Chapter 18

Analysis of Data from Complex Surveys

Laura M. Stapleton
University of Maryland, Baltimore County

The goals of this chapter are to introduce the reader to the problems associated with using traditional statistical procedures with data from complex samples, provide simple strategies to more appropriately estimate statistics and their sampling variances under some common sampling designs, and provide additional resources that can be accessed for further reading.

The initial section of the chapter explains the effects that clustering and stratification in the sampling design can have on estimates of the sampling variance (or standard error). The effect of ignoring the sampling design in calculation of standard errors will be highlighted. The next sections describe several different methods to estimate appropriate statistics, illustrated with the use of a small example data set. These methods include two design effect adjustments, linearization through Taylor Series approximation, jackknife repeated replication, balanced repeated replication, and bootstrapping. The final section walks through the use of these methods via two different analyses using empirical data from the IEA Civic Education study.

The chapter ends with a summary of the software available for analyses with complex sample data and a list of recommended readings to learn more about issues in complex sample data analysis and variance estimation.

GLOSSARY OF KEY CONCEPTS

Bias. How far the average statistic lies from the parameter it is estimating. Random errors cancel each other out in the long run, those from bias will not. Bias can be classified into negative and positive bias. Negatively-biased estimates are estimates that tend to be smaller than the true parameters and positively-biased estimates are estimates that tend to be larger than the true parameters.

Cluster or Multistage Sampling. A sampling technique where the entire population is divided into groups, or clusters, and a random sample of these clusters are selected. When all observations in the selected clusters are included in the sample, the sample is called a cluster sample and when only a sample within the cluster is selected, the sample is called a multistage sample.

Design Effect. The inflation or deflation in the sampling variance of a statistic due to the sampling design.

Intraclass Correlation (ICC). The amount of variance in a response variable that can be attributed to a clustering effect.

Linearization. A method by which sampling variances (and standard errors) are estimated under complex sample designs. Also referred to as Taylor Series approximation, variance propagation, and the Delta method.

Replication Techniques. Methods by which sampling variances (and standard errors) are estimated under complex sample designs. With these methods, replicate samples are created from the original sample and the empirical variability of the statistics across the replicate samples is used to create a measure of the sampling variability for parameter estimates from the original sample. These methods include Jackknife Repeated Replication, Balanced Repeated Replication, and Bootstrapping.

Sampling Variance. The variability in the sample estimates of a population parameter if all possible samples (of the same size) were drawn from a given population. It is the square of the standard error.

Standard Error. The average distance any single sample estimate of a population parameter is expected to be from the true value. It is the standard deviation of the sample estimates of a population parameter, over all possible samples of the same size. It is the square root of the sampling variance.

Stratified Sampling. A stratified sample is obtained by taking samples from each stratum or sub-group of a population.

Variance Estimation. The process by which the sampling variance (or standard error) is estimated. Usually when using complex sample data sets, traditional estimates of sampling variance are found to be biased.