AN EVALUATION OF THE USE OF PERSONAL COMPUTERS FOR VARIANCE ESTIMATION WITH COMPLEX SURVEY DATA

Barbara Lepidus Carlson, Ayah E. Johnson, and Steven B. Cohen, Agency for Health Care Policy and Research
Barbara Lepidus Carlson, AHCPR, Parklawn Building, Room 18A-55, 5600 Fishers Lane Rockville, MD 20857

KEY WORDS: Statistical software, complex survey design, SUDAAN, PC CARP

The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Health Care Policy and Research is intended or should be inferred.

1. INTRODUCTION

Many national surveys have sample designs that deviate from simple random sampling. Stratification is often considered to increase the precision of survey estimates. Clustering is frequently used to make the field work of the survey more efficient. In addition, when greater representation of certain policy-relevant subgroups is necessary, disproportionate sampling is often used. Sampling weights are calculated to reflect the unequal probabilities of selection.

Mainframe computers have been the primary resource used to support federal research and analysis. Since standard statistical computing packages, such as SAS and SPSS, assume simple random sampling, any variance estimates arising from them may not reflect the actual variance achieved by adoption of a more complex design. Specialized software which accounts for complex survey designs when estimating variances has existed for about a decade, but primarily for use on mainframe computers.

With the increased prevalence of personal computers (PCs) and their increasing capacity and speed, the idea of analyzing survey data files on PCs becomes more plausible, particularly moderate-sized data files, say, less than 25,000 records, arising from relatively smaller surveys or subsets of larger surveys. Using PCs rather than mainframe computers has many potential benefits as well as costs.

The benefits generally relate to cost savings, but freedom from some aspects of mainframe computing also make PCs more attractive. While actual execution time on the mainframe may be substantially less than on a PC, the time from submission of a batch job to the receipt of the printout is generally much longer on a mainframe. Mainframes often operate on a time-sharing basis, which may mean waiting in an execution and/or print queue during a busy period. All mainframe computers have down-times, some more often than others. Due to prohibitive expensive, large jobs are often submitted for execution during a discount time, usually overnight or over the weekend, which substantially slows down the entire process.

One generally has immediate access to printouts when using a PC. PC packages often have an interactive or menu format, rather than a batch format, which can make a package easier to learn and use. And furthermore, the freedom from Job Control Language, used to inform the mainframe operating system how to process the job, is an attraction of the PC. There is usually an option with PC packages to output tables and other results from an analysis into a separate textfile, which can be quite helpful later for creating tables without retyping the numbers.

Costs associated with using PCs relate primarily to the run-time issue, as well as space and memory constraints. Obtaining computing equipment well-suited for statistical analysis can become quite expensive, since it must have enough memory, disk space, speed, and often a mathematical co-processor. Unless one has a memory manager which allows for several tasks to be performed simultaneously, a PC can be "tied up" while running a lengthy analysis or downloading a file. The need to download datafiles from the mainframe, along with its potential for introducing transmission errors into the database, is also a consideration.

Since large databases require significant resources in order to be analyzed in a timely and efficient manner, under what circumstances will the PC versions of variance estimation programs be useful in keeping computing costs down without sacrificing the efficiency normally associated with the power of a mainframe computer? Two of the more frequently-used mainframe packages for the analysis of complex survey data now have PC counterparts: SUDAAN and PC CARP.

In this paper, each of these PC programs is evaluated relative to its mainframe version, and the two PC programs are compared to each other. A comparison solely among the mainframe packages is made as well. Features available in these packages as well as issues related to the actual implementation of the programs, including data preparation steps, number of programming statements, time and cost issues, are examined using two data sets from the 1987 National Medical Expenditure Survey (Edwards and Berlin, 1989), which has a complex sample design. Programming statements and samples from resulting output can be obtained from the authors.

