COMPUTERIZATION OF COMPLEX SURVEY ESTIMATES[1]

M.A. Hidiroglou[2]

Survey data collected by statistical agencies is most likely to be processed through to the tabulation stage by these agencies. The computer programs associated with this processing are also most likely tailored to the particular design and variables used. The statistics computed from such surveys typically range from simple descriptive totals and means to these required for analytic studies such as comparison of domains, regression analysis and contingency tables analysis. This paper describes a computer program which computes these statistics and their associated sampling errors for commonly used sampling designs.

1. INTRODUCTION

A variety of statistics are computed for survey data which often arise from large, complex national and regional surveys. The statistics computed from such surveys typically range from simple descriptive totals and means to those required for analytic studies such as comparison of domains, regression analysis, and contingency tables analysis. Domain estimation refers to the estimation of statistics for subgroups of the population of interest which are not explicitly provided for in the design. Yates (1960) contains considerable material on the estimation of domain means and their differences. Hartley (1959) and Rao (1975) provide an excellent account of the methodology used for domain estimation. The variance estimators associated with the domain estimators are easy extensions of variance estimators for simple statistics. This is not, however, the case for more complex statistics. The estimation of regression equations from survey data presents several problems; for example, the definition of the regression equations, the identification of the population for which inferences are desired, and the variance estimation for the regression coefficients (see Konijn (1962), Kish and Frankel (1974) and

---

[1] Presented at the Annual Meetings of the American Statistical Association, Detroit, August 1981.

[2] M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada.

Fuller (1975). The testing of hypotheses for contingency tables given survey design considerations have been studied by Nathan (1969, 1972), Rao and Scott (1981), Garza-Hernandez and McCarthy (1962) and Koch, Freeman and Freeman (1975) to name a few.

Survey data collected by statistical agencies is most likely processed through to the tabulation stage by these agencies. The computer programs associated with this processing are also most likely tailored to the particular design used. It is quite possible that computer programs used to produce estimates of totals (say) and their associated variances must be developed from scratch every time that a new survey design is introduced. This is time consuming, expensive, tedious and in some sense repetitive. Use of statistical software packages such as SPSS or SAS may be considered as an alternative. These packages may be readily used to produce weighted'estimates. However, the variances that they compute do not take sample design factors such as stratification and clustering into account unless they are programmed to do so. A user must therefore be fairly familiar with the language used by these packages if he wants to obtain proper variance estimates for survey estimates.

Recently, there have been attempts to develop programs which compute variances for a general class of designs. Some of these programs are STDERR by Shah (1974), SURREGR by Holt (1975), SUPER CARP and MINI CARP by Hidiroglou, Fuller and Hickman (1980). These programs basically require the specification of the estimator to be used and the variables to be analysed. It will be assumed that the data sets that these programs are being applied to have been edited and that missing observations have been imputed. In this paper, SUPER CARP and MINI CARP will be described. SUPER CARP can be used to construct estimated totals, ratio estimates, the difference of ratio estimates and contingency tables tests for multistage stratified samples. It contains a number of regression procedures appropriate for data observed subject to response (Measurement) error. Covariance matrices can be estimated for sub-population means, and totals and for stratum means and totals. MINI CARP is a smaller program which differs from SUPER CARP in that it does not contain and of SUPER CARP's regression procedures. A comparison of the capabilities of the two programs is given in Table 1.

TABLE 1. Capabilities of SUPER CARP (S) and MINI CARP (M)

Multivariate Estimate of	For		
	Entire Population	Individual Strata	Sub-population
Simple Parameters			
. Means	S,M	S,M	S,M
. Totals	S,M	S,M	S,M
. Ratios	S,M	S,M	S,M
. Difference of Ratios	S,M	S,M	S,M
. Proportions	S,M	S,M	S,M

Complex Parameters		Tests	
. Weighted Least Squares	S	. Regression Coefficent	S
. Weighted Errors-in-the Variables (Known & Estimated error covariances)	S	. Goodness-of-fit	S,M
		. Independence for Two-Way Table	S,M

