What's New in SUDAAN 11

Angela Pitts[1], Michael Witt[1], Gayle Bieler[1]
[1]RTI International, 3040 Cornwallis Rd, RTP, NC 27709

Abstract
SUDAAN 11 is due to be released in 2012. SUDAAN® is a statistical software package for the analysis of complex survey and other cluster correlated data. This paper will highlight several new procedures, statistics, and features in SUDAAN 11. The new procedures include WTADJX, which computes survey weight adjustments (otherwise known as calibration weighting) and extends the capabilities currently available in WTADJUST; VARGEN, which allows users to compute statistics between variables; and IMPUTE, which is an extension of SUDAAN 10's HOTDECK procedure and includes three additional imputation methods beyond weighted sequential hot deck. In addition, new statistics include the Breslow-Day test for homogeneity of odds ratios, the kappa measure of agreement for cross-classified tables, confidence intervals for marginals in regression procedures, and the representativity indicator (*R*-indicator) for measuring potential nonresponse bias. Finally, some important new options in SUDAAN 11 will be noted, including by-group processing and the ability to create and use new variables in a SUDAAN procedure.

Key Words: SUDAAN, weighting, calibration, imputation

1. Introduction

SUDAAN 11 continues its tradition of introducing new enhancements that are not available yet in other statistical software packages. Enhancements to include in each new release are chosen based on our customers' needs, requests and feedback. New features and enhancements in SUDAAN 11 are summarized below.

2. New Procedures

2.1 VARGEN Procedure
VARGEN is a new descriptive statistics procedure. This procedure computes point estimates and their associated design-based variances for user-defined parameters that can be expressed as complex functions of estimated means, totals, ratios, percents, population variances, population standard deviations, and correlations. Examples include estimating differences between two variables; estimating the population covariance and Pearson correlation between two variables; testing the significance of a mean (or any statistic) against a nonzero value; and estimating a ratio of means or a ratio of ratios. Point estimates can be computed within subgroups, and subgroup contrasts can also be estimated in similar fashion to other descriptive procedures in SUDAAN.

In its simplest form, VARGEN serves as a convenient alternative to DESCRIPT and RATIO for obtaining basic descriptive statistics, such as means, totals, percents, and ratios. VARGEN has the basic advantage of allowing the user to request all of these

parameters in the same procedure call. This feature alone represents added flexibility over DESCRIPT and RATIO.

VARGEN's uniqueness lies in its ability to compute design-based point and precision estimates (i.e. variances and standard errors) for user-defined complex functions of statistics—a feature not found readily in any other descriptive procedure.

The following summarizes a few of the unique and intended uses of VARGEN:

- VARGEN directly computes the statistical significance of a difference in estimates created with different analysis variables. For example, if the variable BODYWT1 stores the body weight of each person in a sample at time 1, and BODYWT2 stores the body weight of each person in a sample at time 2, VARGEN can be used to estimate the statistical significance of the difference between the weighted means of BODYWT1 and BODYWT2.

- VARGEN tests the significance of a mean (or any statistic) against a nonzero value. For example, one could test whether the weighted mean of BODYWT1 was significantly different from 125 pounds at some specified level of significance.

- VARGEN estimates the precision of a ratio of estimated ratios, such as the ratio of means. For example, one may be interested in estimating the ratio of the mean body weight among females to the mean body weight among males.

- VARGEN estimates precision for functions of variables.

- VARGEN estimates population variances, population standard deviations, and the precision (*i.e.*, standard error) of these statistics. It can also compare the population variances or standard deviations between groups. Note the population variance and population standard deviation are not the same as the variance and standard error of an estimate. The population variance and standard deviation is a measure of the deviation of a variable from its mean. The variance and standard error of an estimate is a measure of the difference between an estimate and its expected value, given the design of the study under consideration.

- VARGEN estimates the population covariance and Pearson correlation between two variables. The design-based standard error of these statistics is also provided.

2.2 WTADJX Procedure

The WTADJX procedure is new weighting procedure. WTADJX is very similar to the WTADJUST procedure introduced in SUDAAN 10. As with WTADJUST, WTADJX is designed to produce weight adjustments that compensate for unit (i.e., whole-record) nonresponse and coverage errors due to undercoverage or duplications in the frame. It can also improve the efficiency of estimated means and totals through sample balancing. As in WTADJUST, WTADJX computes weight adjustments by fitting a generalized exponential model. Model parameters are estimated by solving calibrations equations.

