SAMPLING ERRORS IN THE INTEGRATED SYSTEM FOR SURVEY ANALYSIS (ISSA)

Guillermo Rojas, Alfredo Aliaga, Macro International
8850 Stanford Boulevard Suite 4000, Columbia, MD 21045

I-INTRODUCTION.

This paper describes the approach followed by the Demographic and Health Surveys Program (DHS) to calculate its sampling errors. DHS is a 15-year large-scale survey program which is providing financial and technical assistance for more than 90 surveys in 50 developing countries in Africa, Asia, the Near East, and Latin America. Approximately 60 surveys have been completed to date, and field work is underway in seven others.

Because of a need for a quick turnaround of data for analysis and report writing. DHS decided to develop an integrated system that would meet its processing needs. In 1985 a team of programmers in collaboration with DHS staff, began developing the Integrated System for Survey Analysis (ISSA).

ISSA is a menu-driven microcomputer package which meets nearly all the data processing needs of any large scale survey. With ISSA questionnaire design, data entry, data editing, secondary file creation, and tabulations, can be accomplished.

A simple ISSA program requires at least two components, an ISSA dictionary and an ISSA program. In the dictionary all the information relevant to the variables are stored. An ISSA program, which in ISSA terminology is known as an "application", uses the variables defined in the dictionary and the ISSA language to execute the desired calculations.

The ISSA programming language is powerful. It includes variable assignation using simple or complex mathematical expressions, arithmetic and string handling functions, recoding (BOX), looping (WHILE-DO-ENDDO), and boolean and conditional statements (IF-THEN-ELSE-ENDIF). The user can define his own functions as well.

Of special interest for sampling errors is the tabulation module of ISSA. A tabulation application is comprised of two sections, a declaration section and a logic section. In the declaration section, tables or means are declared using the TABLE or MEAN declaration. In the logical section, the ISSA language is used for data transformation, selection of cases, definition of the units of analysis, and the actual tallying of the tables.

When ISSA was first designed, no consideration was given to the calculation of sampling errors. DHS have been using the program "clusters" developed by the World Fertility Survey Program (WFS) to calculate its sampling errors. Clusters has some limitations, but perhaps the most important is that sampling errors for fertility and mortality rates, which are one of the most important estimators of our surveys, can't be calculated.

Because of that DHS decided to design and implement an interface with ISSA that would facilitate the calculation of sampling errors. The interface would meet the following specifications.

1-To have an integrated system to carry out all the data processing activities, from the design of the questionnaire to the production of the final report.

2-Three methodologies would be available: The ultimate cluster approach (a linear Taylor series approximation), the Balanced Repeated Replication Model (BRRM), and the Jackknife model.

3-With the BRRM and Jackknife models, sampling errors for statistics such as mortality and fertility rates can be obtained.

4-No new language would need to be learned.

5-No external files would need to be created.

6-Sampling errors for several units of analysis using the same data file, can be calculated. In the case of DHS sampling errors can be calculated for household information, households members, women, and children.

7-The computer requirements are the same as those needed to run a regular ISSA program.

II-ISSA INTERFACE

Two modules are required to calculate sampling errors using ISSA. In the first module ISSA is provided with an application similar to that described for tabulations. The application will create a TBX, TBD, and DIC files as output. The second module takes the files created in the previous step and calculates the sampling errors according to a methodology specified by the user. It is important to note that once the first module has run, the three methodologies can be applied without rerunning the first module.

Files created by an ISSA application

The TBX file has one record for each cluster or ultimate sampling unit (USU). Each record contains the USU code and the disk address where the tables/means for that USU are stored.

The TBD has three different components. The first part contains general information about the application. The second part contains a description of each table/mean generated by the application. The third section has one block for each USU or cluster in the input file. The block holds the cumulative tables or means for each cluster.

The DIC file has basically the same structure as any other dictionary defined in ISSA. It is used for purposes of printing the final results.

In order to implement the interface three new commands were incorporated into the ISSA language. These commands are BREAK BY, SMEAN, and STABLE and their syntax is as follows:

BREAK BY <var expresion>

The BREAK BY command, specifies the variable(s) which holds the ultimate sampling unit or cluster number. Each USU will generate an entry in the TBX file and the tables for that USU will be written down in one area of the TBD file. In almost all DHS surveys this variable is the cluster number, but it could be a list or a concatenation of variables.

