A COMPARISON OF METHODS AND PROGRAMS FOR COMPUTING VARIANCES OF ESTIMATORS FROM COMPLEX SAMPLE SURVEYS

Bruce Kaplan and Ivor Francis, Cornell University
J. Sedransk, SUNY - Albany

1. Introduction

In recent years, there has been an expanded interest in software packages to be used in processing and analyzing data from sample surveys. Francis and Sedransk (1976) initially listed some desirable features of the software needed for such purposes, and subsequently have begun to collect, organize and present information about existing software. In several situations, numerical tests of the capabilities of the software have also been carried out (Kaplan, Francis and Sedransk (1979), and Francis, Sherman, Buhrman and Willard (1979)).

In this paper we consider programs for computing point and variance estimates from survey data, and present information supplied by the program developers. In an earlier paper, Kaplan, Francis and Sedransk (1979) explored the development of benchmark data sets to test the capabilities of portable programs, and presented some of the results obtained by testing three such programs.

We hope that the presentation of information and evaluation of software will be of assistance to those wishing to select a portable program for their own use. In addition, those planning to develop their own programs may wish to enhance their own efforts by considering desirable features of both portable and non-portable programs. Finally, we expect that a state-of-the-art assessment of the capabilities of "variance estimation" programs is of general interest to sample survey practitioners.

2. Identification of Programs

Before summarizing some of the characteristics of the programs useful for survey point estimation, it is useful to describe the procedures which were used to identify programs and packages used in processing and analyzing data from sample surveys (i.e., frame development), and to describe the process of self-evaluation of program capabilities. By using this sequential process we hope to obtain detailed information from the program developers about the capabilities of their programs. This entire procedure is relatively inexpensive, and if the procedure can be validated and combined with user ratings it will provide an efficient way to obtain and to provide broad evaluations of the capabilities of computer packages.

To initiate the frame development process, a questionnaire (Q.1a) was sent in 1977 to all North American members of the International Association of Survey Statisticians and to all members of the Subsection on Survey Research Methods of the American Statistical Association, a total of approximately 2500 questionnaires in all. The 375 respondents named 194 programs used in their statistical computing.

In 1978 an improved version of this questionnaire (Q.1b) was sent to all 123 institutional members of AAPOR (American Association of Public Opinion Research), all 36 Survey Research Centers listed in Survey Research, and 115 statisticians selected purposively from the Federal Statistical Directory, as well as to most of the respondents to Q.1a. The objectives were to identify appropriate computer programs and to obtain very general information about their purported emphases. Thus, if the contacted organization used computer programs in the processing or analysis of data from sample surveys, the respondent was asked to provide the names of the programs and the names and addresses of the developers. These programs were to be classified according to the following five tasks: (1) data management and file building, (2) editing: error detection, correction and imputation, (3) data description, tabulation and plotting, (4) estimation of finite population parameters and associated variances for complex sample surveys, and (5) statistical analyses and model building.

Also in 1978 additional programs were identified from responses to notices placed in journals such as Amstat News, International Statistical Institute Newsletter, SIGSOC Bulletin, the Royal Statistical Society's News and Notes, Newsletter of the Institute of Statisticians, Software World, Computing, V.V.S. Bulletin, Communications in Statistics B, and the American Economic Review.

Developers of programs identified in the frame development phase were re-

quested to complete a lengthy questionnaire (Q.2) about program capabilities. This self-evaluation questionnaire contained a small number of questions pertaining to each of the five tasks listed above. The objectives were: (1) to provide basic summary information about package capabilities over a broad range of tasks, and (2) to identify packages of purported excellence in carrying out specific tasks. Such packages might be queried more intensively in an additional, detailed, questionnaire whose content would be limited to one specific task (e.g., variance estimation); or such packages might be subjected to a controlled test of their capabilities (see, e.g., Kaplan, Francis and Sedransk (1979) for variance estimation).

Using the information from Q.1a and Q2, a questionnaire (Q.3S) concerning variance estimation for complex sample surveys was developed and sent to the developers of the programs listed below. (Note that the institution where the program is currently being maintained is listed in parentheses.)

