

प्रशिक्षण पाठ्यक्रम
(28 अक्टूबर से 17 नवम्बर, 2014)
TRAINING PROGRAMME
(28 Oct. to 17 Nov. , 2014)

उच्च संकाय प्रशिक्षण केन्द्र
CENTRE OF ADVANCED FACULTY TRAINING

सर्वेक्षण अभिकल्पना में आधुनिक
प्रगति एवं सांख्यिकीय सॉफ्टवेयर
द्वारा सर्वेक्षण आँकड़ों का विश्लेषण
**Recent Advances in Survey Design
and
Analysis of Survey Data
Using Statistical Software**

हुकुम चंद्र, पाठ्यक्रम संयोजक
कौस्तव आदित्य, पाठ्यक्रम सह-संयोजक
Hukum Chandra, Course Coordinator
Kaustav Aditya, Course Co-Coordinator

संदर्भ पुस्तिका - 2
REFERENCE MANUAL-2



प्रतिदर्श सर्वेक्षण प्रभाग
DIVISION OF SAMPLE SURVEY

भाकृअनुप-भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली-110 012



ICAR-INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE

LIBRARY AVENUE, PUSA, NEW DELHI - 110 012

www.iasri.res.in

प्राक्कथन


भारतीय कृषि सांख्यिकी अनुसंधान संस्थान (भा.कृ.सां.अ.सं.) कृषि सांख्यिकी एवं संगणक अनुप्रयोग के क्षेत्र में देश का एक अग्रणी संस्थान है। यह संस्थान प्रतिदर्श सर्वेक्षण, परीक्षात्मक अभिकल्पना, पूर्वानुमान एवं कृषि प्रणाली प्रतिकथन, सांख्यिकीय अनुवंशिकी, कृषि जैव-सूचना और संगणक अनुप्रयोग जैसे विभिन्न क्षेत्रों में आधारभूत एवं अनुप्रयुक्त सांख्यिकी में शोध तथा अध्यापन एवं प्रशिक्षण पाठ्यक्रम आयोजित करने का कार्य कर रहा है।

उपयुक्त प्रतिदर्श तकनीक का अनुप्रयोग और आकलन प्रक्रिया कृषि और संबद्ध विज्ञान में अनुसंधान का एक आवश्यक घटक है। नियोजन प्रक्रिया को सुविधाजनक बनाने के लिए फसलों, पशुपालन एवं मत्सय आदि के विभिन्न मापदंडों के विश्वसनीय अनुमान प्राप्त करने के लिए प्रतिदर्श सर्वेक्षण आयोजित किये जाते हैं। संस्थान ने फसलों, पशुधन और मत्सय पालन आदि से सम्बन्धित सर्वेक्षण के क्षेत्र में महत्वपूर्ण योगदान किया है।

शिक्षा प्रभाग, भारतीय कृषि अनुसंधान परिषद, नई दिल्ली के तत्वाधान में कृषि सांख्यिकी एवं संगणक अनुप्रयोग में उच्च संकाय प्रशिक्षण केन्द्र, भारतीय कृषि सांख्यिकी अनुसंधान संस्थान "सर्वेक्षण अभिकल्पना में आधुनिक प्रगति एवं सांख्यिकीय सॉफ्टवेयर द्वारा सर्वेक्षण आँकड़ों का विश्लेषण" नामक प्रशिक्षण कार्यक्रम 28 अक्टूबर से 17 नवम्बर 2014 की अवधि के दौरान आयोजित कर रहा है। यह प्रशिक्षण कार्यक्रम संकाय सदस्यों/वैज्ञानिकों को सर्वेक्षण विधियों के क्षेत्र में आधुनिक विकास पर जोर देते हुए विभिन्न प्रतिचयन विधियाँ, सर्वेक्षण आँकड़ों के विश्लेषण के लिए सॉफ्टवेयर पैकेज का प्रयोग एवं भारत में कृषि एवं बागवानी आँकड़ों के संग्रहण पद्धति की जानकारी प्रदान करने के लिए तैयार किया गया है। इसके अलावा भारत में फसल उत्पादन के पूर्वानुमान से सम्बन्धित कुछ महत्वपूर्ण विषयों को भी शामिल किया गया है। प्रशिक्षण कार्यक्रम को प्रयोगात्मक और क्षेत्रीय दौरों का महत्व देते हुए तैयार किया गया है।

इस पाठ्यक्रम के संकाय में प्रतिदर्श सर्वेक्षण एवं सम्बन्धित क्षेत्रों के व्यापक अनुभवी वैज्ञानिकों एवं प्रख्यात सांख्यिकीविदों को शामिल किया गया है। प्रशिक्षण कार्यक्रम के आरम्भ में वितरित प्रशिक्षण पुस्तिका प्रतिभागियों के ज्ञान एवं उनकी कार्य क्षमता को समृद्ध करने में उपयोगी होगी। यह उम्मीद है कि प्रतिभागी इस प्रशिक्षण कार्यक्रम से प्राप्त अनुभव को अपने निजी कार्यस्थल पर इस ज्ञान का उपयोग करने में समक्ष होंगे। डा. हुकुम चन्द्र, पाठ्यक्रम संयोजक और डा. कौस्तव आदित्य, सह-पाठ्यक्रम संयोजक को इस मूल्यवान दस्तावेज को समय पर तैयार करने के लिए बधाई देता हूँ।

नई दिल्ली-110 012
28 अक्टूबर, 2014


उमेश चन्दर सूद
निदेशक, भा.कृ.सां.अ.सं.

FOREWORD


Indian Agricultural Statistics Research Institute (IASRI) is a premier Institute in India in the discipline of Agricultural Statistics and Computer Applications. The Institute is engaged in conducting research and organizing teaching and training programmes in basic and applied statistics in different areas like Sample Surveys, Design of Experiments, Forecasting and Agricultural Systems Modeling, Statistical Genetics, Agricultural Bioinformatics and Computer Applications.

Application of suitable sampling techniques and estimation procedure is an essential component of research in agriculture and allied sciences. Sample surveys are conducted for developing reliable estimates of various parameters in case of crops, livestock and fisheries etc so as to facilitate the planning process. The Institute has made significant contributions in the field of surveys related to crops, horticulture, livestock and fisheries etc.

The Centre of Advanced Faculty Training (CAFT) in Agricultural Statistics and Computer Application at the IASRI, New Delhi is organizing a training programme on "*Recent Advances in Survey Design and Analysis of Survey Data using Statistical Software*" during October 28-November 17, 2014 under the aegis of Education Division, ICAR, New Delhi. The training programme has been designed to provide exposure to Faculty members/Scientists on different sampling procedures with due emphasis on recent developments and use of software packages for survey data analysis as well as system of collection of agricultural and horticultural statistics in India. In addition, some important topics related to forecasting of crop production in India have also been included. The training programme is practical oriented with emphasis on hands on experience and field visits.

The faculty of this course comprises of scientists and eminent statisticians with vast experience in the field of Sample Surveys and related areas. The training manual being brought out and distributed before the start of the training programme will provide a wealth of knowledge to the participants in enriching their work capabilities. It is expected that the experience gained from this training programme will enable the participants to use this knowledge in their respective work place. I wish to compliment Dr. Hukum Chandra, Course Coordinator and Kaustav Aditya, Course Co- Coordinator for bringing out this valuable document in time.

New Delhi-110012
October 28, 2014


U C Sud
Director, IASRI

आमुख

भारतीय कृषि सांख्यिकी अनुसंधान संस्थान (भा. कृ. सां. अ. सं.) देश में कृषि सांख्यिकी, संगणक अनुप्रयोग तथा जैव-सूचना के क्षेत्र में अन्वेषण करने तथा प्रोत्साहित करने के लिये एक प्रमुख संस्थान है। संस्थान, भारतीय कृषि अनुसंधान परिषद के मानव संसाधन विकास कार्यक्रम के तत्वाधान में कृषि सांख्यिकी एवं संगणक अनुप्रयोग में उच्च संकाय प्रशिक्षण केन्द्र के रूप में कार्यरत है। फसलों, बागवानी फसलों, पशुपालन एवं मत्स्य आदि के विभिन्न मापदंडों के आकलन से सम्बन्धित प्रतिदर्श सर्वेक्षण सहित कृषि सांख्यिकी के विभिन्न क्षेत्रों में मूल तथा प्रायोगिक दोनों प्रकार के अनुसंधान किये जा रहें हैं। प्रतिदर्श सर्वेक्षण प्रभाग प्रतिदर्श सर्वेक्षण के विभिन्न पहलूओं जैसे जटिल सर्वेक्षणों की अभिकल्पना एवं विश्लेषण, लघु क्षेत्र आकलन, प्रतिदर्श आँकड़ों के लिए बूटस्ट्रेप विधि, प्रतिदर्श आँकड़ों के विश्लेषण के लिए सॉफ्टवेयर का विकास, जी. आई. एस तथा रिमोट सेंसिंग तकनीक तथा विचरण अनुमान तकनीक इत्यादि क्षेत्रों के अनुसंधान में शामिल है।

“सर्वेक्षण अभिकल्पना में आधुनिक प्रगति एवं सांख्यिकीय सॉफ्टवेयर द्वारा सर्वेक्षण आँकड़ों का विश्लेषण” नामक इस पाठ्यक्रम का व्यापक उद्देश्य कृषि-विज्ञान के विषयों से सम्बन्धित प्रतिभागियों को विभिन्न प्रतिचयन तकनीक तथा आकलन प्रक्रियाओं, प्रतिदर्श सर्वेक्षणों में आधुनिक विकास, सर्वेक्षण के आँकड़ों के विश्लेषण के लिए उपयोग होने वाले सॉफ्टवेयर पैकेज जैसे आर., एस. ए. एस., एस. पी. एस. एस. और जी. आई. एस. एवं दूरसंवेदी तकनीकों की जानकारी प्रदान करना है। सैद्धान्तिक से ज्यादा प्रयोगात्मक पक्ष पर अधिक जोर दिया गया है। प्रतिभागियों के उपयोग के लिए यह संदर्भ सामग्री सरल रूप में प्रस्तुत की गयी है।

हम संस्थान के संकाय सदस्यों तथा मेहमान संकाय सदस्यों का आभार व्यक्त करते हैं, जिन्होंने अपना समय तथा ऊर्जा लगाकर अपना लेक्चर तैयार किया है। हम विभिन्न समितियों के अध्यक्षों एवं सदस्यों के उनके सहयोग के लिए भी आभारी हैं। हम डॉ. मान सिंह, डा. वेदप्रकाश, डा. आदर्श कुमार मोघा, श्री ए आर पॉल, श्री जी एम पाठक, श्री धर्म पाल सिंह, श्री देवी प्रसाद शर्मा, श्री श्योराज सिंह, श्रीमती एन चन्द्रा, श्री चंद्र पाल सिंह, श्री एस पी सिंह एम कुमार एवं श्रीमती ऊषा रस्तोगी को विशेष धन्यवाद प्रस्तुत करते हैं उनके अथक प्रयासों ने इस व्याख्यान पुस्तिका को समय पर तैयार करने में मदद की है। हम भारतीय कृषि अनुसंधान परिषद के विभिन्न संस्थानों, राज्य कृषि विश्व विद्यालयों इत्यादि के प्रतिभागियों को इस प्रशिक्षण कार्यक्रम हेतु नामित करने के लिए भी आभारी हैं। हम भारतीय कृषि अनुसंधान परिषद के शिक्षा प्रभाग द्वारा इस पाठ्यक्रम को आयोजित करने के लिये संस्थान में विश्वास व्यक्त करने के लिये ऋणी हैं। हम डा. यू. सी. सूद, निदेशक, भा. कृ. सां. अ. सं. एवं प्रधान प्रभाग, प्रतिदर्श सर्वेक्षण के निरन्तर मार्गदर्शन तथा प्रशिक्षण कार्यक्रम को सुचारू रूप से चलाने के लिये सभी आवश्यक सुविधायें प्रदान करने के लिये आभारी हैं। अंत में हम उन सभी सदस्यों का धन्यवाद करते हैं जिन्होंने इस व्याख्यान पुस्तिका को तैयार करने में सहायता की है।

नई दिल्ली
28 अक्टूबर, 2014

हुकुम चन्द्र
कौस्तव आदित्य

हुकुम चन्द्र
कौस्तव

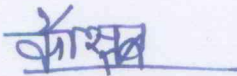
PREFACE

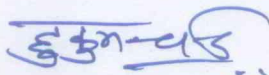
The Indian Agricultural Statistics Research Institute (IASRI) is a premier Institute for promoting and conducting research in the field of Agricultural Statistics, Computer Applications and Bio-informatics. The Institute is also functioning as a Centre of Advanced Faculty Training (CAFT) in Agricultural Statistics and Computer Application under the aegis of Human Resource Development Programme of the Indian Council of Agricultural Research (ICAR). Both basic and applied research is being carried out in various areas of Agricultural Statistics including Sample Surveys relating to estimation of different parameters of interest in case of field crops, horticulture crops, livestock and fisheries etc. The Division of Sample Survey in particular involve in research on various aspects of sample surveys like design and analysis of complex survey, small area estimation, Bootstrap method for sample data, development of software for survey data analysis, GIS and remote sensing techniques and variance estimation techniques etc.

The broader objective of this training programme on "*Recent Advances in Survey Design and Analysis of Survey Data using Statistical Software*" is to provide exposure to the participants belonging to different disciplines of agricultural sciences in proper understanding of various sampling techniques and estimation procedures, recent developments in sample surveys, in use of software packages for survey data analysis like R, SAS, SPSS and MS-Excel and GIS and remote sensing techniques etc. More emphasis is given on the applied aspects rather than theoretical. The reference material has been presented in a simplified way for use by participants.

We take this opportunity to thank all the faculty members from the Institute and the Guest Faculty who have devoted their time and energy in preparing their lectures in time. We are also thankful to the Chairman and Members of various committees for their support. We render special thanks to Dr. Man Singh, Dr. Ved Prakash, Dr. A. K. Mogha, Sh. A. R. Paul, Sh. G. M. Pathak, Sh. D. P. Singh, Sh. D. P. Sharma, Sh. Sheoraj Singh, Smt. N. Chandra, Sh. C. P. Singh, Sh. S. P. Singh, Sh. M. Kumar and Smt. U. Rani. Their sincere efforts helped in bringing out this lecture manual on time. We are also thankful to various ICAR Institutes, State Agricultural Universities etc. for nominating participants to this training programme. We are indebted to the Education Division of ICAR for entrusting the responsibility of organizing this course. We are also grateful to Dr. U.C. Sud, Director, IASRI and Head Division of Sample Survey for his continuous guidance and providing all the necessary facilities for smooth conduct of this training programme. Finally, we are thankful to one and all who helped us in preparing this reference manual.

New Delhi
October 28, 2014


Kaustav Aditya
(Course Co-Coordinator)


Hukum Chandra
(Course Coordinator)

CONTENTS

S.No.	LECTURES	Page No.
21	Statistical Methods for Forestry Dr. Girish Chandra(GF)	21.1-21.11
22	Overview of Small Area Estimation Techniques Dr. Hukum Chandra	22.1-22.12
23	Neural Network for survey data analysis Dr. G. K. Jha	23.1-23.23
24	An Overview of Crop Forecast Dr. Amarender Kumar	24.1-24.17
25	Overview of Design of Experiments Dr. Kishan Lal	25.1-25.15
26	Regression for Sample Surveys Dr. Hukum Chandra	26.1-26.11
27	Small Area Estimation Technique for Crop Yield Estimation at District Level Dr. Hukum Chandra	27.1-27.09
28	Introduction to Remote Sensing Dr. Prachi Mishra Sahoo	28.1-28.10
29	Introduction to Geographical Information System Dr. Anshu Bhardwaj	29.1-29.14
30	Crop Yield Estimation Using Geoinformatics Dr. K.N. Singh	30.1-30.07
31	Introduction to MS Excel Mr. S.B. Lal	31.1-31.11
32	Analysis of Survey Data Using Microsoft Excel Dr. K. Aditya	32.1-32.22
33	SAS-An Overview Dr. Rajender Parsad	33.1-33.54
34	Analysis of Survey Data Using SAS Dr. Anil Rai	34.1-34.24
35	SPSS - An Overview Dr. Seema Jaggi	35.1-35.13
36	Analysis of Survey Data Using SPSS Dr. U.C. Sud	36.1-36.06
37	R Software -An Overview Dr. Hukum Chandra	37.1-37.17
38	Analysis of Survey Data Using R Dr. Hukum Chandra	38.1-38.12
39	Overview of SSDA 2.0 Software and Practical Mr. S. B. Lal	39.1-39.17

ON STATISTICAL METHODS FOR FORESTRY RESEARCH

Girish Chandra

**Division of Forestry Statistics, Indian Council of Forestry Research and Education,
PO-New Forest, Dehradun-248006.
(Email: *gchandra23@yahoo.com*)**

21.1 Introduction

Forestry research, like in any other branch of science, is based on scientific method which commonly consists of a number of statistical methodologies. Starting with forest biometry as the development and application of statistical methods to assess, estimate and evaluate biological characteristic and process of forests, forest biometry has had a pervasive influence on forestry science. The sampling in forest survey in order to estimate the abundance of one or more forest resources has long had a probabilistic basis ranging from simple random sampling to multistage/multiphase designs making use of auxiliary information from aerial photography and satellite imagery. Statistically designed experiments with models are important in field experiments in order to gain a better understanding of trees, stand, and forest responses and minimized the biasness and experimental errors. Recently developed software like SAS, SPSS, R, Minitab etc. are the new advancement in analyzing the collected data in the most efficient away.

In forestry research, the basic statistical methods have their significant role to understand and apply the advanced techniques during and after the experiments. Complete ideas of measures of central tendencies, measures of dispersions are essential. The concept of probability is central to the science of statistics. As a subjective notion, probability can be interpreted as degree of belief in a continuous range between impossibility and certainty, about the occurrence of an event. Hypothesis is a tentative conjecture regarding the phenomenon under consideration. As a case of illustration, one may observe that the growing stock of trees in the borders of a plantation are better than trees inside. A simple hypothesis that could be formed from this fact is that the better growing stock of trees in the periphery is due to increased availability of sun light from the open sides. One may then deduce that by varying the spacing between trees and thereby controlling the availability of light, the trees can be made to grow differently. This would lead to a spacing experiment wherein trees are planted at different espacements and the growth is observed. One may then observe that trees under the same espacement vary in their growth and a second hypothesis formed would be that the variation in soil fertility is the causative factor for the same. Accordingly, a spacing cum fertilizer trial may follow. Further observation that trees under the same espacement, receiving the same fertilizer dose differ in their growth may prompt the researcher to conduct a spacing cum fertilizer cum varietal trial. At the end of a series of experiments, one may realize that the law of limiting factors operate in such cases which states that crop growth is constrained by the most limiting factor in the environment. Analysis of variance (ANOVA) is basically a technique of partitioning the overall variation in the responses observed in an investigation into different assignable sources of variation, some of which are specifiable and others unknown. Further, it helps in testing whether the variation due to any particular component is

significant as compared to residual variation that can occur among the observational units.

In the present note, we are concentrating about important sampling methodologies and models used in forest surveys. The calculation of forest mensurations and some measures for biodiversity are also discussed. Some of the examples for this note are taken from FAO (www.fao.org). Due to the limited scope, other important methods may also be discussed during the presentation.

21.2 Forestry Sampling

In a broad sense all *in situ* studies involving noninterfering observations on nature can be called field surveys. These may be various purposes like estimation of population parameters under interest, comparison of different populations, finding interrelations of variables, studying movement pattern or distribution pattern of organisms etc. An excellent account of the earlier development of applications of different sampling methods in forestry has been provided by Chacko (1965). Some of the methodological issues of sample surveys in forestry are discussed below under the different contexts they arise.

(a) General Field Surveys:

In forestry research, we are typically categorizing a General Field Surveys category which commonly consists of resource surveys. Most of the resource surveys start with a stratification of the population into forest types as per Champion and Seth (1968) or density classes in the context of India. Systematic sampling is easier to execute and has got some theoretical disadvantages, the major one being that a precise estimate of variance is not obtainable unless there are at least two random starts. Plots (may nested) of convenient sizes (say 0.1 ha) are then laid out in the forest area. Sampling intensity may typically vary from 1 to 5 percent depending on the size and nature of the population. A special type of sampling known as point sampling has been in vogue in timber surveys for quite some time though its application is mostly restricted to plantations. This is essentially a probability proportional to size (PPS) sampling scheme. The trees are selected with probability proportional to their basal areas and distances from a fixed sampling point. Presently, ground surveys are increasingly getting replaced by remote sensing techniques and this trend is justifiable when we consider the large effort and time involved in traditional surveys.

(b) Rare and Endangered Species:

When carrying out rare and endangered plant assessment studies, the following situation may commonly encounter. In most of the sampled places, the species richness is low or negligible, but some places have few scattered pockets of high species richness are encountered. Under such situations Adaptive Cluster sampling (Thompson, 1990) is a powerful technique for parameter estimation when characteristics under interest is sparsely distributed but highly aggregated. Examples of such populations can be found in rare and endangered plant species, non timber forest products, mineral investigations (unevenly distributed ore concentrations), pollution concentrations and hot spot investigations, and epidemiology of rare diseases. In adaptive cluster sampling, neighbouring units of the selected initial units

are added to the sample, whenever the value of the characteristic under study satisfies a pre-specified condition. The precision of adaptive cluster sampling can be enhanced by adopting appropriate sampling scheme for selecting the initial sample. Horvitz Thompson and Hansen Hurwitz estimators are generally used under such designs. Roesch (1993) used the design for a survey of forest trees. Thompson and Seber (1996) described some examples of rare species, rare diseases and environmental pollution studies where the use of adaptive sampling scheme can be highly beneficial.

(c) Socio-economic surveys:

In forest surveys, often required to conduct the socioeconomic surveys for the forest dependent communities. Stratified multistage sampling plans are especially suitable for such surveys. The units can be conveniently stratified by size or income levels or even by administrative regions. The advantage is that the sampling frame need be prepared only for the selected subsampling units in a multistage sampling plan. When the sampling units largely differ in their sizes, adjusting the selection probabilities in proportion to the size may result in better estimates. Recently many research projects under this aspect are running by ICFRE, Dehradun including Identification of Extents of Forest Fringe Villages of India.

(d) Pest and disease surveys:

Generally, cluster sampling scheme is helpful for the pest and disease surveys in forests. At the first stage the classes of plantations can be taken. The sampling intensity in terms of number of trees in the case of a survey on the incidence of sal borer in different forest Divisions found to vary from 5 to 50 percent depending upon the variability in the infestation level existed in these geographic units (Mathew, 1989).

(e) Other important sampling techniques as per the conditions:

When the measurements of units are very costly and time consuming and there is heterogeneity between the units of the population, in such situations, Ranked Set Sampling (McIntyre, 1952) is a cost-effective and precise method for sample selection.

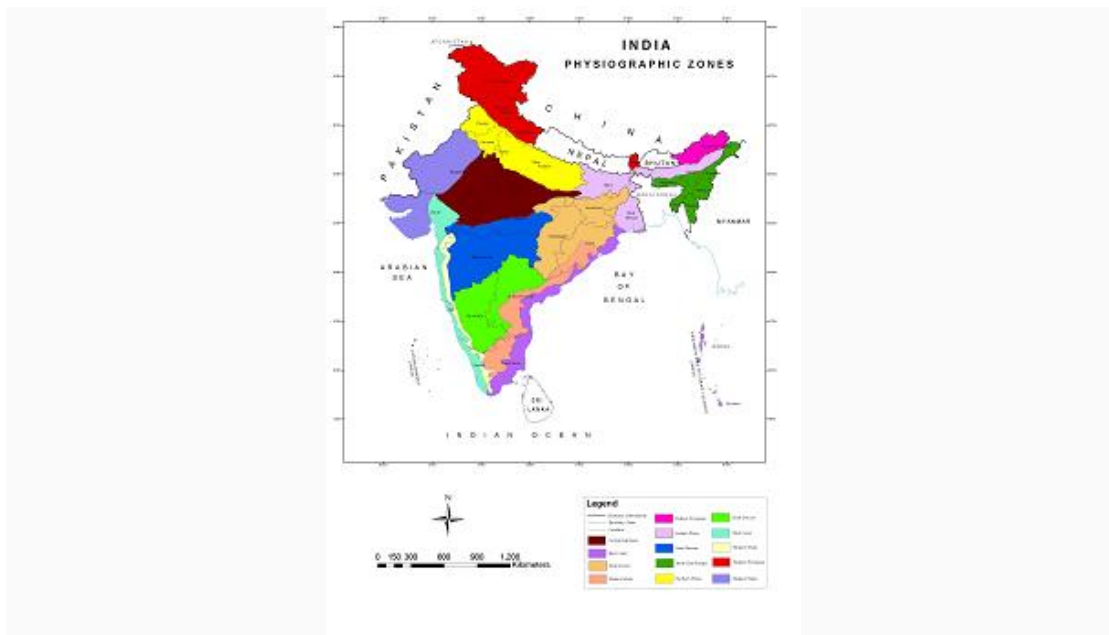
A different sampling method for studying spatial pattern is distance sampling wherein the distance to the nearest neighbouring individual is measured from a randomly located point or individual. But this does not provide an estimate of density unless the pattern is random and so has to be used in conjunction with quadrat sampling to measure and test the degrees of randomness. Spatial patterns arising from the process of dispersal are amenable to be studied through diffusion theory and some of these models possess remarkable predictive ability.

(f) Forest Inventory:

Forest inventory has almost always relied on the sampling design as the basis of statistical inference. Forest inventory typically are intended to provide estimators of multiple attributes of the forest.

Forest Survey of India, Dehradun developed a methodology (available in India State of Forest Report 2013) for a comprehensive assessment of forest resources inside

and outside forest areas at national level by stratifying the country into 14 physiographic zones and to take a sample of 10 percent districts for detailed inventory during a cycle of two years.



21.3 Modelling in Forest Surveys

A set of trees in a forest stand, producers and consumers in an economic system are examples of the components of the model. Growth and yield modelling is an essential prerequisite for evaluating the consequences of a particular management action on the future development of forest ecosystem and is, now a day, has been central theme of Forest Management. Prediction of growth or yield is important because many management decisions depend on it. For instance, consider the question, is it more profitable to grow acacia or teak in central India. The answer to this question depends, apart from the price, on the expected yield of the species concerned in that site. The number of growth and yield models applicable to forests is rapidly increasing due to readily available efficient computing technology. Some of the important base models (height and age) in forest survey mentioned in Tiwari (2013) are as follows:

$$\text{Chapman-Richards } H = b_0 [1 - \exp(-b_1 t)]^{b_2} \text{ solved by } b_2$$

$$\text{Chapman-Richards } H = b_0 [1 - \exp(-b_1 t)]^{b_2} \text{ solved by } b_1$$

$$\text{Korf } H = b_0 \exp(-b_1 / t^{b_2}) \text{ solved by } b_1$$

$$\text{Korf } H = b_0 \exp(-b_1 / t^{b_2}) \text{ solved by } b_2$$

$$\text{Hossfeld IV } H = b_0 / (1 + b_1 / t^{b_2}) \text{ solved by } b_0 \text{ and assuming } b_1 = \beta / S$$

$$\text{Sloboda } H = b_0 \exp(-b_1 \exp[b_2 / (b_3 - 1) t^{(b_3 - 1)}]) \text{ solved by } b_1$$

$$\text{Chapman-Richards } H = b_0 [1 - \exp(-b_1 t)]^{b_2} \text{ solved by } b_0 \text{ and assuming } b_1 = b_3 (H_1 / t_1)^{b_4} t^{b_5}$$

Here, H is the dominant height (m) at age t (years); t ranged from 5 to 20 years to reduce the mean squared error, and β , b_0 , b_1 , b_2 , b_3 , b_4 and b_5 are the parameters to be estimated.

For most of the modelling purposes, 'stand' is considered as a unit of management. Stand is taken as a group of trees associated with a site. Some of the common measures of stand attributes are described here first before discussing the different stand models.

The most common measurements made on trees apart from a simple count are diameter at breast height (DBH) or girth at breast-height (GBH) and total height. Here, a few stand attributes that are derivable from these basic measurements and some additional stand features are briefly mentioned.

Canopy: The cover of branches and foliage formed by the crown of Trees

Canopy Density: It is the percentage area of land covered by the canopy of trees.

Mean diameter: It is the diameter corresponding to the mean basal area of a group of trees or a stand, basal area of a tree being taken as the cross sectional area at the breast-height of the tree.

Stand basal area: The sum of the cross sectional area at breast-height of trees in the stand usually expressed m^2 on a unit area basis.

Mean height: It is the height corresponding to the mean diameter of a group of trees as read from a height-diameter curve applicable to the stand.

Top height: Top height is defined as the height corresponding to the mean diameter of 250 biggest diameters per hectare as read from height diameter curve.

Site index: Projected top height of a stand to a base age which is usually taken as the age at which culmination of height growth occurs.

Stand volume: The aggregated volume of trees in the stand usually expressed in m^3 on a unit area basis.

FAO described the models according to the degree of resolution of the input variables. Stand models can be classified as (i) whole stand models (ii) diameter class models and (iii) individual tree models. Though a distinction is made as models for even-aged and uneven-aged stands, most of the models are applicable for both the cases. Generally, trees in a plantation are mostly of the same age and same species whereas trees in natural forests are of different age levels and of different species. The term even-aged is applied to crops consisting of trees of approximately the same age but differences up to 25% of the rotation age may be allowed in case where a crop is not harvested for 100 years or more. On the other hand, the term uneven-aged is applied to crops in which the individual stems vary widely in age, the range of difference being usually more than 20 years and in the case of long rotation crops, more than 25% of the rotation.

Whole stand models predict the different stand parameters directly from the concerned regressor variables. The usual parameters of interest are commercial volume/ha, crop diameter and crop height. The regressor variables are mostly age,

stand density and site index. Since age and site index determine the top height, sometimes only the top height is considered *in lieu* of age and site index. Diameter class models trace the changes in volume or other characteristics in each diameter class by calculating growth of the average tree in each class, and multiply this average by the inventoried number of stems in each class. The volumes are aggregated over all classes to obtain stand characteristics. Individual tree models are the most complex and individually model each tree on a sample tree list. Most individual tree models calculate a crown competition index for each tree and use it in determining whether the tree lives or dies and, if it lives, its growth in terms of diameter, height and crown size. A few models found suitable for even-aged and uneven-aged stands are described separately in the following.

21.3.1 Models for even-aged stands

Sullivan and Clutter (1972) provided three basic equations which form a compatible set in the sense that the yield model can be obtained by summation of the predicted growth through appropriate growth periods. More precisely, the algebraic form of the yield model can be derived by mathematical integration of the growth model. The general form of the equations is

$$\text{Current yield} = V_1 = f(S, A_1, B_1)$$

$$\text{Future yield} = V_2 = f(S, A_2, B_2)$$

$$\text{Projected basal area} = B_2 = f(A_1, A_2, S, B_1)$$

where S = Site index

V_1 = Current stand volume

V_2 = Projected stand volume

B_1 = Current stand basal area

B_2 = Projected stand basal area

A_1 = Current stand age

A_2 = Projected stand age

Nowadays there are a number of yield equations available for different species, basal area and volume of even aged stands.

21.3.2 Models for uneven-aged stands

Boungiorno and Michie (1980) present a matrix model in which the parameters represent (i) stochastic transition of trees between diameter classes and (ii) in

growth of new trees which depends upon the condition of the stand. The model has the form

$$y_{1t+q} = \beta_0 + g_1(y_{1t} - h_{1t}) + g_2(y_{2t} - h_{2t}) + \dots + g_n(y_{nt} - h_{nt})$$

$$y_{2t+q} = b_2(y_{1t} - h_{1t}) + a_2(y_{2t} - h_{2t})$$

...

...

...

$$y_{nt+q} = b_n(y_{(n-1)t} - h_{(n-1)t}) + a_n(y_{nt} - h_{nt})$$

where y_{it+q} gives the expected number of living trees in the i th size class at time t .

h_{it} gives the number of trees harvested from i th size classes during a time interval.

g_i, a_i, b_i are coefficients to be estimated.

Here the number of trees in the smallest size class is expressed as a function of the number of trees in all size classes and of the harvest within a particular time interval. With the same time reference, the numbers of trees in higher size classes are taken as functions of the numbers of trees in adjacent size classes. It is possible to estimate the parameters through regression analysis using data from permanent sample plots wherein status of the number of trees in different diameter classes in each time period with a specified interval is recorded along with the number of trees harvested between successive measurements.

For an over-simplified illustration, consider the following data collected at two successive instances with an interval of $q = 5$ years from a few permanent sample plots in natural forests. The data given in following Table show the number of trees in three diameter classes at the two measurement periods. Assume that no harvesting has taken place during the interval, implying $h_{it}; i = 1, 2, \dots, n$ to be zero. In actual applications, more than three diameter classes may be identified and data from multiple measurements from a large number of plots will be required with records of number of trees removed from each diameter class between successive measurements.

Table: Data on number of trees/ha in three diameter classes at two successive measurements in natural forests.

Sampleplot number	Number of trees /ha at Measurement - I			Number of trees/ha at Measurement - II		
	dbh class <10cm (y_{1t})	dbh class 10-60 cm (y_{2t})	dbh class >60 cm (y_{3t})	dbh class <10cm (y_{1t+q})	dbh class 10-60 cm (y_{2t+q})	dbh class >60 cm (y_{3t+q})
1	102	54	23	87	87	45
2	84	40	22	89	71	35
3	56	35	20	91	50	30
4	202	84	42	77	167	71
5	34	23	43	90	31	29
6	87	23	12	92	68	20
7	78	56	13	90	71	43
8	202	34	32	82	152	33
9	45	45	23	91	45	38
10	150	75	21	83	128	59

The equations to be estimated are,

$$y_{1t+q} = \beta_0 + g_1 y_{1t} + g_2 y_{2t} + g_3 y_{3t}$$

$$y_{2t+q} = b_2 y_{1t} + a_2 y_{2t}$$

$$y_{3t+q} = b_3 y_{2t} + a_3 y_{3t}$$

Using the multiple linear regression, the following estimates are obtained.

$$y_{1t+q} = 99.8293 - 0.0526 y_{1t} - 0.0738 y_{2t} - 0.1476 y_{3t}$$

$$y_{2t+q} = 0.7032 y_{1t} + 0.2954 y_{2t}$$

$$y_{3t+q} = 0.7016 y_{2t} + 0.2938 y_{3t}$$

Equations such as in above model have great importance in projecting the future stand conditions and devising optimum harvesting policies on the management unit as demonstrated by Boungiorno and Michie (1980).

21.4 Forest Mensuration

In several areas of forestry research the determination of the volume or biomass of trees are essential. Since the measurement of volume or biomass is destructive, one may resort to pre-established volume or biomass prediction equations to obtain an estimate of these characteristics. These equations are found to vary from species to species and for a given species, from stand to stand.

Determination of volume of any specified part of the tree such as stem or branch is usually achieved by cutting the tree part into logs and making measurements on the logs. For research purposes, it is usual to make the logs 3m in length except the top end log which may be up to 4.5m. But if the end section is more than 1.5m in length, it is left as a separate log. The diameter or girth of the logs is measured at the middle portion of the log, at either ends of the log or at the bottom, middle and tip portions of the logs depending on the resources available. The length of individual logs is also measured. The measurements may be made over bark or under bark after peeling the bark as required. The volume of individual logs may be calculated by using one of the formulae given in the following table depending on the measurements available.

Volume of the log	Remarks
$\frac{(b^2 + t^2)l}{8\pi}$	Smalian's formula
$\left(\frac{m^2}{4\pi}\right)l$	Huber's formula
$\frac{(b^2 + 4m^2 + t^2)l}{24\pi}$	Newton's formula

where b is the girth of the log at the basal portion

m is the girth at the middle of the log

t is the girth at the thin end of the log

l is the length of the log or height of the log

The data collected from sample trees on their volume or biomass along with the dbh and height of sample trees are utilized to develop prediction equations through regression techniques. For biomass equations, sometimes diameter measured at a point lower than breast-height is used as regressor variable. Volume or biomass is taken as dependent variable and functions of dbh and height form the independent variables in the regression. Some of the standard forms of volume or biomass prediction equations in use are given below.

$$y = a + b D + c D^2$$

$$\ln y = a + b D$$

$$\ln y = a + b \ln D$$

$$y^{0.5} = a + b D$$

$$y = a + b D^2 H$$

$$\ln y = a + b D^2 H$$

$$y^{0.5} = a + b D^2 H$$

$$\ln y = a + b \ln D + c \ln H$$

$$y^{0.5} = a + b D + c H$$

$$y^{0.5} = a + b D^2 + c H + d D^2 H$$

In all the above equations, y represents tree volume or biomass, D is the tree diameter measured at breast-height or at a lower point but measured uniformly on all the sample trees, H is the tree height, and a , b , c are regression coefficients, \ln indicates natural logarithm.

Usually, several forms of equations are fitted to the data and the best fitting equation is selected based on measures like adjusted coefficient of determination or Furnival index. When the models to be compared do not have the same form of the dependent variable, Furnival index is invariably used.

$$\text{Adjusted } R^2 = 1 - \frac{n-1}{n-p} (1-R^2)$$

n is the number of observations on the dependent variable and p is the number of parameters in the model

The Furnival index for each model is obtained by multiplying the corresponding values of the square root of mean square error with the inverse of the geometric mean. For instance, the derivative of $\ln y$ is $(1/y)$ and the Furnival index in that case would be

$$\text{Furnival index} = \sqrt{MSE} \left(\frac{1}{\text{Geometric mean}(y^{-1})} \right)$$

21.5 Biodiversity Indices

Biodiversity is defined here as the property of groups or classes of living entities to be varied. Biodiversity manifests itself in two dimensions *viz.*, variety and relative abundance of species (Magurran, 1988). The former is often measured in terms of species richness index which is,

$$\text{Species richness index} = \frac{S}{\sqrt{N}}$$

where S = Number of species in a collection

N = Number of individuals collected

The relative abundance is usually measured in terms of diversity indices, a best known example of which is Shannon-Wiener index (H).

$$H = -\sum_{i=1}^s p_i \ln p_i$$

where p_i = Proportion of individuals found in the i th species

The values of Shannon-Wiener index obtained for different communities H_1 and H_2 can be tested using Student's t test where t is defined as

$$t = \frac{|H_1 - H_2|}{\sqrt{\text{Var}(H_1) + \text{Var}(H_2)}}$$

which follows Student's t distribution with n degrees of freedom where

$$v = \frac{(\text{Var}(H_1) + \text{Var}(H_2))^2}{(\text{Var}(H_1))^2 / N_1 + (\text{Var}(H_2))^2 / N_2}$$

$$\text{Var}(H) = \frac{\sum p_i (\ln p_i)^2 - (\sum p_i \ln p_i)^2}{N} + \frac{S-1}{2N^2}$$

21.6 Forestry Statistics of India at a Glance

In India, Division of Forestry Statistics, Indian Council of Forestry Research and Education (ICFRE), Dehradun publishes "Forestry Statistics India", a compendium of the official statistics of the forestry sector of India, since 1994. Generally it covers state wise estimates of important parameters like forest resources: area under forest cover, production of wood and non wood forest products, Protected Areas and Wildlife, Joint Forest Management Committees, International Trade in forest products etc. ICFRE also published First Forest Sector Report India 2010. These are available at www.icfre.org. Further, Forest Survey of India (FSI), a national organization under Ministry of Environment Forest and Climate Change, has mandated for assessing and monitoring the forest resources of the country periodically. The forest cover is being assessed using satellite images and following wall to wall approach every two years since 1980s. The detail on state/zonal/forest type wise forest cover, mangrove cover, tree cover, growing stock, forest inventory etc with methodologies are available in India State of Forest Report of Forest Survey of India (www.fsi.nic.in).

References

- Boungiorno, J. and Michie, B. R. (1980). A matrix model of uneven-aged forest management. *Forest Science*, 26(4): 609-625.
- Chacko, V. J. (1965). *A Manual on Sampling Techniques for Forest Surveys*. The Manager of Publications, Delhi
- Champion, H.G., and Seth, S.K. (1968). *A Revised Survey of the Forest Types of India*. Government of India, New Delhi.
- Magurran, A. E. (1988). *Ecological Diversity and its Measurement*. Croom Helm Limited, London. 179 p.

- Mathew, G, Rugmini, P. and Sudheendrakumar, V. V. (1998). Insect biodiversity in disturbed and undisturbed forests in the Kerala part of Western Ghats. KFRI Research Report No. 135, 113 p.
- McIntyre, G. A. (1952). A Method for Unbiased Selective Sampling Using Ranked Sets. Australian Journal of Agricultural Research 3, 385-390.
- Roesch, F.A., Jr. (1993). Adaptive Cluster Sampling for Forest Inventories. Forest Science 39, 655-669.
- Sullivan, A. D. and Clutter, J. L. (1972). A simultaneous growth and yield model for loblolly pine. Forest Science, 18: 76-86.
- Thompson, S.K. (1990). Adaptive Cluster Sampling. Journal of the American Statistical Association 85, 1050-1059.
- Thompson, S.K. and Seber, G.A.F. (1996). Adaptive Sampling. John Wiley and Sons, Inc.
- Tiwari, V. P. (2013): Modelling Growth and Yield in Trees and Stands. In Proceedings of the national seminar, recent advances in applied statistics and its application in forestry, (in press).

OVERVIEW OF SMALL AREA ESTIMATION TECHNIQUES

Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi-110012

22.1 INTRODUCTION

Most national surveys are planned to give reliable estimates at national and large domain levels and are not appropriate to produce small domain level estimates due to small sample sizes. A sample survey designed for a large population may select a small number of units or even no unit for the small area of interest. Non-response, late response etc., may further reduce sample size for a particular small domain. Consequently, sample sizes within the areas (domains) are too small to warrant the use of direct sample estimates, (estimates that use only the data on the target variable from the domain of study and time period of interest). For example, National Sample Survey Office (NSSO) Survey in India. The NSSO surveys are planned to generate statistics at state and national level. Indeed, the NSSO surveys provide reliable state and national level estimates, they cannot be used to derive reliable direct estimates at the district level owing to small sample sizes which lead to high levels of sampling variability. Due the lack of statistics at this level may suffer a proper planning, fund allocation and also monitoring of several plans at these levels. In Indian context district is a very important domain of planning process. Same time this is also true that conducting any such surveys aimed at this level is going to be very trivial and costly/time consuming. Using the state level survey (e.g., NSSO survey) data to derive the estimates at district or further smaller domain level may end up with very small sample sizes in these domains which results very unstable estimate for these domains. Hence, we need a special technique to produce estimates for such small domains or small areas. The underlying theory which resolves the problem of small sample sizes is referred to as small area estimation (SAE) in the literature of survey sampling.

SAE plays a prominent role in survey sampling due to growing demands for reliable small area statistics from both public and private sectors. Sample surveys, whether conducted by government organisations or by private entities, aim to produce reasonably accurate direct estimators, not only for the characteristics of whole population but also for a variety of subpopulations or domains. Many policymakers and researchers also want to obtain statistics for small domains. These small domains are also called small areas, because the sample size in the area or domain from the survey is small. Due to small sample size domain-specific direct estimators provide unacceptably large coefficient of variation. Therefore, it becomes necessary to employ indirect small area estimators that make use of the sample data from related areas or domains through linking models, and thus increase the effective sample size in the small areas. Such estimators can provide significantly smaller coefficient of variation than direct estimators, provided the linking models are valid. The purpose of this article is to summarize some commonly used SAE techniques and provide some example based on applications to Indian data.

22.2 SMALL AREA ESTIMATION METHODS

Small area typically denotes a subset of the population for which very little information is available from the sample survey. These subsets refer to a small geographic area (e.g., a municipality, a census division, block, tehsil, gram panchayat etc.) or a demographic group (e.g., a specific age-sex-race group of people within a large geographical area) or a cross classification of both. The statistics related to these small areas are often termed as small area statistics. Due to the increasing demand, survey organizations are faced with producing the small area estimates from existing sample surveys. Unfortunately, sample sizes in small areas tend to be too small, sometimes non-existent, to provide domains specific reliable direct estimates for these small areas. Accurate direct estimates for small areas would require a substantial increase in the overall sample size which in turn could overwhelm an already constrained budget and which could further lengthen the data processing time.

The problem of SAE is two-fold. First is the fundamental question of how to produce reliable estimates of characteristics of interest for small areas, based on very small samples taken from these areas. The second related question is how to assess the estimation error. Having only a small sample (and possibly an empty sample) in a given area, the only possible solution to the estimation problem is to borrow information from other related data sets. The SAE methods look at producing estimates with adequate precision for such small areas or domains, through an estimation procedure that ‘borrows strength’ from related areas or time periods (or both) and thus increase the overall (effective) sample size and precision. These estimation procedures are based on either implicit or explicit models that provide a link to related areas or time periods (or both) through the use of supplementary data (auxiliary information) such as recent census counts and current administrative records, see Rao (2003).

The methods used for SAE can be divided accordingly by the related data sources that they employ, whether cross-sectional (from other areas), past data or both. A further division classifies the methods by the type of inference: ‘design-based’, ‘model-dependent’ (with sub-division into the frequentist and Bayesian approaches), or the combination of the two. Based on the level of auxiliary information, available methods can also be divided into area level and unit level small area models. In the sequel we describe briefly these estimators.

22.2.1 Direct estimators

Suppose a linear estimator based on sample weights $\{w_j; j \in s\}$ is used to make inference about population level quantities. Here, s denotes the sample of size n drawn with sampling design $p(s)$ from a population $U = \{1, \dots, N\}$ of size N . Further, if $\pi_j = \sum_{j \in s} p(s)$ are the first order inclusion probabilities then $w_j = \pi_j^{-1}$ defines the design weight of element j . Under simple random sampling, $\pi_j = nN^{-1}$ and $w_j = Nn^{-1}$. For $i = 1, 2, \dots, m$, assume that the population consists of m non-overlapping domains or small areas U_i each with population of size N_i such that $U = \bigcup_{i=1}^m U_i$ and $N = \sum_{i=1}^m N_i$. Let s_i be the part of the sample of size $n_i (\geq 0)$ that falls in small area i and $n = \sum_{i=1}^m n_i$. It may be noted that n_i is a random variable. We denote by y_{ij} the

value of j -th population unit in small area i for the characteristic of interest Y . The population mean of Y in the area i , $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_j$ could be then estimated using the same weights leading to estimator

$$\hat{Y}_i^{H\ddot{a}jek} = \left(\sum_{j \in s_i} w_j \right)^{-1} \left(\sum_{j \in s_i} w_j y_j \right) \quad (2.1)$$

or, if the population size N_i of the small area i is known,

$$\hat{Y}_i^{HT} = N_i^{-1} \left(\sum_{j \in s_i} w_j y_j \right) \quad (2.2)$$

The estimators (2.1) and (2.2) are sometimes referred to as direct estimators of small area mean \bar{Y}_i . More precisely, the estimator (2.1) is referred as the Hajek type of the direct estimator, and (2.2) as the Horvitz-Thompson (HT) type of the direct estimator. These names refer to alternative approaches to estimating finite population means in the classical sampling literature, see Cochran (1977). Irrespective of which form of direct estimator is used, it is easy to see that its variance can be large when the area sample size n_i is small. For example, under simple random sampling, with no auxiliary information, a direct estimator of the mean of Y for small area i ($\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_j$) is

$$\hat{Y}_i = \bar{y}_i,$$

where $\bar{y}_i = \sum_{j \in s_i} w_j y_j / \sum_{j \in s_i} w_j = n_i^{-1} \sum_{j \in s_i} y_j$ is sample mean of Y in area i , and its variance is

$$\text{Var}_p(\hat{Y}_i) = (1 - f_i) S_i^2 / n_i$$

with $f_i = n_i / N_i$ and $S_i^2 = (N_i - 1)^{-1} \sum_{j=1}^{N_i} (y_j - \bar{y}_i)^2$, $N_i \geq 2$. Here E_p and Var_p denotes the expectation and variance respectively under the design-based approach. An unbiased estimator for S_i^2 is

$$s_i^2 = (n_i - 1)^{-1} \sum_{j \in s_i} (y_j - \bar{y}_i)^2.$$

Thus, an unbiased estimator for variance is given by

$$v(\hat{Y}_i) = n_i^{-1} (1 - f_i) s_i^2 \text{ when } N_i \text{ is known.}$$

For unknown N_i , $f_i = n_i / N_i$ is replaced by $f = n / N$ and then the estimator for variance is

$$v(\hat{Y}_i) = (1 - f) s_i^2 / n_i.$$

It is obvious that for small sample size n_i , the variance will be larger unless the variability of the Y values is sufficiently small.

Suppose that in addition to survey variable Y , values of p -auxiliary variables are also known. Let us denote by \mathbf{x}_{ij} a $p \times 1$ vector of auxiliary variable X for the unit j in area i . Then with known auxiliary information, a more efficient design-based direct estimator for \bar{Y}_i is the regression estimator defined as

$$\hat{Y}_i^{reg} = \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)' \boldsymbol{\beta}_i \quad (2.1)$$

where β_i is the vector of regression coefficients in area i , $\bar{\mathbf{x}}_i = n_i^{-1} \sum_{j \in S_i} \mathbf{x}_j$ and $\bar{\mathbf{X}}_i = N_i^{-1} \sum_{j=1}^{N_i} \mathbf{x}_j$ are the vectors of sample mean and population mean of auxiliary variable X in the area i respectively. The variance of (2.1) is

$$\text{Var}_p(\hat{Y}_i^{reg}) \approx n_i^{-1}(1-f_i)S_i^2(1-\rho_i^2) = (1-\rho_i^2)\text{Var}_p(\hat{Y}_i) \quad (2.2)$$

where ρ_i is the multiple correlation between survey variable Y and auxiliary variables X in area i .

An estimate of variance (2.2) is then

$$v(\hat{Y}_i^{reg} | n_i) = (1-\hat{\rho}_i^2)(1-f_i)s_i^2/n_i.$$

We note that by use of auxiliary variables, the variance is reduced by the factor $(1-\rho_i^2)$. This indicates that use of good auxiliary information, in the sense of high correlation with survey variable Y , increases the accuracy in small area estimation. However, the problem with the regression estimator (2.2) is that in practice the regression coefficients β_i are seldom known. Replacing β_i by its ordinary least square (OLS) estimates $\hat{\beta}_i$ is not effective because of small sample sizes in each area i .

22.2.2 Indirect estimators

When the sample size for each small area is sufficiently large to give reasonably accurate estimates, the direct estimator is the most desirable. As the sources of data are usually sample surveys designed to produce larger or higher level statistics, sample sizes for the small areas are usually small. Consequently, the associated variances of these estimators are likely to be unacceptably large. Therefore, for estimating the small areas, it is necessary to employ the estimation methods that ‘borrow strength’ from related areas. These estimators are often referred as the indirect estimators since they use values of survey variables (and auxiliary variables) from other small areas or times, and possibly from both. They borrow information (data) from other small areas or times (or both) by use of statistical models either based on implicit or explicit models that link related small areas through auxiliary information. The traditional indirect estimation techniques based on implicit linking models are synthetic and composite estimation described as below.

22.2.3 Synthetic estimators

In producing the synthetic estimates for small areas, availability of direct estimates for a set of larger domains of the population is assumed. Appropriate weights or proportions are then applied to these large population domain estimates to obtain the desired small area estimates. This class of estimators implicitly assumes that small areas which are being considered are similar, in some sense, to some larger areas which contain them and for which the reliable direct estimate is available. Gonzales (1973) described synthetic estimator as one in which an unbiased estimator of a large area is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the larger areas. The term ‘synthetic’ refers to the fact that an estimator computed from a large domain is used for each of the separate areas comprising that domain, assuming that the areas are ‘homogeneous’ with respect to the quantity that is estimated. Thus, synthetic estimators already borrow information from other ‘similar areas’.

We now describe a synthetic estimation, in particular, a scale down approach for synthetic estimation. We assume that population can be cross classified into G groups or classes and m small areas. Groups or classes are larger in size, that is, larger domains. Let N_{ig} be population of size of cell (i, g) ($g = 1, \dots, G; i = 1, \dots, m$) and y_{jig} denotes the value of unit j ($j = 1, \dots, N_{ig}$) for variable of interest Y in the cell (i, g) . We further assume availability of reliable direct estimates $\hat{T}_{y.g} = \sum_{i=1}^m \hat{T}_{y_{ig}}$ for the totals of larger group g ($g = 1, \dots, G$) that encompass the small areas i ($i = 1, \dots, m$) for a given survey, where $\hat{T}_{y_{ig}}$ is the estimate of population total ($T_{y_{ig}} = \sum_{j=1}^{N_{ig}} y_{jig}$) of Y in the (i, g) -th cell. From the available estimates for population $\hat{T}_{y.g}$, estimates of population means for group g are obtained as $\hat{Y}_{.g} = \left(\sum_{i=1}^m \hat{T}_{y_{ig}} \right) / \left(\sum_{i=1}^m N_{ig} \right) = \hat{T}_{y.g} / N_{.g}$. A suitable auxiliary information available from a census or some other source is used to compute a series of weights or proportions w_{ig} such that $\sum_g w_{ig} = 1$. The weights w_{ig} are then applied to the group means to derive the synthetic estimator for the i^{th} small area mean \bar{Y}_i as $\hat{Y}_i^{syn} = \sum_{g=1}^G w_{ig} \hat{Y}_{.g}$. This estimator is referred to as the design-based synthetic estimator. See Gonzales and Hoza (1978).

