THE USE OF OSIRIS IV TO ANALYZE COMPLEX SAMPLE SURVEY DATA

James M. Lepkowski, The University of Michigan

1. Introduction

Sample surveys, with complex sample designs involving multistage selection and weighting, are often conducted to provide estimates of descriptive measures for a finite population. Increasingly, complex survey sample data are being used in analyses that may require estimates that are not descriptive nor simple in form. The application of analytic techniques to sample survey datasets has raised a number of questions about the appropriate analytical strategy to be followed. The controversy has created confusion for the survey analyst, not only about appropriate strategies, but also about the availability of statistical software to conduct recommended analyses.

The OSIRIS IV Statistical Software System (Survey Research Center Computer Support Group, 1981) offers a variety of statistical software designed for the analysis of survey datasets. Two programs are designed to provide estimates of sampling error for estimates calculated from complex sample survey data. The programs were developed from a classical design-based analysis perspective in which sampling errors are estimated to assess the reliability of survey findings. The programs are flexible allowing the incorporation of sampling errors into analytic strategies for the analysis of complex sample survey data.

The subsequent discussion is a description of the two OSIRIS sampling error programs and the strategies for their design. Two analytic techniques which incorporate the sampling error estimates or other output from these programs directly into an analysis are also considered. The next section briefly reviews the nature of complex sample surveys and the controversy surrounding analytic use of survey data. Section 3 summarizes features of the OSIRIS IV System, including those features useful for the management of survey datasets. The two OSIRIS IV sampling error programs are described in Section 4, and applications of these programs to two analysis problems are given in Section 5. Several additional issues concerning the design of sampling error programs are discussed in Section 6.

2. Complex Sample Surveys and Analytic Perspectives

The term "complex sample survey" is not a well-defined expression, although survey designers and analysts have a common understanding about the general features of such a design. The surveys are typically have large samples (i.e., at least 1,000-1,500 sample elements) selected from large populations. They have multiple purposes (i.e., more than a single characteristic is being investigated) and usually utilize randomized sample selection. The sample designs often employ multistage selections and include features such as stratification of sampling units and unequal probabilities of selection. The final survey dataset is large (i.e., many records and a large number of variables), and it may include some type of sampling weights and special codes associated with each record for the computation of sampling errors.

The consumer of survey data is confronted with two perspectives on the analysis of the survey dataset: design-based or model-based statistical inference. There is considerable debate about the appropriateness of these analysis perspectives with respect to the analysis of data from a complex sample survey (see, for example, Kish and Frankel, 1974, and Smith, 1976).

The design-based perspective incorporates the survey sample design directly into the analysis. An implicit assumption of the design-based perspective is that the finite population from which the sample was selected reflects an underlying structure that is appropriate and important to any inferences drawn from the data. As a result, features of the sample selection such as stratification, clustering, and weighting are incorporated directly into the analysis of survey data.

The model-based perspective, on the other hand, starts with a statistical model. No implicit assumptions are made about the appropriateness or importance of the finite population structure. An analysis under a model-based perspective is conditional only on the data and the model, and the model is modified only when the data demonstrate the need to do so. Randomized selection and other sample design features are considered irrelevant to the analysis (see Kalton, 1981, for a discussion of these issues).

The computation of sampling error estimates for complex sample surveys arises from and gives rise to a design-based approach to statistical inference. Features of the complex sample design are incorporated into sampling error estimation procedures. It is implicitly assumed that the design reflects features of the finite population, and, hence, is informative. The OSIRIS IV sampling error facilities reflect this design-based perspective for complex sample survey data.

3. The OSIRIS IV Statistical Software System

In order to understand the complete capabilities of the OSIRIS sampling error programs, it is useful to review the history and basic features of OSIRIS. The OSIRIS IV system is the most recent version of a package of data management and statistical analysis software that has been under development at the Institute for Social Research at The University of Michigan since the late 1950´s. The first versions of an OSIRIS system (OSIRIS 40 and OSIRIS II) were developed by 1967, and the widely distributed OSIRIS III system was produced in 1973. Further developments in data structures and the use of an interactive, keyword format command environment lead to the development of the current version of OSIRIS in 1981 (Computer Support Group, 1981).

