Complex Survey Design Using Stata

This document provides a basic overview of how to handle complex survey design using Stata. Probability weighting and compensating for clustered and stratified samples are common survey design features that Stata can help you address in your statistical analyses. Stata is publicly available in both the library and residential computer clusters. Stata is also available on Stanford’s UNIX system. To use any of the public computers on campus, or to access Stata on UNIX, you must have a SUNet ID (Stanford University Network Identifier). See the following URL for information on obtaining a SUNet ID:

http://www.stanford.edu/group/itss/services/sunetid/

Note: Stata commands are indicated in bold type. Variable names in italics should be substituted with your actual variable names. For a list of Stata’s default settings, estimation methods and subcommands not discussed here, please consult Stata’s reference manuals (available in the SSRC at Green Library) or its online resources at: http://www.stata.com/links/resources1.html.

Table of Contents

What is Complex Survey Design? ................................................................. 2
Specifying Survey Design Features .......................................................... 2
  Specifying Survey Design Features Using Menus .................................. 2
  Specifying Survey Design Features Using Syntax .................................. 3
SVY Commands ...................................................................................... 2-4
Subpopulation Analysis .......................................................................... 4
For More Information and Assistance ..................................................... 4-5
  Stata Tutorials ..................................................................................... 4
  Documentation and Books .................................................................... 4
  Consulting ............................................................................................ 5

What is Complex Survey Design?

Many common statistical analyses assume that the data being analyzed constitute a simple random sample, in which everyone in the population of interest has exactly the same probability of being chosen during sample selection. In real-life survey research, however, sampling frequently works differently. Samples are often stratified by variables of interest (such as region of the country or racial background of the individual) to ensure a balanced number of respondents for each category of the variable. For example, a national survey in the U.S. may have two strata: urban areas and rural areas and within each stratum, individuals may be sampled independently. Additionally, survey samples are frequently clustered; for example, schools are sampled first, then a certain number of students are chosen from within each sampled school. In this case, schools are the survey’s primary sampling units, even though students may be the units of interest in the analysis.

Both of these sample features can skew the standard errors from results of statistical analyses. Because standard errors affect significance levels, the conclusions you draw from an analysis that does not take these complex survey design features into account may be false.
Weighting is another common feature of survey samples. Many surveys over-sample subpopulations of interest in order to include enough people for meaningful statistical analysis of a subgroup. For example, Cuban Americans may be oversampled in a national survey in order for researchers to be able to analyze their ethnic group on its own. In this example, then, the survey includes disproportionately more Cuban Americans than the U.S. population in general. Therefore, when a survey uses oversampling, results cannot be generalized to the broader population until probability weights are applied. These weights take into account the greater probability that Cuban Americans will be included in the sample compared to other groups. It is important to apply probability weights if you want to generalize your results to a larger population.

Stata commands that can adjust the standard errors to correct for complex survey design factors and simplifies oversampling can be found in the following sections.

### Specifying Survey Design Features

In Stata, you can use the `svyset` command to specify the stratification scheme, sampling weights and primary sampling units (used in clustering) for your data. The keywords used here are `strata` (which specifies the stratification variable) and `pweight` (which specifies the weight variable). You must also specify the primary sampling units (PSUs). Stata allows you to specify the data’s survey design in two ways: using drop-down menus and using syntax.

#### Specifying Survey Design Features Using Menus

If you are using menus to run commands, you should choose the Survey data analysis submenu under the Statistics menu. Select Setup & utilities, then Declare survey design for dataset. Enter the relevant information in the Strata and Sampling units boxes. To specify weights, click on the Weights tab, select the Sampling weight variable button and input the name of the variable identifying the sampling weight. On the Main tab, note that you can increase the number of stages to specify the strata and PSUs for multistage designs.

#### Specifying Survey Design Features Using Syntax

If you are typing syntax into the Command window instead of using drop-down menus, then you can submit this information in a single command line:

```
svyset yourpsu [pweight=yourpweight], strata(yourstrata)
```

You do not need to enter every subcommand in this line; for example, if your survey only uses cluster sampling, then you do not need to include the `pweight` and `strata` subcommands.

Since Stata will remember these specifications and will apply them to every survey analysis procedure (SVY command) that follows it, you will only need to enter it once at the start of your program.

### SVY Commands

The following table lists many of the most commonly used SVY commands for survey data. You can use these commands once you have given Stata the relevant information about the survey design using
the commands above. Each of the SVY commands described in this section produces estimates that are corrected for the complex design of the survey data that you are analyzing. Note that the commands needed to analyze survey data, require the use of the svy: prefix.

For additional information, consult Stata’s help function. For example, for more information on the svy: regress command, type: help svy: regress.

**General/Set-Up**

svyset Specify survey design features
svydescribe Describe strata and PSUs

**Means, Proportions, Ratios and Totals**

svy: mean Estimate population means
svy: proportion Estimate population proportions
svy: ratio Estimate population ratios
svy: total Estimate population totals

**Cross-Tabulations**

svy: tabulate One- and Two-way contingency tables

**Regression Models**

svy: regress Estimate linear regression models
svy: intreg Estimate interval regression models
svy: ivregress Estimate single-equation instrumental-variables regression models
svy: logit Estimate logistic regression models, reporting coefficients
svy: logistic Estimate logistic regression models, reporting odds ratios
svy: mlogit Estimate multinomial logistic regression models
svy: ologit Estimate ordered logistic regression models
svy: probit Estimate probit models
svy: oprobit Estimate ordered probit models
svy: poisson Estimate Poisson regression models
Post-Estimation
These commands are used after using `svy` estimation commands.

- `lincom` Estimate linear combinations of parameters
- `test` Perform hypotheses tests

Survival Analysis
- `svy: streg` Estimate parametric survival models
- `svy: stcox` Estimate cox proportional hazards models

Subpopulation Analysis
Using the `subpop` option allows you to look at subgroups in your data. For example, if you have an indicator variable, `female` (where 1=female and 0=male), you could examine the annual income (`income`) for females using the following command:

```
svy, subpop(female): mean income
```

For More Information and Assistance

Stata Tutorials
Stata’s website (http://www.stata.com) also has online resources such as FAQ, listserver, and online courses, etc.; see Stata users support at http://www.stata.com/support/. Finally, many Stata resources are available at other places on the Web. For a comprehensive list of such resources, see http://www.stata.com/links/resources1.html or do a general web search.

Stata Documentation and Books
This document provides only basic information about handling complex survey design in Stata. To learn more about features of this software package and to find out the syntax of specific statements, consult the Stata manuals available in the Velma Denning Room of the SSRC (Green Library Bing Wing Room 120F). To view the current collection, you can either stop by or consult our website (http://ssds.stanford.edu) for book lists organized by title, subject and call number. The collection is non-circulating, but additional copies of some books and manuals may be available in the Green Library reserves system.
Consulting

If you have questions about using Stata, please contact the software consultants at Social Science Data and Software. Our website is http://ssds.stanford.edu. The software consultants are available during the academic year on a walk-in basis. Please see our website for our current walk-in hours.

Note: This document is based on Stata 11.