Rao and Choudry (1995) suggested the use of a ratio synthetic estimator. Let us consider availability of a single auxiliary variable. The ratio synthetic estimator for the population total of Y in small area i is $\hat{T}_{y_i}^{synR} = \hat{R}_i T_{x_i}$. They assumed that area i population ratios $R_i = T_{y_i} / T_{x_i}$, $T_{y_i} = \sum_{j=1}^{N_i} y_j$ and $T_{x_i} = \sum_{j=1}^{N_i} x_j$ respectively being the population total of the characteristic of interest Y and covariate X for the i^{th} small area, are homogeneous. Thus, $R_i = R_U = T_y / T_x$, where R_U , T_y and T_x are the values for the whole population. Here R_U is estimated by $\hat{R}_U = \bar{y} / \bar{x}$, where \bar{y} and \bar{x} are the overall sample means. We use a subscript of U to denote the population level quantities. The design-variance of a synthetic estimator $\hat{T}_{y_i}^{syn}$ of the population total of Y in area i will be small relative to the design-variance of a direct estimator $\hat{T}_{y_i}^d$ because it depends on the precision of direct estimators at a large area level. This variance can be estimated using standard design-based methods but it is more difficult to estimate the MSE of $\hat{T}_{y_i}^{syn}$ because it is hard to estimate the bias.

We now turn to model-based synthetic estimation. Let us consider the regression model

$$y_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta} + e_{ij} \tag{2.3}$$

where y_{ij} is value of variable of interest for the j^{th} ($j = 1, \dots, n_i$) unit in the small area i ($i = 1, \dots, m$) and \mathbf{x}_{ij} is the $p \times 1$ vector of auxiliary variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. The error term e_{ij} is often assumed to be normally distributed

with mean zero and variance σ^2 . Under model (2.3), two indirect estimators for small areas are:

The regression synthetic estimator for the mean of Y in small area i is defined as

$$\hat{Y}_i^{SynREG} = \bar{y}_i + (\bar{X}_i - \bar{x}_i)' \hat{\beta} \quad (2.4)$$

where $\hat{\beta}$ is the full sample estimate, i.e. calculated using data from entire areas and thus the different from direct regression estimator (2.1). For the areas with no sample data, the model-based synthetic estimator for \bar{Y}_i is defined as

$$\hat{Y}_i^{MSyn} = \bar{X}_i' \hat{\beta} \quad (2.5)$$

The estimator (2.5) will be very efficient when small area i does not exhibit strong individual effect with respect to the regression coefficient.

22.2.4 Composite estimators

As the sample size in a small area increases, a direct estimator becomes more desirable than a synthetic estimator. This is true whether or not the sample was designed to produce estimates for small areas. This motivates the use of a weighted sum of direct estimator and synthetic estimator as a desirable alternative than choosing one over the other. This weighted estimator is termed as the composite estimator. These estimators are of interest because they permit trade-off among the advantages and disadvantages of direct and synthetic estimators through their weighted combination. In general, the composite estimator for the population total of Y in small area i is defined as

$$\hat{T}_{y_i}^c = \phi_i \hat{T}_{y_i}^d + (1 - \phi_i) \hat{T}_{y_i}^{syn} \quad (2.6)$$

where $\hat{T}_{y_i}^d$ is the direct estimator and $\hat{T}_{y_i}^{syn}$ is the synthetic estimator for the population total of Y for small area i , and ϕ_i ($0 \leq \phi_i \leq 1$) is a suitably chosen weight. The estimator (2.6), a weighted sum of two component estimators can have a mean squared error (MSE) smaller than that of either component estimator when an appropriate weighting scheme is used. However, deriving the optimal weighing has generally been a challenging problem in small area estimation since these estimators are surprisingly sensitive to poor estimates of the optimum weight. Ideally, the weights should be selected as to minimise the MSE but this is problematic since the MSE of the synthetic estimator is generally unknown because of its bias. Several methods of weight selection have been proposed in the literature, see Rao (2003).

The traditional indirect estimators such as synthetic and composite have the advantage of being simple to implement. These estimation techniques provide a more efficient estimate than the corresponding design-based direct estimator for each small area through the use of implicit models which ‘borrow strength’ across the small areas. These models assume that all the areas of interest behave similarly with respect to the variable of interest and do not take into account the area specific variability. However, it can sometimes lead to severe bias if the assumption of homogeneity within the larger domain is violated or the structure of the population changed since the previous census. That is area specific variability typically remains even after accounting for the auxiliary information. This limitation is handled by an alternative estimation technique based on an explicit linking model, which provides a better approach to SAE by incorporating random area-specific effects that account for the between area variation beyond that is explained by auxiliary variables included in the model, referred as the

mixed effect model. Note that the random area effects in the mixed model capture the dissimilarities between the areas. In general, estimation methods based on an explicit model are more efficient than traditional methods based on an implicit model.

22.3 MIXED MODELS IN SMALL AREA ESTIMATION

The explicit models used in small area estimation are a special case of the linear mixed model and are very flexible in formulating and handling complex problems in small area estimation. Based on the level of auxiliary information available and utilised, two types of random effects model for small area estimation are described in the literature:

- (i) The area level mixed effect model (or Area level model) which uses area-specific auxiliary information and
- (ii) Unit level mixed effect model (or Unit level model) which uses the unit level auxiliary information.

These are special cases of the linear mixed model, usually referred as area level and unit level small area models.

22.3.1 Area level models

When individual measurements for auxiliary variables are not available and the auxiliary information is available only at the area level, then a small area model extensively discussed in the literature is the area level model. The model, used originally by Fay and Herriot (1979) for the prediction of mean per capita income in small geographical areas (less than 500 persons) within counties is defined as,

$$\tilde{\theta}_i = \theta_i + e_i ; \theta_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i \quad (3.1)$$

where $\tilde{\theta}_i$ denotes the direct sample estimator (for example, the sample mean \bar{y}_i), so that e_i represents in this case the sampling error, assumed to have zero mean and known design variance $Var_D(e_i) = \sigma_{Di}^2$ and \mathbf{x}_i represent the area level information. Here, θ_i is true population parameter (for example, the population mean). The model (3.1) is thus seen to integrate a model dependent random effect u_i and a sampling error e_i with the two errors being independent, that is, $\tilde{\theta}_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i + e_i$.

The Best Linear Unbiased Predictor (BLUP) of θ_i under this model is,

$$\hat{\theta}_i = \gamma_i \tilde{\theta}_i + (1 - \gamma_i) \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{GLS} = \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{GLS} + \gamma_i (\tilde{\theta}_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{GLS}) \quad (3.2)$$

where $\gamma_i = \sigma_u^2 / (\sigma_{Di}^2 + \sigma_u^2)$. In practice, the variances σ_u^2 and σ_{Di}^2 are usually unknown and they are replaced by sample estimates yielding the corresponding Empirical-BLUPs (or EBLUPs).

An important aspect of small area estimation is the assessment of the prediction errors. Assessment of the prediction errors under the EBLUP approach is complicated because of the errors induced by the estimation of the model parameters. To illustrate the problem, consider the model defined by (3.1) and suppose that the design variances σ_{Di}^2 are known. If $\boldsymbol{\beta}$ and σ_u^2 were also known, the variance of the BLUP is,

$$Var[\hat{\theta}_i(\sigma_u^2, \boldsymbol{\beta})] = \gamma_i \sigma_{Di}^2 = g_{li}.$$

In practice, $\boldsymbol{\beta}$ and σ_u^2 are estimated from the sample and substituted for the true values, giving rise to the EBLUP. A naïve variance estimator is obtained by replacing σ_u^2 by $\hat{\sigma}_u^2$ in g_{li} . This estimator ignores the variability of $\hat{\sigma}_u^2$ and hence underestimates the true variance. Prasad and Rao (1990) approximate the true prediction MSE of the EBLUP under normality of the two error terms and for the case where σ_u^2 is estimated by the ANOVA (fitting of constants) method as,

$$MSE[\hat{\theta}_i(\hat{\sigma}_u^2, \hat{\boldsymbol{\beta}})] = E[\hat{\theta}_i(\hat{\sigma}_u^2, \hat{\boldsymbol{\beta}}) - \theta_i]^2 = g_{li} + g_{2i} + g_{3i} \quad (3.3)$$

where ,

$g_{2i} = (1 - \gamma_i)^2 \mathbf{x}_i' \text{Var}(\hat{\boldsymbol{\beta}}_{GLS}) \mathbf{x}_i$ is the excess in MSE due to estimation of $\boldsymbol{\beta}$ and

$g_{3i} = [\sigma_{Di}^4 / (\sigma_{Di}^2 + \sigma_u^2)^3] \times \text{Var}(\hat{\sigma}_u^2)$ is the excess in MSE due to estimation of σ_u^2 .

The neglected terms in the approximation are of order $o(1/m)$ (m is the number of sampled areas). Note that the leading term in (3.3) is $g_{li} = \gamma_i \sigma_{Di}^2$ such that for large m and small values γ_i , $MSE[\hat{\theta}_i(\hat{\sigma}_u^2, \hat{\boldsymbol{\beta}})] \ll \sigma_{Di}^2 = \text{Var}_D(\hat{\theta}_i)$ illustrating the possible gains from using the model dependent estimator. Building on the approximation (3.3), Prasad and Rao (1990) derive a MSE estimator with bias of order $o(1/m)$ as,

$$mse[\hat{\theta}_i(\hat{\sigma}_u^2, \hat{\boldsymbol{\beta}})] = g_{li}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2) \hat{\text{Var}}(\hat{\sigma}_u^2) \quad (3.4)$$

where $g_{ki}(\hat{\sigma}_u^2)$ is obtained from g_{ki} by substituting $\hat{\sigma}_u^2$ for σ_u^2 , $k=1,2,3$.

22.3.2 Unit level mixed effect models

One of the simplest models commonly used in small area estimation is the ‘nested error unit level regression model’, employed originally by Battese *et al.* (1988) for predicting areas under corn and soybeans in 12 counties of the state of Iowa. Suppose that the values of auxiliary variables are known for every unit in the sample and that the true area means of these variables are also known. Denoting by \mathbf{x}_{ij} the auxiliary values for unit j in area i , the model has the form,

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + u_i + e_{ij} \quad (3.5)$$

where y_{ij} denotes the value of variable of interest for sampled unit j ($j=1, \dots, n_i$) in area i ($i=1, \dots, m$), \mathbf{x}_{ij} is a $p \times 1$ vector of unit level auxiliary variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the unknown fixed effects, n_i is the number of sample units in area i , u_i is the area specific random effect associated with area i with mean zero and variance σ_u^2 , and e_{ij} is individual level random error with mean zero and variance σ_e^2 .

The two error terms are mutually independent. The random error u_i represents the joint effect of small areas that are not accounted for by the auxiliary variables, also known as the model error for area i . The normality of u_i and e_{ij} is often assumed. The model (3.5) also holds for non-sampled units and for the whole population too. We assume that samples are drawn independently across small areas according to a specified sampling design so sample design within small areas is ignorable. The model (3.5) can be expressed as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + u_i \mathbf{1}_{n_i} + \mathbf{e}_i \quad (3.6)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$, $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})'$ is a $n_i \times p$ matrix and $\mathbf{e}_i = (e_{i1}, \dots, e_{in_i})'$. The covariance of \mathbf{y}_i is $\text{Var}(\mathbf{y}_i) = \mathbf{V}_i = \sigma_e^2 \mathbf{I}_{n_i} + \sigma_u^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}'$, which depends on a vector of

fixed parameters θ , usually called the variance components of the model. Here $\mathbf{1}_{n_i}$ is the unit vector of length n_i and \mathbf{I}_{n_i} is a identity matrix of order n_i . Population mean of Y in area i is $\bar{Y}_i = \bar{\mathbf{X}}_i' \boldsymbol{\beta} + u_i + \bar{e}_i$, where $\bar{\mathbf{X}}_i = N_i^{-1} \sum_{j=1}^{N_i} \mathbf{x}_j$, is assumed to be known. For sufficiently large N_i , $\bar{e}_i = N_i^{-1} \sum_{j=1}^{N_i} e_j \approx 0$ and then mean of Y in small area i is approximated by $\mu_i = \bar{\mathbf{X}}_i' \boldsymbol{\beta} + u_i$.

For known $\theta = (\sigma_u^2, \sigma_e^2)$, following the Henderson (1975) the BLUP for the mean of Y for small area i , $\hat{\mu}_i$ is

$$\hat{\mu}_i = \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}} + \gamma_i (\bar{y}_i - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}) = \gamma_i \{ \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)' \hat{\boldsymbol{\beta}} \} + (1 - \gamma_i) \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}}, \quad (3.7)$$

where $\hat{\boldsymbol{\beta}} = \left(\sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i \right)$ is the BLUE of $\boldsymbol{\beta}$ and $\gamma_i = \sigma_u^2 \left(\sigma_u^2 + n_i^{-1} \sigma_e^2 \right)^{-1}$. The weight γ_i ($0 \leq \gamma_i \leq 1$) called ‘shrinkage factor’, provides a trade off between the approximately design-unbiased regression estimator and the synthetic estimator and measures the model variance σ_u^2 relative to total variance $\left(\sigma_u^2 + n_i^{-1} \sigma_e^2 \right)$. For a small value of σ_u^2 , weight γ_i will be small and consequently the synthetic part get more weight and vice versa. For $n_i = 0$, i.e. areas with no samples, $\gamma_i \rightarrow 0$ and $\hat{\mu}_i = \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}}$.

Further, $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}_i$ depends on variance components θ that define the covariance matrix \mathbf{V}_i . In practice the variance components are unknown and estimated from sample data using standard method of estimation such as ANOVA, maximum likelihood (ML) or restricted maximum likelihood (REML) methods of estimation (Harville, 1977). We use ‘hat’ to denote an estimate and then, a two stage estimators known as the EBLUP of the mean of Y for small area i is

$$\hat{\mu}_i = \bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}} + \hat{\gamma}_i (\bar{y}_i - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}}). \quad (3.8)$$

where $\hat{\gamma}_i$ and $\hat{\boldsymbol{\beta}}$ are the estimates of γ_i and $\boldsymbol{\beta}$ respectively obtained by replacing θ by $\hat{\theta}$.

The mean squared error (MSE) of the EBLUP is evaluated to observe the variability in the estimator, but no closed form of MSE exists except in some special cases. For known θ , the MSE of the BLUP (3.7) is

$$MSE(\hat{\mu}_i) = g_{1i}(\sigma_u^2, \sigma_e^2) + g_{2i}(\sigma_u^2, \sigma_e^2) \quad (3.9)$$

where

$$g_{1i}(\sigma_u^2, \sigma_e^2) = (1 - \gamma_i) \sigma_u^2 = \gamma_i (\sigma_e^2 / n_i), \text{ and}$$

$$g_{2i}(\sigma_u^2, \sigma_e^2) = (\bar{\mathbf{X}}_i' - \gamma_i \bar{\mathbf{x}}_i') \left(\sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} (\bar{\mathbf{X}}_i' - \gamma_i \bar{\mathbf{x}}_i').$$

Here $g_{1i}(\sigma_u^2, \sigma_e^2)$ is the leading term in (3.9) whereas in MSE of the simple regression estimator leading term is σ_e^2 / n_i . This shows that the BLUP is superior to the simple regression estimator in terms of MSE if the shrinkage factor γ_i is small. This first term $g_{1i}(\sigma_u^2, \sigma_e^2)$ in (3.9) shows the variability of the BLUP when all the parameters

are known and is of order $o(1)$. The second term $g_{2i}(\sigma_u^2, \sigma_e^2)$ due to estimating the fixed effects β is of order $o(m^{-1})$ for large m . See Henderson (1975).

The naïve approximation to the estimate of MSE of EBLUP $\hat{\mu}_i$ is obtained by replacing θ by $\hat{\theta}$ in (3.9) as $mse_{naïve}(\hat{\mu}_i) = g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$. This approximation to MSE seriously underestimates the true MSE because the BLUP assumes known variances and hence the MSE estimator obtained by replacing the unknown variances by their sample estimates $\hat{\theta}$ fails to account for the error resulting from variance estimation. Prasad and Rao (1990) proposed the MSE approximation for the EBLUP as

$$MSE(\hat{\mu}_i) \approx g_{1i}(\sigma_u^2, \sigma_e^2) + g_{2i}(\sigma_u^2, \sigma_e^2) + g_{3i}(\sigma_u^2, \sigma_e^2) \quad (3.10)$$

with bias of order $o(m^{-1})$, where m is the number of small areas. Similarly, the mean squared error of the EBLUP \hat{Y}_i^{EBLUP} is

$$MSE(\hat{Y}_i^{EBLUP}) = (1 - f_i)^2 MSE(\hat{\mu}_i^*) + N_i^{-1}(1 - f_i)\sigma_e^2 \quad (3.11)$$

where $MSE(\hat{\mu}_i^*)$ is obtained from $MSE(\hat{\mu}_i)$ replacing $g_{2i}(\sigma_u^2, \sigma_e^2)$ by $g_{2i}^*(\sigma_u^2, \sigma_e^2)$. Prasad and Rao (1990) proposed an approximately model-unbiased estimator for the MSE as

$$mse(\hat{\mu}_i) \approx g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + 2g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) \quad (3.15)$$

where $g_{1i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$, $g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ and $g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ are obtained from $g_{1i}(\sigma_u^2, \sigma_e^2)$, $g_{2i}(\sigma_u^2, \sigma_e^2)$ and $g_{3i}(\sigma_u^2, \sigma_e^2)$ respectively, replacing $\theta = (\sigma_u^2, \sigma_e^2)$ by $\hat{\theta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)$.

The order of the bias being $o(1/m)$ since $g_{2i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ and $g_{3i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ have biases of order $o(1/m)$. This is an approximately model unbiased estimator in the sense that its bias is of order $o(m^{-1})$ and therefore considered as a second order approximation. This estimator is valid for both the method of fitting constant and REML method of estimating variances under certain regularity conditions and under the normality of random errors u_i and e_{ij} , but not for the MLE.

22.4 WEIGHTING METHODS IN SMALL AREA ESTIMATION

The model (3.5) assumes that samples are drawn independently across areas according to a specified sampling design such that the sample design within small areas is ignorable or alternatively selection bias is absent. The estimation based on such models do not make use of unit level survey weights and the corresponding estimators are not design consistent unless the sampling design is self weighting within small areas. In contrast, the design-based direct estimators are design consistent but fail to borrow strength from the related areas. In recent years, some methods proposed in the literature make use of survey weights in model-based small area estimation. Kott (1989) proposed a design consistent estimator, also model unbiased under the simple random effect model with the same assumption of random errors as in (3.5). He showed that this estimator is robust with respect to model failure under certain conditions and derived an estimator of mean squared error without including the random effect component. Empirical results show the mean squared error estimates are quite unstable and even take negative values. Consequently, this

approach cannot be used to compare proposed design-consistent small area estimator and the conventional design-based direct estimator. Prasad and Rao (1999) and You and Rao (2002) proposed a model assisted estimator for SAE called the pseudo-EBLUP, which depends on the survey weights and remains design consistent as the sample sizes in the small areas increased. Chandra and Chambers (2005, 2009, 2011) introduced the calibrated weighting based approach in SAE and defined the model-based direct estimator (MBDE) for small area means. This approach uses the calibrated sample weights derived under a population level version of the linear mixed model to define weighted linear small area estimators as well as a simple expression for the MSE. In contrast to design-based direct estimators, these estimators borrow strength from other areas via the linear mixed model used in defining the weights. There are many practical advantages associated with this approach, arising from the fact that the estimators are computed as weighted linear combinations of the actual sample data from the small areas of interest. Perhaps the most important of these are the simplicity of both the estimation process and the estimation of the MSE. Further, the MBDE is easy to interpret and to build into a survey processing system. In case of model misspecifications, the MBDE provides more robust small area estimates than the EBLUP or pseudo-EBLUP. Chandra and Chambers (2009) proposed the multipurpose weights to define the MBDE for small areas.

22.5 EXTENSION OF MIXED MODELS IN SMALL AREA ESTIMATION

The models considered so far assume that the random area effects are independent between areas, but in practice, it would be reasonable to assume that area effects associated with neighbouring areas by some distance measure (not necessarily geographical) are correlated, and correlation decays to zero as distance increases. Such models are very common in spatial analysis, but are not in wide use in SAE. An improvement in the EBLUP method can be achieved by including spatial structure in the random area effects. See Singh *et al.* (2005) for the spatial-EBLUP approach in SAE. Chandra *et al.* (2007) compare the EBLUP and MBD approaches for the spatially correlated population.

As noted earlier, in order to increase the overall sample size in small area estimation, we borrow the information from other data sets. This information can be borrowed from ‘similar’ areas or from a previous occasion. In the time series modelling approach, we exploit information in data over time (e.g., repeated surveys) in order to obtain further improvement in efficiency of estimators. In general, empirical studies show that small area estimates that draw upon information across time are more efficient than those that draw upon information across area since the time series data usually represent the same information about the target variable from the past, see for example, Pfeiffermann *et al.* (1998) and Datta *et al.* (1999). Sometimes cross sectional and times series data are combined to obtain further improvement in efficiency of the small area estimators. In general, empirical studies show that for repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small areas and time. See Rao and You (1994). Singh *et al.* (2005) used spatial-temporal models in SAE. They used spatial models for exploitation of spatial auto-correlation amongst the small area units and a spatial temporal model fitted via Kalman filtering for the time series data. Chambers and Tzavidis (2006) introduced the M-quantile approach to SAE.

Commonly used approaches of SAE assume that the underlying relationship between variable of interest y and set of covariates x are linear. However, in practice many survey data for example, agricultural, income expenditure surveys such linear relationship is not valid. Chandra and Chambers (2011) proposed SAE method for the variable which follows linear model under log transformation. Other issue, existing approaches of SAE assumes that relationship between variable of interest y and covariates x are stationary over the study space, that is, same for all areas, however, such assumption is not correct for many survey data for example agricultural, environmental data etc. Chandra *et al.* (2012) described the small area estimation for such data using geographical weighted regression approach to capture the spatial nonstationarity in the data.

22.6 APPLICATIONS IN INDIAN SURVEY DATA

22.6.1 Small area estimation in agricultural and allied sector data

An early development in small area estimation for crop-yield can be dated back to 1966 and 1968 when Panse *et al.* (1966) made an attempt to estimate the crop-yields at Block level using double sampling approach. Eye's estimates of crop yield from large number of plots prior to harvest based on crop-cutting experiments on a sub-sample of plots were used as supplementary information to build up estimates of crop-yields at Block level. However, this technique could not succeed due to physical constraints and it could not be pursued further at that time. Srivastava *et al.* (1999) used a synthetic method for crop-estimation at Block level. The population was classified into two dimensions with small area on one side and post-strata (homogeneous groups) on the other side. For crop-yield, the cell weights were estimated by raking ratio methods using the data collected in the crop-cutting approach. In fact, many auxiliary information collected during crop-cutting experiments were used in conjunction with small area level data for crop-area for estimating the cell-weights. This approach was applied for estimation of crop-yields at Block level for wheat and paddy crops on the basis of data from crop estimation surveys in Haryana State of India during 1987-88. The results were quite consistent and satisfactory. However, the effect of estimating cell-weights could not be taken into account. Moreover, the results were based on certain assumptions and efficiencies were based on variances which did not account for the biases. If assumptions fail, the biases could be serious. The synthetic approach of estimation was also applied by Singh and Goel (2000) for estimation of crop yields for wheat crop at Tehsil level, using remote sensing data. Post-strata were formed using vegetation index derived from remote sensing satellite data. Wheat crop data from GCES during 1995-96 in Rohtak district of Haryana State in India while the spectral data of IRS-IBLISS-II for February 17, 1996 was taken for vegetation index. The method improved the efficiency of the estimators to some extent in terms of standard error. However, neglecting the bias remains a serious limitation. Within a framework of sampling design conforming to General Crop Estimation Survey (GCES) approach, Sud *et al.* (2001) attempted to develop crop-yield estimates at Blocks level using farmers estimates. These estimates were, in fact, direct estimates and were based on usual sample survey techniques for improvement of estimators. The methods did not fit into the SAE approach.

A National Agricultural Insurance Scheme (NAIS) replacing comprehensive Crop Insurance Scheme (CCIS) was launched during 1999-2000 in India and area unit level

was identified as Gram Panchayat (GP) level in place of Blocks. This necessitated immediate need of crop yield at GP level for finalization of premium, claim for indemnity etc, by the Insurance companies. An alternative approach was suggested by Sharma *et al.* (2004) for scaling down Block level crop yields to GP level by developing correction factors based on the information on crop yields on selected fields through enquiries from farmers. This approach has a number of limitations: (i) it is not cost-effective (ii) subjective assessment of crop yields by farmers, which could be underestimation and/or overestimation and, finally may affect correction factors and (iii) at large level, the approach is not physically feasible. Sud *et al.* (2008) described an alternative approach for estimation of crop yield at the GP level as against the direct approach of estimation based on crop cutting experiments. They tested this approach via a pilot study and concluded that using this technique it is possible to obtain crop yield estimates at the GP level by incurring a little extra expenditure on farmers' appraisal data on crop production (farmer appraisal data is to be collected at the rate of 10 farmers per GP) and the already available CCEs data as obtained out of the GCES. The farmer appraisal data is used to scale down estimates of crop yield at the Tehsil/CDB/Mandal level as obtained through CCEs approach.

Sisodia and Chandra (2011) used small area estimation techniques for crop yield estimation at block level. It is noteworthy that in the applications described above, all are based on synthetic assumption based approach for small area estimation. In contrast, Sud *et al.* (2011) considered an application of small area estimation techniques based on random effect model which take care of dissimilarities between the small areas too. They derived district level estimates of crop yield for paddy in the State of Uttar Pradesh using the data on crop cutting experiments supervised under Improvement of Crop Statistics scheme and the secondary data from Population Census. The results show considerable improvement in the estimates generated by using small area estimation method.

Singh *et al.* (1993) and Bhatia *et al.* (2000) used small area estimation approach in estimation of milk production. In the first study, estimates of milk production at district level were developed using small area estimation approach on the milk yield records of cows in Himachal Pradesh. However, in the second study, estimates of milk production were developed from cows and buffaloes at districts level in the state of Haryana. Both of these studies examined the estimation of milk production at district level using synthetic type of estimators.

22.6.2 Small area estimation in NSSO data

Srivatsava *et al.* (2007) applied small area estimation method based on area level linear mixed to derive District level estimates of amount of loan outstanding per household using data from the 2002-2003 Debt-Investment Survey of National Sample Surveys Organization (NSSO) for the rural areas of Uttar Pradesh. In particular these authors used area level Fay and Herriot model (Fay and Herriot, 1979) to obtain the model-based District level estimates. Srivatsava (2009) used the small area estimation to obtain the district level estimates for some poverty parameters for the state of Uttar Pradesh using the NSSO survey data. Chandra *et al.* (2011a) used the Debt-Investment Survey 2002-03 of NSSO and the Population Census 2001 and the Agriculture Census 2003 and estimated the proportion of indebted households at district as well as at district by land holding levels in the State of Uttar Pradesh. Similarly, Chandra *et al.* (2011b) employ small area estimation approach to derive the

estimates of proportion of poor households at district level in the State of Uttar Pradesh in India by linking data from the Household Consumer Expenditure Survey 2006-07 of NSSO 63rd round and the Population Census. Both of these studies show that the small area estimates are precise and representative for the area they belong. Moreover, they have acceptable value of coefficient of variation.

REFERENCES

- Cochran, W.G. (1977). *Sampling Techniques*. John Wiley and Sons, New York.
- Battese, G. E., Harter, R.M., and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* 83, 28-36.
- Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika* 93(2), 225-268.
- Chandra, H. and Chambers, R. (2011). Small area estimation under transformation to linearity. *Survey Methodology* 37 (1), 39-51.
- Chandra, H. and Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistic* 25 (3), 379-395.
- Chandra, H. and Chambers, R. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Statistics in Transition* 7, 637-648.
- Chandra H., Salvati N. and Chambers R. (2007). Small area estimation for spatially correlated populations. A comparison of direct and indirect model-based methods. *Statistics in Transition* 8, 887-906.
- Chandra, H., Salvati, N. and Sud, U.C. (2011a). Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique. *Journal of Applied Statistics* 38(11), 2413-2432.
- Chandra, H., Sud, U. C. and Salvati, N. (2011b). Estimation of district level poor households in the state of Uttar Pradesh in India by combining NSSO survey and census data. *Journal of the Indian Society of Agricultural Statistics* 65(1), 1-8.
- Chandra, H., Salvati, N., Chambers, R. and Tzavidis, N. (2012). Small area estimation under spatial nonstationarity. *Computational Statistics and Data Analysis* 56, 2875-2888.
- Fay, R. E., and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74, 269-277.
- Gonzalez, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.
- Gonzalez, M.E. and Hoza, C. (1978). Small area estimation with applications to unemployment and housing estimates. *Journal of the American Statistical Association* 73, 7-15.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72, 320-338.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423-447.
- Kott, P. (1989). Robust small domain estimation using random effects modelling. *Survey Methodology* 15, 1-12.
- Panse, V.G., Rajagopalan, M. and Pillai, S.S. (1966). Estimation of crop yields for small areas. *Biometrics* 66, 374-388.

- Pfeffermann, D. and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* 16, 217-237.
- Prasad, N.G.N., and Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology* 25, 67-72.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.
- Rao, J.N.K. and Choudry, G. H. (1995). *Small area estimation: overview and empirical study*. In B.G. Cox et al. (editors). *Business Survey Methods*, John Wiley and Sons, New York, 527-542.
- Rao, J.N.K. and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics* 22, 511-528.
- Sharma, S.D., Srivastava, A.K. and Sud, U.C. (2004). Small area crop estimation for crop yield estimates at Gram Panchayat level. *Journal of the Indian Society of Agricultural Statistics* 57, 26-37.
- Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatial-temporal models in small area estimation. *Survey Methodology* 31, 183-195.
- Sisodia, B.V.S. and Chandra, H. (2012). Estimation of crop production at smaller geographical level in India. *Journal of the Indian Society of Agricultural Statistics*, 66(2), 313-319.
- Srivastava, A.K., Sud, U. C. and Chandra, H. (2007). Small Area Estimation- An Application to National Sample Survey Data. *Journal of the Indian Society of Agricultural Statistics* 61(2), 249-254.
- Sud, U.C., Chandra, H. and Srivastava, A.K. (2011). Crop yield estimation at district level using improvement of crop statistics scheme data - an application of small area estimation technique. *Journal of the Indian Society of Agricultural Statistics*, 66(2), 321-326.
- You, Y. and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics* 30, 431-439.

Some project reports and popular articles / lectures on small area estimation from IASRI, New Delhi

- Bhatia, D.K., Arya, S.N. and Gupta, H.C. (2000). *Small area estimation of milk production*. Project Report, Indian Agricultural Statistics Research Institute, New Delhi.
- Singh R. and Goel, R.C. (2000). *Use of remote sensing satellite data in crop surveys*. Technical Report, Indian Agricultural Statistics Research Institute, New Delhi.
- Singh, S., Jain, J.P., and Bhatia, D.K. (1993). *Small area estimation of milk production*. Project Report, Indian Agricultural Statistics Research Institute, New Delhi.
- Srivastava, A.K., Ahuja, D.L., Bhatia, D.K. and Mathur, D.C. (1999). *Small area estimation techniques in agriculture*. In Souvenir on Advances in Agricultural Statistics and Computer Application, Indian Agricultural Statistics Research Institute, New Delhi, 32-45.

OVERVIEW OF SMALL AREA ESTIMATION TECHNIQUES

- Sud, U. C., Bathla, H. V. L., Khatri, R. S., Mahajan, V. K., Mathur, D. C., Chandra, H. (2008). *Pilot Study on Small Area Crop Estimation Approach for Crop Yield Estimates at Gram Panchayat Level*. Project Report, Indian Agricultural Statistics Research Institute, New Delhi.
- Sud, U.C., Mathur, D.C., Srivastava, A.K., Bathla, H.V.L. and Jha, G.K. (2001). *Crop yield estimation at small area level using farmers' estimates*, Project Report, Indian Agricultural Statistics Research Institute, New Delhi.

NEUEAL NETWORKS FOR EDITING AND IMPUTATION

Girish Kumar Jha

Indian Agricultural Research Institute, New Delhi-110012

23.1 Introduction

Producers of statistics have always been concerned about the quality of their statistics. Editing and imputation are often necessary to improve the quality of data collected from a survey or a census. The aim of editing is to identify the records that are unacceptable, then identify the values of such records that need to be corrected and then correct those values using imputation. When computers are used in this process, this is called automated editing and imputation. Fellegi and Holt (1976) developed methods for automatic editing and imputation of survey data using high-speed computers. They assumed that the edit specifications would be given explicitly by subject matter expert so that as few values as possible should be changed and that the imputation is of the hot deck type. It is estimated that 20 to 40 per cent of the total cost of a survey or a census are still used for editing even after extensive use of computerized editing. Hence, to look for new opportunities to reduce the resources and time required for editing is a continuous challenge. Recent progress in the field of artificial neural networks (ANN) and continued development in capacity and speed of computers indicate that editing can be reformulated with the help of neural networks which can be trained to edit and impute from a sample of records edited by experts rather than use of explicit edit and imputation rules.

In recent years neural computing has emerged as a practical technology, with successful applications in many fields as diverse as finance, medicine, engineering, geology, physics and biology. The excitement stems from the fact that these networks are attempts to model the capabilities of the human brain. From a statistical perspective neural networks are interesting because of their potential use in prediction and classification problems.

Artificial neural networks (ANNs) are non-linear data driven self adaptive approach as opposed to the traditional model based methods. They are powerful tools for modelling, especially when the underlying data relationship is unknown. ANNs can identify and learn correlated patterns between input data sets and corresponding target values. After training, ANNs can be used to predict the outcome of new independent input data. ANNs imitate the learning process of the human brain and can process problems involving non-linear and complex data even if the data are imprecise and noisy. Thus they are ideally suited for the modeling of survey data which are known to be complex and often non-linear.

A very important feature of these networks is their adaptive nature, where “learning by example” replaces “programming” in solving problems. This feature makes such

computational models very appealing in application domains where one has little or incomplete understanding of the problem to be solved but where training data is readily available.

These networks are “neural” in the sense that they may have been inspired by neuroscience but not necessarily because they are faithful models of biological neural or cognitive phenomena. In fact majority of the network are more closely related to traditional mathematical and/or statistical models such as non-parametric pattern classifiers, clustering algorithms, nonlinear filters, and statistical regression models than they are to neurobiology models.

Neural networks (NNs) have been used for a wide variety of applications where statistical methods are traditionally employed. They have been used in classification problems, such as identifying underwater sonar currents, recognizing speech, and predicting the secondary structure of globular proteins. In time-series applications, NNs have been used in predicting stock market performance. As statisticians or users of statistics, these problems are normally solved through classical statistical methods, such as discriminant analysis, logistic regression, Bayes analysis, multiple regression, and ARIMA time-series models. It is, therefore, time to recognize neural networks as a powerful tool for data analysis.

The purpose of this lecture note is to provide an overview of ANNs and discuss how neural networks can be used effectively and efficiently in control and imputation of individual records of a data set. A detailed discussion on this topic is given by Roddick (1993) and Nordbotten (1995).

23.2 Characteristics of neural networks

- The NNs exhibit mapping capabilities, that is, they can map input patterns to their associated output patterns.
- The NNs learn by examples. Thus, NN architectures can be ‘trained’ with known examples of a problem before they are tested for their ‘inference’ capability on unknown instances of the problem. They can, therefore, identify new objects previously untrained.
- The NNs possess the capability to generalize. Thus, they can predict new outcomes from past trends.
- The NNs are robust systems and are fault tolerant. They can, therefore, recall full patterns from incomplete, partial or noisy patterns.
- The NNs can process information in parallel, at high speed, and in a distributed manner.

23.3 Basics of artificial neural networks

An artificial neural network is a set of simple computational units that are highly interconnected. The units are also called nodes and loosely represent the biological neuron. A graphical presentation of neuron is given in Figure 1. A neuron is an information processing unit that is fundamental to the operation of a neural network. The connections between nodes are unidirectional and are represented by arrows in the figure. These connections model the synaptic connections in the brain. Each connection has a weight called the synaptic weight, denoted as w_{kj} , associated with it. The synaptic weight, w_{kj} , is interpreted as the strength of the connection from the j th unit to the k th unit. Unlike a synapse in the brain, the synaptic weight of an artificial neuron may lie in a range that includes negative as well as positive values. If a weight is negative, it is termed inhibitory because it decreases the net input. If the weight is positive, the contribution is excitatory because it increases the net input.

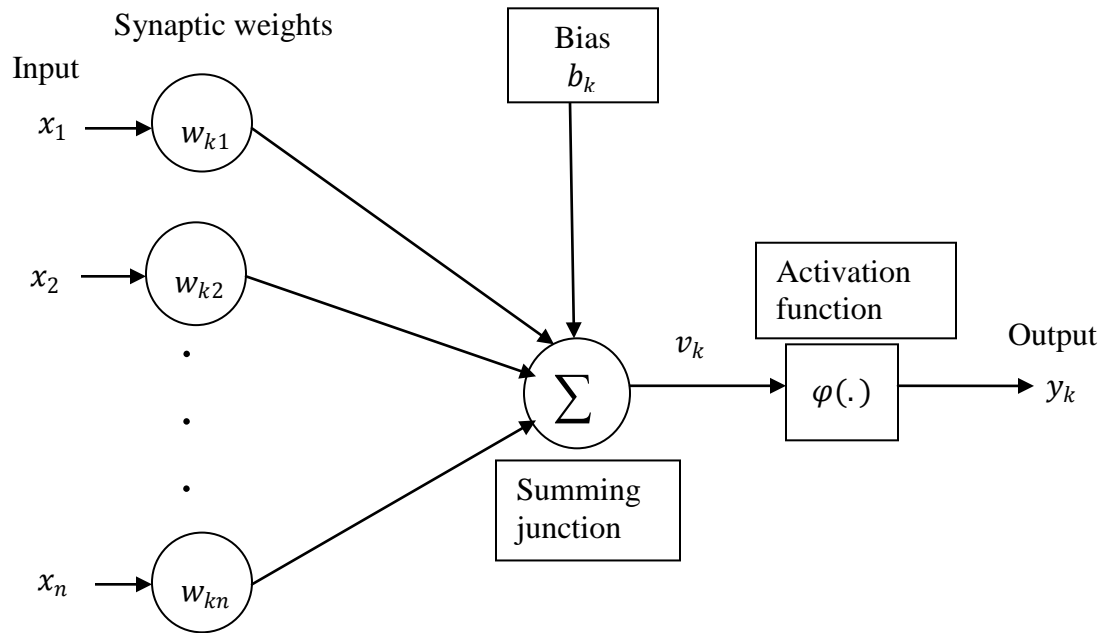


Figure 1: Nonlinear model of a neuron

The input into a node is a weighted sum of the outputs from nodes connected to it. Each unit takes its net input and applies an activation function to it. An activation function which is also known as squashing function, squashes or limits the amplitude range of the output of a neuron. The neuronal model of Figure 1 also includes an externally applied bias, denoted by b_k . The bias b_k has the effect of increasing or lowering the net input of the activation function depending on whether it is positive or negative respectively.

In mathematical terms, we may describe a neuron k by the following equations

$$y_k = \varphi(v_k) = \varphi \left(\sum_{j=1}^n w_{kj} x_j + b_k \right)$$

where x_1, x_2, \dots, x_n are the input patterns, $w_{k1}, w_{k2}, \dots, w_{kn}$ are the synaptic weights of neuron k, b_k is the bias, $\varphi(\cdot)$ is the activation function and y_k is the output of the neuron. The sigmoid function, whose graph is s-shaped, is by far the most common form of activation function used in the construction of artificial neural networks. The neural networks are built from layers of neurons connected so that one layer receives input from the preceding layer of neurons and passes the output on to the subsequent layer.

23.4 Neural networks architectures

An artificial neural network is defined as a data processing system consisting of a large number of simple highly inter connected processing elements (artificial neurons) in an architecture inspired by the structure of the cerebral cortex of the brain. There are several types of architecture of neural networks. However, the two most widely used NNs are discussed below:

Feed forward networks

In a feed forward network, information flows in one direction along connecting pathways, from the input layer via the hidden layers to the final output layer. There is no feedback (loops) i.e., the output of any layer does not affect that same or preceding layer.

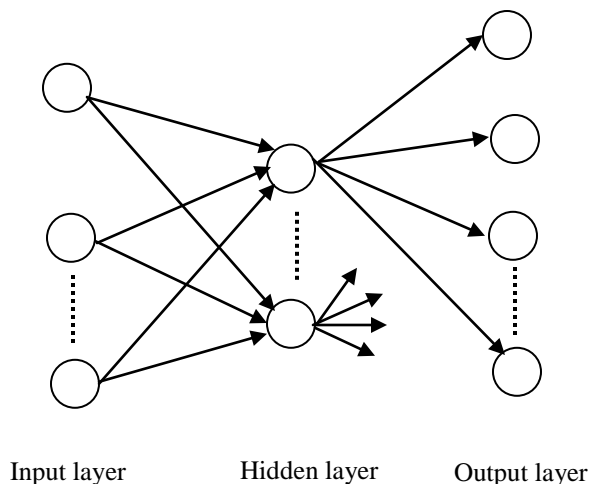


Figure 2: A multi-layer feed forward neural network

Recurrent networks

These networks differ from feed forward network architectures in the sense that there is at least one feedback loop. Thus, in these networks, for example, there could exist one layer with feedback connections as shown in figure below. There could also be neurons with self-feedback links, i.e. the output of a neuron is fed back into itself as input.

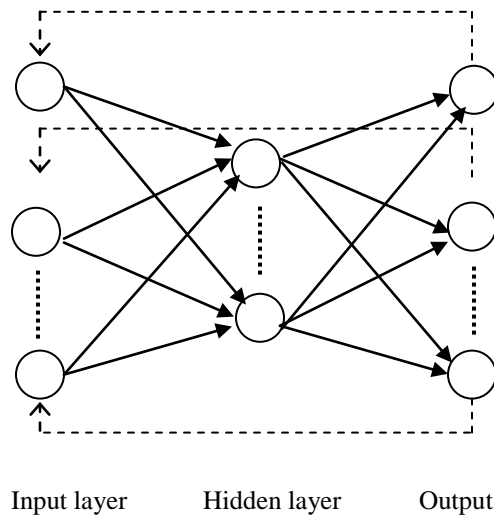


Figure 3: A recurrent neural network

Learning/Training methods

Learning methods in neural networks can be broadly classified into three basic types: supervised, unsupervised and reinforced.

Supervised learning

In this, every input pattern that is used to train the network is associated with an output pattern, which is the target or the desired pattern. A teacher is assumed to be present during the learning process, when a comparison is made between the network's computed output and the correct expected output, to determine the error. The error can then be used to change network parameters, which result in an improvement in performance.

Unsupervised learning

In this learning method, the target output is not presented to the network. It is as if there is no teacher to present the desired patterns and hence, the system learns of its own by discovering and adapting to structural features in the input patterns.

Reinforced learning

In this method, a teacher though available, does not present the expected answer but only indicates if the computed output is correct or incorrect. The information provided helps the network in its learning process. A reward is given for a correct answer computed and a penalty for a wrong answer. But, reinforced learning is not one of the popular forms of learning.

Types of neural networks

The most important class of neural networks for real world problems solving includes

- Multilayer Perceptrons
- Radial Basis Function Networks
- Kohonen Self Organizing Feature Maps

Multilayer Perceptrons

The most popular form of neural network architecture is the multilayer perceptrons (MLP) which is a generalization of the single-layer perceptron. Typically, the MLP network consists of a set of source nodes that constitute the input layer, one or more hidden layers of computation nodes and an output layer of computation nodes. The input signal propagates through the network in a forward direction on a layer by layer basis. MLP have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the error back-propagation algorithm. A multilayer perceptron has three distinctive characteristics:

- The model of each neuron in the network includes a nonlinear activation function which should also be a differentiable everywhere. A commonly used form of nonlinearity that satisfies this requirement is a sigmoidal nonlinearity. The presence of nonlinearities is important because otherwise the input-output relation of the network could be reduced to that of a single layer perceptron.
- The network contains one or more layers of hidden neurons that are not part of the input or output of the network. These hidden neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input patterns.
- The network exhibits a high degree of connectivity determined by the synapses of the network. A change in the connectivity of the network requires a change in the population of synaptic connections or their weights.

Given enough data, enough hidden units, and enough training time, an MLP with just one hidden layer can learn to approximate virtually any function to any degree of accuracy. (A statistical analogy is approximating a function with n th order polynomials.) For this reason MLPs are known as universal approximators and can be used when we have little prior knowledge of the relationship between inputs and targets. Although one hidden layer is always sufficient provided we have enough data, there are situations where a network with two or more hidden layers may require fewer hidden units and weights than a network with one hidden layer, so using extra hidden layers sometimes can improve generalization.

23.5 Radial Basis Function Networks

Radial basis function (RBF) networks have a very strong mathematical foundation rooted in regularization theory for solving ill-conditioned problems. RBF networks, almost invariably, consists of three layers: a transparent input layer, a hidden layer with sufficiently large number of nodes and an output layer. As its name implies, radially symmetric basis function is used as activation function of hidden nodes. The transformation from the input nodes to the hidden nodes is non-linear one and training of this portion of the network is generally accomplished by an unsupervised fashion. The training of the network parameters between the hidden and output layers occurs in asupervised fashion based on target outputs.

MLPs are said to be distributed-processing networks because the effect of a hidden unit can be distributed over the entire input space. On the other hand, Gaussian RBF networks

are said to be local-processing networks because the effect of a hidden unit is usually concentrated in a local area centered at the weight vector.

23.6 Kohonen Neural Network

Self Organizing Feature Map (SOFM, or Kohonen) networks are used quite differently to the other networks. Whereas all the other networks are designed for supervised learning tasks, SOFM networks are designed primarily for unsupervised learning (Patterson, 1996).

The principal goal of the SOFM is to transform an incoming signal pattern of arbitrary dimension into a one-or two dimensional discrete map, and to perform this transformation adaptively in a topologically ordered fashion. Figure 4 shows the schematic diagram of a two-dimensional lattice of neurons commonly used as the discrete map. Each neuron in the lattice is fully connected to all the source nodes in the input layer. This network represents a feedforward structure with a single computational layer consisting of neurons arranged in rows and columns. A one-dimensional lattice is a special case of the configuration depicted in Fig. 4: in this special case the computational layer consists simply of a single column or row of neurons.

Each input pattern presented to the network typically consists of a localized region or “spot” of activity against a quiet background. The location and nature of such a spot usually varies from one realization of the input pattern to another. All the neurons in the network should therefore be exposed to a sufficient number of different realizations of the input pattern to ensure that the self-organization process has a chance to mature properly.

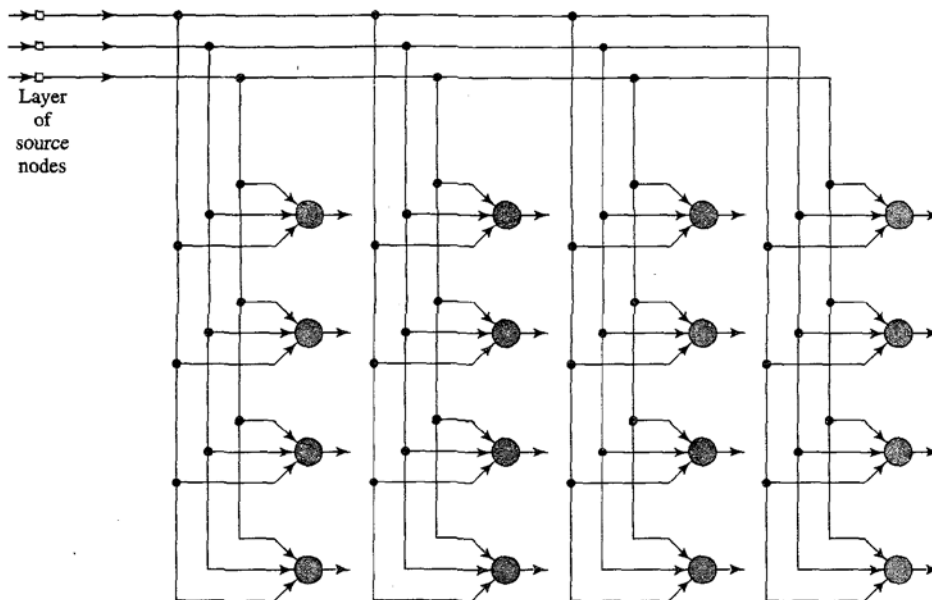


Fig. 4: Two-dimensional lattice of neurons

The algorithm responsible for the formation of the self-organizing map proceeds first by

initializing the synaptic weights in the network. This can be done by assigning them small values picked from a random number generator, in so doing, no prior order is imposed on the feature map. Once the network has been properly initialized, there are three essential processes involved in the formation of the self-organizing map, as summarized here:

- *Competition*: For each input pattern, the neurons in the network compute their respective values of a discriminant function. This discriminant function provides the basis for competition among the neurons. The particular neuron with the largest value of discriminant function is declared winner of the competition.
- *Cooperation*: The winning neuron determines the spatial location of a topological neighborhood of excited neurons, thereby providing the basis for cooperation among such neighboring neurons.
- *Synaptic Adaptation*: This last mechanism enables the excited neurons to increase their individual values of the discriminant function in relation to the input pattern through suitable adjustments applied to their synaptic weights. The adjustments made are such that the response of winning neuron to the subsequent application of similar input pattern is enhanced.

The processes of competition and cooperation are in accordance with two of the four principles of self-organization. As for the principle of self-amplification, it comes in a modified form of Hebbian learning in the adaptive process. The presence of redundancy in the input data is needed for learning since it provides knowledge.

One possible use of SOFM is in exploratory data analysis. A second possible use is in novelty detection. SOFM networks can learn to recognize clusters in the training data, and respond to it. If new data, unlike previous cases, is encountered, the network fails to recognize it and this indicates novelty.

23.6 Neural networks for editing and imputation

Neural networks have proven to be effective mapping tools for a wide variety of problems and consequently they have been used extensively by practitioners in almost every application domain ranging from agriculture to zoology. Since neural networks are best at identifying patterns or trends in data, they are well suited for editing and imputation applications.

Different types of sampling and non-sampling errors are introduced in preparation of statistics. The survey design introduces design errors, data collection introduces register, sampling, interviewer and response errors while the aggregation and dissemination processes are the sources of processing and presentation errors. Statistical data editing plays a special role in preparation of statistics because it aims at reducing the effect of errors in other statistical processes. Editing can be of two different types, logical (pre-defined rules must be obeyed) and statistical (a value is unlikely). In this case, we are concerned with evaluation of overall editing performance that is detection of data fields with errors. There are two performance requirements for editing. They include efficient error detection and influential error detection. Error detection should be evaluated in terms of both the number of errors correctly identified and the number of incorrect

detections it makes. Imputation is the process by which missing or suspicious values are replaced. Here we concern ourselves with assessing the imputation of identifiable missing values. Ideally an imputation procedure should be capable of effectively reproducing the key outputs that would have been obtained from “complete data”.

The multi-layer feed forward neural networks are mainly used for editing and imputation. This network can be used for both detecting errors and imputing missing values. There are two approaches for error detection. The first one considers the presence or absence of an error as target variable. For this approach the presence of both clean and perturbed datasets are required for training the networks. By comparing clean and perturbed data, an indicator of presence/absence of errors for each variable is calculated. The network is trained on the perturbed data with the indicator variable. The other approach consists in considering as target variable the variable itself. If the predicted value differs from the actual value then it can be considered erroneous. As far as the imputation process is concerned, neural networks model with target variable equal to the variable itself are trained on those records for which the target value is not missing, and the networks thus generated are applied for imputing missing values.

The large number of parameters that must be selected to develop a neural network model for any application indicates that the design process still involves much trial and error. The next section provides a practical introductory guide for designing a neural network model.

23.7 Development of an ANN model

The various steps in developing a neural network model are

A. Variable selection

The input variables important for modeling variable(s) under study are selected by suitable variable selection procedures.

B. Formation of training, testing and validation sets

The data set is divided into three distinct sets called training, testing and validation sets. The training set is the largest set and is used by neural network to learn patterns present in the data. The testing set is used to evaluate the generalization ability of a supposedly trained network. A final check on the performance of the trained network is made using validation set.

C. Neural network architecture

Neural network architecture defines its structure including number of hidden layers, number of hidden nodes and number of output nodes etc.