OSIRIS IV has software for data management that includes copying and sorting data, creating OSIRIS type datasets, data cleaning (i.e., wild code and consistency checking), merging, and data modification and recoding. Two data structures are available: the standard rectangular dataset and the hierarchical dataset, a more compact form of storage for some types of survey data. OSIRIS analytic facilities include frequency distributions, correlation and regression analysis, analysis of variance, multivariate analysis with ordinal and nominal predictors, factor analysis and multidimensional scaling, cluster analysis, and sampling error estimation.

OSIRIS IV programs can be accessed interactively, or by batch processing, through the OSIRIS monitor, a control system which calls the data management or analysis software that the user specifies. Program control statements use a keyword format and the data records are accessed and computations performed sequentially. This sequential processing eliminates the need for large computer "core" storage requirements during the program operation, and facilitates the processing of large datasets, often reducing the costs of processing. Case selection can be achieved through the application of filters, and weighted data can be handled by many OSIRIS programs. The system also has facilities for specification and use of missing data values, a variety of data storage modes (e.g., character, floating point binary), and interactive error processing.

2. Estimation of Sampling Errors

Two OSIRIS programs are available for the estimation of sampling errors: PSALMS and REPERR. The PSALMS program (an acronym for "Paired Selection Algorithm for Multiple Subclasses") is designed to estimate sampling errors for ratio means, means of subclasses, and differences between subclass means. The REPERR program (Repeated Replications for Sampling Errors) estimates sampling errors for regression statistics such as regression and correlation coefficients. Both programs are revised versions of programs contained in the Sampling Error Program Package developed and distributed by the Survey Research Center in the early 1970's (Kish, Frankel, and Van Eck, 1972).

Sample designs for complex sample surveys are typically so complicated that to provide formula for, let alone direct estimates of, sampling errors for many different survey estimators would be a difficult and expensive task. Assumptions are often made to simplify the statistical and computational task in order that manageable software can be written, processing costs may be reduced, and sampling errors calculated routinely from survey data. The PSALMS and REPERR programs use two basic assumptions to simplify the sampling error estimation task.

First, it is assumed that two or more primary sampling units have been selected from each sampling stratum. Stratified multistage sample selection is not essential since even a simple random sample of elements can be considered a selection with the elements as primary sampling units drawn from a single stratum comprising the entire population. Complex samples are frequently selected in which only one primary sampling unit is selected from a stratum such as in a controlled selection of primary units. A collapsed stratum technique is employed for such designs to allow variance estimation in which "neighboring" strata are collapsed into a single stratum with at least two primary selections (see Cochran, 1963). It is believed that collapsing strata leads to slight overestimates of the actual variance, since the collapsing procedure does not fully recover additional gains due to the "deeper" stratification actually employed in the selection.

A second assumption is that the primary selections are made independently or with replacement. In practice, with replacement sampling is not often used at a primary stage of selection. On the other hand, sampling error estimation formulae for without replacement sampling can be quite difficult to use in routine processing. For samples using without replacement selection at the first stage, the complexities of without replacement sampling error formulae are avoided, and sampling error estimates are calculated using the simpler with replacement formulae. This use of with replacement formulae leads to an overestimate of the actual variance.

A variety of estimation techniques can be used to estimate sampling errors for multistage sample designs under these assumptions. Simple replicated methods using squared differences among totals for first stage selections, or simplifying forms of those squared differences such as paired selection designs (Keyfitz, 1957; Kish and Hess, 1959), are often used. Deming (1960) suggests combining strata or "thickening zones" to simplify variance calculations, while the method of random groups has also been advocated (Hansen, Hurwitz, and Madow, 1953). Repeated replication methods, including balanced half samples (McCarthy, 1966) and jackknifing (Kish and Frankel, 1970) are another practical and flexible approach to variance estimation.

These techniques must be applied in a typical sample survey to a variety of estimates, some of which may be nonlinear functions of the sample values. The repeated replication techniques are appropriate for variance estimation of nonlinear functions. Although they provide biased variance estimates for nonlinear estimators, the bias is generally considered to be small.

A second approach is the linearization of the estimate through a first order Taylor series expansion and subsequent use of a large sample approximation to the variance. The Taylor series approximation is known to perform well for variance estimates of ratio means, but it can be complicated to implement for other statistics, and it may be subject to substantial bias for certain distributions of the data.

