

Subpopulation Command

Many analyses focus on specific subpopulations, such as high school graduates, public school students, postsecondary students, or STEM majors. When analyzing data based on complex samples, each case in the dataset contains important information needed to properly calculate variance estimates. Rather than deleting cases from your dataset that are not part of your subpopulation of interest, or using common subsetting commands such as “if” to limit cases entering your analysis, be sure to use the subpopulation selection command that is part of your software package. In Stata, this command is “subpop”, while in SAS the command is “domain”. This portion of the Education Longitudinal Study of 2002 (ELS:2002) Third Follow-Up Data File Documentation (NCES 2014-364) will help you use subsetting commands in SAS and Stata to properly analyze ELS data. Consult your software package’s technical documentation in conjunction with this resource document for additional information.

6.3.1 Standard Errors¹

For probability-based sample surveys, most estimates are nonlinear statistics. For example, a mean or proportion, which is expressed as $\Sigma wy / \Sigma w$,² is nonlinear because the denominator is a survey estimate of the (unknown) population total. In this situation, the variances of the estimates cannot be expressed in closed form. One common procedure for estimating variances of survey statistics is the Taylor series linearization procedure. This procedure takes the first-order Taylor series approximation of the nonlinear statistic and then substitutes the linear representation into the appropriate variance formula based on the sample design. Woodruff (1971) presented the mathematical formulation of this procedure. The variance estimation must also take into account stratification and clustering. There are other variance estimation procedures, such as jackknife and BRR. Taylor series and BRR estimation were supported for the base year, first follow-up, and second follow-up and is also supported for the third follow-up.

Variance estimation procedures assumed a with-replacement design at the first stage of sampling. Because school sampling rates were moderately low, this assumption yields estimates that are only slightly biased in the positive direction. For stratified multistage surveys and a with-replacement sample design, the Taylor series procedure requires the specification of analysis strata and analysis PSUs. In the base year, 361 analysis

¹ Taken from Ingels, S.J., Pratt, D.J., Alexander, C.P., Jewell, D.M., Lauff, E. Mattox, T.L., and Wilson, D. (2014). Education Longitudinal Study of 2002 Third Follow-up Data File Documentation (NCES 2014-364). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved [date] from <http://nces.ed.gov/pubsearch>.

² Σw is the estimated population and y is the corresponding value of the variable for which a mean or proportion is calculated. In the case of estimation of a proportion, y is a 0/1 variable indicating whether a certain characteristic is present for the sample member. In the case of estimation of a mean, y is either a continuous or ordinal value reported by or associated with the sample member.

Considerations for Analysis of ELS:2002 Data > Analyzing Subpopulations > Slide 17 of 19

strata were formed from the sampling strata used in the first stage of sampling, and the analysis PSUs were the individual schools. Given that the school sample was selected using probability with minimum replacement (pmr), for variance estimation in the base year, variance estimation strata were formed consisting of two PSUs per stratum (Chromy 1981). However, when there were an odd number of schools in a sampling stratum, one of the analysis strata formed had three PSUs. The same analysis strata and PSUs as in the base year were used in the first follow-up, in the second follow-up, and in the third follow-up.

As described in chapter 3, the ELS:2002 base-year sampling design was a stratified two-stage design. A stratified sample of schools was selected with probabilities proportional to a composite measure of size at the first stage, and a stratified systematic sample of students was selected from sample schools at the second stage. At the first stage, the school sampling rates varied considerably by school sampling strata. At the second stage, Asian and Hispanic students were sampled at higher rates than other students. Because of this complex sampling design, statistical analyses should be conducted using software that properly accounts for the complex survey design.

Many commonly used statistical computing packages assume that the data were obtained from a simple random sample; that is, they assume that the observations are independent and identically distributed. When the data have been collected using a complex sampling design, the simple random sampling assumption usually leads to an underestimate of the sampling variance, which would lead to artificially small confidence intervals and liberal hypothesis test results (i.e., rejecting the null hypothesis when it is in fact true more often than indicated by the nominal Type I error level) (Carlson, Johnson, and Cohen 1993).

Statistical strategies that have been developed to address this issue include first-order Taylor series expansion of the variance equation, BRR, and the jackknife approach (Wolter 2007). Special-purpose software packages that have been developed for analysis of complex sample survey data include SUDAAN, WesVar, Stata, SPSS, SAS, R, and AM. Evaluations of the relative performances of these packages are reported by Cohen (1997).

- SUDAAN is a commercial product developed by RTI; information regarding the features of this package and its lease terms is available from the website <http://www.rti.org/sudaan>.
- WesVar is a product of Westat, Inc.; information regarding the features of this package and its lease terms is available from the website <http://www.westat.com/our-work/information-systems/wesvar-support/download-wesvar>
- Information regarding the features of Stata and its lease terms is available from the website <http://www.stata.com>.

Considerations for Analysis of ELS:2002 Data > Analyzing Subpopulations > Slide 17 of 19

- SPSS is a product of IBM; information regarding the features of this package may be found on the website <http://www-01.ibm.com/software/analytics/spss/products/statistics>.
- SAS information (SAS-STAT User's Guide) can be found at <http://support.sas.com/documentation/>.
- R information is available at <http://www.r-project.org>.
- AM software, a product developed by the American Institutes for Research (AIR), can be downloaded for free from the website <http://am.air.org/>.

The following pseudo code is an example of generic SUDAAN code used to produce estimates and standard errors using Taylor series. The symbols /* and */ in the code indicate the beginning and end of a comment. These comments contain **words that are meant to be replaced** by file or variable names relevant to your specific analyses. Note that the dataset must be sorted by analysis strata and analysis PSUs before analyzing the data in SUDAAN.

```
proc descript data=/* insert filename34*/ design=wr;
nest analstr analpsu; /* these variables are the analysis strata and analysis PSUs,
respectively */
weight F3QWT;
var /*insert variables*/;
subpopn /* insert domain of interest if domain is a subset of students*/;
print nsum mean semean ;
run;
```

Corresponding Stata version 9 code is:

```
drop _all
set memory 18000
use "/* insert filename */", clear
sort analstr analpsu /* these variables are the analysis strata and analysis PSUs,
respectively */
svyset analpsu [pweight=f3qwt], strata(analstr)
svy: tab /*insert variables*/, subpop (name of domain) row se
Earlier versions of Stata require the following syntax:
svyset [pweight=f3qwt], strata(analstr) psu(analpsu)
svytab /*insert variables*/, subpop (name of domain) row se
```

References

- Carlson, B.L., Johnson, A.E., and Cohen, S.B. (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, 9(4): 795–814.
- Chromy, J.R. (1981). Variance Estimators for a Sequential Sample Selection Procedure. In D. Krewski, R. Platek, and J.N.K. Rao (Eds.), *Current Topics in Survey Sampling* (pp. 329–347). New York: Academic Press.
- Cohen, S.B. (1997). An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data. *The American Statistician*, 57(13): 285–292.
- Wolter, K. (2007). *Introduction to Variance Estimation*. Second Edition. New York: Springer-Verlag.
- Woodruff, R.S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association*, 66, 411–414.