- Number of hidden layers: The hidden layer(s) provide the network with its ability to generalize. In theory, a neural network with one hidden layer with a sufficient number of hidden neurons is capable of approximating any continuous function. In practice, neural network with one and occasionally two hidden layers are widely used and have to perform very well.
- Number of hidden nodes: There is no magic formula for selecting the optimum number of hidden neurons. However, some thumb rules are available for calculating

number of hidden neurons. A rough approximation can be obtained by the geometric pyramid rule proposed by Masters (1993). For a three layer network with n input and m output neurons, the hidden layer would have $\sqrt{n*m}$ neurons.

- Number of output nodes: Neural networks with multiple outputs, especially if these outputs are widely spaced, will produce inferior results as compared to a network with a single output.
- Activation function: Activation functions are mathematical formulae that determine the output of a processing node. Each unit takes its net input and applies an activation function to it. Non linear functions have been used as activation functions such as logistic, tanh etc. The purpose of the transfer function is to prevent output from reaching very large value which can ‘paralyze’ neural networks and thereby inhibit training. Transfer functions such as sigmoid are commonly used because they are nonlinear and continuously differentiable which are desirable for network learning.

D. Evaluation criteria

The most common error function minimized in neural networks is the sum of squared errors. Other error functions offered by different software include least absolute deviations, least fourth powers, asymmetric least squares and percentage differences.

E. Neural network training

Training a neural network to learn patterns in the data involves iteratively presenting it with examples of the correct known answers. The objective of training is to find the set of weights between the neurons that determine the global minimum of error function. This involves decision regarding the number of iteration i.e., when to stop training a neural network and the selection of learning rate (a constant of proportionality which determines the size of the weight adjustments made at each iteration) and momentum values (how past weight changes affect current weight changes).

23.8 Conclusion

The computing world has a lot to gain from neural networks. Their ability to learn by example makes them very flexible and powerful. A large number of claims have been made about the modeling capabilities of neural networks, some exaggerated and some justified. Hence, to best utilize ANNs for different problems, it is essential to understand the potential as well as pitfalls of neural networks. Despite experiments with change of parameters and topologies of the networks employed, it is difficult to suggest general approach for optimizing the networks. The same type of network may have relatively good performance with respect to some variables and very poor performance with respect to others for editing and imputation purposes. For some tasks, neural networks will never replace conventional methods, but for a growing list of applications, the neural architecture will provide either an alternative or a complement to these existing techniques. Finally, I would like to state that even though neural networks have a huge potential we will only get the best of them when they are integrated with Artificial Intelligence, Fuzzy Logic, Particle Swarm Optimization and related subjects.

23.9 Practical Exercise using SPSS

The Neural Network add-on module of SPSS provides two type of network architecture namely multilayer perceptron (MLP) and radial basis function (RBF) networks.

Creating a Multilayer Perceptron Network

From the menus choose:

Analyze

Neural Networks

Multilayer Perceptron...

Creating a Radial Basis Function Network

From the menus choose:

Analyze

Neural Networks

Radial Basis Function...

The details of construction and training of a feed forward neural network will be illustrated through a predictive application using the Neural Network add-on module of SPSS in the class.

References

- Cheng, B. and Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science*, 9, 2-54.
- Fellegi, I. P. & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.
- Jha, G. K. and Sinha, K. (2014). Time-delay neural networks for time series prediction: An application to the monthly wholesale price of oilseeds in India. *Neural Computing and Applications*, 24, 563-571.
- Haykin, S. (2006). *Neural Networks: A comprehensive foundation*, Pearson Prentice Hall.
- Kaastra, I. and Boyd, M.(1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10, 215-236.
- Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data, *Survey Methodology*, 12, 1-16.
- Nordbotten, S. (1995). Editing statistical records by neural networks, *Journal of Official Statistics*, 11, 391-411.
- Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics*, 12, 385-401.
- Rao, J.N.K (1999): Some current trends in sample survey and methods. *Sankhya*, 61, Series B, 1-57.
- Roddick, L.H. (1993). Data editing using neural networks. Technical Report, Systems Development Division, Statistics Canada.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65, 386-408.

- Rumelhart, D.E., Hinton, G.E and Williams, R.J. (1986). "Learning internal representation by error propagation", in *Parallel distributed processing: Exploration in microstructure of cognition*, Vol. (1) (D.E. Rumelhart, J.L. McClelland and the PDP research group, edn.) Cambridge, MA: MIT Press, 318-362.
- Warner, B. and Misra, M. (1996). Understanding neural networks as statistical Tools. *The American Statistician*, 50, 284-293.
- Zhang, G., Patuwo, B. E. and Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14, 35-62.

CROP FORECASTING TECHNIQUES– AN OVERVIEW

Amrender Kumar

ICAR-Indian Agricultural Research Institute (IARI), New Delhi

E-mail: akjha@iari.res.in

24.1 Introduction

Reliable and timely forecasts provide important and useful input for proper, foresighted and informed planning, more so, in agriculture which is full of uncertainties. Agriculture now-a-days has become highly input and cost intensive. Without judicious use of fertilizers and plant protection measures, agriculture no longer remains as profitable as before. Uncertainties of weather, production, policies, prices, etc. often lead to mass suicides by farmers. New pests and diseases are emerging as an added threat to the production. Under the changed scenario today, forecasting of various aspects relating to agriculture are becoming essential. But in spite of strong need for reliable and timely forecasts, the current status is far from satisfactory. For most of the sectors, there is no organized system of forecasting.

24.2 Crop production forecast system

Crop yield is affected by technological change and weather variability. It can be assumed that the technological factors will increase yield smoothly through time and, therefore, year or some other parameter of time can be used to study the overall effect of technology on yield. Weather variability both within and between seasons is the second and the only uncontrollable source of variability in yields. Weather variables affect the crop differently during different stages of development. Thus extent of weather influence on crop yield depends not only on the magnitude of weather variables but also on the distribution pattern of weather over the crop season which, as such, calls for the necessity of dividing the whole crop season into fine intervals. This will increase number of variables in the model and in turn a large number of constants will have to be evaluated from the data. This will require a long series of data for precise estimation of the constants which may not be available in practice.

Fisher (1924) and Hendricks and Scholl (1943) have done pioneering work in crop weather relationship. They have given models which require small number of parameters to be estimated while taking care of distribution pattern of weather over the crop season.

Fisher assumed that the effect of change in weather variable in successive periods would not be an abrupt or erratic change but an orderly one that follows some mathematical law. He assumed that these effects are composed of the terms of a polynomial function of time. Further, the value of weather variable in w-th week, X_w was also expressed in terms of orthogonal functions of time.

$$A_w = a_0[f_0(w)] + a_1[f_1(w)] + \dots + a_k[f_k(w)]$$

$$X_w = \rho_0 [f_0(w)] + \rho_1 [f_1(w)] + \dots + \rho_k [f_k(w)]$$

Where ρ_i 's are distribution constants.

Substituting these in usual regression equation

$$Y = A_0 + A_1X_1 + A_2X_2 + \dots + A_nX_n + e$$

(here Y denoted yield and X_w rainfall in w-th week, $w = 1,2,\dots,n$) and utilising the properties of orthogonal and normalised functions, he obtained

$$Y = A_0 + a_0\rho_0 + a_1 \rho_1 + a_2\rho_2 + \dots + a_k\rho_k + e$$

where $A_0, a_0, a_1, a_2, \dots, a_k$ are constants to be determined and ρ_i ($i=1, \dots, k$) are distribution constants of X_w . Fisher has suggested to use $k = 5$ for most of the practical situations. In fitting this equation for $k = 5$, the number of constants to be evaluated will remain 7, no matter how finely growing season is divided. This model was used by Fisher for studying the influence of rainfall on the yield of wheat.

Hendricks and Scholl (1943) have modified Fisher's technique. They divided the crop season into n weekly intervals and have assumed that a second degree polynomial in week number would be sufficiently flexible to express the effect of weather on yield in successive periods. Further, they used values of weather variables as such. Mathematically

$$A_w = a_0 + a_1w + a_2w^2$$

In particular, $A_1 = a_0 + 1.a_1 + 1^2.a_2$
 $A_2 = a_0 + 2.a_1 + 2^2.a_2$

 $A_n = a_0 + n.a_1 + n^2.a_2$

Substituting the expression for A_w in regression equation, the model was obtained as

$$Y = A_0 + a_0 \sum_w X_w + a_1 \sum_w w X_w + a_2 \sum_w w^2 X_w + e$$

In this model number of constants to be determined reduces to 4, irrespective of n.

This model was extended for two weather variables to study joint effects.

The model obtained was

$$Y = A_0 + a_0 \sum_w X_{1w} + a_1 \sum_w w X_{1w} + a_2 \sum_w w^2 X_{1w} +$$

$$b_0 \sum_w X_{2w} + b_1 \sum_w w X_{2w} + b_2 \sum_w w^2 X_{2w} +$$

$$c_0 \sum_w X_{1w} X_{2w} + c_1 \sum_w w X_{1w} X_{2w} + c_2 \sum_w w^2 X_{1w} X_{2w} + e$$

Since the data for such studies extended over a long period of years, an additional variate T representing the year was included to make allowance for time trend.

Another important contribution in this field is by Baier (1977). He has classified the crop-weather models in three basic types.

1. Crop growth simulation models
2. Crop-weather Analysis models
3. Empirical statistical models

24.3 Crop-growth simulation models

A crop growth simulation model may be defined as a simplified representation of the physical, chemical and physiological mechanisms underlying plant growth processes. If the basic plant processes - production and distribution of dry matter and water relations are properly understood and modelled, the entire response of the plant to the environmental conditions can be simulated. Therefore, there is no need to differentiate between climatic regions, since the simulation model itself will show the limiting factors for growth. In humid climates with low temperature and radiation levels, the model will generally show the greatest response of yields to increase in total radiation received. In an arid and hot climate it will show the greatest response to the distribution and total amount of precipitation. Various time intervals can be introduced in simulation models, for example, in view of the daily cycle of many plant processes, hourly intervals are most practical. It is then assumed that the rate calculated for a particular moment does not change appreciably over a period of one hour. It is possible to evaluate thereby specific processes such as photosynthesis, transpiration or respiration for an hour and then accumulate the hourly rates over the day and the daily rates over the growing season in order to arrive at the total seasonal dry matter production or yield of economic products. Simulation programme must be regarded more as a guide to research into the behaviour of biological systems rather than as a final solution. Simulation can be most useful if the model accounts for most relevant phenomena and contains no false assumptions. Simulation provides an insight into crop-weather relationships, explains why some factors are more important for yield than others, suggests factors likely to have statistical significance and provides the basis for new experiments on processes which appear to be important but are not yet sufficiently understood. Thus, the simulation approach does not replace the statistical approach, but is complementary to it.

24.4 Crop-weather analysis models

Crop-weather analysis models are defined here as the product of two or more factors, each representing the (simplified) functional relationship between a particular plant response (e.g. yield) and the variations in selected variables at different plant developmental phases. The overall effects, as expressed by the numerical values of the factors modify each other but are not additive as in the case of a multivariate linear regression equation. Such models do not require a formulated hypothesis of the basic plant and environmental process; thus, the input requirements are less stringent but the output information is more dependent on the input data and less detailed than in the case of simulation models. Therefore, crop-weather analysis models are a practical research tool for the analysis of crop responses to weather and climate variations when only climatological data are available. Conventional statistical procedures are used in such models to evaluate the coefficients relating crop responses to climatological or derived agrometeorological data. A convenient time interval is one day, but in practice shorter or longer periods can also be used, provided the response characteristics of the crops do

not change appreciably over the selected period in relation to the variable taken into consideration.

24.5 Empirical Statistical Models

In the empirical approach, one or several variables (representing weather or climate, soil characteristics or a time trend) are related to crop responses such as yield. The weighting coefficients in these equations are by necessity obtained in an empirical manner using standard statistical procedures, such as multivariable regression analysis. This statistical approach does not easily lead to an explanation of the cause and effect relationships but it is a very practical approach for the assessment or prediction of yields. The coefficients in such empirical models and the validity of the estimates depend to a large extent on the design of the model, as well as on the representiveness of the input data. If the soil and climate conditions and the cropping practices are fairly homogeneous over the area represented by the input data, or if soil and geography are properly weighted in the equations, then it can be expected that the coefficients and the estimates have practical significance for the assessment of the crop conditions or prediction of yields for the specific area in question.

Several Empirical Statistical models were developed all over the world. The independent variables included weather variables, agrometeorological variables, soil characteristics or some suitably derived indices of these variables. Water Requirement Satisfaction Index (WRSI), Thermal Interception Rate Index (TIR), Growing Degree Days (GDD) are some agroclimatic indices used in models. Southern Oscillation Index (SOI) has also been used with other weather variables to forecast crop yield (Ramakrishna et al. 2003). To account for the technological changes year variable or some suitable function of time trend was used in the model. Some workers have also used two time trends. Moving averages of yield were also used to depict the technological changes.

In contrast to empirical regression models, the Joint Agricultural Weather Information Centre employs the crop weather analysis models that simulate accumulated crop responses to selected agrometeorological variables as a function of crop phenology. Observed weather data and derived agrometeorological variables are used as input data.

M Frere and G.F. Popov (1979) used the method which utilises actual rainfall and climatological information for the calculation of water requirement of crops and in turn crop water balance. The method is based on a cumulative water balance established over the whole growing season for the given crop and for successive periods of 10 days or a week. The water balance is difference between precipitation received by the crop and the water lost by the crop and the soil. Based on water surplus and deficit they have calculated index. Initially the index is taken to be 100 and is modified in successive decades/weeks depending on the water surplus or deficits. This index has been shown to be directly related to yield and can give a very satisfactory and early qualitative estimation of yields in rainfed crops. It may be possible to derive quantitative estimations of yields also but these estimates will have to be based on the potential yield of crops which will depend on local environmental conditions and will vary from place to place. It may also be mentioned that the method is intended mainly for utilisation in developing countries, where in rainfed agriculture the main constraint is generally inadequate availability of water to the crop. Therefore, the method does not directly involve the temperature which conditions the growth of the crop. However, the

temperature intervenes indirectly in three ways in the method of crop water balance assessment. Firstly, the effect of air temperature may be noticed in the length of the growing cycle which is generally directly dependent on temperature. Further air temperature intervenes directly in the calculation of potential evapotranspirations and in this respect influences the whole water balance. Finally the external temperatures may be important in some climatic zones, particularly as regards frosts.

In India, major organisations involved in developing methodology for forecasting crop yield based on weather are IMD and IASRI. The methodology adopted by IMD involves identification of significant correlations between yield and weather factors during successive overlapping periods of 7 to 60 days of the crop growing season. By analysing the correlation coefficients for statistical and phenological significance, the critical periods when the weather variables have significant effect on yield are identified. The weather variables in critical periods are used through multiple regression analysis to obtain forecast equations. Using this methodology models were developed for principal crops on meteorological subdivisions basis. Data from various locations are averaged to get the figures for meteorological sub-divisions and these are utilized alongwith time trend to develop the forecast model. Monthly forecasts are issued from these models by taking the actual data upto time of forecast and normal for the remaining period. In some models yield Moisture Index, Generalised Monsoon Index, Moisture Stress, aridity anomaly Index are also used (Sarwade, 1988; Sarkar, 2002).

24.6 Weather indices based models

At IASRI, the model suggested by Hendricks and Scholl has been modified (Agrawal et al 1980; 1983; Jain et al 1980) by expressing effects of changes in weather variables on yield in the successive periods as second degree polynomial in respective correlation coefficients between yield and weather variables. This will explain the relationship in a better way as it gives appropriate weightage to different periods. Under this assumption, the models were developed for studying the effects of weather variables on yield using complete crop season data whereas forecast model utilised partial crop season data. These models were found to be better than the one suggested by Hendricks and Scholl.

These models were further modified (Agrawal et al 1986) by expressing the effects of changes in weather variables on yield in successive periods as a linear function of respective correlation coefficients between yield and weather variables. As trend effect on yield was found to be significant, its effect was removed from yield while calculating correlation coefficients of yield with weather variables to be used as weights. Effects of second degree terms of weather variables were also studied. The results indicated that (i) the models using correlation based on yield adjusted for trend effect were better than the ones using simple correlations, (ii) inclusion of quadratic terms of weather variables and also the second power of correlation coefficients did not improve the model. This suggests that the following models can be used to study effects of weather on yield and its forecasts.

$$Y = A_0 + a_0Z_0 + a_1Z_1 + cT + e$$

$$\text{where } Z_j = \sum_{w=1}^n r_w^j X_w \quad ; \quad j = 0, 1$$

Here Y is yield r_w is correlation coefficient between the weather variable in w -th period (X_w) with yield (adjusted for trend effect), and e is error term. The models were further extended for studying joint effects.

The forecast model has been developed using partial crop season data considering all weather variables simultaneously. The model finally recommended was of the form

$$Y = A_0 + \sum_{i=1}^p \sum_{j=0}^1 a_{ij} Z_{ij} + \sum_{i \neq i'=1}^p \sum_{j=0}^1 a_{ii'j} Z_{ii'j} + cT + e$$

where

$$Z_{ij} = \sum_{w=1}^m r_{iw}^j X_{iw} \quad \text{and} \quad Z_{ii'j} = \sum_{w=1}^m r_{ii'w}^j X_{iw} X_{i'w}$$

$r_{iw}/r_{ii'w}$ is correlation coefficient of Y with i -th weather variable/product of i -th and i' -th weather variable in w -th period. m is period of forecast and p is number of weather variables used.

In this approach, for each weather variable, two types of indices were developed, one as simple total values of weather variable in different periods [un-weighted index $-Z_{i0}$] and the other one as weighted total [weighted index Z_{i1}] weights being correlation coefficients between yield /de-trended yield (if trend is present) and weather variable in respective periods. On similar lines, for studying joint effects, un-weighted & weighted indices for interactions were computed with products of weather variables (taken two at a time). Stepwise regression technique was used to select important indices in the model.

These models were used to forecast yield of rice and wheat in different situations, viz (i) rainfed area having deficient rainfall (rice), (ii) rainfed area having adequate rainfall (rice) and (iii) irrigated area (wheat). The results revealed that reliable forecasts can be obtained using this approach when the crops are 10-12 weeks old. This approach was also used to develop forecast model for sugarcane (Mehta, et al. 2000). However, these studies were carried out at district level and required a long series data of 25-30 years which are not available for most of the locations. Therefore, the study has been undertaken to develop the model on agro-climatic zone basis for rice and wheat by combining the data of various districts within the zone so that a long series could be obtained in a relatively shorter time. Previous years yield, moving averages of yield and agricultural inputs were taken as the variables taking care of variation between districts within the zone. Year was included to take care of technological changes. Different strategies for pooling district level data for the zone were adopted. Results revealed that reliable forecasts can be obtained using this methodology at 12 weeks after sowing i.e. about 2 months before harvest. The data requirement reduced to 10-15 years as against 25-30 years for district level models. The study also revealed that forecast model will be appropriate to forecast the yield of zone even if data for some districts within the zone are not available at model development stage or at forecasting stage (Agrawal et al. 2001). The approach has been successfully used for forecasting yields of rice, wheat, sugarcane and potato for Uttar Pradesh. (Agrawal, et al. 2005, Mehta, et al. 2010).

24.7 Complex Polynomial through GMDH technique

This methodology has been successfully applied by Mustafi and Chaudhuri (1981) for forecasting monthly tea crop production. At IASRI use of this technique was explored for forecasting potato yield in Uttar Pradesh (Mehta, et al. 2010).

The main feature of this technique is that it itself selects the structure of the model without using a prior information about relationship of dependent variable (y) with independent variables (x_1, x_2, \dots, x_p).

The fitted polynomial is of the form

$$y = a + \sum_{i=1}^p b_i x_i + \sum_{j=1}^p c_{ij} x_i x_j + \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p d_{ijk} x_i x_j x_k + \dots$$

The technique involves fitting of quadratic equations for all pairs of independent variables and identifying a few best performers in terms of predictive ability (using appropriate statistic); converting entire set of independent variables (called zero generation variables) to new variables (first generation variables) which are obtained as predicted values from these selected quadratic equations (of zero generation variables). The process of fitting and identifying best quadratic equations is repeated using first generation variables and second generation variables are obtained. The whole process is repeated with every new generation of variables till appropriate model is obtained (using certain criteria). At final stage, one best quadratic equation is selected as the final model.

Two approaches are followed for identifying few best performers. In the first approach, available data set is divided into two non-overlapping sets - training set and checking set. From training set, quadratic equations are fitted and checking set is used to test the predictability of different quadratic equations using root mean square error. In the other approach, PRESS statistic (predicted sum of square) is used wherein the whole data set is used as fit and check set. For evaluation of PRESS, each data point (one by one) is taken as a testing set (of size 1) and is predicted from the quadratic equation fitted from the remaining n-1 data points PRESS is calculated as

$$PRESS = \sum_{s=1}^n [y_s - \hat{y}_s^{(s)}]^2$$

where y_s is the value of the dependent variable for s^{th} observation and $\hat{y}_s^{(s)}$ is the predicted value of y_s computed from an equation based on remaining n -1 data points.

Theoretically the generation of variables is continued till the decreasing trend of PRESS statistic ceases i.e. PRESS statistic starts increasing. Some times even after number of generations the decreasing trend continues. This increases complexity of the model. In practical situations when there is abrupt decrease in value of PRESS

statistic and at the same time the coefficient of determination is quite high, the procedure is terminated.

This approach was used to obtain district as well as agroclimatic zone level models for potato in Uttar Pradesh using weather indices (unweighted and weighted) as explanatory variables. The performance of this model was found to be better than indices based regression approach for Bareilly district and north eastern zone. For remaining districts and zones the performance was worse or at par with the indices based regression approach.

24.8 Discriminant Function Technique

Discriminant function technique is a linear / quadratic function that discriminates different groups the best. Use of this technology has been explored for obtaining quantitative forecast of yields for rice in Raipur district. This methodology involved grouping the long series of years into three groups - congenial, normal and adverse with respect to crop yields (adjusted for trend effect, if any). Using weather data of these groups, linear / quadratic discriminant functions were obtained using phasewise weather data. Weather scores for each year at different phases of crop growth obtained through these discriminant functions were used alongwith inputs and time trend as regressors in model development through stepwise regression. Quadratic discriminant function was found appropriate and the methodology could provide reliable forecast two months before harvest. (Rai and Chandrahas 2000). The methodology was further modified using weekly weather data. Various strategies were proposed to solve the problem of number of variables more than number of observations. The study was carried out to forecast wheat yield in Kanpur district. The finally recommended strategy involved following steps. Discriminant functions were developed using weather variables data of first week. These discriminant functions were used to compute scores for each year. Taking data on weather variables in the second week and discriminant scores computed from the first week, discriminant function analysis was carried out which provided scores for each year based on data upto second week. The process was repeated for successive weeks data till the time of forecast and finally discriminant scores based on entire data were obtained for each year which alongwith trend were used to develop the model through stepwise regression technique. In contrast to earlier model by Rai et al. this model was based on complete data upto the time of forecast and relative importance of weather variables in different weeks (Aditya 2008). A detailed study using the recommended strategy has been carried out forecasting rice, wheat and sugarcane in Uttar Pradesh. Methododlogy was found to be successful for obtaining district / agroclimatic zone/state level forecasts. Performance was found to be better than weather indices based regression models for some cases whereas in some cases reverse trend was found (Chandrahas et al. 2010). However, this approach requires larger data base as compared to weather indices based regression approach.

24.9 Principal Component Regression

Principal component regression is a well known technique to reduce number of explanatory variables in the model. The technique involves conversion of explanatory variables into a set of uncorrelated variables with variances in descending order (known as principal components). The whole variation of the system explained by explanatory variables is explained by first few principal components which are used as regressors in the model in place of original variables. Besides solving the problem

of number of explanatory variables more than number of observations, the technique also solves the problem of multicollinearity. The approach has been attempted for forecasting yields of rice, wheat and sugarcane in Uttar Pradesh but the approach was not found to be successful (Chandrabhas et al. 2010).

24.10 Water Balance Technique

The concept of potential evapotranspiration (PET) was introduced by Thornthwaite and Penman in 1948 which use of agrometeorological variables in yield assessment models. Performance of models using agrometeorological variables was found to be better than the models using only meteorological variables (Baier and Robertson 1968). This is due to the fact that agrometeorological models use variables like soil moisture, actual and potential evapotranspiration, crop water requirement, effective rainfall etc. which take into account the soil properties and crop characteristics in addition to meteorological variables. These variables are estimated using Water Balance technique. This technique is useful for rainfed crops.

The water balance technique or model is a simple equation keeping an account of receipt and expenditure of water from the soil reserves. Mathematically

$$S_i = S_{i-1} + R_i - E_i$$

where S_i , R_i and E_i represent soil moisture, water received and expenditure respectively at the end of i -th period. The receipt (R) is in the form of rainfall and expenditure (E) is in the form of evapotranspiration from soil and crop cover. The stress depends on the demand and availability. When demand is more than the availability the stress occurs. Degree of stress depends on the gap between demand and availability. The variation in the model used by different research workers is in the form of variables used for indicating receipt and expenditure and the limitation imposed on these variables. For example in this equation, i represents the period after which balancing is done. It can be a day, week, fortnight or month. The number of periods depend on the length of the crop season, starting from the time of sowing. The depth of root zone can be kept constant for the entire season or can vary with the age of the crop. Receipt can be in the form of rainfall assuming that it is absorbed by the soil at a constant rate irrespective of the intensity of rain and antecedent moisture condition of soil. It can be made more realistic by using the actual infiltration rate of water for a given soil under a given intensity of rain and antecedent moisture condition. Similarly, expenditure which depends on demand can be estimated in different ways. It can simply be PET, the amount that can potentially evapotranspire under a given climatic condition or a fraction of it i.e. PET/2 or PET/4, irrespective of crop, and its stage of growth. This form of expenditure/demand is used when the objective is crop planning, estimation of drought proneness or agroclimatic classification of an area. For the purpose of monitoring and assessing yield of a crop, actual water requirement (WR) of the crop is taken as demand. The water requirement (at different time points) is estimated by multiplying the crop coefficients k (representing the crop characteristic and stage of the crop growth) and the value of evaporation (representing the loss of water to climatic condition such as temperature, wind velocity, humidity and sunshine at that stage). Estimates of evaporation are obtained either directly from an evaporimeter or from different formulae developed by Thornthwaite (1948), Penman(1948) and

Christiansen(1966). For estimates obtained from different formulae different values of k are used. The values of k for different crops or groups of crops under given climatic conditions are developed by agronomists and water technology scientists by conducting field experiments.

If the receipt is more than the expenditure, the excess water is stored in the soil depending upon the water holding capacity and water already stored in the soil. If excess water is more than the retention capacity of the soil, it goes waste as runoff. Thus, in the process of estimating soil moisture an estimate of runoff is also obtained. The amount of rain water used by plants and stored in the soil is taken as effective rainfall (ER), when rainfall is not enough to meet the crop water requirement, the requirement is met from the soil moisture stored in the root-zone. When rain and soil moisture together are not enough to meet the requirement the amount actually available is extracted by the plants and part of requirement remains unfulfilled. In this manner an estimate of actual evapotranspiration (AE) is obtained. When AE is less than WR a stress to the crop occurs. $1 - (AE/WR)$ is used as estimate of stress to the crop. The ratio AE/PE is also sometimes used as variable for yield assessment. Since the final yield is the outcome of the aggregate of water/moisture availability or non-availability through its life cycle an accumulated stress or satisfaction index is prepared. As the deficiency at critical stages causes greater damage, weights are assigned to stress at different stages according to their importance. In this manner an accumulated weighted stress index (SI) or water requirement satisfaction Index (WRSI) is obtained for each season. The index is related to yield through a regression model. The accumulated index helps in crop monitoring right from the date of sowing and provides pre-harvest estimate of yield during any time of the season. Therefore, pre-requisite for a seemingly simple model are estimates of many parameters of soil like depth, water holding capacity, wilting point, field capacity, infiltration rate under different antecedent moisture condition and crop parameters like crop coefficients, rooting pattern, and water extraction pattern of roots. Also a knowledge of critical stages of growth, an insight into effect of stress at different stages and an estimate of initial soil moisture are necessary. The success of the model depends on the accuracy of estimates of these parameters.

In a study conducted at IASRI a water balance model of the following form was used

$$S_i = S_{i-1} + ER_i - WR_i$$

for estimating moisture stress to the pearl millet crop of IARI, New Delhi. Balancing was done at the end of each day of the crop season. Depth of root zone varied with the age of the crop. A new method was developed to calculate effective rainfall (ER) on the basis of amount of rain and antecedent moisture condition. ER was used in the model in place of rainfall. Estimates of moisture stress and moisture surplus were obtained from the model. Weights were assigned to stress at different stages. Detrimental effect of excess water in the root-zone was also taken into account and weights were assigned to surplus water also depending upon the time of its occurrence. An accumulated stress and surplus moisture index (SI) was prepared. SI was related to yield through a regression equation. The fitted equation explained 91% variation in pearl millet yield. A reduction of 42.7 kg/ha was expected due to per unit of stress in the potential yield of 3000 kg/ha. The error in predicted yield of two years were was 3.2% and 0.5% respectively (Saksena and Bhargava 1995).

In another study, models were developed for rainfed sorghum, maize and rice using agrometeorological indices. Water balance was carried out at weekly intervals. Weighted stress index was prepared phase-wise by applying weights to surplus as well as deficit moisture depending upon the stage at which it occurred. Stress index of phase 2 i.e. 4 to 7th weeks after sowing played an important role in determining the yield of sorghum both at Delhi and Parbhani (Maharashtra) district. Models with phase 2 Index and trend variables as regressors could forecast yield 6 weeks before harvest. Deviations in forecast and observed yield varied between 3.5% to 11%. Similarly for maize crop at Delhi model with trend, index of surplus moisture at phase 1 and 2 and deficit moisture at phase 3 and 4 as predictor variables could forecast yield 4 weeks before harvest. Deviation in the forecast and observed yield was only 4.8 %. Model for rice in Raipur district included trend and accumulated index for the five phases up to maturity as explanatory variables .

24.11 Artificial Neural Network Technique

In contrast to regression approach, Artificial Neural Network (ANN) technique has been explored for forecasting yields of rice, wheat and sugarcane in Uttar Pradesh. (Kumar et al. 2010). This is an attractive tool under machine learning techniques for forecasting and classification purposes. ANNs are data driven self-adaptive methods in that there are few apriori assumptions about the models for problems under study. These learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe. After learning from the available data, ANNs can often correctly infer the unseen part of a population even if data contains noisy information. As forecasting is performed via prediction of future behaviour (unseen part) from examples of past behaviour, it is an ideal application area for ANNs, at least in principle. (Dewolf et al. 1997, 2000). However, the technique requires a large data base.

24.12 Forewarning systems for crop pests and diseases

Pests and diseases are one of the major causes of reduction in crop yield. Timely application of remedial measures may reduce the yield loss. For application of these measures one must have prior knowledge of the time and severity of the outbreak of these pests and diseases. Forecasting system can help in this direction.

In pests and diseases forewarning system, the variables of interest may be maximum pest population / disease severity, pest population / disease severity at most damaging stage of the crop, pest population / disease severity at different stages of crop growth or at various standard weeks, time of first appearance of pests / diseases, time of maximum pest population / disease severity, time of pest population / disease severity crossing threshold limit, extent of damage, weekly monitoring of pests and diseases progress, occurrence / non-occurrence of pests and diseases. If data are available at periodic interval for 15-20 years, the detailed study can be carried out for different variables of interest. However, depending upon the data availability, different types of models can be utilized for developing forewarning system. The models could be of two types, 'Between year model' and 'Within year model'.

24.12.1 Between year models

These models are developed using previous years' data. An assumption is made that the present year is a part of the composite population of the previous years and accordingly the relationships developed on the basis of previous years' data will be applicable for the present year. The forecast for pests and diseases are obtained by substituting the current year data into the model developed upon the previous years. Various methods have been attempted when data are available in quantitative form. Some of the important techniques are discussed below :

24.12.1.1 Thumb rule

This approach is the most common and extensively used. It is a simple system which describes the forecasting of the pests and diseases based on past experience. For example for potato late blight, a day is favorable if

- the 5 day temperature average is $< 25.5^{\circ}\text{C}$
- the total rainfall for the last 10 days is $> 3.0\text{ cm}$
- the minimum temperature on that day is $> 7.2^{\circ}\text{C}$

When this situation arises, there is a possibility of potato late blight appearance.

24.12.1.2 Regression Model

The regression model taking pest / disease variable as dependent and suitable independent variables such as weather variables, crop stages, population of natural enemies/predators etc. is used. The form of the model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + e$$

where $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are regression coefficients, X_1, X_2, \dots, X_p are independent variables and e is error term. These variables are used in original scale or on a suitable transformed scale such as cos, log, exponential etc. (Coakley et al 1985; Trivedi et al. 1999).

24.12.1.3 Fuzzy regression

In regression analysis, the unfitted errors between a regression model and observed data are generally assumed as observation error that is a random variable having a normal distribution, constant variance, and a zero mean. In fuzzy regression analysis, the same unfitted errors are viewed as the fuzziness. Fuzzy regression can be quite useful in estimating the relationship among variables where the availability data are imprecise and fuzzy.

Fuzzy regression analysis gives a fuzzy functional relationship between dependent and independent variables where vagueness is present in some form. There are three situations where the fuzzy analysis can be viewed *viz.* Crisp parameters and fuzzy data, Fuzzy parameters and crisp data and Fuzzy parameters and fuzzy data. Fuzzy regression method is based on minimizing fuzziness as an optimal criterion, which can be achieved by linear programming procedures.

24.12.1.4 Growing Degree Day Approach

This method is based on the assumption that the pest becomes inactive below a certain temperature known as base temperature. Growing degree day is worked out as

$$\text{GDD} = \Sigma (\text{mean temp.} - \text{base temp.})$$

GDD is used in the model as explanatory variable. This method requires proper knowledge of base temperature and initial time from which accumulation is to start.

24.12.1.5 Model based on weather indices

In this approach, using weekly and fortnightly weather variables suitable indices are worked out which are used as regressors in the model. The model is of the form

$$Y = a_0 + \sum_{i=1}^p \sum_{j=0}^1 a_{ij} Z_{ij} + \sum_{i \neq i'}^p \sum_{j=0}^1 b_{ii'j} Z_{ii'j} + e$$

where

$$Z_{ij} = \sum_{w=n_1}^{n_2} r_{iw}^j X_{iw} \quad \& \quad Z_{ii'j} = \sum_{w=n_1}^{n_2} r_{ii'w}^j X_{iw} X_{i'w}$$

Y variable to forecast; X_{iw} is value of i^{th} weather variable in w^{th} period; $r_{iw} / r_{ii'w}$ is suitable weight given to i^{th} weather variable / product of i^{th} and i'^{th} weather variable in w^{th} period; p is number of weather variables considered; n_1 and n_2 are the initial & final periods for which weather data were included in the model and e is error term.

If information on favourable weather conditions is known, subjective weights based on this information can be used for constructing weather indices. In absence of such information correlation coefficients between Y and respective weather variable/product of weather variables can be used [Agrawal *et al.* (2004), Chattopadhyay *et al.* (2005-a), Chattopadhyay *et al.* (2005-b), Desai *et al.* (2004) and Dhar *et al.* (2007)]

24.12.1.6 Principal component regression

Forewarning models can be developed using the principal component technique as normally relevant weather variables are large in number and are expected to be highly correlated among themselves. Using the first few principal components of weather variables as independent variables forecast models can be developed.

24.12.1.7 Discriminant function analysis

Forewarning models of pests and diseases based on time series data on weather variables can be developed using the discriminant function analysis. For this analysis, a series of data for 25-30 years are required. Based on the pest and diseases variables, data can be divided into different groups – low, medium and high etc. and using weather data in these groups, linear or quadratic discriminant functions can be fitted which can be used to find discriminant scores. Considering these discriminant scores as independent variables and diseases / pest as a dependent variable, regression analysis can be performed. Johnson *et al.* (1996) used discriminant analysis for forecasting potato late blight.

24.12.1.8 Complex polynomial [Group Method of Data Handling (GMDH)]

It provides complex polynomial in independent variables. It selects the structure of the model itself without prior information about relationship. Form of the model :

$$Y = a + \sum_{i=1}^m b_i X_i + \sum_{i=1}^m \sum_{j=1}^m c_{ij} X_i X_j + \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m d_{ijk} X_i X_j X_k + \dots$$

The technique involves fitting of quadratic equations for all pairs of independent variables and identifying a few best performers in terms of predictive ability (using appropriate statistics); converting entire set of independent variables (called zero generation variables) to new variables (first generation variables) which are obtained as predicted values from these selected quadratic equations (of zero generation variables). The process of fitting and identifying best quadratic equations is repeated using first generation variables and second generation variables are obtained. The whole process is repeated with every new generation of variables till appropriate model is obtained (using certain criteria). At final stage, one best quadratic equation is selected as the final model. (Bahuguna et al 1992; Trivedi et al 1999).

24.12.1.9 Machine Learning Techniques

Machine learning techniques offer many methodologies like decision tree induction algorithms, genetic algorithms, neural networks, rough sets, fuzzy sets as well as many hybridized strategies for the classification and prediction (Han and Kamber, 2001; Pujari, 2000; Komorowski *et al.*, 1999; Witten and Frank, 1999). Decision tree induction represents a simple and powerful method of classification that generates a tree and a set of rules, representing the model of different classes, from a given dataset. Decision Tree (DT) is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node represents the class. The top most node in a tree is the root node. For decision tree ID3 algorithm and its successor C4.5 algorithm by Quinlan (1993) are widely used. One of the strengths of decision trees compared to other methods of induction is the ease with which they can be used for numeric as well as nonnumeric domains. Another advantage of decision tree is that it can be easily mapped to rules. Artificial Neural Networks (ANNs) is another attractive tool under machine learning techniques for forecasting and classification purposes. ANNs are data driven self-adaptive methods in that there are few a priori assumptions about the models for problems under study. These learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe. After learning from the available data, ANNs can often correctly infer the unseen part of a population even if data contains noisy information. As forecasting is performed via prediction of future behaviour (unseen part) from examples of past behaviour, it is an ideal application area for ANNs, at least in principle. (Agrawal et al. 2004; Dewolf et al. 1997, 2000; Kumar, et al. 2010). However, the technique requires a large data base.

24.12.1.10 Deviation Method

This method can be utilized when periodical data at different intervals during the crop season are available for only 5-6 years. The pest population at a given point of crop stage is assumed to be due to two reasons – natural cycle of the pest and weather.

To identify the natural cycle, data at different intervals is averaged over years and a suitable model is fitted to these averaged data points. Then the entire data is converted as deviations from the predicted natural cycle. Appropriate model is fitted using these deviations as dependent and weather as independent variables. [Mehta *et al.*(2001)]

24.12.1.11 Ordinal logistic model – model for qualitative data

The timely control measures to prevent pest / disease outbreak can be taken even if the information on the extent of severity is not available but merely the epidemic status is accessible. This information could be obtained through modeling qualitative data. Such models have added advantage that these could be obtained even if the detailed and exact information on pest count / disease severity is not available but only the qualitative status such as epidemic or no epidemic / low, medium or high is known. Such a situation arises quite often in pest / disease data. In such cases, the data are classified as 0/1 (2 categories); 0,1,2 (three categories). The logistic regression is used for obtaining probabilities of different categories. For example, for two categories, the model is of the form :

$$P(E=1) = \frac{1}{1 + \exp(-z)} + e$$

where z is a function of weather variables.

Forecast / Prediction rule :

- If $P \geq .5$ more chance of occurrence of epidemic
- If $P < .5$ probability of occurrence of epidemic is minimum

(Mehta *et al.* 2001; Mishra *et al.* 2004; Johnson *et al.* 1996; Agrawal, *et al.* 2004)

24.12.2 Within year model

Sometimes, past data on pests and diseases are not available but the pests and diseases status at different points of time during the crop season are available for the current season only. In such situations, within years growth model can be used for forewarning maximum disease severity / pest population, provided there are 10-12 data points between time of first appearance of pest / disease and maximum or most damaging stage.

The methodology consists in fitting appropriate growth pattern to the pests and diseases data based on partial data and using this growth curve for forecasting the maximum value of variable of interest. A number of growth models such as logistic, Gompertz etc. can be used for this purpose (Agrawal *et al.* 2004). Prajneshu (1998) developed a non linear statistical model for describing the dynamic population growth.

References

- Agrawal, Ranjana; Jain, R.C.; Jha, M.P. and Singh, D. (1980). Forecasting of rice yield using climatic variables. *Ind.J.Agr. Sci.* 50(9) : 680-84.
- Agrawal, Ranjana; Jain, R.C. and Jha, M.P. (1983). Joint effects of weather variables on rice yield, *Mausam*, 34 (2) :189-94.

- Agrawal, Ranjana; Jain, R.C. and Jha, M.P. (1986). Models for studying rice crop-weather relationship, *Mausam*, 37(1): 67-70.
- Agrawal, Ranjana; Jain, R.C. and Mehta, S.C. (2001). Yield forecast based on weather variables and agricultural inputs on agroclimatic zone basis. *Indian Journal of Agricultural Science*, 71(7).
- Agrawal, Ranjana; Ramakrishna, Y.S.; Kesava Rao, A.V.R.; Kumar, Amrender; Bhar, Lal Mohan; Madan Mohan and Saksena, Asha (2005). Modeling for forecasting of crop yield using weather parameter and agricultural inputs. (AP Cess Fund project report).
- Agrawal, Ranjana, Mehta, S.C., Bhar, L.M. and Kumar, Amrender (2004). Development of weather based forewarning system for crop pests and diseases - Mission mode project under NATP.
- Baier, W. (1977). Crop weather models and their use in yield assessments. Tech. note no. 151, WMO, Geneva, 48 pp.
- Baier, W., Robertson, G.W. (1968). The performance of soil moisture estimates as compared with direct use of climatological data for estimating crop yields. *Agric. Meteorol.*, 5,7-31.
- Chandrasah; Agrawal, Ranjana and Walia, S.S. (2010). Use of discriminant function and principal component techniques for weather based crop yield forecasts. (IASRI, New Delhi).
- Chattopadhyay, C, Agrawal, R., Kumar, A., Bhar, L.M., Meena, P.D., Meena, R.L., Khan, S.A., Chattopadhyay, A.K., Awasthi, R.P., Singh, S.N., Chakravarthy, N.V.K., Kumar, A., Singh, R.B. and Bhunia, C.K. (2005-a). Epidemiology and forecasting of Alternaria blight of oilseed Brassica in India – a case study. *Zeitschrift für Pflanzenkrankheiten und Pflanzenschutz (Journal of Plant Diseases and Protection)*, 112(4), 351-365.
- Chattopadhyay, C., Agrawal, R., Kumar, Amrender, Singh, Y.P., Roy, S.K., Khan, S.A., Bhar, L.M., Chakravarthy, N.V.K., Srivastava, A., Patel, B.S., Srivastava, B., Singh, C.P. and Mehta S.C. (2005-b). Forecasting of *Lipaphis erysimi* on oilseed Brassicas in India – a case study. *Crop Protection*, 24, 1042-1053.
- Christiansen, J.E. (1966). Estimating pan-evaporation and evapotranspiration from climatic data. *Irrigation and drainage conference. Las Vegas, Nev. Nov. 2-4.*
- Dewolf, E.D. and Francl, L.J. (1997). Neural network that distinguish in period of wheat tan spot in an outdoor environment. *Phytopathology*, 87(1) : 83-7.
- Dewolf, E.D. and Francl, L.J. (2000). Neural network classification of tan spot and stagonospore blotch infection period in wheat field environment. *Phytopathology*, 20(2) :108-13.
- Dhar, Vishwa, Singh, S.K., Kumar, M., Agrawal, R. and Kumar Amrender (2007). 'Prediction of pod borer (*Helicoverpa armigera*) infestation in short duration pigeonpea (*Cajanus cajan*) in central Uttar Pradesh', *Indian Journal of Agricultural Sciences*, 77(10), 701-04.
- Fisher, R.A. (1924). The influence of rainfall on the yield of wheat at Rothamsted. Roy. Soc. (London), *Phil. Trans. Ser.B.*,213 : 89-142.
- Frere, M. and Popov, G.F. (1979). Agrometeorological crop monitoring and forecasting. FAO plant production and protection paper, plant production and protection division, FAO, Rome.
- Hendrick, W.A. and Scholl, J.C. (1943). Technique in measuring joint relationship. The joint effects of temperature and precipitation on crop yield. N. Carolina Agric. Exp. Stat. Tech. Bull. 74.

- Jain, R.C.; Agrawal, Ranjana and Jha, M.P. (1980). Effects of climatic variables on rice yield and its forecast, *Mausam* 31(4) : 591-96.
- Kumar, Amrender; Ramasubramanian, V. and Agrawal, Ranjana (2010). Neural network based forecast modeling crops. (IASRI Publication).
- Mehta, S.C.; Agrawal, Ranjana; Singh V.P.N. (2000): Strategies for composite forecast. *JISAS*, 53(3).
- Mehta, S.C.; Satya Pal; Kumar, Vinod (2010). Weather based models for forecasting potato yield in Uttar Pradesh. (IASRI Publication).
- Mustafi, A. and Chaudhary, A. S. (1981), Use of Multilayer Group Method of Data Handling for prediction of tea crop production, *Jour. Ind. Soc. Agril. Stat*, 33(1), 89–101.
- Prajneshu (1998). A non-linear statistical model for aphid population growth. *Jour. Ind. Soc. Agril. Stat.*, 51(1), 81-93.
- Pujari, A.K. (2000). *Data Mining Techniques*. Universities Press.
- Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufman.
- Rai, T. and Chandrahas (2000). Use of discriminant function of weather parameters for developing forecast model for rice crop. (IASRI Publication).
- Ramakrishna, Y.S.; Singh, H.P. and Rao, G. Nageswara (2003). Weather based indices for forecasting foodgrain production in India. *Journal of Agrometeorology* 5(1) : 1–11.
- Sarkar, J. (2002). Forecasting rice and wheat yield over different meteorological sub-divisions of India using statistical models. 56th Annual Conference of ISAS.
- Sarwade, G.S. (1988). Meteorological yield models. IRS-UP/SAC/CYM/TN/17/58, SAC, Ahmedabad.
- Saksena, Asha; Jain, R.C.; Yadav, R.L. (2001). Development of early warning and yield assessment model for rainfed crops based on agrometeorological indices. (IASRI publication).
- Stynes (1980). Crop loss assessment. *Mics. Publ. Univ. Minn. Agric. Exp. Stn.* 7.
- Thorntwaite, C.W. (1948). An approach towards rational classification of climate. *Geographical Review.* 38:55-94.
- Trivedi, T.P., Paul Khurana, S.M., Jain, R.C., Mehta, S.C. and Bhar, L.M. (1999), Development of Forewarning system for aphids (*Myzus persicae*) on potato, Annual Report NCIPM, N. Delhi.
- Witten, I.H. and Frank E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

OVERVIEW OF DESIGN OF EXPERIMENTS

Krishan Lal
Former Principal Scientist

Indian Agricultural Statistics Research Institute, New Delhi-110012
(klkalra@gmail.com)

25.1 Introduction

Design of Experiments is a formal mathematical method for systematically planning and conducting scientific studies that change experimental variables together in order to determine their effect of a given response. It makes controlled changes to input variables in order to gain maximum amounts of information on cause and effect relationships with a minimum sample size. It is more efficient than a standard approach of changing "one variable at a time" in order to observe the variable's impact on a given response. It generates information on the effect various factors have on a response variable and in some cases may be able to determine optimal settings for those factors.

In an experiment, we deliberately change one or more process variables (or factors) in order to observe the effect changes have on one or more response variables. Thus the design of experiments (DEO) is an efficient procedure for planning experiments so that the data obtained can be analyzed to yield valid and objective conclusions. DOE begins with determining the objectives of an experiment and selecting the process factors for the study. An *Experimental Design* is the laying out of a detailed experimental plan in advance of doing the experiment. Well chosen experimental designs maximize the amount of "information" that can be obtained for a given amount of experimental effort.

The basic principles of experimental designs are randomization, replication and local control. These principles make a valid test of significance possible. Each of them is described briefly in the following subsections.

25.2. Randomization

The first principle of an experimental design is randomization, which is assigning treatments to the experimental units randomly. It implies that every possible allotment of treatments has the same probability. An experimental unit (plot, animal, etc.) is the smallest division of the experimental material and a treatment means an experimental condition whose effect is to be measured and compared. The purpose of randomization is to remove bias and other sources of extraneous variation, which are not controllable. Another advantage of randomization (accompanied by replication) is that it forms the basis of any valid statistical test. Randomization is usually done by drawing numbered cards from a well-shuffled pack of cards, or by drawing numbered balls from a well-shaken container or by using tables of random numbers.

25.2.1 Replication

The second principle of an experimental design is replication. It is the repetition of the treatments under investigation to different experimental units. Replication is essential for obtaining a valid estimate of the experimental error and to some extent increasing the precision of estimating the pairwise differences among the treatment effects.

25.2.2 Local Control

Local Control (Blocking) is the simplest technique to take care of the variability in response because of the variability in the experimental material. To block an experiment is to divide, or partition, the observations into groups called blocks in such a way that the observations in each block are collected under relatively similar experimental conditions. The main purpose of the principle of local control is to increase the efficiency of an experimental design by decreasing the experimental error. If blocking is done well, the comparisons of two or more treatments are made more precisely than similar comparisons from an unblocked design.

The data generated through designed experiments exhibit a lot of variability. Even experimental units (plots) subjected to same treatment give rise to different observations thus creating variability. The statistical methodologies, in particular the theory of linear estimation, enable us to partition this variability into two major components. The first major component comprises of that part of the total variability to which we can assign causes or reasons while the second component comprises of that part of the total variability to which we cannot assign any cause or reason. This variability arises because of some factors unidentified as a source of variation. Howsoever careful planning is made for the experimentation; this component is always present and is known as experimental error. The observations obtained from experimental units identically treated are useful for the estimation of this experimental error. Ideally one should select a design that will give experimental error as small as possible. A popular measure of the experimental error is the Coefficient of Variation (CV). The other major component of variability is the one for which the causes can be assigned or are known. There is always a deliberate attempt on the part of the experimenter to create variability by the application of several treatments. So treatment is one component in every designed experiment that causes variability and that is the parameter of interest of the experimenter. The variability caused by other sources may occur in the experiment. The experimenter is always interested to reduce these variability(s) so as to reduce the experimenter error. Thus on the basis of variability caused by other sources, the designs may be classified as follows:

25.2.3 Designs of Zero-way elimination of heterogeneity

If the experimental material is homogeneous and does not exhibit any variability then the treatments are applied randomly to the experimental units. Such designs are known as the zero-way elimination of heterogeneity designs or completely randomized designs (CRD). Besides the variability arising because of the application of treatments, the variability present in the experimental material (plots) is the other

major known source of variability. These designs are used when all the experimental units are homogeneous. Generally experiments conducted in pots or industry experiments used as the completely randomized designs.

25.2.4 Designs of One-way elimination of heterogeneity

When the variability in experimental units is in one direction (slope) so that homogeneous groups (blocks) can be made perpendicular to this slope, the designs of one-way elimination of heterogeneity are used. These designs are called block designs. The most common block design is the randomized complete block (RCB) design. In this design all the treatments are applied randomly to the plots within each block. However, for large number of treatments the blocks become large if one has to apply all the treatments in a block, as desired by the RCB design. Therefore, there is a strong need to effectively control the variation through blocking. This necessitates the use of incomplete block designs. A block design is said to be an incomplete block design if the design has at least one block that does not contain all the treatments. Some common incomplete block designs are balanced incomplete block (BIB) design, partially balanced incomplete block (PBIB) design including Lattice designs – square and rectangular, cyclic designs, alpha designs, etc.

25.2.5 Designs of Two-way elimination of heterogeneity

Further, the variability in the experimental material may be in two directions and forming rows and columns can control this variability and the treatments are assigned to the cells. Each cell is assigned one treatment. Such designs are termed as two-way elimination of heterogeneity setting designs or the row-column designs. The most common row-column design is the Latin square design (LSD). The other row-column designs are the Youden square designs, Youden type designs, Generalized Youden designs, Pseudo Youden designs, etc.

The designs discussed above are for the varietal experiments with single factor at two or more than two levels. Sometimes we have the experimental situations when there are two or more than two factors each at two or more than two levels. These experiments are called factorial experiments.

25.2.5.1 Factorial Experiments

Experiments in which the effects (main effects and interactions) of more than one factor are studied together are called factorial experiments. In general if there are n factors, say, F_1, F_2, \dots, F_n and the i^{th} factor has s_i levels, $i=1, \dots, n$, then the total number of treatment combinations is $\prod_{i=1}^n s_i$. Factorial experiments are of two types.

1. **Symmetrical Factorial Experiments:** In these experiments the number of levels of all factors is same *i.e.*, $s_i = s \quad \forall i = 1, \dots, n$.
2. **Asymmetrical Factorial Experiments:** In these experiments the number of levels of all the factors are not same *i.e.* there are at least two factors for which the number of levels s_i 's are different.