Examples,

```
BREAK BY cluster;
BREAK BY provin+cluster;
BREAK BY provin+depto+cluster;
```

SMEAN name Vx[/Vy] <dcl expresion> [<dcl expresion>]

The SMEAN (Sampling Means) declaration was created to calculate sampling errors for statistics such as means, proportions, and ratios. This declaration is exactly the same as the regular MEAN ISSA declaration where Vx is the dependent variable. However, SMEAN can have the variable Vy which is used to calculate ratios.

Examples,

```
SMEAN TS01 evborn urbrur
SMEAN TS02 second urbrur
SMEAN TS03 hcard/child59 urbrur
```

STABLE name <dcl expresion> [<dcl expresion>]..

This declaration works exactly the same way as the regular tables declaration of ISSA. With this command sampling errors for statistics such as mortality, fertility, medians, and indices, can be calculated. These statistics cannot be obtained using SMEAN, because there is no one to one relationship between numerators and denominators. Example, to calculate fertility

rates it is necessary to tally separately the numerator (years of exposure of a woman in a specific group of age) from the denominator (number of births a woman had given birth in a specific age group). In order to properly calculate the rates, some post-processing manipulations to the resulting tables are required after the entire file has been tallied.

III-SAMPLING ERRORS EXECUTION

Once an application with the previous commands runs the sampling errors can be calculated. In order to do that the user must provide a parameters file. The first instruction in the parameters file is the type of model that is going to be used. The names of the models are CLUSTERS for the Taylor series approximation, BRRM for the balanced repeated replication method, and JACKKNIFE.

Clusters Model

To use clusters the user needs to specify in the parameters file how the strata (the way clusters will be paired) is going to be conformed. To tell the program which clusters belong to the same stratum, it is only necessary to type the cluster number separated by commas and ending with a semicolon. The dash (-) can be used to specify a range of clusters.

Example,

```
1, 3;
2, 4, 7;
9, 12;   10-11, 14;
16, 18;  17, 19-20;  21,24;
```

To guarantee that all the USUs of the sample are included in the run, the program will check that all the clusters and only those found in the TBX file have been defined in the parameters file as part of a stratum, otherwise the program will abort.

The BRRM Model

The BRRM method is based on the calculation of sampling errors using half samples. These half samples are determined by an ORTHOGONAL MATRIX which is made up of rows named replications and columns called pairs. The number of pairs is determined by the number of clusters divided by two. For this reason the BRRM approach requires that the number of clusters must be an even number. Each pair is composed of two clusters (one positive and one negative). Because of this peculiarity, the first pair is composed of clusters 1 and 2, the second of clusters 3 and 4, and so on. The odd clusters are positive and the even ones are negative.

The optimal number of half samples (replications) is a number that is greater than or equal to the number of pairs and at the same time multiple of four. The way the BRRM model operates, is to pass across the TBD file for each replication, selecting only the clusters specified by the orthogonal matrix. After passing through all the replications, the model will calculate the of variance among replications to produce the sampling errors.

It is important to know that the program will automatically generate the orthogonal matrix based on the number of clusters of the survey. In order to do so, the system is provided with a file containing primitive Hadamard matrices. This primitive matrices can be used to generate other matrices based on the Hadamard property that states that a matrix of grade H can be used to generate a matrix of grade 2 times H. This can be done by copying the primitive matrix to its right, bottom, and diagonal. In the diagonal portion, the sign of the primitive matrix is changed.

```
H    H
H   -H
```

For example, the matrix of 4 can be used to generate the matrices of 8, 16, 32, 64, 128, 256, 512, 1024. The matrix of grade 12 will generate the matrices of 24, 48, 96, 192, 384, 778. Currently the file has the primitive matrices of 4, 12, 20, 44, 60, 68, 76, 84, 100.

In the case of SMEAN declarations (means, proportions, ratios), it is known in advance which are the numerators and denominators. However that is not the case with rates and medians which are tallied via the STABLE declaration. In general the table declaration requires manipulation of the tables generated in the first module in order to calculate the desired estimator. One of the peculiarities of ISSA is that it allows the user to manipulate the resulting matrices of a table declaration after the input file is read. Essentially with this calculations, the user can modify the table or create new tables based on the original ones. This type of manipulations are essential to calculate estimators as fertility and mortality rates.

However, once the sampling errors program comes into play, there is no access to the ISSA language. To solve this problem, an interpreter, which is a subset of the post-processing operations allowed by ISSA, was written.