The OSIRIS IV sampling error programs use these two variance estimation techniques for computing sampling error estimates for nonlinear estimators. For sampling errors of ratio means, the PSALMS program uses a first order Taylor series approximation to the variance. The approximation for ratio means presents a relatively straightforward programming task, and it provides good estimates of variance. In addition, the approximation is readily adapted to variance estimation for subclass means.

For other nonlinear estimators, the repeated replication variance estimation technique is often preferred because of its simple form for even the most complicated of estimators. The REPERR program employs repeated replication techniques, including both the balanced half-sample and the jackknife methods, to estimate variances for regression statistics. The balanced repeated replication procedure (BRR) in REPERR can be used for designs with exactly two primary units per stratum, and either fully balanced designs (i.e., the number of strata is a multiple of four) or partially balanced designs (i.e., the number of strata is not a multiple of four). Designs with two or more primary units are handled by the jackknife repeated replication option (JRR) or by a user-specified repeated replication scheme.

In order to implement both PSALMS and REPERR, a specification of sampling strata and sampling error computing units must be available for every sample element in the dataset. Stratification information is usually available directly from the design description, and it may involve collapsed strata or, in unstratified designs, a single stratum for the entire sample. The sampling error computing unit, or SECU, corresponds in many designs to the primary selection, but it may also represent a grouping of several primary selections to reduce the number of distinct units to be used in the estimation process.

For example, in area probability sample surveys, some primary selections will enter the sample with certainty, and these are designated as self-representing units. Since they are "selected with certainty," the selection of self-representing units does not contribute to the variance of estimates, and the self-representing units are no longer considered to be primary sampling units. The first stage selections in self-representing units often consist of a large number of blocks or clusters of population elements. Rather than using each selected block or cluster as a separate primary selection or SECU for sampling error estimation (an expensive computing task), these primary units are grouped into SECU´s within self-representing units to create two or more SECU´s per self-representing unit. The self-representing unit is then treated in variance estimation as a separate stratum with two or more SECU´s. Such a procedure, of course, reduces the precision of estimates of variance, but the loss in precision is more than justified by the substantial reduction in computing costs.

In addition to the specification of strata and SECU´s for a sample, the PSALMS and REPERR programs utilize similar keyword syntax to specify program features. Weighted analyses are allowed through the use of a weight variable (different weights for a numerator and denominator of a ratio mean can be handled through an appropriate recoding of the data). Temporary recoding of variables during program operation, filters to select a subset of records, retrieval of records from structured datasets, and specifications for bad data items (e.g., illegal characters) are handled through the same standard OSIRIS IV conventions for each program. There are, of course, important differences between the programs.

PSALMS

The PSALMS program employs three basic models for sample designs to simplify the sampling error calculations. The paired selection model requires exactly two SECU´s in each stratum in the dataset; the Keyfitz (1957) estimation formulae are applied under this model. For systematically samples of primary units selected from a deliberately ordered list, a successive differences model may be used, where it is assumed that there are at least two SECU´s per stratum and the ordering of the list places units similar to one another next to each other in the list. A multiple differences model is also available which allows at least two SECU´s per stratum. The program allows combinations of the models to be specified for the different strata, thus providing considerable flexibility for handling a variety of sample designs.

The estimated means for which PSALMS computes sampling errors can be based on the entire sample as well as for two types of subgroups of the population. The first subgroup type is termed a domain, and domains are defined in terms of groups of strata. Estimated domain sampling errors are computed for the subgroup of records in the designated strata. The second type of subgroup is referred to as a subclass, a subgroup of the population which typically crosses stratum and SECU boundaries. For example, racial categories which cut across geographically defined strata and SECU´s may be used to create separate subclasses for which estimates are desired.

The specification of the estimate for which sampling errors are desired requires a numerator and denominator for each estimate, or two numerators and two denominators if the sampling error for the difference between two ratio means is desired. Subclasses for each numerator and denominator are also needed (the default subclass specification is all records passing the filter). The denominator is typically a count variable indicating the number of cases in the base of the mean. The numerator may be a continuous or count measure, or it may be an indicator variable used to obtain counts of particular subpopulations to obtain the numerator for computing a proportion.