Factorial experiments have many advantages over single factor experiments.

- These experiments provide an opportunity to study not only the individual effects of the factors but also their interactions.
- These experiments have the further advantage of economizing the experimental resources. When the experiments are conducted factor by factor a large number of experimental units are required for getting the same precision of estimation as one would have got when all the factors are experimented together in the same experiment, *i.e.*, factorial experiment. There is thus a considerable amount of saving of resources. Moreover, factorial experiments also enable us to study interactions which the experiments conducted factor by factor do not allow us to study.

25.1.5.2 Symmetrical factorial experiments

The simplest symmetrical factorial experiments are 2^n factorial experiments in which all the n factors have 2 levels each. Consider the 2^2 factorial experiment with 2 factors say A and B each at two levels, say 0 and 1 . There will be 4 treatment combinations that can be written as

- $00 = a_0 b_0 = (1)$; A and B both at first levels
- $10 = a_1 b_0 = a$; A at second level and B at first level
- $01 = a_0 b_1 = b$; A at first level and B at second level
- $11 = a_1 b_1 = ab$; A and B both at second level.

We denote the treatment combinations by small letters (1), a , b , ab indicating the presence of low or high level of the factor and treatment totals by $[1]$, $[a]$, $[b]$, $[ab]$. The following table gives the responses due to Factor A and Factor B .

Factor A → Factor B ↓	a_0 or 0	a_1 or 1	Response due to A
b_0 or 0	$[1]$ or $[a_0 b_0]$	$[a]$ or $[a_1 b_0]$	$[a] - [1]$ or $[a_1 b_0] - [a_0 b_0]$
b_1 or 1	$[b]$ or $[a_0 b_1]$	$[ab]$ or $[a_1 b_1]$	$[ab] - [b]$ or $[a_1 b_1] - [a_0 b_1]$
Response Due to B	$[b] - [1]$ or $[a_0 b_1] - [a_0 b_0]$	$[ab] - [a]$ or $[a_1 b_1] - [a_1 b_0]$	

The responses $[a] - [1]$ and $[ab] - [b]$ are called simple effects of the factor A at 0 and 1 levels, respectively of the factor B . If the factors A and B are independent, the responses $[a] - [1]$ and $[ab] - [b]$, both provide the estimate of the response due to A (except for the experimental error). The average of these two simple effects is known as Main Effect of factor A . Thus the main effect of factor A is

$$A = \frac{1}{2} \{ [a_1 b_1] - [a_0 b_1] + [a_1 b_0] - [a_0 b_0] \} \text{ or } A = \frac{1}{2} \{ [ab] - [b] + [a] - [1] \} \quad (1)$$

This is simplified by writing it in the form $A = \frac{1}{2} (a - 1)(b + 1)$, where the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by corresponding treatment totals. From (1) we find that A is a linear function of the four treatments totals with the sum of the coefficients of the linear function equal to zero $(\frac{1}{2} - \frac{1}{2} + \frac{1}{2} - \frac{1}{2} = 0)$. Such a linear function among the treatment totals with sum of coefficients equal to zero is called a contrast (or a comparison) of the treatment totals. Similarly the main effect of factor B is

$$B = \frac{1}{2} \{ [a_1b_1] + [a_0b_1] - [a_1b_0] - [a_0b_0] \} \text{ or } B = \frac{1}{2} \{ [ab] + [b] - [a] - [1] \} \quad (2)$$

This is simplified by writing it in the form $B = \frac{1}{2} (a + 1)(b - 1)$ where the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by corresponding treatment totals. From (2), we find that B is a linear function of the four treatments totals with the sum of the coefficients of the linear function equal to zero $(\frac{1}{2} + \frac{1}{2} - \frac{1}{2} - \frac{1}{2} = 0)$, hence a contrast.

Consider now the difference of two simple effects of A

$$= \{ [ab] - [b] - [a] + [1] \} \quad (3)$$

Had the two factors been independent, then (3) would be zero. If not then this provides an estimate of interdependence of the two factors and it is called the interaction between A and B . The interaction between A and B is defined as

$$AB = \frac{1}{2} (a - 1)(b - 1)$$

where the expression on the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by the corresponding treatment totals. It is easy to verify that AB is a contrast of the treatment totals. The coefficients of the contrasts A and AB are such that the sum of the products of the corresponding coefficients of the contrasts A and AB is equal to zero *i.e.* $(\frac{1}{2})(\frac{1}{2}) + (-\frac{1}{2})(-\frac{1}{2}) + (\frac{1}{2})(-\frac{1}{2}) + (-\frac{1}{2})(\frac{1}{2}) = 0$. Thus the contrasts A and AB are orthogonal contrasts. It is easy to verify that the interaction of the factor B with factor A , *i.e.*, BA is the same as the interaction AB and hence the interaction does not depend on the order of the factors. It is also easy to verify that the main effect B is orthogonal to both A and AB .

The above three orthogonal contrasts defining the main effects and interaction can be easily obtained from the following table, which gives the signs with which to combine the treatment totals and also the divisor for obtaining the corresponding sum of

squares. Main effects and interactions are expressed in terms of individual treatment totals.

<i>Treatment Totals</i> → <i>Factorial Effect</i> ↓	[1]	[a]	[b]	[ab]	Divisor
<i>M</i>	+	+	+	+	4 <i>r</i>
<i>A</i>	-	+	-	+	4 <i>r</i>
<i>B</i>	-	-	+	+	4 <i>r</i>
<i>AB</i>	+	-	-	+	4 <i>r</i>

Here *r* denotes the replication number. The rule to write down the signs of the main effect is to give a plus sign to the treatment combinations containing the corresponding small letter and a minus sign where the corresponding small letter is absent. The signs of interaction are obtained by multiplying the corresponding signs of the two main effects. The first line gives the general mean

$$M = \frac{1}{4} \{[ab] + [a] + [b] + [1]\}$$

Consider now the 2^3 factorial experiment with 3 factors *A*, *B*, and *C* each at two levels, say 0 and 1. The 8 treatment combinations are written as

- 000 = $a_0 b_0 c_0 = (1)$; *A*, *B* and *C*, all three at first level
- 100 = $a_1 b_0 c_0 = a$; *A* at second level and *B* and *C* at first level
- 010 = $a_0 b_1 c_0 = b$; *A* and *C* both at first level and *B* at second level
- 110 = $a_1 b_1 c_0 = ab$; *A* and *B* both at second level and *C* at first level
- 001 = $a_0 b_0 c_1 = c$; *A* and *B* both at first level and *C* at second level.
- 101 = $a_1 b_0 c_1 = ac$; *A* and *C* both at second level and *B* at first level
- 011 = $a_0 b_1 c_1 = bc$; *A* at first level and *B* and *C* both at second level
- 111 = $a_1 b_1 c_1 = abc$; *A*, *B* and *C*, all three at second level

In a three factor experiment there are 3 main effects *A*, *B*, and *C*; 3 first order or two factor interactions *AB*, *AC*, and *BC*; and *one* second order or three factor interaction *ABC*. The main effects and interactions may be written as

$$A = \frac{1}{4}(a-1)(b+1)(c+1), \quad B = \frac{1}{4}(a+1)(b-1)(c+1), \quad C = \frac{1}{4}(a+1)(b+1)(c-1)$$

$$AB = \frac{1}{4}(a-1)(b-1)(c+1), \quad AC = \frac{1}{4}(a-1)(b+1)(c-1), \quad BC = \frac{1}{4}(a+1)(b-1)(c-1)$$

$$ABC = \frac{1}{4}(a-1)(b-1)(c-1).$$

These main effects and interactions are mutually orthogonal.

Incidentally, it may be remarked that the method of representing the main effects and interactions, which is due to Yates, is very useful and quite straightforward. For example, if the design is 2^4 then

$$A = \frac{1}{2^3}(a-1)(b+1)(c+1)(d+1), AB = \frac{1}{2^3}(a-1)(b-1)(c+1)(d+1),$$

$$ABC = \frac{1}{2^3}(a-1)(b-1)(c-1)(d+1), \text{ and } ABCD = \frac{1}{2^3}(a-1)(b-1)(c-1)(d-1)$$

By this rule the main effect or interaction of any design of the series 2^n can be written out directly without first obtaining the simple effects and then expressing the main effects or interactions. For example,

$$A = \frac{1}{2^{n-1}}(a-1)(b+1)(c+1)(d+1)(e+1) \dots,$$

$$AB = \frac{1}{2^{n-1}}(a-1)(b-1)(c+1)(d+1)(e+1) \dots,$$

$$ABC = \frac{1}{2^{n-1}}(a-1)(b-1)(c-1)(d+1)(e+1) \dots,$$

$$\text{and } ABCD = \frac{1}{2^{n-1}}(a-1)(b-1)(c-1)(d-1)(e+1) \dots$$

In case of a 2^n factorial experiment, there will be $2^n (=v)$ treatment combinations. We shall have n main effects; $\binom{n}{2}$ first order or two factor interactions; $\binom{n}{3}$ second order or three factor interactions; $\binom{n}{4}$ third order or four factor interactions and so on, $\binom{n}{r}$, $(r-1)^{th}$ order or r factor interactions and $\binom{n}{n}$, $(n-1)^{th}$ order or n factor interaction.

Using these v treatment combinations, the experiment may be laid out using any of the suitable experimental designs viz. completely randomized design or block designs or row-column designs, etc.

The analysis of experimental designs can be performed by using the technique of Analysis of Variance.

25.3 Analysis of Variance (ANOVA)

Analysis of Variance is a technique of partitioning the overall variation in the responses into different assignable sources of variation, some of which are specifiable and others unknown. Total variance in the sample data is partitioned is expressed as the sum of its non-negative components is a measure of the variation due to some specific independent source or factor or cause. ANOVA consists in estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to ascribable factors (causes) with the estimate due to chance factor (causes), the latter being known as experimental error or simply the error.

Total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components associated with the nature

of classification of the data. The systematic procedure for achieving this is called *Analysis of Variance*.

ANOVA is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance.

Thus, ANOVA is a statistical technique that can be used to evaluate whether there are differences between the average value, or mean, across several population groups. With this model, the response variable is continuous in nature, whereas the predictor variables are categorical. For example, in a clinical trial of hypertensive patients, ANOVA methods could be used to compare the effectiveness of three different drugs in lowering blood pressure. Alternatively, ANOVA could be used to determine whether infant birth weight is significantly different among mothers who smoked during pregnancy relative to those who did not. In the simplest case, where two population means are being compared, ANOVA is equivalent to the independent two-sample t -test.

In any ANOVA model, general mean is always taken as fixed effect and error is always taken as random effect. Thus class of model can be classified on the basis of factors, other than these two factors. ANOVA can be viewed as a generalization of t -tests: a comparison of differences of means across more than two groups.

The ANOVA is conducted under some assumptions. These assumptions are:

- Samples have been drawn from the populations that are normally distributed.
- Observations are independent and are distributed normally with mean zero and variance σ^2 .
- Effects are additive in nature.
- Populations have equal variance.
- Samples are randomly and dependently distributed $e_{ij} \sim N(0, \sigma^2)$.

The ANOVA is performed as One-way or Two-way ANOVA when the number of factors is one or two, respectively. In general if the number of factors is more than we perform multi-factor ANOVA.

25.3.1 One-Way ANOVA

A one-way analysis of variance is used when the data are divided into groups according to only one factor. The questions of interest are usually: (a) Is there a significant difference between the groups? and (b) If so, which groups are significantly different from which others?

If the observations of the response variable can be assumed to be normal distributed, a model for an observation in the sample is given by

$$X_{ij} = \mu + e_{ij}$$

where μ = fixed but unknown constant,

e_{ij} = random variable distributed as $N(0, \sigma^2)$.

We add the difference between groups (treatments) into model

$$X_{ij} = \mu + \tau_i + e_{ij}$$

where τ_i is the effect of I level of factor A.

In one-way ANOVA, the hypothesis to be tested is

$$H_0: \mu_1 = \mu_2 = \dots = \mu_v$$

All population means are equal i.e. no treatment effect (no variation in means among groups) against the alternate hypothesis

$H_A: \mu_i \neq \mu_j$ for all $i, j = 1, \dots, v$ i.e. at least one population mean is different.

If the variances in the groups (treatments) are similar, we can divide the variation of the observations into the variation of the groups (variation of the means) and the variation within the groups. The variation is measured with the sum of the squares

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ijk} - \bar{X})^2 \quad SSA = \sum_{i=1}^k (\bar{X}_i - \bar{X})^2 \quad SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

$$SST = SSA + SSE$$

Further, $MSA = \text{Mean square factor A} = \frac{SSA}{k - 1}$ and $MSE = \text{Mean square error} = \frac{SSE}{n - k}$

The idea is to measure the variation of the group means. If the variation of the means is large compared to the variation in the groups (Within), the groups differ. We must take care of the number of the observations dividing the sum of squares with the degrees of freedom.

The ANOVA Table is

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Statistic
Among Groups	SSA	$c - 1$	MSA	MSA/MSE
Within Groups	SS E	$n - 1$	MS E	
Total	SS T	$n - c$		

ANOVA helps in testing whether the variation due to any particular component is significant as compared to residual variation that can occur among the observational units. The ANOVA looks at the variance within classes relative to the overall variance. The dependent variable must be metric, and the independent variables, which can be many, must be nominal. ANOVA is used to uncover the main and interaction effects of categorical independent variables (called "factors") on an interval dependent variable.

The key statistic in ANOVA is the F-test of difference of group means, testing if the means of the groups formed by values of the independent variable (or combinations of values for multiple independent variables) are statistically significant. If the F test shows that overall the independent variable(s) is (are) related to the dependent variable, then *multiple comparison tests* of significance are used to explore just which values of the independent(s) have the most to do with the relationship.

25.3.2 Two-Way ANOVA

When we study two factors, say Factor A and Factor B, on the dependent variable we go for Two-way ANOVA. In it the model used is

$$X_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where τ_i is the effect of i level of factor A, β_j is the effect of Factor j level of Factor B and γ_{ij} is the effect of interaction of Factor A and Factor B.

In it the sum of squares is

$$\begin{aligned} SST &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{X})^2 & SSA &= cn' \sum_{i=1}^r (\bar{X}_{i..} - \bar{X})^2 \\ SSB &= rn' \sum_{j=1}^c (\bar{X}_{.j.} - \bar{X})^2 & SSAB &= n' \sum_{i=1}^r \sum_{j=1}^c (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2 \\ SSE &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n'} (X_{ijk} - \bar{X}_{ij.})^2 \end{aligned}$$

Further the mean squares can be calculated as

$$\begin{aligned} MSA &= \text{Mean square factor A} = \frac{SSA}{r-1} \\ MSB &= \text{Mean square factor B} = \frac{SSB}{c-1} \\ MSAB &= \text{Mean square interaction} = \frac{SSAB}{(r-1)(c-1)} \\ MSE &= \text{Mean square error} = \frac{SSE}{rc(n'-1)} \end{aligned}$$

The ANOVA Table is

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Statistic
Factor A	SSA	$r - 1$	$MSA = SSA / (r - 1)$	MSA/MSE
Factor B	SSB	$c - 1$	$MSB = SSB / (c - 1)$	MSB/MSE
Interaction AB	SSAB	$(r - 1)(c - 1)$	$MSAB = SSAB / (r - 1)(c - 1)$	$MSAB/MSE$
Error	SSE	$rc(n - 1)$	$MSE = SSE/rc(n' - 1)$	
Total	SST	$n - 1$		

Now we will discuss some examples of design of experiments.

Example 1: {Nigam, A.K. and Gupta V.K., 1979, *Handbook on Analysis of Agricultural experiments*, First Edition, I.A.S.R.I. Publication, New Delhi, pp16-20}.A feeding trial with 3 feeds namely (i) Pasture(control), (ii) Pasture and Concentrates and (iii) Pasture, Concentrates and Minerals was conducted at the Yellachihalli Sheep Farm, Mysore, to study their effect on wool yield of Sheep. For this purpose twenty-five ewe lambs were allotted at random to each of the three treatments and the three treatments and the weight records of the total wool yield (in gms) of first two clipping were obtained. The data for two lambs for feed 1, three for feed 2 and one for feed 3 are missing. The details of the experiment are given below:

Yield (in gms)

FEED 1	FEED 2	FEED 3
850.50	510.30	992.25
453.60	963.90	850.50
878.85	652.05	1474.20
623.70	1020.60	510.30
510.30	878.85	850.50
765.45	567.00	793.80
680.40	680.40	453.60
595.35	538.65	935.55
538.65	567.00	1190.70
850.50	510.30	481.95
850.50	425.25	623.70
793.80	567.00	878.85
1020.60	623.70	1077.30
708.75	538.65	850.50
652.05	737.10	680.40
623.70	453.60	737.10
396.90	481.95	737.10
822.15	368.55	708.75
680.40	567.00	708.75
652.05	595.35	652.05
538.65	567.00	567.00
850.50	595.35	453.60
680.40	.	652.05
.	.	567.00
.	.	.

where Feed 1- Pasture (control),
 Feed 2- Pasture and Concentrates and
 Feed 3- Pasture, Concentrates and Minerals.

1. Perform the analysis of variance of the data to test whether there is any difference between treatment effects.
2. Perform all possible pair wise treatment comparisons and identify the best treatment i.e. the treatment giving highest yield.

SAS Code for analysis

```
data crd;
input trt yld;
cards;
1      850.5
1      453.6
1      878.85
.
.
.

3      652.05
3      567
3      .
;
proc glm;
Class trt;
Model yld = trt/ss3;
Means trt;
Means trt/LSD;
lsmeans trt/pdiff;
Run;
```

Example 2: An initial varietal trial (Late Sown, irrigated) was conducted to study the performance of 20 new strains of mustard vis-a-vis four checks (Swarna Jyoti: ZC; Vardan: NC; Varuna: NC; and Kranti: NC) using a Randomized complete Block Design (RCB) design at Bhatinda with 3 replications. The seed yield in kg/ha was recorded. The details of the experiment are given below:

Yield in kg/ha

Strain	Code	Replications		
		1	2	3
RK-04-3	MCN-04-110	1539.69	1412.35	1319.73
RK-04-4	MCN-04-111	1261.85	1065.05	1111.36
RGN-124	MCN-04-112	1389.19	1516.54	1203.97

HYT-27	MCN-04-113	1192.39	1215.55	1157.66
PBR-275	MCN-04-114	1250.27	1203.97	1366.04
HUJM-03-03	MCN-04-115	1296.58	1273.43	1308.16
RGN-123	MCN-04-116	1227.12	1018.74	937.71
BIO-13-01	MCN-04-117	1273.43	1157.66	1088.20
RH-0115	MCN-04-118	1180.82	1203.97	1041.90
RH-0213	MCN-04-119	1296.58	1458.65	1250.27
NRCDR-05	MCN-04-120	1122.93	1065.05	1018.74
NRC-323-1	MCN-04-121	1250.27	926.13	1030.32
RRN-596	MCN-04-122	1180.82	1053.47	717.75
RRN-597	MCN-04-123	1146.09	1180.82	856.67
CS-234-2	MCN-04-124	1574.42	1412.35	1597.57
RM-109	MCN-04-125	914.55	972.44	659.87
BAUSM-2000	MCN-04-126	891.40	937.71	798.79
NPJ-99	MCN-04-127	1227.12	1203.97	1389.19
SWARNA JYOTI(ZC)	MCN-04-128	1389.19	1180.82	1273.43
VARDAN(NC)	MCN-04-129	1331.31	1157.66	1180.82
PR-2003-27	MCN-04-130	1250.27	1250.27	1296.58
VARUNA(NC)	MCN-04-131	717.75	740.90	578.83
PR-2003-30	MCN-04-132	1169.24	1157.66	1111.36
KRANTI(NC)	MCN-04-133	1203.97	1296.58	1250.27

Note: Strains of mustard in bold are the four checks.

1. Perform the analysis of variance of the data to test whether there is any difference between treatment effects.
2. Perform all possible pair wise treatment comparisons and identify the best treatment i.e. the treatment giving highest yield. Also identify the other treatments which are non-significantly different from this treatment.

SAS Code

```
data rbd;
Input trt tep yield;
Cards;
1 1 1539.69
2 1 1261.85
3 1 1389.19
4 1 1192.39
```

OVERVIEW OF DESIGN OF EXPERIMENTS

```

.
.
.
23 3 1111.36
24 3 1250.27
;
PROC GLM;
Class rep trt;
Model yield = rep trt;
Means trt/LSD;
Means trt/DUNCAN;
Run;

```

Example 3: An experiment was conducted at Agricultural Research Station, Kopurgaon, Maharashtra on cotton during the year 1969-1970 using a Latin Square Design to study the effects of foliar application of urea in combination with insecticidal sprays on the cotton yield. The 6 treatments were {**T₁** : Control (*i. e.* no N and no insecticides), **T₂** :100kg N/ha applied as urea (half at final thinning and half at flowering as top dressing), **T₃**: 100kg N/ha applied as urea(80 kg N/ha In 4 equal split doses as spray and 20 kg N/ha at final thinning), **T₄**:100 kg. N/ha applied as CAN (half at final thinning and half at flowering as top dressing), **T₅** : T₂ + six insecticidal sprays, **T₆** : T₄ + six insecticidal sprays}. There were 6 replication, and the data of cotton in kg per plot is:

T ₃ 3.10	T ₆ 5.95	T ₁ 1.75	T ₅ 6.40	T ₂ 3.85	T ₄ 5.30
T ₂ 4.80	T ₁ 2.70	T ₃ 3.30	T ₆ 5.95	T ₄ 3.70	T ₅ 5.40
T ₁ 3.00	T ₂ 2.95	T ₅ 6.70	T ₄ 5.95	T ₆ 7.75	T ₃ 7.10
T ₅ 6.40	T ₄ 5.80	T ₂ 3.80	T ₃ 6.55	T ₁ 4.80	T ₆ 9.40
T ₆ 5.20	T ₃ 4.85	T ₄ 6.60	T ₂ 4.60	T ₅ 7.00	T ₁ 5.00
T ₄ 4.25	T ₅ 6.65	T ₆ 9.30	T ₁ 4.95	T ₃ 9.30	T ₂ 8.40

- i) Perform the analysis of the data and identify the best treatment.
- (ii) Test whether the average effect of T₃(100kg N/ha applied as urea) and T₄ (100 kg N/ha) is same as the average effect of T₅(T₂ + six insecticidal sprays) and T₆(T₄ + six insecticidal sprays).

SAS Code

```

data latin;
input row col trt yield;
cards;
1      1      3      3.10
1      2      6      5.95
1      3      1      1.75
1      4      5      6.40
.
.
.
6      5      3      9.30
6      6      2      8.40
;
proc glm;
class row col trt;
model yield = row col trt;
means trt/tukey;
means trt/lsd;
lsmeans trt/pdiff;
contrast 'T3 T4 vs T5 T6' trt 0 0 1 1 -1 -1;
run;

```

References

- Cochran, W. G., and Cox, G. M. (1957). *Experimental Design*, 2nd edition. New York: Wiley.
- Fisher, R.A. and Yates, F. (1963). *Statistical Tables For Biological, Agricultural and Medical Research*. Longman Group Ltd., England.
- Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York.

LOGISTICS REGRESSION FOR SAMPLE SURVEYS

Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi-110012

26.1 INTRODUCTION

Researchers use sample survey methodology to obtain information about a large aggregate or population by selecting and measuring a sample from the population. Categorical outcomes such as binary, ordinal, and nominal responses occur often in survey research. *Logistic regression analysis* is often used to investigate the relationship between these discrete responses and a set of explanatory variables. Discussions of logistic regression in sample surveys include Binder (1981, 1983), Roberts, Rao, and Kumar (1987), Skinner, Holt, and Smith (1989), and Lehtonen and Pahkinen (1995). Due to the variability of characteristics among items in the population, researchers apply scientific sample designs in the sample selection process to reduce the risk of a distorted view of the population, and they make inferences about the population based on the information from the sample survey data. In order to make statistically valid inferences for the population, we must incorporate the sample design in the data analysis. That is, the fact that survey data are obtained from units selected with complex sample designs needs to be taken into account in the survey analysis: weights need to be used in analyzing survey data and variances of survey estimates need to be computed in a manner that reflects the complex sample design. This write up discusses the logistic regression model for sample survey data.

Any analysis that ignores the sample design and the weights must be based on assumptions. If the sample is designed to generate equal probability sample, then the weights for estimating means, rates, or relationships among variables may be safely ignored. Kish (1965, pp. 20-21) called these designs *epsem* designs and noted that even complex multi-stage samples can be designed to be *epsem* for sampling units at the final or near final stage of the design. As noted later, adjustments for non-response may create unequal weights even if the design was initially *epsem*. If post-stratification or multi-dimensional calibration is applied to the data through adjustments to the weights, these processes will almost always create unequal weights adjustments and, therefore, unequal weights.

Standard methods on statistical analysis assume that survey data arise from a simple random sample of the target population. Little attention is given to characteristics often associated with survey data, including missing data, unequal probabilities of observation, stratified multistage sample designs, and measurement errors. Most standard statistical procedures in software packages commonly used for data analysis do not allow the analyst to take most of these properties of survey data into account unless specialized survey procedures are used. Failure to do so can have an important impact on the results of all types of analysis, ranging from simple descriptive statistics to estimates of parameters of multivariate models. Examples of logistic regression in surveys can be found in Korn and Graubard (1999).

In the *maximum likelihood* (ML) estimation for simple random samples, we work with unweighted observations and appropriate likelihood equations can be

constructed, based on standard distributional assumptions, to obtain the maximum likelihood estimates of the model coefficients and the corresponding covariance matrix estimate. Using these estimates, standard likelihood ratio and binomial based Wald test statistics can be used for testing model adequacy and linear hypothesis on the model coefficients. Under more complex designs involving element weighting and clustering, a ML estimator of the model coefficients and the corresponding covariance matrix estimator are *not consistent* and, moreover, the standard test statistics are not asymptotically chi-square with appropriate degree of freedom. For consistent estimation of model coefficients, the standard likelihood equations are *modified* to cover the case of weighted observations. A widely used method for fitting models for binary, polytomous and continuous response variables in complex surveys is based on a *modification* of ML estimation. The traditional ML estimation of model parameters is for response variables from simple random sampling, for which appropriate likelihood functions can be derived under standard distributional assumptions. But, for more complex designs no convenient likelihood functions are available, and therefore, a method called *pseudo likelihood estimation* is used instead. The method is henceforth called the *PML* method.

The *PML* method of pseudo likelihood is often used on complex survey data for *logit* analysis in similar analysis situations to the weighted least square (WLS) method. But the applicability of the PML method is wider, covering not only models on domain proportions of a binary or polytomous response but also the usual regression type settings with continuous measurements as the predictors. In PML estimation of model coefficients and their asymptotic covariance matrix we use a modification of a ML method. In addition, a consistent covariance matrix estimator of the PML estimator is constructed such that the clustering effects are properly accounted for.

26.2 STANDARD LOGISTIC REGRESSION MODEL

Often, though, we have data where the interest is in predicting or “explaining” a binary (or dichotomous) outcome variable by a set of explanatory variables. This type of outcome variable is a two-category variable. Examples are:

- Yes/No or Agree/Disagree responses to questionnaire items in a survey.
- Success/failure of a treatment, explained by dosage of medicine administered, patient’s age, sex, weight and severity of condition.

Let Y be a dichotomous outcome variable, coded as $Y = 1$ for the outcome of interest, denoted a “success”, and $Y = 0$ for the other possible outcome, denoted a “failure”. We use the Greek character π to represent the probability that the “success” outcome occurs in the population (i.e. $Y = 1$). The probability of a “failure” outcome (i.e. $Y = 0$) is then $1 - \pi$. We also note here that the mean of Y in the population, which the statisticians call the “expected value” of Y and denote $E(Y)$, is just equal to π . So we have $\Pr(Y = 1) = E(Y) = \pi$ and $\Pr(Y = 0) = 1 - \pi$.

Suppose that we have a variable, Y , denoting cholesterol level, so that $Y = 1$ if cholesterol level is “high”, and $Y = 0$ if cholesterol level is “low”. If we have n outcomes of this form, we might be interested in modelling the variation in Y using a set of explanatory variables. Can we identify characteristics associated with the likelihood of high/low cholesterol level – which subgroups in the population are more

likely to have high cholesterol than others? We could try to fit the usual *linear regression model* of the form

$$E(Y) = \pi = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k, \tag{1}$$

where X_1, X_2, \dots, X_k are the explanatory variables. Here we are modelling the probability of a “success” (π). For example, we might want to model the probability, π , of having “high” cholesterol as a linear function of the single explanatory variable age, X , $\pi = \beta_0 + \beta_1 X$.

It is possible to fit this model (1) using the ordinary least squares (OLS) method. Indeed, such a model might produce sensible results. However,

- The predicted values of π obtained from fitting this model may be outside the $[0,1]$. Since π is a probability, its value must lie within the interval $[0,1]$. However, the right-hand side (RHS) of equation (3.1) is unbounded so that, theoretically, the RHS can take on values from $-\infty$ to ∞ . This means we could get a predicted probability of, for example, 2.13 from our fitted model, which is rather non-sensical! It turns out that if $0.25 < \pi < 0.75$ the linear probability model produces fairly sensible results though.
- The usual regression assumption of normality of Y is not satisfied, since Y is either 0 or 1.

What is the solution? Instead of fitting a model for π , we use a transformation of π . The transformation we shall use is the odds of a “success” outcome, i.e. we shall model $\pi/(1-\pi)$.

The odds are defined as the probability of a “success” divided by the probability of a “failure”

$$\text{odds} = \frac{\text{Pr}(\text{success})}{\text{Pr}(\text{failure})} = \frac{\text{Pr}(\text{success})}{1 - \text{Pr}(\text{success})} = \frac{\pi}{1 - \pi}. \tag{2}$$

It is easy to convert from probabilities to odds and back again. Notice that the odds can take values between 0 to ∞ . For examples,

1. If $\pi = 0.8$, then the odds are equal to $0.8/(1-0.8) = 0.8/0.2 = 4$.
2. If $\pi = 0.5$, then the odds are equal to $0.5/(1-0.5) = 0.5/0.5 = 1$. So, if the odds are equal to 1 the chance of “success” is the same as the chance of “failure”.
3. If the odds are equal to 0.3 then solving $\pi/(1 - \pi) = 0.3$ gives $\pi = 0.3/1.3 = 0.2308$.

So, we can think of the odds as another scale for representing probabilities. We also note here that since division by zero is not allowed, the odds will be undefined when the probability of “failure” (i.e. $1-\pi$) is 0. The logistic regression model for the odds is of the form

$$\frac{\pi}{1 - \pi} = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}. \tag{3}$$

With this model, the range of values that the right-hand side can take is now between 0 and ∞ , which is the same range as that of the left-hand side. An alternative and equivalent way of writing the logistic regression model in (3) is in terms of the log of the odds, called the logit form of the model

$$\log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k . \tag{4}$$

The logit is just another transformation of the underlying probability π . It is the (natural) logarithm of the odds. Now, we have something that looks more familiar. We have a linear model on the logit scale. This is the more common form of the logistic regression model. We can also write the model in terms of the underlying probability of a “success” outcome,

$$\pi = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}} . \tag{5}$$

Note that all three forms of the model (3) – (5) are equivalent.

26.3 Performing Logistic Regression

Our aim is to quantify the relationship between the probability of a “success” outcome, π , and the explanatory variables X_1, X_2, \dots, X_k based on some sample data. For now, we assume that in the population there is a relationship between π and a single continuous explanatory variable X and that this relationship is of the form

$$\text{logit}(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 X . \tag{6}$$

This model can be estimated using SPSS (or practically any other general-purpose statistical software) as

$$\text{logit}(\hat{\pi}) = b_0 + b_1 X , \tag{7}$$

where b_0 and b_1 are the estimated regression coefficients. The estimation for logistic regression is commonly performed using the statistical method of maximum likelihood estimation. The steps to run the *logistic regression* procedure in SPSS are:

- [1] First call up the logistic regression dialogue box by selecting Analyze | Regression | Binary Logistic, and then transfer the y (response variable) variable to the Dependent box and the x (covariates) variable to the Covariates box.
- [2] Click on the Options button, check the options as per requirement.
- [3] Click on OK to run the logistic regression.

Example - Suppose that we are interested in whether or not the gestational age (GAGE) of the human foetus (number of weeks from conception to birth) is related to the birth weight. The dependent variable is birth weight (BWGHT), which is coded as 1 = normal, 0 = low. The data for 24 babies of whom were classified as having low weight at birth) are shown in Table 1.

Table 1: Gestational Ages (in Weeks) of 24 Babies by Birth Weight

Normal Birth Weight (BWGHT = 1)	Low Birth Weight (BWGHT = 0)
40, 40, 37, 41, 40, 38, 40, 40	38, 35, 36, 37, 36, 38, 37
38, 40, 40, 42, 39, 40, 36, 38, 39	

The model we shall fit is:

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{GAGE} . \tag{8}$$

The output given by SPSS from fitting this model is shown in Figure 2. Note that under Block 0, SPSS produces some output which corresponds to fitting the “constant

model” – this is the model $\text{logit}(\pi) = \beta_0$. Under Block 1, SPSS gives the estimated regression coefficients (see final table in the output labelled “Variables in the Equation”)

$b_0 = -48.908$, $b_1 = 1.313$, and so our fitted model is

$$\text{logit}(\hat{\pi}) = -48.908 + 1.313 \text{ GAGE.} \quad (9)$$

Two obvious questions present themselves at this stage: (1) How do we know if this model fits the data well? and (2) How do we interpret this fitted model?

We shall consider several approaches to assess the “fit” of the model. Note that in practice, reporting two or three of these is normally sufficient. **SPSS** produces a **classification table**. This is a simple tool which is sometimes used to assess how well the model fits. Consider the model (8) for BWGHT explained by GAGE. First, we choose a “cut-off” value c (usually 0.5). For each individual in the sample we “predict” their BWGHT condition as 1 (i.e. normal) if their fitted probability of being normal birth weight is greater than c , otherwise we predict it as 0 (i.e. low). SPSS gives a table showing how many of the observations we have predicted correctly.

In this example, we have 24 cases altogether. Of these, 7 were observed as having low birthweight (BWGHT = 0) and 5 of these 7 we correctly predict, i.e. they have a fitted probability of less than 0.5. Similarly, 15 out of the 17 observed as having normal birthweight are correctly predicted. Generally, the higher the overall percentage of correct predictions (in this case $20/24 = 83\%$) the better the model. However, there is no formal test to decide whether a certain percentage of correct predictions is adequate. Also, it is easy to construct a situation where the logistic regression model is in fact the correct model and therefore will fit, but the classification will be poor.

The Likelihood Ratio Test

We can formally test to see whether a variable is significant in explaining some of the variability in the response. Suppose we have to evaluate two models. For example,

$$\text{Model 1:} \quad \text{logit}(\pi) = \beta_0 + \beta_1 X_1$$

$$\text{Model 2:} \quad \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Here, Model 1 is said to be “nested” within Model 2 – all the explanatory variables in Model 1 (X_1) are included in Model 2. We are interested in whether the additional explanatory variable in Model 2 (X_2) is required, i.e. does the simpler model (Model 1) fit the data just as well as the fuller model (Model 2). In other words, the null hypothesis is that $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 \neq 0$.

The likelihood ratio test is based on a statistic which SPSS produces for each model it fits. This statistic is $-2 \text{ Log Likelihood}$ (also called the scaled deviance) and it measures the degree of discrepancy between the observed and fitted values. A model that fits the data exactly would have this value equal to zero. The value of this statistic for the “smaller” model (Model 1) will be larger than the value for the “larger” model (Model 2). The Likelihood Ratio (LR) test statistic is the difference in the value of the $-2 \text{ Log Likelihood}$ statistic between Model 1 and Model 2. For the example of gestational age and birth weight, the following table gives the values for $-2 \text{ Log Likelihood}$:

Model	-2 Log Likelihood
Model 1 (constant model)	28.975
Model 2 (with GAGE)	16.298

SPSS gives the value of -2 Log Likelihood for the model that includes GAGE in the “Model Summary” table. The LR test statistic is $28.975 - 16.298 = 12.676$, and SPSS gives this value in the “Omnibus Tests of Model Coefficients” table.

The statistical theory underlying the use of this statistic is beyond the scope of this course. We just note that if the null hypothesis (that the coefficient of GAGE is 0) is true, the LR test statistic should have a chi-squared distribution with one degree of freedom, as long as the sample size is “large”. So, the procedure is to observe the value of the LR test statistic and compare with the table value from a chi-squared distribution with one degree of freedom. If the LR test statistic is too large relative to the table value, then this will imply that the null hypothesis should be rejected, i.e. the simpler model does not fit the data as well as the fuller model. At the 5% level of significance, the cut-off value from the table is 3.84, which indicates that in the example we should reject the null hypothesis in favour of the alternative, i.e. GAGE is significant. [Again, SPSS just provides a p-value (in the “sig” column) – if this is less than 0.05 then we reject the null hypothesis at the 5% level.] That is, the model with gestational age is better than the model with just the constant term.

As already mentioned above, estimation of the coefficients (i.e. the β 's) in logistic regression is performed in SPSS using the method of maximum likelihood estimation (MLE). The standard errors are also computed by SPSS, and their estimation also relies on MLE theory.

Consider our model in the example of gestational age and birth weight,

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{ GAGE} \tag{10}$$

Another way of checking if the variable GAGE should be in the model or not is to calculate the ratio of the estimate to its standard error (this is a bit like the t statistic in linear regression)

$$b_1 / s_{b_1} \tag{11}$$

If the null hypothesis that $\beta_1 = 0$ is true, then this statistic has an approximate standard normal distribution. So, we can compare this to values in the normal tables for a given level of significance. Equivalently, we can calculate the Wald statistic,

which is the square of this ratio, i.e. $\left(\frac{b_1}{s_{b_1}}\right)^2$. If the null hypothesis that $\beta_1 = 0$ is true,

then this statistic has a *chi-squared* distribution with one degree of freedom. This is what SPSS calculates and displays in the “Variables in the Equation” table, along with an associated p-value. In the example of gestational age and birth weight, the value of the Wald test statistic for the coefficient corresponding to the variable GAGE is 5.890 ($= [1.313/0.541]^2$). The p-value is given as 0.015, indicating that the coefficient is significant at the 5% level (but not at the 1% level).

Wald’s test or the Likelihood Ratio Test – do they lead to the same conclusion?

The simple answer is not always! In most cases, both tests would lead you to the same decision. However, in some cases the Wald test produces a test statistic that is

insignificant when the Likelihood Ratio test indicates that the variable should be kept in the model. This is because sometimes the estimated standard errors are “too large” (this happens when the absolute value of the coefficient becomes large) so that the ratio becomes small. The Likelihood Ratio test is the more robust of the two.

26.4 Interpreting the Model

The logistic regression model can be written in three different scales (logit, odds, or probability). It can therefore be interpreted in these different scales. Consider the example of gestational age and birth weight. The interpretation of the coefficient for GAGE in the model

$$\text{logit}(\hat{\pi}) = -48.908 + 1.313 \text{ GAGE}, \quad (12)$$

is that a unit change in GAGE increases the log odds of normal birth weight by 1.313, on average, i.e. a one-week increase in gestational age increases the log odds of normal birth weight by 1.313, on average. The model (12) can also be written as the odds model by taking the exponent of both sides, i.e.

$$\text{odds} = \frac{\hat{\pi}}{1 - \hat{\pi}} = e^{-48.908 + 1.313 \text{ GAGE}} = e^{-48.908} e^{1.313 \text{ GAGE}}. \quad (13)$$

Thus, a one-week increase in gestational age changes the odds of normal birth weight multiplicatively by a factor equal to $e^{1.313}$, i.e. by 3.716. This factor is called an odds ratio (more about these later), and is computed for us by SPSS and displayed in the final column (labelled Exp(B)) of the “Variables in the Equation” table. Equivalently, we may say that a unit increase in GAGE increases the odds of normal birth weight by $[3.716 - 1] \times 100\%$, i.e. 272%. Finally, the model (12) can also be written in the probability scale, i.e.

$$\hat{\pi} = \frac{e^{-48.908 + 1.313 \text{ GAGE}}}{1 + e^{-48.908 + 1.313 \text{ GAGE}}}. \quad (14)$$

Thus, we can estimate the probability of normal birth weight for any given gestational age. For the general fitted model equation

$$\text{logit}(\hat{\pi}) = b_0 + b_1 X, \quad (15)$$

the value of X when $\hat{\pi} = 0.5$ is called the median effective level, i.e. the outcome of interest has a 50% chance of occurring. This happens when the odds = 1, i.e. when the logit (log odds) = 0, and this occurs when $X = -b_0/b_1$. For the birth weight data, this is when gestational age = $48.908/1.313 = 37.25$ weeks. At this age, the baby has a 50% chance of being of normal birth weight.

What about interpretation of the constant term?

Some statisticians interpret the exponent of the constant term as the baseline odds. In our example this is the value of the odds when $\text{GAGE} = 0$. This interpretation is not really applicable here and I would think of the constant as simply a nuisance parameter to be kept in the model so that the odds are of the correct scale.

Note: When there are two or more explanatory variables, say X_1 and X_2 , the interpretation of the coefficient β_1 as the change in the log odds when X_1 changes by one unit is correct only if X_1 and X_2 are unrelated (i.e. when X_1 changes, X_2 is unaffected).

26.5 COMPLEX SURVEY DATA – PROBLEMS

- Standard method discussed above assumes that data arise from a SRS.
- No attention to survey design used to collect the data.
- *Maximum likelihood* (ML) estimation- unweighted observations and likelihood equations was constructed, based on standard distributional assumptions, to obtain the ML estimates of the model coefficients and the corresponding covariance matrix estimate.
- Complex survey involving element weighting, stratification and clustering etc, a ML estimator of the model coefficients and the corresponding covariance matrix estimator are *not consistent*.
- The standard test statistics are not asymptotically chi-square with appropriate degree of freedom.

Solution to the problem

- For consistent estimation, the standard likelihood equations are *modified* to cover the case of weighted observations - a *modification* of ML estimation.
- For complex design: no convenient likelihood functions are available, and therefore, a method called pseudo likelihood estimation (PML) is used
- In PML estimation we use a modification of a ML method
- The PML estimator: the clustering effects etc are properly accounted for.

26.6 IMPLEMENTING LOGISTIC REGRESSION FOR COMPLEX SURVEYS DATA

SAS procedures for analyzing survey data

- **SURVEYSELECT** procedure selects probability samples using various sample designs, including stratified sampling and sampling with probability proportional to size.
- **SURVEYMEANS** procedure computes descriptive statistics for sample survey data, including means, totals, ratios, and domain statistics.
- **PROC SURVEYLOGISTIC** -This procedure performs a logistic regression taking into account the survey design variables.

STATA COMMANDS

- **logistic** for fitting a logistic regression in Stata, followed by the variable list, the first of which must be the dependent variable.
- **svylogit** command performs logistic regression model that consider the impact of survey weights
- **STATA** allows for three aspects of a sample design: stratification, clustering, and weighting. These are defined by strata, psu, and pweight in Stata.
- **svyset** command to “set” the sample design.
- **svyset [pweight=weight], strata(name) psu(name)**
- **svydes** command provides summary information for the survey design as specified.

SPSS syntax

- **CSLOGISTIC** performs logistic regression analysis for samples that are drawn by **complex sampling methods**.

- The procedure estimates variances by taking into account the sample design that is used to select the sample
- The basic specification is a variable list and a PLAN subcommand with the name of a complex sample analysis plan file, which may be generated by the **CSPLAN** procedure.
- The default model includes the intercept term, main effects for any factors, and any covariates.
- **CSLOGISTIC** performs logistic regression analysis for sampling designs that are supported by the CSPLAN and CSSELECT procedures.

REFERENCES

- Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.
- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.
- Binder, D. A. (1981), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *Survey Methodology*, 7, 157–170.
- Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279–292.
- Chambers, R.L. and C.J. Skinner, 2003. *Analysis of Survey Data*. Wiley, Chichester, UK.
- Dobson, A. J. (1990). *An Introduction to Generalized Linear Models*. London: Chapman and Hall.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied Regression Analysis*. New York: Wiley.
- Kish, Leslie, 1965. *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Korn, E. and Graubard B. (1999), *Analysis of Health Survey*, New York: John Wiley & Sons, Inc.
- Lehtonen, R. and Pahkinen E. (1995), *Practical Methods for Design and Analysis of Complex Surveys*, Chichester: John Wiley & Sons, Inc.
- McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman Hall.
- Morel, G. (1989) "Logistic Regression under Complex Survey Designs," *Survey Methodology*, 15, 203–223.
- Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein and J. Rasbash, 1998. Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society B*, 60, 23-40.
- Roberts, G., Rao, J. N. K., and Kumar, S. (1987), "Logistic Regression Analysis of Sample Survey Data," *Biometrika*, 74, 1–12.
- Skinner, C. J., Holt, D., and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, New York: John Wiley & Sons, Inc.

Figure 1: SPSS Output from Logistic Regression (Gestational Age Data)

Case Processing Summary

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	24	100.0
	Missing Cases	0	.0
	Total	24	100.0
Unselected Cases		0	.0
Total		24	100.0

a. If weight is in effect, see classification table for the total number of cases.

Block 0: Beginning Block

Classification Table^{a,b}

Observed			Predicted		Percentage Correct
			BWGHT		
			.0000	1.0000	
Step 0	BWGHT	.0000	0	7	.0
		1.0000	0	17	100.0
Overall Percentage					70.8

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	.887	.449	3.904	1	.048	2.429

Variables not in the Equation

	Score	df	Sig.
Step 0 Variables GAGE	10.427	1	.001
Overall Statistics	10.427	1	.001

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step 1 Step	12.676	1	.000
Block	12.676	1	.000
Model	12.676	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	16.298	.410	.585

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.626	5	.898

Contingency Table for Hosmer and Lemeshow Test

		BWGHT = .0000		BWGHT = 1.0000		Total
		Observed	Expected	Observed	Expected	
Step 1	1	1	.951	0	.049	1
	2	2	2.518	1	.482	3
	3	2	1.753	1	1.247	3
	4	2	1.372	3	3.628	5
	5	0	.185	2	1.815	2
	6	0	.213	8	7.787	8
	7	0	.009	2	1.991	2

Classification Table^a

Observed		Predicted			
		BWGHT		Percentage Correct	
		.0000	1.0000		
Step 1	BWGHT	.0000	5	2	71.4
		1.0000	2	15	88.2
Overall Percentage					83.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	GAGE	1.313	.541	5.890	1	.015	3.716
	Constant	-48.908	20.338	5.783	1	.016	.000

a. Variable(s) entered on step 1: GAGE.

SMALL AREA ESTIMATION TECHNIQUE FOR DISTRICT LEVEL CROP YIELD ESTIMATION

Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi, India

27.1 INTRODUCTION

Crop area and crop production forms the backbone of any agricultural statistics system. In India, crop area figures are, by and large, compiled on the basis of complete enumeration while the crop yield is estimated on the basis of sample survey approach. The yield rate estimates are developed on the basis of scientifically designed crop cutting experiments (CCEs) conducted under the scheme of General Crop Estimation Surveys (GCES). A crop cutting experiment consists of randomly locating a field growing a specific crop, location and marking, as per specified instructions, a plot of given size and shape in the selected field, harvesting, threshing and winnowing the produce within the plot and weighing the grains obtained. Since the grain on the harvested day contains moisture, it is stored and reweighted after drying to determine the marketable form of produce. The GCES covers 68 crops (52 food and 16 non-food) in 25 States and 4 Union Territories. More than 500,000 CCEs are conducted annually for this purpose. This much sample size is sufficient to provide precise estimates of crop yield (i.e., production per hectare of land) at the district level. Although the CCE technique is an objective method of assessment of crop yield, the procedure of conduct of CCE is tedious and time consuming. Due to this and some other factors, a tendency has been seen that the enumerators do not follow the prescribed procedure for the conduct of CCE in a number of cases. As a result of this, the data quality under the GCES is observed to be below the desirable limit. To improve the quality of data collected under the GCES, a scheme titled 'Improvement of Crop Statistics (ICS)' has been introduced by the Directorate of Economics and Statistics, Ministry of Agriculture, Government of India and implemented by the National Sample Survey Office (NSSO) and the State Agricultural Statistics Authority (SASA) jointly. Under this scheme, quality check on the field operation of GCES is carried out by supervising around 30,000 CCE by NSSO and State Government supervisory officers. The findings of the ICS results reveal that the crop cutting experiments are generally not carried out properly resulting in data which lacks desired quality.

In view of limitation of infrastructure and constraints of resources, there is a felt need to reduce the sample size under GCES drastically so that volume of work of the enumerator is reduced and also better supervision of the operation of CCE becomes possible leading to improvement in data quality. However, reduction in sample size will have a direct bearing on the standard error of the estimator. The reduced sample size is more alarming when used for producing estimates at district level since estimators based on the sample data from any particular district can be unstable. This small sample size problem can be easily resolved provided auxiliary information is available to strengthen the limited sample data from the district. The underlying theory is referred to as the small area estimation (SAE). The SAE techniques aim at producing reliable estimates for such districts/areas with small (or even no) sample sizes by borrowing strength from data of other areas. The SAE techniques are

generally based on model-based methods, see for example, Pfeffermann (2002) and Rao (2003). The idea is to use statistical models to link the variable of interest with auxiliary information, e.g. Census and Administrative data, for the small areas to define model-based estimators for these areas. Such small area models can be classified into two broad types:

- (i) Area level random effect models, which are used when auxiliary information is available only at area level. They relate small area direct estimates to area-specific covariates (Fay and Herriot, 1979) and
- (ii) Unit level random effect models, proposed originally by Battese, Harter and Fuller (1988). These models relate the unit values of a study variable to unit-specific covariates.

In this article we explore an application of SAE techniques to derive model-based estimates of average yield for paddy crop at small area levels in the State of Uttar Pradesh in India by linking data generated under ICS scheme by NSSO (data collected with much reduced sample size, however, the quality of data is very high) and the Population Census 2001. Small areas are defined as the districts of State of Uttar Pradesh in India. It is noteworthy that we adopt the area level model since covariates for our study are available only at the area level. The paper illustrates how the ICS data and Census data can be combined to derive reliable district level estimates of crop yield. The rest of the paper is organised as follows. Section 2 introduces the data used for the analysis and Section 3 describes the methodology applied for the analysis. In Section 4 we present the diagnostic procedures for examining the model assumptions and validating the small area estimates and discuss the results. Section 5 finally sets out the main conclusions.

27.2 DATA DESCRIPTION

In this study we use data pertaining to supervised CCE on paddy crop under ICS scheme for kharif season for the State of Uttar Pradesh in India collected during the year 2009-10. The variable of interest for which small area estimates are required is yield for paddy crop. We are interested in estimating the average yield at the district level. In the State of Uttar Pradesh there are 70 districts however supervision, on a sub-sample, of crop cutting experiments work under ICS scheme is carried out in 58 districts only and there is no sample data for the remaining 12 districts. In what follows, we refer these 12 districts as the out of sample districts. These 70 (58 in sample and 12 out of sample) districts are the small areas for which we are interested in producing the estimates. The area specific sample sizes for these 58 sample districts range from minimum of 4 to maximum of 28 CCE with average of 11 (see Figure 1). A total of 655 CCE were supervised for recording yield data in the State of Uttar Pradesh for paddy crop for the year 2009-10. We see that in a few districts the sample size is small so the traditional sample survey estimation approaches lead to unstable estimate. In addition, in 12 districts due to non availability of sample under ICS, we can not estimate paddy yield. Indeed, there is no design based solution to provide estimates for these 12 out of sample districts (Pfeffermann, 2002). The SAE is an obvious choice for such cases. The covariates (auxiliary variables) known for the population are drawn from the Population Census 2001. Note that use of covariates from the 2001 Population Census to model yield data of paddy crop from the 2009-10 ICS scheme data may raise issues of comparability. However, the covariates used in this study are not expected to change significantly

over a short period of time. There were 121 covariates available from these sources to consider for modeling. However, we did some exploratory data analysis, for example, first we segregated group of covariates with significant correlation with target variable and subsequently we implemented step wise regression analysis. Finally we choose model with two significant variables, average household size (HH_SIZE) and female population of marginal household (MARG_HH_F) with 26 per cent R^2 . The residual diagnostic plots in Figure 2 indicate that fitted model is reasonable. For SAE analysis we therefore used these two covariates. Note that for SAE of 12 out of sampled districts we used the same two covariates since we assume that the underlying model for sample areas also holds for out of sample districts.

