SSDA – A WINDOWS BASED SOFTWARE FOR SURVEY DATA ANALYSIS

S. B. Lal*

Anu Sharma*

Abstract:

Analysis of survey data requires the use of specialized software that incorporates most of the available survey designs. Although, a large number of software are available for the analysis of survey data with extensive features and provide good estimation, but at the same time they have been found to be costly, requires extensive coding and domain expertise to get the results. Therefore, a low cost, less complex and user friendly windows based software has been developed for the analysis of survey data. In this paper, Software for Survey Data Analysis (SSDA) has been described. SSDA provides the estimates of population mean and variance for data collected using stratified multistage sampling design and the descriptive statistics. SSDA has modules for data management, analysis, reporting and HTML help. C# (C-Sharp) programming language available under .NET programming environment has been used for developing SSDA. It is anticipated that people involved in teaching sampling methodology, statisticians and agricultural research workers will be immensely benefited by the use of this software.

Key-words: C#, Object oriented programming, .NET technology, Sampling, Survey data analysis, Crystal reports.

* Scientists, Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Pusa, New Delhi-110012

1. Introduction:

Researchers often use sample survey methodology to get information about a population by selecting a random sample from that population. Capturing variability of the characteristics under study among the sampling units in the population is an important factor to be considered while making statistically valid inferences about the population. The estimation of population parameters and their standard errors are affected by the choice of sampling design. Large amount of data are often collected from sample survey designs which needs specialized software for its analysis.

Several softwares are commercially available for survey data analysis such as PC-CARP, SUDAAN, STATA, CENVAR, CLUSTERS. Most of them are generic in nature and therefore, they require domain expertise for the analysis of survey data. Keeping in view the complexity, efficiency in terms of cost and time, a low cost user friendly software for survey data analysis was needed to be developed. This paper describes Software for Survey Data Analysis (SSDA) for analyzing the survey data collected using important sampling designs. SSDA analyzes the data collected through most common sampling designs applied in fields for data collection such as simple random sampling (SRS), systematic, probability proportional to size with replacement (PPSWR), stratified, cluster, two stage and stratified two stage sampling schemes. This software has been developed keeping in view of its user friendliness and easily operable. Further, the management of data, analysis and reporting provide most commonly used formats and normally satisfies basic user requirements. This software is very useful not only for teaching sampling design methodologies to students but also for statisticians in analyzing the data collected from complex surveys.

2. Software Organization:

SSDA is developed using object oriented C# programming language under .NET framework 2.0 (Haertle, 2002). The software contains reusable dynamic link libraries (.dll) incorporating various methods for sampling designs included in the software. These class libraries can be utilized in any other .NET application requiring similar computations by just adding their references. The graphical user interface (GUI) for the software is menu driven and

requires minimal key board input. The software has four modules namely data management, analysis, report and Hyper Text Markup Language (HTML) help. Data Management has sub modules for input data management and imputation of missing data. Fig. 1 is the hierarchical structure chart displaying the software design.

