Analyzing survey Data with available Statistical software packages

Mohammad Ehsanul Karim <wildscop@yahoo.com>,
Institute of Statistical Research and Training (6th batch), University of Dhaka, Dhaka – 1000, Bangladesh

Before processing survey data, it is really important to know the type of data and analysis we have to handle according to our pre-specified aim / objective. Selecting appropriate Statistical package depends on what sort of analysis we are doing on what sort of data. Therefore first we need to fix detail rules relative to what we are trying to do and what specific techniques we might be looking to apply. However, sometimes the statistics package used most often comes down to personal preference. Here we discuss some packages with specific features of analyzing survey data.

(a) SAS

The biggest package in this space is probably SAS. SAS is more common than SPSS, and it is easier (although perhaps more expensive) to find decent SAS programmers. SAS is better than any other package at handling complex data structures (e.g. multiple time points, clustered data, large numbers o related and unrelated questions etc.). And SAS can do lots of statistics, and do them correctly. May be this is the reason why most of the articles in popular journals are analyzed with SAS. If we receive our data in the form of SAS datasets, then SAS is the obvious choice. If not, SAS has the ability to read in virtually any kind of data source we would possibly encounter, and transform it according to our will. It can handle huge (gigabyte) datasets that choke many other packages. And we can get free help from e-mailing support list SAS-L, some of the finest and smartest folks on the 'net.

However, all other things aside, there are a few things that other packages may be better at (or easier to use) - the scatterplot matrix in S-plus may be one example; we can do the same thing via various macros in SAS, but it's not as easy. Staggered-entry Kaplan-Meier survival curves are not easy to do in SAS either (proc lifetest doesn't accept different start dates). If the data and analysis are very simple, SAS may be overkill e.g., if our analysis is something like a poll of who one would vote for president, and all we want to report are the percent saying "Bush", 'Clark', 'Dean', etc.... then we don't need SAS.

SAS still doesn't have survey sample versions of PROC FREQ and PROC LOGISTIC ready for release - they're scheduled for 9.1 (although PROC SURVEYFREQ is experimental in 9.0). Other stat packages already have these. SAS implements a

SURVEYLOGIT procedure in 9.1, SAS will be missing a SURVEYMLOGIT procedure (for fitting generalized or cumulative logit models) that SUDAAN offers. The ability to perform tests on linear combinations of the parameters is offered in MIANALYZE with SAS version 9.0, but not with previous versions. SAS does Taylor series linearization for error estimates, too. SAS is too much expensive for single user or small organizations. Some suggests that if we want to analyze complex surveys in SAS, then we should buy the SAS add-on version of SUDAAN for the cost considerations.

If we want to focus on sample surveys, we have to define the type of survey, and the type of analysis. SAS makes its error estimates based on the primary sampling units, and variation between observations within PSUs. For more complex analyses, say, a generalized additive model using sample survey data, SAS is at least as good as the other options. We can build a variance-covariance model using SAS/IML or a DATA step, and then use that matrix when fitting the model. The high-end sample survey packages aren't designed to handle every possible statistical analysis. SAS sample survey PROCs can work with samples even if they are based on continuous - population extensions of the classical discrete-population sampling theory. Samples such as the adaptive samples developed by Steve Thompson can be analyzed, if we can work out the sampling weights for ourselves.

To summarize, SAS has some of the fanciest file-reading abilities, and it has some of the most detailed report-writing abilities, including fancy control of the printer. Also, it has been pretty fast to install new statistical procedures applicable for surveys.

(b) STATA

STATA's strength is that we get a comprehensive stats package, including survey facilities, in one powerful, reasonably-priced, sanely-licensed package written by a highly responsive, user-oriented company. It is a general purpose package, not a single purpose, dedicated survey package.

STATA includes a number of commands designed to handle the special requirements of complex survey data. The commands will handle any or all of the following survey-design features: probability sampling weights, stratification, and cluster sampling. There are commands for estimating means, totals, ratios, and proportions, and commands for linear regression, logistic regression, probit models, and survey estimators for sampling designs. Variance estimates are produced using Taylor-series linearization methods. Finite-population corrections for simple random sampling without replacement of primary sampling units (PSUs) can optionally be computed. Variance estimation for multistage sample data is carried out through the customary between-PSU-differences calculation. In addition, many other estimation commands in STATA have features that make them suitable for certain limited survey designs. For example, STATA's Cox regression routine (stcox) handles sampling weights properly when sampling weights are specified, and it also handles clustering. The svyset command allows us to set the variables that contain the sampling weights, strata, and any PSU identifiers at the outset. Estimating the difference between two subpopulation means can

be done by running svymean with a by() option to produce subpopulation estimates, and then running the command lincom. The svymean, svyprop, svyratio, and svytotal commands will produce estimates for multiple subpopulations. Options give the user control over what is displayed in the output. We can fit linear regressions, logistic regressions, and probit models using svy estimators. svylogit can display estimates as coefficients or as odds ratios. After we run regression, we can use lincom to compute odds ratios for any covariate group relative to another. Linear regressions, logistic regressions, and probit models can also be fit for a subpopulation.