27.3 SMALL AREA ESTIMATION METHODOLOGY

In this Section we describe the underlining theory of SAE used in the paper. In particular, we elaborate SAE based on the area level model (Fay and Herriot, 1979). It was proposed to estimate the per-capita income of small places with population size less than 1000. This model relates small area direct survey estimates to area-specific covariates. The SAE under this model is one of the most popular methods used by private and public agencies because of its flexibility in combining different sources of information and explaining different sources of errors. To start with, we first fix our notation. Throughout, we use a subscript d to index the quantities belonging to small area or district d ($d = 1, \dots, D$), where D is the number of small areas (or districts) in the population. Let $\hat{\theta}_d$ denotes the direct survey estimate of unobservable population value θ_d for area d ($d = 1, \dots, D$). Let \mathbf{x}_d be the p -vector of known auxiliary variable, often obtained from various administrative and census records, related to the population mean θ_d . The simple area specific two stage model suggested by Fay and Herriot (1979) has the form

$$\hat{\theta}_d = \theta_d + e_d \text{ and } \theta_d = \mathbf{x}_d^T \boldsymbol{\beta} + u_d, d = 1, \dots, D. \quad (1)$$

We can express model (1) as an area level linear mixed model given by

$$\hat{\theta}_d = \mathbf{x}_d^T \boldsymbol{\beta} + u_d + e_d; d = 1, \dots, D. \quad (2)$$

Here $\boldsymbol{\beta}$ is a p -vector of unknown fixed effect parameters, u_d 's are independent and identically distributed normal random errors with $E(u_d) = 0$ and $Var(u_d) = \sigma_u^2$, and e_d 's are independent sampling errors normally distributed with $E(e_d | \theta_d) = 0$, $Var(e_d | \theta_d) = \sigma_d^2$. The two errors are independent of each other within and across areas. Usually, σ_d^2 is known while σ_u^2 is unknown and it has to be estimated from the data. Methods of estimating σ_u^2 include maximum likelihood (ML) and restricted maximum likelihood (REML) under normality, the method of fitting constants without normality assumption, See Rao (2003, Chapter 5). Let $\hat{\sigma}_u^2$ denotes estimate of σ_u^2 . Then under model (2), the Empirical Best Linear Unbiased Predictor (EBLUP) of θ_d is given by

$$\hat{\theta}_d^{EBLUP} = \mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\gamma}_q (\hat{\theta}_d - \mathbf{x}_d^T \hat{\boldsymbol{\beta}}) = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) \mathbf{x}_d^T \hat{\boldsymbol{\beta}} \quad (3)$$

where $\hat{\gamma}_d = \hat{\sigma}_u^2 / (\sigma_d^2 + \hat{\sigma}_u^2)$ and $\hat{\boldsymbol{\beta}}$ is the generalized least square estimate of $\boldsymbol{\beta}$. It may be noted that $\hat{\theta}_d^{EBLUP}$ is a linear combination of direct estimate $\hat{\theta}_d$ and the model based regression synthetic estimate $\mathbf{x}_d^T \hat{\boldsymbol{\beta}}$, with weight $\hat{\gamma}_d$. Here $\hat{\gamma}_d$ is called ‘shrinkage factor’ since it ‘shrinks’ the direct estimator, $\hat{\theta}_d$ towards the synthetic estimator, $\mathbf{x}_d^T \hat{\boldsymbol{\beta}}$. For out of sample areas (i.e. areas with $n_d = 0$), the EBLUP predictor (3) leads to synthetic predictor of the form $\hat{\theta}_d^{SYN} = \mathbf{x}_d^T \hat{\boldsymbol{\beta}}$.

Prasad and Rao (1990) proposed an approximately model unbiased (i.e. with bias of order $o(1/D)$) estimate of mean squared error (MSE) of the EBLUP (3) given by

$$\overline{MSE}(\hat{\theta}_d^{EBLUP}) = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2) \hat{V}ar(\hat{\sigma}_u^2), \quad (4)$$

where

$$\begin{aligned} g_{1d}(\hat{\sigma}_u^2) &= \hat{\gamma}_d \sigma_d^2, \\ g_{2d}(\hat{\sigma}_u^2) &= (1 - \hat{\gamma}_d)^2 \mathbf{x}_d^T \hat{V}ar(\hat{\boldsymbol{\beta}}) \mathbf{x}_d, \text{ and} \\ g_{3d}(\hat{\sigma}_u^2) &= \left\{ \sigma_d^4 / (\sigma_d^2 + \hat{\sigma}_u^2)^3 \right\} \hat{V}ar(\hat{\sigma}_u^2) \end{aligned}$$

with $\hat{V}ar(\hat{\sigma}_u^2) \approx 2D^{-2} \sum_{d=1}^D (\sigma_d^2 + \hat{\sigma}_u^2)^2$ when estimating $\hat{\sigma}_u^2$ by method of fitting constants. See Rao (2003, Chapter 5) for details about various theoretical developments. Under model (2), the MSE estimate for the synthetic predictor $\hat{\theta}_d^{SYN}$ is given by $\overline{MSE}(\hat{\theta}_d^{SYN}) = \mathbf{x}_d^T \overline{V}ar(\hat{\boldsymbol{\beta}}) \mathbf{x}_d + \hat{\sigma}_u^2$.

27.4 EMPIRICAL RESULTS

This Section presents the results from data and theory described in previous Sections. We carry out some diagnostics to examine the reliability of small area estimates. We used the bias diagnostics and coefficient of variation to validate the reliability of the model-based small area estimates. We also computed the 95 percent confidence (CI) intervals for both direct and model-based estimates. The bias diagnostics is used to investigate if the model-based estimates are less extreme when compared to the direct survey estimates. In addition, if direct estimates are unbiased, their regression on the true values should be linear and correspond to the identity line. If model-based estimates are close to the true values the regression of the direct estimates on the model-based estimates should be similar (Ambler *et al.*, 2001 and Chandra *et al.*, 2011). We plot direct estimates on Y-axis and model-based estimates on X-axis and look for divergence of regression line from $Y = X$ and test for intercept = 0 and slope = 1. The bias scatter plots of the direct estimates against the model-based estimates are given in Figure 3. From the bias diagnostic we find that the intercept fails this diagnostic (i.e., intercept is different from the zero). The plots show that the model-based estimates are less extreme when compared to the direct estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average. We computed the coefficient of variation (CV) to assess the improved precision of the model-based estimates compared to the direct estimates. The CVs show the sampling variability as a percentage of the estimate. Although, there are no internationally acceptable tables for judging what CV is too high,

estimates with large CVs are considered unreliable. Figure 4 shows the CVs for the direct survey estimates and model-based. The figure shows that the estimated CVs for the model-based estimates have a higher degree of reliability when compared to the direct survey estimates. Table 1 presents the district-wise model-based estimates, 95 percent confidence interval (CI) limits and percentage coefficient of variation for paddy crop yield for all 70 (i.e. both for 58 sample and 12 out of sample) districts. In right hand side part of Table 1, results for last 12 districts correspond to out of sample districts. The CV results in Table 1 reveal that average CV of these out of sample districts is 20.10 per cent. Figure 5 shows the 95% CI of the model-based and the direct survey estimates. It is apparent that the standard errors of the direct estimates are large and therefore the estimates are unreliable.

27.5 CONCLUSIONS

This paper illustrates that the small area estimation technique can be satisfactorily applied to produce reliable district level estimates of crop yield using CCE supervised under ICS scheme. Although the ICS supervised crop cutting experiments number only 30,000 in the entire country i.e. the sample size is very low, the collected data is of very high quality. The estimates generated using this data are expected to be relatively free from various sources of non-sampling errors. Further small area estimation technique provides estimates for those districts where there is no sample information under ICS and so direct estimates can not be computed. It is, therefore, recommended that wherever it is not possible to conduct adequate number of crop cutting experiments due to constraints of cost or infrastructure or both, small area estimation technique can be gainfully used to generate reliable estimates of crop yield based on a smaller sample.

REFERENCES

- Ambler, R., Caplan, D., Chambers, R. Kovacevic, M. and S. Wang (2001). Combining Unemployment Benefits Data and LFS Data to Estimate ILO Unemployment for Small Areas: An Application of a Modified Fay-Herriot Method. *Proceedings of the International Association of Survey Statistician*, Meeting of the ISI, Seoul, August 2001.
- Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistics Association*, **83**, 28-36.
- Census of India (2001). *Registrar General and Census Commissioner*, New Delhi, India.
- Chandra, H., Salvati, N. and Sud, U.C. (2011). Disaggregate-level Estimates of Indebtedness in the State of Uttar Pradesh in India-An Application of Small Area Estimation Technique. *Journal of Applied Statistics*, Forthcoming issue.
- Fay, R. E. and Herriot, R. A. (1979). Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistics Association*, **74**, 269-277.
- Pfeffermann, D. (2002). Small Area Estimation: New Developments and Directions. *International Statistical Review*, **70**, 125-143.
- Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, New York.

Table 1. Districts wise values of model-based estimate, 95 percent confidence interval limits and coefficient of variation (CV) for paddy crop yield.

Districts	Estimate	Lower	Upper	CV, %	Districts	Estimate	Lower	Upper	CV, %
Saharanpur	17759	13667	21851	11.52	Ambedkar Nagar	16667	13652	19681	9.04
Muzaffarnagar	17208	11735	22681	15.90	Sultanpur	16793	13899	19688	8.62
Bijnor	18927	16306	21547	6.92	Bahraich	14735	13606	15865	3.83
Moradabad	16781	12329	21232	13.26	Shrawasti	15168	10783	19553	14.46
Rampur	17174	16148	18200	2.99	Balrampur	12338	9206	15470	12.69
Jyotiba Phule Nagar	11622	8894	14351	11.74	Gonda	16708	14611	18805	6.28
Ghaziabad	16726	11101	22351	16.82	Siddharthnagar	12921	9808	16033	12.05
Bulandshahar	18116	14555	21677	9.83	Basti	14165	10331	17999	13.53
Aligarh	14278	10277	18280	14.01	Sant Kabir Nagar	13273	11626	14920	6.20
Mathura	12688	8322	17054	17.20	Mahrajganj	18640	14465	22815	11.20
Etah	12508	10274	14742	8.93	Gorakhpur	12437	9608	15266	11.37
Mainpuri	13711	9065	18357	16.94	Kushinagar	16699	12301	21096	13.17
Budaun	13307	9961	16652	12.57	Deoria	8866	6143	11588	15.35
Bareilly	14140	10976	17305	11.19	Azamgarh	12033	10073	13993	8.14
Pilibhit	14687	10207	19166	15.25	Mau	10489	7090	13888	16.20
Shahjahanpur	18411	16184	20638	6.05	Ballia	7763	5056	10470	17.44
Kheri	15079	12023	18135	10.13	Jaunpur	16418	13286	19549	9.54
Sitapur	16422	12836	20007	10.92	Ghazipur	11279	8606	13953	11.85
Hardoi	19315	16665	21965	6.86	Chandauli	12229	8333	16125	15.93
Unnao	14005	11188	16821	10.05	Varanasi	17063	12659	21468	12.91
Lucknow	18242	13196	23289	13.83	Sant Ravidas Nagar	7133	2939	11327	29.40
Rae Bareli	19287	16128	22446	8.19	Mirzapur	15052	11815	18290	10.76
Farrukhabad	10446	7420	13471	14.48	Sonbhadra	16328	11079	21578	16.08
Kannauj	30450	27119	33782	5.47	Meerut [#]	14984	8898	21069	20.31
Etawah	15431	13899	16964	4.97	Baghpat [#]	12442	6182	18702	25.16
Auraiya	21021	17121	24922	9.28	Gautam Buddha Nr [#]	16704	10436	22973	18.76
Kanpur Dehat	19547	15717	23378	9.80	Hathras [#]	15258	9158	21357	19.99
Kanpur Nr	16315	12090	20539	12.95	Agra [#]	14803	8716	20890	20.56
Banda	13375	8039	18711	19.95	Firozabad [#]	14391	8289	20492	21.20
Fatehpur	15881	11406	20355	14.09	Jalaun [#]	15186	9048	21325	20.21
Pratapgarh	16437	12543	20331	11.84	Jhansi [#]	17378	11209	23547	17.75
Kaushambi	16624	11363	21884	15.82	Lalitpur [#]	16928	10684	23172	18.44
Allahabad	20218	16164	24272	10.03	Hamirpur [#]	16520	10273	22767	18.91
Barabanki	18756	15176	22336	9.54	Mahoba [#]	16285	10030	22540	19.21
Faizabad	16556	12690	20422	11.68	Chitrakoot [#]	14948	8773	21122	20.65

[#] Districts with no sample information under ICS, Nr denotes Nagar

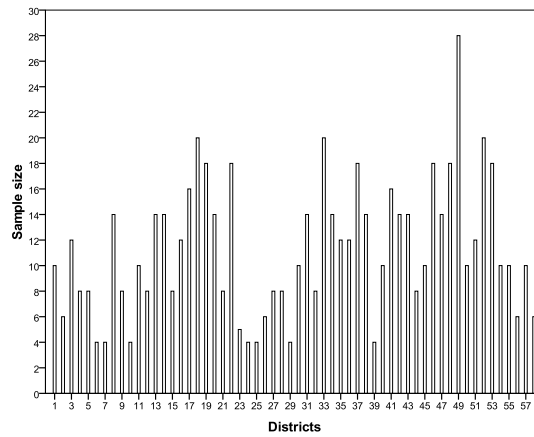


Figure 1. Distribution of district-specific sample sizes in sample districts.

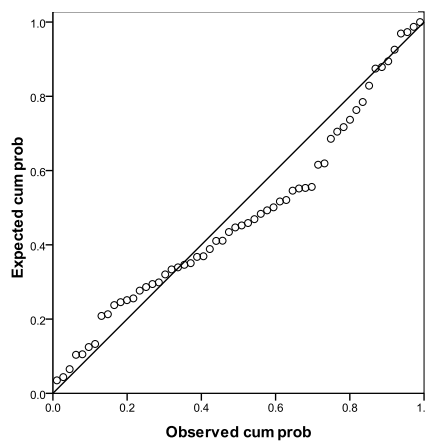
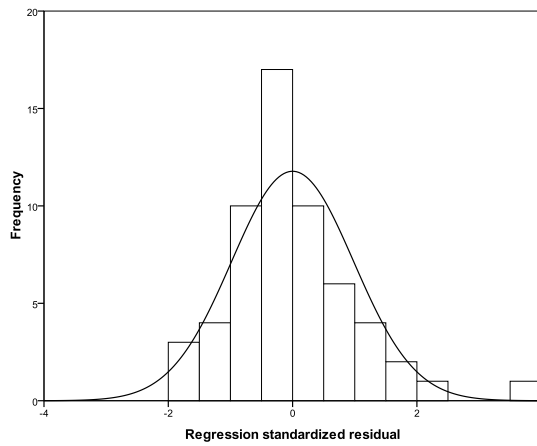


Figure 2. Histogram and normal P-P plot of regression standardized residual.

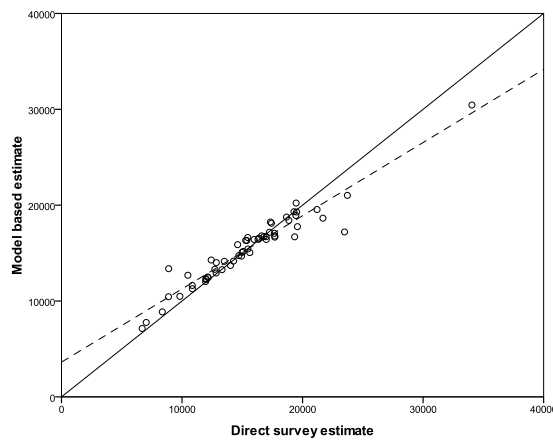


Figure 3. Bias diagnostic plots for sample districts. Direct estimates versus model based estimates, $y=x$ line (Solid) and linear regression fit line (dash).

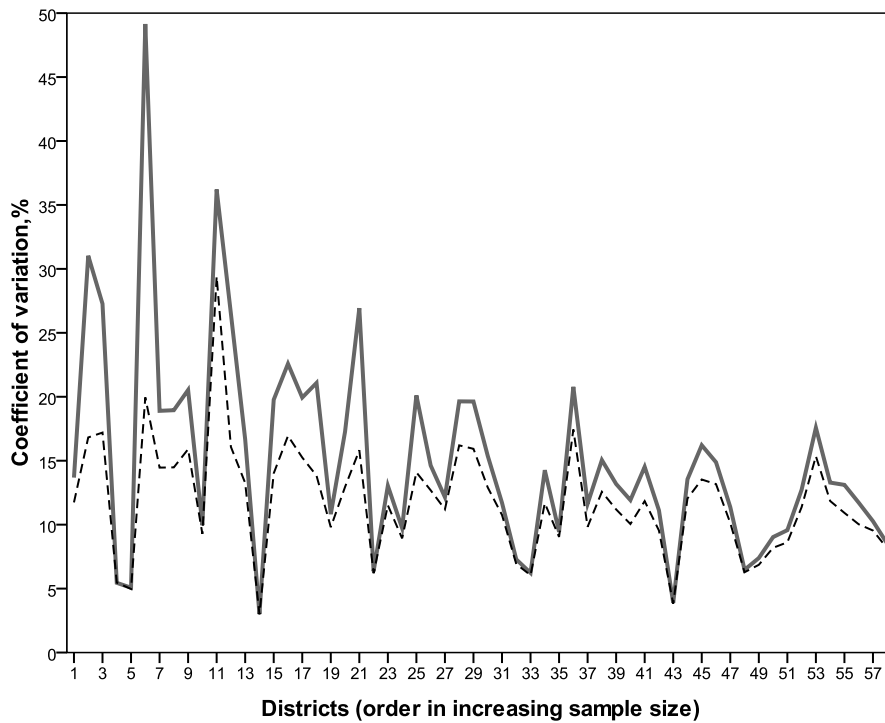


Figure 4. Coefficient of variations of direct estimates (solid line) and model based estimates (dash line) for sampled districts.

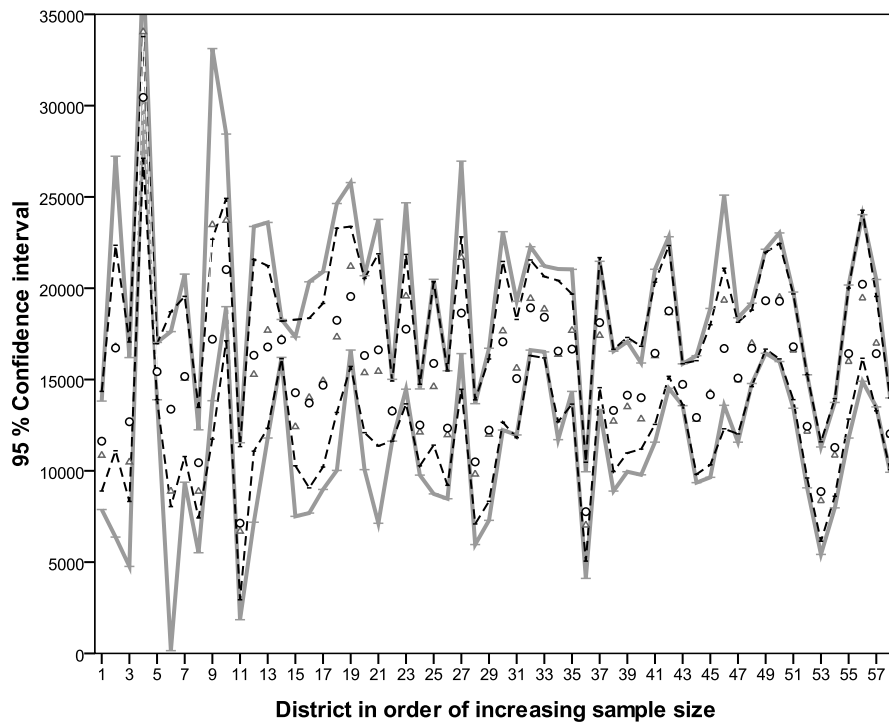


Figure 5. 95 per cent confidence interval (CI) for direct estimates (Δ) and model based estimates (O) for sampled districts. CI for direct estimates (solid line) and CI for model based estimates (dash line).

INTRODUCTION TO REMOTE SENSING TECHNIQUES

Prachi Mishra Sahoo

Indian Agricultural Statistics Research Institute, New Delhi-110012

28.1 A DEFINITION OF REMOTE SENSING

The transport of information from an object to a receiver (observer) by means of radiation transmitted through the atmosphere. The interaction between the radiation and the object of interest conveys information required on the nature of the object (eg. reflection coefficient, emittance, roughness).

Examples

- (i) The reflection of sunlight from vegetation will give information on the reflection coefficient of the object and its spectral variation, and thus on the nature of the object (green trees, etc.).
- (ii) Microwave radiation transmitted from a radar system and scattered from a rain cloud in the back direction to a receiver will give information on the raindrop size and intensity.

28.1.1 PASSIVE AND ACTIVE SENSING

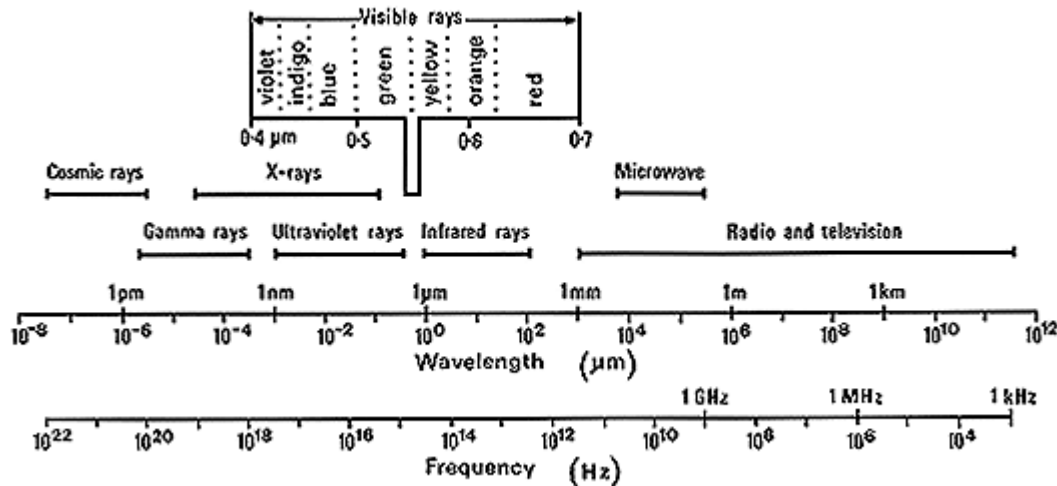
The first example above is an example of passive remote sensing, where the reflected radiation observed originates from a natural source - the sun. The second example is an example of active remote sensing, where the scattered radiation originates from a specially designed active radar system.

28.1.2 ELECTROMAGNETIC RADIATION

Radiation can be observed either as a wave motion, or as single discrete packets of energy, photons. Normally, one is dealing with a large number of photons arriving in a short time, and the radiation can be treated physically as a wave motion. However, in the visible and ultraviolet regions, very weak sources are typified by the detection of single photons. The wave theory of radiation has been developed extensively. It impacts on remote sensing in the way that radiation is reflected at a surface and transmitted, absorbed and scattered in a medium.

28.1.3 THE ELECTROMAGNETIC SPECTRUM

Electromagnetic radiation covers a very large range of wavelengths. In Remote Sensing we are concerned with radiation from the ultraviolet (UV), which has wavelengths of from 0.3 to 0.4 μm (10^{-6} m) to radar wavelengths in the region of 10 cm (10^{-1} m) (see Figure below). Thus the phenomena observed in the various wavelength regions differ considerably.



Range of electromagnetic wavelengths and the transmission through the atmosphere.

28.1.4 DATA RESOLUTIONS

Resolution refers to the intensity or rate of sampling, and extent refers to the overall coverage of a data set. Extent can be seen as relating to the largest feature, or range of features, which can be observed, while resolution relates to the smallest. For a feature to be distinguishable in the data, the resolution and extent of the measurement dimensions of the data set need to be appropriate to the measurable properties of the feature. For a feature to be separable from other features, these measurements must also be able to discriminate between the differences in reflectance from the features

28.1.5 SPECTRAL

As indicated in the preceding sections, different materials respond in different, and often distinctive, ways to EM radiation. This means that a specific spectral response curve, or spectral signature, can be determined for each material type. Basic categories of matter (such as specific minerals) can be identified on the basis of their spectral signatures alone, but may require that the spectra be sufficiently detailed in terms of wavelength intervals and covers a wide spectral range. Composite categories of matter (such as soil which contains several different minerals) however, may not be uniquely identifiable on the basis of spectral data alone.

28.1.6 SPATIAL

Spatial resolution defines the level of spatial detail depicted in an image. This may be described as a measure of the smallness of objects on the ground that may be distinguished as separate entities in the image, with the smallest object necessarily being larger than a single pixel. In this sense, spatial resolution is directly related to image pixel size. In terms of photographic data, an image pixel may be compared to grain size while spatial resolution is more closely related to photographic scale. In practical terms, the 'detectability' of an object in an image involves consideration of spectral contrast as well as spatial resolution. Feature shape is also relevant to visual discrimination in an image with long thin features such as roads showing up more

readily than smaller symmetric ones. Pixel size is usually a function of the platform and sensor, while the detectability may change from place to place and time to time.

28.1.7 RADIOMETRIC

Radiometric resolution in remotely sensed data is defined as the amount of energy required to increase a pixel value by one quantisation level or 'count'. The radiometric extent is the dynamic range or the maximum number of quantisation levels that may be recorded by a particular sensing system. Most remotely sensed imagery is recorded with quantisation levels in the range 0­255, that is, the minimum 'detectable' radiation level is recorded as 0 while the 'maximum' radiation is recorded as 255. This range is also referred to as 8 bit resolution since all values in the range may be represented by 8 bits (binary digits) in a computer. Radiometric resolution in digital imagery is comparable to the number of tones in a photographic image ­ both measures being related to image contrast.

28.1.8 TEMPORAL

The temporal resolution of remotely sensed data refers to the repeat cycle or interval between acquisition of successive imagery. This cycle is fixed for spacecraft platforms by their orbital characteristics, but is quite flexible for aircraft platforms. Satellites offer repetitive coverage at reduced cost but the rigid overpass times can frequently coincide with cloud cover or poor weather. This can be a significant problem when field work needs to coincide with image acquisition. While aircraft data are necessarily more expensive than satellite imagery, these data offer the advantage of user-defined flight timing, which can be modified if necessary to suit local weather conditions. The off-nadir viewing capability of the SPOT­HRV provides some flexibility to the usual repeat cycle of satellite imagery by imaging areas outside of the nadir orbital path. This feature allows daily coverage of selected regions for short periods and has obvious value for monitoring dynamic events such as flood or fire.

28.1.9 DIGITAL IMAGE PROCESSING

The roots of remote sensing reach back into ground and aerial photography. But modern remote sensing really took off as two major technologies evolved more or less simultaneously: 1) the development of sophisticated electro-optical sensors that operate from air and space platforms and 2) the digitizing of data that were then in the right formats for processing and analysis by versatile computer-based programs. Today, analysts of remote sensing data spend much of their time at computer stations, but nevertheless still also use actual imagery (in photo form) that has been computer-processed. Now it can be seen that the individual bands and color composites that have introduced in the previous lectures and it is interesting to investigate the power of computer-based processing procedures in highlighting and extracting information about scene content, that is, the recognition, appearance, and identification of materials, objects, features, and classes (these general terms all refer to the specific spatial and spectral entities in a scene).

Processing procedures fall into three broad categories: *Image Restoration (Preprocessing)*; *Image Enhancement*; and *Classification and Information Extraction*. Apart from preprocessing the techniques of contrast stretching, density slicing, and spatial filtering will be discussed. Under Information Extraction, ratioing and

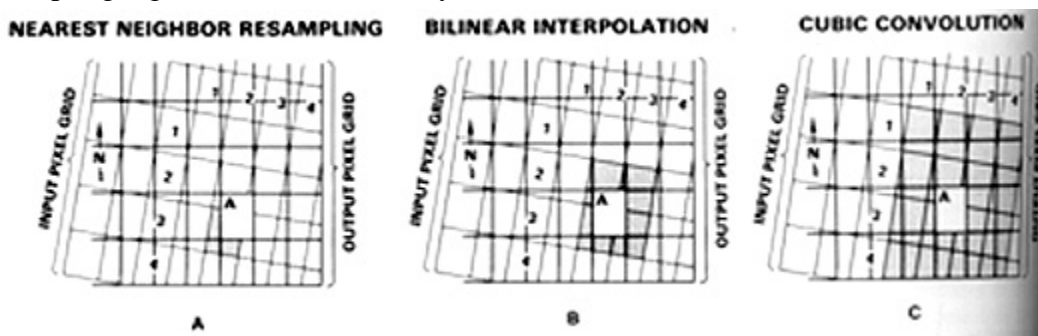
principal components analysis have elements of Enhancement but lead to images that can be interpreted directly for recognition and identification of classes and features. Also included in the third category but treated outside this lecture is Change Detection and Pattern recognition.

The data in satellite remote sensing is in the form of *Digital Number* or DN. It is said that the radiances, such as reflectance and emittances, which vary through a continuous range of values are digitized onboard the spacecraft after initially being measured by the sensor(s) in use. Ground instrument data can also be digitized at the time of collection. Or, imagery obtained by conventional photography is capable of digitization. A DN is simply one of a set of numbers based on powers of 2, such as 2^6 or 64. The range of radiances, which instrument-wise, can be, for example, recorded as varying voltages if the sensor signal is one which is, say, the conversion of photons counted at a specific wavelength or wavelength intervals. The lower and upper limits of the sensor's response capability form the end members of the DN range selected. The voltages are divided into equal whole number units based on the digitizing range selected. Thus, a IRS band can have its voltage values - the maximum and minimum that can be measured - subdivided into 2^8 or 256 equal units. These are arbitrarily set at 0 for the lowest value, so the range is then 0 to 255.

28.1.10 PREPROCESSING

Preprocessing is an important and diverse set of image preparation programs that act to offset problems with the band data and recalculate DN values that minimize these problems. Among the programs that optimize these values are atmospheric correction (affecting the DNs of surface materials because of radiance from the atmosphere itself, involving attenuation and scattering); sun illumination geometry; surface-induced geometric distortions; spacecraft velocity and attitude variations (roll, pitch, and yaw); effects of Earth rotation, elevation, curvature (including skew effects), abnormalities of instrument performance (irregularities of detector response and scan mode such as variations in mirror oscillations); loss of specific scan lines (requires destriping), and others. Once performed on the raw data, these adjustments require appropriate radiometric and geometric corrections.

Resampling is one approach commonly used to produce better estimates of the DN values for individual pixels. An estimate of the new brightness value (as a DN) that is closer to the B condition is made by some mathematical re-sampling technique. Three sampling algorithms are commonly used:



In the Nearest Neighbor technique, the transformed pixel takes the value of the closest pixel in the pre-shifted array. In the Bilinear Interpolation approach, the average of

the DN's for the 4 pixels surrounding the transformed output pixel is used. The Cubic Convolution technique averages the 16 closest input pixels; this usually leads to the sharpest image.

28.2 FALSE COLOR COMPOSITE

The first example of a color composite, made by combining (either photographically or with a computer-processing program) any three bands of images with some choice of color filters, usually blue, green, and red. The customary false color composite made by projecting a green band image through a blue filter, a red band through green, and the photographic infrared image through a red filter.

28.2.1 TRUE COLOR VIEW

By projecting IRS Bands 1, 2, and 3 through blue, green, and red filters respectively, a quasi-true color image of a scene can be generated.

28.2.2 OTHER COLOR COMBINATIONS

Other combinations of bands and color filters (or computer assignments) produce not only colorful new renditions but in some instances bring out or call attention to individual scene features that, although usually present in more subtle expressions in the more conventional combinations, now are easier to spot and interpret.

28.2.3 CONTRAST STRETCHING:

Almost without exception, the image will be significantly improved if one or more of the functions called Enhancement are applied. Most common of these is contrast stretching. This systematically expands the range of DN values to the full limits determined by byte size in the digital data. *For IRS this is determined by the eight-bit mode or 0 to 255 DN's.* Examples of types of stretches and the resulting images are shown. Density slicing is also examined. The contrast stretching, which involves altering the distribution and range of DN values, is usually the first and commonly a vital step applied to image enhancement. Both casual viewers and experts normally conclude from direct observation that modifying the range of light and dark tones (gray levels) in a photo or a computer display is often the single most informative and revealing operation performed on the scene. When carried out in a photo darkroom during negative and printing, the process involves shifting the gamma (slope) or film transfer function of the plot of density versus exposure (H-D curve). This is done by changing one or more variables in the photographic process, such as, the type of recording film, paper contrast, developer conditions, etc. Frequently the result is a sharper, more pleasing picture, but certain information may be lost through trade-offs, because gray levels are "overdriven" into states that are too light or too dark.

28.2.4 SPATIAL FILTERING

Just as contrast stretching strives to broaden the image expression of differences in spectral reflectance by manipulating DN values, so spatial filtering is concerned with expanding contrasts locally in the spatial domain. Thus, if in the real world there are boundaries between features on either side of which reflectance (or emissions) are quite different (notable as sharp or abrupt changes in DN value), these boundaries can be emphasized by any one of several computer algorithms (or analog optical filters). The resulting images often are quite distinctive in appearance. Linear features, in particular, such as geologic faults can be made to stand out. The type of filter used,

high- or low-pass, depends on the spatial frequency distribution of DN values and on what the user wishes to accentuate.

Another processing procedure falling into the enhancement category that often divulges valuable information of a different nature is *spatial filtering*. Although less commonly performed, this technique explores the distribution of pixels of varying brightness over an image and, especially detects and sharpens boundary discontinuities. These changes in scene illumination, which are typically gradual rather than abrupt, produce a relation that we express quantitatively as "spatial frequencies". The spatial frequency is defined as the number of cycles of change in image DN values per unit distance (e.g., 10 cycles/mm) along a particular direction in the image. An image with only one spatial frequency consists of equally spaced stripes (raster lines). For instance, a blank TV screen with the set turned on has horizontal stripes. This situation corresponds to zero frequency in the horizontal direction and a high spatial frequency in the vertical.

28.3 PRINCIPAL COMPONENTS ANALYSIS

There is a tendency for multi-band data sets/images to be somewhat redundant wherever bands are adjacent to each other in the (multi-)spectral range. Thus, such bands are said to be correlated (relatively small variations in DNs for some features). A statistically based program, called Principal Components Analysis, decorrelates the data by transforming DN distributions around sets of new multi-spaced axes. The underlying basis of PCA is described in this section. Color composites made from images representing individual components often show information not evident in other enhancement products. Canonical Analysis and Decorrelation Stretching are also mentioned.

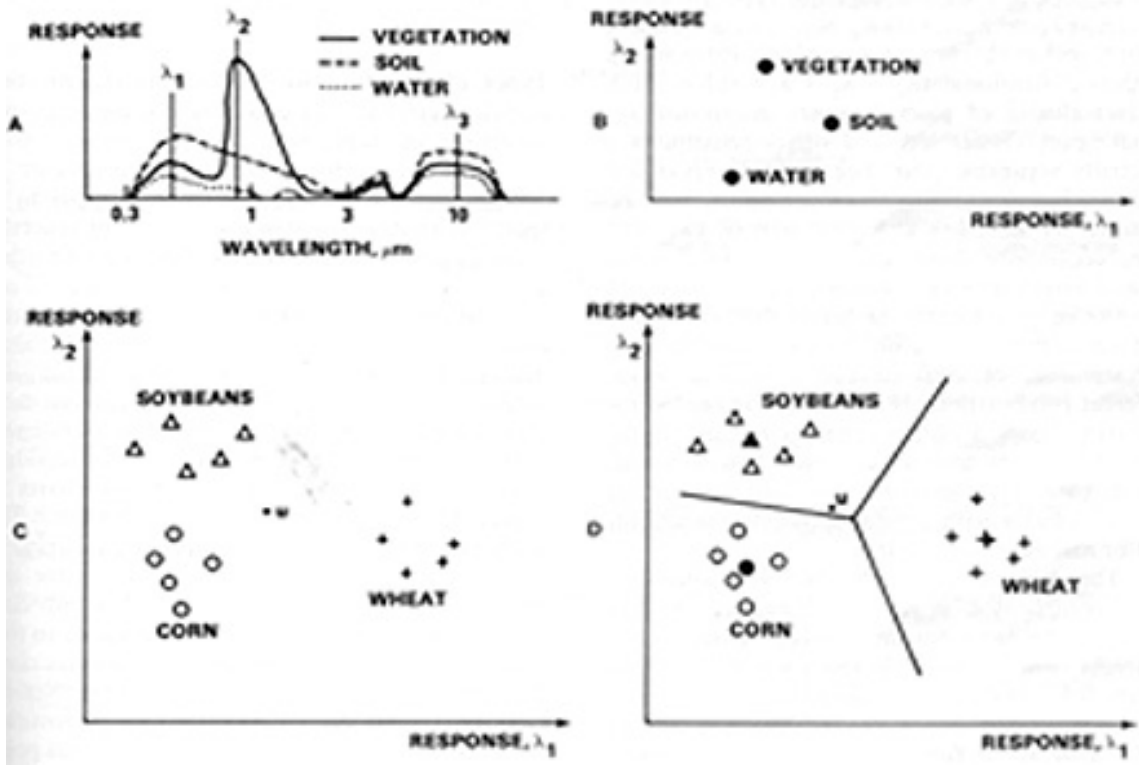
28.4 RATIOING

Ratioing is an enhancement process in which the DN value of one band is divided by that of any other band in the sensor array. If both values are similar, the resulting quotient is a number close to 1. If the numerator number is low and denominator high, the quotient approaches zero. If this is reversed (high numerator; low denominator) the number is well above 1. These new numbers can be stretched or expanded to produce images with considerable contrast variation in a black and white rendition. Certain features or materials can produce distinctive gray tones in certain ratios. Three band ratio images can be combined as color composites, which highlight certain features in distinctive colors. Ratio images also reduce or eliminate the effects of shadowing.

28.5 CLASSIFICATION

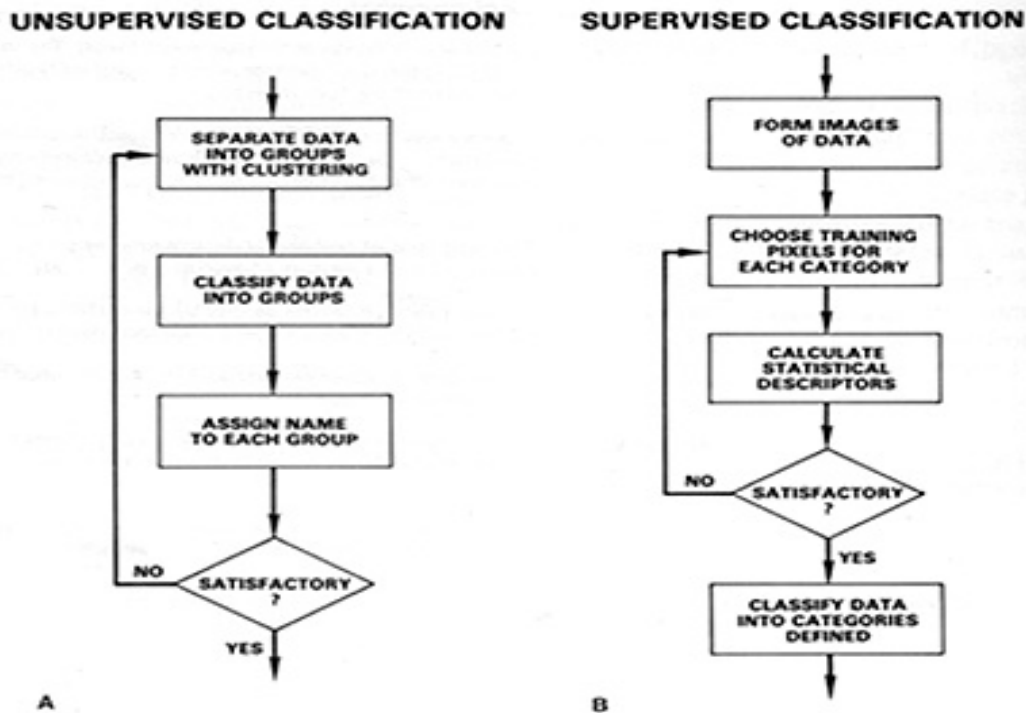
There are two of the common methods for identifying and classifying features in images: *Unsupervised* and *Supervised Classification*. Closely related to Classification is the approach called *Pattern Recognition*.

Before starting, it is well to review several basic principles, with the aid of this diagram:



In the upper left are plotted spectral signatures for three general classes: Vegetation; Soil; Water. The relative spectral responses (reflectance in this spectral interval), in terms of some unit, e.g., reflected energy in appropriate units or percent (as a ratio of reflected to incident radiation, times 100), have been sampled at three wavelengths. (The response values are normally converted [either at the time of acquisition on the ground or aircraft or spacecraft] to a digital format, the DNs or Digital Numbers cited before, commonly subdivided into units from 0 to 255 [2^8]).

Two methods of classification are commonly used: Unsupervised and Supervised. The logic or steps involved can be grasped from these flow diagrams:



In *unsupervised classification* any individual pixel is compared to each discrete cluster to see which one it is closest to. A map of all pixels in the image, classified, as to which cluster each pixel is most likely to belong, is produced (in black and white or more commonly in colors assigned to each cluster). This then must be interpreted by the user as to what the color patterns may mean in terms of classes, etc. that are actually present in the real world scene; this requires some knowledge of the scene's feature/class/material content from general experience or personal familiarity with the area imaged. In *supervised classification* the interpreter knows beforehand what classes, etc. are present and where each is in one or more locations within the scene. These are located on the image, areas containing examples of the class are circumscribed (making them training sites), and the statistical analysis is performed on the multiband data for each such class. Instead of clusters then, one has class groupings with appropriate discriminant functions that distinguish each (it is possible that more than one class will have similar spectral values but unlikely when more than 3 bands are used because different classes/materials seldom have similar responses over a wide range of wavelengths). All pixels in the image lying outside training sites are then compared with the class discriminants, with each being assigned to the class it is closest to - this makes a map of established classes (with a few pixels usually remaining unknown) which can be reasonably accurate (but some classes present may not have been set up; or some pixels are misclassified).

28.5.1 SUPERVISED CLASSIFICATION

Supervised classification is much more accurate for mapping classes, but depends heavily on the cognition and skills of the image specialist. The strategy is simple: the specialist must recognize conventional classes (real and familiar) or meaningful (but somewhat artificial) classes in a scene from prior knowledge, such as, personal experience with the region, by experience with thematic maps, or by on-site visits. This familiarity allows the specialist to choose and set up discrete classes (thus supervising the selection) and then, assign them category names. The specialists also

locate training sites on the image to identify the classes. **Training Sites** are areas representing each known land cover category that appear fairly homogeneous on the image (as determined by similarity in tone or color within shapes delineating the category). Specialists locate and circumscribe them with polygonal boundaries drawn (using the computer mouse) on the image display. For each class thus outlined, mean values and variances of the DNs for each band used to classify them are calculated from all the pixels enclosed in the site. More than one polygon can be established for any class. When DNs are plotted as a function of the band sequence (increasing with wavelength), the result is a **spectral signature** or spectral response curve for that class. In reality the spectral signature is for all of the materials within the site that interact with the incoming radiation. Classification now proceeds by statistical processing in which every pixel is compared with the various signatures and assigned to the class whose signature comes closest. A few pixels in a scene do not match and remain unclassified, because these may belong to a class not recognized or defined).

Many of the classes in general are almost self-evident ocean water, waves, beach, marsh, shadows. In practice, we could further sequester several such classes. For example, we might distinguish between ocean and bay waters, but their gross similarities in spectral properties would probably make separation difficult. Other classes that are likely variants of one another, such as, slopes that faced the morning sun as IRS flew over versus slopes that face away, might be warranted. Some classes are broad-based, representing two or more related surface materials that might be separable at high resolution but are inexactly expressed in the IRS image. In this category we can include trees, forests, and heavily vegetated areas (the golf course or cultivated farm fields).

Note that software does not name them during the stage when the signatures are made. Instead, it numbers them and names are assigned later. Several classes gain their data from more than one training site. Most of the software has a module that plots the signature of each class.

28.5.2 MINIMUM DISTANCE CLASSIFICATION

One of the simplest supervised classifiers is the parallelepiped method. But on we employ a (usually) somewhat better approach (in terms of greater accuracy) known as the Minimum Distance classifier. This sets up clusters in multidimensional space, each defining a distinct (named) class. Any pixel is then assigned to that class if it is closest to (shortest vector distance).

We initiate our exemplification of Supervised Classification by producing one using the `Minimum_Distance` routine. The software program acts on DNs in multidimensional band space to organize the pixels into the classes we choose. Each unknown pixel is then placed in the class *closest* to the mean vector in this band space. We can elect to combine classes to have either color themes (similar colors for related classes) and/or to set apart spatially adjacent classes by using disparate colors

28.5.3 MAXIMUM LIKELIHOOD CLASSIFICATION

The most powerful classifier in common use is that of Maximum Likelihood. Based on statistics (mean; variance/covariance), a (Bayesian) Probability Function is calculated from the inputs for classes established from training sites. Each pixel is then judged as to the class to which it most probably belongs. This is done with the IRS data, using three reflected radiation bands. The result is a pair of quite believable classification maps whose patterns (the classes) seem to closely depict reality but keep in mind that several classes are not normal components of the actual ground scene, e.g., shadows.

In many instances the most useful image processing output is a classified scene. This is because you are entering a partnership with the processing program to add information from the real world into the image you are viewing, in a systematic way, in which you try to associate names of real features or objects with the spectral/spatial patterns evident in individual bands, color composites, or PCI images. The most of the software are capable of producing both unsupervised and supervised classifications.

REFERENCES

1. Lillesand, T.M. and Kiefer.R .W (1987): Remote Sensing and Image Interpretation. Jhon Wiley & Sons, New York.
2. Sabins, F.F..Jr (1987): Remote Sensing – Principles and Interpretations, W.H.Freeman and Company, New York.

INTRODUCTION TO GEOGRAPHICAL INFORMATION SYSTEM (GIS)

Anshu Bharadwaj

Indian Agricultural Statistics Research Institute, New Delhi-110012

29.1 Introduction

The most effective means of depicting events or phenomena over space and time is through spatial representation or a map. A Map explains relationship between different objects or processes. At the beginning of our civilization, information was represented as an artistic depiction. Today, with the advent of remote sensing, Global positioning system, organization of databases around Geographic Information System as well as advances in computing and communication technologies and digital cartography has revolutionized the map making.

Geographical Information Systems (GIS) are systems (of hardware, software, data, applications and policies) that deal with spatially referenced and geographically tagged or linked data. Over the past 40-50 years, GIS technology has evolved into an “integrating” technology that encompasses surveying & positioning, map-making and cartography, imaging and image interpretation, databases, computing and networking technology. Applications of GIS are varied and support natural resources management, disaster management, planning and development, environmental management, land and water management, ocean and marine research, climate change and many other areas where people, society are involved. Thus, GIS has become not only an important technology but is also becoming a tool that assists in governance, development of society and supports citizen activities. GIS allows integration of multiple maps/image with geo-tagged tabular data and enables determine spatial patterns and choice based on spatial criterions. For example, GIS allows to determine the spatial distribution of features/objects (say, distribution of hospitals or distribution of flooding in a city and so on), the relationship among entities in a spatial distribution (say, distribution of hospitals to roads or flooding spread to Emergency Centres and so on) and the correlation of multiple spatial variables (does population correlate with land use or the correlation between soils, slopes, land use to determine the sediment yield in a reservoir and so on) in geographic space (say, in a district, a watershed or a nation or the whole Earth itself). Further, today’s GIS systems allow creating map visualization of and making amenable the spatial or map representation of tabular data – say, population data, migration data, consumer data, financial transactions, and beneficiary data and so on, thereby allowing creation of population maps, consumer maps and their visualisation.

Applications of GIS has seen a quantum jump with its integration on the Web platform - which now provides a GIS engine and front-end GIS interface to any users on a simple browser. From a technology perspective, the GIS Web component can interface with any type of client - desktop, mobile or Web and serving GIS maps and GIS Applications prolifically to a large community of users. GIS users can now create pervasive

geographic knowledge – their own maps, their GIS models and their own workflows and decision-rules and Geo-web services can deliver this GIS knowledge to everyone and, thereby, help better understand data correlations in spatial format and therefore help in better decisions to be made.

29.2 Geoinformatics and GIS

Spatial data handling involves many disciplines. We can distinguish disciplines that develop spatial concepts, provide means for capturing and processing of spatial data, provide a formal and theoretical foundation, are application-oriented, and support spatial data handling in legal and management aspects. Table given below shows a classification of some of these disciplines. They are grouped according to how they deal with spatial information. The list is not meant to be exhaustive.

Table 1: Disciplines involved in spatial data handling

Characteristics of disciplines	Sample disciplines
Development of spatial concepts	Geography Cognitive Science Linguistics Psychology
Means for capturing and processing spatial data	Remote Sensing Surveying Engineering Cartography Photogrammetry
Formal and theoretical foundation	Computer Science Expert Systems Mathematics Statistics
Applications	Archaeology Architecture Forestry Geo-Sciences Regional and Urban Planning Surveying
Support	Legal Sciences Economy

The discipline that deals with all aspects of spatial data handling is called geoinformatics. It is defined as:

Geoinformatics is the integration of different disciplines dealing with spatial information.

Geoinformatics has also been described as “the science and technology dealing with the structure and character of spatial information, its capture, its classification and qualification, its storage, processing, portrayal and dissemination, including the

infrastructure necessary to secure optimal use of this information”. It is also defined as “the art, science or technology dealing with the acquisition, storage, processing production, presentation and dissemination of geoinformation.”

Terminology

Frequently used technical terms in spatial data handling related to GIS are:

- geographic (or geographical) information system (GIS),
- geo-information system,
- spatial information system (SIS),
- land information system (LIS), and
- multi-purposecadastre.

Geographic information systems are used by various disciplines as tools for spatial data handling in a geoinformatics environment. There are many definitions for a geographic information system. The main characteristics, however, are the analytical functions that provide means for deriving new information from existing data.

Just as we use a word processor to write documents and deal with words on a computer, we can use a GIS application to deal with spatial information on a computer. GIS stands for 'Geographical Information System'. A GIS consists of:

- Digital Data - the geographical information that you will view and analyse using computer hard- ware and software.
- Computer Hardware - computers used for storing data, displaying graphics and processing data.
- Computer Software - computer programs that run on the computer hardware and allow you to work with digital data. A software program that forms part of the GIS is called a GIS Application.

With a GIS application you can open digital maps on your computer, create new spatial information to add to a map, create printed maps customised to your needs and perform spatial analysis.

A GIS is defined as follows (ARONOFF, 1989):

“A GIS is a computer-based system that provides the following four sets of capabilities to handle geo-referenced data:

1. input,
2. data management (data storage and retrieval),
3. manipulation and analysis, and
4. output.”

Depending on the interest of a particular application, a GIS can be considered to be a data store (application of a spatial database), a tool- (box), a technology, an information source or a science (spatial information science).

Definitions: Few more definitions of GIS are:

The *common ground* between information processing and the many fields using spatial analysis techniques. (Tomlinson, 1972)

A powerful *set of tools* for collecting, storing, retrieving, transforming, and displaying spatial data from the real world. (Burroughs, 1986)

A computerised *database management system* for the capture, storage, retrieval, analysis and display of spatial (locationally defined) data. (NCGIA, 1987)

A *decision support system* involving the *integration* of spatially referenced data in a problem solving environment. (Cowen, 1988)

An intelligent definition of GIS has been given as:

A system of integrated computer-based tools for end-to-end processing (capture, storage, retrieval, analysis, display) of data using location on the earth's surface for interrelation in support of operations management, decision making, and science.

- set of integrated tools for spatial analysis
- encompasses end-to-end processing of data
 - capture, storage, retrieval, analysis/modification, display
- uses explicit location on earth's surface to relate data
- aimed at decision support, as well as on-going operations and scientific inquiry

Like in any other discipline, the use of tools for problem solving is one thing; to produce these tools is something different. Not all are equally well suited for a particular application. Tools can be improved and perfected to better serve a particular need or application. The discipline that provides the background for the production of the tools in spatial data handling is spatial information theory (or SIT) or geographic information science.

29.3 Characteristics of Spatial Data

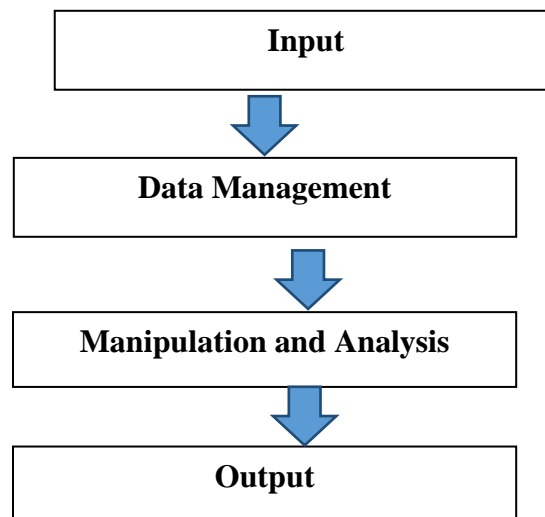
A GIS deals with spatial data or objects (e.g., parcels, rivers, wells, ...), their attributes and characteristics (e.g., location, area, length, name, depth, ...) and the relationships between the objects (e.g., a parcel boundary follows a river, a well is located in a certain parcel, ...). The objects are stored in the database with geometric primitives (volumes, areas, lines, and points) and the relationships between them (topology). Spatial data have the following characteristics:

Table 2: Characteristics of spatial data

Spatial reference (geographic location, coordinates)	where?
Attributes (non-spatial)	what?
Spatial relationships (topology, metric, order)	in what relationship?
Temporal component (different concepts of time)	when?

