Implementing Balanced Replicated Subsampling Designs in SAS™ Software

Roland K. Hawkes, Southern Illinois University

ABSTRACT

The resampling method called balanced replicated subsampling for assessing error variances from complex survey samples is implemented in SAS software. Methods for structuring data files and SAS macros for calculations are discussed. The methods are illustrated with data on livestock ownership, school attendance and children's health, including incidence of malaria, from a survey in rural Zimbabwe.

INTRODUCTION

Most practical household surveys do not use the simple random samples that are discussed in our statistics texts and programmed into the analyses done by our statistical software. Practical survey sample designs involve elements of clustering, stratification and systematic selection that introduce sources of variation that are not accounted for in the theory of simple random samples.

Cluster sampling is the random selection of groups or clusters of cases and the subsequent measurement of all or samples of their members. Observations within clusters are generally more homogeneous or similar to each other than randomly selected observations. So error is introduced into estimators beyond that from sampling individual observations.

Stratification potentially reduces the variability of estimators by sampling independently from groups known to be more homogeneous than the whole population being sampled.

Systematic sampling is the method of selecting one in n observations from a list. It is equivalent to simple random sampling if the list itself in random order. If the list is ordered by a variable that is correlated with the variable whose parameters are being estimated, the variability of estimators will be reduced.

There are textbook formulae for estimating error variances from various kinds of samples - cluster, stratified, multistage designs, &c., But these require calculations that do not accommodate themselves to the abilities of the software we use for analysis. Also, our practical sampling designs incorporate so many kinds of pure design that there are no textbook formulae that pertain. A technical statistician can readily work out expressions for error variances but that is beyond the skills of most researchers and the computational barriers remain formidable.

These difficulties lead many researchers to calculate statistical tests and interval estimates based on assumption of simple random sampling enshrined in the software. Some then warn the reader that they are not really accurate and are included only as rough guides to variability of estimators. Usually no guidance is offered about the degree of the roughness and the reader is left to guess or to ignore altogether these parts of the reports.

An alternative to the analytical methods of the texts is the use of subsampling designs. These methods were pioneered a half century ago by Mahalanobis (1946a, 1946b). The central idea is that no matter what the sample design, multiple samples should be independently selected using that design. The differences between estimates from these independent samples may be used to estimate the error variance. Then these error variances are straightforwardly projected to the entire sample formed from the combined subsamples.

Subsample designs require a minimum of two subsamples else there is no differences between samples. It is often not practical to include many more than two in the design. Hence the resulting estimates of sampling errors are themselves subject to excessive sampling error. However recent developments have led to the idea of *pseudoreplication* which involves subsampling the data *after* they are collected. The sample design requires that two PSUs are selected in each stratum. Then after the data are in, many successive pairs of subsamples are selected. One sample of the pair includes one randomly selected PSU from each stratum while the other subsample is composed of the remaining PSUs. For each pair of samples, the relevant estimates are computed. The estimated variability of these "half samples" can then be straightforwardly projected to estimated the variability of estimators of the total sample.

Methods of replicated subsampling have a long history. Mahalanobis's (1946a, 1946b) basic insight was that multiple independent samples using the same complex design (he called them interpenetrating subsamples) could be used to compare the work of independently operating field teams. The idea was expanded by Deming (1956) who used the variation among subsamples to estimate sample variances. McCarthy (U.S. National Center for Health Statistics 1969a, 1969b) proposed and developed the balanced replication methods used in this paper. Kish and Frankel (1970) reported an evaluation of balanced replication designs which helped to establish them in the statistical armamentarium of sample designers.

A good recent introduction (along with bootstrap and jackknife methods) is chapter 7 "Resampling and Variance Estimation in Complex Surveys" in Chaudhuri and Stenger (1992).

ISSUES IN THE PRACTICAL IMPLEMENTATION OF RESAMPLING METHODS

The existence of these methods has been of no help to most researchers. Large research groups conducting ongoing institutionalized surveys have the advantage of staffs of statisticians and programmers and of specialized software that combine to produce estimates of the variability of their estimators. The individual researcher or the small group conducting one-time projects do not have these resources.

Implementing the analysis of complex sample designs as a routine part of data analysis by the individual researcher and the small organization should meet a number of criteria.

