Chapter 11

Introduction to Survey Sampling and Analysis Procedures

Chapter Table of Contents

Chapter 11
Introduction to Survey Sampling and Analysis Procedures

Overview

This chapter introduces the SAS/STAT procedures for survey sampling and describes how you can use these procedures to analyze survey data.

Researchers often use sample survey methodology to obtain information about a large population by selecting and measuring a sample from that population. Due to variability among items, researchers apply scientific probability-based designs to select the sample. This reduces the risk of a distorted view of the population and allows statistically valid inferences to be made from the sample. Refer to Cochran (1977), Kalton (1983), and Kish (1965) for more information on statistical sampling. You can use the SURVEYSELECT procedure to select probability-based samples from a study population.

Many SAS/STAT procedures, such as the MEANS and GLM procedures, can compute sample means and estimate regression relationships. However, in most of these procedures, statistical inference is based on the assumption that the sample is drawn from an infinite population by simple random sampling. If the sample is actually selected from a finite population using a complex design, these procedures generally do not calculate the estimates and their variances correctly. The SURVEYMEANS and SURVEYREG procedures do 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.

The following table briefly describes the sampling and analysis procedures in SAS/STAT software.

SURVEYSELECT	
Design Accommodated	stratification clustering replication multistage sampling unequal probabilities of selection

Sampling Methods	simple random sampling unrestricted random sampling (with replacement) systematic sequential selection probability proportional to size (PPS) with and without replacement PPS systematic PPS for two units per stratum sequential PPS with minimum replacement
SURVEYMEANS	
Design Accommodated	stratification clustering unequal weighting
Available Statistics	population total population mean proportion standard error confident limit *t* test
SURVEYREG	
Design Accommodated	stratification clustering unequal weighting
Available Analysis	fit linear regression model regression coefficients covariance matrix significance tests estimable functions contrasts

The following sections contain brief descriptions of these procedures.

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. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population.

PROC SURVEYSELECT provides methods for both equal probability sampling and sampling with probability proportional to size (PPS). In PPS sampling, a unit's selection probability is proportional to its size measure. PPS sampling is often used in cluster sampling, where you select clusters (groups of sampling units) of varying size in the first stage of selection. Available PPS methods include without replacement, with replacement, systematic, and sequential with minimum replacement. The procedure can apply these methods for stratified and replicated sample designs. See Chapter 63, "The SURVEYSELECT Procedure," for more information.

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.

You can use the SURVEYMEANS procedure to compute the following statistics:

- population total estimate and its standard deviation and corresponding *t* test
- population mean estimate and its standard error and corresponding *t* test
- proportion estimate for a categorical variable and corresponding *t* test
- $(1 - \alpha)\%$ confidence limits for the population total estimates, the population mean estimates, and the proportion estimates
- data summary information

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.

PROC SURVEYMEANS presently does not perform domain analysis (subgroup analysis). However, note that you can produce a domain analysis with PROC SURVEYREG (see Example 62.7 on page 3269). This capability will be available in a future release of the SURVEYMEANS procedure.

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 a multistage sample design, the variance estimation method depends only on the first stage of the sample design. Thus, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input

design information about any additional stages of sampling. This variance estimation method assumes that the first-stage sampling fraction is small or that the first-stage sample is drawn with replacement, as it often is in practice.

For more information on variance estimation for sample survey data, refer to Lee, Forthoffer, and Lorimor (1989), Cochran (1977), Kish (1965), Särndal, Swenson, and Wretman (1992), Wolter (1985), and Hansen, Hurwitz, and Madow (1953).

In addition to the traditional Taylor expansion method, other methods for variance estimation for survey data include balanced repeated replication and jackknife repeated replication. These methods usually give similar, satisfactory results (Wolter 1985, Särndal, Swenson, and Wretman 1992); the SURVEYMEANS and SURVEYREG procedures currently provide only the Taylor expansion method.

See Chapter 61, "The SURVEYMEANS Procedure," and Chapter 62, "The SURVEYREG Procedure," for complete details.

Design Information for Survey Procedures

Survey sampling is the process of selecting a probability-based sample from a finite population according to a sample design. You then collect data from these selected units and use them to estimate characteristics of the entire population.