Functional Components of a GIS

According to the definition a GIS always consists of modules for input, storage, analysis, display and output of spatial data. Figure 2 shows a diagram of these modules with arrows indicating the flow of data in the system. For a particular system, each of these modules may provide more or less functions. However, if one would be completely missing the system should not be called a geographic information system. The most important component of GIS software is the functional complex of functions for spatial analysis. If this functionality is poorly developed, we cannot call such a system a GIS.

**Figure 1: Functional modules of a GIS**

The functions for data input are closely related to the disciplines of surveying engineering, photogrammetry, remote sensing, and the processes of digitizing, i.e., the conversion of analog data into a digital representation. Today, digital data on different media or on a computer network are used increasingly. Following table lists the methods and devices used in the data input process.

Table 3: Data input

Method	Devices
manual digitizing	coordinate entry via keyboard
	digitizing tablet with cursor or mouse (digital) photogrammetry
semi-automatic digitizing	line following devices
automatic digitizing	scanner
input of available digital data	magnetic tape CD ROM networks

Data output is closely related to the disciplines of cartography, printing and publishing. Table 4 lists different methods and devices used for the output of spatial data. Cartography and scientific visualization make use of these methods and devices to produce their products. The importance of digital products (datasets) is increasing and data dissemination on digital media or on computer networks becomes extremely important. In both processes, data input and data output, the Internet, and Internet technology have a major share. The World Wide Web plays the role of an easy to use interface to repositories of large datasets. Aspects of data dissemination, security, copyright, and pricing require special attention. Spatial information infrastructure deals with these issues.

Table 4: Data output and visualization

Method	Devices
Hardcopy	printerplotter (pen plotter, ink-jet printer, thermal transfer printer, electrostatic plotter)film writer
Softcopy	computer screen (CRT)
Output of digital datasets	magnetic tape CD ROM network

Why GIS?

A general motivation for the use of GIS can be illustrated with the following example. For a planning task usually different maps and other data sources are needed. Assuming a conventional analogue procedure we would have to collect all the maps and documents needed before we can start the analysis. The first problem we encounter is that the maps and data have to be collected from different sources at different locations (e.g., mapping agency, geological survey, soil survey, forest survey, census bureau, etc.), and that they are in different scales and projections. In order to combine data from maps they have to be converted into working documents of the same scale and projection. This has to be done manually, and it requires much time and money.

With the help of a GIS, the maps can be stored in digital form in a database in world coordinates (meters or feet). This makes scale transformations unnecessary, and the

conversion between map projections can be done easily with the software. The spatial analysis functions of the GIS are then applied to perform the planning tasks. This can speed up the process and allows for easy modifications to the analysis approach.

Who would use a GIS?

Simply put, anybody who needs to work with spatially referenced data. A small number of examples of potential users are as follows. Municipalities maintain large and complex databases that contain the street locations, building footprints, height contours, sewer lines, land use designations, and much more. Hydro and phone companies use them to record locations of their lines, both above and below ground, and for deciding where to put new ones. Geologists use them to record locations of rock formations and for use in resource prospecting operations. Anthropologists use them to record locations of current sites and perhaps to predict where new ones could be found. The military maintains very large, comprehensive, and usually highly classified databases on everything that could be useful to them. And emergency services like 911 have to have a very detailed municipal address database in order to route the vehicles to the emergency as quickly as possible. Cemeteries could use a GIS to store the locations and occupants of the burial plots. Mount Pleasant Cemetery in the heart of Toronto is renowned for its collection of trees and shrubs, the locations of which could also be stored in a GIS. To my knowledge, they have not yet done so. This is not an exhaustive list!

How this relates to us!

We are all GIS, since we use and make decisions based on spatial data all the time. For example, the locations of your dwelling, work place, school, nearby stores, banks, and local landmarks are all included in your personal spatial database and are normally what you would think of when asked about spatial data. However, don't forget the less obvious things, like computer keyboards, remote controls, locations of items in a store, and the location of your furniture (important for the 3 a.m. bathroom run). We pose questions, called queries in the jargon, to our spatial databases, like where is the nearest grocery store, how do I get there, or perhaps in idle speculation like what is the average income in Rosedale? When we move to a new part of town (or even a new town), our queries often come up blank and we have to update our neighbourhood databases with the locations of stores, bus stops, parks, and so on. We also make decisions using spatial data, some of which are quite complex, on a daily basis. Perhaps the most common is route planning, usually from your home to some other place. This can be made more complex by your significant other calling and asking that you stop by a grocery store on the way home and pick up some broccoli for dinner. If the store is significantly out of your way, you may have to adjust the route for your trip home. Others that you might not immediately consider include how to pack stuff in boxes and where to put the boxes in the truck, designing a flower garden, and even interior decorating. The point is that a GIS is a tool we use to help us to store and manipulate large datasets and to perform complex operations that would take a human a long time (with plenty of opportunity for errors) to do. However, the algorithms and storage techniques that it uses are usually analogous to

human thought processes. The purpose of this document is to explain a number of the common processes used by GIS to provide an idea of how they work.

29.4 Modeling and Structuring the Spatial Data: How we represent features or spatial elements?

The classic example of a database that is not spatially referenced is a telephone directory. In it are stored the subscriber's name, address and telephone number, sorted by last name. Although it contains spatial data (the address) the referencing is by the person's name. You cannot use the phone book to get the numbers of everyone on your street (at least, not easily), or everyone in your neighbourhood. The biggest headache for designers and maintainers of GIS is that there are many different ways in which data can be locationally referenced. Any GIS worthy of its name should be able to handle any, or any combination, of the following types of data:

- Point: Addresses, elevation spot heights, locations of malls, banks, cities, volcanoes, etc.
- Line: Contours, geological faults, streets, highways, rivers, etc.
- Areas: Forests, climatic zones, lakes, soil types, land use, nations, counties, etc.
- Networks: Streets, highways, rivers (which are directed networks, an extra complication!)
- Tessellations: Census districts, postal codes, electoral boundaries. (A tessellation completely divides a region into non-overlapping areas.)
- Overlapping regions: Newspaper circulation areas, telephone exchanges.

The GIS must be able to store all the data for the geographical entities, along with whatever non-spatial attributes that are attached to them, in a way that can minimize disk file size and retrieval time. Methods fall into three basic data models, or structures, described below. So how can we store geographic data? In Layers In order to better organize geographical data in a region, data that describe similar themes are stored separately. For example, a standard topographic map sheet shows contours, road networks, stream networks, power lines, forested areas, buildings, and spot heights, among other things. The descriptions for each would be stored in different files, and these are referred to as layers. The concept is analogous to drawing each on a transparency and then overlaying them at your will.

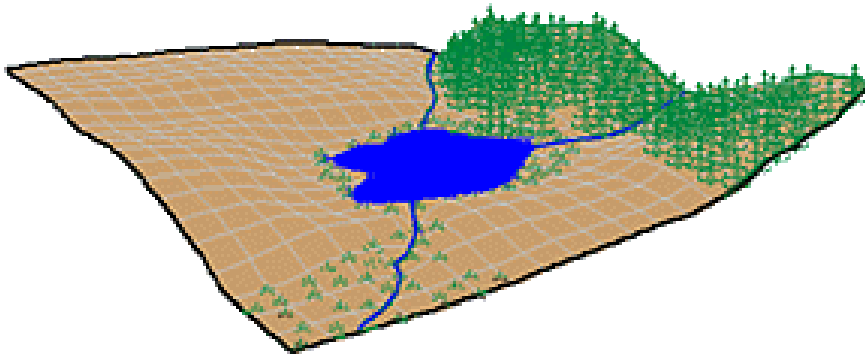


Figure 2: Real World

29.4.1 Raster data model

The region of interest is divided up into small regular blocks (usually squares), with each block having a specific value attached to it. Each variable in the data set will be defined in a different layer. Even locations where the variable (e.g. forest) is not present must be given a value, usually zero. It's easy to see that for a large area with a large number of variables, the data set can get very large very quickly. Stores images as rows and columns of numbers with a Digital Value/Number (DN) for each cell.

Units are usually represented as square grid cells that are uniform in size.

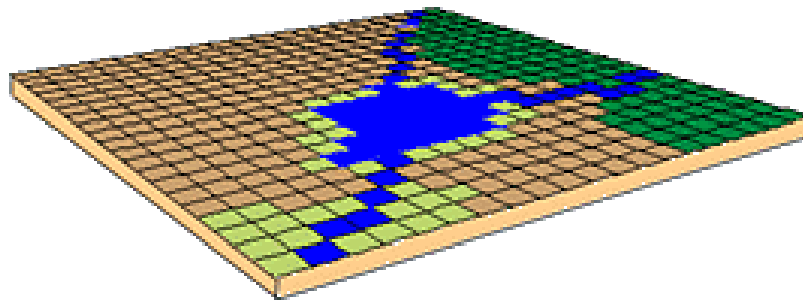


Figure 3: Raster Representation of Real World

Data is classified as “*continuous*” (such as in an image), or “*thematic*” (where each cell denotes a feature type. Numerous data formats (TIFF, GIF, ERDAS.imgetc)

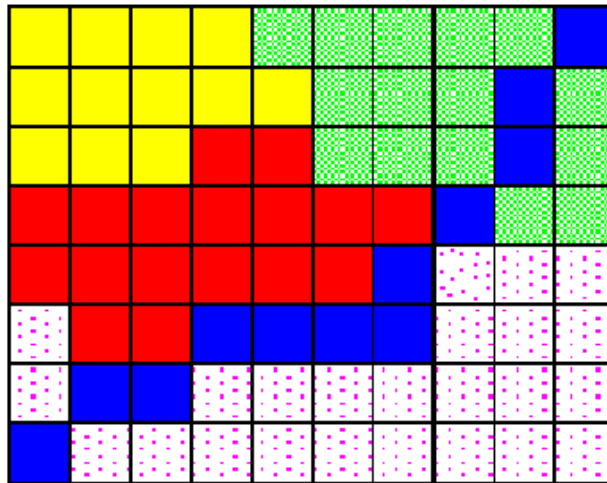


Figure 4: Grid Cell Structure

Advantages: Layer overlays are really simple, since all layers are defined with the same grid over the region. Topology is implicitly defined, since the location of each cell relative to all the others can be easily found. Disadvantages: If you want to increase the resolution (that is, decrease the cell size) by a factor of two, the data set size will

quadruple! In order to reduce this problem, various compression techniques, such as quadtrees and run-length encoding, are employed. Resolution is also problematic because the discretization process has an effect analogous to rounding of numbers, but in a spatial sense -- that is, what you see in the raster image is usually larger or smaller than the real-world equivalent. Objects smaller than one cell may not appear at all! Uses: All satellite and aerial photograph data come in raster form. Each pixel represents the amount of light received by the sensor at a particular wavelength at the location. All satellites collect data from more than one wavelength, so a particular satellite pass will create an instant multilayer raster map of an area, as well as business for the data storage industry. Common GIS packages using the raster model are GRASS and IDRISI. Raster data are best used for representing variables that vary continuously in space, such as elevations.

29.4.2 Vector Data Model

All of the geographic objects of interest are described in terms of geometric elements: points, lines, polygons, and volumes if data are three-dimensional. All similar entities are grouped together and stored in different layers, as described above. Advantages: Much greater precision in the definition of objects is possible by defining the geometric extent of the regions in which they occur. This means that one can draw far better maps with vector data than with raster data. Much less space is required to store all the information, since empty space on the map can be ignored. Allows user to specify specific spatial locations and assumes that geographic space is continuous, not broken up into discrete grid squares

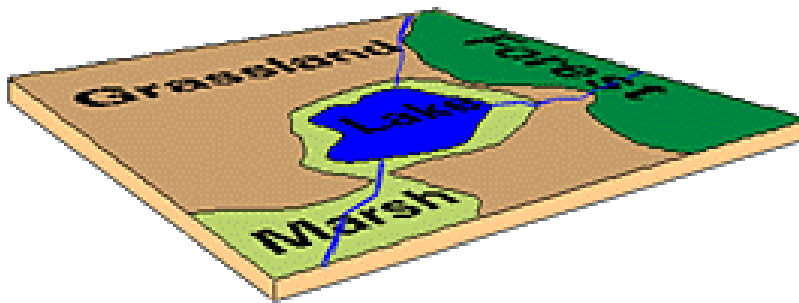


Figure 5: Vector Representation of Real World

We store features as sets of X,Y coordinate pairs.

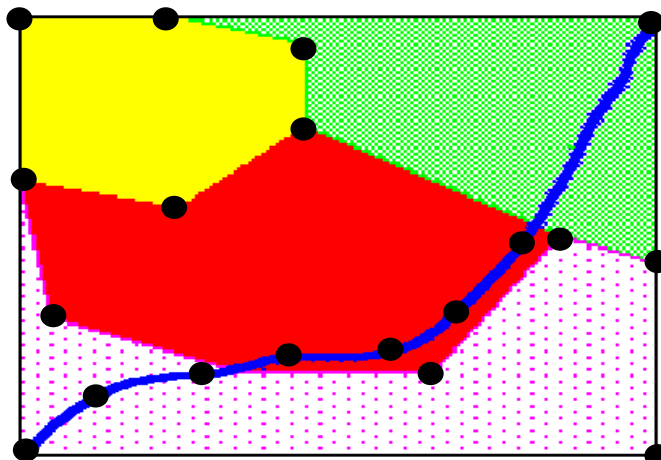


Figure 6: Point-Line (Coordinate Structure)

Disadvantages: Topology between the geometric objects must be explicitly defined, though it can be done quite efficiently. The file structures required are more complex than the raster data files, and layer overlay operations can be very complex to perform. Spatial variability can be represented, using a Triangulated Irregular Network, but it is still not as effective as the use of regularly gridded data, and mathematical operations, such as derivatives, on layers or between two or more layers are all but impossible to perform.

Uses: Very widely used in such fields as computer cartography, analysis of networks, municipal databases that contain descriptions of building footprints, streets, etc. Common GIS packages that are vector-oriented include ARC/GIS and MapInfo.

29.4.3 Object-oriented Models

Also called semantic models, object-oriented models organize geographic objects into different classes, on both a general level and to more specific levels. The more specific classes inherit certain properties from their "parent" class. For example, a class called "wetland" could be a parent class of "bog", "marsh", "swamp", and "lake". Each of the subclasses would inherit properties such as area, perimeter, and streams that drain into it, from the parent class. Advantages: All data pertaining to each object are encapsulated within the definition of the object, which protect them better from external tampering. Objects are a more natural way of looking at spatial data and are easier to conceptualize. Disadvantages: They are quite complicated to set up, and the theory behind them is rather difficult for the novice to get a grip on.

Uses: Not widely used at the moment. SYSTEM 9, which has had work done on it here at University of Toronto, and TIGRIS, are two GIS that use object-oriented models.

Objects in space are typically represented as three distinct spatial elements:

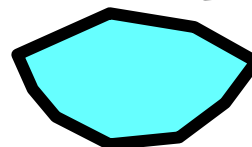
Points - simplest element



Lines (arcs) - set of connected points



Polygons - set of connected lines



These three spatial elements are represented to represent real world features and attach locational information to them.

Attributes

- In the raster data model, the cell value (Digital Number) is the attribute. Examples: brightness, landcover code, SST, etc.
- For vector data, attribute records are linked to point, line & polygon features. Can store *multiple* attributes per feature. Vector features are linked to attributes by a *unique feature number*.

Key Functions of GIS

In GIS the data can be used in many ways to get useful and meaningful information as per requirement of the user. Few of the functions that GIS can perform with the data are enumerated as follows:

1. Data can be positioned by its known spatial coordinates.
2. Data can be input and organized (generally in **layers**).
3. Data can be stored and retrieved.
4. Data can be analyzed (usually via a Relational DBMS).
5. Data can be modified and displayed

Various GIS Functionality includes:

- Data Assembly
- Data Storage
- Spatial Data Analysis and Manipulation
- Spatial Data Output

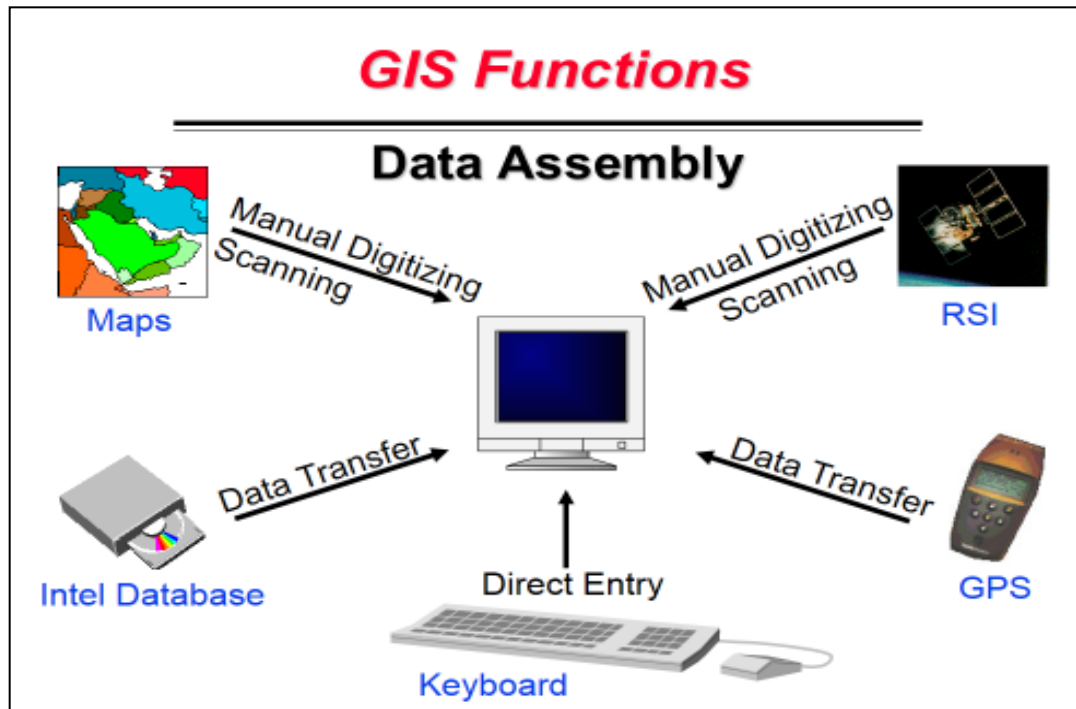


Figure 7: GIS functions

29.5 Major Areas of Practical Application of GIS Technology

Given below are the major potential areas of applications where GIS has been practically used and is being employed to get better information leading to policy planning and decision making.

1. Street Network-Based

- address matching
- vehicle routing and scheduling
- location analysis, site selection
- development of evacuation plans

2. Facilities Management

- locating underground pipes, cables
- balancing loads in electrical networks
- planning facility maintenance
- tracking energy use

3. Land Parcel-Based

- zoning, subdivision plan review
- land acquisition
- environmental impact statements
- water quality management
- ownership of maintenance

4. Natural Resource-Based

- forest management
- wildlife habitat, migration routes management
- wild and scenic rivers preservation
- recreation resources planning
- floodplain management
- wetland preservation
- agricultural lands management
- groundwater modeling and contamination tracking
- environmental impact analysis
- viewshed analysis

Conclusion

Applications of GIS have great social and national relevance and can support activities of government, enable enterprises to better manage business processes and bring important geographical knowledge to citizens. GIS applicability in varied areas like agriculture, forestry, land use, urban planning, environmental planning, etc has made it emerged as the potential tool for better planning and policy making as well as its implementation. Thus, GIS has considerable impact on the economies of local, regional, and national governance and development - by creating greater efficiency in information understanding, more visual communication for better comprehension of information and better decision making by information integration.

CROP YIELD ESTIMATION USING GEO-INFOMATICS**K. N. SINGH****Indian Agricultural Statistics Research Institute, New Delhi-110012****30.1 Introduction**

The soil fertility changes occur due to cropping, manure and fertilizer applications. Soil test results of one farm need to have scope to be connected with the broader population of all farms in a given area. But we may not be able to sample each farm in the population, because it is too costly, troublesome and time consuming, especially with the multiple small farm holdings as in India. We thus need to generalize results over an entire area. For the periods between 1975 to 1980, soil fertility maps for nitrogen (N), phosphorous (P) and potassium (K) were prepared using soil test data generated by soil testing laboratories functioned throughout the country (Ghosh and Hasan, 1979). Till date there is no major up-gradation in these maps. Singh et al. (2004) used point estimates for districts to prepare soil fertility maps of N, P and K for the states of Andhra Pradesh and Maharashtra. Further, Singh et al. (2006) have interlinked fertilizer recommendations for targeted yields of crops with these maps. Soil fertility maps have been prepared for 12 agriculturally important states using Soil Index Values (IV) for each district. Index values were calculated using standard procedure (Biswas and Mukherjee, 1987).

30.2 Methodology

The IVs were classified in to three categories viz. (Low 0- 1.5, Medium 1.5-2.5 and High >2.5). Soil Test Crop Response (STCR) approach was used to prescribe optimum doses of nutrients, based on available soil nutrients. From available nutrient Index Values and STCR equations the backward calculation for soil test values (STV) were obtained as follows:

Low	:	0.0 - 1.5	::	0-a (a>0)
Medium	:	1.5 - 2.5	::	a-b (b>a)
High	:	>2.5	::	>b

If $IV \leq 1.5$

$$STV = a \times (IV) / (1.5)$$

If $IV > 1.5$ and ≤ 2.5

$$STV = a + [(b-a) \times (IV - 1.5)]$$

If $IV > 2.5$

$$STV = (b/2.5) \times IV$$

Where a and b were positive coefficients used for describing the range of different nutrients. The a and b values depend on soil characteristics and are different for different soils. These denote the fertility of a soil with respect to N, P or K and are determined through soil test crop response correlation experiments. If a soil sample has available nutrient (N, P or K) below 'a' that means it is low, between 'a' and 'b', it is medium and above 'b', it is high. The district wise index values have been assigned from the database generated on N, P and K index values to the corresponding district layer of the state in GIS and generated the thematic maps accordingly.

The calculated soil test values were incorporated into the fertility maps to prescribe nutrients for targeted yields. This online application Software was developed to recommend fertilizer doses for the targeted yield at the District level. This system has the facility to input actual soil test values at the farmer's fields to obtain optimum doses. The application is a user-friendly tool to help the farmer in improving the efficiency (appropriate dose) of fertilizer use to achieve a specific crop yield.

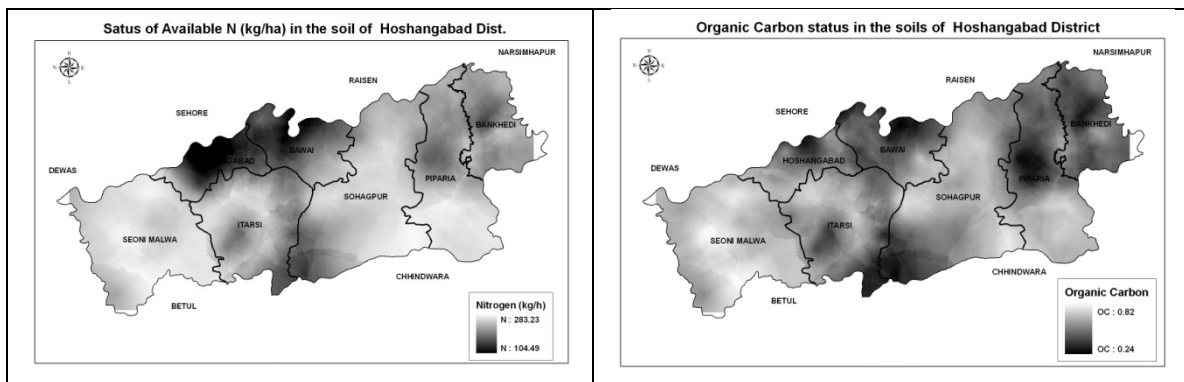
Remote Sensing, Geographic Information Systems (GIS), and the Agricultural Non-point Source Pollution (AGNPS) model have been used to assess runoff and sediment yield from various sub-watersheds above Cheney Reservoir in Kansas, USA (Bhuyan et al. (2002)). Ray and Dadhwal (2001) used satellite-based remote sensing data and GIS tools for estimating seasonal crop evapotranspiration in Mahi Right Bank Canal (MRBC) command area of Gujarat, India.

The recent technologies like GIS and GPS thus have much to offer for preparing soil fertility maps. Once the soil fertility maps are created, it is possible to transform the information from Soil Test Crop Response models into Spatial fertilizer recommendation maps. Such maps provide site-specific recommendation, validation for soil fertility over the following years. The fertilizer doses for targeted yield can be prescribed to the

farmers by locating his field/ area on the map with the help of latitude/longitude information.

To cover complete district Stratified Multistage Stratified Random Sampling has been adapted. To select the soil samples from different categories (big, small and marginal) of farmers it was essential to select farmers and to select farmers first to select village which is first stage unit. To select villages from a tehsil Simple Random Sampling without Replacement (SRSWOR) has been used. There is problem of spatial estimation, sometimes called spatial prediction. This arises in case a spatial field is partially observed at selected sites and the goal is to infer the field at unobserved sites. An example of spatial random field is soil nutrient concentrations over an agricultural domain. Among different methods of spatial interpolation of soil properties, kriging is an optimal interpolation method (Issak and Srivastava, 1989). To select the best model Akaike's (1973) information criterion (AIC) has been used.

In case of N, spherical method had the least AIC value. Hence for N, spherical Variogram method was used for kriging. Similarly linear, spherical, exponential, linear and linear Variogram methods of kriging were used for P, K, OC, EC and pH respectively.



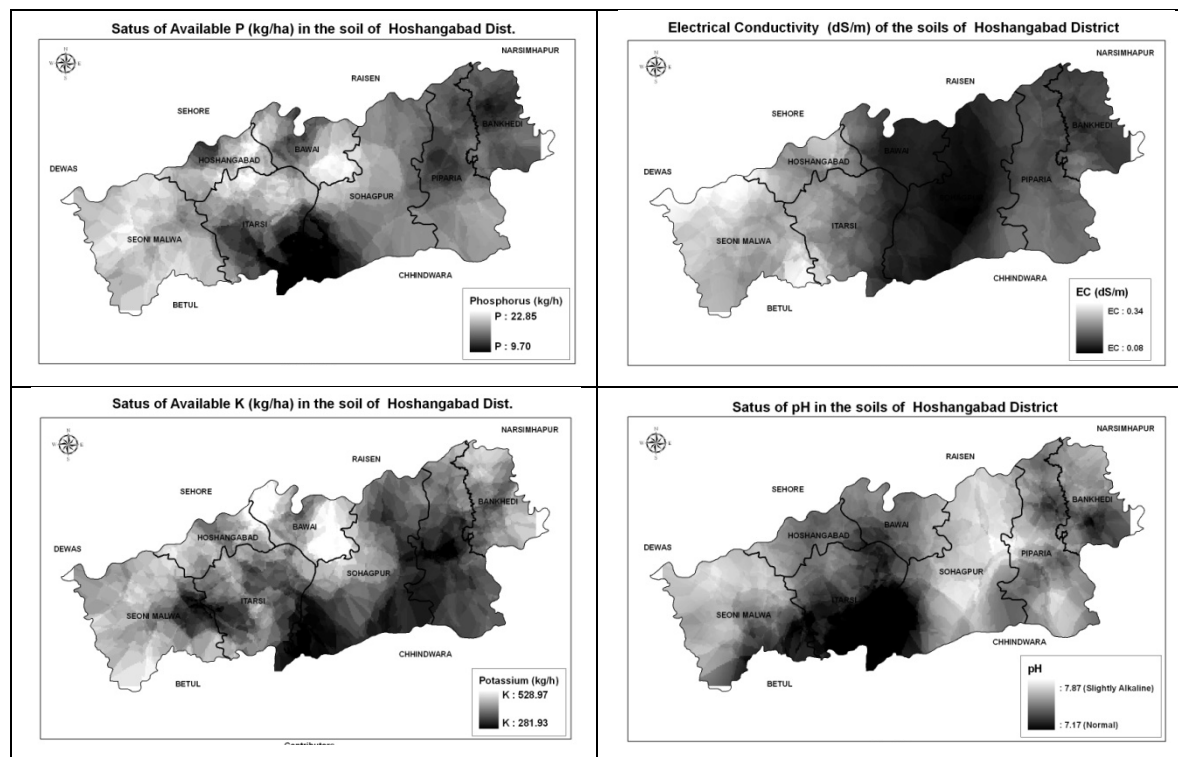


Fig 1. Kriged raster images (response surface) of different soil nutrients of Hoshangabad district.

Estimated response surface (Fig. 1) clearly showed that in Hoshangabad district OC in soil ranged between 0.28% to 0.81%, available soil N was in the range of 104 to 279 kg/ha, available soil P was in the range of 10 to 22.9 kg/ha, available soil K was in the range of 282 to 529 kg/ha. The EC was in the range of 0.08 to 0.34 desi siemens (dS/m) and pH was in the range of 7.2 to 7.9. With the help of these raster images all the ground points (pixel) was assigned with unique estimated value of respective nutrients. It was observed that calculated Abs (t) was less than that of tabulated t (for $P < 0.05$) for all the nutrients in 2007. This showed that in subsequent year there was no significant change in these nutrients. The results of year 2008 showed that only pH changed. For other nutrients there was no significant difference. Therefore, it is inferred that observed soil parameters for Hoshangabad district did not change significantly for at least two consecutive years except for pH. Again a web based on line spatial fertilizer recommendation system has been developed where farmers can get information up to field level if he has knowledge of

Longitude and Latitude, otherwise all the villages have been included in the system and village wise recommendation can be obtained.

The crop yield can be forecasted using the equations of the form:

$FN=3.92T-0.46SN$, $4.26T-0.59SN$, $3.47T-0.37SN$, $4.00T-0.44SN$, $3.78T-0.48SN$, etc.

$FP_2O_5=2.61T- 2.45SP$, $2.35T-3.16SP$, $2.53T- 2.12SP$, $2.32T-2.09SP$, $2.39T-2.90SP$, etc.

$FK_2O =2.47T-0.25SK$, $1.89T-0.20SK$, $2.12T-0.20SK$, $1.82T-0.17SK$, $1.24T-0.12SK$, etc.

Where,

FN , FP_2O_5 , FK_2O are fertilizer applied..

SN , SP , SK are soil test values for N, P and K and T is forecast yield (qt/ha)

Singh et al. (2006) utilized remote sensing data for preparing land productivity maps using simple linear relationship between Normalized Difference Vegetative Index (NDVI) values and Land Productivity Index (LPI) values. Satellite data for selected areas of Hoshangabad and Guna Districts have been used to obtain relationship between NDVI values and soil nutrients (Singh et al. (2009)). To obtain frequency data for each nutrient polygon to carry out statistical analysis the union of polygons was performed. Relationships between nutrients and NDVI values have been obtained for different months. The results indicate satisfactory relationship between nutrients and NDVI values. There is a good agreement between available nitrogen and maximum of maximum of NDVI values and it is best in the month of December and February. Available phosphorus can be estimated using average of average of NDVI values in the month of February and Potassium can be estimated using average of average of NDVI in the month of December.

REFERENCES

1. Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaksi, editors), Akademiai Kiado, Budapest, Hungary, Pp. 267-281
2. Biswas T. D. and Mukherjee S. K. (1987). Text Book of Soil Science. Tata Mc Graw-Hill Publishing Company Limited, New Delhi. P.193.
3. Bhuyan, SJ; Marzen, LJ; Koelliker, JK; Harrington, JA Jr.; Barnes, PL (2002). Assessment of runoff and sediment yield using remote sensing, GIS, and AGNPS. *Journal of Soil and Water Conservation Ankeny*. 57(6), 351-364.
4. Ghosh, A. B. and Hasan, R. (1979). *Bulletin of Indian Society of Soil Science*, 12, 1-8.
5. Issak, E. H. and Srivastava, R. M. (1989). An introduction to Applied Geostatistics, Oxford Univ. Press, New York, p.561
6. K. N. Singh. N. S. Raju and A. Subba Rao (2006). Land Productivity Assessment using Remote Sensing (RS) and Geographic Information System (GIS). *Indian Journal of Agricultural Sciences*, 76 (2) 81-84.
7. K. N. Singh, N. S. Raju, A. Subba Rao, Abhishek Rathore, Sanjay Srivastava, R. K. Samanta and A. K. Maji (2006). Prescribing optimum doses of nutrients for targeted yield through soil fertility maps in Andhra Pradesh (AP). *Jour. Ind. Soc. Agril. Stat.* 59(2): 131-140.
8. K. N. Singh, Abhishek Rathore, A. K. Tripathi, A. Subba Rao, Salman Khan and Bharat Singh (2009). Use of geographic information system, remote sensing and global positioning system in the application of precise fertilizer to maintain soil productivity of the farmers fields. *New Technology for rural development having potential of commercialization*. Allied Publishers Pvt. Ltd., New Delhi. PP 183-195
9. Ray, S. S., Dadhwal, V. K. (2001). Estimation of crop evapotranspiration of irrigation command area using remote sensing and GIS. *Agricultural-Water-Management*. 49(3), 239-249.

10. Singh, K. N., Raju, N. S., Subba Rao A., Srivastava Sanjay and Maji A. K. (2004). GIS based system for prescribing optimum dose of nutrients for targeted yield through soil fertility maps in Andhra Pradesh (AP). In the Proceedings of National workshop on recent trends in earth resources mapping, MANIT Bhopal. pp. 67-72.
11. Singh, K. N., Raju, N. S., Subba Rao A., Srivastava Sanjay and Maji A. K. (2004). GIS based system for prescribing optimum dose of nutrients for targeted yield through soil fertility maps in Maharashtra. In the Proceedings of National Seminar on Information and Communication Technology for Agriculture and Rural Development NAARM, Hyderabad. pp. 167-174.

Microsoft Excel

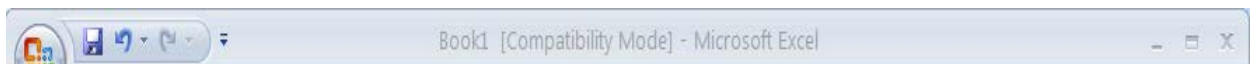
S. B. Lal

Indian Agricultural Statistics Research Institute, New Delhi-110012

31.1 WHAT IS A SPREADSHEET

Excel is a spreadsheet program. A **spreadsheet** is a grid of rows and columns that helps organize, summarize, and calculate data. Spreadsheets are an everyday part of many professions, including accounting, statistical analysis, and project management. You can use Excel to create business forms, such as invoices and purchase orders, among many other useful documents.

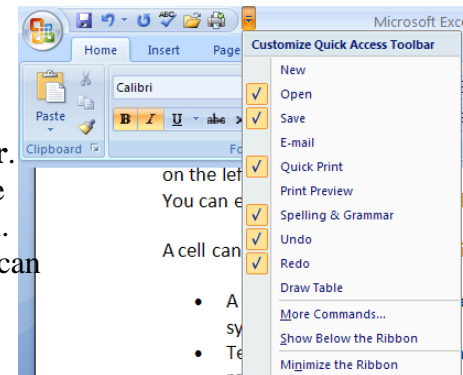
To open Microsoft Excel click on **Start, All Programs, and Microsoft Excel**. Let's look at the toolbars.



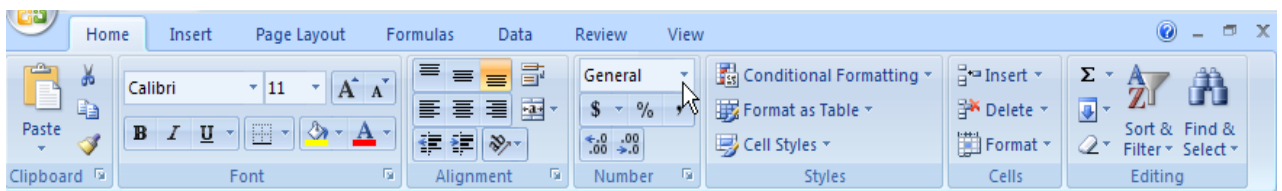
This is the **Title Bar**. It gives the name of the program and the title of the workbook you are using. Since we have just opened up a new workbook and have not saved it with a name, the default title is Book1.



The **Office Logo Button** is located at the upper left of the **Title Bar**. Clicking on it will open a dropdown menu that contains many of the menu items that used to appear under **File** in prior versions of Excel. Right beside the **Office Button** is the **Quick Access Toolbar**. You can Add or subtract commands to the toolbar by simply clicking/ on them in the list.



Next we have the **Ribbon**. The **Ribbon** has seven **Tabs** that give instructions to the software. The **Ribbon Tabs** begin with **Home** and continue with **Insert, Page Layout, Formulas, Data, Review, and View**. On the right-hand end, there is an icon for the Help Menu, Minimize, Restore Down, and Close.



Clicking on one of these Tabs will open the **Group**. The **Group** that belongs to each **Tab** shows related Command items together. You may then choose a Command.

31.2 Workbooks and Worksheets

When Excel is opened, a workbook appears with three worksheets. Each worksheet contains columns and rows. There are **1,048,575 rows and 16,384 columns**. The combination of a column coordinate and a row coordinate make up a cell address. For example, the cell located in the upper left corner of the worksheet is cell A1, meaning column A, row 1. The cell address is visible in the **Name Box**.

Place your cursor in the first cell, A1. The formula bar will display the cell address in the **Name Box** on the left side of the **Formula** bar. Notice that the address changes as you move around the sheet. You can easily move from cell to cell by pressing tab or using the arrow keys.

A cell can contain any of the following:

- A number (and any associated punctuation, such as decimal points, commas, and currency symbols).
- Text (including any combination of letters, numbers, and symbols that aren't number-related).
- A **formula**, which is a math equation.
- A **function**, which is a named equation that shortcuts an otherwise complex operation.

Creating a New Workbook

It is easy to create a new workbook! Simply, click on *Office Button – New* and click on *Blank Workbook* to create a new workbook.

Creating a New Worksheet

Creating a new worksheet is just as easy. By default, each Excel workbook contains three worksheets. Three tabs displaying *Sheet 1*, *Sheet 2*, and *Sheet 3* will be displayed at the bottom of the workbook to indicate the separate sheets. To add a new worksheet, simply click on the tab after the tab that says Sheet 3.

Navigating and Selecting

Moving around a worksheet is easy! You can easily move from cell to cell by using the arrow keys or pressing tab (will move the cursor to the right) or shift-tab (shift-tab will move you to the left). You can also use your mouse to click within a cell which will select that cell. Sometimes you will want to select a **range** of cells.

A **range** is a group of one or more cells. If you select more than one cell at a time, you can then perform actions on the group of them at once, such as applying formatting or clearing the contents. A range can even be an entire worksheet.

A range is referenced by the upper left and lower right cells. For example, the range of cells B1, B2, C1, and C2 would be referred to as B1:C2.

To select a range:

- **With the mouse:** Drag across the desired cells with the left mouse button held down. Be careful when you're positioning the mouse over the first cell (before pressing the mouse button). Position the pointer over the **center** of the cell, and not over an edge.

If you drag while the pointer is on the edge of the cell, Excel interprets the selection as a move operation and whatever is in the cell(s) is dragged to a different spot.

- **With the keyboard:** Select the first cell, and then hold down the **Shift** key while you press the arrow keys to expand the selection area.

To select a nonrectangular or noncontiguous range, select the first portion of the range (that is, the first rectangular piece), and then hold down the **Ctrl** key while you select additional cells/ranges with the mouse.

To select an entire column, click the column header (where the letter is). To select an entire row, click the row header (where the number is). You can click one row or column and then drag to select additional columns, or hold down **Ctrl** as you click on the headers for noncontiguous rows and/or columns.

Entering and Editing Data

Let's learn how to enter data into your worksheet. First, you place the cursor in the cell in which you would like to enter data. Then you type the data and press Enter.

You can also edit information in a cell by double-clicking in a cell or by clicking in the formula bar. Try these two options.

Inserting Columns and Rows

If you don't plan your worksheet layout correctly, you might end up with too many or too few rows or columns in a certain area. You can always move data around in the sheet to help with this, but sometimes it's easier to simply insert or remove columns or rows.

Formatting Columns and Rows

Often you will need to change your columns and rows in order for text to fit or for the text to fit on the page correctly. There are a number of different methods one can use to do this. Let's start with columns.

Column Width: The formatting that is unique to columns is **Column Width**. **Column Width** is measured in characters. A column's width can be from 0 to 255 characters, which is a **really** wide column! Decimal values are allowed. In fact, the default size is 8.43 characters.

A width of 12, for example, means the column is wide enough for 12 average characters, using whatever you chose as the Standard font. The default is Calibri 11 pts. To change the font from the default, go to *Tools-Options-General-Standard font*.

Column Width

Be careful when you set a column's width with *AutoFit*. The column may wind up wider than you expected. Any text will be on a *single line* in its cell. No matter how long the text is! If you accidentally find you've widened a cell out of sight to the right, use *Undo*. (my favorite button!) Then resize the column with another method.

Column Width - Drag

Dragging is a natural method of adjusting column width. But since you can't see the change until you release the mouse button, it may take you several attempts to get a satisfactory width.

Row Height

The only unique formatting for rows is *Row Height*. *Row Height* is measured in points, like font size, from 0 to 409 points. A row height of zero hides the row.

The default setting for **Row Height** is *AutoFit*. The row height adjusts to the largest font size in the row.

AutoFit will leave a little white space, called the cell padding, between the text in the cell and the cell edges. When Arial 10 pt. is the Standard Font, the *Row Height* is **12.75 points**. You may find that this looks a bit crowded when the gridlines are shown. If you don't print the gridlines, your paper version will look OK.

Moving to a New Worksheet

In Microsoft Excel, each workbook is made up of several worksheets. Before moving to the next topic, let's move to a new worksheet. You can move from worksheet to worksheet by clicking on the tabs at the bottom of the worksheet. Let's move to **Sheet 2**.

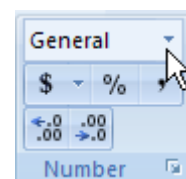
Formatting Text and Data

Once information has been entered into a cell, you might want to change or enhance the way the information is displayed. Text can be formatted in the same way that one uses in Microsoft Word or PowerPoint. Most of the formatting choices can be found in the **Font** grouping under the **Home** tab. There are numerous ways to format data. Let's look at some. First remember to always make sure that the cell you want to format is selected.

Using Formatting Buttons – On the **Ribbon**, make sure the Home tab is selected. In the **Number Group** box, there are several buttons which allow one-click formatting.

Notice how each number changes depending on the formatting.

Formatting Numbers



Let's look at other formatting options.

After formatting

Let's change it to a dollar amount.

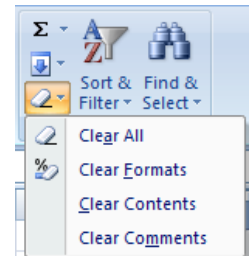
1. Make sure that the cursor is in cell A5.
2. Right-click again.
3. Click on *Format Cells*.
4. Click Currency in drop down menu.
5. Look at the options available including currency symbols.

Deleting vs Clearing a Cell

Many beginners get confused about clearing versus deleting in Excel, so let's look at this concept briefly. When you clear the content from a cell, the formatting for that cell is still there. It may be helpful to think of an Excel worksheet as a stack of empty cardboard boxes, each one with its open side facing you. You can put something into a cell or take something out. When you take something out of a cell, it's called clearing its content. The cell itself remains in the "stack," but it's now empty.

To clear the content from a cell:

1. Press Delete on the keyboard.
2. Right-click the cell and then select Clear Contents.
3. On the Home tab, in the Editing group, select Clear > Clear Contents.



Unfortunately, clearing a cell's content doesn't clear its formatting.

To clear formatting:

1. On the Home tab, in the Editing group, select Clear > Clear Formats
2. To clear both contents and formats at once, select Clear All.

In contrast, deleting the cell removes the cell itself from the stack and makes the surrounding cells shift. Think about what happens when you pull a box out of a stack of boxes -- the boxes above it fall down one position, right? It's the same thing with Excel cells, except it's reverse-gravity (cells fall up rather than down), and you have the choice of making the remaining cells shift up or to the left. Let's look at how this works.

Filling Cells Automatically

You can use Microsoft Excel to fill cells automatically with a series. For example, you can have Excel automatically fill in times, the days of the week or months of the year, years, and other types of series. Days of the week and months of the year fill in a similar fashion.

Filling Time

Merging Cells

Sometimes, rather than having text wrap in a cell, you will actually want the text to run across the width of the data. Usually when making a spreadsheet, you need to create a heading for

the sheet. This heading should run across the width of your data. To do this, one must merge the cells across the width of the data.

31.3 Performing Mathematical Calculations

Let's add a column of numbers using the **AutoSum** Button Σ . To select the **AutoSum** button choose **Home > Editing > Σ** and automatically add a column of numbers.

What's a formula?

A **formula** is an equation that performs some type of operation and issues a result. In Excel, formulas always begin with an equal sign. Here are some formula examples:

- **=2+6:** This formula is strictly math. If you place this formula in a cell, the cell displays **8**.
- **=A1+6:** Same as the preceding, but this time you're adding 6 to whichever value is in cell A1 and displaying the result in the cell into which you enter this formula. This formula does not change A1's contents.
- **=A1+A2:** Same thing again, but you're adding the contents of cell A1 to the contents of cell A2.
- **=A1+A2-A3:** In this example, multiple cells are referenced.

Here are the symbols you can use in formulas to indicate mathematical operations:

- **+: Addition**
- **-: Subtraction**
- ***: Multiplication**
- **/: Division**

More Formula Examples

The math operators in Excel have an order of operation, just like in regular math. The order of operation is the order in which they're processed when multiple operators appear in the same formula. Here are the rules that determine the order:

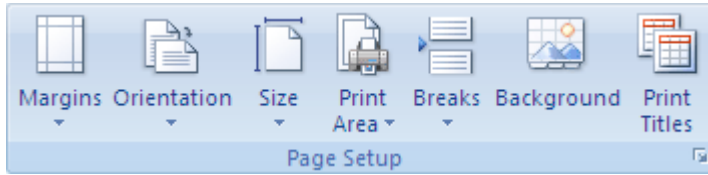
1. Any operations that are in parentheses, from left to right
2. Multiplication (*) and division (/)
3. Addition (+) and subtraction (-)

Parentheses override everything and go first. So, if you need to execute an operation out of the normal order, you place it in parentheses. Now let's try some formula examples that refer to cells and use math operations. For this exercise, enter the following values in cells in a blank worksheet:

A1: 12 A2: 6 A3: 4 A4: 9

Printing

Let's prepare to print! If your worksheet is more than one printed page, it is possible to have the heading on each page by going to the **Page Layout** tab, in the **Page Setup** group and click **Print Titles**.



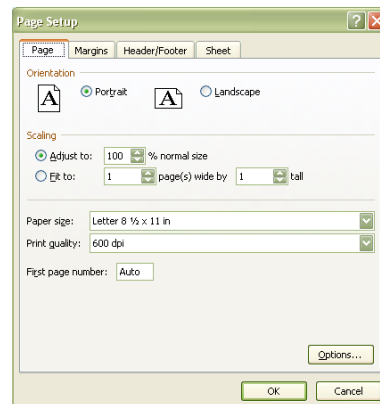
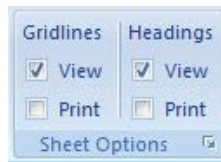
On the **Sheet** tab, under **Print Titles**, do one or both of the following:

In the **Rows to repeat at top** box, type the reference of the rows that contain the column labels if you want the heading repeated on each page.

In the **Columns to repeat at left** box, type the reference of the columns that contain the row labels if you want those to show.




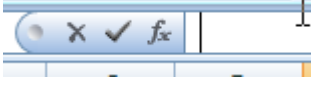
We want our sheet to print with no gridlines, and centered horizontally across the page, but not vertically. Let's go the *Page Layout > Sheet Options*. There should not be a check under **Print** in the Gridline section.

Make sure that you have checked your spelling and made any necessary corrections. Click on the **Office Button** and **Print>Print Preview** (Always do a print preview in Excel!). Click on *Page Setup>Margins* and make sure that there is a check under **“Center on Page”** > **horizontally**. Now let's print!!!



Recognizing Cursor Styles

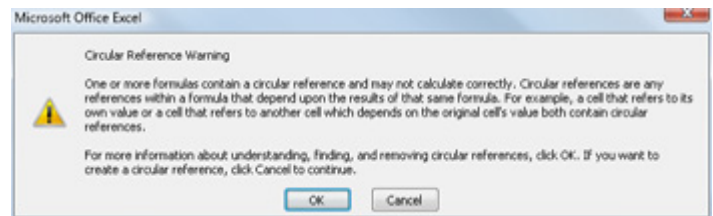
There are four common cursor styles used in Excel.

	<p>Click and drag to highlight multiple cells with this cursor, or click in a cell to select the single cell</p>		<p>Click and drag with this cursor to fill cell contents into cells below or to the right.</p>
	<p>Click and drag the contents of the selected cell to any other cell.</p>		<p>Click to place the cursor into the Formula bar so that you can edit an equation or function.</p>

Common formula errors

Here are some of the most common mistakes people make when entering formulas and functions:

- **Not putting in all the required arguments:** If a function is expecting more arguments than you have entered, and you get a dialog box, be sure you've placed commas between the arguments and that you haven't overlooked any.
- **Circular references:** If you refer to the cell's own address in a function, you create a circular error, which is like an endless loop. Suppose that you enter `=A1+1` into cell A1. You'll get an error message like the one below. If you click **OK** at this message, a Help window appears to help you find the problem.
- **Text in an argument:** Most functions require numeric arguments. If you enter text as an argument, for example, `=SUM(text)`, the word `#NAME?` appears in the cell. This happens because Excel allows you to name ranges of cells using text, so technically `=SUM(text)` isn't an invalid function. It is invalid only if there's no range that has been assigned the name "text."
- **Hash marks (###) in a cell:** This happens when the cell isn't wide enough to display its value. Widen the column to fix this.



If you receive an error when copying a formula, don't panic; it happens to everyone. Use the skills you learned earlier in this chapter to display the formulas and then check them for the common errors discussed here.

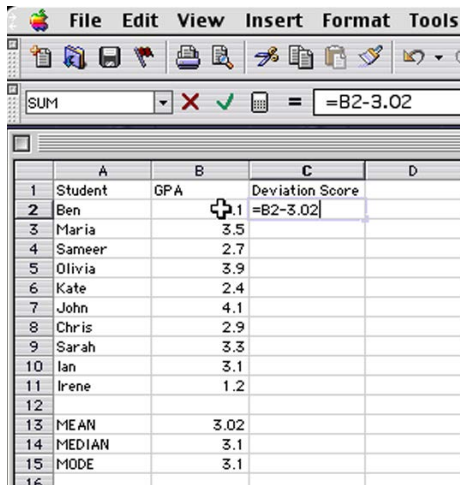
31.4 Using functions in Excel

Let us see the use of Excel to get sample statistics and how to make a histogram. You will learn how to use Excel formulas and take advantage of the formula lookup function.

There are two ways to start Excel. The first is to open an Excel file (extension .xls). The second is to open the program directly from the menu bar. Open Excel and type the data below into the spreadsheet. Save the file as the “testdata” in a folder.

	A	B
1	Student	GPA
2	Ben	3.1
3	Maria	3.5
4	Sameer	2.7
5	Olivia	3.9
6	Kate	2.4
7	John	4.1
8	Chris	2.9
9	Sarah	3.3
10	Ian	3.1
11	Irene	1.2
12		
13		

1. Get the average of these scores by typing this formula into the cell B12: “=average(B2:B11)” and hitting enter.
2. Get the median of the scores. This time use the “Insert” -> “Function” menus to select the median function. These menus can be helpful when you don’t remember the exact command for a function.
3. Get the mode. Can you guess the formula?
4. Calculate a column of deviation scores in column C. First, type the formula “=B2-3.02” into cell C2. Then highlight the cell, click “copy”, highlight the rest of the column, and click “paste.”
5. Calculate a column of squared deviation scores in column D.



A	B	C	D	E
Student	GPA	Deviation Score	Squared Deviation	
Ben	3.1	0.08	=(C2)^2	
Maria	3.5	0.48		
Sameer	2.7	-0.32		
Olivia	3.9	0.88		
Kate	2.4	-0.62		
John	4.1	1.08		
Chris	2.9	-0.12		
Sarah	3.3	0.28		
Ian	3.1	0.08		
Irene	1.2	-1.82		

6. You can then get the sum of squared deviations, the variance, and the standard deviation from column D.

In cell D13 put “=sum(D2:D11)”
 In cell D14 put “=D13/9”
 In cell D15 put “=sqrt(D14)”

Make sure you understand *why* those formulas yield the SS, variance, and standard deviation.