> Analysis should be routinized in procedures that can be used by research assistants that have only an elementary knowledge of statistics.
>
> The programming for analysis should be at a level that can be handled inside the project and should not require the assistance of outside personnel.

Analysis should be integrated into the statistical program used for data management and analysis and not require passing off problems to supplementary software.

SAS software seems to have the potential to meet these criteria.

The macro facility puts programming within the reach of anyone that has a basic grasp of the structure of SAS.

The ability to save results to files and to use them in subsequent procedures allows the maximum use of existing procedures and the minimization of new programming.

The IML™ procedure for matrix operation allows the programming of statistical calculations by anyone trained to an intermediate level in statistics.

Of the many packages that will handle the computing of the newer resampling methods of analysis, SAS provides the most far reaching and flexible facility for managing large and messy data sets. Files can be combined and split. Missing values are routinely dealt with.

These abilities have prompted much recent interest in the use of SAS software to implement resampling methods as evidenced by the work of Chenier and Vos (1996), Hur, *et al.* (1966), Hamada (1995) and Hood (1995). At the 1996 SUGI conference, Thompson (1996) offered a tutorial in the use of SAS software with bootstrap methods. Hallinan (1995) provides a good general introduction to using SAS software with modern resampling methods.

So SAS is a strong candidate for implementing the routine use of replicated subsampling and other resampling procedures. This paper explores how that can be done.

AN APPLICATION - A HOUSEHOLD SURVEY IN RURAL ZIMBABWE

To illustrate the implementation of the methods I use data from survey research done in 1991 by the Centre for Applied Social Sciences of the University of Zimbabwe. The work was done for the USAID Natural Resources Management Project (NRMP) (United States Agency for International Development project number 690-0251).

The aim of NRMP was to implement locally based wildlife management programs in four administrative districts of the Matabeleland provinces of western Zimbabwe. A baseline socioeconomic survey was mandated by the project design. The task of that survey was to develop information about the communities involved in the project.

Each district is divided into wards. Each ward is in turn divided into villages. The boundaries of wards and villages were established following independence, in 1980, in accord with a central government mandate that villages should be of a size to contain about 100 households and that a ward should contain about six villages. Those instructions were carried out reasonably accurately and can be rough guides to the sizes of our units. However, these recently introduced political boundaries may not correspond to demarcations that are otherwise meaningful in the lives of the places.

Twenty six wards were selected from the four districts with local advice about their centrality to the project[1]. In each of the wards, interviews were done in each village. In each village, two independent samples of households - here I will call them clusters - were selected. Interviews were conducted in more that three thousand (3241) households. Information was recorded on almost thirty thousand (29148) residents of those households (including part time and seasonal residents).

The method of selecting clusters varied from district to district. It depended on the kind of information that was locally available, on the structure of traditional leadership and on the size of potential sampling units within villages. Very often the necessary information was not available without site visits and extensive local consultation. This meant that design decisions were made as the work proceeded rather in ideal textbook fashion. However, the pattern of two independently and identically selected clusters per village was constant throughout the entire exercise.

In this paper, I select two of these wards to illustrate the method of sample analysis. These are Ndolwane ward in Bulilima Mangwe District and Sinakoma ward in Binga District.

In Ndolwane ward as with other wards in the district, there were no central lists of households to sample from. However, The traditional leadership structure is largely intact and well understood as are ward and village boundaries. The traditional official at the lowest tier is the *sabhuku*. The research team was able to list sabhukus in each village in a fairly accurate way. Then two were randomly selected and all the households in their domains were included in the sample. So for each village in Ndolwane ward cluster represents all the households in the domain of a randomly selected *sabhuku*. This design yielded a sample size of about twenty percent of the population of households.

In Sinakoma (and the rest of Binga District) the ward councillor held very complete records of the households in the villages. Here the sampling was easy (once the travel to some very remote places was accomplished). We simply took one in ten systematic samples from the village lists. To preserve the basic design of selecting two independent clusters from each, two such samples were selected resulting in a twenty percent sample of households[2].

The Ndolwane sample is technically a *stratified* (by village) *cluster* (the sabhuku) *sample*. In practice, we expect estimators from cluster samples to be more variable than those from simple random samples of the same size. That is because observations within clusters are likely to be more similar to each other than randomly and independently selected observations. There is intracluster correlation among the observations. For example, if we find one household under a sabhuku to be cattle poor, we would expect the others also to tend that way. Similarly, we would expect the reverse - that cattle wealthy households will be clustered together. So in selecting two cluster we could easily get two poor groups or two rich groups and be quite wide of the mark in our estimates compared to randomly selecting the same number of households individually[3].