A *sample design* encompasses the rules and operations by which you select sampling units from the population and the computation of sample statistics, which are estimates of the population values of interest. The objective of your survey often determines appropriate sample designs and valid data collection methodology. A complex sample design often includes stratification, clustering, multiple stages of selection, and unequal weighting.

For more detailed information, refer to Cochran (1977), Kalton (1983), Kish (1965), and Hansen, Hurwitz, and Madow (1953).

To select a sample with the SURVEYSELECT procedure and analyze your survey data with the SURVEYMEANS and SURVEYREG procedures, you need to specify sample design information to those procedures. This information includes design strata, clusters, and sampling weights.

Population

Population refers to the target population or group of individuals of interest for study. Often, the primary objective is to estimate certain characteristics of this population, called *population values*. A *sampling unit* is an element or an individual in the target population. A sample is a subset of the population that is selected for the study.

Before you use the survey procedures, you should have a well-defined target population, sampling units, and an appropriate sample design.

In order to select a sample according to your sample design, you need to have a list of sampling units in the population. This is called a *sampling frame*. PROC SURVEYSELECT selects a sample using this sampling frame.

Stratification

Stratified sampling involves selecting samples independently within strata, which are nonoverlapping subgroups of the survey population. Stratification controls the distribution of the sample size in the strata. It is widely used in practice to meet a variety of survey objectives. For example, with stratification you can ensure adequate sample sizes for subgroups of interest, including small subgroups, or you can use stratification to improve the precision of overall estimates. To improve precision, units within strata should be as homogeneous as possible for the characteristics of interest.

Clustering

Cluster sampling involves selecting clusters, which are groups of sampling units. For example, clusters may be schools, hospitals, or geographical areas, and sampling units may be students, patients, or citizens. Cluster sampling can provide efficiency in frame construction and other survey operations. However, it can also result in a loss in precision of your estimates, compared to a nonclustered sample of the same size. To minimize this effect, units within clusters should be as heterogeneous as possible for the characteristics of interest.

Multistage Sampling

In *multistage sampling*, you select an initial or first-stage sample based on groups of elements in the population, called *primary sampling units* or *PSUs*.

Then you create a second-stage sample by drawing a subsample from each selected PSU in the first-stage sample. By repeating this operation, you can select a higher-stage sample.

If you include all the elements from a selected primary sampling unit, then the two-stage sampling is a cluster sampling.

Sampling Weights

Sampling weights, or *survey weights*, are positive values associated with each unit in your sample. Ideally, the weight of a sampling unit should be the "frequency" that the sampling unit represents in the target population. Therefore, the sum of the weights over the sample should estimate the population size N. If you normalize the weights such that the sum of the weights over the sample equals the population size N, then the weighted sum of a characteristic y estimates the population total value Y.

Often, sampling weights are the reciprocals of the selection probabilities for the sampling units. When you use PROC SURVEYSELECT, the procedure generates the sampling weight component for each stage of the design, and you can multiply these sampling weight components to obtain the final sampling weights. Sometimes, sampling weights also include nonresponse adjustments, post-sampling stratification, or regression adjustments using supplemental information.

When the sampling units have unequal weights, you must provide the weights to the survey analysis procedures. If you do not specify sampling weights, the procedures use equal weights in the analysis.

Population Totals and Sampling Rates

The ratio of the sample size (the number of sampling units in the sample) n and the population size (the total number of sampling units in the target population) N is written as

$$f = \frac{n}{N}$$

This ratio is called the *sampling rate* or the *sampling fraction*. If you select a sample without replacement, the extra efficiency compared to selecting a sample with replacement can be measured by the *finite population correction* (fpc) factor, $(1 - f)$.

If your analysis should include a finite population correction factor, you can input either the sampling rate or the population total. Otherwise, the procedures do not use the fpc when computing variance estimates. For fairly small sampling fractions, it is appropriate to ignore this correction. Refer to Cochran (1977) and Kish (1965).

As stated in the section "Variance Estimation" on page 151, for a multistage sample design, the variance estimation method depends only on the first stage of the sample design. Therefore, if you are specifying the sampling rate, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the target population.