2. GENERAL DESCRIPTION

2.1 Notation

In general, SUPER CARP and MINI CARP can accept data from a multistage stratified design. Assuming that the design has s stages, a g dimensional data vector is read in for each observation. We denote this data vector as

$$(Z_{hi_{\sim s}1}, Z_{hi_{\sim s}2}, \ldots, Z_{hi_{\sim s}g}),$$

where h = 1, 2, ..., L denotes strata; $i_{\sim s} = (i_1, i_2, \ldots, is)$ represent the stages; $i_1 = 1, 2, \ldots, n_h$ represents the first stage identification; $i_2 = 1, 2, \ldots, n_{hi_1}$ represent the second stage identification; ... ; $i_s = 1, 2, \ldots, n_{hi_{\sim s-1}}$ represents the last s-th identification. $Zhi_{\sim s}k$ is the $hi_{\sim s}$-th observation for the k-th variable of interest. Weights associated with the $n_{hi_{\sim s}}$-th observation will be referred to as $w_{hi_{\sim s}}$. These weights would be inversely proportional to the selection probabilities of each ultimate sampled

unit. The specification of the variables to be used in the analysis (be it total or ratio estimation or regression estimation) is done by using a selection vector $\underset{\sim}{y} = (v_1, v_2, \ldots, v_{p+1})$ where $1 \leq v_k \leq g$ for $k = 1, 2, \ldots, p+1$. Given that the type of analysis and the identification of the variables has been decided upon, let the chosen vector for the $hi_{\sim s}$-th observation be

$$(Y_{hi_{\sim s}}, X_{hi_{\sim s}1}, X_{hi_{\sim s}2}, \ldots, X_{hi_{\sim s}g}),$$

where Y denotes the dependent variable and X denotes the independent variables if regression analysis is specified. Note that v_1 is always the index for the dependent variable in the case of regression. For other types of analyses, the ordering within the selection vector is not important.

2.2 Types of Computations

The simple statistics and a partial list of the regression options available in the program are outlined. A complete descritpion of all the available options is written up in the SUPER CARP or MINI CARP manuals (1980).

(i) Total Estimator, e.g.

$$\hat{X}_{(k)} = \sum_h \sum_{i_1} \cdots \sum_{i_s} w_{hi_{\sim s}} X_{hi_{\sim s}(k)}, \quad k = 1, 1, 2, \ldots, p.$$

The estimated covariance matrix for

$$\hat{X} = \{\hat{X}_{(1)}, \hat{X}_{(2)}, \ldots, \hat{X}_{(p)}\} \text{ is}$$

$$v_1(\hat{X}) = \sum_{h=1}^{L} (n_h - 1)^{-1} n_h(1-f_h) \sum_{i_1=1}^{n_h} (\hat{d}_{hi_{\sim 1}.} - \hat{\underline{d}}_{h..})^T (\hat{d}_{hi_{\sim 1}.} - \hat{\underline{d}}_{h..}) \quad (2.2.1)$$

where

$$\hat{d}_{\sim hi_1.} = \{\hat{d}_{hi_1(1)}, \hat{d}_{hi_1(2)}, \ldots, \hat{d}_{hi_1(p)}\}$$

$$\hat{d}_{hi_1(k)} = \sum_{i_2=1}^{n_{hi_1}} \cdots \sum_{i_s=1}^{n_{hi_{s-1}}} w_{hi_s} X_{hi_s(k)}$$

$$\hat{\bar{d}}_{\sim h..} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{d}_{\sim hi_1.} \quad .$$

Note that the above variance formula may be applied to pps schemes with and without replacement. For with replacement schemes, only the first stage variance needs to be computed (Des Baj, 1968, pg. 120) and the correction factors f_h are set to zero. In large scale surveys, it is often assumed that the first stage clusters have been selected without replacement even though the actual selection scheme may have been without replacement. This assumption inconjunction with small sampling fractions implies that resulting variance is fairly close to the one which would have been obtained by taking all stages and selection prodcedure sinto account. If the sampling fractions are not negligible at each stage and that the sampling has been performed using without replacement S.R.S. at each stage, Des Raj's rule (1966) can be used to advantage to compute each stage component of covariance. The covariance matrix accounting for s stages is:

$$v(\hat{\underset{\sim}{X}}) = \sum_{r=1}^{s} v_r(\hat{\underset{\sim}{X}})$$

where for $r \geq 2$

$$v_r(\hat{\underset{\sim}{X}}) = \sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \cdots \sum_{i_{r-1}=1}^{n_{hi_{r-2}}} \left[\prod_{j=0}^{r-2} \frac{n_{hi_j}}{N_{hi_j}} \right]$$