2. BACKGROUND

When data have been collected from a survey which has a complex sampling design, the simple random sample assumption made by most statistical computing packages can often lead to an underestimate of the variance, which can therefore lead to artificially small confidence intervals and anticonservative hypothesis testing; i.e., rejecting the null hypothesis when it is in fact true.

A few different statistical strategies have been developed to address this issue. Among them are: a first-order Taylor series expansion of the variance equation; a balanced-repeated replication method (BRR); and the jackknife approach (Wolter, 1985). Several software packages have been developed which incorporate one or more of these strategies into their variance calculations.

The current evaluation focused only on those software packages which currently have PC counterparts: SESUDAAN, SUDAAN, and SUPER CARP on the mainframe, SUDAAN and PC CARP on the PC. Other programs which are designed to analyze data from complex surveys exist, but have no PC counterparts to date, and are therefore not pertinent to the subject matter of this paper. These mainframe packages have been evaluated elsewhere (Cohen et al., 1986, Cohen et al., 1988).

SUDAAN (Shah, 1990) and SESUDAAN (Shah et al., 1981) are two programs appropriate for estimating variances from complex survey designs and have been developed by the Research Triangle Institute (RTI). Among many other capabilities, these programs can compute weighted means and proportions and their corresponding standard errors

(Table 1). SESUDAAN is a mainframe package only, while SUDAAN has PC, mainframe, and VAX/VMS versions. It should be noted that SUDAAN on the mainframe is still in test mode and is not yet available to the public. Any comments regarding the mainframe SUDAAN should be regarded in this context. SUDAAN will accept both SAS and text datafiles. SESUDAAN will accept only the SAS data format.

Table 1. Comparison of Software Capabilities

	SESU-DAAN	SUDAAN	SUPER CARP	PC SUDAAN	PC CARP
Totals	Yes	Yes	Yes	Yes	Yes
Means	Yes	Yes	Yes	Yes	Yes
Proportions	Yes	Yes	Yes	Yes	Yes
Quantiles	No	Yes	No	Yes	Yes
Ratios	Yes	Yes	Yes	Yes	Yes
Differences	No	Yes	Yes	Yes	Yes
Cross-tabs	Yes	Yes	No	Yes	Yes
Design effects	Yes	Yes	No	Yes	Yes
Regressions	(SURREGR)	*	Yes	No*	Yes
Chi-square test of independence	No	Yes	Yes (no p-value)	Yes	Yes (no p-value)

*Future plans include regression analysis.

A Taylor Series approximation is used in SESUDAAN and SUDAAN to compute variance estimates. RTI has produced a family of such programs, including regression and logistic regression packages. While most of these programs were written in SAS language (SAS Institute, Inc., 1985), the newest of them, SUDAAN, is written in "C". SUDAAN incorporates the features of SESUDAAN, RATIOEST (ratio estimation package), and RTIFREQS (frequencies), and has many improvements over these older versions. Future plans include incorporating the regression packages into SUDAAN as well. Although RTI refers to both the mainframe and PC versions as "SUDAAN", for purposes of clarity, the PC version will henceforth be referred to as "PC SUDAAN" in this paper.

SUPER CARP (Hidirouglou et al., 1980) and PC CARP (Fuller et al., 1988) are products of the Statistical Laboratory at Iowa State University. SUPER CARP is a mainframe package, the latest version of which is approximately ten years old. PC CARP, its PC counterpart, is relatively recent, and has many improvements over its mainframe parent. These programs are written in FORTRAN G and also make use of the Taylor Series approximation method. Two supplemental programs, logistic regression and post-stratification, are also available. SUPER CARP and PC CARP will only accept text datafiles.

3. THE SURVEY DATA

The 1987 National Medical Expenditure Survey (NMES), sponsored by the Agency for Health Care Policy and Research, included two distinct household surveys. The first is a national probability sample of the civilian, noninstitutionalized U.S. population. The second is a survey of the American Indian and Alaska Native population living on or near reservations and eligible for services from the Indian Health Service (IHS). Both household survey components were designed to provide statistically unbiased national estimates of health care utilization, expenditures, and access to care, and health insurance coverage for their respective target populations for calendar year 1987. The Household Survey (HHS) and Survey of American Indians and Alaska Natives (SAIAN) sample designs can be characterized as stratified multi-stage area probability designs with three stages of sample selection.