The sample design used in Sinakoma (and the rest of Binga district) is a different matter. Usually, systematic samples are treated as simple random samples. This is justified if the order of the list from which selections are made is essentially random with respect to the parameters being estimated. If the ordering of the list is correlated with the values of the variables whose parameters are being estimated, the assumption of simple random sampling is conservative - that is, estimators will be less variable than the model implies. In any case, the design used here will permit the

estimation of the variance of estimators regardless of the nature of the lists. We expect in advance that estimators will be about as variable as those defined on a simple random sample of equivalent size[4].

When the characteristics of individual household members are the focus of attention, then the household also becomes a cluster. This makes the analysis even more prohibitive with textbook methods. However the subsampling method is as easy to apply and as appropriate as with the households themselves as the units of analysis.

THE LOGIC OF RESAMPLING METHODS

The central idea of these resampling methods is that the variability of parameter estimates between the independently selected subsamples may be used to estimate the variance of estimators defined on the entire combined sample. This neatly avoids the problem of figuring out exact expressions sampling variances from complex designs and for intractable nonlinear estimators. Of course, it introduces its own problems of variance in the variance estimators themselves. I do not address that issue here.

The development here closely follows Sudman (1976).

B is the parameter to be estimated.

b is the statistic used by the researcher to estimate B.

$b^*_j=(b_j+b_j')/2$ is the mean of b for the jth replicate and its complement. (The complement of any replicate consists of all the PSUs not included in the replicate.

$var(b^*_j)=(b_j-b_j')^2/4$ denotes the estimate of the variance of b^*_j.

$var^*_k(b^*)$ denotes the average over k replications of the $var(b^*_j)$.

The next step in the logic is that of *pseudoreplication.* This involves resampling the data after they are collected but in a way that reflects the sampling procedures. The simplest application involves samples which are stratified and in which two PSUs have been selected from each stratum. This is exactly the sample design that was followed in the NRMP survey. The method consists of selecting half samples or *replicates* - the observations in one randomly selected PSU of each pair. Each replicate is compared to its complement - observations in the remaining PSUs.

One replication may be used in the simplest case. But in the practical case many random replications are generated to increase the precision of the variance estimates. A further step is the use of balanced replications. This method selects subsamples that are balanced in the sense that the successive replicates are orthogonal. Designs for 4, 8, 12 ,16, 20 and 24 following Placket and Burman (1946) are given by Sudman (1976, pp. 179 ff.). If the number of strata is not a multiple of

Table 1. Balanced Replication Designs for Computing Sampling Errors When N=8.

	Pair							
Replication	1	2	3	4	5	6	7	8
1	+	-	-	+	-	+	+	-
2	+	+	-	-	+	-	+	-
3	+	+	+	-	-	+	-	-
4	-	+	+	+	-	-	+	-
5	+	-	+	+	+	-	-	-
6	-	+	-	+	+	+	-	-
7	-	-	+	-	+	+	+	-
8	-	-	-	-	-	-	-	-

Table 2. Cluster Codes and Replication Patterns for Villages of Two Wards in the Natural Resources Management Program Socioeconomic Survey.

Ward and Village	Cluster codes A	Cluster codes B	Replication Patterns Cluster A	Replication Patterns Cluster B
Ndolwane Ward[1]				
Tematema	11	12	1,1,1,2,1,2,2,2	2,2,2,1,2,1,1,1
	9*	10*	.,.,.,.,.,.,.,.	.,.,.,.,.,.,.,.
Ndolwane	1	4	2,1,1,1,2,1,2,2	1,2,2,2,1,2,1,1
Khuphula	17	18	2,2,1,1,1,2,1,2	1,1,2,2,2,1,2,1
Khame	7	8	1,2,2,1,1,1,2,2	2,1,1,2,2,2,1,1
Tshemahale	5	6	2,1,2,2,1,1,1,2	1,2,1,1,2,2,2,1
Hungwa	2	3	1,2,1,2,2,1,1,2	2,1,2,1,1,2,2,1
not recorded	.	.	.,.,.,.,.,.,.,.	.,.,.,.,.,.,.,.
Sinakoma Ward[2]				
Nampande	200	201	1,1,1,2,1,2,2,2	2,2,2,1,2,1,1,1
Matala	202	203	2,1,1,1,2,1,2,2	1,2,2,2,1,2,1,1
Bundimba	204	205	2,2,1,1,1,2,1,2	1,1,2,2,2,1,2,1
Chinga	206	207	1,2,2,1,1,1,2,2	2,1,1,2,2,2,1,1
Dangamuseye	208	209	2,1,2,2,1,1,1,2	1,2,1,1,2,2,2,1
Luzava	210	211	1,1,1,2,1,2,2,2	2,2,2,1,2,1,1,1
not recorded	.	.	.,.,.,.,.,.,.,.	.,.,.,.,.,.,.,.