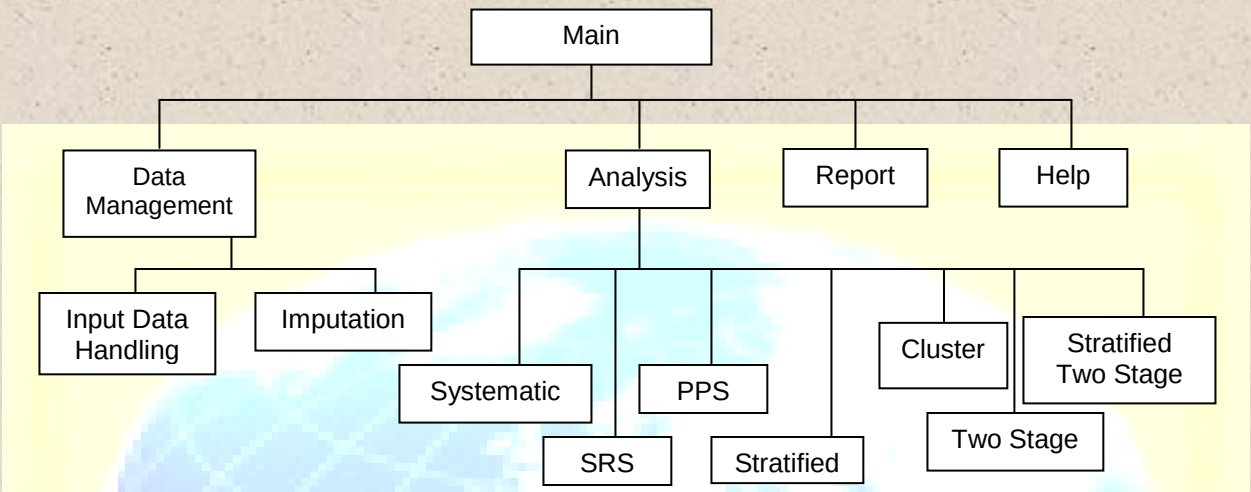

Fig. 1: Hierarchical structure chart displaying the software design

The software also provides HTML based content, index and search facility for using the software. The help also contains details about sampling schemes available in SSDA with their estimation methods. SSDA runs on windows 2000/XP/2003/Vista/2008 (Server or Professional) with .NET Framework 1.1/2.0/3.5 or Higher and MS Access database installed on it. The minimum hardware required for running SSDA are Intel Pentium IV processor, 256 MB RAM and 70 MB of hard disk space.

3. Data Management Capabilities:

This module supports features for reading and managing the input data in a spreadsheet such as creating a new file, saving a data file, importing data from MS-Excel (.xls), plain text file (.txt) and MS-Access (.mdb) file, renaming columns, jumping to a given row, filtering the data column and row, insert or delete columns or rows and printing a data file. Fig. 2 and 3 shows the data import from a MS-Excel file and other features available under data management module.

