RECENT DEVELOPMENTS AND FUTURE PLANS FOR SUDAAN

Babubhai V. Shah and Beth G. Barnwell, Research Triangle Institute
P.O. Box 12194, Research Triangle Park, NC 27709

Introduction

RTI has a long history of developing and maintaining state of the art software for the design based analysis of survey data. In the 1970's, RTI first developed its STDERR procedure. This first general use package estimated means, totals, proportions and their design based variance estimates for user specified subgroups for a variety of stratified, unequally weighted, multi-stage sample designs. This was soon followed by the SURREGR procedure, which added the estimation of linear regression models. During the 1980's, additional data analysis procedures (RTIFREQS, SESUDAAN, RATIOEST, and RTILOGIT) were added to include cross-tabulations, standardized subgroup comparisons, estimation of general ratios ($\Sigma y_i/\Sigma x_i$), and linear logistic regression models. All of the above procedures were implemented as supplemental SAS (copyright SAS Institute Inc.) version 5 procedures running on IBM mainframe computers under MVS.

In recent years with support from the Public Health Service (PHS) to develop a comprehensive software package that meets the needs of statistical analysts at the National Center for Health Statistics (NCHS) and PHS, we embarked on the task of developing a system that incorporates many of the features of RTI's existing survey data analysis system but also includes significant enhancements.

The objective was to develop survey data analysis software that is flexible enough to meet the needs of survey statisticians, as well as research scientists, not necessarily schooled in survey sampling. As a result, the package contains several procedures for estimating descriptive statistics as well as model parameters. When plans were first being laid for what SUDAAN would do and how it would do it, many excellent data management systems were already available. Therefore, we decided that SUDAAN would focus solely on statistical computations leading to estimates and their appropriate standard errors, starting with input data that were already in a "standard" form for SUDAAN. The decision was made to limit development to survey data analysis not offered by other statistical packages and not duplicate capabilities of these packages. The design of these procedures includes many features to make it convenient for the users, while staying within the scope defined for SUDAAN.

The objective of the paper is to briefly describe these features, that are available currently and those that are planned for the future.

Current Features

To make SUDAAN useful we have introduced several procedures, all of which permit a variety of sample designs. For all of these procedures we have implemented a number of user-friendly features including syntax compatibility across procedures and operating systems, substantial user control over printed output, limited recoding and variable subsetting capabilities, and consistent input and output file specifications across procedures and operating systems. We will describe each of these features in detail.

SUDAAN Procedures

Currently, SUDAAN offers three procedures for descriptive statistics and four procedures for statistical modeling. Together they provide the user with a powerful variety of analytical tools. The numbers of observations, of variables per observation, of tables, of variables per table, and of independent variables per model are limited only by your computer's storage.

The three descriptive procedures are:

CROSSTAB computes weighted frequencies, percentage distributions, and their standard errors for user-specified cross-tabulations

RATIO computes estimates and standard errors of generalized ratios of the form Swy/Swx, where x and y are observed variables and w is the analysis weight

DESCRIPT computes estimates of means, totals, proportions, geometric means, quantiles, and their standard errors.

In addition, CROSSTAB produces both chi-square tests of independence for two-way contingency tables and the covariance matrix of estimates within a table. Both RATIO and DESCRIPT produce:

- standardized estimates
- post-stratified estimates
- contrasts among domain estimates

SUDAAN offers four statistical modeling procedures:

REGRESS computes linear regression models and hypothesis tests concerning the model parameters

LOGISTIC computes logistic regression models for binary outcomes and hypothesis tests.

CATAN is a general categorical data analysis procedure for log-linear modeling of a contingency table

SURVIVAL computes discrete and continuous proportional hazards models for survival time data.

All procedures except CATAN allow continuous and categorical independent variables. CATAN is limited to categorical variables. They supply estimates of the coefficients and analysis of variance table tests of multiple degree of freedom hypotheses. They may also be used to specify and test general linear contrasts of the coefficients. Finally, coefficient estimates and their covariance matrix may be saved for additional analysis.

Variance Estimation

Variances and covariances in SUDAAN are computed according to the sample design specified by the user. Options are provided for three basic types of sample designs. These are:

- With replacement sampling at the first stage

- Equal probability sampling without replacement at the first stage, and with or without replacement at subsequent stages

- Without replacement unequal probability sampling at the first stage and probability sampling, with or without replacement, at subsequent stages

These three broad design groups allow the user to specify a wide variety of sample designs often found in sample surveys.

