Providing software solutions since 1976

Sample Survey Design and Analysis

Overview

Researchers often use sample survey methodology to obtain information about a large aggregate or population by selecting and measuring a sample from that population. Due to the variability of characteristics among items in the population, researchers apply scientific sample designs in the sample selection process to reduce the risk of a distorted view of the population, and they make inferences about the population based on the information from the sample survey data. In order to make statistically valid inferences for the population, they must incorporate the sample design in the data analysis.

Traditional SAS® procedures, such as the MEANS procedure and the GLM procedure, compute statistics under the assumption that the sample is drawn from an infinite population by simple random sampling. These procedures generally do not correctly estimate the variance of an estimator if they are applied to a sample drawn by a complex sample design. SAS users have requested procedures that analyze data from complex sample surveys. In response to this request, SAS/STAT® software now provides the SURVEYFREQ, SURVEYLOGISTIC, SURVEYMEANS, SURVEYPHREG, SURVEYREG, and SURVEYSELECT procedures.

To select probability-based random samples from a study population, you can use the SURVEYSELECT procedure, which provides a variety of methods for probability sampling. To analyze sample survey data, you can use the SURVEYFREQ, SURVEYLOGISTIC, SURVEYMEANS, SURVEYPHREG, and SURVEYREG procedures, which incorporate the sample design into the analyses. These procedures can be used for multistage designs or for single-stage designs, with or without stratification, and with or without unequal weighting.

Procedure	Features
PROC SURVEYFREQ	• estimates of population means and totals • estimates of population proportions • standard errors • confidence limits • hypothesis tests (t tests) • domain analysis • ratio estimates
PROC SURVEYLOGISTIC	• cumulative logit regression model fitting • logit, complementary log-log and probit link functions • generalized logit regression model fitting • estimates of regression coefficients • estimates of covariance matrices • hypothesis tests • model diagnostics • estimates of odds ratios • confidence limits • estimable functions • estimates and standard errors for contrasts • domain analysis
PROC SURVEYMEANS	• estimates of population means and totals • estimates of population proportions • standard errors • confidence limits • hypothesis tests (t tests) • domain analysis • ratio estimates
PROC SURVEYPHREG	• regression analysis based on the Cox proportional hazards model • hazard ratio estimates • predicted values and their standard errors • martingale, Schoenfeld, score, and deviance residuals • significance tests • confidence limits • estimable functions • domain analysis
PROC SURVEYREG	• linear regression model fitting • estimates of regression coefficients • estimates of covariance matrices • significance tests • confidence limits • estimable functions • estimates and standard errors for contrasts • domain analysis
PROC SURVEYSELECT	• simple random sampling • unrestricted random sampling (with replacement) • systematic sampling • sequential sampling • selection probability proportional to size (PPS) with and without replacement • PPS systematic sampling • PPS for two units per stratum • sequential PPS with minimum replacement

The SURVEYFREQ Procedure

The SURVEYFREQ procedure produces one-way to n-way frequency and crosstabulation tables from sample survey data. These tables include estimates of population totals, population proportions (overall proportions, and also row and column proportions), and corresponding standard errors. Confidence limits, coefficients of variation, and design effects are also available. The procedure provides a variety of options to customize your table display.

For one-way frequency tables, PROC SURVEYFREQ provides Rao-Scott chi-square goodness-of-fit tests, which are adjusted for the sample design. You can test a null hypothesis of equal proportions for a one-way frequency table, or you can input other null hypothesis proportions for the test. For two-way frequency tables, PROC SURVEYFREQ provides design-based tests of no association between the row and column variables. These tests include the Rao-Scott chi-square test, the Rao-Scott likelihood-ratio test, the Wald chi-square test, and the Wald log-linear chi-square test.

The following statements illustrate the syntax of PROC SURVEYFREQ:

```
proc surveyfreq data=SIS_Survey;
   tables Response;
   strata  State  NewUser;
   cluster School;
   weight SamplingWeight;
run;
```

The SURVEYLOGISTIC Procedure

The SURVEYLOGISTIC procedure fits linear logistic regression models for discrete response survey data by the method of pseudo-maximum likelihood, incorporating the sample design into the analysis. The SURVEYLOGISTIC procedure enables you to use categorical classification variables as explanatory variables, using the familiar syntax for main effects and interactions employed in the GLM and LOGISTIC procedures. The following link functions are available in PROC SURVEYLOGISTIC: the cumulative logit function (CLOGIT), the generalized logit function (GLOGIT), the probit function (PROBIT), and the complementary log-log function (CLOGLOG). Design-based variances of the estimated regression parameters and the odds ratios are estimated using a Taylor series expansion approximation.

The following statements illustrate the syntax of PROC SURVEYLOGISTIC:

```
proc surveylogistic data=SampleStrata;
   strata state type/list;
   model Rating (order=internal) = Usage;
   weight SamplingWeight;
run;
```

The SURVEYMEANS Procedure

The SURVEYMEANS procedure provides estimates of population means and population totals from sample survey data. The sample design can be a complex sample design with stratification, clustering, and unequal weighting. PROC SURVEYMEANS also provides domain analysis (subgroup or subpopulation analysis). The procedure uses the Taylor series expansion method to provide estimates of design-based variances for the quantities of interest.

