WESVARPC: SOFTWARE FOR COMPUTING VARIANCE ESTIMATES FROM COMPLEX DESIGNS

David Morganstein and J. Michael Brick
Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850

ABSTRACT

This paper discusses a Windows-based software program, WesVarPC, for computing sampling errors for statistics collected from complex survey designs. The program uses replication methods (either Balanced Repeated Replication or Jackknife) to compute estimated sampling errors. WesVarPC produces estimated sampling errors for a wide variety of statistics, including totals, percents, ratios, and more complex functions of totals. Estimated sampling errors can be computed for subdomains defined by cross-tabulations of as many as eight variables. Sampling errors of functions of cell statistics, such as differences or log odds-ratios, can also be produced easily. WesVarPC reads SAS and flat ASCII file formats. The program can use replicate weights attached to the incoming survey data file or it can produce post-stratified replicate weights if control totals are provided. The paper begins with a discussion of replication methods. The main portion of the paper is a description of the WesVarPC program features. WesVarPC is available to all interested persons and organizations at no charge.

KEYWORDS: Replication methods; jackknife methods; statistical software; linear and logistic modeling

1. INTRODUCTION

This paper describes a method for estimating sampling errors of estimates derived from complex surveys using replication methods. The software for computing these estimates is WesVarPC, a stand-alone software package developed by Westat, Inc. for use in the PC environment running under Windows 3.1.

It is well known that the sampling errors of estimates made from survey data collected with designs involving stratification, clustering, or unequal selection probabilities are not the same as those from simple random samples. The estimators of sampling errors computed in most statistical packages, such as SAS and SPSS, assume simple random sampling and most often underestimate the standard errors of the estimates. Since the assumption is inappropriate for data collected using complex sample designs, special purpose computer programs are needed. WesVarPC can be used to correctly compute estimators of sampling errors for such data by using replication.

Replication is one of three approaches for estimating sampling errors for complex surveys. The first approach is to use the exact formula for the sampling errors of each statistic. Several problems make using this method impractical for application in a general purpose solution: some of the formulas for the standard errors are intractable; the formulas depend on the specific sample design; and, the formulas depend on the estimator. For these reasons, approximate methods of computing sampling errors have been considered for most practical problems.

The second approach is to approximate the statistic using a Taylor series linearization, and then to compute the sampling error for the linear statistics. This method has been implemented in several software packages for the PC, such as SUDAAN (Shah *et al.*, 1989) and PCCARP (Fuller *et al.*, 1989). See Carlson *et al.* (1993) and Cohen *et al.* (1988) for reviews on software available for use on a PC.

The third method, used in WesVarPC, is replication. Replication has some advantages that make it an important technique for estimating sampling errors from complex samples. Replication can be applied to a wide variety of sample designs. It is easily applied to many estimators without users modifying the software. When done properly, replication can take into account frequently made adjustments in the survey weights to account for nonresponse, post-stratification, raking adjustments, or other calibration estimation methods.

WesVarPC is a successor to earlier mainframe software produced by Westat as a SAS user-written procedure. The main-frame software was developed in the late 1970's and the 1980's by Westat for main-frame IBM and mini-VAX computers using the SAS engine. The three SAS procedures were called PROC WESVAR (for computing sampling errors of estimates in tables), PROC WESREG (for linear regression modeling), and PROC WESLOG (for logistic regression modeling). See Flyer *et al.* (1989) for a discussion of this software. While WesVarPC is a descendant of these programs, it greatly enhances their capabilities. Some of the new features of WesVarPC are described below.

2. USER FEATURES

WesVarPC was developed for operation in the Windows-PC environment. It is a stand-alone program written in the C language. It will operate on a 386 computer as long as the Windows operating system is present (a math/coprocessor is recommended). At least 4 megs of RAM are required. Tests indicate that WesVarPC running on a 486 PC produces results almost as quickly as PROC WESVAR did on a mainframe.

The choice to develop WesVarPC within the Windows environment was a deliberate one. Unlike all other sampling error programs that use a text editor to type in a program in an arcane language, the WesVarPC user specifies actions by selecting from menu items. The user does not have to refer to a dictionary or codebook to assist in correctly typing variable names. The environment makes it easy to learn how to specify requests in WesVarPC for users familiar with Windows. Functions of variables and functions of table cells (such as log odds ratios) are specified in a calculator-like environment. Regression models are formulated easily by clicking on the dependent and independent variables in the models. Once a request is constructed, it is stored in a retrievable format, allowing changes to be made at a later time. WesVarPC is the first sampling errors program that provides the familiar interface of all Windows programs.