4. METHODS

The computer packages were evaluated with respect to efficiency, accuracy, and ease of use on both a mainframe and a personal computer, since each package has a version for both environments. For each package, weighted means and weighted proportions were estimated for the same set of variables on each of two similar data sets: the HHS data and the SAIAN data. Along with these estimates, standard errors and design effects were also computed. The evaluation of these software packages is done by examining: programming effort, comparison of features and flaws, handling of missing data, execution time, computing costs, computational accuracy, and documentation.

4.1 Computing Environment

The mainframe computer used is an IBM 3090 Model 300J located at the National Institutes of Health in Bethesda, Maryland. It runs under the OS/MVS/ESA operating system. The personal computer used is an AGI 3000D (a 386 IBM clone) with an 80386 Processor, 80387 Numeric Co-Processor, 40 mb hard disk, and 3 mb extended RAM (not expanded). It runs under the MS-DOS (version 3.30) operating system.

4.2 Data

Each of the two data sets (HHS and SAIAN) consisted of stratum and primary sampling unit (PSU) indicators, a sampling weight, and 35 variables for which estimates of weighted means or proportions, as appropriate, were computed. There are 11 continuous variables and 24 categorical variables with up to 7 categories each, half of which are dichotomous. Missing data existed in the file for most of the variables, due to nonresponse (at most 10%) as well as intentional skip patterns in the questionnaire. The inclusion of missing data enabled evaluation of how each software package handled these cases.

In order to test the effect on computer time of number of observations and strata, and to reduce the number of observations to a suitable size, a subset of the HHS population was selected. In this case, the file being tested included only non-whites. Furthermore, the SAIAN and HHS files were limited to round-one respondents. Those who did not respond to the questionnaires containing the variables used were excluded from consideration.

Variables were chosen such that roughly one-third of them were dichotomous, one-third categorical (more than two categories), and one-third continuous. An enumeration of these variables, along with a description of the variable type, can be obtained from the authors.

4.3 File Description

The SAIAN data set had 5,584 observations, and the HHS data subsample had 8,310 observations. For SAIAN, there were 11 strata, with 2 PSUs per stratum. Each PSU had between 53 and 788 observations. For HHS, there were 71 strata, with 2 PSUs per stratum, each PSU having between 2 and 298 observations. Three strata had to be collapsed with adjoining strata due to empty PSUs for the HHS subsample

chosen.

The SAIAN textfile had a logical record length of 77, and took up about 450 kb. The SAS version of the SAIAN datafile took about 950 kb of space on the mainframe, 1.7 mb of space on the PC. The HHS textfile had a logical record length of 71, and took up approximately 610 kb of space. The SAS version of the HHS datafile took up 1.4 mb on the mainframe, 2.6 mb on the PC.

4.4 Procedures

On the mainframe, SAS data sets sorted by stratum and PSU were created, keeping only the variables and observations relevant to the evaluation, and converting missing values (originally coded as negative numbers) to dots, the SAS convention for missing numeric data. All of the programs being evaluated require that the data be sorted by the nesting variables. These SAS data files were used for SESUDAAN and SUDAAN runs. Text files were then created from these SAS files using SAS and then WYLBUR, the mainframe's online interactive text editor, to change dots to blanks for missing data (the FORTRAN convention for missing data), and were used for SUDAAN and SUPER CARP runs. SUDAAN will accept both text and SAS datafiles. The text files were downloaded to the PC using MS-DOS Kermit software. These text files were then used on the PC for PC SUDAAN and PC CARP runs. The files were input into the PC version of SAS (SAS Institute, Inc., 1988) and then used for one more PC SUDAAN run.