7. Of course, there is a faster way to calculate the SS, the variance, and the standard deviation directly using Excel commands for these statistics. See if you can do this using the “Insert->Function” menus. (Hint: look for sums of squared deviations under D not S).
8. What if you wanted to get other measures of position for each score? (You’ve already got deviation scores). Let’s get percentiles for each score. In E2, type this formula: “=percentrank(\$A\$2:\$A\$11, A2)”. The range of cells before the comma tells Excel which cells contain the full dataset. The cell after the comma tells Excel the value for which you’d like to get a percentile. The dollar signs tell Excel that when you copy and paste the formula, those cell references should remain absolute.
9. Let’s get z-scores for each score. Figure out two different ways to do this on your own (step by step using multiple columns and quickly using a single formula).
10. Show that the z-scores do not change when you add a constant to each score or multiply each score by a constant.
11. What if we wanted to look at the interquartile range? To do this, we need to identify the 25th and 75th percentiles. In (8) we went from scores to percentiles. Now we want to do the reverse. Somewhere in your worksheet, type “=percentile(A2:A11, .25)”. This will give you Q1 (the 25th percentile). Use the formula to get Q2 (50th percentile or median) and Q3 (75th percentile).

Of course, in Excel there are multiple ways to accomplish the same thing. Check out the MEDIAN() and QUARTILE() functions.

12. How would the box-plot look for this data?

13. Use Excel to make a histogram

Click on Tools → Data Analysis. A new window should open, scroll down to Histogram, click OK. The dialogue box will ask you for an input range, a bin range, and other stuff. For the input range, highlight the column of datapoints in column A (or type the cell range directly). Leave the bin range blank for now – Excel will determine the bin sizes for you. Don't forget to check off "Chart output" so that it gives you a chart, that's the whole point. Hit OK and see what happens.

Get rid of the spaces between the bars by highlighting the chart, then double clicking on the bars themselves. A new window should open, use the tabs to choose "Options" and it will ask for overlap and gap width. *Overlap* is usually set to 0 and *Gap width* is usually set to 150. Change *Gap width* to zero. Resize the histogram by clicking on the white area surrounding the chart, then dragging one of the corners down.

How many intervals would be best for this data? A number that neither hides too much information nor presents too much information. Too many intervals produce a histogram that is cumbersome and a poor summary of the data; too few produce a histogram that fails to summarize because it loses too much important detail. Remember: it's a judgment call.

To find the optimal bin width, consider the range of data by sorting (highlight the data, go to the Data menu, and choose Sort). Check out the minimum and the maximum. Think about what intervals could be used for this data set. Take these points into consideration:

- all intervals should be the same width
- the entire range should be covered
- the bottom score of each interval should be a multiple of the interval width (e.g. if the interval width is 5 You should not start at 12 but rather at 10)
- the interval should reflect the conventions of the research area and make sense (especially make sense)

Now, try setting the intervals yourself. To do this , you need to create a column that contains the values you want on your x-axis (the "bin bottoms"). Then find the Histogram dialogue box again, enter the input range, and then move the cursor to "Bin range" and highlight your bin column. Hit return. Did it work? Try again with different bins. What information are you losing? What are you gaining? Find an optimal bin interval and describe why you think it's so good.

Online Tutorials on Excel

1. HP Learning Center: <http://h30187.www3.hp.com/>
2. Microsoft: <http://office.microsoft.com/en-us/excel/FX100646951033.aspx>
3. Goodwill Industries: <http://www.gcfllearnfree.org/>
4. Baycon Group: <http://www.baycongroup.com/el0.htm>
5. Internet4Classrooms: http://www.internet4classrooms.com/on-line_excel.htm

ANALYSIS OF SURVEY DATA USING MICROSOFT EXCEL

Kaustav Aditya

Indian Agricultural Statistics Research Institute, New Delhi-110012

32.1 INTRODUCTION

The program *SampleCalc* (short for Sample Calculator) calculates unbiased point estimates and approximate confidence intervals of the principal population characteristics for each specified variable and each category of specified attributes. *SampleCalc* is a Microsoft Excel Add-In developed by Peter Tryfos, Professor of Management Science, Faculty of Administrative Studies, at York University, Ontario, Canada.

It is assumed that the observations are entered in an Excel worksheet and arranged in the form of a table, the columns of which correspond to the variables and attributes of interest and the rows to the sampled elements. The editing of the data, and the coding of the variables, attributes, and missing values, should take place before *SampleCalc* is used.

Before using *SampleCalc*, be aware that the following additional information is required by each sampling method. Make sure that this information is entered on the worksheet containing the observations or is otherwise available before *SampleCalc* is called.

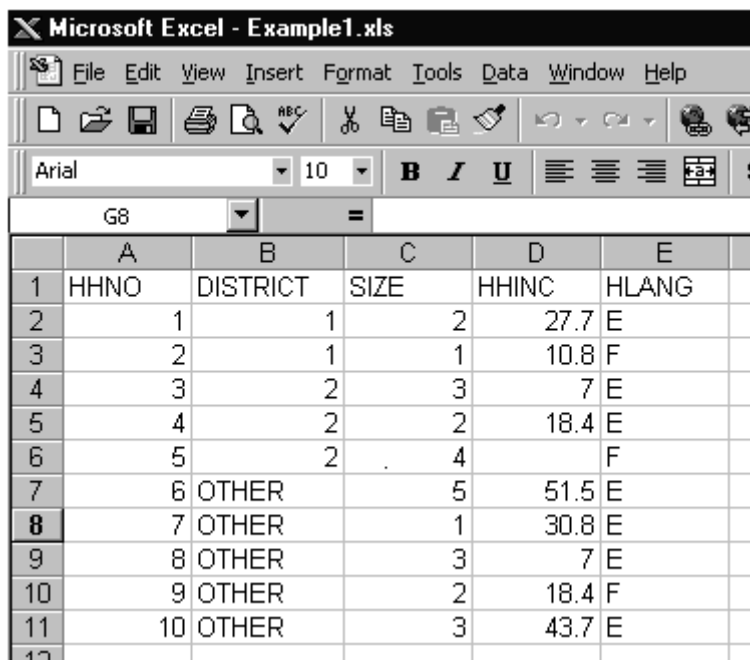
1. Simple Number of elements in the population
Confidence level
2. Stratified Labels identifying the groups (strata)
Number of elements in each group (stratum) in the population
Confidence level
3. Two-Stage Labels identifying the selected groups (strata)
Number of elements in each selected group (stratum) in the population
Number of elements in the population
Number of groups (strata) in the population
Number of selected groups (strata)
Confidence level
4. Cluster Labels identifying the selected groups (strata)
Number of elements in the population
Number of groups (strata) in the population

Number of selected groups (strata)

Confidence level

It should be noted that the confidence intervals generally require large samples; however, *SampleCalc* does not check that this requirement is satisfied. Also, the number of categories of an attribute cannot exceed 50.

The application of *SampleCalc* will be illustrated with the help of a simple example. It will be assumed that a random sample without replacement of 10 households was selected from the population of households in a city. The observations are entered in an Excel spreadsheet in the manner shown below



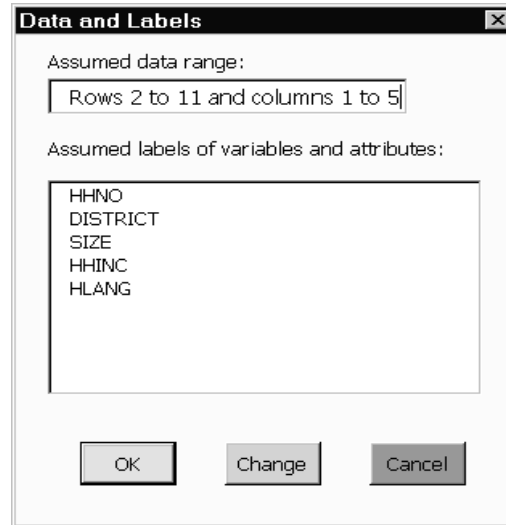
	A	B	C	D	E
1	HHNO	DISTRICT	SIZE	HHINC	HLANG
2	1	1	2	27.7	E
3	2	1	1	10.8	F
4	3	2	3	7	E
5	4	2	2	18.4	E
6	5	2	4		F
7	6	OTHER	5	51.5	E
8	7	OTHER	1	30.8	E
9	8	OTHER	3	7	E
10	9	OTHER	2	18.4	F
11	10	OTHER	3	43.7	E

The labels of the variables and attributes are entered in the first row. *HHNO* stands for household number, *DISTRICT* for the district identification (1, 2, or *OTHER*), *SIZE* for the number of persons in the household, *HHINC* for household income, and *HLANG* for language spoken at home (E: English, F: French). There is a missing value (an empty cell) in cell D6.

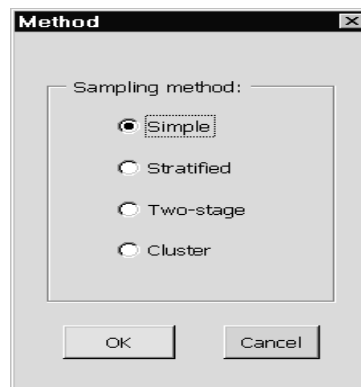
32.2 COMMON FIRST STEPS

Enter the sample observations (data) in the form of a compact table in an Excel worksheet. The rows must correspond to the sampled elements and the columns to the variables and attributes of the study. It is strongly recommended that the first row of this table contain the labels of the variables and attributes. A missing value should be represented either by a blank cell or one containing a period.

1. Click on any one cell of this compact table.
2. In the Tools menu, click on *SampleCalc*. A dialog box entitled **Data and Labels** will appear.



- If the displayed data range and labels are correct, click the OK button and go to Step 3.
 - If the displayed data range or the labels are not correct, or if the data do not have the recommended format, click the Change button. In the two ensuing dialog boxes, select the proper ranges for the data and the labels. The **Data and Labels** dialog box will reappear. Click the OK button and go to Step 3.
3. The **Method** dialog box now appears. Select the method by which the sample was selected and click either the OK button to proceed, or the Cancel button to abort *SampleCalc*.

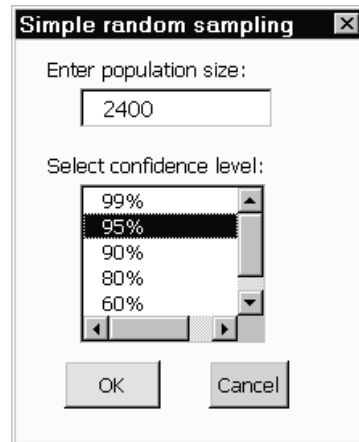


The subsequent conversation with *SampleCalc* depends on the sampling method selected. Refer to the section corresponding to the method by which the sample was selected.

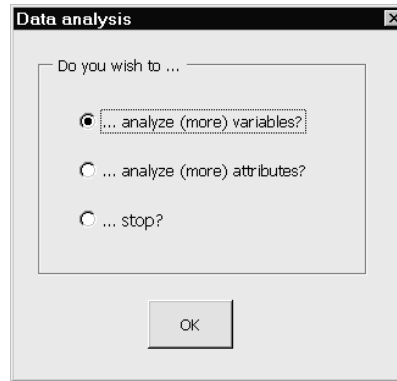
32.3 METHOD: SIMPLE RANDOM SAMPLING

To illustrate the application of SampleCalc, it will be assumed that the observations were selected by drawing a simple random sample without replacement of 10 households from among the 2400 households in the city.

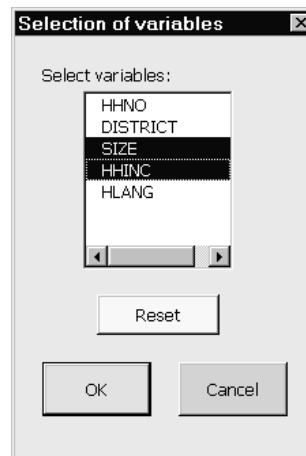
After selecting Simple in the **Method** box, the **Simple random sampling** dialog box appears.



- 1 In the dialog box entitled **Simple random sampling**, enter the number of elements in the population and select the confidence level. Then, click either the OK button to proceed, or the Cancel button to abort *SampleCalc*.
- 2 The program will display a message for your information to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK. (VRESULTS will contain the results of calculations concerning the selected variables, while ARESULTS will contain the results regarding the selected attributes.)
- 3 The dialog box entitled **Data analysis** will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. Click OK.



If you chose to analyze one or more variables, a dialog box entitled **Selection of variables** will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed

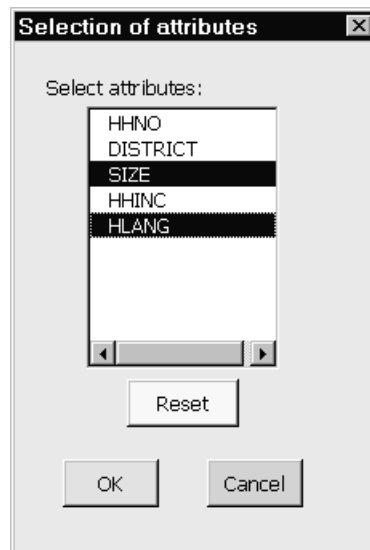


- 4 If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated variance of the variable, the estimated standard deviation of the estimate of the population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.

ANALYSIS OF SURVEY DATA USING MICROSOFT EXCEL

Summary of results for variables, simple random sampling										
95% confidence interval										
Label	No. Obs.	Est. Mean	Est. Total	Est. Varia	Est. StDo	Mean, low	Mean, up	Total, low	Total, upper limit	
SIZE	10	2.6	6240	1.44	0.399166	1.817635	3.382365	4362.324	8117.676	
HHINC	9	23.92222	57413.33	224.3528	5.28573	13.56219	34.28225	32549.26	82277.41	

If you chose to analyze one or more attributes, a dialog box entitled **Selection of attributes** will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed.



- If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.

Summary of results for attributes, simple random sampling										
95% confidence interval										
Label	Category	No. Obs.	Est. Prop.	Est. Num	Est. StDo	Proportio	Proportio	Number, I	Number, upper	limit
SIZE		2	10	0.3	720	0.152434	0.001229	0.598771	2.950668	1437.049
SIZE		1	10	0.2	480	0.133055	-0.06079	0.460788	-145.892	1105.892
SIZE		3	10	0.3	720	0.152434	0.001229	0.598771	2.950668	1437.049
SIZE		4	10	0.1	240	0.099791	-0.09559	0.295591	-229.419	709.419
SIZE		5	10	0.1	240	0.099791	-0.09559	0.295591	-229.419	709.419
HLANG	E		10	0.7	1680	0.152434	0.401229	0.998771	962.9507	2397.049
HLANG	F		10	0.3	720	0.152434	0.001229	0.598771	2.950668	1437.049

- 6 If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to leave *SampleCalc* and return to Excel. You may want to adjust the column widths of VRESULTS and ARESULTS in order to see the results more clearly. (If you wish to delete the worksheets VRESULTS and ARESULTS, right-click on the tab of the worksheet and select Delete from the pop-up menu.)

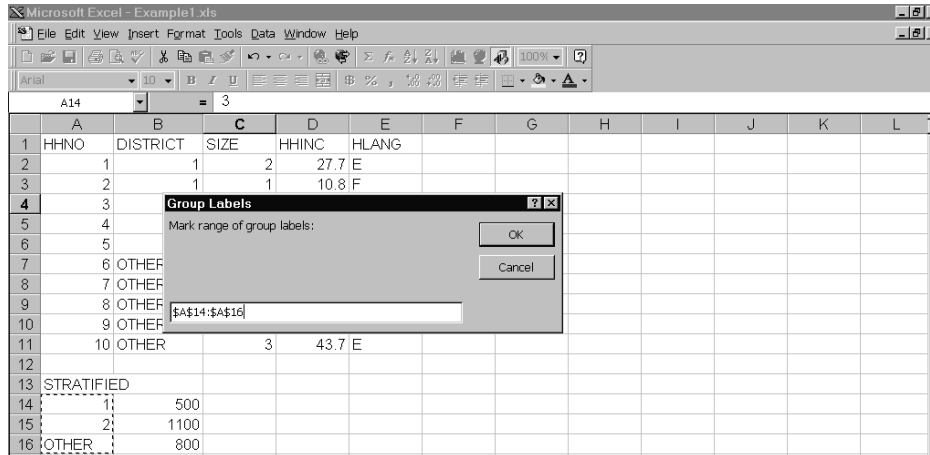
32.4 METHOD: STRATIFIED RANDOM SAMPLING

To illustrate the application of SampleCalc, it will be assumed that the city households are grouped into three districts and that the observations were selected by drawing a simple random sample without replacement of 2 households from among the 500 households in District 1, one of 3 households from among the 1100 households in District 2, and one of 5 households from among the 800 households in the district labeled Other. The labels and the number of households in each district are entered in the same worksheet as the sample observations.

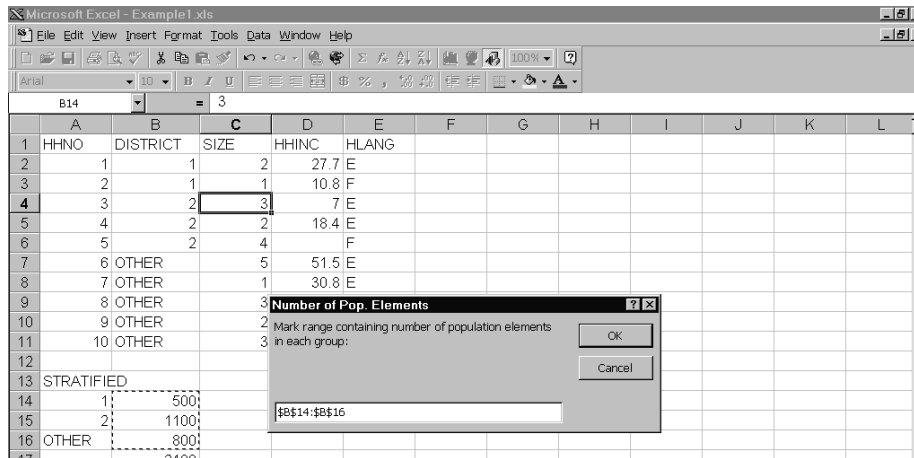
After selecting Stratified in the **Method** box, the **Group labels** dialog box appears.

- 1 In the dialog box entitled **Group labels**, select the range containing the labels of the groups (strata). Click OK to proceed, or Cancel to abort the program.

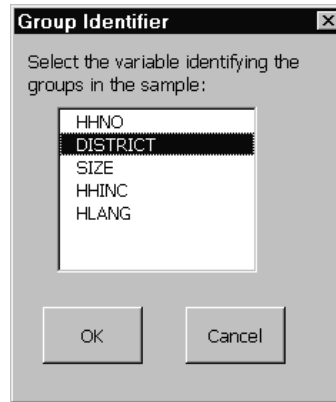
ANALYSIS OF SURVEY DATA USING MICROSOFT EXCEL



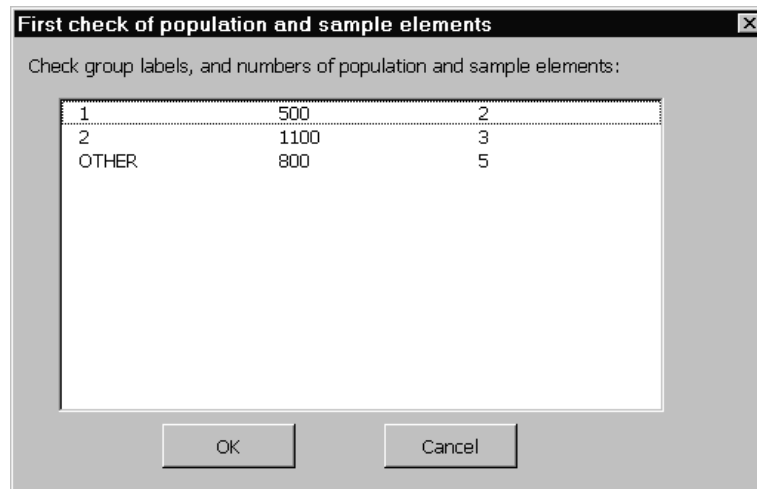
- 2 In the dialog box entitled Number of Pop. Elements, select the range containing the number of elements in each group (stratum) in the population. Click OK to proceed, or Cancel to abort the program.



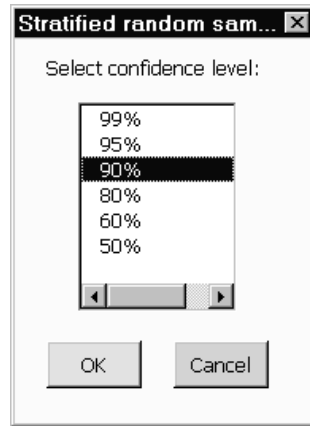
- 3 A dialog box entitled Group Identifier will now appear. Select the label of the column in the table of data that identifies the group (stratum) to which each sampled element belongs. Click OK to proceed, or Cancel to abort.



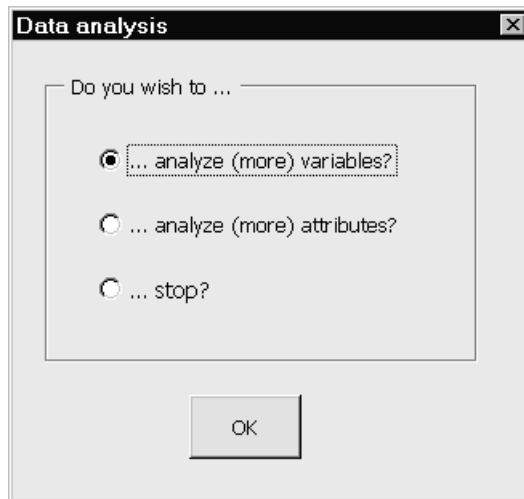
- 4 A dialog box entitled First check of population and sample elements will now appear. It shows the program's understanding of the group (stratum) labels, and of the corresponding number of elements in the population and sample. If you observe an anomaly, click Cancel to abort SampleCalc, check the data, and begin anew. If the program's understanding appears correct, click OK to proceed.



- 5 In the dialog box entitled Stratified random sampling, select the desired confidence level, and click OK to proceed or Cancel to abort.

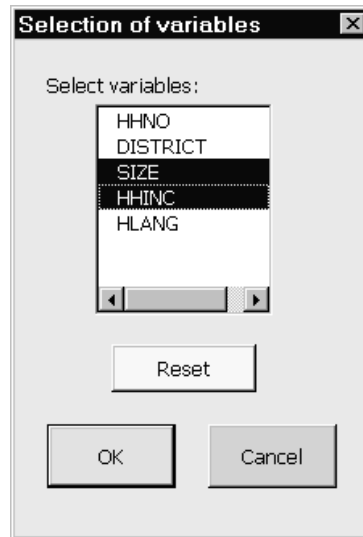


- 6 The program will display a message for your information to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK. (VRESULTS will contain the results of calculations concerning the selected variables, while ARESULTS will contain the results regarding the selected attributes.)
- 7 The dialog box entitled Data analysis will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. Click OK.



- 8 If you chose to analyze one or more variables, a dialog box entitled Selection of variables will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the

entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box.



- 9 If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated standard deviation of the estimate of the population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.

Summary of results for variables, stratified random sampling										
90% confidence interval 90% confidence interval										
Label	No. Obs.	Est. Mean	Est. Total	Est. Varia	Est. StDo	Mean, low	Mean, up	Total, low	Total, upper limit	
SIZE	10	2.620833	6290		0.359476	2.029496	3.212171	4870.79	7709.21	
HHINC	9	19.92458	47819		4.141546	13.11174	26.73743	31468.18	64169.82	

- 10 If you chose to analyze one or more attributes, a dialog box entitled Selection of attributes will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box. If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.

Summary of results for attributes, stratified random sampling										
90% confidence interval 90% confidence interval										
Label	Category	No. Obs.	Est. Prop.	Est. Numi	Est. StDo	Proportio	Proportio	Number, I	Number, upper	limit
SIZE	2	10	0.323611	776.6667	0.196218	0.000833	0.646389	1.999544	1551.334	
SIZE	1	10	0.170833	410	0.123385	-0.03214	0.373802	-77.1256	897.1256	
SIZE	3	10	0.286111	686.6667	0.172923	0.001653	0.57057	3.966343	1369.367	
SIZE	4	10	0.152778	366.6667	0.152569	-0.0982	0.403754	-235.677	969.0103	
SIZE	5	10	0.066667	160	0.066458	-0.04266	0.17599	-102.376	422.3762	
HLANG	E	10	0.676389	1623.333	0.196218	0.353611	0.999167	848.6662	2398	
HLANG	F	10	0.323611	776.6667	0.196218	0.000833	0.646389	1.999544	1551.334	

11 If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to return to Excel. You may want to adjust the column width of VRESULTS and ARESULTS in order to see the results more clearly

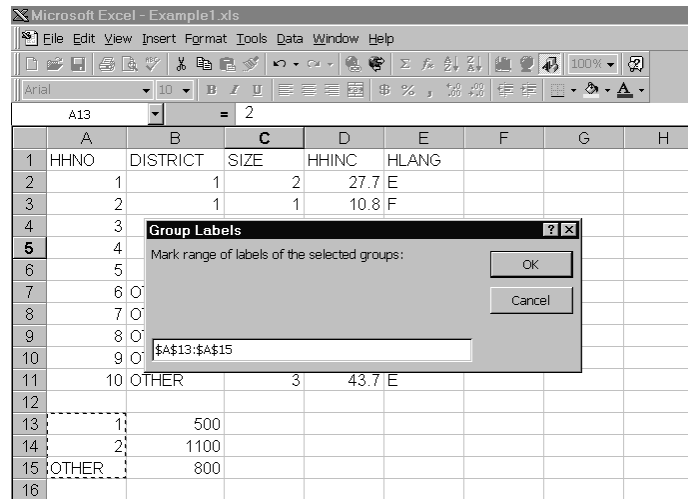
32.5 METHOD: TWO-STAGE RANDOM SAMPLING

To illustrate the application of SampleCalc, it will be assumed that the 5000 city households are grouped into six districts and that the observations were selected in two stages. In the first, three of the six districts were selected at random and without replacement; these were the districts 1, 2, and other. In the second stage, a simple random sample without replacement of 2 households was drawn from among the 500 households in District 1, one of 3 households from among the 1100 households in District 2, and one of 5 households from among the 800 households in the district labeled Other. The labels and the number of households in each district are entered in the same worksheet as the sample observations.

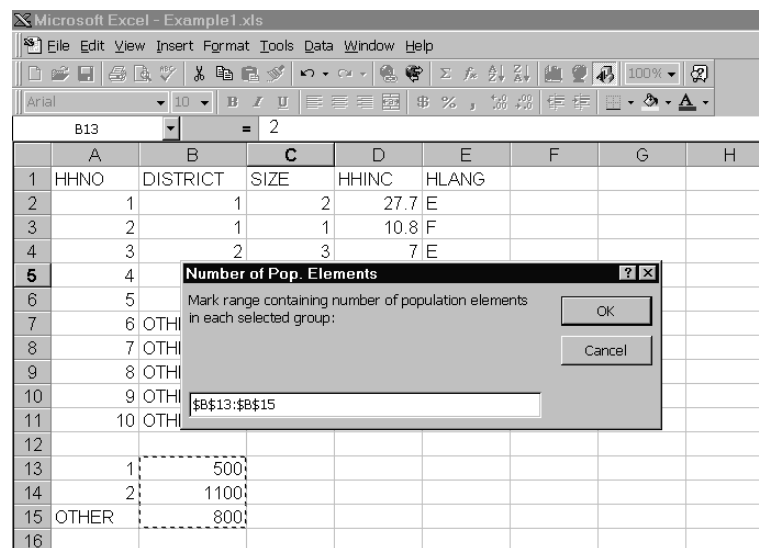
After selecting Two-stage in the **Method** box, the **Group labels** dialog box appears.

- 1 In the dialog box entitled Group labels, select the range containing the labels of the selected groups (strata). Click OK to proceed, or Cancel to abort the program.

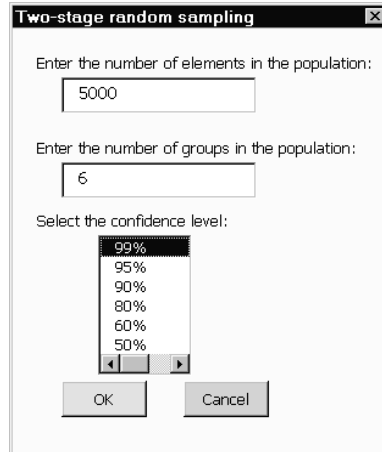
ANALYSIS OF SURVEY DATA USING MICROSOFT EXCEL



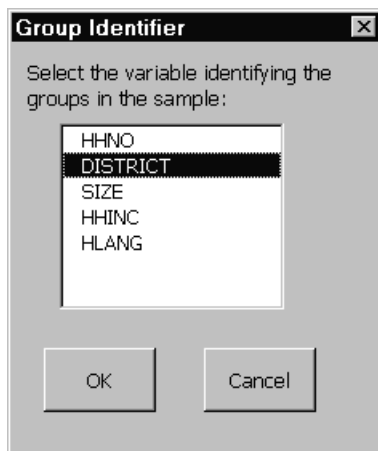
- 2 In the dialog box entitled Number of Pop. Elements, select the range containing the number of elements in each selected group (stratum) in the population. Click OK to proceed, or Cancel to abort the program.



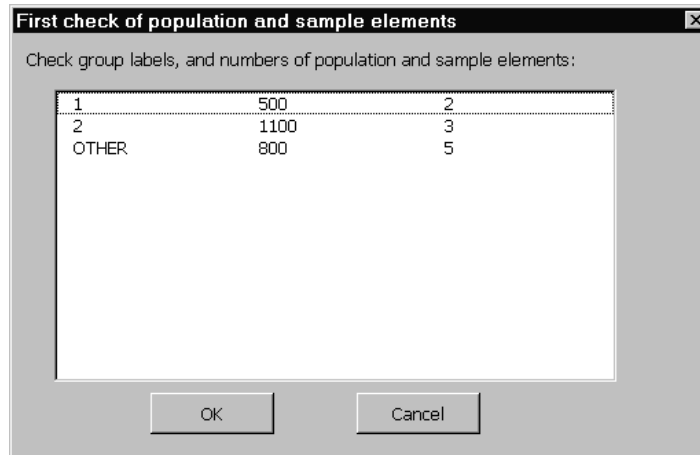
- 3 In the dialog box entitled Two-stage random sampling, enter the number of elements and the number of groups (strata) in the population, and select the desired confidence level. Click OK to proceed or Cancel to abort.



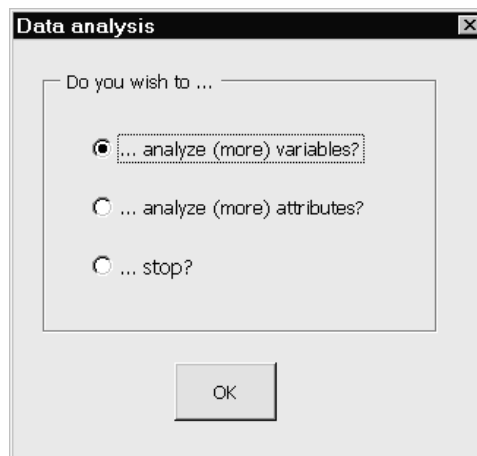
- 4 A dialog box entitled Group Identifier will now appear. Select the label of the one column in the table of data that identifies the group (stratum) to which each sampled element belongs. These labels should be consistent with those in Step 5; otherwise, an error message will eventually appear. Click OK to proceed, or Cancel to abort.



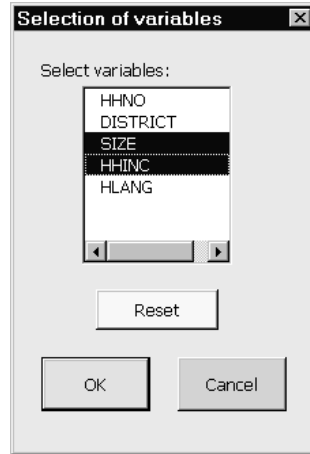
- 5 A dialog box entitled First check of population and sample elements will now appear. It shows the program's understanding of the labels of the selected groups (strata), and of the corresponding number of elements in the population and sample. If you observe an anomaly, click Cancel to abort SampleCalc, check the data, and begin anew. If the program's understanding appears correct, click OK to proceed.



- 6 The program will display for your information a message to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK.
- 7 The dialog box entitled Data analysis will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. You can change the selection at any time before clicking OK. Click OK to confirm your selection and proceed.



- 8 If you chose to analyze one or more variables, a dialog box entitled Selection of variables will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box.



- 9 If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated standard deviation of the estimate of the population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.

Microsoft Excel - Example1.xls

File Edit View Insert Format Tools Data Window Help

Summary of results for variables, two-stage random sampling

	A	B	C	D	E	F	G	H	I	J	K	L
1	Summary of results for variables, two-stage random sampling											
2	99% confidence interval 99% confidence interval											
3	Label	No. Obs.	Est. Mean	Est. Total	Est. Varia	Est. StDo	Mean, low	Mean, up	Total, low	Total, upper	limit	
4	SIZE	10	2.516	12580		0.673345	0.781463	4.250537	3907.313	21252.69		
5	HHINC	9	19.1276	95638		4.624932	7.213774	31.04143	36068.87	155207.1		
6												
7												

- 10 If you chose to analyze one or more attributes, a dialog box entitled Selection of attributes will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box. If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.

1	Summary of results for attributes, two-stage random sampling										
2	99% confidence interval										
3	Label	Category	No. Obs.	Est. Prop.	Est. Numl	Est. StDo	Proportio	Proportio	Number, I	Number, upper limit	
4	SIZE	2	10	0.310667	1553.333	0.142542	-0.05652	0.677856	-282.612	3389.278	
5	SIZE	1	10	0.164	820	0.104227	-0.10449	0.432488	-522.44	2162.44	
6	SIZE	3	10	0.274667	1373.333	0.152774	-0.11888	0.668213	-594.396	3341.063	
7	SIZE	4	10	0.146667	733.3333	0.146567	-0.23089	0.524222	-1154.44	2621.112	
8	SIZE	5	10	0.064	320	0.0639	-0.10061	0.228606	-503.031	1143.031	
9	HLANG	E	10	0.649333	3246.667	0.18308	0.177719	1.120948	888.5954	5604.738	
10	HLANG	F	10	0.310667	1553.333	0.142542	-0.05652	0.677856	-282.612	3389.278	
11											

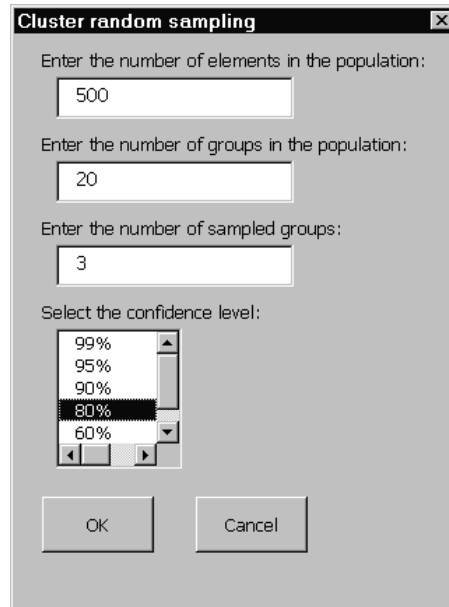
- 11 If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to return to Excel. You may want to adjust the column widths of VRESULTS and ARESULTS in order to see the results more clearly.

32.6 METHOD: CLUSTER RANDOM SAMPLING

To illustrate the application of SampleCalc, it will be assumed that a population consists of 500 city households grouped into 20 districts, and that the observations were selected by drawing a simple random sample without replacement of three of the six districts; these were the districts 1, 2, and Other. All the households in the selected districts were interviewed.

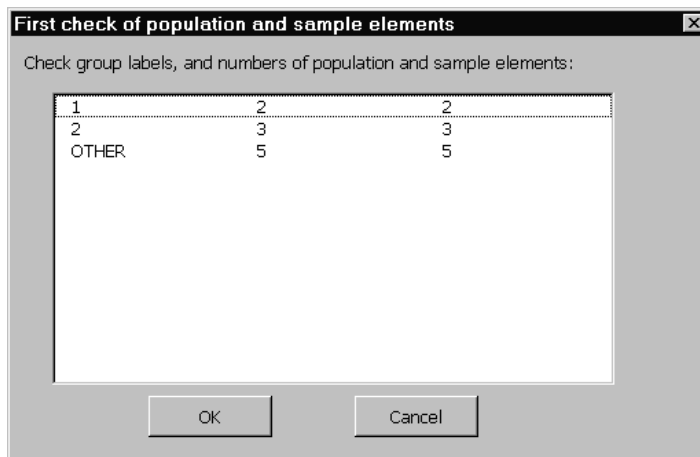
After selecting Cluster in the **Method** box, the **Cluster random sampling** dialog box appears.

- 1 In the dialog box entitled **Cluster random sampling**, enter the number of elements in the population, the number of groups (strata) in the population, and the number of selected groups, and select the desired confidence level. Click OK to proceed or Cancel to abort.



- 2 A dialog box entitled Group Identifier will now appear. Select the label of the one column in the table of data that identifies the group (stratum) to which each sampled element belongs. Click OK to proceed, or Cancel to abort.

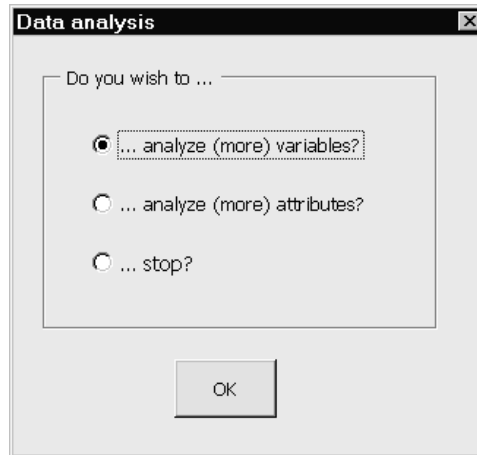
- 3 A dialog box entitled First check of population and sample elements will now appear. It shows the program's understanding of the labels of the selected groups (strata), and of the corresponding number of elements in the population and sample. If you observe an anomaly, click Cancel to abort SampleCalc, check the data, and begin anew. If the program's understanding appears correct, click OK to proceed.



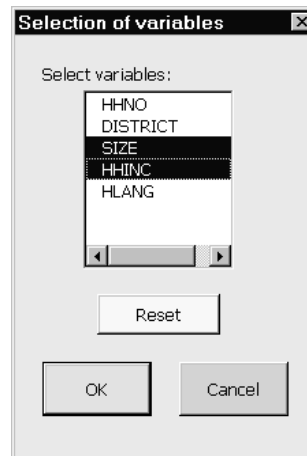
- 4 The program will display for your information a message to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK. (VRESULTS will contain the results of

calculations concerning the selected variables, while ARESULTS will contain the results regarding the selected attributes.)

- 5 The dialog box entitled Data analysis will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. You can change the selection at any time before clicking OK. Click OK to confirm your selection and proceed.



- 6 If you chose to analyze one or more variables, a dialog box entitled **Selection of variables** will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the **Data analysis** dialog box.



ANALYSIS OF SURVEY DATA USING MICROSOFT EXCEL

- If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated standard deviation of the estimate of the population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.

Summary of results for variables, cluster random sampling											
Label	No. Obs.	Est. Mean	Est. Total	Est. Varia	Est. StDo	Mean, low	Mean, upp	Total, low	Total, upper limit	80% confidence intervals	
SIZE	10	0.346667	173.3333		0.117265	0.196333	0.497001	98.16638	248.5003		
HHINC	9	3.04	1520		1.391248	1.25642	4.82358	628.2099	2411.79		

- If you chose to analyze one or more attributes, a dialog box entitled **Selection of attributes** will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the **Data analysis** dialog box. If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.

Summary of results for attributes, cluster random sampling											
Label	Category	No. Obs.	Est. Prop	Est. Num	Est. StDo	Proportio	Proportio	Number, l	Number, upper limit	80% confidence intervals	
SIZE	2	10	0.04	20	1.18E-18	0.04	0.04	20	20		
SIZE	1	10	0.026667	13.33333	0.012293	0.010907	0.042426	5.453696	21.21297		
SIZE	3	10	0.04	20	0.021292	0.012704	0.067296	6.352068	33.64793		
SIZE	4	10	0.013333	6.666667	0.012293	-0.00243	0.029093	-1.21297	14.5463		
SIZE	5	10	0.013333	6.666667	0.012293	-0.00243	0.029093	-1.21297	14.5463		
HLANG	E	10	0.093333	46.66667	0.032523	0.051638	0.135028	25.81911	67.51423		
HLANG	F	10	0.04	20	1.18E-18	0.04	0.04	20	20		

- If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to return to Excel. You may want to adjust the column widths of VRESULTS and ARESULTS in order to see the results more clearly

32.7 GLOSSARY

Sampling With and Without Replacement: Sampling is said to be with or without replacement according to whether or not an element or group of elements can appear more than once in the sample.

Simple Random Sampling: One method of selecting a simple random sample is by a series of draws, in every one of which each eligible element has the same chance of being selected, and each selection is unrelated to (independent of) other selections. Such a sample is without replacement if the element selected in any draw is not eligible for selection in any subsequent draw.

Stratified random sampling: Requires that the elements of the population be grouped according to one or more criteria. A stratified random sample consists of simple random samples without replacement from each and every group (stratum).

Two-Stage Random Sampling: Requires that the elements of the population be grouped according to one or more criteria. As the name implies, the sample is selected in two stages. In the first, a simple random sample of groups (strata) is selected. Then, in the second stage, a simple random sample of elements is selected from each group (stratum) that was selected in the first stage.

Cluster Random Sampling: Cluster sampling can be viewed as a special case of two-stage sampling. A cluster sample requires that the elements of the population be grouped according to one or more criteria. In the first stage, a simple random sample of groups (strata) is selected. Then, in the second stage, *all* the elements are selected from each group (stratum) that was selected in the first stage.

Variables and Attributes: A variable is a feature or aspect of an element that lends itself naturally to a numerical description. For example, the age of a person, the income of a household, the maximum temperature in a day, etc. Unlike a variable, an attribute is a feature or aspect of an element that lends itself only to a categorical, qualitative---not numerical---description. For example, a person's gender (male, female), a household's location (in, say, district A, B, or C), the industrial classification of a manufacturing company, etc.

Population Characteristics: The population characteristics of greatest interest in practice are the mean or total of a variable, and the proportion or number in a category of an attribute.

Point and Interval Estimates: Rather than, or in addition to, saying that a population characteristic is estimated to have such and such a value (the point estimate of the characteristic), it may be informative to state that a certain interval is estimated to contain a population characteristic with a certain probability. For example, instead of saying that the population mean of a variable is estimated to be 7.3, it may be informative to state that the population mean of the variable is in the interval from 6.5 to 8.1, where this statement is correct with probability 95%. The probability is often referred to as the confidence level and the interval estimate as a confidence interval.

Unbiased estimates: An estimate (more precisely, an estimator) of a population characteristic is said to be unbiased if its average value in a very large number of identical samples is equal to the population characteristic.

REFERENCES

<http://www.yorku.ca/ptryfos/index.htm>

<http://www.yorku.ca/ptryfos/compprog.htm>

<http://office.microsoft.com/en-in/excel-help/about-add-in-programs-HP005238607.aspx>

SAS FOR STATISTICAL PROCEDURES

Rajender Parsad

I.A.S.R.I., Library Avenue, New Delhi-110 012

rajender@iasri.res.in

33.1 Introduction

SAS (Statistical Analysis System) software is comprehensive software which deals with many problems related to Statistical analysis, Spreadsheet, Data Creation, Graphics, etc. It is a layered, multivendor architecture. Regardless of the difference in hardware, operating systems, etc., the SAS applications look the same and produce the same results. The three components of the SAS System are Host, Portable Applications and Data. Host provides all the required interfaces between the SAS system and the operating environment. Functionalities and applications reside in Portable component and the user supplies the Data. We, in this course will be dealing with the software related to perform statistical analysis of data.

Windows of SAS

1. Program Editor : All the instructions are given here.
2. Log : Displays SAS statements submitted for execution and messages
3. Output : Gives the output generated

Rules for SAS Statements

1. SAS program communicates with computer by the SAS statements.
2. Each statement of SAS program must end with semicolon (;).
3. Each program must end with run statement.
4. Statements can be started from any column.
5. One can use upper case letters, lower case letters or the combination of the two.

Basic Sections of SAS Program

1. DATA section
2. CARDS section
3. PROCEDURE section

Data Section

We shall discuss some facts regarding data before we give the syntax for this section.

Data value: A single unit of information, such as name of the specie to which the tree belongs, height of one tree, etc.

Variable: A set of values that describe a specific data characteristic e.g. diameters of all trees in a group. The variable can have a name upto a maximum of 8 characters and must begin with a letter or underscore. Variables are of two types:

Character Variable: It is a combination of letters of alphabet, numbers and special characters or symbols.

Numeric Variable: It consists of numbers with or without decimal points and with + or - ve signs.

Observation: A set of data values for the same item i.e. all measurement on a tree.

Data section starts with Data statements as

DATA NAME (it has to be supplied by the user);

Input Statements

Input statements are part of data section. This statement provides the **SAS** system the name of the variables with the format, if it is formatted.

List Directed Input

- Data are read in the order of variables given in input statement.
- Data values are separated by one or more spaces.
- Missing values are represented by period (.).
- Character values are followed by \$ (dollar sign).

Example 1:

```
Data A;  
INPUT ID SEX $ AGE HEIGHT WEIGHT;  
CARDS;  
1 M 23 68 155  
2 F . 61 102  
3. M 55 70 202  
;
```

Column Input

Starting column for the variable can be indicated in the input statements for example:

```
INPUT ID 1-3 SEX $ 4 HEIGHT 5-6 WEIGHT 7-11;  
CARDS;  
001M68155.5  
2F61 99  
3M53 33.5  
;
```

Alternatively, starting column of the variable can be indicated along with its length as

```
INPUT @ 1 ID 3.  
@ 4 SEX $ 1.  
@ 9 AGE 2.  
@ 11 HEIGHT 2.  
@ 16 V_DATE MMDDYY 6.  
;
```

Reading More than One Line Per Observation for One Record of Input Variables

```
INPUT # 1 ID 1-3 AGE 5-6 HEIGHT 10-11
```

```
# 2 SBP 5-7 DBP 8-10;
```

```
CARDS;
```

```
001 56 72
```

```
140 80
```

```
;
```

Reading the Variable More than Once

Suppose id variable is read from six columns in which state code is given in last two columns of id variable for example:

```
INPUT @ 1 ID 6. @ 5 STATE 2.;
```

```
OR
```

```
INPUT ID 1-6 STATE 5-6;
```

Formatted Lists

```
DATA B;
```

```
INPUT ID @1(X1-X2)(1.)
```

```
@4(Y1-Y2)(3.);
```

```
CARDS;
```

```
11 563789
```

```
22 567987
```

```
;
```

```
PROC PRINT;
```

```
RUN;
```

Output

Obs.	ID	x1	x2	y1	y2
1	11	1	1	563	789
2	22	2	2	567	987

```
DATA C;
```

```
INPUT X Y Z @;
```

```
CARDS;
```

```
1 1 1 2 2 2 5 5 5 6 6 6
```

```
1 2 3 4 5 6 3 3 3 4 4 4
```

```
;
```

```
PROC PRINT;
```

```
RUN;
```

Output

Obs.	X	Y	Z
1	1	1	1
2	1	2	3

```
DATA D;
```

```
INPUT X Y Z @@;  
CARDS;  
1 1 1 2 2 2 5 5 5 6 6 6  
1 2 3 4 5 6 3 3 3 4 4 4  
;  
PROC PRINT;  
RUN;
```

Output:

Obs.	X	Y	Z
1	1	1	1
2	2	2	2
3	5	5	5
4	6	6	6
5	1	2	3
6	4	5	6
7	3	3	3
8	4	4	4

33.2 DATA FILES

SAS System Can Read and Write

- A. Simple ASCII files are read with input and infile statements
- B. Output Data files

Creation of SAS Data Set

```
DATA EX1;  
INPUT GROUP $ X Y Z;  
CARDS;  
T1 12 17 19  
T2 23 56 45  
T3 19 28 12  
T4 22 23 36  
T5 34 23 56  
;
```

Creation of SAS File From An External (ASCII) File

```
DATA EX2;  
INFILE 'B:MYDATA';  
INPUT GROUP $ X Y Z;  
OR  
DATA EX2A;  
FILENAME ABC 'B:MYDATA';  
INFILE ABC;  
INPUT GROUP $ X Y Z;  
;
```

Creation of A SAS Data Set and An Output ASCII File Using an External File

```

DATA EX3;
FILENAME IN 'C:MYDATA';
FILENAME OUT 'A:NEWDATA';
INFILE IN;
FILE OUT;
INPUT GROUP $ X Y Z;
TOTAL =SUM (X+Y+Z);
PUT GROUP $ 1-10 @12 (X Y Z TOTAL)(5.);
RUN;

```

This above program reads raw data file from 'C: MYDATA', and creates a new variable TOTAL and writes output in the file 'A: NEWDATA'.

Creation of SAS File from an External (*.csv) File

```

data EX4;
infile'C:\Users\Admn\Desktop\sccnars.csv' dlm=',' ;
      /*give the exact path of the file, file should not have column headings*/
input sn loc $ year season $ crop $ rep trt gyield syield return kcal;
/*give the variables in ordered list in the file*/
/*if we have the first row as names of the columns then we can write in the above
statement firstobs=2 so that data is read from row 2 onwards*/
biomass=gyield+syield; /*generates a new variable*/
proc print data=EX4;
run;

```

Note: To create a SAS File from a *.txt file, only change csv to txt and define delimiter as per file created.

Creation of SAS File from an External (*.xls) File

Note: it is always better to copy the name of the variables as comment line before Proc Import.

```

/* name of the variables in Excel File provided the first row contains variable name*/
proc import datafile = 'C:\Users\Desktop\DATA_EXERCISE\descriptive_stats.xls'
/*give the exact path of the file*/
out = descriptive_stats replace; /*give output file name*/
proc print;
run;

```

If we want to make some transformations, then we may use the following statements:

```

data a1;
set descriptive_stats;
x = fs45+fw;
run;

```


Here **proc import** allows the SAS user to import data from an EXCEL spreadsheet into SAS. The **datafile** statement provides the reference location of the file. The **out** statement is used to name the SAS data set that has been created by the import procedure. **Print procedure** has been utilized to view the contents of the SAS data set **descriptive_stats**. When we run above codes we obtain the output which will same as shown above because we are using the same data.

Creating a Permanent SAS Data Set

```
LIBNAME XYZ 'C:\SASDATA';
DATA XYZ.EXAMPLE;
INPUT GROUP $ X Y Z;
CARDS;
.....
.....
.....
RUN;
```

This program reads data following the cards statement and creates a permanent SAS data set in a subdirectory named \SASDATA on the C: drive.

33.3 Using Permanent SAS File

```
LIBNAME XYZ 'C:\SASDATA';
PROC MEANS DATA=XYZ.EXAMPLE;
RUN;
```

TITLES

One can enter upto 10 titles at the top of output using TITLE statement in your procedure.

```
PROC PRINT;
TITLE 'HEIGHT-DIA STUDY';
TITLE3 '1999 STATISTICS';
RUN;
```

Comment cards can be added to the SAS program using
/* COMMENTS */;

FOOTNOTES

One can enter upto 10 footnotes at the bottom of your output.

```
PROC PRINT DATA=DIAHT;
FOOTNOTE '1999';
FOOTNOTE5 'STUDY RESULTS';
RUN;
```

For obtaining output as RTF file, use the following statements
Ods rtf file='xyz.rtf' style =journal;
Ods rtf close;

For obtaining output as PDF/HTML file, replace rtf with pdf or html in the above statements.

If we want to get the output in continuous format, then we may use

```
Ods rtf file='xyz.rtf' style =journal bodytitle startpage=no;
```

LABELLING THE VARIABLES

```
Data dose;
```

```
title 'yield with factors N P K';
```

```
input N P K Yield;
```

```
Label N = "Nitrogen";
```

```
Label P = " Phosphorus";
```

```
Label K = " Potassium";
```

```
cards;
```

```
...
```

```
...
```

```
...
```

```
;
```

```
Proc print;
```

```
run;
```

We can define the linesize in the output using statement OPTIONS. For example, if we wish that the output should have the linesize (number of columns in a line) is 72 use Options linesize =72; in the beginning.

33.4 Statistical Procedure

SAS/STAT has many capabilities using different procedures with many options. There are a total of 73 PROCS in SAS/STAT. SAS/STAT is capable of performing a wide range of statistical analysis that includes:

1. Elementary / Basic Statistics
 2. Graphs/Plots
 3. Regression and Correlation Analysis
 4. Analysis of Variance
 5. Experimental Data Analysis
 6. Multivariate Analysis
 7. Principal Component Analysis
 8. Discriminant Analysis
 9. Cluster Analysis
 10. Survey Data Analysis
 11. Mixed model analysis
 12. Variance Components Estimation
 13. Probit Analysis
- and many more...