(2.2.2)

$$\times\ n_{hi_{r-1}} (n_{hi_{r-1}} - 1)^{-1} (1-f_{hi_{r-1}})$$

$$\times \sum_{i_r} (\hat{d}_{hi_r}(.) - \hat{\tilde{d}}_{hi_{r-1},.}(.))^T$$

$$\times\ \hat{d}_{hi_r}(.) - \hat{\tilde{d}}_{hi_{r-1},.}(.)$$

where

$$f_{hi_{r-1}} = n_{hi_{r-1}} N_{hi_{r-1}}^{-1},\ n_{hi_o} = n_h,\ N_{hi_o} = N_h,$$

$$\hat{d}_{hi_r}(.) = \hat{d}_{hi_r}(1), \hat{d}_{hi_r}(2), \ldots, \hat{d}_{hi_r}(p)$$

$$\hat{d}_{hi_r}(k) = \sum_{i_{r+1}} \cdots \sum_{i_s} w_{hi_s} X_{hi_s}(k)$$

$$\hat{\tilde{d}}_{hi_{r-1},.}(.) = n_{hi_r}^{-1} \sum_{i_r} \hat{d}_{hi_r}(.) \quad .$$

The variance estimation for an r-stage design can therefore be done by estimating the components at each stage ($v_r(\hat{X})$) and summing them up. This can be done by passing over the data set r separate times. The first time around, strata and first stage units are read into the program to give $v_1(\hat{X})$. The second time around, the original primary sampling units are read into the program as "strata" and the secondary units are identified as clusters to give $v_2(\hat{X})$. The r-th time around, the original ($r \geq 2$) (r-1)-th stage units are read into the program as "strata" and the r-th stage units are identified as clusters to give $v_r(\hat{X})$.

On each pass a sampling rate $g_{hi_{r-1}}$ must be read in for the hi_{r-1}-th unit where

$$g_{hi_{r-1}} = 1 - \left[\prod_{j=0}^{r-2} \frac{n_{hi_j}}{N_{hi_j}}\right]\left[1 - \frac{n_{hi_{r-1}}}{N_{hi_{r-1}}}\right] .$$

Using this procedure, the program will be computing $v_r(\hat{X})$ in the format given by $v_1(\hat{X})$.

If the sampling factors are not negligible at each stage and that sampling has been performed using without replacement p.p.s. schemes at each stage, the variance expression at each stage must take into account joint selection probabilities. SUPER CARP and MINI CARP do not compute joint selection probabilities. For the case where two units per stratum have been selected without replacement and unequal probability, the variance of the estimator for total can be obtained using formula (2.2.1) with a correction factor for each stratum which includes the joint probabilities of selection. This correction factor is given by

$$f_h = \frac{2\pi_{h12} - \pi_{h1}\pi_{h2}}{\pi_{h12}} \quad , h=1, 2, \ldots, L$$

where π_{h12} is the joint probability of selection for the selected units 1 and 2. If $n_h \geq 2$ and that the joint probabilities of selection are not available, an approximation to the without replacement variance has been given by Gray

(1975). Gray shows that the variances of an unequal without replacement sample may be partitioned into a "with replacement" variance component times a finite population correction factor which depends on the joint probabilities. This correction factor has been found to be roughly equal to one minus the inverse of the sampling fraction for populations which have more than 15 elements within each stage. Using Gray's approximation, variances for multistage unequal without replacement schemes can be computed.

