ESTIMATING SAMPLING VARIABILITY IN A MULTI-SURVEY ENVIRONMENT

Robert H. Finch, Jr., Social Security Administration

The Office of Research and Statistics conducts many research projects about the social and economic conditions of people who are or could be eligible for benefits under one or more of the social security programs. In the conduct of this research both sample and universe data are obtained from a myriad of sources. To process these data in a timely manner, it was necessary to develop a computer software package which would be relatively easy to use by nonprogrammers but computationally efficient.

While the Social Security Administration has several computer systems, varying by type of design and vendor, the time sharing facility available for research use is an UNIVAC 1108. Thus, a software package was developed to analyze statistical surveys using this facility.

The research analysts in the Office of Research and Statistics prefer to look at survey data in tabular form, i.e., cross tabulations. With this in mind, we felt a survey analysis software package should provide cross tabulations with relative ease. In addition, many algorithms for estimating sampling variances from complex surveys (e.g., random group, collapsed strata and replication) require data in summary form and a cross tabulation package is a good vehicle to provide these summaries. Thus we have the rationale for the development of the software package called "PASS" (Processor for the Analysis of Statistical Surveys).

PASS is designed to be operated by analysts or skilled statistical assistants rather than computer programmers, thus putting the burden of data processing on the computer rather than on the user. PASS is designed as a two step processor: a preprocessor and a post processor. The preprocessor scans a pseudo-English control stream and generates a custom built program (which is the post processor) to produce the requested output.

Through the use of the pseudo-English control statements PASS is able to perform the following tasks for a data file:

1. Create class intervals for data cells.
2. Select cases for inclusion or exclusion in the entire run or in specific tables.
3. Obtain multivariate frequency counts (from one to five variables nested in the stub of a table, and from one to three variables nested in the columns).
4. Obtain totals of variables in addition to or instead of case counts.
5. Obtain tables of mean, percentiles, standard deviations, and horizontal and vertical percentage distributions.
6. Perform cumulative ("AND OVER" and/or "AND UNDER") summing of variables.
7. Produce tabulations which are unweighted, case weighted and/or inflated by a constant.
8. Perform cell by cell operations on tables (i.e., adding two tables, multiplying two tables, subtracting two tables and/or dividing two tables) for tables of the same dimensions.
9. Provides an open FORTRAN routine for special data input requirements (e.g., multiple files with differing formats).
10. Provides an open FORTRAN routine using PASS macro-instructions to access tables by row, column, cell or entire table matrix for special output procedures.
11. Optionally print selected or all data cells for specified or all cases.
12. Provide estimates of sampling variability for a number of sample designs.

The preprocessor consists of modules (sections) to perform the various tasks to produce the desired output. The "INPUT" section describes the input file and the location of the data items to be included in the system. This is completely flexible since it allows card, tape or disk files. The "USER" section allows the inclusion of a user coded FORTRAN procedure for special input processes such as reading composite files.

The "RECODE" section defines how the user wishes to recode his data. We allow three types of recodes: equal interval, range and specified value. Thus a user can input raw data and create new variables defining one or more sets of class intervals for the same variable, without modifying the original variable.

The "TRANSFORM" section allows the user to insert his own FORTRAN code to modify the input data or to create new variables from one or more previously defined variables. In addition the user can selectively print out selected variables for some or all cases. The transform section can appear either before or after the recode section depending on what is desired. If the transform section appears before the recode sections, cases can be selectively skipped before extensive processing and/or newly created transform variables can then be recoded. If the recode section appears before the transform section, recode variables can be used in the transform section (e.g., to selectively print values for selected cases).

The "LABELS" section provides the capability for users to input labels for variables and values of variables for identifying table cells.

The "TABLES" section is where the user specifies the multivariate cross tabulations he wishes. The user can specify up to 99 unique multivariate cross tabulations. In addition within each cross tabulation a user can specify a number of tables of various types, (i.e., attribute, variable, mean, sum of squares, standard deviation, percentile, horizontal percents, columnar percents, and/or generated). Many of the table types can be case-weighted, scalar weighted and/or be rounded. In addition tables may be integer, real or double precision. Some types of tables can use repeated entry (i.e., a case can

be tallied in a table several times). Also many types of tables can be specified as "AND OVER" or "AND UNDER". The printing of "zero" rows, columns, or both can be eliminated. In addition, the printing of entire tables can be suppressed for tables used only to generate other tables.

The "OUTPUT" section provides the user with FORTRAN capability through the use of "Macro calls" to retrieve copies of the tables created--entire tables, rows, columns or individual cells. With this capability the user can build composite tables, call routines to create publication quality tables or write the table cells on tape or mass storage for later use.

The "VARIANCE" section enables the user to obtain estimates of sampling variability for many commonly used sample designs. There are parameters to define the appropriate variance algorithm from among the many available. The system assumes that the data file is in ascending order of a stratum variable.

The "PROGRAM" section is optional and is used in conjunction with the variance section. The "BEFORE" control card enables the user to modify the data laid up in the table arrays. This is especially useful when doing replication variances, (e.g., applying a complex estimation process to each replicate). The "AFTER" control card enables the user to access the variance arrays at the end of the job in order to calculate variance ratios or create specialized outputs.

This completes a summary of PASS. Now I would like to discuss the variance section in more detail, since this section is what makes PASS especially unique.