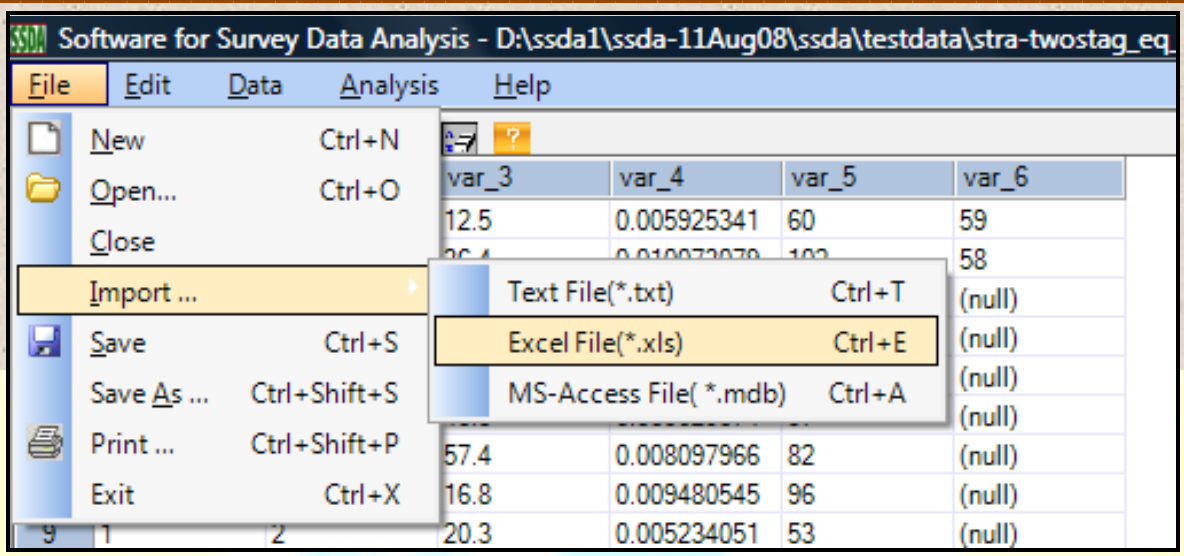

Fig. 2: Importing an excel file

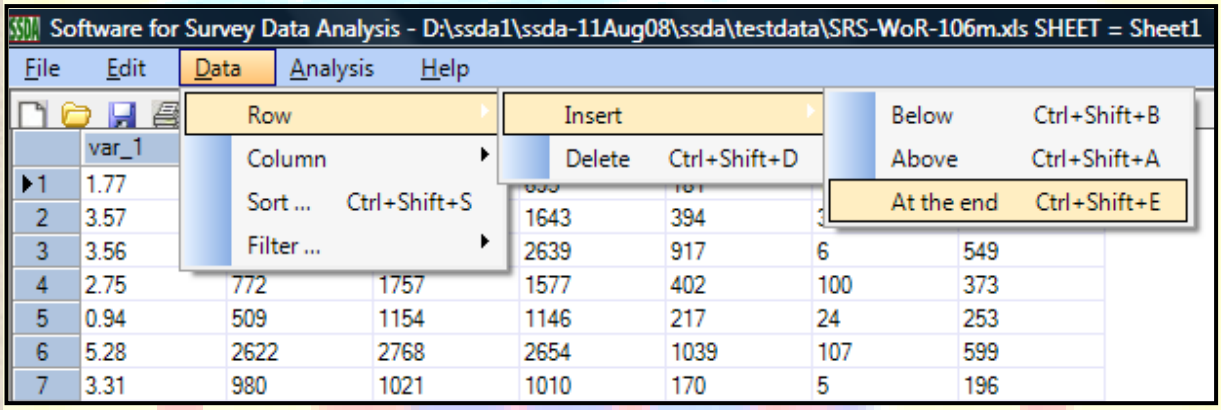

Fig. 3: Other data management features

In many cases, it is found that the data records contain some missing observations which need to be filled before proceeding for analysis. SSDA has a module to impute the missing data using three methods namely zero substitution, mean substitution and mean of neighboring units as shown in Fig. 4.

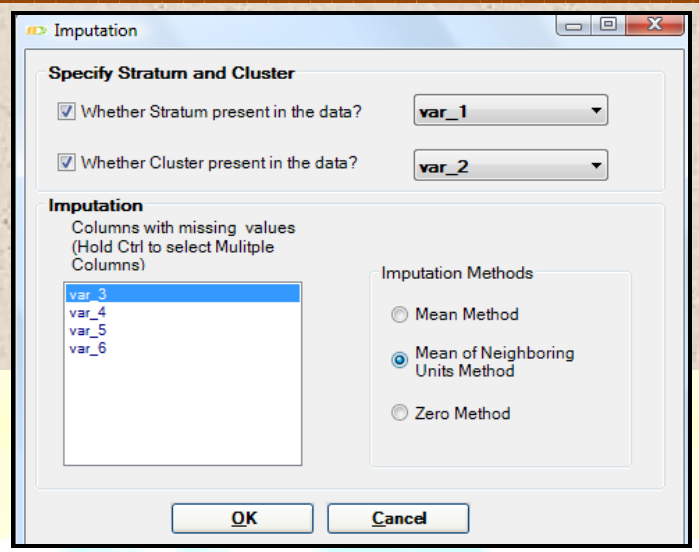

Fig. 4: Imputing missing values using SSDA

4. Analytical Features:

SSDA analyzes the data collected using simple random sampling (SRS), systematic, probability proportional to size (PPS), stratified, cluster, two stage and stratified two stage sampling schemes with equal/unequal probabilities and with/without replacements. Table 1 shows all the combinations available in SSDA for analysis.

Table 1: Sampling designs and other options available with SSDA

S. No.	Sampling Scheme	Replacement (With/Without)	Probability (Equal/Unequal)	Ratio Method Available (Yes/No)
1.	SRS	With	Equal	Yes
		Without	Equal	Yes
2.	Systematic	-	Equal	No
3.	PPS	With	Unequal	No
4.	Stratified	With	Equal	Yes (with separate & combined)
		Without	Equal	Yes (with separate & combined)

		With	Unequal	No
5.	Cluster	With	Equal	No
		Without	Equal	No
		With	Unequal	No
6.	Two Stage	With (both stages)	Equal (both stages)	No
		Without (both stages)	Equal (both stages)	No
		With (first stage)	Unequal (first stage)	No
7.	Stratified Two Stage	Without (both stages)	Equal (both stages)	No
		With (both stages)	Equal (both stages)	No
		With (first stage)	Unequal (first stage)	No

SSDA provides the estimates of population mean, variance and efficiency of the sampling design in comparison to simple random sampling without replacement. It also provides the descriptive statistics of the data without considering the sampling design. The standard procedures (Sukhatme *et. al* 1984) have been followed for estimation of population parameters for various sampling schemes.