In addition to the data file and the program code, several other files were necessary in order to run SUDAAN in both mainframe and PC environments. For SUDAAN, when using a SAS data file, one need only create in addition a "LEVEL" file, which contains labels for different values of any or all categorical variables, as would appear in a SAS PROC FORMAT value statement. This file is optional. When using a text data file in SUDAAN, this "LEVEL" file is also optional, but an "LBL" file and a "LABELS" file are always required, regardless of whether there are categorical variables, listing the variables in the order in which they appear on the data set, and variable descriptors (length, type, filename), respectively. These auxiliary files were downloaded to the PC as well, using Kermit. No auxiliary files were necessary to run SUPER CARP or PC CARP.

Programming effort was measured by the number of statements required to run the program. For the SESUDAAN programs, a statement was defined as ending with a semicolon. For SUDAAN programs, the number of statements was defined as the number of program statements ending with a semicolon plus the number of lines required in all of the necessary auxiliary textfiles. For SUPER CARP, the number of statements was equal to the number of lines in the program, excluding the extra lines needed to enumerate the variable names. Since PC CARP was not run from a batch program, the number of responses to PC CARP prompts was recorded.

When writing the programs to execute all of the packages being evaluated, an attempt was made to minimize the number of steps needed to execute the program and to make the runs on the various software packages as similar as possible. However, although SUPER CARP and PC CARP do not allow for formats (i.e., labeling of variable response categories), use was made of this capability for the SUDAAN and SESUDAAN runs. In SUPER CARP and PC CARP, extra steps became necessary due to the way in which the program handles missing data (described later in more detail).

Execution times and computing costs, two of the outcomes of interest, were automatically computed and recorded on the printed output from the mainframe runs. A precise execution time for PC CARP is inappropriate due to the interactive, rather than batch, nature of the program and its dependence on user response and key-entry speed. Therefore, for all of the PC runs, approximate run times from the first keystroke to the final result were recorded, minimizing hesitation at the prompted questions for PC CARP. Computational accuracy was evaluated by examining the output from the various programs and determining at which decimal point discrepancies began to occur.

5. RESULTS

5.1 Programming effort

As described in the Methods section, programming effort was measured here in terms of number of programming statements required. Disregarding Job Control Language, for mainframe packages, and statements used to label categorical variable responses, SUDAAN and PC SUDAAN using a SAS datafile and SESUDAAN required the fewest programming statements (see Table 2). Using a text datafile in SUDAAN requires several extra files to help describe the data, which increases the programming effort at least six-fold, both on the mainframe and the PC. SUPER CARP requires a large number of statements since every estimate requires an extra set of specification lines. Since PC CARP does not have a batch mode, the number of necessary responses to prompts for the estimates run were recorded instead. This puts PC CARP at the top of the scale, with 353 "statements" versus the SUDAAN program (SAS data) with 11 programming statements.

Table 2. Number of Programming Statements[1] Required for 35 Estimates[2]

Mainframe packages	
SESUDAAN	15
SUDAAN (SAS datafile)	11
SUDAAN (text datafile)	87
SUPER CARP	82
Microcomputer packages	
PC SUDAAN (SAS datafile)	11
PC SUDAAN (text datafile)	87
PC CARP	353[3]

[1]Excluding format-related statements, and JCL for mainframe runs.
[2]"Estimate" here refers to one mean (continuous variables) or a set of proportions (for all values of a categorical variable).
[3]Number of keyboard responses to prompts.

Labeling values of categorical variables becomes a bit more cumbersome when one leaves the SAS environment of SESUDAAN, and its PROC FORMAT, for the SUDAAN environment, which requires an extra text file with one line per value. Neither SUPER CARP nor PC CARP have the ability to format categorical variables.