Survey data have some special data-management needs. The svydes command can be used to examine the strata and PSU structure of the dataset. It can also be used to see the number of missing and nonmissing observations per strata (or optionally per PSU) for one or more variables. Programmers can use the _robust command to compute survey design-based variance estimates for their own estimators. For example, a maximum likelihood estimator can be implemented using STATA's ml optimizer, and then the _robust command can be used to compute appropriate variance estimates for survey data. STATA's survey facilities will get us through almost any survey analyses, but it would be somewhat unreasonable to expect STATA to do everything that the best single purpose survey packages can do. If we need package for complex survey methods, then official STATA does not include replication-based methods (which SUDAAN and WesVar do). However, -svr- package adds BRR and Jackknife capabilities.

STATA as a mix of provided commands and a programming language gives an excellent balance for researchers. STATA books are nice and readable. SPSS has 50 routines which seem to have been written by 50 different people each of whom had their own unique ideas about what syntax and features should be like. STATA is far more internally consistent in its commands. STATA's commitment to quality control is higher. Find one small mistake or imperfection and they want to know and will correct it as fast as possible. In the short, middle and especially long run this really counts. Some statisticians says : "Use SPSS for recodes; use STATA for everything else!" STATA seems faster than SPSS, at least for some tasks. That is because it stores so much in memory. On the other hand, sometimes, data sets that are so monstrous, we can't get them to work in STATA even though SPSS can slowly but surely tackle them. This basically depends on system requirement.

STATA's graphs were comprehensive in earlier versions. But with much improved graphics in STATA v8, users are quite contented now with STATA package only. The new STATA v8 also offers some more features for survey analysis:

1. STATA's ml user-programmable likelihood-estimation routine has new options that automatically handle the production of survey estimators, including stratification and estimation on a subpopulation.
2. Survey estimation is now available for the Heckman selection model and the Heckman selection model applied to probit.
3. Survey estimation is now available for negative-binomial regression and generalized negative-binomial regression.

4. Constraints may now be applied to equations using survey estimators in the same way as with STATA's other estimators.

5. Point estimates, standard errors, and confidence intervals are now available for linear combinations of estimated parameters using the same procedure as with STATA's other estimators.

6. Point estimates, standard errors, and confidence intervals are now available for nonlinear combinations of estimated parameters.

7. Nonlinear combinations estimators and generalized predictions are available.

(c) SPSS

SPSS is extremely popular among people doing "social statistics": sociologists, political scientists, psychologists etc. A spectrum of analyses offered by SPSS is standard and sufficient for most applications in those fields. Unfortunately, as many other statistical packages comparable to SPSS, its flexibility is very limited. If something isn't already built in the package, we cannot do it easily. Until very recently, SPSS had no survey facilities at all and was very frustrating for complex survey analysis, especially with regard to weights. As of this month, SPSS has released an add-on package for the new SPSS 12.0 called Complex Samples. However, the cost to license this add-on package alone is very high. Some statisticians use SPSS for survey data.

SPSS's help and user interface is better; on the other hand, SPSS has to be better because there is no way anybody could remember all of the wildly differing syntax across commands. The syntax of SPSS is consistent; professionals like to keep old command files created on other systems to use them again. Also, SPSS used to be much better about display.

It can be used on many different types of computers and operating systems. It was a very good point for SPSS to be available on many different systems (mainframes), SPSS is quite good for big data sets. If we have a lot of multiple response variables or the need to quickly analyze subgroups, usually SPSS is recommended, also SPSS is one of the industry standards.

For the options "variable label" and "value label" in SPSS, it is very easy to use and quite useful for survey data. Many appreciate recoding which is simple and powerful, there are quite a lot of useful (and easy to use) commands for data manipulation.
Error and warning messages are often easy to understand and informative.
Documentation is well written, in English. SPSS is very easy to use especially on micro with menus. It is so easy that many non statistician research people using SPSS without knowing what they do. Statistical consultants frequently receive students with results produced with SPSS asking "What I have done ?".

(d) S-plus / R

Graphics in general seem to be easy in S-plus. R is parallel version of S-plus which seems to have a lot of handy macros available. R is available for download and

fully functional use at no cost. R is nice because it is powerful and free, but it is different enough to render comparison with other packages difficult. There is an add on package called 'survey' written by Thomas Lumley that offers a set of techniques for that particular domain. It just came out in version 2 this past August,2003. The survey Package is with title: “analysis of complex survey samples”. Its functions are: Summary statistics, generalized linear models, and general maximum pseudolikelihood estimation for stratified, cluster-sampled, unequally weighted survey samples. Variances by Taylor series linearization or replicate weights. It favors a GLM approach to survey analysis, and implements replications and jackknife estimates. And this new survey package is quite general and is an excellent model for how new algorithms for complex sample survey analyses can be added to R.