If domain estimation is required for some of the variables, a new variable ${}_d Y_{hi_{\sim s}}(k)$ is defined for all elements in the population, where

$$ {}_d y_{hi_{\sim s}}(k) = \begin{cases} Y_{hi_{\sim s}}(k) & \text{if the } hi_{\sim s}\text{-th element belongs to the domain d (say } D_d) \\ 0 & \text{otherwise} \end{cases} $$

An alternative way of defining ${}_d Y_{hi_{\sim s}}(k)$, is

$$ {}_d Y_{hi_{\sim s}}(k) = {}_d a_{hi_{\sim s}} \; Y_{hi_{\sim s}}(k) \quad \text{where} $$

$$ {}_d a_{hi_{\sim s}} = \begin{cases} 1 & \text{if the } hi_{\sim s}\text{-th element belongs to } D_d \\ 0 & \text{otherwise} \end{cases} $$

Note that if $\hat{Y}$ and $v(\hat{\underset{\sim}{Y}})$ are unbiased for Y and $v(\hat{\underset{\sim}{Y}})$ respectively, then the corresponding domain estimators ${}_d\hat{\underset{\sim}{Y}}$ and $v({}_d\hat{\underset{\sim}{Y}})$ are unbiased for ${}_d\underset{\sim}{Y}$ and $V({}_d\hat{\underset{\sim}{Y}})$. The standard formulae for $\hat{Y}$ and $v(\hat{Y})$ can now be applied to the "synthetic" variables ${}_d Y_{hi_{\sim s}}$. Stratum totals can be computed individually by treating the strata as classification variables.

(ii) Ratio Estimator

The vector $\{Y_{hi_{\sim s}}(1), X_{hi_{\sim s}}(1), \ldots Y_{hi_{\sim s}}(p), X_{hi_{\sim s}}(p)\}$ is used in the analysis and the estimated ratios are:

$$ \hat{R}(t) = \hat{X}_{(t)}^{-1} \hat{Y}_{(t)}, \; t = 1, 2, \ldots, p \; ; $$

where $\hat{Y}_{(t)}$ and $\hat{X}_{(t)}$ are of the form given in the previous section. The estimated covariance matrix for $\hat{\underset{\sim}{R}} = \{ \hat{R}(1), \hat{R}(2), \ldots, \hat{R}(p) \}$ is as given in the previous section with

$$\hat{d}_{hi_{\sim r}(t)} = \hat{X}^{-1}_{(t)} \sum_{i_{r+1}} \cdots \sum_{i_s} w_{hi_{\sim s}} \{ Y_{hi_{\sim s}(1)} - \hat{R}(t)\, X_{hi_{\sim s}(t)} \};\ t = 1, \ldots, p.$$

The ratio estimator can be used for computing the mean for each variable of interest by setting all X-variables to 1. Domain means can be computed by using $_{d}y_{hi_{\sim s}(t)}$ in the place of $Y_{hi_{\sim s}(t)}$ and $_{d}a_{hi_{\sim s}}$ in the place of $X_{hi_{\sim s}(t)}$. If subpopulation proportions of Y for a domain D_d are required, the numerator of the ratio is the sum of weighted $_{d}Y_{hi_{\sim s}(t)}$ and the denominator is the sum of weighted $Y_{hi_{\sim s}(t)}$. The estimated ratio for two variables defined over a domain D_d may similarly be obtained. Stratum proportions and ratios may be computed with the strata serving as the classification variables.

(iii) Regression Estimation

Some considerable attention has been paid recently to regression concepts in survey sampling. There are several explanations for this. First, there is an increased emphasis on analytic surveys, with partly unresolved questions of proper weighting of observations. Secondly, modeling in general, especially in the regression context, has attracted widespread interest, as well as criticism, as a tool in making survey estimates. SUPER CARP properly weights the observations and computes the variances of the estimated regression coefficients using a method given by Fuller (1975).