Variance estimation for the nonlinear statistics computed in the SUDAAN series of procedures for survey data analysis is based on a first-order Taylor series approximation of the deviations of estimates from their expected values. This approximation for large samples is well-known (see Kendall and Stuart, 1973). Woodruff (1971) presented applications of this technique to sample surveys. Asymptotic results for Taylor series linearization have been derived by Binder (1983) and Godambe and Thompson (1985). These calculations are fully described in a technical report titled, "Statistical Methods and Mathematical Algorithms Used in SUDAAN" (Shah et al, 1993).

User-friendly Features

We have incorporated several features in SUDAAN to provide users with considerable control over input and output and to provide other features that make it convenient to specify user requirements.

Design statements

The SUDAAN design statements are the same for all procedures. This means that they only have to be set up once for a given data set, then they may be used in any SUDAAN procedure.

Syntax compatibility

SUDAAN, which uses a free-format SAS-like syntax, maintains uniform statement syntax and semantics across all procedures. Statements and options with the same names in multiple procedures behave the same way in all the procedures. Furthermore the SUDAAN syntax is the same for all the different operating systems to which we have ported SUDAAN. Currently, supported operating systems include

- PC-DOS
- VAX/VMS
- VAX/ULTRIX
- SUN-OS
- MVS

We are working toward adding CMS for the IBM mainframe and OS2 for the IBM PC in the future. We intend to maintain this compatibility as we develop new procedures and port to new environments in the future. This makes it very easy for a user familiar with one procedure or operating system to use additional procedures, or to use SUDAAN in a different environment.

Control Over Printed Output

SUDAAN gives users substantial control over the appearance of printed table and output data sets. We offer two different table styles. The BOX style has a single row and column dimension with separate subtables for each combination of levels of any additional dimensions. All the statistics are collected in a box for each table cell. The NCHS style allows multiple dimensions to be nested for rows. If there is a single statistic, multiple column dimensions may also be nested. Otherwise there is a column for each of the statistics.

For either style the user may specify widths, number of decimal places, units, and scientific or standard format for each statistic. In addition the user may specify any number of titles and footnotes to document tables. There are also a number of table features uses can control. These include line size, page length, column widths, top and left margins, widths for label columns, nesting for rows and columns, spaces between columns, spaces between rows, and spaces for indenting. All of these items have reasonable default values, so that users are not required to specify a great deal of information in order to obtain printed results.

The user also has substantial control over the choice and order of the statistics to be printed together in a table. Each statistic is associated with a keyword to which the user may refer in PRINT and OUTPUT statements. No explicit keyword lists are required, however, since SUDAAN has built-in lists of default statistics to print for each procedure.

RECODE and SUBPOPN Statements

With Release 6.0 SUDAAN has added limited facilities for recoding or subsetting data on the fly. The RECODE statement allows for recoding of one or more variables to consecutive values 0, 1, ...n.. The SUBPOPN statement allows users to define a subpopulation of the data for analysis by specifying a logical expression composed of variable names, numeric literals, and logical and relational operators.

SUDAAN Files

SUDAAN currently accepts text input files on all operating systems. This means that the same data files can be analyzed by SUDAAN on many different types of computers without any changes at all. SUDAAN also creates text output data sets in all the supported environments. These data sets are identical for all systems, and can easily be imported into SAS or other packages for further processing, printing or graphing. In addition SUDAAN can read SAS data sets on some systems. These include SAS Version 5 data sets on the VAX under VMS and on the IBM mainframe under MVS and CMS, and SAS Version 6 data sets on the IBM-PC under DOS. We are planning for a new binary data file structure which will be available on all operating systems. We will provide utility programs to convert between this format and the other supported formats. This file type will have the advantages of maintaining complete numeric accuracy as well as requiring less storage space.

Improvements Already Underway

We have already begun a number of improvements to SUDAAN to make it more convenient to use. Among the most important are a new version of SUDAAN which will run under SAS on some systems, and a menu-driven front-end to build the file of SUDAAN input statements.