(e) MATLAB

Some statisticians prefer MATLAB while building custom statistical tests, such as implementation of custom models, bootstrapping, or unusual equation fitting. This is may be because the programming interface for such things are fairly straightforward. However, packages like SAS, SPSS and NCSS have considerably large libraries of statistical routines available, and have routines to deal with all types of anomalies (like missing data, unequal sample sizes, etc.). MATLAB, even with the Statistics toolbox, is no nearly as complete a package. Some prefer MATLAB's visualization and data transformation capabilities, but that is probably more opinion and familiarity than specific features.

(f) SUDAAN

SUDAAN uses a SAS-like language. There are two versions of SUDAAN with different data interfaces:
a. "SAS-callable" SUDAAN: SUDAAN is called directly as a SAS procedure.
b. "Standalone SUDAAN": Independent program that reads external file formats, including SAS files or SPSS files.
In either case, the same programming language is used. Release 8.0 of SUDAAN is the current version. Some say that SUDAAN is very powerful but inflexible. Some think the documentation is poorly written but we can get by with what's there.

Types of designs that can be accommodated with SUDAAN are Multiple design options allow users to analyze data from stratified, cluster sample, or multistage sample designs. Sample members may have been selected with unequal probabilities, and either with or without replacement. Any number of strata and stages can be specified. In addition, different design options may be combined in one study if different sampling methods were used for parts of the population.

The Taylor series linearization method (GEE for regression models) is used combined with variance estimation formulas specific to the sample design. The user does not need to develop special replicate weights since the sample design can be specified directly to the program. Jackknife and Balanced Repeated Replication (BRR) variance estimation is also supported.

(g) WesVar

Types of designs that can be accommodated are:

1. All variance estimates are based on replicate weights, either generated within the program user-provided.

2. Stratified or unstratified, single- or multistage designs. Finite population corrections can be accomodated. 2/stratum designs using BRR or Jackknife, >2/stratum using Jackknife.

3. If replicate weights are generated within the program, external control totals may be provided to be provided to perform post-stratification or raking of the weights, and nonresponse weighting adjustment.

4. Multiply-imputed datasets can be analyzed.

(h) Epi Info

Free; includes some ability to specify complex designs, but only rather limited analyses are possible. It provides for easy form and database construction, data entry, and analysis with epidemiologic statistics, maps, and graphs. Types of designs that can be accommodated with Epi Info (CSAMPLE procedure) are Stratified sampling, with or without clustering; multistage samples; unequal-probability (e.g. pps) samples. Does not calculate finite population corrections, so either sampling fraction must be small or sampling must be with replacement.

(i) CENVAR

Developed by U.S. Bureau of the Census (International Programs Center) and freely distributed. Sample designs ranging from simple random samples of elements to more complex stratified, multistage cluster designs are its specialty.[1]

---

[1] I am indebted to the following people for directions, instructions, suggestions, comments, references for writing this documentation. Without their help it was not possible for me to write this paper in such a short time. Thanks to all of them.

Ken\ B. Water\ Zalek Bloom\ Mark Lamias\ Stas Bekman\ MichalB\ Wuzzy\ Rich Ulrich\ Stan Brown\ DePuy, Venita\ Frank E Harrell Jr\ Jim McGowan\ Marc Schwartz\ Joseph Saint Pierre\ Arthur Tabachneck\ Will; WMB; Statistical Services\ Richard Williams, Associate Professor\ Nick Cox; Executive Editor, STATA Journal\ Dale McLerran; Fred Hutchinson Cancer Research Center\ Lee Sieswerda, Epidemiologist; Thunder Bay District Health Unit\ Stephen Soldz; Director of Research; Health & Addictions Research\ David Cassell, CSC; Senior computing specialist; mathematical statistician\ Nicholas Winter; Assistant Professor; Department of Government; Cornell University \ Bob Abelson; KAI Research, Inc.; 6001 Montrose Rd.; Suite 920; Rockville, MD 20852\ Marianne Mueller; Institute of Social & Preventive Medicine; University of Bern, Switzerland\ Peter L. Flom, PhD; Assistant Director, Statistics and Data Analysis Core; Center for Drug Use and HIV Research; National Development and Research Institutes; 71 W. 23rd St; New York, NY 10010\

Also, if readers have any comments, corrections, updates; the author can be reached at any of the electronic mail addresses provided here - <wildscop@yahoo.com> or <ehsan@isrt.ac.bd> with appropriate words in the subject field. This should be noted that this stuff was first documented at Friday, October 31, 2003 and certain parts of this document may be subject to change at any later period.