1. Ndolwane ward is in Bulilima Mangwe district. Each cluster represents all the households in the domain of a randomly selected sabhuku - the lowest level of the traditional leadership structure.

2. Sinakoma ward is in Binga district. Each cluster is a one in ten systematic sample of households from the lists kept by village development committee chairmen.

. Indicates unrecorded data.

*. Clusters 9 and 10 in Ndolwane ward were added to the list of clusters to allay a local political dispute that ensued when the list of local leaders was alleged to exclude at least one important group. They are not included in the analysis. They are retained in the presentation to remind us that we are dealing with real and fallible data with all its flaws and imperfections.

four, then the next largest design are used and the final columns are omitted. Table 1 gives the balanced design for n=8 which is the proper choice for five through eight strata. It is the one I use in the present application where the number of villages per ward varies from five to seven. In the two wards I use for illustration, the number of villages in each is six. Hence the last two columns of the design matrix are ignored.

For each replication, + in the design matrix indicates that the first PSU of the pair is included in the replicate while - indicates that the second PSU is selected. The remaining PSUs are included in the complement. For example, in the first replicate with our six strata the first replicate is composed of the first PSU from the first stratum, the second PSU from the second stratum and the second, first, second and first PSUs from the third through sixth strata respectively.

STRUCTURING DATA FILES FOR BALANCED REPLICATED SUBSAMPLING

These replication patterns may be coded into the files of data to be analyzed. Each observation may be coded on eight variables that indicate which half sample - the replicate or the complement - it is to be assigned on each of the eight replications.

Table 2 shows how observations in each cluster are coded on each of the eight replications. The codes are:

1 - include in the replicate
2 - include in the complement
. - exclude from analysis because of missing data, etc..

Eight new variables - REP1 through REP8 - containing this information were created and included in the data files. Values were assigned that corresponded to the replication patterns associated with the cluster identification numbers[5]. For example, the first cluster in Tematema village (in column A and coded as 11) is included in replicates in replications 1,2,3 and 5 and in the complement in the others. The second cluster (in column B and coded as 12) necessarily follows the exactly opposite pattern.

A SAS MACRO FOR ESTIMATING STANDARD ERRORS OF MEANS

Once these variables are included in the data, it is a straightforward matter to sort by each of them in turn and to use any SAS procedure that supports the BY instruction to generate statistics for the various subsamples. Then the SAS IML procedure may be used to do the statistical calculations over the various replications. The process may be written into a SAS macro.

The most basic SAS macro for calculating standard errors of means is Program 1. It provides a skeletal structure than can be used for implementing the calculation of standard errors for any statistics that can be output from SAS procedures that use the BY instruction. It may easily be expanded to provide more extensive and sophisticated output.

Step A begins the macro with LIST being the names of the variables to be included in the analysis and DSN being the name of the SAS dataset to be used. The macro presumes that REP1 to REP8 with the assignment information for the eight replications are also in the dataset. These are read into a temporary dataset and REP0 is computed to be 0 if an observation is included in the total sample to be analyzed.