A. HES Variance and Crosstabulation Program (National Center for Health Statistics)

B. STDERR-SESUDAAN (Research Triangle Institute)

C. Consumer Expenditure Survey Variance Program (Bureau of Labor Statistics)

D. KOTAB2-VTAB (National Central Bureau of Statistics, Sweden)

E. Current Population Survey Variance Program (Bureau of the Census)

F. SUPER CARP (Iowa State University)

G. OSIRIS IV: PSALMS (ISR, University of Michigan)

H. OSIRIS IV: PREP (ISR, University of Michigan)

I. Variance/Covariance System for the Canadian Labour Force Survey (Statistics Canada)

J. Generalized Variance Estimation Program (GENVAR) (Bureau of the Census)

K. CLUSTERS (World Fertility Survey)

L. Consistent System (CS) (Laboratory of Architecture and Planning, M.I.T.)

M. JES Summary System (E.S.C.S. - U.S.D.A.)

N. SYSTEM 4204 (E.S.C.S. - U.S. Department of Agriculture)

O. Table Producing Language (TPL) (Bureau of Labor Statistics)

P. Generalised Survey System - Estimation Package (GSS) (Australian Bureau of Statistics)

Q. SPSS - Splithalf Procedure (Australian Bureau of Statistics)

R. Rothamsted General Survey Program (RGSP) (Rothamsted Experimental Station)

3. Results

In Table 1, we indicate which of the above programs have stated that they provide (1) estimates of means, ratios and differences of ratios; (2) estimates of variances for estimates of means, ratios and differences of ratios; and (3) estimates of coefficients of variation and design effects for (a) the entire population, (b) individual strata, and (c) subpopulations. For each type of estimate, and for each grouping of (population) elements, the presence of a program's identifying letter (see the list above) indicates that the specified capability has been stated as being available.

In Table 2, we present information about the availability of the program, the generality of the program, the documentation and the user command language. A program is considered to be available (and labelled "A") if, and only if, it can be acquired, and is currently being used in at least two institutions. For the generality of a program, it is classified as: (a) specific to a particular survey, (b) useful for a particular kind of sample design, or (c) useful for several kinds of sample designs. The user command language is classified as : (a) fixed position alpha or numeric codes, (b) codes in fixed order with punctuation to indicate omitted codes, (c) free field alphanumeric commands with specified syntax, or (d) English-like verbs and nouns or sentences.

Additional questions in Q.3S requested detailed information about: (1) type of variance estimation procedure used (e.g., Taylor series, jackknife), (2) types of sample design accommodated, (3) use of finite population corrections, (4) incorporation of all sources of variation in variance estimators, (5) treatment of very small sample sizes, and (6) limitations on number of variables, number of clusters,

1. Capabilities of Variance Estimation Programs

Estimates of	Estimates for: Entire population	Individual strata	Subpopulations
means	ABCDEFGHIJKLMNO QR	BCD FGHIJKLMNO QR	ABCDEFGH JKLMNO QR
ratios	A CDEFG IJKLMNOPQR	CD FG IJKLMNOPQR	A CDEFG JKLMNOPQR
differences of ratios	EFG IJKL O R	FG IJKL O R	EFG JKL O R
variances of means	ABCDEFGHIJKLMNO QR	BCD FGHIJKLMNO QR	ABCDEFGH JKLMNO QR
variances of ratios	A CDEFG IJKLM OPQR	CD FG IJKLM OPQR	A CDEFG JKLM OPQR
variances of differences of ratios	EFG IJKL O	FG IJKL O	EFG JKL O
coefficients of variation	CDE G I K MNO R	CD G I K MNO R	CDE G K MNO R
design effects	E GH K R	GH K R	E GH K R

NOTE: The presence of a program's identifying letter (see text) indicates that the specified capability has been stated as being available.

number of strata, etc. The responses to these questions will be summarized in a subsequent report.

Acknowledgment

This work was partially supported by the National Science Foundation under Grant Number MCS 75-13373-A02

References

1. Francis, I. and Sedransk, J. (1976). "Software Requirements for the Analysis of Surveys," Proceedings, International Biometric Conference, Boston.

2. Francis, I., Sherman, S., Buhrman, and Willard, J. (1979). "A Comparison of Software for Tabulating Survey and Census Data," Proceedings of the Statistical Computing Section, American Statistical Association.

3. Kaplan, B., Francis, I., and Sedransk, J. (1979). "Criteria for Comparing Programs for Computing Variances of Estimators From Complex Sample Surveys," Proceedings of the 12th Annual Symposium on the Interface of Computer Science and Statistics.

2. Selected Characteristics of Variance Estimation Programs

Program	Availability	Generality	User command language	Documentation
A	A	b	ad	Users' guide
B	A	c	c	Users' guide
C		b	a	"Very little"
D		c	a	In Swedish
E		a	ac	Technical report, internal documentation
F	A	c	a	Manual
G	A	c	c	Part of OSIRIS IV manual
H	A	c	c	Part of OSIRIS IV manual
I		b	a	Technical report
J	A	c	d	"Draft"
K	A	c	a	Users' manual
L	A	c	d	Manuals (reference, tutorial, technical)
M		a	a	Users' manual
N		c	a	Users' manual
O	A	c	d	Users' manual
P		c	a	Manuals (users', operations, external reference specification)
Q		b	c	Users' and systems documentation
R	A	c	c	Manuals, introductory guide

NOTE: See text for program identification, and for definitions of availability, generality, and classification of user command language.