The interface with the user to write this commands is through the parameters file. All the relevant syntax is checked and comments are allowed as in any ISSA application. With this language operations like the ones used to calculate fertility and mortality rates can be accomplished.

The JACKKNIFE Model

The Jackknife approach works the same way as the BRRM model. However, in this case there is no orthogonal matrix; instead, the number of strata and replications equals the number of clusters in the sample.

In this model there are no restrictions on the number of clusters. Replication number one will include all the clusters except the first one; replication number two will exclude the second cluster, and so on. Jackknife operates basically the same way as BRRM and everything said for BRRM also applies for Jackknife. The formula to calculate the variance is slightly different of that used by BRRM.

IV-COMPARATIVE RESULTS

A number of studies have been conducted to compare the quality, both theoretical and empirical, of the linearization(Clusters approach) and the replication(BRRM and Jackknife) methodologies. Kish and Frankel simulated a number of multistage sample designs. They compared properties of those methodologies for a number of estimators. They found a similarity in the sampling variance calculation for the three methods and concluded that there was evidence that linearization gave greater accuracy in variance estimation. They also discovered that the replication approach, and in particular the BRRM, gave confidence interval coverage which was slightly closer to the nominal coverage rate. Several other studies have indicated that the linearization and Jackknife are superior to the BRRM in terms of the resulting quality of variance estimation.

To probe what the studies have shown, in this section we are going to present sampling errors according to the three methodologies for the estimators listed below (see Appendix). These errors will be presented for six DHS countries.

- Mean of children ever born for all women.
- Contraceptive prevalence rate for currently married women.
- Proportion of children 12 to 23 who have a health card.
- Infant mortality rate.
- Total fertility rate.

Differences in sampling errors among the three methodologies are relatively small. However, for every estimate in every country, the clusters model consistently gives the smallest sampling error, and the BRRM always gives the highest.

Looking at the average of the relative errors for each variable through all six countries, the same pattern described in the previous paragraph can be observed. The following table shows that pattern.

Average Relative Error (%)

	BRRM	Jackk	Clusters
Children ever born	3.2	2.2	1.8
Contraception use	6.0	4.2	3.2
Children with cards	5.8	4.3	4.1
Infant mortality rate	11.6	8.5	
Total fertility rate	4.1	2.9	

When sampling errors are analyzed according to subpopulation and type of variable, it can be concluded that there is a relationship between the combination of such factors and the sampling errors. When the variable is universal in a large subpopulation the differences of the average relative error between methodologies become smaller. This can be seen with the "children ever born" variable which has the lowest sampling error as well as the lowest range of variation for the three methodologies. This range of variation goes from 1.8% to 3.2% with an absolute difference of 1.4%. In contrast, the sampling error for "contraception use" is higher because the subpopulation in this case is restricted to currently married women. The difference between methodologies in this case ranges from 3.2% to 6% with an absolute difference of 2.8%.

There is also a relationship between the type of event and the magnitude of the sampling error. If the event is not of common occurrence, then the sampling error is likely to be high. This can be seen for the "infant mortality rate", which has the highest sampling error among the estimators included in this analysis. The average relative error ranges from 8.5% to 11.6% between methodologies. This is mainly due to the fact that death in the first year of life is rather a rare event.

Since the linear (Clusters model) approach gives the smallest sampling errors, calculation of the relative loss of precision for the Jackknife and BRRM with respect to clusters show that Jackknife loses less precision than BRRM and that the loss is reduced when the estimator has been post-stratified. This is the case with the estimator "children with health cards" for the population 12 to 23 months old which has been post-stratified out of all children under five years old.

REFERENCES

1. Frankel, M.R. (1971). Inference From Survey Samples, Ann Arbor: Institute for Social Research, University of Michigan.

2. Kish, L. and Frankel, M.R.(1974). Inferences from complex samples. J.R. Statistical Soc. B 36, 1-22

3. Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag

Table 4. Total fertility rate for women 15-49

Country	Estimate	Standard Error			Relative Error			Relative Loss	
		BRRM	JACKK	CLUSTERS	BRRM	JACKK	CLUSTERS	BRRM	JACKK
Colombia	3.204	0.195	0.153	NA	0.061	0.048	NA	1.271	NA
Botswana	4.863	0.209	0.143	NA	0.043	0.029	NA	1.483	NA
Bolivia	5.037	0.196	0.132	NA	0.039	0.026	NA	1.500	NA
Ghana	6.406	0.198	0.152	NA	0.031	0.024	NA	1.292	NA
Egypt	4.540	0.129	0.079	NA	0.028	0.017	NA	1.647	NA
Zimbabwe	5.428	0.246	0.167	NA	0.045	0.031	NA	1.452	NA
Mean		**0.196**	**0.138**	**NA**	**0.041**	**0.029**	**NA**	**1.441**	**NA**