Program 1. Standard Errors of Means from Balanced Replicated Subsamples

A basic SAS macro to calculate standard errors of means.

```
/*Standard Errors of Means from Replicated Subsampling Design.         */
/*                      R.K.H. - March, 1997                           */

/* A - Macro invocation and definition of temporary dataset.           */

    %MACRO REPMEANS (LIST, DSN=); /* LIST = variables, DSN = dataset.  */
       DATA REPTEMP;SET &DSN;     /* Create temporary dataset          */
       KEEP REP1^QREP8 &LIST;      /*  with replications and variables.*/
       IF (REP1 NE .) THEN REP0=0;/* REP0=0 for included observations. */

/* B - Calculate means for observations, replicates and complements.   */

   %DO I= 0 %TO 8;                /*For all replication patterns:      */
      DATA ;SET REPTEMP;          /*   Use temporary dataset.          */
      REP=REP&I;                  /*   Use i'th replication.           */
      PROC SORT; BY REP;          /*   Sort by replication.            */

      PROC MEANS NOPRINT; BY REP; /*Calculate means of variables by    */
        VAR &LIST;                /*   replication and save to a       */
        OUTPUT OUT=TEMP MEAN=;    /*   temporary dataset.              */

     PROC APPEND BASE=REPSTATS    /*Add means to REPSTATS, the entire  */
        DATA=TEMP;                /*   collection of means.            */

  %END;                           /*Loop to next replication.          */

/* C - Use matrix processing procedure for calculations.               */
  PROC IML;                       /*Invoke matrix processor.           */

  PRINT 'Balanced Replicated Subsampling for Means'.; /* Title         */

  USE REPSTATS;                                 /*Read means for:      */
  READ ALL VAR {&LIST} WHERE (REP=0) INTO B;  /*  whole sample.       */
  READ ALL VAR {&LIST} WHERE (REP=1) INTO BR; /*  replicates.         */
  READ ALL VAR {&LIST} WHERE (REP=2) INTO BC; /*  complements.        */

  IF ((NROW(BR) NE 8) OR (NROW(BC) NE A))     /*Halt execution if some */
     THEN GOTO FINISH;                        /*  groups are empty.   */

  DIFF=BR^QBC;                     /*Calculate differences,           */
  VARCO= (DIFF'*DIFF)/(4*8);      /*  covariance matrix               */
  STERR= (VECDIAG(VARCO))##.5;    /*  and standard errors.            */

  ROW={&LIST};                    /*Make row labels                    */
  COL={"MEAN" "ST.ERROR"};        /* and column labels.                */
  STATS=B'||STERR;                /*Construct output matrix.           */
  PRINT STATS                     /*Print labelled output.             */
     [ ROWNAME=ROW COLNAME=COL FORMAT=10.2];

  %FINISH:;                /*Label for abend.                          */
  QUIT;                    /*Exit IML procedure.                       */

/* D - End macro.                                                      */
  %MEND;
```

Some statistics about ownership of livestock by households in Ndolwane ward generated by the SAS macro.

```
                        The SAS System

       Balanced Replicated Subsampling for Means        .


                  STATS          MEAN   ST.ERROR

                  CATTLE        5.969      0.599
                  DONKEYS       2.352      0.313
                  GOATS         9.227      0.644
```

The tabulation is of households. The variables are:

CATTLE = number of cattle owned.
DONKEYS = number of donkeys owned.
GOATS = number of goats owned.

Otherwise it is missing.

Step B iterates through the replication patterns, calculates and outputs means by replicate and complement and adds them to the ones already calculated. The calculation includes the use of REP0 which has the effect of producing means for all the observations which are included in the analysis.

Program 2. Modifications to Program 1 to Extend Output and Generate Statistics for Estimates of Proportions.

This program presumes that the input data is coded as 1 if an attribute is present and 0 if it is absent. Hence the MEANS procedure is used to calculate proportions.

Step B is modified to save the sample sizes on which statistics are based.

```
PROC MEANS NOPRINT; BY REP;
  VAR &LIST;
  OUTPUT OUT=TEMP MEAN=;
  OUTPUT OUT=TEMPN N=;                          */Output sample sizes.   /*
  PROC APPEND BASE=REPSTATS
  DATA=TEMP;
  PROC APPEND BASE=N DATA=TEMPN;                */Save sample sizes.     /*
```