The primary difference between WTADJUST and WTADJX is that in WTADJUST, the vector of model explanatory variables and the vector of calibration variables must be the same. These two vectors are the independent variables specified on the MODEL statement in WTADJUST. In WTADJX, the two vectors are allowed to differ. In WTADJX, model explanatory variables are specified on the MODEL statement. And

calibration variables are specified on the CALVARS statement. This difference is important. With WTADJX, variables that are known for respondents only (e.g., survey questionnaire items) can be used as model explanatory variables in the weight adjustment process. Among other things, this allows researchers to assess the potential for bias in estimates when nonrespondents are not missing at random.

2.3 IMPUTE Procedure

IMPUTE is the new item imputation procedure and replaces the HOTDECK procedure introduced in SUDAAN 10. IMPUTE extends the capabilities of the previous HOTDECK procedure by including four methods of item imputation:

- The **Cox-Iannacchione Weighted Sequential Hot Deck** imputation for item nonresponse, described in Cox (1980) and Iannacchione (1982). This methodology is based on a weighted sequential sample selection algorithm developed by Chromy (1979).

- **Cell mean** imputation, described by Korn and Graubard (1999). This method is new in SUDAAN 11.

- **Linear regression** imputation for continuous variables. This method is new in SUDAAN 11.

- **Logistic regression** imputation for binary variables. This method is new in SUDAAN 11.

The procedure also provides numerous summary statistics for PRINT tables and OUTPUT datasets that were previously not available with HOTDECK, including pre- and post-imputation statistics and imputation class statistics.

3. Enhancements

3.1 CROSSTAB Procedure

3.1.1 Cohen's Kappa Measure of Inter-Rater Agreement

The new AGREE statement in CROSSTAB allows one to estimate the kappa measure of agreement in square tables. Cohen's κ (kappa) Coefficient is a statistical measure of inter-rater reliability. It is generally thought to be a more robust measure than a simple percent agreement calculation, since κ takes into account the agreement occurring by chance. Cohen's κ measures the agreement between two raters who each classify a set of items into N mutually exclusive and exhaustive categories. CROSSTAB provides design-based standard errors of the kappa and tests the hypothesis that H_0: Kappa=0 in each of the analytic strata.

3.1.2 Breslow-Day Test for Homogeneity of Odds Ratios

The new BDTEST statement in CROSSTAB provides the Breslow-Day Test for homogeneity of odds ratios in stratified 2x2 tables. Recall that the RISK statement estimates the design-based adjusted odds ratio (aka *Mantel-Haenszel common odds ratio*) for stratified 2x2 tables, along with their design-based standard errors and confidence intervals (Mantel and Haenszel, 1959; Graubard, Fears, and Gail, 1989). The BDTEST statement extends these capabilities to produce a test of the Breslow-Day hypothesis for homogeneity of odds ratios (Breslow and Day, 1980; Breslow, 1996; Agresti, 2002)

across the epidemiologic stratification variables, accounting for the complex sample design. It tests the null hypothesis that the Mantel-Haenszel odds ratios across the strata are all equal. All 5 test statistics in CROSSTAB are available for testing the null hypothesis (*i.e*., WALDCHI, WALDF, ADJWALDF, SATADJCHI and SATADJF.)

3.2 WTADJUST Procedure

3.2.1 Additional Summary Statistics

Several additional summary statistics have been added to the WTADJUST procedure. These statistics can be referenced in PRINT or OUTPUT statements using the following keywords in the BETAS output group:

Keyword	Description
TOTORIG	Sum of the original sample weights over respondent records.
TOTTRIM	Sum of the trimmed sample weights over respondent records.
TOTFINAL	Sum of the final adjusted weights over respondent records.
CNTLTOTAL	These are the control totals that the adjusted weights should sum to.
DIFFWT	Difference between the final adjusted sample weights and the control totals.

3.3 Logistic Modeling and Weight Adjustment Procedures

3.3.1 Descriptive Statistics for Weight Adjustment and Response Propensity

The LOGISTIC, WTADJUST, and WTADJX procedures will now produce descriptive statistics for the model-predicted response propensity and the weight adjustment. Descriptive statistics that can be obtained include the mean, population variance, population standard deviation, and relative standard deviation of the response propensity and associated weight adjustment. These new statistics can be obtained using the new PREDSTAT statement in SUDAAN. Standard errors associated with these estimates can also be obtained.