The PSALMS program provides two types of printed output. The summary output displays on a single line for each estimate or difference estimate the estimate itself, the estimated standard error, the square root of the design effect (i.e., the ratio of the estimated standard error to the standard error computed under the assumption that a simple random sample of the same size was used to select the sample), a measure of intracluster homogeneity (labelled "roh"), and an indicator or flag for estimates with a coefficient of variation of the denominator exceeding 0.15. The full output provides detailed results for each estimate including variances and covariances of numerators and denominators, weighted and unweighted totals, and the covariance between the individual means in differences of means.

An undocumented output option allows the full output to be placed in machine readable form in a separate file for later processing. This option is particularly convenient for summarizing sampling error results or using them as input to other analytic programs.

Missing data in the numerator or denominator for a particular record causes the record to be deleted from estimation for that estimate. If, through definition of a subclass, or by design, a SECU has no observations, a zero is substituted in estimation formulae for the estimate. Strata without observations are simply not included in the calculations. The program does not estimate components of variance, and finite population corrections are not included in the estimated sampling variances.

REPERR

The REPERR program provides for two basic sampling error estimation methods, BRR and JRR. The BRR method requires exactly two SECU's per stratum and allows up to a total of 88 strata (176 SECU's). Designs with more than two SECU's in some or all strata may use the JRR method in REPERR, or users may consider procedures such as combining strata or, through the user specified replicate option, specifying their own replication scheme. Although the JRR option allows an unlimited number of SECU's, computational costs may rapidly become excessive with a large number of SECU's. Some combining of SECU's, with consequent losses of precision for variance estimates, is recommended. The user specified replicate option may also be used to estimate sampling errors for simple replicated sample designs, as well as other designs.

The regression problem is specified in REPERR exactly as in the OSIRIS regression analysis program REGRESSN. A correlation matrix and estimates are computed for each replicate and for the total sample for the specified problem. Variances for means, regression coefficients, standardized regression coefficients, correlation coefficients, and multiple correlation coefficients are computed as a function of the sum of the squared difference between replicate estimates and total sample estimate. No subclass specifications are provided in the program set-up, although an appropriate filter may be provided to select a subclass of interest for regression analysis.

Printed output for REPERR includes the estimates, their standard errors, and the square of the design effects, all presented in a format similar to the REGRESSN program output. Optional printout includes regression analyses for each replicate, sums and sums of squares and cross-products by replicate, and the definition of the replicates used in the calculation.

Special input and output provisions allow the computation of sampling errors for estimates other than for regression statistics. The replicate sums and sums of squares and cross-products or the replicate regression estimates may be written to an external output device. They may also serve as the program input. Thus, replicate sums of squares and cross-products may be computed and stored on one analysis for a large number of variables and used as input to later runs using subsets of variables for specific regression analyses. Replicate estimates other than for regression analysis may be used as input to the program, perhaps computed from replicate sums of squares and cross-products, to obtain sampling errors for other statistics.

Missing data can be handled in two ways in REPERR. A case-wise deletion deletes an entire record or case if it is missing on any one of the dependent or independent variables specified in the regression analysis. Variable-wise deletions remove only the missing value of a single variable prior to computing sums and sums of squares, or a pair of variables for sums of cross-products. Empty SECU's or strata are handled in the same way as in the PSALMS program.

The similarities and differences between the PSALMS and REPERR programs are summarized in Figure 1.

5. Analytic Applications

THE PSALMS and REPERR programs have been used in a variety of analytic contexts other than the simple estimation of sampling errors.[1]

The Weighted Least Squares method for the analysis of categorical data using linear models described by Grizzle, Starmer, and Koch (1969) has become an increasingly useful strategy for the analysis of contingency tables and other data. Koch and Lemeshow (1972) and Koch, Freeman, and Freeman (1976) have demonstrated the use of this strategy for the general analysis of complex sample survey data using linear models.

Essentially the method fits a linear model to functions or estimates derived from survey data. The functions may be formulated through a compounded series of linear, exponential, or logarithmic operators that are applied to various estimates from the survey data. A linear model is specified for the final set of functions or estimates, and the parameters of the model are estimated and the goodness of fit of the model to the data is evaluated using Weighted Least Squares methods.

