



Cornell University
Office of Statistical Consulting

Colleges of Human Ecology and Agriculture and Life Sciences
 Division of Nutritional Sciences

StatNews #12: Analysis of Complex Surveys

October 22, 1996

A complex survey samples observations disproportionately within strata and has multiple stages of sampling. This newsletter discusses the consequences of these features for the analysis of complex surveys, presents software options, and provides references.

The consequence of disproportionate sampling within strata is that estimates of means, population totals, regression coefficients, and other statistics should be made using the sampling weights if the estimates are to accurately reflect the population. Also, the variances (or standard errors) of those estimates should be calculated using the sampling weights. If the sampling weights cover a broad range of probabilities, then use of the sampling weights will result in larger variances of estimates than if a simple survey had been done. Software is needed that will allow the incorporation of sampling weights. Most statistical software such as SAS, SYSTAT, and SPSS can do this.

The consequence of multi-stage cluster sampling is that variances of estimates will typically be larger than in a simple survey. We usually assume that the cluster sampling does not affect estimates themselves, only their variances. The effect on variances of complex sampling is quantified as the design effect, which is the ratio of the variance under complex sampling divided by the variance that would have been obtained if a simple survey of the same size could have been done. Design effects can differ within the same survey markedly, depending upon the variable of interest, the sub-group of the population, and, in regression analyses, the variables in the regression model. For example, across the means of four anthropometric variables and six population subgroups in the second National Health and Nutrition Examination Survey, estimated design effects ranged from 1.0 to 5.7.

Two commonly used software programs for analyzing data from complex surveys are SUDAAN and WesVarPC. SUDAAN uses a linearization method, and WesVarPC offers a choice of resampling methods including balanced repeated replication and jackknife. Depending on the survey to be analyzed, one program may be more suitable than the other. The Office of Statistical Consulting can provide access to both programs, and WesVarPC may be installed free of charge on any computer running Windows. Contact Karen Grace-Martin at kg47@cornell.edu for more information.

A potential new approach to analyzing data from some complex surveys is to use software for multi-level modeling. Levels in the model are specified that correspond to the stages of sampling; this accounts for the cluster sampling. Sample weights must be used to account for unequal sampling probabilities. Both SAS PROC MIXED and MLN can incorporate sample weights in a multi-level analysis. However, the two programs use weights differently, PROC MIXED produces different results depending upon how the model is specified, variances in PROC MIXED will be severely inflated if the weights are not normalized properly, and both programs are documented poorly with regard to the use of weights. Therefore, we recommend that you do not use this approach until it is better understood, and advise extreme caution if you do decide to use it. Cara Olsen has developed a short document detailing some of these issues.

References (given in order of increasing technical complexity)

1. Lee ES, Forthofer RN, Lorimor RJ. Analyzing Complex Survey Data. Newbury Park, CA: Sage University Paper #71, 1989.
2. Lehtonen R, Pahkinen EJ. Practical Methods for Design and Analysis of Complex Surveys. New York: John Wiley and Sons, 1995.
3. Skinner CJ, Holt D, Smith TMF (eds). Analysis of Complex Surveys. New York: John Wiley and Sons, 1989.

Author: Edward A. Frongillo, Jr.

This newsletter was distributed to faculty and graduate students in the Division of Nutritional Sciences and the College of Human Ecology, and faculty in the College of Agriculture and Life Sciences, by the Office of Statistical Consulting. Please forward it to any interested colleagues and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should e-mail statcons@cornell.edu.

Information about the Office of Statistical Consulting can be obtained at World Wide Web address

<http://www.human.cornell.edu/admin/statcons/>.