3.3.2 Weighted Response Rates

The new PREDSTAT statement can also be used to obtain estimates of the ***weighted response rate*** and the ***R-indicator*** (Representativity Indicator) statistic. The *R*-indicator provides a measure of the representativity of the respondents with respect to the sample or population from which they were drawn. For further information, see Schouten and Cobben (2007), Schuten, Cobben, and Bethlehem (2009), Skinner, Shlomo, Schouten, Zang, and Bethlehem (2009), and Cobben and Schouten (2007). The PREDSTAT

statistics in SUDAAN, including the *R*-indicator statistic, are also discussed in Witt (2010).

3.3.3 Precision Estimates That Properly Account for the Estimated Weight Adjustment

Beginning in SUDAAN 11, LOGISTIC, WTADJUST, and WTADJX can now properly account for the sample weight adjustment when estimating descriptive statistics (means, totals, percents, ratios) and their standard errors for any user-supplied variable. For each respondent record on the input file, the adjusted sample weight used in the computations is the product of the base weight (supplied on the WEIGHT statement) and the adjustment factor computed in the procedure. New statements associated with the weight-adjusted descriptive statistics include the VAR, NUMER, DENOM, TABLES, VCONTRAST, VDIFFVAR, VPAIRWISE, and VPOLYNOMIAL statements.

3.3.4 Additional Design Effects in LOGISTIC, WTADJUST, and WTADJX to Measure Impact of Weight Adjustments

In addition to providing standard error estimates that account for the weight adjustment, LOGISTIC, WTADJUST, and WTADJX also provide two sets of design effect-like statistics. In other words, the realized gains in statistical efficiency (decreases in standard error) from using one of SUDAAN's calibration-weighting procedures can now be measured:

1. One set of design effects measures the potential bias from ignoring the estimation of the weight adjustment. These design effects are called MDEFF statistics in SUDAAN 11 and are defined as the variance of an estimate that accounts for the estimation of the nonresponse adjustment relative to the variance of an estimate that ignores this estimation. The variance estimates in both the numerator and denominator of the MDEFF statistics also account for the complex sample design.

2. The second set of design effects provides a measure of the effect of the weight adjustment on the variance of estimated means, percents, ratios, and totals. These are referred to as the ADEFF statistics. The numerator of the ADEFF design effect is a variance estimate that properly accounts for the estimation of the weight adjustment from the model in LOGISTIC, WTADJUST, and WTADJX. The denominators of these design effects are variance estimates that assume the weight adjustment is equal to 1.00. The variance estimates in both the numerator and denominator account for the complex sample design.

3.4 Modeling Procedures

3.4.1 Confidence Intervals for Predicted and Conditional Marginals

Beginning in SUDAAN 11, all modeling procedures produce 100(1-α)% confidence limits for predicted and conditional marginals, in addition to standard errors and associated *t*-tests. You can use SUDAAN 11 to estimate and compare increasingly popular model-adjusted risks, risk differences, and risk ratios from survey data, including data with multiple imputations.

3.5 Iterative Procedures

3.5.1 Model Parameter Estimates Available at each Iteration

For all modeling and weighting procedures (except REGRESS) the model parameters at each iteration of the Newton-Raphson algorithm used to estimate model parameters can be printed or output to a data file using the new output group ITBETAS or keyword ITBETA. This feature is provided to help researchers detect problematic variables that may cause the iterative algorithms to not converge.

4. Statements

4.1 NEWVAR

The NEWVAR statement allows users to recode existing variables, store the recoded variable in a new variable, and use the new variable in the same procedure for processing (e.g., on a CLASS, MODEL, VAR, or TABLES statement). The NEWVAR statement is available in all procedures and more than one NEWVAR statement can be included in the same procedure call. NEWVAR can create new variables via direct assignment or using IF-THEN-ELSE logic.

4.2 BY and BYV

The BY statement (RBY in SAS-Callable SUDAAN) allows users to request output by the values of the variables specified on the BY statement. The new BY statement in SUDAAN 11 is very similar to the BY statement in SAS. Note the one important difference—SAS requires the dataset to be sorted by the variables listed on the BY statement, while SUDAAN does not. As always, SUDAAN only requires the dataset to be sorted by the variables listed on the NEST statement, unless the NOTSORTED option is specified on the PROC statement. Using the BY statement in a SUDAAN procedure is not equivalent to subsetting the file numerous times by the BY variable values. Instead, the BY statement in SUDAAN is simply a shorthand notation for performing multiple runs of a procedure with a SUBPOPN or SUBPOPX statement for each level of the BY variables.