The variance section is intended to be used under the direction of a sampling statistician. The sampling statistician will have to ensure that the data file contains the necessary sampling control variables and is in the proper sequence for estimating sampling variability. In addition the sampling statistician will have to select the variance type, the type of estimator used for the sample (where necessary), the type of quantity being estimated, the operation (if any) being performed on the estimates, the algorithm to be used (where there is a choice) and whether or not to fit a variance curve.

The "variance types" presently included in PASS are: Simple or Stratified Random; Random Group; Collapsed Strata; Keyfitz (a special routine for nonself representing PSU's for the full sample of the "Current Population Survey"); Paired Selection; Replication and combined types of Random Group-Collapsed Strata or Random Group-Keyfitz.

The "type of estimator" is the least developed of the parameters used in PASS to select variance algorithms. The standard inflation (or Horvitz-Thompson) estimator is the system default and at the present time the only other estimator allowed is a special case of a Stratified Ratio Estimate for stratified random sampling of attributes, where the independent total is the stratum population.

The "type of quantity" parameter is basically required to select the appropriate algorithm for Random Sampling. Attributes are the most frequently used quantities and are available for all "variance types." Variables are also available for all "variance types." Ratios and products are available for all "variance types" except random (which will be available later).

Operations on data items are also permitted. The operations are: sum, difference, product and ratio. These operations are permitted for all "variance types" except random and for all "quantities." At this time, the variance components algorithms are the only ones available, however, we plan to include direct and high order terms where possible.

The use of the variance components algorithm has distinct advantages however. Since the variance of sums, differences, products and ratios all require the same variance components and because these components are combined differently at the end, only one variance routine is required for each of the "variance types." The combining operations can be performed in the routine which prints the results of the variance calculations, with little extra computer overhead.

The "VAR TABLES" parameter allows the user to define which tables are to be used as a basis for the variance calculations. Since the variance calculations are based on a transparent third dimension defined in the variance section, which is not available in the "tables" section, the variance tables must be in different cross-tabulation. I have been unable to define a term for the third dimension variable but can only define it by example: for "Random Group" variances it is the random group to which the case has been assigned; for "Collapsed Strata" variances it is the stratum code within the superstratum in which the code occurs.

In addition to putting out tables of standard errors at the end of the computer run, PASS has some additional options available both for diagnostic checking and summarization. The standard variance output for a typical attribute table consists of a variance table which gives for each table cell: an item number (for identification purposes); the estimate; the standard error; the variance and the relvariance. Additional output can be requested: i.e., variance cards to be used to generalize the data later; and/or the user can request that the variance data be put into a curve fitting program to provide generalized tables of standard errors.

For diagnostic information, the "TAB-PRINT" option can be used. This will cause the basic cross-tabulations to be printed for each stratum as well as the cumulative variance table--normally these are not printed. These additional tables are useful in checking the variance calculations as well as examining the contributions to the variances of each stratum.

Additional informational output can be obtained by requesting "DESIGN EFFECTS." This will produce an additional table of design effects (the ratio of the variance for the particular sample design to the variance of a simple random sample of the same size). At a later time, we plan to provide for running two variance types in the same run and obtaining the variance ratios for the two types as an analytical tool.

The curve fitting routine that we use assumes a curve of the form $Vx^2 = A + Bx^{-1} + Cx^{-1/2}$ where Vx^2 is the relvariance of the estimated quantity x. The function minimized is $F = \left[\frac{Vo^2 - Vc^2}{Vc^2}\right]^2$ where Vo^2 is the relvariance observed (computed in the variance section) and Vc^2 is the relvariance from the curve. This is an iterative process and continues until the difference between the curve parameters (A, B, and C) change by less than 1 percent between two successive iterations or when the number of iterations reaches 25. (For simple random the curve is $V^2 = A + Bx^{-1}$).

These curves usually give excellent results for attribute data. Variables do not fit quite as well but it is the best we have found. We also use the curve for ratios (usually subpopulation means), but the results are less satisfactory.

The output of the curve fitting is quite extensive. It includes the curve parameters for each iteration, the percent difference of each parameter from the preceding iteration and the number of digits of accuracy for the coefficients. It also provides a table giving the item number, the estimate, the relvariance input, the relvariance from the curve, the residual and the relative residual for each input item.

For checking the curve, both multiple R and R^2 are given. At the end of the fitting process the model is given. In addition a plot is given for the fitted curve and the original scatter gram on the same graph as well as a plot of the residuals. The resulting curve parameters are then fed into a routine which produces a generalized table of standard errors for fixed estimates for publication purposes and general use. In addition, using the approximation: $V^2x/y = V^2x - V^2y$ it produces a table of standard errors of percentages.

These tables are produced for attributes and/or variables. If the quantity is ratio, the curve fitting is done for the numerator of the ratio, the denominator of the ratio and the ratio itself. However, the generalized table for the ratio is different--it is a table of relative standard errors for the fixed estimates (the denominator of the ratio).

In conclusion let me say that PASS fulfills many requirements for analyzing data from diverse sources. It is an extremely useful tool for analysts who want a quick look at their data. PASS is also useful to mathematical statisticians who want estimates of sampling variability. Managers can relate to PASS since they only need one software package to satisfy both the analytical and statistical needs required of survey data.