For data from a complex sample survey, the method allows the incorporation of the survey design features directly into the parameter estimation and model fitting procedures. For example, a vector of estimates such as proportions from a contingency table may be computed from a survey dataset where the proportions are computed separately for several subclasses. The estimated variance-covariance matrix for the vector can be computed from the survey data as well by using the PSALMS program, and thus incorporating the survey design features into the estimated variances and covariances. This vector and variance-covariance matrix may then be used as the initial function of data values to which a linear model is fit in the Weighted Least Squares methodology.

Lepkowski (1980) and Lepkowski, Bromberg, and Landis (1981) describe the use of the PSALMS program as the computational tool to provide estimates of the required vector and variance - covariance matrix. The estimates and estimated variances and covariances can all be obtained in PSALMS through specification of subclass means or differences between subclass means (covariances for two means are computed and displayed on the full output when a difference of subclass means is specified). If the full output is written in a machine readable format to an external file, the appropriate estimates can be read into a computer program that formats the estimates and estimated sampling errors so that they can be used as input to a Weighted Least Squares analysis program such as GENCAT (Landis, Standish and Koch, 1976). The subsequent model specification and analysis based on the estimates from PSALMS take full account of the complex sample design in the analysis, including the effects of stratified multistage selection, weights, and covariance between subclasses due to clustered sample selection.

[1] One of the most frequent applications is to provide the data for summary models of sampling errors for various survey estimates using the sampling errors themselves (Gonzalez, et. al., 1975) or design effects and intracluster homogeneities (Kish, Groves, and Krotki, 1976).

The REPERR input and output facilities may also be used to provide specialized design-based analytic capabilities. For example, in a recent application at the Survey Research Center it was of interest to examine the sampling error of differences between regression coefficients computed for nonoverlapping subclasses, a problem similar to the analysis of covariance. The program REPERR was used to estimate regression coefficients and other statistics for two subclasses by running the program twice and using filters to pass only appropriate subclass records to the program each time. The replicate estimates were written to separate external files for each subclass. Another program called REPVAR, and currently available only at the University of Michigan, was written to run under the OSIRIS monitor to read the separate subclass replicate estimates, to compute differences of estimates for each replicate, and to compute replicated variance estimates for the subclass differences of regression coefficients.

Other similar application of REPERR can be formulated, as well as the use of the REPERR output option of sums and sums of squares and cross-products for selected variables.

6. Discussion

The design of computer software for the analysis of complex sample survey data requires consideration of many factors not encountered with other types of data. The sheer size and complexity of many survey datasets requires the availability of data management facilities and the use of computing strategies that can process large numbers of cases. Development of special purpose programs for data management is expensive, and may limit the ability to provide datasets that are readily transmitted to other computing facilities. Sequential processing of records from large survey samples may be the only alternative to storing entire datasets, or substantial portions of them, in computer "core" storage during processing.

Given the complexity of many survey designs, practical methods of providing fairly accurate estimates of sampling error for a variety of survey designs are needed. Approximations and assumptions of the type described previously, with consequent small overestimates of sampling error, are acceptable when the simplicity of programming requirements and reduced processing costs are considered.

Other considerations in the design of computer software for complex sample data analysis include the ability to assess threats to the validity of approximations and the flexibility of the program to provide estimates or other input to various analytic applications. The validity of the Taylor series approximation to the variance of a ratio mean can be examined directly from the sample data. For other nonlinear statistics, the assessment of the approximation´s validity is not as straightforward, and other, simpler methods of variance estimation are available. Program flexibility is illustrated by the REPERR facility for reading as input and writing as output replicate sums and sums of squares and cross-products or the replicate estimates. Taking advantage of the program´s flexibility allows users to perform other analyses without additional major programming requirements, and still have access to the flexible and robust repeated replication estimation procedures used in REPERR.

The present OSIRIS sampling error estimation facilities have developed historically to address problems of sampling error estimation and summarization for large scale sample surveys. The design-based perspective has not been extended computationally to analytic strategies such as the analysis of linear models for complex sample survey data. Further developments may lead to an integrated sampling error estimation and linear model fitting program in OSIRIS as well. A monograph documentary the features and applications of PSALMS and REPERR is also planned.

References

Cochran, W. (1963). Sampling Techniques. New York: Wiley and Sons, Inc.

Deming, W.E. (1960). Sample Design in Business Research. New York: Wiley and Sons, Inc.