Step C is modified to read sample sizes and to calculate and print 95% confidence intervals and design effects.

```
USE REPSTATS;
READ ALL VAR {&LIST} WHERE (REP=0) INTO B;
READ ALL VAR {&LIST} WHERE (REP=1) INTO BR;
READ ALL VAR {&LIST} WHERE (REP=2) INTO BC;
USE N;                                  */Input total sample    /*
READ ALL VAR {&LIST} WHERE (REP=0) INTO N;      */ sizes into N.        /*
IF ((NROW(BR) NE 8) OR (NROW(BC) NE 8))
   THEN GOTO FINISH;

DIFF=BR-BC;
VARCO= (DIFF'*(DIFF)/(4*8));
STERR= (VECDIAG(VARCO))##.5;

ROW={&LIST};
COL={"PROPORTION" "N" "ST.ERROR" "LOWER" "UPPER" DESIGN};
LOWER=B'-1.96#STERR;                    */Calculate lower and           /*
UPPER=B'+1.96#STERR;                    */  upper confidence limits.    /*
SIMPLERR=(B#((B/B)-B))/N;               */Calculate standard error for  /*
                                        */ simple random sampling.      /*
DEFF= (STERR##2)/SIMPLERR';             */Calculate design effect.      /*

STATS=B']]n']]STERR]]LOWER]]UPPER]]DEFF;*/Construct and print           /*
PRINT STATS                             */ formatted output matrix.     /*
  [ROWNAME=ROW COLNAME=COL FORMAT=8.3]; */                              /*
```

Some characteristics of school age children in Sinakoma ward produced by the SAS macro.

```
                          The SAS System

       BALANCED REPLICATED SUBSAMPLING FOR PROPORTIONS         .

STATS     PROPORTION          N   ST.ERROR      LOWER      UPPER     DESIGN

MALE           0.554    112.000      0.048      0.459      0.648      1.055
INSCHOOL       0.616    112.000      0.033      0.552      0.680      0.502
BADHEALTH      0.179    112.000      0.058      0.065      0.292      2.560
MALARIA        0.107    112.000      0.027      0.053      0.161      0.882
```

The tabulation is of children aged 7 through 14 - approximately primary school age - in Sinakoma ward in Binga district. The variables are proportions indicating that the child:

MALE = is male.
INSCHOOL = is currently in school.
BADHEALTH = has many health problems chronic illness.
MALARIA = has had malaria within the past year.

Step C invokes the IML procedure. The matrices B, BR, and BC are the collection of means for the entire sample, for the eight replicates and the eight complements respectively. These are read from the file of means calculated in step B. If there are not eight replicates and eight complements (perhaps because small subsets of data are being analyzed) the procedure is halted. Otherwise the standard errors are estimated from differences between replicates and complements[6]. Finally, labeled output of means and standard errors is produced. Then control is passed back to the program that invoked the macro.

Output is shown for the numbers of various kinds of livestock owned by households in Ndolwane ward. It is a bare minimum of useful information but it gets the job done and serves to illustrate the procedure.

A SAS MACRO FOR ESTIMATING VARIANCES OF PROPORTIONS: AN EXAMPLE WITH MORE EXTENSIVE OUTPUT.

Program 2 illustrates changes that extend the basic structure to report proportions, to report confidence intervals and design effects. The design effect of an estimator is the ratio of its variance to the variance of the estimator for a simple random sample of the same size.

The step B is modified to save the sizes of the various samples which are required in the calculation of design effects. The IML procedure in step C is expanded to read sample sizes, and to calculate lower and upper bounds for a 95% confidence interval (with z=1.96 assuming approximate normality). Further it calculates the design effect. The output is appropriately reconstructed to accommodate the new statistics.

Output is shown for an application of the new macro. The observations are all children of primary school age from households in Sinakoma ward. The statistics are about sex, school enrollment and health problems including malaria which is common and life threatening in the area. The output explains itself but the design effects deserve comment. The very low (0.502) estimated effect for school attendance suggests that there may be negative intra-household correlation. That is if one child attends school, the others are likely not to. The high (2.560) effect for health problems may indicate that frequent illness may be common to all children within households. This is not the place to examine these questions. However, the outcomes illustrate that once the calculations are done the results pose interesting and pursuable questions.

CONCLUSION

This paper has illustrated the use of SAS software to analyze sample survey information using the resampling method known as balanced replicated subsampling. It has shown that SAS provides structures that lend themselves to practically implementing this analysis which requires both the use of large and messy datasets and the ability to do sophisticated calculations.

The single researcher or the small research group doing one time sample surveys can have access to balanced replicated subsampling analysis using SAS. It does not require investment in extra software or the services of technical statisticians and programmers.

NOTES

1. Wards were used because the project was implemented at the ward level. For another approach to sampling in Zimbabwe, using census enumeration areas, see Adamchak and Mbizvo (1991).

