Analyze Complex Survey Data Using SAS

Anthony An and Donna Watts
SAS Institute Inc., SAS Campus Drive, Cary, NC 27513, USA
Anthony.An@sas.com, Donna.Watts@sas.com

1. Overview

Researchers often use sample survey methodology to obtain information about a large population by selecting and measuring a sample from that population. By applying scientific probability-based designs to select the sample, researchers reduces the risk of a distorted view of the population and allow statistically valid inferences to be made from the sample (Cochran (1977) and Kish (1965)). The SURVEYSELECT procedure can be used to select probability-based samples from a study population.

When a sample is selected from a finite population using a complex sample design, one should calculate the estimates and their variances incorporating the sample design. The SURVEYMEANS and SURVEYREG procedures properly analyze survey data, taking into account the sample design. These procedures use the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs.

2. Survey Sampling

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or a sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of selection. The SURVEYSELECT procedure provides methods for both equal probability sampling and probability proportional to size (PPS) sampling. Systematic and sequential sampling are methods are also available in PROC SURVEYSELECT.

3. Survey Data Analysis

The SURVEYMEANS and SURVEYREG procedures perform statistical analysis for survey data. These analytical procedures take into account the design used to select the sample. The sample design can be a complex sample design with stratification, clustering, and unequal weighting. To analyze survey data with these procedures, you need to specify sample design information, for example, design strata, clusters, and sampling weights.

You can use the SURVEYMEANS procedure to compute the population mean, total and proportion estimates, domain mean, total and proportion estimates, and estimates of ratios. The standard errors, confident limits, and corresponding t-tests will also available in the procedure.

PROC SURVEYREG fits linear models for survey data and computes regression coefficients and their variance-covariance matrix. The procedure also provides significance tests for the model effects and for any specified estimable linear functions of the model parameters.

4. Variance Estimation

The SURVEYMEANS and SURVEYREG procedures use the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or primary sampling units (PSUs), in the sample design, the procedures estimate the variance from the variation among the PSUs. When the design is stratified, the procedures pool stratum variance estimates to compute the overall variance estimate.

For more information about these survey anlysis SAS procedures, please visit the website http://www.sas.com/statistics

REFERENCES

Cochran, W. G. (1977). *Sampling Techniques, Third Edition, New York: John Wiley & Sons, Inc.*

Fuller, W. A. (1975). Regression Analysis for Sample Survey. *Sankhyā*, **37 (3), Series C**, 117-132.

Kish, L. (1965). *Survey Sampling. New York: John Wiley & Sons, Inc.*

Woodruff, R. S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association*, **66**, 411-414.

RESUME

Pour obtenir des informations sur une population finie, les représentations d'échantillonnage complexes basées sur la probabilité sont souvent utilisées dans de nombreuses disciplines scientifiques pour sélectionner les échantillons. Pour tirer des conclusions statistiquement valides sur la population étudiée, la représentation d'échantillonnage doit être prise en compte dans l'analyse des données du sondage.

Dans ce document, il est question de la méthode d'estimation de la variance basée sur la représentation qui est utilisée dans SAS pour l'analyse des données du sondage. Nous discuterons également la possibilité de sélectionner un échantillon de probabilité en utilisant des représentations variées grâce à la PROC SURVEYSELECT, la possibilité de calculer des statistiques simples pour les données du sondage grâce à la PROC SURVEYMEANS et la possibilité d'appliquer des modèles de régression linéaire aux données grâce à la PROC SURVEYREG.