An important feature of any sampling errors program is the formats of survey data files it can read and write. In WesVarPC, survey data files can be read from a variety of file formats: SAS for the PC (version 6.04), SAS Transport from any platform including Windows, SPSS for Windows, dBase, and ASCII text format can be imported into WesVarPC. The program will read files with hundreds of variables and thousands of records. WesVarPC has been tested with files of 80,000 records and 200 variables. WesVarPC also provides a simple means of exporting files so that they can be merged with the original data set. This is especially valuable if replicate weights are created using WesVarPC.

Two user interfaces are provided to help improve the presentation of the results and the use of the software for production volume work. The first is an Excel macro that reads the output listing produced in WesVarPC and formats the output for presentation using the facilities of that spreadsheet software. The macro is included with the WesVarPC software and can be used as a template for developing procedures for other spreadsheet packages. The second interface is a description of how to use the software without having to specify the requests in the usual Windows point-and-click environment. This is especially helpful for production work where hundreds of tables are needed and a graphical interface is not as efficient as a C program that specifies the requests. A C program prototype that can be used to produce request files is supplied.

The documentation for the program is provided in two formats: an easy to read User's Guide (Brick *et al.*, 1996) with appendices containing technical details; and, internal help screens. The User's Manual was written for a researcher with a limited background in sampling error theory. The Manual assumes very little knowledge, outside of using Windows, on the part of the user and contains example screens of nearly everything the user will encounter when running the program. Every screen includes a Help option to access an explanation of the currently available options.

The software is available free-of-charge and can be downloaded from the Internet. The URL for Westat is http://www.westat.com. From there the user can navigate to download the software and documentation. The documentation is in Adobe Acrobat reader files so that the user can print a copy of the manual. The latest changes in the software are always available from this site. A listserve is available (send an email to *listserv@listserv.westat.com* with a text message of *subscribe WESVAR-L*) for discussion of the software and applications of replication. For more information, send an email to *wesvar@westat.com*.

3. COMPUTING SAMPLING ERRORS

There are three major components of using WesVarPC to compute sampling errors for survey statistics. The first is Prep, used to import the survey data file into a WesVarPC formatted file. Tables is the second component, used to prepare estimates and sampling errors for the estimates in table format. The third component is Regression. It is used to estimate the sampling errors for linear and logistic models. In addition to these components, WesVarPC has limited ability to label and recode variables (Format component) as well as display and print the output listings (Browse component). These features are described in detail in the User's Manual but this article will concentrate on the Prep, Tables, and Regression components.

3.1 Prep

The main functions of the Prep component are to import data from other file formats, create replicate weights, and poststratify the full sample and replicate weights. As mentioned earlier, WesVarPC can import a variety of file formats making it relatively simple to bring the data directly into the program. Since WesVarPC is based on replication, the key to computing the sampling errors is developing replicate weights. If replicate weights are already on the data file, these can be imported during the same process. For those survey files without replicate weights, a limited capacity for creating them is available in Prep.

WesVarPC supports both balanced repeated replication (BRR) and jackknife replication methods. If the sample design can be viewed as a stratified cluster design with two Primary Sampling Units (PSUs) selected per stratum, then either one of two forms of BRR replicates can be defined. The two forms are the ordinary BRR, in which half the sample is used to form replicate estimates of the statistic, and Fay's variant of BRR, in which the weights are perturbed using a different factor (see Judkins, 1990). A stratified jackknife procedure is also supported by the software for this commonly occurring design. The other procedure supported is the delete-one jackknife, which is more appropriate for systematic random samples.

Many designs can be approximated by these designs. For example, although they do not contribute a between-PSU component to sampling error, certainty PSUs do contribute a within-PSU component of variance. The units within the certainty PSU can be divided into two groups that can be viewed as a pair of PSUs for variance estimation purposes. Single PSUs per stratum designs can be collapsed to form a pseudo-stratum consisting of two strata. In designs with more than two PSUs per stratum, the major strata can be divided into variance computation strata, each with a pair of PSUs.

Although there is a facility for creating replicate weights in WesVarPC, this may not be sufficient for all designs. For example, if you have three PSUs per strata, it may be best to set up the replicate weights before you import the file into WesVarPC. Similarly, nonresponse adjustments can only be reflected in replicates created in WesVarPC if they are a part of a poststratification adjustment to external totals. However, if the replicate weights are created outside of WesVarPC and imported, a wide range of designs and adjustments can be properly reflected in the replicate weights.

3.2 Tables

One of the most useful components of WesVarPC Tables allows users to specify a multi-way table with up to eight dimensions within which statistics can be calculated. For example, sampling errors can be computed within a 3-way table defined by region of the country, age (in categories) and gender of the respondent by simply clicking on the three variable names that define the table. The output listing contains the sum of weights for each cell of the three-way table, row, column, and total percents along with their sampling errors, and confidence intervals.

