StatNews #35

Software for Analyzing Complex Surveys
October 1999
Updated: October 2003

This newsletter briefly reviews the issues that must be addressed when analyzing data from a complex survey, and provides an update on the available software. Two previous newsletters (StatNews #11 and #12, October 1996) introduced the features of complex surveys and the special procedures required to analyze data from complex surveys.

The most basic type of survey uses a simple random sample, in which each person has an equal probability of being chosen. This type of survey does not require any special statistical methods, but it has a few drawbacks. First, it may not include enough subjects from particular subgroups of the population to allow for accurate estimates for these subgroups. "Oversampling", that is, including disproportionately more members of particular subgroups, can solve this problem, but then the estimate for the total population must be adjusted for the fact that some groups are over-represented in the sample. Performing a weighted analysis using weights that represent unequal probabilities of inclusion in the sample results in more accurate estimates.

Another drawback of simple random sampling is that it may not be practical to implement on a large scale. It is usually much cheaper to conduct a large survey in stages. First, the population is divided into large groups (strata), often on the basis of geography. Then, large clusters of subjects are selected from each stratum. Individual subjects are then selected from within the selected clusters. Because subjects are clustered, subjects within a cluster are more similar to each other than one would expect if they were chosen by simple random sampling. As a result, standard errors for estimates from the survey will usually be too small if the data are analyzed as if from a simple random sample. Using a special procedure that incorporates information about the clustering results in more accurate standard errors.

Most statistical procedures, including those in SAS, SPSS, and STATA, can perform a weighted analysis, although the standard errors will likely be incorrect. Special designed procedures should be used for analyzing complex samples. Such procedures are now available in these three software packages and statisticians in the Office of Statistical Consulting have experience with them. More information is available at http://www.stat.cornell.edu/statcom/. StatNews #35: Software for Analyzing Complex Surveys SAS version 8.0 has procedures that can select samples, compute descriptive statistics, and run regressions for complex sampling designs with stratification, clustering, and unequal weighting (SurveySelect, SurveyMeans, and SurveyReg. New in SAS 9.0 are two experimental procedures, SurveyFreq and SurveyLogistic. SurveyFreq can produce frequency and cross-tabulation tables for complex samples. SurveyLogistic fits logistic regression for discrete response survey data. More information on these procedures can be found at: http://www.sas.com/rnd/app/papers/papers_da.html and http://www2.sas.com/proceedings/sugi28/265-28.pdf. SPSS Complex Samples, an add on module available with SPSS 12.0, can do complex samples selection and planning, descriptive statistics (including t-tests) and cross-tabulate analyses for samples drawn by complex sampling methods. Currently, the Cornell SPSS license does not include the Complex Samples module.

STATA is an all-purpose statistical package that has a command-line interface. Much like SAS, Stata provides analytic tools in a programming environment, through a number of survey procedures (*svy estimators*). The survey procedures available in STATA are more comprehensive than those in SAS, including procedures such as probit and poisson regression, among others.

More information on this topic is available at http://www.ats.ucla.edu/stat/stata/topics/Survey.htm. Two stand-alone packages are available as well: SUDAAN and WesVarPC. SUDAAN uses a SAS-like programming language and can handle a wide variety of survey designs and analyses. SUDAAN can be used in conjunction with SAS, but is expensive to purchase. WesVarPC is a Windows-based program with a graphical user interface which can be downloaded free of charge at http://www.westat.com/wesvar/licensing/index.html. SPSS sells a special version of WesVarPC as an add-on module. WesVarPC uses replication methods that may not be straightforward to implement for some survey designs.

Contact the CSCU office if you need help analyzing complex survey data.

Authors: Cara Olsen and Karen Grace Martin

Note: the licensing link does not work. Contact westat for licensing information (www.westat.com)
April 21, 2012

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).