An Example of Using the Survey Procedures

This section demonstrates how you can use the survey procedures to select a probability-based sample, compute descriptive statistics from the sample, perform regression analysis, and make inferences about income and expenditures of a group of households in North Carolina and South Carolina. The goals of the survey are to

- estimate total income and total basic living expenses
- investigate the linear relationship between income and living expenses

Sample Selection

To select a sample with PROC SURVEYSELECT, you input a SAS data set that contains the sampling frame, or list of units from which the sample is to be selected. You also specify the selection method, the desired sample size or sampling rate, and other selection parameters.

In this example, the sample design is a stratified simple random sampling design, with households as the sampling units. The sampling frame (the list of the group of the households) is stratified by State and Region. Within strata, households are selected by simple random sampling. Using this design, the following PROC SURVEYSELECT statements select a probability sample of households from the HHSample data set.

```
proc surveyselect data=HHSample out=Sample
                  method=srs n=(3, 5, 3, 6, 2);
   strata State Region;
run;
```

The STRATA statement names the stratification variables State and Region. In the PROC SURVEYSELECT statement, the DATA= option names the SAS data set HHSample as the input data set (the sampling frame) from which to select the sample. The OUT= option stores the sample in the SAS data set named Sample. The METHOD=SRS option specifies simple random sampling as the sample selection method. The N= option specifies the stratum sample sizes.

The SURVEYSELECT procedure then selects a stratified random sample of households and produces the output data set Sample, which contains the selected households together with their selection probabilities and sampling weights. The data set Sample also contains the sampling unit identification variable Id and the stratification variables State and Region from the data set HHSample.

Survey Data Analysis

You can use the SURVEYMEANS and SURVEYREG procedures to estimate population values and to perform regression analyses for survey data. The following example briefly shows the capabilities of each procedure. See Chapter 61, "The SURVEYMEANS Procedure," and Chapter 62, "The SURVEYREG Procedure," for more detailed information.

To estimate the total income and expenditure in the population from the sample, you specify the input data set containing the sample, the statistics to be computed, the variables to be analyzed, and any stratification variables. The statements to compute the descriptive statistics are as follows:

```
proc surveymeans data=Sample sum clm;
   var Income Expense;
   strata State Region;
   weight Weight;
run;
```

The PROC SURVEYMEANS statement invokes the procedure, specifies the input data set, and requests estimates of population totals and their standard deviations for the analysis variables (SUM), and confidence limits for the estimates (CLM).

The VAR statement specifies the two analysis variables, Income and Expense. The STRATA statement identifies State and Region as the stratification variables in the sample design. The WEIGHT statement specifies the sampling weight variable Weight.

You can also use the SURVEYREG procedure to perform regression analysis for sample survey data. Suppose that, in order to explore the relationship between the total income and the total basic living expenses of a household in the survey population, you choose the following linear model to describe the relationship.

$$\text{Expense} = \alpha + \beta * \text{Income} + \text{error}$$

The following statements fit this linear model.

```
proc surveyreg data=Sample;
   strata State Region ;
   model  Expense = Income;
   weight Weight;
run;
```

In the PROC SURVEYREG statement, the DATA= option specifies the input sample survey data as Sample. The STRATA statement identifies the stratification variables as State and Region . The MODEL statement specifies the model, with Expense as the dependent variable and Income as the independent variable. The WEIGHT statement specifies the sampling weight variable Weight.

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.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory*, Volumes I and II, New York: John Wiley & Sons, Inc.

Kalton, G. (1983), *Introduction to Survey Sampling*, SAGE University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills and London: SAGE Publications, Inc.

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

Lee, E. S., Forthoffer, R. N., and Lorimor, R. J. (1989), *Analyzing Complex Survey Data*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-071, Beverly Hills and London: Sage Publications, Inc.

Särndal, C.E., Swenson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag Inc.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag 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.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

Copyright © 1999 by SAS Institute Inc., Cary, NC, USA.

ISBN 1–58025–494–2

All rights reserved. Produced in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of the software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

The Institute is a private company devoted to the support and further development of its software and related services.