SUDAAN Running Under SAS

We are currently developing a version of SUDAAN that will run as a set of SAS procedures on the VAX under VMS and the IBM mainframe under MVS. This version will be able to read and write SAS data sets of any type. Users will be able to integrate their SUDAAN jobs with SAS, taking full advantage of the data management capabilities available in that system. SAS has recently announced the availability of the TOOLKIT for development of SAS procedures on additional platforms. Eventually we may be able to offer SUDAAN as a set of SAS procedures on some of these additional platforms.

Menu-Driven Front-End

We are developing a menu-driven program to build the file of SUDAAN input statements. For users familiar with the software, this will be a quick and easy way to develop SUDAAN input files without having to refer to the manual. We plan to include at least some on-line help in this program initially. As we get user feedback we will improve the program and its help facility.

Plans for the Future

The current procedures are in use at several sites. We have received many comments from the users at these sites. We plan to make improvements based on the suggestions received so far. The suggestions can be divided into three categories as follows:

- General improvements
- Additional statistics in existing procedures.
- New procedures

We shall discuss each one in turn, the list presented here is indicative of the types of suggestions received and being considered. The list not exhaustive and no time table for their implementation is implied or assured.

General Improvements

Documentation

Documentation of our software is not as easy to use as we would like. We provide complete technical details on the formulae used in each of the procedures (see the technical report by Shah et al (1993)). The current user's manual is primarily a reference manual and we are planning to add a new simpler tutorial manual that will teach new users the first steps in using SUDAAN software.

New Statement

A REPARM statement for reparametrizing parameters related to categorical variables on the MODEL statement may be added in the future. This statement will permit users to reparametrize the coefficients corresponding to levels of a categorical variable. The current method defaults to treating the last level as the reference level. The user will be able to specify any other level as a reference level. Alternately, the user may use the unweighted (weighted) mean over all levels as reference, or specify orthogonal polynomial type (linear, quadratic, cubic) functions of levels.

Additional Statistics in Existing Software

The following is the suggested list of additional statistics for the existing procedures:

- In CROSSTAB: Mantel-Haenszal statistics, odds ratio, stratified odds ratio, measure of relative risk, etc.

- In DESCRIPT: T-statistic and P-value for the contrast or differences

- In REGRESS and LOGISTIC equivalent (psuedo) likelihood and equivalent large sample Chi-square for goodness of to allow one to compare different models fitted to the same data.

- Computation and output of variance covariance matrix for all the cell estimates in a single table produced by CROSSTAB, DESCRIPT, and RATIO

- Relative risk and odds ratio for the LOGISTIC model

Additional Procedures

We have received many suggestions for adding procedures to SUDAAN. We list a few serious candidates:

- Procedure to fit logistic model to polychotomus data.

- Procedure to fit generalized linear model

- Procedure to compute variance components

- Procedure to account for nonresponse and imputation

- Procedure to save tables and PRINT them in different formats or style without need for rerunning the jobs

Long Range Plans

We hope to develop SUDAAN language now used internally to develop current procedures to the extent that other users can use it and write their own procedures. The research on the development of the language is continuing at a slow pace. A trial version is being tested in house on a few simple tasks, and we have encountered a few problems that require further research to develop a generally useful version. A detailed description of the Language is presented in LaVange et al. (1991). Once the SUDAAN language is developed, the source listing (in SUDAAN) for each of the procedures will be available to the users. This will facilitate writing of new procedures or special modifications of existing procedures by the users.

Bibliography

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review 51, pp. 279-292.

Godambe, V. P. and M. E. Thompson (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. International Statistical Review 54, pp. 127-138.

Kendall, M. G. and A. Stuart (1973), The Advanced Theory of Statistics, Hafner Publishing Company: New York.

LaVange, Lisa M., Babubhai V. Shah, Beth G. Barnwell, Joyce F. Killinger (1991) "SUDAAN: A comprehensive package for SUrvey DAta ANalysis", Data Quality Control pp. 209-226 edited by Liepins and Uppuluri, Marcel Dekker, Inc.

Shah, B. V., B. G. Barnwell, P. N. Hunt and L. M. LaVange (1992). SUDAAN User's Manual. Research Triangle Institute, Research Triangle Park, NC 27709.

Shah, B. V., Ralph E. Folsom, Lisa M. LaVange, Sara C. Wheeless, Kerrie E. Boyle, and Rick L. Williams (1993). Statistical Methods and Mathematical Algorithms Used in SUDAAN. Research Triangle Institute, Research Triangle Park, NC 27709.

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