Table 5. Infant mortality rate for children born in the five years preceding the survey

Country	Estimate	Standard Error			Relative Error			Relative Loss	
		BRPM	JACKK	CLUSTERS	BRPM	JACKK	CLUSTERS	BRPM	JACKK
Colombia	32.304	5.853	3.702	NA	0.181	0.115	NA	1.574	NA
Botswana	38.424	4.154	4.405	NA	0.108	0.115	NA	0.939	NA
Bolivia	81.939	6.736	5.308	NA	0.082	0.065	NA	1.262	NA
Ghana	77.188	8.671	5.351	NA	0.112	0.069	NA	1.623	NA
Egypt	73.153	5.856	4.065	NA	0.080	0.056	NA	1.429	NA
Zimbabwe	49.078	6.487	4.262	NA	0.132	0.087	NA	1.517	NA
Mean		**6.293**	**4.516**	**NA**	**0.116**	**0.085**	**NA**	**1.391**	**NA**

APPENDIX

Relative errors, design effects, and relative loss under different approaches

Table 1. Mean children ever born to women 15-49

		Relative Error			Design Effect			Relative Loss	
Country	Estimate	BRRM	JACKK	CLUSTERS	BRRM	JACKK	CLUSTERS	BRRM	JACKK
Colombia	2.179	0.049	0.035	0.033	2.891	2.072	1.984	1.472	1.056
Botswana	2.580	0.031	0.022	0.019	2.114	1.474	1.332	1.600	1.120
Bolivia	2.789	0.028	0.020	0.015	2.425	1.719	1.278	1.902	1.341
Ghana	3.168	0.027	0.019	0.014	1.975	1.412	1.024	1.911	1.356
Egypt	4.017	0.024	0.014	0.011	3.138	1.833	1.456	2.159	1.273
Zimbabwe	2.953	0.032	0.022	0.015	2.084	1.434	0.991	2.111	1.467
Mean		**0.032**	**0.022**	**0.018**	**2.438**	**1.657**	**1.344**	**1.859**	**1.269**

Table 2. Contraceptive prevalence rate among married women 15-49

		Relative Error			Design Effect			Relative Loss	
Country	Estimate	BRRM	JACKK	CLUSTERS	BRRM	JACKK	CLUSTERS	BRRM	JACKK
Colombia	0.648	0.029	0.022	0.020	2.150	1.554	1.442	1.462	1.077
Botswana	0.330	0.073	0.052	0.044	2.130	1.482	1.289	1.659	1.182
Bolivia	0.303	0.053	0.040	0.030	2.416	1.844	1.388	1.778	1.333
Ghana	0.129	0.093	0.062	0.054	2.058	1.370	1.157	1.714	1.143
Egypt	0.378	0.071	0.045	0.032	5.140	3.228	2.267	2.250	1.417
Zimbabwe	0.431	0.042	0.030	0.026	1.844	1.320	1.135	1.636	1.182
Mean		**0.060**	**0.042**	**0.032**	**2.623**	**1.800**	**1.478**	**1.768**	**1.230**

Table 3. Availability of a health card among children 12-23 months

		Relative Error			Design Effect			Relative Loss	
Country	Estimate	BRRM	JACKK	CLUSTERS	BRRM	JACKK	CLUSTERS	BRRM	JACKK
Colombia	0.549	0.060	0.042	0.039	1.532	1.093	1.007	1.538	1.077
Botswana	0.742	0.035	0.028	0.028	1.469	1.173	1.179	1.238	1.000
Bolivia	0.231	0.100	0.074	0.065	1.785	1.327	1.199	1.533	1.133
Ghana	0.403	0.089	0.067	0.065	1.997	1.519	1.432	1.385	1.038
Egypt	0.605	0.031	0.023	0.020	1.544	1.113	0.987	1.583	1.167
Zimbabwe	0.776	0.032	0.026	0.026	1.472	1.166	1.198	1.250	1.000
Mean		**0.058**	**0.043**	**0.041**	**1.633**	**1.232**	**1.199**	**1.398**	**1.068**