By clicking on variable names, the user can compute estimated totals for analytic variables. Similarly, statistics such as the average income or the ratio of income to assets can be computed by simply clicking a few icons on the screen. Functions of cell estimates can also be requested easily. For example, the sampling error of a log-odds ratio can be defined by naming the cells of the table and specifying that a log is to be taken of the odds ratio for the estimated cells counts. These functions of cell estimates are especially useful for testing the statistical significance of various user-specified contrasts.

As the user prepares a variety of computations by selecting variables, functions and tables, WesVarPC records the request in a data file. This permanent request file can be re-opened at a later time, modified and re-submitted. Requests for sampling errors of specific variables can be added or subtracted easily. By double clicking on variables that were previously selected, they are removed. Similarly, new tables can be added to the request and old ones deleted, simply by clicking on the specified table. This method of defining a request is particular easy to learn and use as there is no need for the user to learn and to type a special purpose, arcane language.

3.3 Regression

Linear and logistic modeling are supported in the Regression component in WesVarPC. The methods of preparing requests in Regression are much like those used in Tables. The user specifies the dependent variable being modeled as well as the independent variables used to predict it. If logistic regression is used, then the dependent variable must be binary. Transformations of the independent variables, such as the log or square root of the variable can be specified using a calculator pad and clicking on the variables.

A variety of regression models are supported by this procedure. An important class of models that is supported is ANOVA models, where the dependent variable is predicted using categorical independent variables. Rather than requiring the user create a series of zero-one variables for each level of the independent variables, each variable with fewer than 10 response levels is automatically defined as a Class variable. If the user clicks on the Class variable, then all of the zero-one variables are automatically entered into the model (except the last level which is the reference cell and is set equal to zero). This makes the specification of the models with categorical variables simple.

Two-way interactions can also be used with Regression. The user simply clicks on both the variables in the interaction along with a click for the interaction symbol. This option enables users to analysis two-way factorial models, in addition to main effects models. All of these procedures are available for both linear and logistic regression models.

Two other powerful features in Regression are the ability to test hypotheses about the estimated parameter values and the ability to generate a covariance matrix of the estimated parameters. An example of the first feature is a test whether or not a particular categorical variable is a statistically important item for the model by simultaneously testing if any of the levels of the variable is equal to zero. The second feature, the covariance matrix of the estimated parameters, can be used with a variety of other multivariate procedures that are require the correct covariance matrix for solution.

As in Tables, users specify a request and then submit it to obtain the output. The request file can be saved and resubmitted later, making whatever modifications might be necessary. A number of defaults, such as whether or not to include an intercept term in the model, can also be used to broaden the types of models and the format of the output listings.

4. Future plans

This paper describes some of the attributes of WesVarPC version 2.0. At this time, we are discussing the utility of the program with users to determine if there are features that would make it simpler to use in practice. In addition, we are now in development of an expansion of the software to compute sampling errors for medians and quantiles. This effort requires careful consideration because some theoretical findings suggest that replication may not provide consistent estimates for quantiles from complex samples; although, some evidence shows that BRR does well in many circumstances. Other areas of expansion and improvement that are raised by users will also be explored.

REFERENCES

BRICK, J.M., BROENE, P., JAMES, P., AND SEVERYNSE, J. (1996) "A User's Guide to WesVarPC," Westat, Inc., Rockville, MD.

CARLSON, B. L., JOHNSON, A.E., and COHEN, S.B. (1993) "An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data," *Journal of Official Statistics*, Vol. 9, 4, 795-814.

COHEN, S.B., XANTHOPOULOS, J.A., and JONES, G.K. (1988) "An Evaluation of Statistical Software Procedures Appropriate for the Regression Analysis of Complex Survey Data," *Journal of Official Statistics*, 4, 17-34.

FLYER, P., MORGANSTEIN, D., and RUST, K. (1989) "Complex Survey Variance Estimation and Contingency Tables Analysis Using Replication," Proceedings of the Survey Research Section of American Statistical Association.

FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J., (1989) "PC CARP," Statistical Laboratory, Iowa State University, Ames, IA.

JUDKINS, D. R., (1990) “Fay’s Method for Variance Estimation,” *Journal of Official Statistics*, Vol. 6, 3, 223-239.

MORGANSTEIN, D.R., (1995) "WesVarPC," Proceedings of the Meetings of the International Statistics Institute.

SHAH, B.V., LAVANGE,L.M., BARNWELL, B.G., KILLINGER, J.E., and WHEELESS, S.C., (1989) "SUDAAN: Procedures for Descriptive Statistics User's Guide," Research Triangle Institute, Research Triangle

Park, N.C.