The regression coefficients estimated from a stratified cluster sample are given by

$$\underset{\sim}{b} = (\underset{\sim}{X}_{n}' \, W \, \underset{\sim}{X}_{n})^{-1} \ \underset{\sim}{X}_{n}' \, W \, \underset{\sim}{Y}_{n}$$

where the (rs)-th element of $(\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)$ is

$$\sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} \; X_{hi_1i_2s} \; W_{hi_1i_2}$$

and the r-th element of $\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{Y}_n$ is

$$\sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} \; X_{hi_1i_2} \; W_{hi_1i_2} \; .$$

The estimated covariance matrix of $\underset{\sim}{b}$ is computed as

$$v(\underset{\sim}{b}) = (\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)^{-1} \hat{\underset{\sim}{G}}_n (\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)^{-1} \; ,$$

where the (rs)-th element of $\hat{\underset{\sim}{G}}_n$ is

$$\hat{g}_n(r,s) = \frac{n-1}{n-p} \sum_{h=1}^{L} \frac{n_h(1-f_h)}{(n_{h-1})} \sum_{i_1=1}^{n_h} (\hat{d}_{hi_1.r} - \bar{d}_{h..r}) \times (\hat{d}_{hi_1.s} - \bar{d}_{h..s})$$

where

$$\hat{d}_{hi_1i_2r} = X_{hi_1i_2r} \; \hat{v}_{hi_1i_2} \; w_{hi_1i_2} \; ,$$

$$\hat{v}_{hi_1i_2} = Y_{hi_1i_2} - \sum_{r=1}^{P} \hat{b}(r) \; X_{hi_1i_2} \; ,$$

$$\hat{d}_{hi_1.r} = \sum_{i_2=1}^{n_{hi_1}} \hat{d}_{hi_1 i_2 r} ,$$

$$\bar{d}_{h..r} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{d}_{hi_1.r} ,$$

$$n = \sum_{h=1}^{L} \sum_{i_1=1}^{n_h} n_{hi_1} .$$

The variance estimation procedure is based on an asymptotic Taylor expansion of the sample regression coefficient vector. This method has several advantages over the Balanced Repeated Replication and Jack-Knife Replication methods. Firstly, it is relatively easy to program, and it can be adopted to multistage sample designs. Secondly, no restrictions are placed on the sample design (two replicates per stratum, for instance) and the assumptions used require some well-behaved moments in the population of interest. Thirdly, it requires the least number of computations.

Data is quite frequently measured with error. Theory for regression models which takes measurement error into account has been given by Fuller (1980a), Fuller (1980b) and Fuller and Hidiroglou (1978). SUPER CARP also has the flexibility to compute tests of hypothesis for any subsets of the regression parameters.

(iv) Contingency Tables

SUPER CARP and MINI CARP perform the goodness-of-fit test and the independence test for data resulting from complex surveys. These two tests take the stratification and the clustering of the design into account. As pointed out by Rao and Scott (1981), pratictioners using traditional Pearson chi-quare statistics for those two tests, given that there may be serious design effects can be seriously misled.

For the goodness-of-fit test, SUPER CARP and MINI CARP use the modified Wald Statistic given by

$$F_{WG} = [(k-1)d]^{-1} (d-k+2) (\hat{p}-p_o)^T \hat{V}^{-1} (\hat{p}-p_o)$$

where

$\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{k-1})^T$ is the vector of estimated proportions given in in the stratum and cluster configurations,

$p_o = (p_{o1}, p_{o2}, \ldots, p_{o,k-1})^T$ is the vector of hypothesized proportions,

$\hat{V}$ = the covariance matrix of $\hat{p}$ given the stratum and cluster configuration,

k = number of categories considered,

$$d = \sum_{h=1}^{L} (n_h - 1),$$

L is the number of strata in the sample and n_i is the number of clusters in the i-th stratum. The covariance matrix $\hat{V}$ is computed using the methods given for ratio estimation. In large samples, F is approximately distributed as a central F with k-1 and d-k+2 degrees of freedom when the null hypothesis is true.

For the test of independence, Fuller (SUPER CARP p. 65-69) has developed a test which takes the design into account. Given that the contingency table which splits the population according to two criteria is made up of R rows and C columns, the null hypothesis to be tested is H_o: $p_{ij} = p_{i+} p_{+j}$ or $p_{+j} = p_{i+}^{-1} p_{ij}$. where p_{ij} = ij-th cell proportion in the population, $p_{+j} = \Sigma_i p_{ij}$ and $p_{i+} = \Sigma_j p_{ij}$.