5.2 Comparison of features and flaws

5.2.1 SESUDAAN vs. SUDAAN vs. PC SUDAAN

SUDAAN has many advantages over its predecessor SESUDAAN, although some flexibility was lost in the transition. SESUDAAN had the benefit of being a procedure within a SAS program. Data could be easily manipulated within the same program using SAS DATA steps, whereas the dataset running under SUDAAN has to be a permanent data set. Titles and formats could be specified in the same program. With the change from SAS to C, these options disappeared, making it more difficult to describe the data as well as to modify variables and re-run the program.

SESUDAAN was more limited, however, in many other respects. When proportions were being estimated, the standard output showed the unweighted denominator but not the numerator. The weighted numerator was available, but had to be calculated from the given information. In addition, one must divide variables into "report" versus "analysis" variables for each run; it is not necessary to make this distinction with SUDAAN. SUDAAN's new features include Chi-square testing, quantiles, and contrasts. It allows for post-stratified estimates and has much more flexibility with respect to the specification of the sampling design. SUDAAN also produces column and row totals automatically, something which SESUDAAN had to be "tricked" into doing.

IBM mainframe computers are very particular when reading textfiles. One problem encountered when running SUDAAN was resolved when all of the files being read by the program had a logical record length set equal to the last column in which there was information. Using a text datafile with more than 80 columns proved impossible to use, yielding error messages which did not pinpoint the problem in SUDAAN's interpretation of the data it was reading.

5.2.2 SUPER CARP vs. PC CARP

PC CARP has several improvements over its predecessor, SUPER CARP. In SUPER CARP, instructions must be specified in particular columns, in a somewhat scattered set pattern. The order of the rows is not at all flexible. In PC CARP, that is not an issue, due to the menu-driven mode of specification. In addition, PC CARP computes design effects, cross-tabulations, and quantiles, and automatically prints relative standard errors (C.V.s), none of which SUPER CARP produced. Both PC CARP and SUPER CARP offer tests of independence for categorical variables. While not the case with SUPER CARP, PC CARP's categorical variable responses are now displayed in numeric order, but for univariate proportions only.

5.2.3 PC SUDAAN vs. PC CARP

PC CARP and PC SUDAAN are comparable in their space and memory requirements. The PC SUDAAN software takes up roughly 790 kb of disk space, and requires 640 kb RAM. It can run on any IBM-compatible PC. The PC CARP software takes up roughly 530 kb, including PRE CARP, and requires 450 kb RAM. It can run on an IBM-compatible machine which must have a mathematical co-processor.

An advantage of PC CARP over PC SUDAAN is the ability to automatically collapse strata when necessary. A disadvantage is the limitation of only being able to read in a textfile. The FORTRAN system requires an explicit description of the input format (in FORTRAN notation), while SAS does not, and the results are printed out in scientific notation. In addition, there are limitations to the number of variables that can be read into the program.

The PC CARP software does not accommodate categorical variable formats, and for straightforward mean computation, a ratio must be computed with the denominator being the sum of the sampling weights, by specifying an all-ones variable. For proportions, the all-ones variable was the dependent variable. Outputting data files with the computed estimates is not an option with the CARP software. Although significance tests are often available, the test statistic is printed without a p-value.

Other than these observations, the main difference between the two PC packages is in their mode of execution. Having to specify multiple analyses via the interactive mode of PC CARP became tedious, since most of the 35 analyses run during our evaluation had to be specified individually with separate runs. One potential improvement to PC CARP would be a batch mode option, which would be especially helpful for large numbers of estimates, when one tends to write out in advance the types of details which correspond to PC CARP's prompts. Two problems with PC CARP's interactive mode are that keyboard entry errors are not readily changeable, and it is also difficult to determine when PC CARP is actually executing, rather than waiting for your response to the previous prompt, since the screen does not change after response to the last question.

5.2.4 SESUDAAN and SUDAAN vs. SUPER CARP

With the SUDAAN programs, most statements can be in virtually any order, compared to the rigid structure of SUPER CARP commands. One exception is that the environment-setting command, SETENV, apparently has to be specified before the PRINT command. This is not specified in the manual (in fact, the examples given show the reverse order), and should be revised.

