grep("bycw1", names(eclskdata_nomis))
grep("bycw90", names(eclskdata_nomis))
```

You find that bycw1 is first found in column 10, while the rest listed are e.g. bycw14. The last replicate value, bycw90, is found in column 99.

Now implement the survey design:

```
jk1_eclskdata_nomis <-
   svrepdesign(
      data = eclskdata_nomis,
      repweights = eclskdata_nomis[, 10:99],
      type = "JK1",
      weights = ~bycw0,
      rscales = eclsk_rscales
   )
```

Ignore the warning that you get upon implementing this design. Check that the design is as you specified by running the following:

```
jk1_eclskdata_nomis
```

Descriptive statistics in 'survey'

Obtain the counts of all children in this population and then again by gender. Notice that unwtd.count in svyby gives SRS counts for these.

```
svytable(~countn, jk1_eclskdata_nomis)
svytable(~gender, jk1_eclskdata_nomis)
svyby(~mathgain, ~gender, jk1_eclskdata_nomis, unwtd.count, na.rm=TRUE)
```

Next examine the mean math gain scores of all children in this population and then by gender.

```
svymean(~mathgain, jk1_eclskdata_nomis, na.rm=TRUE)
svyby(~mathgain, ~gender, jk1_eclskdata_nomis, svymean, na.rm=TRUE)
```

Finally, we will run a t-test to examine the difference in mean gain scores now that we are accurately accounting for the study design and nonresponse.

```
svyttest(gender~mathgain, jk1_eclskdata_nomis)
```

Once we run this analysis, we will see that there is **not** a statistically significant difference in mathematics achievement gain from fall to spring of kindergarten between boys and girls. This exercise has shown that applying complex survey analysis techniques may yield different results than those that incorrectly assume SRS and do not utilize the study weights.

Many more types of analyses can be performed using the package 'survey' for R. For an in depth explanation of the theory behind the analysis types available in this package see Thomas Lumley's 2010 book *Complex Surveys: A Guide to Analysis Using R*, Wiley Series in Survey Methodology.

Appendix A

Summary of results from the three analysis types on mathematics scores demonstrated in this document.

	all	1	2
unweighted n	17703	9000	8702
unweighted mean	10.35	10.53	10.18
unweighted se mean	0.051	0.075	0.068
weighted n	3804806	1959828	1844869
weighted mean	10.26	10.37	10.15
weighted se mean	0.003	0.007	0.006
complex weighted n	3863204	1993056	1870148
complex weighted mean	10.26	10.37	10.15
complex weighted se mean	0.108	0.120	0.125