Given that $P_{ij|i}$ is defined as $p_{i+}^{-1} p_{ij}$ and that the corresponding sample estimators are $\hat{p}_{ij|i} = \hat{p}_{i+}^{-1} \hat{p}_{ij}$, estimates for $(p_{+1}, p_{+2}, \ldots, p_{+,c-1})$ can be obtained by regressing $\hat{p}_{ij|i}$ $(i = 1, 2, \ldots, R;\ j = 1, 2, \ldots, C-1)$ on $(C-1)$-dimensional row vectors whose elements are one for the j-th entry corresponding to $\hat{p}_{ij|i}$ and zero otherwise. The regression is of a generalized least-squares nature because the $\hat{p}_{ij|i}$ do not have the same error structure. An estimator for the covariance matrix of the $p_{ij|k}$'s, incorporating the sample design, is obtained using the ratio estimator formulae. The test statistic for H_o is then based on the residual sums of squares for this regression.

3. INPUT

In a typical survey situation, the data associated with a given selected unit is characterized by stratum, first stage, second stage up to s-th stage identification and a sampling weight. The data must be ordered hierarchically with respect to this identification in order to produce estimates of variance which reflect the stratified and clustered of the data.

SUPER CARP and MINI CARP are run using command language specified in numeric codes in fixed card positions. For both programs, there are six mandatory control cards to be input at all times. A number of optional control cards may also be input if more information is required by options specified in the mandatory cards. The mandatory cards are the parameter card, the variable name card, the format card, the screening card, the analysis card and the variable identification card. The parameter card provides overall preliminary information to the program such as, problem identification, number of observations to be read in, input service identification (tape, disk or cards), data identification structure, data output and stratum collapsing controls. The format card specifies the input format for the data as well at its identification and the associated weight. the variable name card assigns chosen names to input data fields in the order that they are read in. The screening card specifies tolerance limits for given variables provided that screening is required. The analysis card specifies the type of analysis to be performed (see table 1). Finally, the variable identification card identifies the variables to be used in the chosen analyses. The optional cards include such

cards as the sampling rate card (sampling rates by stratum can be read in), the errors-in-the variable cards for supplying the program with covariance matrices for variables measured with error, the hypothesis testing card for specifying coefficients in a regression analysis to be tested equal to zero.

4. COMPUTATIONS

4.1 For Means and Corrected Sums of Squares and Cross-Products

The means, corrected sums of squares and cross-products are statistics routinely computed in a survey package. The choice of algorithms for computing these statistics should take into consideration precision, speed and storage requirements. Beaton, Rubin and Barone (1976) have noted that a "concern about highly accurate computation methods must be tempered with a concern for whether the data are accurate enough to make the results meaningful". Different variations of one-pass and two-pass algorithms have been studied by Ling (1974). Ling's conclusion is that there is no universally best algorithm. The best algorithm for a given data set depends on the numbers in that data set. One of his recommendations is to use double precision arithmetic to be beyond the accuracy attainable in single-precision arithmetic. One-pass recursive algorithms should be chosen over the usual one-pass 'desk-machine' method because they have a higher tendency to produce less computational errors. This is especially the case for subroutines programmed in single precision. In SUPER CARP and MINI CARP one-pass recursive algorithms programmed in double precision have been chosen.

4.2 Inversion of Matrices

Matrix inversion is required for regression and contingency table analysis. the choice for inversion algorithms is quite important in packages. This has been reported by Longley's (1967) paper in which he examined the accuracy of some inversion algorithms and found serious computational inaccuracies. He reported that the most accurate results were obtained by using the orthonormalization procedure. Kopitze, Boardman and Graybill (1975) recommend the use of the Cholesky decomposition as an inversting algorithm. They point out that as compared to the Gaussian elimination schemes, it does not require

pivoting to stabilize symmetric positive definite matrices. This means less time for inverting. The Cholesky decomposition does not need much core storage and is easier to program than the Gaussian elimination scheme. One of its other advantages, as Wilkinson's (1965) analysis shows, is that it is quite accurrate. Another of its advantages is that it can be used to find eigenvalues for systems of equations of the form $\underset{\sim}{A}\,\underset{\sim}{x} = \lambda\,\underset{\sim}{B}\,\underset{\sim}{x}$ where $\underset{\sim}{A}$ is a positive matrix and $\underset{\sim}{B}$ is a positive semi-definite matrix. Computations of eigenvalues are required in SUPER CARP for some of the errors-in-the variables regression analyses. It is for this reason and the precision considerations that the Cholesky decomposition has been adopted for inversion purposes in SUPER CARP.