The BYV (RBYV in SAS-Callable SUDAAN) statement allows users to specify processing options for individual variables specified on the BY statement. For example, a user can specify which direction (ascending or descending) a subset of BY-variable values should be processed with. Similar options can also be specified on the BY statement, but the options specified on the BY statement apply to all the variables listed on the BY statement. If different options are desired for subsets of BY-variables, specify those options via the BYV statement.

Additional requirements include the following:

- All variables on the BY statement are treated as categorical for the purpose of BY-group processing (however, you do not need to include them on a CLASS statement, and missing values can be considered valid levels of a BY variable).

- BY variables can be those created with the NEWVAR statement.

- BY variables can also appear on SUBPOPX and SUBPOPN statements.

- SUDAAN performs the analysis specified in the SUDAAN procedure for each combination of the BY variables. SUDAAN treats each combination as if it were an extension to any SUBPOPN/SUBPOPX statement already specified.

4.3 SUBPOPX

The new SUBPOPX statement in SUDAAN 11 is used to define a subpopulation in a more flexible way than SUBPOPN. SUBPOPX is available in all procedures.
SUBPOPX offers many advantages over SUBPOPN, among them:

- Relational operators can be cascaded. For example:
  SUBPOPX 12 <= AGE <= 65;

- The variable in (value1, value2, …) and variable in (value1:value2, value3:value4, …) syntax is allowed. For example:
  SUBPOPX RACE in (1,2,3);

- Arithmetic operators can be included. For example,
  SUBPOPX WEIGHT <= MAXWEIGHT * 1.20;

- The following logical and arithmetic operators are allowed: NOT, EQ, LT, LE, GT, GE, +, -, * and /.

References

Agresti, A. (2002). Categorical Data Analysis, 2nd Ed. New York: Wiley.

Breslow, N.E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91, 14–26.

Breslow, N.E. and N.E. Day (1980). Statistical Methods in Cancer Research, Vol 1., The Analysis of Case-Control Studies (IARC Scientific Publications No. 32), Lyon, International Agency for Research on Cancer.

Chromy, J.R. (1979) Sequential sample selection methods: Proceedings of the Survey Research Methods Section.

Cobben, F. and Schouten, B. (2007) "An Empirical Validation of R-indicators." Discussion Paper, CBS, Voorburg. Available at http://www.cbs.nl/NR/rdonlyres/A789F70B-AF96-4635-9E3D-050EB2B3FFB0/0/200806x10pub.pdf

Cox, Brenda G. (1980). "The Weighted Sequential Hot Deck Imputation Procedure." Proceedings of the American Statistical Association, Section on Survey Research Methods.

Graubard, B., Fears, T., and Gail, M. (1989). "Effects of cluster sampling on epidemiologic analysis in population-based case-control studies." *Biometrics* 45, 1053-1071.

Iannacchione, Vincent G. (1982). "Weighted Sequential Hot Deck Imputation Macros." Presented at the Seventh Annual SAS User's Group International Conference.

Korn, E. L. and B. I. Graubard (1999). Analysis of Health Surveys. New York: Wiley.

Mantel, N. and Haenszel, W. (1959). "Statistical aspects of the analysis of data from retrospective studies of disease." *Journal of the National Cancer Institute* 22, 719-748.

Research Triangle Institute (2012). SUDAAN Language Manual, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.

Schouten, B., Cobben, F. and Bethlehem, J. (2009) "Indicators for the Representativeness of Survey Response." *Survey Methodology, June 2009*. Vol 35, No. 1 pp 101-113.

Schouten, B. and Cobben, F. (2007), "R-indexes for the comparison of different fieldwork strategies and data collection modes." Discussion Paper 07002, Statistics Netherlands, Voorburg, The Netherlands. Available at http://www.risq-project.eu/papers/schouten-cobben-2007-a.pdf

Witt, M. B. (2010) Estimating the R-indicator, Its Standard Error and Other Related Statistics with SAS and SUDAAN. In JSM Proceedings, Section on Survey Research Methods. Alexandria, VA: American Statistical Association