International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.2, Issue.1, Jan-Feb 2012 pp-242-246 ISSN: 2249-6645

SSDA.Analysis - A Class Library for Analysis of Sample Survey Data

Anu Sharma[1] and S. B. Lal[1]
[1](Indian Agricultural Statistics Research Institute) Pusa, New Delhi-110012

ABSTRACT

Agricultural researchers frequently use sample survey methodologies for estimation of various parameters in crops, livestock, fisheries production and allied fields. Analysis of data generated from these surveys requires the use of specialized software for survey data analysis. Most of the software for survey data analysis has proprietary source code and libraries which are not available to the users and cannot be utilized for the development of new applications. SSDA.Analysis is an objected oriented C# class library for analysis of survey data. The main goal of this library is to support fast development of MS-Windows based applications requiring readymade procedures for survey data analysis. It implements the logic of standard procedures for the estimation of parameters for various sampling designs within a framework designed to be easy to use, extend, and integrate with other .NET compatible software tools. This reusable library is highly useful for programmers and statistician involved in statistical software development. A windows based software named, Software for Survey Data Analysis (SSDA), has been developed using this library for the survey data analysis. We also reports here the results obtained after analyzing the mushroom data using this software.

Keywords: **Analysis, C#, library, sample survey.**

I. INTRODUCTION

Agricultural researchers frequently use sample survey methodologies for estimation of various parameters in crops, livestock, fisheries production and allied fields. Analysis of crop performance over a defined production area requires the identification and quantification of yield determining variables coupled with measurements of yield. Because of the inherent complexity of most crop production ecosystems, analyses of crop performance usually necessitate the collection and interpretation of large amounts of data describing yield and yield constraints and proponents. Analysis of this data requires the use of specialized software for survey data analysis.

Most of the statistical packages available worldwide either have more extensive features or are expensive. Softwares used for survey data analysis are SUDAAN, STATA, WesVarPC, PC-CARP, CENVAR and CLUSTERS etc. (Lepkowski J. and Bowles J. 1996). Some of these packages are windows based and others are DOS-based. SUDAAN is a commercial statistical software package for the analysis of correlated data, including correlated data encountered in complex sample surveys. SUDAAN is a single program consisting of a family of eleven analytic procedures used to analyze data from complex sample surveys and other observational and experimental studies involving repeated measures and cluster-correlated data (Wikepedia, 2008).

Most of the software has proprietary source code and libraries which are not available to the users. So their procedures can not be utilized in developing new software that requires similar types of computations. Also, this makes it impossible for users to extend functionality to these packages and prohibits the user for experimentation with or customization of sampling algorithms. At present no such ready made library and associated source code seems to be available for the analysis of sample survey data. So, there is a strong need of development of reusable library that implements the standard procedures for the estimation of parameters for various sampling schemes.

A set of C# library classes for, survey data analysis based software development, seems to be the proper solution to these problems, exploiting the modularity, reusability and versatility of C# design and coding.

The objective of this work was to investigate the use of object-oriented programming (OOP) as a tool for creation and management of survey data analysis procedures for agricultural applications. Specifically, we sought a programming approach that would allow (i) incremental application building without rewriting existing code, (ii) construction of a user-friendly interface from which all parameters can be assigned and software runs.

In this paper we presents SSDA.Analysis, an object oriented library for estimation of parameters of parameters of interest for Stratified Multistage Sampling Design. A windows based software and a sample web application have been developed using this library for the analysis of survey data.

II. ABOUT SSDA LIBRARY

SSDA.Analysis has been developed using C#.NET that uses .NET framework 2.0. C# is an object-oriented programming language and utilizes various key features of object oriented technologies such as its ability to program in an event driven operating system with great ease, write code for events automatically, optimize code capability for native platform etc (Haertle, R. 2002; Robinson, S. et al. 2004). C# programming language was selected as it provides features that reduce common programming errors and is supported by a wealth of standard libraries and programming tools. SSDA.Analysis provides the methods for calculating the estimates of population mean, variance and design efficiency of the sampling scheme in comparison to the simple random sampling without replacement for the sampling designs shown in Table 1. The standard procedures as given in (Sukhatme, P.V et al. 1984) have been followed for estimation of population parameters for various sampling schemes.