4.3 Stratum Collapsing

If a sampled population is highly heterogeneous and several criteria are available for stratification, it is quite possible that some strata my contain only one cluster. For such strata, it is not possible to estimate the variability. In such cases, the user may request that the one cluster strata be collapsed with neighbouring strata. If such a request if not made, SUPER CARP or MINI CARP exclude with one cluster from variance computations but include them for estimation purposes. The program lists those strata with only one unit. This information may lead the user to collapse those strata in a subsequent pass. If collapsing is to be done, the strata which are to be collapsed should be similar to neighbouring strata. A suggested method for collapsing which is easily amenable to programming is as follows. If a stratum is encountered that contains only one cluster, that stratum is combined with the following stratum in the file sequence. If the last stratum contains only one element, the last stratum is combined with the next to last stratum. A stratum with a sampling rate of one is not collapsed because such a stratum makes no contribution to the between primary component of the sampling variance. Strata with a sampling rate of one should never appear after a stratum with only one cluster. One way to ensure this condition is to place all observations with a sampling rate of one at the beginning of the file sequence. If two strata are collapsed, the resulting sampling rate for the new stratum is computed as a function of the old sampling rates and and number of elements in the previous strata.

4.4 Clusters of Size One

If clusters of size one within a stratum at the first stage, collapsing of adjoining strata ensures that variance estimates will be computed. For a multi-stage design, some of the stages may contain single element clusters. For those clusters, no within cluster variation can be computed. There are several ways for handling this situation. One is to assume a zero-variance contribution from those single-element clusters. Another is to collapse them with neighbouring clusters. An alternative is to assume that they contribute a variance equal to the overall within variation of the clusters for which the within variation can be computed. The variance contribution for those stages where some of the clusters are of size one would incorporate this approximation.

5. SOME DESIRABLE FEATURES OF A VARIANCE ESTIMATION PROGRAM

Francis, Heiberger and Velleman (1975) listed criteria useful in evaluating programs in general. In this section, some of the desirable features of a computer program for estimating variance from complex surveys will be listed. These include user's documentation, input controls, printed output and statistical effectiveness. These desirable features will be related to those provided by SUPER CARP and MINI CARP.

User's documentation should consist of a manual which basically tells the user how to use the program. SUPER CARP and MINI CARP both have manuals which explain to the user how to use them. These manuals are structured as follows. They contain an introduction which summarizes the various available statistical options. Data input and command statements used to specify procedures, variables and options are explained and examples are provided to illustrate their use. Since data input and command statements are to be entered in a specific sequence, a flow diagram is provided. The program procedures are described in terms of the formulea used, the numberical techniques employed and some references to the literature.

As stated earlier, the command language used for SUPER CARP and MINI CARP is in the form of code number or alphanumeric codes in fixed card columns. As pointed out by Francis, Heiberger and Velleman (1975), the most computationally efficient command languages employ code number in fixed card columns. The disadvantage of this method is that users may make excessive references to the manual to identify the commands. Procedures and options could have been specified with the addition of a control statement translator which the addition of a control statement translator which would have allowed English like commands. The advantage of this input method is that it is relatively easy to learn. The disadvantage is that the time and effort required for programming this translator can be prohibitive.

The printed output in SUPER CARP and MINI CARP identifies the statistical procedure used and labels the variables used in the analysis. Part of the output refers to the program's version number, name and date it was last updated. This identification can be used to trace and fix bugs in the stated program version. Some informative diagnostic messages are also printed out. These include messages referring to input controls such as attempting to read in more variables than the program has been dimensioned to handle, trying to read too many cluster, improper input format. If some strata contain one cluster, the program will print out list of such strata. If the user requests collapsing of single cluster strata, the resulting strata will be printed out.