As with their PC counterparts, the SUDAAN programs differ from SUPER CARP with respect to the ability to collapse data and the type of datafiles and missing value indicators they will accept. SUPER CARP requires the specification of the data formats in FORTRAN and outputs its numeric data in scientific notation. SUPER CARP can provide no output datafile, does not allow categorical variable formatting, and does not display p-values. The SUDAAN programs have these capabilities on the mainframe, as they do in the PC version. Virtually the same variable limits apply within the packages.

5.3 Handling of missing data

How these packages deal with missing data turned out to be one of the most difficult issues. SESUDAAN and SUDAAN treat blanks and dots as missing values for continuous variables. Values of categorical variables outside the range specified earlier in the program were also treated as missing in both of these packages, including values less than or equal to zero.

As mentioned in the Methods section, SUPER CARP and PC CARP are quite limited in how they deal with missing data. Basically, they just don't accept missing data. If it is there, a missing value is treated as a zero. In order to get around this problem, missing value indicator variables had to be created for each variable containing missing data, more than doubling the size of the text files. In both SUPER CARP and PC CARP, these indicator variables were treated as subgroup variables.

In SUPER CARP, one can specify screening operations, which will screen out observations based on equality or inequality statements. This would have worked fine had the negatively-coded missing values not been converted to blanks, but one cannot specify the deletion of values equal

to "blank". The major problem with the screening option in SUPER CARP is that whole records are deleted, rather than excluding one value; i.e., it is a list-wise deletion. When multiple variables are being considered, one doesn't want a whole record deleted from all analyses simply because one of the variables has a missing value.

PC CARP does not have a screening procedure, but comes with a hot-deck imputation program, called PRE CARP, which can be run prior to the PC CARP run. Unfortunately, this hot-deck imputation procedure is not always desired, especially in this evaluative situation, where comparability with other programs was important. If one chooses not to use the PRE CARP, PC CARP will treat missing values as zeros.

5.4 Execution Time

On the mainframe, CPU time varied widely among the packages. As one can see from Table 3, SESUDAAN required the least time (7 seconds for the SAIAN data, 11 seconds for the HHS data), and SUDAAN (using a text datafile) took the most time (71 seconds and 115 seconds). The other two packages were comparable to each other. The disparity between the two data files seems to be explained more by the number of observations than the number of strata and PSUs. Thus, the increase in the number of strata did not affect the execution time.

Table 3. Comparison of Mainframe Packages: CPU Time and Cost

	SESUDAAN	SUDAAN (using SAS datafile)	SUDAAN (using text datafile)	SUPER CARP
CPU time (in seconds)				
SAIAN data set	7.22	52.04	71.21	50.71
HHS data set	10.58	85.07	114.81	81.80
Cost (in dollars)				
SAIAN data set	11.09	110.75	147.57	55.82
HHS data set	16.16	166.53	217.51	82.82

N.B. The SAIAN data consist of 5584 observations and 11 strata. The HHS data (non-whites only) consist of 8310 observations and 71 strata.

On the PC, a different story emerges (Table 4). SUDAAN, regardless of datafile type, was significantly faster than PC CARP. Again, it must be stated that approximate run-time was recorded, from first keystroke to final result, since PC CARP is an interactive program and response and keying speed affect the overall time. PC CARP took more than seven times longer to execute (59 minutes for SAIAN data, 126 minutes for HHS) than did SUDAAN (8 minutes and 17 minutes), primarily due to the interactive nature of PC CARP. As with the mainframe packages, the disparity in time between the two data sets is more dependent on the number of observations than the number of strata.

5.5 Computing Costs

Mainframe computing costs for these runs were often substantial (Table 3). On the NIH mainframe computer, costs are a function of CPU time, region used, I/O count, and number of tape drives used. SESUDAAN was the least expensive to run, costing less than $20 for each dataset. SUPER CARP was next, costing $56 for the SAIAN data and $83 for the HHS data. SUDAAN cost at least ten times more than SESUDAAN, with the text datafile run being the most expensive ($148 SAIAN, $218 HHS). Number of observations seems to be accounting for the cost disrepancies within package for the two data files. Aside from the initial costs of hardware and software, there is no cost for the PC runs.