2. In two other districts, not discussed here, the sizes of potential sampling units within villages turned out to be quite variable. A method was contrived to combine the smaller ones and to subsample the very very large ones when they were chosen. Records were kept that allow the calculation of weighting factors in constructing estimates. The same method of balanced replication may be applied to those wards.

3. If there is no intracluster correlation, each cluster "looks like" a simple random sample and a cluster design is as efficient as simple random sampling. If there is heterogeneity within clusters - negative intracluster correlation - then estimates from a cluster design are even less variable than those from simple random sampling.

4. In the case of lists that are cyclically ordered, estimates may be more variable than those from simple random samples. I do not discuss that case here except to note that the replicated subsample design will take that kind of ordering into account also.

5. There are many ways to do this in the data preparation process. I will not discuss them here.

6. Note that the entire covariance matrix of the errors is calculated as an intermediate step. This is done to permit the easy expansion of the algorithm to the construction of simultaneous confidence intervals.

REFERENCES

Adamchak, Donald T. and Michael T. Mbizvo (1991), "Family Planning Information and Media Exposure among Zimbabwean Men," *Studies in Family Planning*, 22, 326-331.

Chaudhuri, Arijit and Horst Stenger (1992), *Survey Sampling: Theory and Methods*, New York: Marcel Dekker.

Chenier, Thomas and Paul Vos (1996), "Beating Student's T and the Bootstrap -- Using SAS Software to Generate Conditional Confidence Intervals for Means," *Proceedings of the Twenty First Annual SAS Users Group International Conference*, 21, Paper 245.

Deming, W. Edwards (1956), "On Simplification of Sampling Design Through Replication With Equal Probabilities and Without Stages," *Journal of the American Statistical Association ,* 51, 24-53.

Hallinan, Charlie (1995), "Data Analysis Using SAS." *Sociological Methods and Research*, 23, 373-391.

Hamada, Chikuma (1995), "Bootstrap Cox Regression Using SAS Software," *Proceedings of the Twentieth Annual SAS Users Group International Conference*, 20, Paper 282.

Hood, Kelley and Sylvia Miller (1995), "Resampling Using the SAS System," *Proceedings of the Twentieth Annual SAS Users Group International Conference*, 20, Paper 272.

Hur, Kwan, Charles A. Oprian, William G. Henderson and Bharat Thakkar (1996), "A SAS Macro for Validating a Logistic Model with Split Sample and Bootstrap Methods," *Proceedings of the Twenty First Annual SAS Users Group International Conference*, 21, Paper 167.

Kish, Leslie and Martin Frankel (1970), "Balanced Repeated Replications for Standard Errors," *Journal of the American Statistical Association*, 65, 1071-1091.

Mahalanobis, P. C. (1946a), "On Large-scale Sample Surveys," *Philosophical Transactions of the Royal Society of London, Series B*, 231, 329-451.

Mahalanobis, P. C. (1946b), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society*, 109, 326-370.

Placket, R.L. and P.J. Burman (1946), "The design of Optimum Multifactorial Experiments," *Biometrika*, 33, 305-325.

Sudman, Seymour (1976), *Applied Sampling*, New York: Academic Press.

Thompson, Paul A. (1996), "A Tutorial on Bootstrapping in the SAS System," *Proceedings of the Twenty First Annual SAS Users Group International Conference*, 21, Paper 243.

U.S. National Center for Health Statistics (1969a), *Replication: An Approach to the Analysis of Data From Complex Surveys*, Vital and Health Statistics, Series 2, No. 14. Washington,D.C.:U.S. Government Printing Office.

U.S. National Center for Health Statistics (1969b), *Pseudoreplication: Further Evaluation and Application of the Balanced Half-sample Technique*, Vital and Health Statistics, Series 2, No. 31. Washington,D.C.:U.S. Government Printing Office.

ACKNOWLEDGEMENTS

Data used in this paper were gathered by The Centre for Applied Social Sciences, University of Zimbabwe for the Natural Resources Management Project. That project was supported by the United States Agency for International Development (USAID) project number 690-0251.

SAS computing for this paper used the computer facilities of Southern Illinois University, Carbondale.

AUTHOR INFORMATION

Roland K. Hawkes
231 Krysher Road
Makanda, Illinois 62958

phone/fax: 618 549 5885
email: hawkes@siu.edu
Internet: http://www.siu.edu/~hawkes/