The results of the analysis using SSDA were tested by taking a sample data from Singh and Mangat (1996). In this illustration, the procedure for analysis of survey data using stratified two stage sampling (without replacement and with equal probability at both stages) to estimate the average amount of loan per society against the defaulters (farmer) in a state have been presented here. The state was divided into two zones called as 'strata'. From each zone (strata), 6 blocks from each stratum were selected as primary stage units (PSUs). From each selected block (PSU), approximately 10% of the societies were selected as second stage units (SSUs).

Total no. of blocks (for each stratum) = 59 and 58 respectively

Total no. of societies (as SSUs) in selected (i^{th}) PSUs = M_i

The stratum ID, PSU, Mi and observations of SSUs has been given in Table 2.

Table 2: Input data to estimate the average amount of loan per society against the defaulters (farmer)

Stratum ID	PSUs	M_i	Amount due from defaulters (no. of SSUs)
1	1	60	12.5, 36.4, 26, 55.6, 58.1, 40.8 (6 units)
1	2	102	57.4, 16.8, 20.3, 70.1, 34.6, 22.6, 44.9, 28.4, 17.5, 33.7 (10 units)
1	3	48	12.9, 41.6, 34.7, 30.8, 61.1 (5 units)
1	4	113	28.7, 82.4, 37.3, 41.9, 24.7, 36.6, 39.3, 49.6, 26, 76.8, 51.6 (11 units)
1	5	92	44.8, 42.9, 51.7, 28.8, 36.4, 40.1, 61.6, 47.8, 77.4 (9 units)
1	6	57	31.6, 24.8, 69.9, 44.9, 59.7, 38.6 (6 units)
2	1	82	49.6, 36.9, 27.3, 63.6, 73, 44.9, 87.1, 61.2 (8 units)
2	2	96	53.7, 34.9, 41.5, 43.4, 56.6, 28.9, 23.4, 32.8, 60.2, 47.6 (10 units)
2	3	53	41.7, 54.9, 33.9, 27.9, 46.3 (5 units)
2	4	71	24.4, 38.9, 47.8, 45, 32.6, 66.5, 58.3 (7 units)
2	5	77	42.9, 37.3, 30.8, 51.9, 60.1, 34.6, 28.4, 38.3 (8 units)
2	6	56	44.7, 34.9, 61.7, 74.6, 37.4, 49.2 (6 units)

The software accepts the data entered in MS-Excel/ Text format or user can create a new data file in the software. In order to analyze the data, a mouse click on "Analyze" menu option and selecting "Stratified two stage sampling design" option is required. A window will appear wherein user may enter other required parameters as shown in Fig. 5.

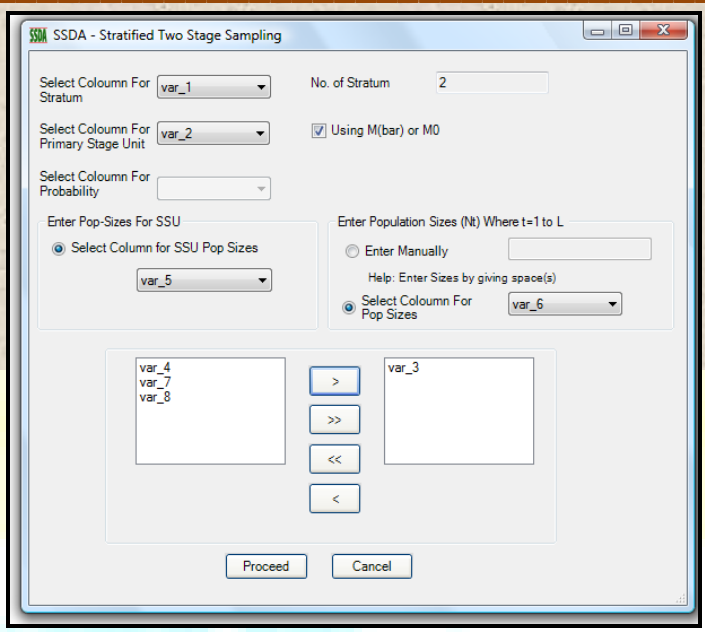

Fig. 5: A Screenshot of stratified two-stage sampling in SSDA

5. REPORTING INTERFACE:

The report module provides details of input parameters entered by user, estimate of parameters based on the selected sampling design and descriptive statistics for sampled data without considering the sampling design after analysis.