Gonzalez, M., Ogus, J., Shapiro, G., and Tepping, B. (1975). "Standards for Discussion and Presentation of Errors in Survey and Census Data," Journal of the American Statistical Association, 70, part II.

Grizzle, J., Starmer, F. and Koch, G. (1969). "Analysis of Categorical Data by Linear Models," Biometrics, 25: 489-503.

Kalton, G. (1980). "Practical Methods for Estimating Survey Sampling Errors," paper presented to the 42nd Session, International Statistical Institute, New Delhi, India.

Kalton, G. (1981). "Models in the Practice of Survey Sampling," invited paper presented to the 43rd Session, International Statistical Institute, Buenos Aires, Argentina.

Keyfitz, N. (1957). "Estimates of Sampling Variance Where Two Units Are Selected from Each Stratum," Journal of the American Statistical Association, 52: 503-510.

Kish, L. and Frankel, M. (1970). "Balanced Repeated Replications for Standard Errors." Journal of the American Statistical Association, 65: 1071-1094.

Kish, L. and Frankel, M. (1974). "Inference from Complex Samples," Journal of the Royal Statistical Society, Series B, 36: 1-37.

Kish, L. and Hess, I. (1959). "On Variance of Ratios and Their Differences in Multistage Samples," Journal of the American Statistical Association, 54: 416-446.

Kish, L., Frankel, M., and Van Eck, N. (1972). SEPP: Sampling Error Program Package. Ann Arbor: Institute for Social Research, University of Michigan.

Kish, L., Groves, R., and Krotki, K. (1976). Sampling Errors for Fertility Surveys. Occasional Paper No. 17. London: World Fertility Survey.

Koch, G. and Lemeshow, S. (1972). "An Application of Multivariate Analyses to Complex Sample Survey Data," Journal of the American Statistical Association, 67: 780-782.

Koch, G., Freeman, D., and Freeman, J. (1975). "Strategies in the Multivariate Analyses of Data from Complex Sample Surveys," International Statistical Review, 43: 59-78.

Landis, J., Stanish, W., and Koch, G. "A Computer Program for the Generalized Chi-Square Analysis of Categorical Data Using Weighted Least Squares (GENCAT)," *Computer Programs in Biomedicine*, 6: 196-231.

Lepkowski, J. (1980). "Design Effects for Multivariate Categorical Interactions," unpublished Ph. D. dissertation, University of Michigan.

Lepkowski, J., Bromberg, J., and Landis, J.R. (1981). "A Program for the Analysis of Multivariate Categorical Data from Complex Surveys." *Proceedings of the Section on Statistical Computing of the American Statistical Association*, pp. 8-12.

McCarthy, P.J. (1966). "Replication: An Approach to the Analysis of Data from Complex Surveys," *Vital and Health Statistics*, Series 2, No. 14. U.S. Department of Health, Education, and Welfare. Washington: U.S. Government Printing Office.

Smith, T.M.F. (1976). "The Foundations of Survey Sampling: A Review (with Discussion)," *Journal of the Royal Statistical Society*, Series A, 139 (2): 183-204.

Survey Research Center Computer Support Group (1981). *Osiris IV User´s Manual, Seventh Edition*. Ann Arbor: Institute for Social Research, University of Michigan.

Figure 1.
Comparison of Features of the OSIRIS Sampling Error Programs PSALMS and REPERR.

	Program	
Feature	PSALMS	REPERR
Estimators	-Means,Subclass Means -Differences of Subclass Means	-Regression Statistics
Subgroups	-Domains -Subclasses	(None)
Estimation Technique	First Order Taylor Series Approximation	Repeated Replications
Estimation Methods	-Paired Selections -Successive Differences -Multiple Selections	-BRR -JRR -User Specified
Missing Data	Case-wise	Case-wise or Variable-wise
Input	-OSIRIS Datasets (Rectangular or Structured)	-OSIRIS Datasets (Rectangular or Structured) -Vector of Replicate Estimates and Labels
Printed Output	-Summary -Full	-Summary -Replicate Regressions -Replicate Specifications -Replicate Sum of Squares and Cross Products Matrix
Machine Readable Output	-Full Output (undocumented)	-Replicate Sum of Squares and Cross Product Matrix