SSDA.Analysis also provides methods for imputation of missing data using zero substitution, mean and mean of neighboring units. It also contains methods to provide descriptive statistics of the sampled data. Measures of central tendency included are arithmetic mean, median and measures of dispersion included are variance, co-efficient of variation, skewness and kurtosis.

III. DESIGN OF SSDA.ANALYSIS LIBRARY

SSDA.Analysis library is divided into following group of classes:

3.1 Simple Random Sampling (SRS)

Two classes namely SRS_EQ and ratio_random are available for SRS. SRS_EQ class implements the procedures for calculating mean, variance, relative standard error with/without replacement along with common methods like calculating total sum, sum of squares, total number of observations etc. in case of simple random sampling with equal/unequal probabilities. ratio_random class implements the procedure for estimating mean, bias, mean squared error with/without replacement with equal probabilities for the cases where auxiliary variable is available.

3.2 Stratified Sampling:

Three classes namely STRATIFIED, Stratified_UEQ_PR and ratio_random are available for stratified sampling design. STRATIFIED and Stratified_UEQ_PR classes implements the procedures for calculating mean, variance, relative standard error with/without replacement along with common methods like calculating total sum, sum of squares, total number of observations, number of stratum, total sum for each stratum etc. in case of stratified sampling with equal and unequal probabilities respectively. ratio_random class also contains the procedures for calculating mean, bias, mean squared error with/without replacement for stratified sampling design with equal probabilities for the cases where auxiliary variable is available.

3.3 Systematic Sampling

SYS_RANDOM implements the procedures for calculating mean, variance, relative standard error with/with replacement along with common methods like calculating total sum, sum of squares, total number of observations etc.

3.4 Cluster Sampling

Cluster class implements the procedures for calculating mean, variance, relative standard error with/with replacement along with common methods like calculating total sum, sum of squares, total number of observations, number of clusters, total sum for each cluster etc. in case of cluster sampling with equal/unequal probabilities.

3.5 Two Stage Sampling

two_stage class implements the procedures for calculating mean, variance, relative standard error with/without replacement with equal/ unequal probabilities for two stage sampling scheme.

3.6 Stratified Two Stage Sampling

STR_TWO_STAGE class implements the procedures for calculating mean, variance and relative standard error for stratified two stage scheme for without replacement with equal probabilities and for with replacement with unequal probabilities. Fig. 1 shows the class diagram of the SSDA.Analysis library.

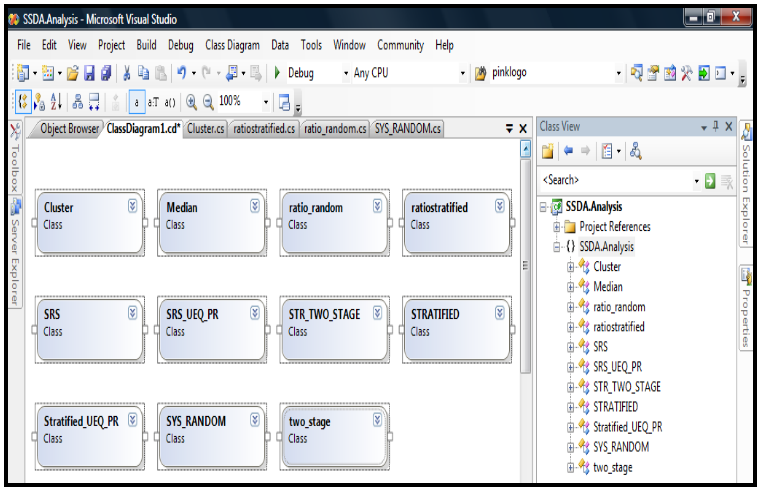


Fig.1: Class Diagram of SSDA.Analysis Library

IV. USING SSDA.ANALYSIS LIBRARY

Steps for using this library in .NET Applications are:

i. Create a Console/Window/Web Application using Visual Studio.NET

ii. Click on Add Reference

iii. Select the SSDA.Analysis library. After adding reference to SSDA.Analysis library, it appears in namespace references.

iv. Use namespace by Adding using ssda.Analysis in the beginning for your project.

v. using ssda.analysis;

vi. Create an Object of ssda.

vii. ssda.Analysis cls = new ssda.Analysis ();

viii. Call Methods and Properties

V. APPLICATIONS OF SSDA.ANALYSIS

5.1 Windows based application development

A windows based Software for Analysis of Survey Data (SSDA) has been developed using the set of classes under SSDA.Analysis library. SSDA estimates the parameters for stratified multistage sampling design. It has four modules namely data management, analysis, report and HTML help. It has been developed using C# (C-Sharp) programming language available under .NET programming environment. This software is available on request at IASRI (www.iasri.res.in).