It also provides facilities to view results page by page, zooming, printing and exporting results to various common file formats such as portable document format (.pdf), microsoft word file (.doc), excel file (.xls) and rich text format (.rtf). This module has been implemented using crystal reports under .NET framework. The various processes involved in getting results are illustrated in Fig. 6. A screen shot of the estimation results produced by SSDA has been shown in Fig. 7.

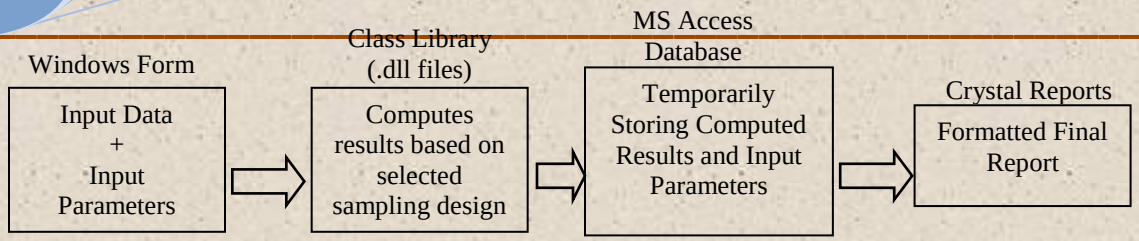

Fig. 6: Processes for generating results

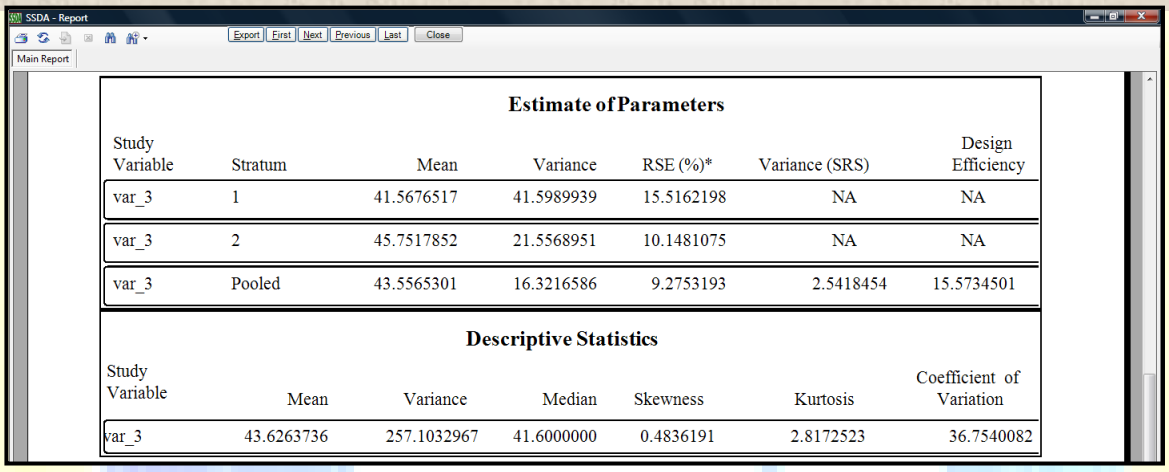

Fig. 7: Report for stratified two stage sampling

REFERENCES:

- Center for Survey Statistician and Methodology (CSSM) Software. (2005). *PC-CARP*. Iowa State University. http://cssm.iastate.edu/software/pccarp.html.
- Haertle, R. (2002). *OOP with Microsoft Visual Basic .NET and Microsoft Visual C# Step by Step*. Microsoft Press.
- Singh, R. and Mangat, N. S. (1996). *Elements of Survey Sampling.* Kluwer Academic Publishers, The Netherlands.
- Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S.and Asok, C. (1984). *Sampling Theory of Surveys with Application.* Iowa State Univ. Press, Ames, Iowa and Indian Society of Agricultural Statistics, New Delhi.