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

- estimates of population means, with corresponding standard errors and *t* tests
- estimates of population totals, with corresponding estimated standard deviations and *t* tests
- estimates of proportions for categorical variables and corresponding *t* tests
- ratio estimates population means and proportions, and their standard errors
- confidence limits for estimated population total, population mean, and proportions
- data summary information
- estimates of domain means, domain totals, and domain proportions and their standard errors

The following statements illustrate the syntax of PROC SURVEYMEANS:

```
proc surveymeans data=Company mean sum;
   var Asset Sale Value Profit;
   weight Weight;
run;
```

The SURVEYPHREG Procedure

PROC SURVEYPHREG performs regression analysis based on the Cox proportional hazards model for sample survey data. The procedure provides design-based variance estimates, confidence intervals, and hypothesis tests concerning the parameters and model effects.

The following statements illustrate the syntax of PROC SURVEYPHREG:

```
proc surveyphreg data = LibrarySurvey;
     weight SamplingWeight;
     strata Branch;
     model lenBorrow*Returned(0) = Age;
 run;
```

The SURVEYREG Procedure

PROC SURVEYREG fits linear regression models for survey data and computes regression coefficients and their variance-covariance matrix. The procedure enables you to specify classification effects with the same syntax as in the GLM procedure. The procedure also provides hypothesis tests for the model effects, for any specified estimable linear functions of model parameters, and for custom hypothesis tests of linear combinations of the regression parameters. The procedure computes design-based confidence limits of the parameter estimates and their linear estimable functions.

The following statements illustrate the syntax of PROC SURVEYREG:

```
proc surveyreg data=Farms total=TotalInStrata;
   strata State Region / list;
   model CornYield = FarmArea/ covb;
   weight Weight;
run;
```

The SURVEYSELECT Procedure

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or can 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 survey population.

In addition to methods for equal probability selection, PROC SURVEYSELECT also provides probability proportional to size (PPS) selection methods. 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 sizes 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.

To select a sample with PROC SURVEYSELECT, you input a SAS data set that contains the sampling frame or list the units from which the sample will be selected. You also specify the selection methods, the desired sample size or sampling rate, and other parameters. The SURVEYSELECT procedure selects the sample, producing an output data set that contains the selected units, their selection probabilities, and sampling weights.

The SURVEYSELECT procedure uses fast, efficient algorithms for these sample selection methods. Thus, it performs well even for very large input data sets or sampling frames, which can occur in practice for large scale sample surveys.

The following statements illustrate the syntax of PROC SURVEYSELECT:

```
proc surveyselect data=Customers
   method=srs n=15
   seed=1953 out=SampleStrata;
   strata State Type;
run;
```

A Note on Variance Estimation

The survey analysis procedures provide a choice of variance estimation methods for complex survey designs. In addition to the Taylor series linearization method, the procedures offer two replication-based (resampling) methods—balanced repeated replication (BRR) and the delete-1 jackknife. These variance estimation methods usually give similar, satisfactory results. The choice of a variance estimation method can depend on the sample design used, the sample design information available, the parameters to be estimated, and computational issues.

The Taylor series linearization method is appropriate for all designs in which the first-stage sample is selected with replacement, or in which the first-stage sampling fraction is small, as it often is in practice. The Taylor series 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. When there are clusters (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 Taylor series method uses only the first stage of the sample design. Therefore, 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.

Replication methods for variance estimation draw multiple replicates (or subsamples) from the full sample by following a specific resampling scheme. Commonly used resampling schemes include balanced repeated replication (BRR) and the jackknife. The parameter of interest is estimated from each replicate, and the variability among the replicate estimates is used to estimate the overall variance of the parameter estimate.

The BRR variance estimation method requires a stratified sample design with two PSUs in each stratum. Each replicate is obtained by deleting one PSU per stratum according to the corresponding Hadamard matrix and adjusting the original weights for the remaining PSUs. The adjusted weights are called replicate weights. The survey procedures also provide Fay’s method, which is a modification of the BRR method.

The jackknife method deletes one PSU at a time from the full sample to create replicates, and modifies the original weights to obtain replicate weights. The total number of replicates equals the number of PSUs. If the sample design is stratified, each stratum must contain at least two PSUs, and the jackknife is applied separately within each stratum.

Instead of having the survey procedures generate replicate weights for the analysis, you can directly input your own replicate weights. This can be useful if you need to do multiple analyses with the same set of replicate weights or if you have access to replicate weights without complete design information.

Documentation

For more information, see the chapters "Introduction to Survey Procedures," "The SURVEYFREQ Procedure," "The SURVEYLOGISTIC Procedure," "The SURVEYMEANS Procedure," "The SURVEYPHREG Procedure," "The SURVEYREG Procedure," and "The SURVEYSELECT Procedure" in the SAS/STAT User's Guide.

Statistics and Operations Research Home Page