5.2 Web based application development

This reusable library may be utilized for the development of web based application and web services for analysis of survey data. Implementation of these functions allows separation of code from client applications. So, this allows us to develop many types of client applications like window or web based or web services.

VI. A CASE STUDY - USING SSDA SOFTWARE FOR ANALYZING MUSHROOM DATA OBTAINED FROM A REAL LIFE SURVEY

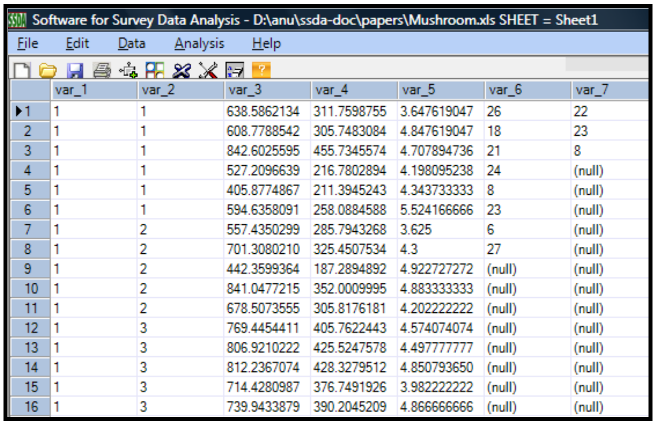
For illustration, we have considered the data obtained from a real life survey to estimate mushroom production (Gupta, Sud and Mathur 2009). The primary data have been collected in Sonepat district of Haryana state pertaining to Button Mushroom (Agaricus Bisporus) crop from November 2007 to April 2008. The data pertaining to shed area for raising mushroom, number of beds in each shed, weight of wet compost used, production of mushroom (q/ha), spawn consumed, wheat/paddy straw used in preparation of compost, processing of mushroom after picking, disposal of produce etc. have been collected from the selected mushroom growers in each of the selected village by enquiry method.

Stratified two-stage random sampling design with blocks/group of blocks as strata, mushroom–growing villages as primary stage sampling units and mushroom growing cultivar as the ultimate stage unit of selection. All the 6 blocks in the district were stratified into 3 strata by suitably combining the adjoining blocks. The three strata thus considered were Ganaur, Sonepat and Rai. A total of 8 villages, 3 from each of Ganaur and Sonepat and 2 from Rai were selected by simple random sampling without replacement. All the mushroom growers are in each of the selected villages were categorized into three categories as small, medium and large. Six cultivars were selected from these categories. Details are given in Table 1.

TABLE 1: Stratum-wise total number of mushroom growing villages, number and name of selected villages, total number of mushroom growers in the selected village (category-wise) and number of selected growers in each of the selected villages