SUPER CARP and MINI CARP are written in FORTRAN and in double precision. They can be run on installations that have a FORTRAN compiler with minor modifications to the job control language. They can both be extented to accommodate new statistical procedures. These can be placed in the program in the form of new subroutines which can be connected to existing software in the program.

REFERENCES

[1] Beaton, A.E., Rubin, D.B., and Barone, J.L. (1976), The Acceptability of Regression Solutions: Another Look at Computational Accuracy, Journal of the American Statistical Association, 71, 158-168.

[2] Raj, Des (1968), Sampling Theory. McGraw-Hill Inc.

[3] Raj, D. (1966), Some Remarks on a Simple Procedure of Sampling Without Replacement, Journal of the American Statistical Association, 61, 393-397.

[4] Francis, I., Heiberger, R.M., and Velleman, P.F. (1975), Criteria and Considerations in the Evaluation of Statistical Program Packages, The American Statistician, 29, 52-56.

[5] Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhya, 37, 117-132.

[6] Fuller, W.A. and Hidiroglou, M.A. (1968), Regression Estimation After Correcting for Attenuation, Journal of the American Statistical Association, 73, 99-104.

[7] Fuller, W.A. (1980a), Properties of Some Estimators for the Errors-in-Variables Model, Annals of Statistics, 8, 407-422.

[8] Fuller, W.A. (1980b), Estimation of Measurement Error Models from Cluster Samples. Paper presented at the meetings of the American Educational Research Association, Boston, Massachussetts.

9] Garza-Hernandez, T. and McCarthy, P.J. (1962), A Test of Homogeneity for a Stratified Sample, Proceedings of the Social Statistics Section of the American Statistical Association, 200-202.

[10] Gray, C.B. (1975), Components of Variance Model in Multistage Stratified Samples, Survey Methodology, 1, 27-43.

[11] Hartley, H.O. (1959), Analytic Studies of Survey Data, Instituto de Statistica, Rome, Volume in onore di Corrado Gini.

[12] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D., (1980), SUPER CARP, Survey Section, Iowa State University, Ames, Iowa.

[13] Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1980), MINI CARP, Survey Section, Iowa State University, Ames, Iowa.

[14] Holt, Mary M. (1977), SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, unpublished report, Research Triangle Institute, Research Triangle Park, North Carolina.

[15] Kish, L., and Frankel, M.R. (1974), Inference from Complex Samples, Journal of the Royal Statistical Society B, 36, 1-37.

[16] Konijn, H.S. (1962), Regression Analysis in Sample Surveys, Journal of the American Statistical Association 57, 590-606.

[17] Kopitze, R., Boardman, T.J. and Graybill, F.A. (1975), Least Square Programs - A Look at the Square Root Procedure, The American Statistician, 29, 64-66.

[18] Ling, R.G. (1974), Comparison of Several Algorithms for Computing Sample Means and Variances, Journal of the American Statistical Association, 69, 859-866.

[19] Longley, James (1967), An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User, Journal of the American Statistical Association, 62, 819-841.

[20] Nathan, G. (1972), On the Asymptotic Power for Tests for Independence in Contingency Tables from Stratified Samples, Journal of the American Statistical Association, 67, 917-920.

[21] Rao, J.N.K. (1975), Analytic Studies of Sample Survey Data, Survey Methodology Vol 1, supplement, Statistics Canada.

[22] Rao. J.N.K. (1975), Unbiased Variance Estimation for Multistage Designs, Sankhya C, 37, 133-139.

[23] Rao, J.N.K. and Scott, A.J. (1981), The Analysis of Categorical Data from Complex Sample Surveys: Chi-square Tests for Goodness-of-Fit and Independence in Two-Way Tables, Journal of the American Statistical Association, 76, 221-230.

[24] Shah, B.V. (1974), STDERR: Standard Errors Program for Sample Survey Data, Research Triangle Institute, Research Triangle Park, North Carolina.

[25] Wilkinson, J.H. (1975), The Algebraic Eigenvalue Problem, Oxford: Clarendon Press, 229-233.

[26] Yates, F. (1960), Sample Methods for Censuses and Surveys, Charles Griffin and Sons, London, Third Edition.