Table 4. Approximate Execution Times (in Minutes) for Microcomputer Packages

	PC SUDAAN (using SAS datafile)	PC SUDAAN (using text datafile)	PC CARP
SAIAN data set	8	8	59
HHS data set	17	17	126

5.6 Computational Accuracy

The estimates from the various packages were compared and resulted in exactly the same mean and proportional estimates as well as standard errors out to the available decimal places, usually at least four places after the decimal point. This is not surprising given that all of the packages evaluated use the Taylor approximation to compute variances. The only exceptions were found in four means and associated standard errors computed in SUPER CARP; two in the HHS data, two in the SAIAN data. However, even those means and standard errors agreed at an acceptable level (at least one place after the decimal). PC CARP means for these variables agreed with the SUDAAN packages, not with SUPER CARP. The design effects did not agree as precisely between the SUDAAN programs and the CARP programs, although they did agree at the integer level. It should be noted that the PC handled the same desired level of precision as was acquired on the mainframe.

5.7 Documentation

For the most part, the software documentation was quite good for all of the packages being evaluated. Examples are used to some extent in all of the manuals, and are quite helpful when one is using one of the packages for the first time. All contain algorithms for the available analyses, for those interested in the technical aspects.

The SESUDAAN manual is concise, but adequate. It has a SAS-manual-like style, with clear instructions on how to structure the program statements. It is lacking somewhat with respect to why and when to use the various commands and options.

The current SUDAAN manual is designed for use with the PC and the VAX, not for the mainframe. Although the program commands are the same for the mainframe, a mainframe section would need to be added to the manual to give information on Job Control Language and file-naming, and other details related to the interaction with the mainframe system. These latter issues caused the greatest problem when attempting to execute SUDAAN on the IBM mainframe. The existing SUDAAN manual, for its intended environment, is organized and clearly-written.

The SUPER CARP manual, although clearly written, is quite dated. It is written in terms of "punched cards", which can be interpreted as lines of code. The manual presumes some prior knowledge of FORTRAN and its data formats and the way it reads data files. Although the authors seem to have abandoned this package temporarily to develop PC CARP, some attention should be paid to updating the SUPER CARP

manual, since it still is useful in cases where a file is too big for the PC to handle efficiently and effectively. Some information on SUPER CARP's interaction with the mainframe (e.g., Job Control Language and file-naming/numbering) should be added as well.

The PC CARP manual is very well-written and demonstrates how to use the package primarily through the use of examples. Screen displays are shown throughout the examples.

6. SUMMARY

All of the packages evaluated for this paper proved relatively straightforward to use, after some practice experience. Most frustrating was the effort to correct problems which occurred in the interface between the mainframe computer and the mainframe software packages, due to the unhelpful error messages and the high computing costs even for jobs which did not run successfully.

One must be very knowledgeable about the dataset which is used for any of these packages prior to running the packages. While each package has different specification requirements, one should know in advance the number of observations in the file, the number of categories for each of the categorical variables, the existence of empty PSUs for the subfile being used, the variable lengths, and the extent of missing data. Data preparation must take place ahead of time, with missing values converted to the appropriate format, and sorting by nesting variables. If one is using a text datafile, any further changes to the data would require importation back into SAS or a database manager for modifications, then exportation back to a textfile.

In comparing the efficiency of the packages, one sees great variation in time, cost, and number of programming statements not only between the mainframe and personal computer packages, but between packages within each of the two environments. SESUDAAN is clearly the most efficient of the mainframe packages evaluated in terms of CPU time, dollars, and data preparation. One must keep in mind that SUDAAN is still in test mode and has not yet been optimized. SUPER CARP's performance was neither the best nor the worst with respect to the evaluation criteria. SUPER CARP's attraction, however, is its vast analytical capabilities. PC SUDAAN ran much more quickly than PC CARP, which required substantially more "statements" than PC SUDAAN, but was in a menu-driven format, which may be appealing to some users of this type of software.