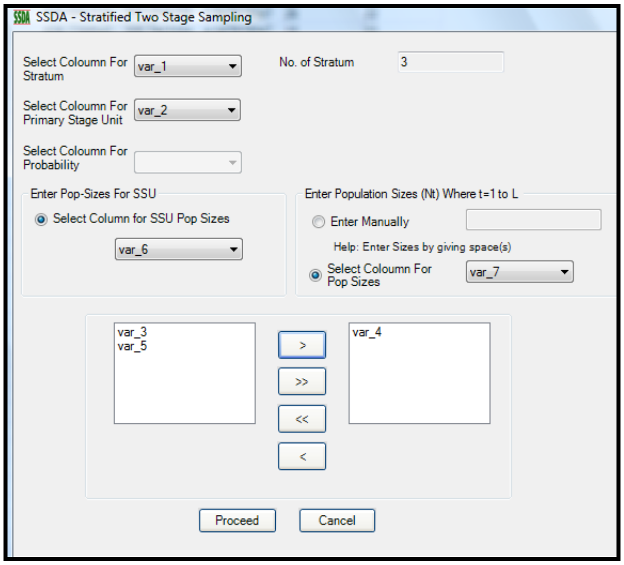
Stratum	Total Number of Villages	Selected Villages	Name of the selected villages	Total number of mushroom growers	Selected mushroom growers
Ganaur (stratum – 1) (Ganaur & Gohana)	23	3	Ahirmajra	24 (5,9, 10)	6 (1,2,3)
			Ganaur	08 (7, 0,1)	6 (5, 0, 1)
			Rajlugarhi	23 (23, 0, 0)	6 (6, 0, 0)
Sonepat (stratum – 2) (Sonepat & Kharkhoda)	22	3	Rohat	26 (23, 0, 3)	6 (4, 0, 2)
			Kakroi	18 (18, 0, 0)	6 (6, 0, 0)
			Baiyapur	21 (6, 6, 9)	6 (2, 2, 2)
Rai (stratum – 3) (Rai & Mudlana)	08	2	Sersa	06 (0, 0, 6)	6 (0, 0, 6)
			Aterna	27 (27, 0, 0)	6 (6, 0, 0)
Total	53	08		153 (109,15, 26)	48 (30, 4,14)

The data collected is entered in MS-Excel file. This data is imported into SSDA software using file→Import option under data management module as shown Fig. 2. Stratified two stage sampling without replacement and with equal probability of selection was selected from various options provided in “select sampling design” dialogue box and input data was entered as shown in Fig. 3.


Software for Survey Data Analysis - D:\anu\ssda-doc\papers\Mushroom.xls SHEET = Sheet1
File
Edit
Data
Analysis
Help

	var_1	var_2	var_3	var_4	var_5	var_6	var_7
▸1	1	1	638.5862134	311.7598755	3.647619047	26	22
2	1	1	608.7788542	305.7483084	4.847619047	18	23
3	1	1	842.6025595	455.7345574	4.707894736	21	8
4	1	1	527.2096639	216.7802894	4.198095238	24	(null)
5	1	1	405.8774867	211.3945243	4.343733333	8	(null)
6	1	1	594.6358091	258.0884588	5.524166666	23	(null)
7	1	2	557.4350299	285.7943268	3.625	6	(null)
8	1	2	701.3080210	325.4507534	4.3	27	(null)
9	1	2	442.3599364	187.2894892	4.922727272	(null)	(null)
10	1	2	841.0477215	352.0009995	4.883333333	(null)	(null)
11	1	2	678.5073555	305.8176181	4.202222222	(null)	(null)
12	1	3	769.4454411	405.7622443	4.574074074	(null)	(null)
13	1	3	806.9210222	425.5247578	4.497777777	(null)	(null)
14	1	3	812.2367074	428.3279512	4.850793650	(null)	(null)
15	1	3	714.4280987	376.7491926	3.982222222	(null)	(null)
16	1	3	739.9433879	390.2045209	4.866666666	(null)	(null)

FIG. 2 Button Mushroom (Agaricus Bisporus) crop data in Sonepat district of Haryana from November 2007 to April 2008



Select Coloumn For Stratum: var_1

No. of Stratum: 3

Select Coloumn For Primary Stage Unit: var_2

Select Coloumn For Probability:

Enter Pop-Sizes For SSU

● Select Column for SSU Pop Sizes: var_6

Enter Population Sizes (Nt) Where t=1 to L

○ Enter Manually

Help: Enter Sizes by giving space(s)

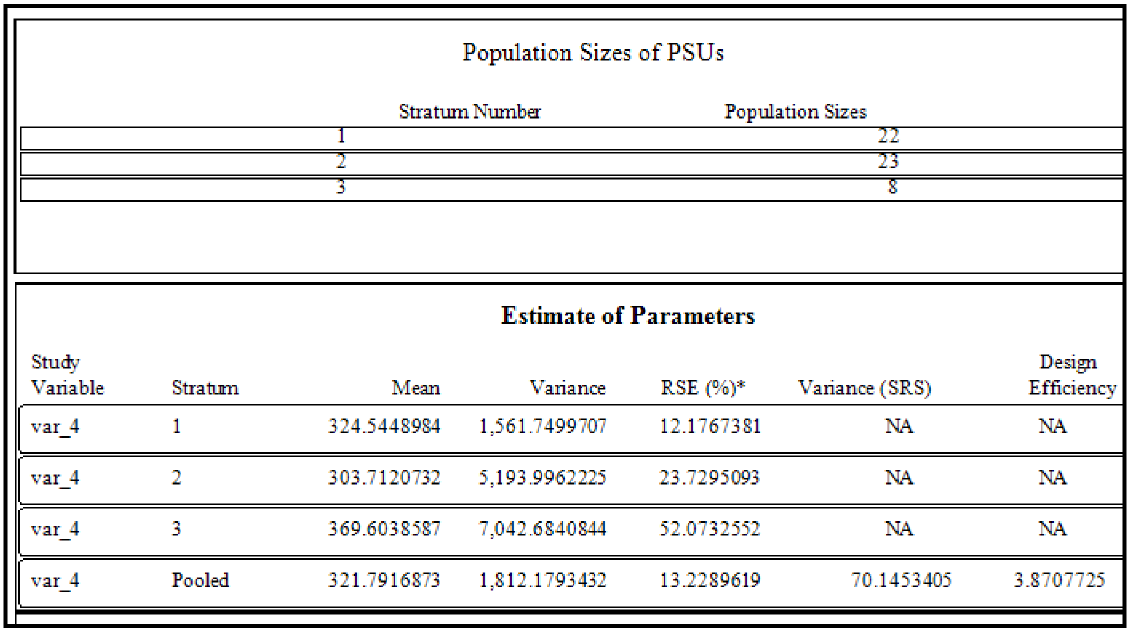
● Select Coloumn For Pop Sizes: var_7

var_3
var_5

var_4

FIG. 3 Screen for Entering Input Data

Results of analysis obtained using software for the variable production of mushroom (q/ha) are shown in Fig. 4. These results are tested with the results obtained after analyzing the same data using MS-Excel using same estimators. Both the results were found same. This ensures the reliability of the software for survey data analysis.



Population Sizes of PSUs

Stratum Number	Population Sizes
1	22
2	23
3	8

Estimate of Parameters

Study Variable	Stratum	Mean	Variance	RSE (%)*	Variance (SRS)	Design Efficiency
var_4	1	324.5448984	1,561.7499707	12.1767381	NA	NA
var_4	2	303.7120732	5,193.9962225	23.7295093	NA	NA
var_4	3	369.6038587	7,042.6840844	52.0732552	NA	NA
var_4	Pooled	321.7916873	1,812.1793432	13.2289619	70.1453405	3.8707725

FIG. 4: Results of Analysis

VII. CONCLUSIONS

Analysis of survey data is an important part of all the surveys. Many software are available for analyzing the data obtained from agricultural surveys. Most of the software has proprietary source code and libraries which are not available to the users. So their procedures can not be utilized in developing new software that requires similar types of computations. Reusable class libraries developed here can be utilized for developing many other types of applications for analyzing survey data like web applications, web services. This library would be useful to computer scientists involved in statistical software development.

REFERENCES

[1] Gupta A.K., Sud U. C. and Mathur D.C. 2009, *Pilot Study to develop sampling methodology for estimation of production of mushroom crop*. Project Report, IASRI publication.

[2] R. Haertle, *OOP with Microsoft Visual Basic .NET and Microsoft Visual C# Step by Step* (Microsoft Press, 2002)

[3] Lepkowski J. and Bowles J. 1996, Sampling Error Software for Personal Computers. *The Survey Statistician*, 35, 10-17.

[4] Robinson, S., Nagel, C., Glynn, J., Skinner, M., Watson, K. and Evjen, B. 2004, *Professional C#*. 3rd Edition, Wiley Publishing.

[5] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S.and Asok, C., *Sampling Theory of Surveys with Application.* (Iowa State Univ. Press, Ames, Iowa and Indian Society of Agricultural Statistics, New Delhi, 1984).

[6] Wikipedia(2008), SUDAAN, available at http://en.wikipedia.org/wiki/SUDAAN.