The analyses that were run on the mainframe had to be run in several passes on the PC. The approximate run-times reflect the sum of the discrete times of each of the runs. PC SUDAAN ran out of memory when computing the proportions; therefore, the means of the 11 continuous variables were calculated in one run, the proportions of the first 12 categorical variables were the second run, and the proportions of the last 12 categorical (dichotomous) variables were the third run. PC SUDAAN is not currently equipped to make use of extended or expanded RAM. The package's authors are currently looking into this issue for future versions. PC CARP required two runs, one for the means and one for the proportions, due to the limitation of a maximum of 50 variables that can be read into the program. PC CARP is also unable to make use of more than 640kb of RAM, a restriction of the FORTRAN compiler. Although expanded memory is currently of no use with these two programs, a memory manager allowing for simultaneous tasks to be performed might make using a PC more palatable for this type of analysis. Although the actual PC SUDAAN runs did not take much time, the downloading of the data and importing into SAS did occupy the PC for a significant period of time.

The scope of this evaluation is limited in that only univariate statistics (weighted means and proportions) were computed. Although many of the analyses carried out on the NMES data only require these types of estimates, the packages have many more capabilities than were examined here. In addition, only one type of personal computer was used, and there are many other configurations being used among those who might use these packages. The evaluations are based on two datasets, with roughly 6,000 and 8,000 records. It appears that the limits have not yet been tested for these packages, with respect to file size. It should also be noted that SUPER CARP and PC CARP were not used as their authors intended, with respect to missing values. The use of missing value indicator variables was intended to get around their limitation of no missing data, but was in essence tricking the system. However, missing data often exist and using their imputation procedure, PRE CARP, might not always be appropriate.

Overall, it appears that using a PC for complex survey data analysis is certainly feasible, and may be desirable in many circumstances. One can look forward to future versions of these and other PC packages which will make even better use of the increasing capabilities of personal computers.

7. REFERENCES

Cohen SB, Burt VL, Jones GK: "Efficiencies in Variance Estimation for Complex Survey Data", The American Statistician, May 1986.

Cohen SB, Xanthopoulos JA, Jones GK: "An Evaluation of Statistical Software Procedures Appropriate for the Regression Analysis of Complex Survey Data", Journal of Official Statistics, 1988.

Edwards W and Berlin M (1989, September). Questionnaires and Data Collection Methods for the Household Survey and the Survey of American Indians and Alaska Natives (DHHS Publication No. (PHS) 89-3450). National Medical Expenditure Survey Methods 2, National Center for Health Services Research and Health Care Technology Assessment. Rockville, MD: Public Health Service.

Fuller WA, Kennedy W, Schnell D, Sullivan G, Park HJ. PC CARP. Statistical Laboratory, Iowa State University, Ames, IA, June 1989, version 1.3 (1988).

Hidiroglou MA, Fuller WA, Hickman RD. SUPER CARP. Survey Section, Statistical Laboratory, Iowa State University, 6th edition, October 1980.

SAS Institute, Inc. SAS User's Guide: Basics, Version 5 Edition. Cary, NC: SAS Institute, Inc., 1985. 1290 pp. (Release 5.18)

SAS Institute, Inc. SAS Language Guide for Personal Computers, Release 6.03 Edition. Cary, NC: SAS Institute, Inc., 1988. 558 pp.

Shah BV. SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC, April 1981.

Shah BV, LaVange LM, Barnwell BG, Killinger JE, Wheeless SC. SUDAAN: Procedures for Descriptive Statistics User's Guide. Research Triangle Institute, Research Triangle Park, NC, March 1989. (Used version 5.02 April 1990)

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