A brief on SAS/STAT Procedures is available at

<http://support.sas.com/rnd/app/da/stat/procedures/Procedures.html>

Example 2.1: To Calculate the Means and Standard Deviation:

```
DATA TESTMEAN;
  INPUT GROUP $ X Y Z;
  CARDS;
  CONTROL 12 17 19
  TREAT1 23 25 29
  TREAT2 19 18 16
  TREAT3 22 24 29
  CONTROL 13 16 17
  TREAT1 20 24 28
  TREAT2 16 19 15
  TREAT3 24 26 30
  CONTROL 14 19 21
  TREAT1 23 25 29
  TREAT2 18 19 17
  TREAT3 23 25 30
  ;
  PROC MEANS;
  VAR X Y Z;
  RUN;
```

The default output displays mean, standard deviation, minimum value, maximum value of the desired variable. We can choose the required statistics from the options of PROC MEANS. For example, if we require mean, standard deviation, median, coefficient of variation, coefficient of skewness, coefficient of kurtosis, etc., then we can write

```
PROC MEANS mean std median cv skewness kurtosis;
  VAR X Y Z;
  RUN;
```

The default output is 6 decimal places, desired number of decimal places can be defined by using option maxdec=.... For example, for an output with three decimal places, we may write

```
PROC MEANS mean std median cv skewness kurtosis maxdec=3;
  VAR X Y Z;
  RUN;
```

For obtaining means group wise use, first sort the data by groups using

```
Proc sort;
  By group;
  Run;
```

And then make use of the following

```
PROC MEANS;
  VAR X Y Z;
  by group;
```

```
RUN;
```

Or alternatively, me may use

```
PROC MEANS;
CLASS GROUP;
VAR X Y Z;
RUN;
```

For obtaining descriptive statistics for a given data one can use PROC SUMMARY. In the above example, if one wants to obtain mean standard deviation, coefficient of variation, coefficient of skewness and kurtosis, then one may utilize the following:

```
PROC SUMMARY PRINT MEAN STD CV SKEWNESS KURTOSIS;
CLASS GROUP;
VAR X Y Z;
RUN;
```

Most of the Statistical Procedures require that the data should be normally distributed. For testing the normality of data, PROC UNIVARIATE may be utilized.

```
PROC UNIVARIATE NORMAL;
VAR X Y Z;
RUN;
```

If different plots are required then, one may use:

```
PROC UNIVARIATE DATA=TEST NORMAL PLOT;
/*plot option displays stem-leaf, boxplot & Normal prob plot*/
VAR X Y Z;
/*creates side by side BOX-PLOT group-wise. To use this option first sort the file on by
variable*/
BY GROUP;
HISTOGRAM/KERNEL NORMAL; /*displays kernel density along with normal curve*/
PROBPLOT; /*plots probability plot*/
QQPLOT X/NORMAL SQUARE; /*plot quantile-quantile QQ-plot*/
CDFPLOT X/NORMAL; /*plots CDF plot*/
/*plots pp plot which compares the empirical cumulative distribution function (ecdf) of a
variable with a specified theoretical cumulative distribution function. The beta,
exponential, gamma, lognormal, normal, and Weibull distributions are available in both
statements.*/
PPPLOT X/NORMAL;
RUN;
```

Example 2.2: To Create Frequency Tables

```
DATA TESTFREQ;
INPUT AGE $ ECG CHD $ CAT $ WT; CARDS;
<55 0 YES YES 1
<55 0 YES YES 17
<55 0 NO YES 7
<55 1 YES NO 257
```

SAS FOR STATISTICAL PROCEDURES

```
<55  1    YES  YES  3
<55  1    YES  NO   7
<55  1    NO   YES  1
55+  0    YES  YES  9
55+  0    YES  NO   15
55+  0    NO   YES  30
55+  1    NO   NO   107
55+  1    YES  YES  14
55+  1    YES  NO   5
55+  1    NO   YES  44
55+  1    NO   NO   27
```

```

;
PROC FREQ DATA=TESTFREQ;
TABLES AGE*ECG/MISSING CHISQ;
TABLES AGE*CAT/LIST;
RUN;
```

SCATTER PLOT

```
PROC PLOT DATA = DIAHT;
PLOT HT*DIA = '*';
/*HT=VERTICAL AXIS DIA = HORIZONTAL AXIS.*/
RUN;
```

CHART

```
PROC CHART DATA = DIAHT;
VBAR HT;
RUN;
```

```
PROC CHART DATA = DIAHT;
HBAR DIA;
RUN;
```

```
PROC CHART DATA = DIAHT;
PIE HT;
RUN;
```

Example 2.3: To Create A Permanent SAS DATASET and use that for Regression

```
LIBNAME FILEX 'C:\SAS\RPLIB';
```

```
DATA FILEX.RP;
```

```
INPUT X1-X5;
```

```
CARDS;
```

```
1  0  0  0  5.2
.75 .25 0  0  7.2
.75 0  .25 0  5.8
.5  .25 .25 0  6.3
.75 0  0  .25 5.5
.5  0  .25 .25 5.7
```

```
.5 .25 0 .25 5.8
.25 .25 .25 .25 5.7
;
RUN;
```

```
LIBNAME FILEX 'C:\SAS\RPLIB';
PROC REG DATA=FILEX.RP;
MODEL X5 = X1 X2/P;
MODEL X5 = X1 X2 X3 X4 / SELECTION = STEPWISE;
TEST: TEST X1-X2=0;
RUN;
```

Various other commonly used PROC Statements are PROC ANOVA, PROC GLM; PROC CORR; PROC NESTED; PROC MIXED; PROC RSREG; PROC IML; PROC PRINCOMP; PROC VARCOMP; PROC FACTOR; PROC CANCELL; PROC DISCRIM, etc. Some of these are described in the sequel.

PROC TTEST is the procedure that is used for comparing the mean of a given sample. This PROC is also used for compares the means of two independent samples. The paired observations t test compares the mean of the differences in the observations to a given number. The underlying assumption of the t test in all three cases is that the observations are random samples drawn from normally distributed populations. This assumption can be checked using the UNIVARIATE procedure; if the normality assumptions for the t test are not satisfied, one should analyze the data using the NPAR1WAY procedure. PROC TTEST computes the group comparison t statistic based on the assumption that the variances of the two groups are equal. It also computes an approximate t based on the assumption that the variances are unequal (the Behrens-Fisher problem). The following statements are available in PROC TTEST.

```
PROC TTEST <options>;
CLASS variable;
PAIRED variables;
BY variables;
VAR variables;
FREQ Variables;
WEIGHT variable;
```

No statement can be used more than once. There is no restriction on the order of the statements after the PROC statement. The following options can appear in the PROC TTEST statement.

ALPHA= p : option specifies that confidence intervals are to be $100(1-p)\%$ confidence intervals, where $0 < p < 1$. By default, PROC TTEST uses ALPHA=0.05. If p is 0 or less, or 1 or more, an error message is printed.

COCHRAN: option requests the Cochran and Cox approximation of the probability level of the approximate t statistic for the unequal variances situation.

$H_0=m$: option requests tests against m instead of 0 in all three situations (one-sample, two-sample, and paired observation t tests). By default, PROC TTEST uses $H_0=0$.

A CLASS statement giving the name of the classification (or grouping) variable must accompany the PROC TTEST statement in the two independent sample cases. It should be omitted for the one sample or paired comparison situations. The class variable must have two, and only two, levels. PROC TTEST divides the observations into the two groups for the t test using the levels of this variable. One can use either a numeric or a character variable in the CLASS statement.

In the statement PAIRED *PairLists*, the *PairLists* in the PAIRED statement identifies the variables to be compared in paired comparisons. You can use one or more *PairLists*. Variables or lists of variables are separated by an asterisk (*) or a colon (:). Examples of the use of the asterisk and the colon are shown in the following table.

The PAIRED Statements	Comparisons made
PAIRED A*B;	A-B
PAIRED A*B C*D;	A-B and C-D
PAIRED (A B)*(C B);	A-C, A-B and B-C
PAIRED (A1-A2)*(B1-B2);	A1-B1, A1-B2, A2-B1 and A2-B2
PAIRED (A1-A2):(B1-B2);	A1-B1 and A2-B2

PROC ANOVA performs analysis of variance for balanced data only from a wide variety of experimental designs whereas PROC GLM can analyze both balanced and unbalanced data. As ANOVA takes into account the special features of a balanced design, it is faster and uses less storage than PROC GLM for balanced data. The basic syntax of the ANOVA procedure is as given:

```
PROC ANOVA < Options>;
  CLASS variables;
  MODEL dependents = independent variables (or effects)/options;
  MEANS effects/options;
  ABSORB variables;
  FREQ variables;
  TEST H = effects E = effect;
  MANOVA H = effects E = effect;
        M = equations/options;
  REPEATED factor - name levels / options;
  By variables;
```

The PROC ANOVA, CLASS and MODEL statements are must. The other statements are optional. The CLASS statement defines the variables for classification (numeric or character variables - maximum characters =16).

The MODEL statement names the dependent variables and independent variables or effects. If no effects are specified in the MODEL statement, ANOVA fits only the intercept. Included in the ANOVA output are F-tests of all effects in the MODEL

statement. All of these F-tests use residual mean squares as the error term. The MEANS statement produces tables of the means corresponding to the list of effects. Among the options available in the MEANS statement are several multiple comparison procedures viz. Least Significant Difference (LSD), Duncan's New multiple - range test (DUNCAN), Waller - Duncan (WALLER) test, Tukey's Honest Significant Difference (TUKEY). The LSD, DUNCAN and TUKEY options takes level of significance ALPHA = 5% unless ALPHA = options is specified. Only ALPHA = 1%, 5% and 10% are allowed with the Duncan's test. 95% Confidence intervals about means can be obtained using CLM option under MEANS statement.

The TEST statement tests for the effects where the residual mean square is not the appropriate term such as main - plot effects in split - plot experiment. There can be multiple MEANS and TEST statements (as well as in PROC GLM), but only one MODEL statement preceded by RUN statement. The ABSORB statement implements the technique of absorption, which saves time and reduces storage requirements for certain type of models. FREQ statement is used when each observation in a data set represents 'n' observations, where n is the value of FREQ variable. The MANOVA statement is used for implementing multivariate analysis of variance. The REPEATED statement is useful for analyzing repeated measurement designs and the BY statement specifies that separate analysis are performed on observations in groups defined by the BY variables.

PROC GLM for analysis of variance is similar to using PROC ANOVA. The statements listed for PROC ANOVA are also used for PROC GLM. In addition; the following more statements can be used with PROC GLM:

```
CONTRAST 'label' effect name< ... effect coefficients > </options>;
ESTIMATE 'label' effect name< ... effect coefficients > </options>;
ID variables;
LSMEANS effects </ options >;
OUTPUT < OUT = SAS-data-set>keyword=names< ... keyword = names>;
RANDOM effects </ options >;
WEIGHT variables
```

Multiple comparisons as used in the options under MEANS statement are useful when there are no particular comparisons of special interest. But there do occur situations where preplanned comparisons are required to be made. Using the CONTRAST, LSMEANS statement, we can test specific hypothesis regarding pre - planned comparisons. The basic form of the CONTRAST statement is as described above, where label is a character string used for labeling output, effect name is class variable (which is independent) and effect - coefficients is a list of numbers that specifies the linear combination parameters in the null hypothesis. The contrast is a linear function such that the elements of the coefficient vector sum to 0 for each effect. While using the CONTRAST statements, following points should be kept in mind.

How many levels (classes) are there for that effect. If there are more levels of that effect in the data than the number of coefficients specified in the CONTRAST statement, the

PROC GLM adds trailing zeros. Suppose there are 5 treatments in a completely randomized design denoted as T_1, T_2, T_3, T_4, T_5 and null hypothesis to be tested is

$$H_0: T_2 + T_3 = 2T_1 \text{ or } -2T_1 + T_2 + T_3 = 0$$

Suppose in the data treatments are classified using TRT as class variable, then effect name is TRT CONTRAST 'TIVS 2&3' TRT -2 1 1 0 0; Suppose last 2 zeros are not given, the trailing zeros can be added automatically. The use of this statement gives a sum of squares with 1 degree of freedom (d.f.) and F-value against error as residual mean squares until specified. The name or label of the contrast must be 20 characters or less.

The available CONTRAST statement options are

E: prints the entire vector of coefficients in the linear function, i.e., contrast.

E = effect: specifies an effect in the model that can be used as an error term

ETYPE = n : specifies the types (1, 2, 3 or 4) of the E effect.

Multiple degrees of freedom contrasts can be specified by repeating the effect name and coefficients as needed separated by commas. Thus the statement for the above example

```
CONTRAST 'All' TRT -2 1 1 0 0, TRT 0 1 -1 0 0;
```

This statement produces two d.f. sum of squares due to both the contrasts. This feature can be used to obtain partial sums of squares for effects through the reduction principle, using sums of squares from multiple degrees of freedom contrasts that include and exclude the desired contrasts. Although only $t-1$ linearly independent contrasts exists for t classes, any number of contrasts can be specified.

The ESTIMATE statement can be used to estimate linear functions of parameters that may or may not be obtained by using CONTRAST or LSMEANS statement. For the specification of the statement only word CONTRAST is to be replaced by ESTIMATE in CONTRAST statement.

Fractions in effects coefficients can be avoided by using DIVISOR = Common denominator as an option. This statement provides the value of an estimate, a standard error and a t-statistic for testing whether the estimate is significantly different from zero.

The LSMEANS statement produces the least square estimates of CLASS variable means i.e. adjusted means. For one-way structure, there are simply the ordinary means. The least squares means for the five treatments for all dependent variables in the model statement can be obtained using the statement.

```
LSMEANS TRT / options;
```

Various options available with this statement are:

STDERR: gives the standard errors of each of the estimated least square mean and the t-statistic for a test of hypothesis that the mean is zero.

PDIFF: Prints the p - values for the tests of equality of all pairs of CLASS means.

SINGULAR: tunes the estimability checking. The options E, E=, E-TYPE = are similar as discussed under CONTRAST statement.

Adjust=T: gives the probabilities of significance of pairwise comparisons based on T-test.

Adjust=Tukey: gives the probabilities of significance of pairwise comparisons based on Tukey's test

Lines: gives the letters on treatments showing significant and non-significant groups

When the **predicted** values are requested as a MODEL statement option, values of variable specified in the ID statement are printed for identification besides each observed, predicted and residual value. The OUTPUT statement produces an output data set that contains the original data set values alongwith the predicted and residual values.

Besides other options in PROC GLM under MODEL statement we can give the option: 1. solution 2. xpx ($=\mathbf{X}^{\prime}\mathbf{X}$) 3 . I (g-inverse)

PROC GLM recognizes different theoretical approaches to ANOVA by providing four types of sums of squares and associated statistics. The four types of sums of squares in PROC GLM are called Type I, Type II, Type III and Type IV.

The Type I sums of squares are the classical sequential sums of squares obtained by adding the terms to the model in some logical sequence. The sum of squares for each class of effects is adjusted for only those effects that precede it in the model. Thus the sums of squares and their expectations are dependent on the order in which the model is specified.

The Type II, III and IV are 'partial sums of squares' in the sense that each is adjusted for all other classes of the effects in the model, but each is adjusted according to different rules. One general rule applies to all three types: the estimable functions that generate the sums of squares for one class of squares will not involve any other classes of effects except those that "contain" the class of effects in question.

For example, the estimable functions that generate SS (AB) in a three- factor factorial will have zero coefficients on main effects and the (A × C) and (B × C) interaction effects. They will contain non-zero coefficient on the (A × B × C) interaction effects, because A × B × C interaction "contains" A × B interaction.

Type II, III and IV sums of squares differ from each other in how the coefficients are determined for the classes of effects that do not have zero coefficients - those that contain the class of effects in question. The estimable functions for the Type II sum of squares impose no restriction on the values of the non-zero coefficients on the remaining effects; they are allowed to take whatever values result from the computations adjusting for effects that are required to have zero coefficients. Thus, the coefficients on the higher-order interaction effects and higher level nesting effects are functions of the number of observations in the data. In general, the Type II sums of squares do not possess of equitable distribution property and orthogonality characteristic of balanced data.

The Type III and IV sums of squares differ from the Type II sums of squares in the sense that the coefficients on the higher order interaction or nested effects that contain the effects in question are also adjusted so as to satisfy either the orthogonality condition (Type III) or the equitable distribution property (Type IV).

The coefficients on these effects are no longer functions of the n_{ij} and consequently, are the same for all designs with the same general form of estimable functions. If there are no empty cells (no $n_{ij} = 0$) both conditions can be satisfied at the same time and Type III and Type IV sums of squares are equal. The hypothesis being tested is the same as when the data is balanced.

When there are empty cells, the hypotheses being tested by the Type III and Type IV sums of squares may differ. The Type III criterion of orthogonality reproduces the same hypotheses one obtains if effects are assumed to add to zero. When there are empty cells this is modified to “the effects that are present are assumed to be zero”. The Type IV hypotheses utilize balanced subsets of non-empty cells and may not be unique. For a 2x3 factorial for illustration purpose adding the terms to the model in the order A, B, AB various types sums of squares can be explained as follows:

Effect	Type I	Type II	Type III	Type IV
General Mean	R(μ)	R(μ)		
A	R(A/ μ)	R(A/ μ ,B)	R(A/ μ ,B,AB)	
B	R(B/ μ ,A)	R(B/ μ ,A)	R(B/ μ ,A,AB)	
A*B	R(A*B/ μ ,A,B)	R(A*B/ μ ,A,B)	R(AB/ μ ,A,B)	

R (A/ μ) is sum of squares adjusted for μ , and so on.

Thus in brief the four sets of sums of squares Type I, II, III & IV can be thought of respectively as sequential, each - after-all others, Σ -restrictions and hypotheses.

There is a relationship between the four types of sums of squares and four types of data structures (balanced and orthogonal, unbalanced and orthogonal, unbalanced and non-orthogonal (all cells filled), unbalanced and non-orthogonal (empty cells)). For illustration, let n_{IJ} denote the number of observations in level I of factor A and level j of factor B. Following table explains the relationship in data structures and Types of sums of squares in a two-way classified data.

33.5 Data Structure Type

	1	2	3	4
Effect	Equal n_{IJ}	Proportionate n_{IJ}	Disproportionate non-zero n_{IJ}	Empty Cell
A	I=II=III=IV	I=II,III=IV	III=IV	
B	I=II=III=IV	I=II,III=IV	I=II,III=IV	I=II
A*B	I=II=III=IV	I=II=III=IV	I=II=III=IV	I=II=III=IV

In general,

I=II=III=IV	(balanced data); II=III=IV	(no interaction models)
I=II, III=IV	(orthogonal data); III=IV	(all cells filled data).

Proper Error terms: In general F-tests of hypotheses in ANOVA use the residual mean squares in other terms are to be used as error terms. For such situations PROC GLM provides the TEST statement which is identical to the test statement available in PROC ANOVA. PROC GLM also allows specification of appropriate error terms in MEANS LSMEANS and CONTRAST statements. To illustrate it let us use split plot experiment involving the yield of different irrigation (IRRIG) treatments applied to main plots and cultivars (CULT) applied to subplots. The data so obtained can be analysed using the following statements.

```
data splitplot;
input REP IRRIG CULT YIELD;
cards;
...
...
...
;
PROC print; run;
PROC GLM;
class rep, irrig cult;
model yield = rep irrig rep*irrig cult irrig* cult;
test h = irrig e = rep * irrig;
contrast 'IRRIGI Vs IRRIG2' irrig 1 -1 / e = rep* irrig;
run;
```

As we know here that the irrigation effects are tested using error (A) which is sum of squares due to rep* irrig, as taken in test statement and contrast statement respectively.

In Test statement	H	=	numerator for - source of variation and
	E	=	denominator source of variation

It may be noted here that the PROC GLM can be used to perform analysis of covariance as well. For analysis of covariance, the covariate should be defined in the model without specifying under CLASS statement.

PROC RSREG fits the parameters of a complete quadratic response surface and analyses the fitted surface to determine the factor levels of optimum response and performs a ridge analysis to search for the region of optimum response.

```
PROC RSREG < options >;
MODEL responses = independents / <options >;
RIDGE < options >;
WEIGHT variable;
ID variable;
By variable;
run;
```

The PROC RSREG and model statements are required. The BY, ID, MODEL, RIDGE, and WEIGHT statements are described after the PROC RSREG statement below and can appear in any order.

The PROC RSREG statement invokes the procedure and following options are allowed with the PROC RSREG:

DATA = SAS - data-set : specifies the data to be analysed.
 NOPRINT : suppresses all printed results when only the output data set is required.
 OUT : SAS-data-set: creates an output data set.

The model statement without any options transforms the independent variables to the coded data. By default, PROC RSREG computes the linear transformation to perform the coding of variables by subtracting average of highest and lowest values of the independent variable from the original value and dividing by half of their differences. Canonical and ridge analyses are performed to the model fit to the coded data. The important options available with the model statement are:

NOCODE : Analyses the original data.
 ACTUAL : specifies the actual values from the input data set.
 COVAR = n : declares that the first n variables on the independent side of the model are simple linear regression (covariates) rather than factors in the quadratic response surface.
 LACKFIT : Performs lack of fit test. For this the repeated observations must appear together.
 NOANOVA : suppresses the printing of the analysis of variance and parameter estimates from the model fit.
 NOOPTIMAL (NOOPT): suppresses the printing of canonical analysis for quadratic response surface.
 NOPRINT : suppresses both ANOVA and the canonical analysis.
 PREDICT : specifies the values predicted by the model.
 RESIDUAL : specifies the residuals.

A RIDGE statement computes the ridge of the optimum response. Following important options available with RIDGE statement are

MAX: computes the ridge of maximum response.
 MIN: computes the ridge of the minimum response.

At least one of the two options must be specified.

NOPRINT: suppresses printing the ridge analysis only when an output data set is required.

OUTR = SAS-data-set: creates an output data set containing the computed optimum ridge.

RADIUS = coded-radii: gives the distances from the ridge starting point at which to compute the optimum.

PROC REG is the primary SAS procedure for performing the computations for a statistical analysis of data based on a linear regression model. The basic statements for performing such an analysis are

PROC REG;

MODEL list of dependent variable = list of independent variables/ model options;

RUN;

The PROC REG procedure and model statement without any option gives ANOVA, root mean square error, R-squares, Adjusted R-square, coefficient of variation etc.

The options under model statement are

P: It gives predicted values corresponding to each observation in the data set. The estimated standard errors are also given by using this option.

CLM: It yields upper and lower 95% confidence limits for the mean of subpopulation corresponding to specific values of the independent variables.

CLI: It yields a prediction interval for a single unit to be drawn at random from a subpopulation.

STB: Standardized regression coefficients.

XPX, I: Prints matrices used in regression computations.

NOINT: This option forces the regression response to pass through the origin. With this option total sum of squares is uncorrected and hence R-square statistic are much larger than those for the models with intercept.

However, if no intercept model is to be fitted with corrected total sum of squares and hence usual definition of various statistic viz R^2 , MSE etc. are to be retained then the option RESTRICT intercept = 0; may be exercised after the model statement.

For obtaining residuals and studentized residuals, the option 'R' may be exercised under model statement and Cook's D statistic.

The 'INFLUENCE' option under model statement is used for detection of outliers in the data and provides residuals, studentized residuals, diagonal elements of HAT MATRIX, COVRATIO, DFFITS, DFBETAS, etc.

For detecting multicollinearity in the data, the options 'VIF' (variance inflation factors) and 'COLLINOINT' or 'COLLIN' may be used.

Besides the options for weighted regression, output data sets, specification error, heterogeneous variances etc. are available under PROC REG.

PROC PRINCOMP can be utilized to perform the principal component analysis.

Multiple model statements are permitted in PROC REG unlike PROC ANOVA and PROC GLM. A model statement can contain several dependent variables.

The statement model $y_1, y_2, y_3, y_4 = x_1 x_2 x_3 x_4 x_5$; performs four separate regression analyses of variables y_1, y_2, y_3 and y_4 on the set of variables x_1, x_2, x_3, x_4, x_5 .

Polynomial models can be fitted by using independent variables in the model as $x1=x$, $x2=x**2$, $x3=x**3$, and so on depending upon the order of the polynomial to be fitted. From a variable, several other variables can be generated before the model statement and transformed variables can be used in model statement. LY and LX gives Logarithms of Y & X respectively to the base e and LogY, LogX gives logarithms of Y and X respectively to the base 10.

TEST statement after the model statement can be utilized to test hypotheses on individual or any linear function(s) of the parameters.

For e.g. if one wants to test the equality of coefficients of $x1$ and $x2$ in $y=\beta_0+\beta_1x1+\beta_2 x2$ regression model, statement

```
TEST 1: TEST  $x1 - x2 = 0$ ;
```

```
Label: Test < equation ..., equation >;
```

The fitted model can be changed by using a separate model statement or by using DELETE variables; or ADD variables; statements.

The PROC REG provides two types of sums of squares obtained by SS1 or SS2 options under model statement. Type I SS are sequential sum of squares and Types II sum of squares are partial SS are same for that variable which is fitted at last in the model.

For most applications, the desired test for a single parameter is based on the Type II sum of squares, which are equivalent to the t-tests for the parameter estimates. The Type I sum of squares, however, are useful if there is a need for a specific sequencing of tests on individual coefficients as in polynomial models.

PROC ANOVA and PROC GLM are general purpose procedures that can be used for a broad range of data classification. In contrast, PROC NESTED is a specialized procedure that is useful only for nested classifications. It provides estimates of the components of variance using the analysis of variance method of estimation. The CLASS statement in PROC NESTED has a broader purpose than it does in PROC ANOVA and PROC GLM; it encompasses the purpose of MODEL statement as well. But the data must be sorted appropriately. For example in a laboratory microbial counts are made in a study, whose objective is to assess the source of variation in number of microbes. For this study n_1 packages of the test material are purchased and n_2 samples are drawn from each package i.e. samples are nested within packages. Logarithm transformation is to be used for microbial counts. PROPER SAS statements are:

```
PROC SORT; By package sample;
```

```
PROC NESTED;
```

```
CLASS package sample;
```

```
Var logcount;
```

```
run;
```

Corresponding PROC GLM statements are

```
PROC GLM;
```

```
Class package sample;
```

```
Model Logcount= package sample (package);
```

The F-statistic in basic PROC GLM output is not necessarily correct. For this RANDOM statement with a list of all random effects in the model is used and Test option is utilized to get correct error term. However, for fixed effect models same arguments for proper error terms hold as in PROC GLM and PROC ANOVA. For the analysis of the data using linear mixed effects model, PROC MIXED of SAS should be used. The best linear unbiased predictors and solutions for random and fixed effects can be obtained by using option 's' in the Random statement.

33.6 PROCEDURES FOR SURVEY DATA ANALYSIS

PROC SURVEYMEANS procedure produces estimates of population means and totals from sample survey data. You can use PROC SURVEYMEANS to compute the following statistics:

- estimates of population means, with corresponding standard errors and *t* tests
- estimates of population totals, with corresponding standard deviations and *t* tests
- estimates of proportions for categorical variables, with standard errors and *t* tests
- ratio estimates of population means and proportions, and their standard errors
- confidence limits for population means, totals, and proportions
- data summary information

PROC SURVEYFREQ procedure produces one-way to *n*-way frequency and crosstabulation tables from sample survey data. These tables include estimates of population totals, population proportions (overall proportions, and also row and column proportions), and corresponding standard errors. Confidence limits, coefficients of variation, and design effects are also available. The procedure also provides a variety of options to customize your table display.

PROC SURVEYREG procedure fits linear models for survey data and computes regression coefficients and their variance-covariance matrix. The procedure allows you to specify classification effects using the same syntax as in the GLM procedure. The procedure also provides hypothesis tests for the model effects, for any specified estimable linear functions of the model parameters, and for custom hypothesis tests for linear combinations of the regression parameters. The procedure also computes the confidence limits of the parameter estimates and their linear estimable functions.

PROC SURVEYLOGISTIC procedure investigates the relationship between discrete responses and a set of explanatory variables for survey data. The procedure fits linear logistic regression models for discrete response survey data by the method of maximum likelihood, incorporating the sample design into the analysis. The SURVEYLOGISTIC procedure enables you to use categorical classification variables (also known as CLASS variables) as explanatory variables in an explanatory model, using the familiar syntax for main effects and interactions employed in the GLM and LOGISTIC procedures.

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or a sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of selection. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability

sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population.

PROC SURVEYSELECT provides methods for both equal probability sampling and sampling with probability proportional to size (PPS). In PPS sampling, a unit's selection probability is proportional to its size measure. PPS sampling is often used in cluster sampling, where you select clusters (groups of sampling units) of varying size in the first stage of selection. Available PPS methods include without replacement, with replacement, systematic, and sequential with minimum replacement. The procedure can apply these methods for stratified and replicated sample designs.

1. Exercises

Example 3.1: An experiment was conducted to study the hybrid seed production of bottle gourd (*Lagenaria siceraria (Mol) Standl*) Cv. Pusa hybrid-3 under open field conditions during Kharif-2005 at Indian Agricultural Research Institute, New Delhi. The main aim of the investigation was to compare natural pollination and hand pollination. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination on number of fruit set for the period of 45 days, fruit weight (kg), seed yield per plant (g) and seedling length (cm). The data obtained is as given below:

Group	No. of fruit	Fruit weight	Seed yield/plant	Seedling length
1	7.0	1.85	147.70	16.86
1	7.0	1.86	136.86	16.77
1	6.0	1.83	149.97	16.35
1	7.0	1.89	172.33	18.26
1	7.0	1.80	144.46	17.90
1	6.0	1.88	138.30	16.95
1	7.0	1.89	150.58	18.15
1	7.0	1.79	140.99	18.86
1	6.0	1.85	140.57	18.39
1	7.0	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.50	18.07
2	7.3	2.58	230.34	19.07
2	8.0	2.62	217.05	19.00
2	8.0	2.68	233.84	18.00
2	8.0	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7.0	2.45	199.87	19.31
2	7.3	2.44	214.30	19.36

{Here 1 denotes natural pollination and 2 denotes the hand pollination}

1. Test whether the mean of the population of Seed yield/plant (g) is 200 or not.
2. Test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different.
3. Test whether hand pollination is better alternative in comparison to natural pollination.

Procedure:

For performing analysis, input the data in the following format. {Here Number of fruit (45 days) is termed as nfs45, Fruit weight (kg) is termed as fw, seed yield/plant (g) is termed as syp and Seedling length (cm) is termed as sl. It may, however, be noted that one can retain the same name or can code in any other fashion}.

```
data ttest1; /*one can enter any other name for data*/
input group  nfs45  fw      syp  sl;
cards;
.....
.....
.....
;
```

*To answer the question number 1 use the following SAS statements

```
proc ttest H0=200;
```

```
var syp;
```

```
run;
```

*To answer the question number 2 use the following SAS statements;

```
proc ttest;
```

```
class group;
```

```
var nfs45 fw syp sl;
```

```
run;
```

To answer the question number 3 one has to perform the one tail t-test. The easiest way to convert a two-tailed test into a one-tailed test is take half of the p-value provided in the output of 2-tailed test output for drawing inferences. The other way is using the options sides in proc statement. Here we are interested in testing whether hand pollination is better alternative in comparison to natural pollination, therefore, we may use Sides=L as

```
proc ttest sides=L;
```

```
class group;
```

```
var nfs45 fw syp sl;
```

```
run;
```

Similarly this option can also be used in one sample test and for right tail test Sides=U is used.

Exercise 3.2: A study was undertaken to find out whether the average grain yield of paddy of farmers using laser levelling is more than the farmers using traditional land levelling methods. For this study data on grain yield in tonne/hectare was collected from 59 farmers (33 using traditional land levelling methods and 26 using new land leveller) and is given as:

Traditional	Laser		Traditional	Laser
3.67	3.6		3.79	3.95
4.04	3.7		3.17	5.3
3.49	5.3		3.58	5.8
2.75	4.4		4.08	2.8
2.63	5.4		4.25	3.0
2.46	3.4		5.21	4.78
2.50	3.5		5.63	4.07
2.88	8.2		3.42	4.88
2.45	7.5		3.88	4.37
2.46	7.6		3.29	
2.67	7.0		3.92	
2.38	7.4		2.25	
2.42	3.4		2.58	
2.54	3.6		3.25	
3.88	5.6		3.46	
3.88	5.6		3.79	
3.42	5.4			

Test whether the traditional land levelling and laser levelling give equivalent yields or are significantly different.

Procedure:

For performing analysis, input the data in the following format. {Here traditional land levelling is termed as LL, laser levelling as LL, method of levelling as MLevel and grain yield in t/ha as gyld. It may, however, be noted that one can retain the same name or can code in any other fashion}.

```
data ttestL; /*one can enter any other name for data*/
input MLevel gyld;
cards;
.....
.....
.....
;
```

*To answer the question number 1 use the following SAS statements

```
proc ttest data =ttestL;
var gyld;
run;
```

Exercise 3.3: The observations obtained from 15 experimental units before and after application of the treatment are the following:

Unit No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	80	73	70	60	88	84	65	37	91	98	52	78	40	79	59
After	82	71	95	69	100	71	75	60	95	99	65	83	60	86	62

1. Test whether the mean score before application of treatment is 65.
2. Test whether the application of treatments has resulted into some change in the score of the experimental units.
3. Test whether application of treatment has improved the scores.

Procedure:

```
data ttest;
```

```
input sn preapp postapp;
```

```
cards;
```

```
1 80 82
```

```
2 73 71
```

```
3 70 95
```

```
4 60 69
```

```
5 88 100
```

```
6 84 71
```

```
7 65 75
```

```
8 37 60
```

```
9 91 95
```

```
10 98 99
```

```
11 52 65
```

```
12 78 83
```

```
13 40 60
```

```
14 79 86
```

```
15 59 62
```

```
;
```

```
*For objective 1, use the following;
```

```
PROC TTEST H0=65;
```

```
VAR PREAPP;
```

```
RUN;
```

```
*For objective 2, use the following;
```

```
PROC TTEST;
```

```
PAIRED PREAPP*POSTAPP;
```

```
RUN;
```

```
*For objective 3, use the following;
```

```
PROC TTEST sides=L;
```

```
PAIRED PREAPP*POSTAPP;
```

```
RUN;
```

Exercise 3.4: In F_2 population of a breeding trial on pea, out of a total of 556 seeds, the frequency of seeds of different shape and colour are: 315 rounds and yellow, 101 wrinkled and yellow, 108 round and green, 32 wrinkled and green. Test at 5% level of significance whether the different shape and colour of seeds are in proportion of 9:3:3:1 respectively.

Procedure:

```
/*rndyel=round and yellow, rndgrn=round and green, wrnkyel=wrinkled and yellow,
wrnkgrn=wrinkled and green*/;
```

```
data peas;
input shape_color $ count;
cards;
rndyel 315
rndgrn 108
wrnkyel 101
wrnkgrn 32
;
proc freq data=peas order=data;
weight count ;
tables shape_color / chisq testp=(0.5625 0.1875 0.1875 0.0625);
exact chisq;
run;
```

Exercise 3.5: The educational standard of adoptability of new innovations among 600 farmers are given as below:

Educational standard			
Adoptability	Matric	Graduate	Illiterate
Adopted	100	60	80
Not adopted	50	20	290

Draw the inferences whether educational standard has any impact on their adoptability of innovation.

Procedure:

```
data innovation;
input edu $ adopt $ count;
cards;
Matric adopt 100
Matric Noadopt 50
grad adopt 60
grad Noadopt 20
illit adopt 80
illit Noadopt 290
;
proc freq order=data;
weight count ;
tables edu*adopt / chisq ;
run;
```

Exercise 3.6: An Experiment was conducted using a Randomized complete block design in 5 treatments a, b, c, d & e with three replications. The data (yield) obtained is given below:

Treatment(TRT)					
Replication(REP)	a	b	c	d	e
1	16.9	18.2	17.0	15.1	18.3
2	16.5	19.2	18.1	16.0	18.3
3	17.5	17.1	17.3	17.8	19.8

1. Perform the analysis of variance of the data.
2. Test the equality of treatment means.
3. Test $H_0: 2T_1=T_2+T_3$, where as T_1, T_2, T_3, T_4 and T_5 are treatment effects.

Procedure:

Prepare a SAS data file using

DATA Name;

INPUT REP TRT \$ yield;

Cards;

...

...

...

;

Print data using PROC PRINT. Perform analysis using PROC ANOVA, obtain means of treatments and obtain pairwise comparisons using least square differences, Duncan's New Multiple range tests and Tukey's Honest Significant difference tests. Make use of the following statements:

PROC Print;

PROC ANOVA;

Class REP TRT;

Model Yield = REP TRT;

Means TRT/lst;

Means TRT/duncan;

Means TRT/tukey;

Run;

Perform contrast analysis using PROC GLM.

Proc glm;

Class rep trt;

Model yld = rep trt;

Means TRT/lst;

Means TRT/duncan;

Means TRT/tukey

Contrast '1 Vs 2&3' trt 2 -1 -1;

Run;

Exercise 3.7: In order to select suitable tree species for Fuel, Fodder and Timber an experiment was conducted in a randomized complete block design with ten different trees and four replications. The plant height was recorded in cm. The details of the experiment are given below:

Plant Height (Cms): Place – Kanpur

Name of Tree	Spacing	Replications			
		1	2	3	4
A. Indica	4x4	144.44	145.11	104.00	105.44
D. Sisso	4x2	113.50	118.61	118.61	123.00
A. Procer	4x2	60.88	90.94	80.33	92.00
A. Nilotic	4x2	163.44	158.55	158.88	153.11
T. Arjuna	4x2	110.11	116.00	119.66	103.22
L. Loucoc	4x1	260.05	102.27	256.22	217.80
M. Alba	4x2	114.00	115.16	114.88	106.33
C. Siamia	4x2	91.94	58.16	76.83	79.50
E. Hybrid	4x1	156.11	177.97	148.22	183.17
A. Catech	4x2	80.2	108.05	45.18	79.55

Analyze the data and draw your conclusions.

Exercise 3.8: An experiment was conducted with 49 crop varieties (TRT) using a simple lattice design. The layout and data obtained (Yld) is as given below:

REPLICATION (REP)-I

BLOCKS(BLK)						
1	2	3	4	5	6	7
22(7)	10(12)	45(22)	37(25)	18(33)	30(33)	5(28)
24(20)	14(26)	44(21)	41(23)	19(17)	34(31)	6(74)
28(25)	8(42)	43(16)	40(11)	21(13)	35(10)	7(14)
27(68)	9(13)	47(37)	42(24)	17(10)	32(12)	2(14)
25(4)	13(10)	49(13)	36(30)	15(36)	29(22)	1(16)
26(11)	12(21)	48(21)	39(34)	20(30)	33(33)	3(11)
23(45)	11(11)	46(12)	38(15)	16(14)	31(18)	4(7)

REPLICATION (REP)-II

BLOCKS(BLK)						
1	2	3	4	5	6	7
22(29)	18(64)	20(25)	23(45)	5(19)	3(13)	14(60)
8(127)	25(31)	27(71)	16(22)	19(47)	24(23)	49(72)
43(119)	46(85)	13(51)	2(13)	47(86)	17(51)	21(10)
1(24)	11(51)	48(121)	37(85)	40(33)	10(30)	42(23)
36(58)	4(39)	41(22)	9(10)	12(48)	31(50)	35(54)
29(97)	39(67)	6(75)	30(65)	33(73)	38(30)	28(54)
15(47)	32(93)	34(44)	44(5)	26(56)	45(103)	7(85)

1. Perform the analysis of variance of the data. Also obtain Type II SS.
2. Obtain adjusted treatment means with their standard errors.
3. Test the equality of all adjusted treatment means.
4. Test whether the sum of 1 to 3 treatment means is equal to the sum of 4 to 6 treatments.
5. Estimate difference between average treatment 1 average of 2 to 4 treatment means.
6. Divide the between block sum of squares into between replication sum of squares and between blocks within replications sum of squares.
7. Assuming that the varieties are a random selection from a population, obtain the genotypic variance.
8. Analyze the data using block effects as random.

PROCEDURE

Prepare the DATA file.

DATA Name;

INPUT REP BLK TRT yield;

Cards;

....

....

....

;

Print data using PROC PRINT. Perform analysis of 1 to 5 objectives using PROC GLM.

The statements are as follows:

Proc print;

Proc glm;

Class rep blk trt;

Model yld= blk trt/ss2;

Contrast 'A' trt 1 1 1 -1 -1 -1;

Estimate 'A' trt 3 -1 -1 -1/divisor=3;

Run;

The objective 6 can be achieved by another model statement.

Proc glm;

Class rep blk trt;

Model yield= rep blk (rep) trt/ss2;

run;

The objective 7 can be achieved by using the another PROC statement

Proc Varcomp Method=type1;

Class blk trt;

Model yield = blk trt/fixed = 1;

Run;

The above obtains the variance components using Hemderson's method. The methods of maximum likelihood, restricted maximum likelihood, minimum quadratic unbiased estimation can also be used by specifying method =ML, REML, MIVQE respectively.

Objective 8 can be achieved by using PROC MIXED.

```
Proc Mixed ratio covtest;
Class blk trt;
Model yield = trt;
Random blk/s;
Lsmeans trt/pdiff;
Store lattice;
Run;
PROC PLM SOURCE = lattice;
LSMEANS trt /pdiff lines;
RUN;
```

Exercise 3.9: Analyze the data obtained through a Split-plot experiment involving the yield of 3 Irrigation (IRRIG) treatments applied to main plots and two Cultivars (CULT) applied to subplots in three Replications (REP). The layout and data (YLD) is given below:

Replication-I			Replication -II			Replication-III		
I1	I2	I3	I1	I2	I3	I1	I2	I3
C1	C1	C1	C1	C1	C1	C1	C1	C1
(1.6)	(2.6)	(4.7)	(3.4)	(4.6)	(5.5)	(3.2)	(5.1)	(5.7)
C2	C2	C2	C2	C2	C2	C2	C2	C2
(3.3)	(5.1)	(6.8)	(4.7)	(1.1)	(6.6)	(5.6)	(6.2)	(4.5)

Perform the analysis of the data. (HINT: Steps are given in text).

Remark 3.9.1: Another way proposed for analysis of split plot designs is using replication as random effect and analyse the data using PROC MIXED of SAS. For the above case, the steps for using PROC MIXED are:

```
PROC MIXED COVTEST;
CLASS rep irrig cult;
MODEL yield = irrig cult irrig*cult / DDFM=KR;
RANDOM rep rep*irrig;
LSMEANS irrig cult irrig*cult / PDIFF;
STORE spd;
run;
/* An item store is a special SAS-defined binary file format used to store and restore information with a hierarchical structure*/

/* The PLM procedure performs post fitting statistical analyses for the contents of a SAS item store that was previously created with the STORE statement in some other SAS/STAT procedure*/
PROC PLM SOURCE = SPD;
LSMEANS irrig cult irrig*cult /pdiff lines;
RUN;
```

Remark 3.9.2: In Many experimental situations, the split plot designs are conducted across environments and a pooled is required. One way of analysing data of split plot designs with two factors A and B conducted across environment is

```
PROC MIXED COVTEST;
CLASS year rep a b;
MODEL yield = a b a*b / DDFM=KR;
/* DDFM specifies the method for computing the denominator degrees of freedom for the tests of fixed
effects resulting from the MODEL*/
RANDOM year rep(year) year*a year*rep*a year*a*b;
LSMEANS a b a*b / PDIF;
STORE spd1;
run;
PROC PLM SOURCE = SPD1;
LSMEANS a b a*b/pdiff lines;
RUN;
```

Exercise 3.10: An agricultural field experiment was conducted in 9 treatments using 36 plots arranged in 4 complete blocks and a sample of harvested output from all the 36 plots are to be analysed blockwise by three technicians using three different operations. The data collected is given below:

Block-1				Block-2			
Technician				Technician			
Operation	1	2	3	Operation	1	2	3
1	1(1.1)	2(2.1)	3(3.1)	1	1(2.1)	4(5.2)	7(8.3)
2	4(4.2)	5(5.3)	6(6.3)	2	2(3.2)	5(6.7)	8(9.9)
3	7(7.4)	8(8.7)	9(9.6)	3	3(4.5)	6(7.6)	9(10.3)
Block-3				Block-4			
Technician				Technician			
Operation	1	2	3	Operation	1	2	3
1	1(1.2)	6(6.3)	8(8.7)	1	1(3.1)	9(11.3)	5(7.8)
2	9(9.4)	2(2.7)	4(4.8)	2	6(8.1)	2(4.5)	7(9.3)
3	5(5.9)	7(7.8)	3(3.3)	3	8(10.7)	4(6.9)	3(5.8)

1. Perform the analysis of the data considering that technicians and operations are crossed with each other and nested in the blocking factor.
2. Perform the analysis by considering the effects of technicians as negligible.
3. Perform the analysis by ignoring the effects of the operations and technicians.

Procedure:

Prepare the data file.

```
DATA Name;
INPUT BLK TECH OPER TRT OBS;
Cards;
....
....
....
```

;

Perform analysis of objective 1 using PROC GLM. The statements are as follows:

```
Proc glm;
Class blk tech oper trt;
Model obs= blk tech (blk) oper(blk) trt/ss2;
Lsmeans trt oper(blk)/pdiff;
Run;
```

Perform analysis of objective 2 using PROC GLM with the additional statements as follows:

```
Proc glm;
Class blk tech oper trt;
Model obs= blk oper(blk) trt/ss2;
run;
```

Perform analysis of objective 3 using PROC GLM with the additional statements as follows:

```
Proc glm;
Class blk tech oper trt;
Model obs = blk trt/ss2;
run;
```

Exercise 3.11: A greenhouse experiment on tobacco mosaic virus was conducted. The experimental unit was a single leaf. Individual plants were found to be contributing significantly to error and hence were taken as one source causing heterogeneity in the experimental material. The position of the leaf within plants was also found to be contributing significantly to the error. Therefore, the three positions of the leaves *viz.* top, middle and bottom were identified as levels of second factor causing heterogeneity. 7 solutions were applied to leaves of 7 plants and number of lesions produced per leaf was counted. Analyze the data of this experiment.

Plants							
Leaf Position	1	2	3	4	5	6	7
Top	1(2)	2(3)	3(1)	4(5)	5(3)	6(2)	7(1)
Middle	2(4)	3(3)	4(2)	5(6)	6(4)	7(2)	1(1)
Bottom	4(3)	5(4)	6(7)	7(6)	1(3)	2(4)	3(7)

The figures at the intersections of the plants and leaf position are the solution numbers and the figures in the parenthesis are number of lesions produced per leaf.

Procedure:

```
Prepare the data file.
DATA Name;
INPUT plant posi $ trt count;
Cards;
```

.....

 ;

Perform analysis using PROC GLM. The statements are as follows:

```
Proc glm;
Class plant posi trt count;
Model count= plant posi trt/ss2;
Lsmeans trt/pdiff; Run;
```

Exercise 3.12: The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

1. Obtain correlation coefficient between each pair of the variables PP, PH, NGL and yield.
2. Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables. Print the matrices used in the regression computations.
3. Test the significance of the regression coefficients and also equality of regression coefficients of a) PP and PH b) PH and NGL
4. Obtain the predicted values corresponding to each observation in the data set.
5. Identify the outliers in the data set.
6. Check for the linear relationship among the biometrical characters.
7. Fit the model without intercept.
8. Perform principal component analysis.

No.	PP	PH	NGL	Yield
1	142.00	0.5250	8.20	2.470
2	143.00	0.6400	9.50	4.760
3	107.00	0.6600	9.30	3.310
4	78.00	0.6600	7.50	1.970
5	100.00	0.4600	5.90	1.340
6	86.50	0.3450	6.40	1.140
7	103.50	0.8600	6.40	1.500
8	155.99	0.3300	7.50	2.030
9	80.88	0.2850	8.40	2.540
10	109.77	0.5900	10.60	4.900
11	61.77	0.2650	8.30	2.910
12	79.11	0.6600	11.60	2.760
13	155.99	0.4200	8.10	0.590
14	61.81	0.3400	9.40	0.840
15	74.50	0.6300	8.40	3.870
16	97.00	0.7050	7.20	4.470
17	93.14	0.6800	6.40	3.310
18	37.43	0.6650	8.40	1.570

19	36.44	0.2750	7.40	0.530
20	51.00	0.2800	7.40	1.150
21	104.00	0.2800	9.80	1.080
22	49.00	0.4900	4.80	1.830
23	54.66	0.3850	5.50	0.760
24	55.55	0.2650	5.00	0.430
25	88.44	0.9800	5.00	4.080
26	99.55	0.6450	9.60	2.830
27	63.99	0.6350	5.60	2.570
28	101.77	0.2900	8.20	7.420
29	138.66	0.7200	9.90	2.620
30	90.22	0.6300	8.40	2.000
31	76.92	1.2500	7.30	1.990
32	126.22	0.5800	6.90	1.360
33	80.36	0.6050	6.80	0.680
34	150.23	1.1900	8.80	5.360
35	56.50	0.3550	9.70	2.120
36	136.00	0.5900	10.20	4.160
37	144.50	0.6100	9.80	3.120
38	157.33	0.6050	8.80	2.070
39	91.99	0.3800	7.70	1.170
40	121.50	0.5500	7.70	3.620
41	64.50	0.3200	5.70	0.670
42	116.00	0.4550	6.80	3.050
43	77.50	0.7200	11.80	1.700
44	70.43	0.6250	10.00	1.550
45	133.77	0.5350	9.30	3.280
46	89.99	0.4900	9.80	2.690

Procedure:

Prepare a data file

Data mlr;

Input PP PH NGL Yield;

Cards;

....

....

;

For obtaining correlation coefficient, Use PROC CORR;

Proc Corr;

Var PP PH NGL Yield;

run;

For fitting of multiple linear regression equation, use PROC REG

Proc Reg;

Model Yield = PP PH NGL/ p r influence vif collin xpx i;

Test 1: Test $PP = 0$; Test 2: Test $PH = 0$;

Test 3: Test $NGL = 0$;

Test 4: Test $PP - PH = 0$;

Test 4a: Test $PP = PH = 0$;

Test 5: Test $PH - NGL = 0$;

Test 5a: Test $PH = NGL = 0$;

Model Yield = PP PH NGL/noint;

run;

Proc reg;

Model Yield = PP PH NGL;

Restrict intercept =0;

Run;

For diagnostic plots

Proc Reg plots(unpack)=diagnostics;

Model Yield = PP PH NGL;

run;

For variable selection, one can use the following option in model statement:

Selection=stepwise sls=0.10;

For performing principal component analysis, use the following:

PROC PRINCOMP;

VAR PP PH NGL YIELD;

run;

Example 3.13: An experiment was conducted at Division of Agricultural Engineering, IARI, New Delhi for studying the capacity of a grader in number of hours when used with three different speeds and two processor settings. The experiment was conducted using a factorial completely randomised design in 3 replications. The treatment combinations and data obtained on capacity of grader in hours given as below:

Replication	speed	Processor setting	trt	cgrader
1	1	1	1	1852
1	1	2	2	1848
1	1	3	3	1855
1	2	1	4	2270
1	2	2	5	2279
1	2	3	6	2272
1	3	1	7	3035
1	3	2	8	3042
1	3	3	9	3028
2	1	1	1	1845
2	1	2	2	1855
2	1	3	3	1860
2	2	1	4	2276
2	2	2	5	2275
2	2	3	6	2248

2	3	1	7	3036
2	3	2	8	3033
2	3	3	9	3038
3	1	1	1	1851
3	1	2	2	1840
3	1	3	3	1840
3	2	1	4	2265
3	2	2	5	2280
3	2	3	6	2278
3	3	1	7	3040
3	3	2	8	3028
3	3	3	9	3040

Experimenter was interested in identifying the best combination of speed and processor setting that gives maximum capacity of the grader in hours.

Solution: This data can be analysed as per procedure of factorial CRD and one can use the following SAS steps for performing the analysis:

Data ex1a;

Input rep speed proset cgrader;

/*here rep: replication; proset: processor setting and cgrader: capacity of the grader in hours*/

Cards;

```

1 1 1 1852
1 1 2 1848
1 1 3 1855
. . . .
. . . .
. . . .
3 3 1 3040
3 3 2 3028
3 3 3 3040

```

;

Proc glm data=ex1;

Class speed proset;

Model cgrader=speed proset speed*proset;

Lsmmeans speed proset speed*proset/pdiff adjust=tukey lines;

Run;

The above analysis would identify test the significance of main effects of speed and processor setting and their interaction. Through this analysis one can also identify the speed level (averaged over processor setting) {Processor Setting (averaged over speed levels)} at which the capacity of the grader is maximum. The multiple comparisons

between means of combinations of speed and processor setting would help in identifying the combination at which capacity of the grader is maximum.

Exercise 3.14: An experiment was conducted with five levels of each of the four fertilizer treatments nitrogen, Phosphorus, Potassium and Zinc. The levels of each of the four factors and yield obtained are as given below. Fit a second order response surface design using the original data. Test the lack of fit of the model. Compute the ridge of maximum and minimum responses. Obtain predicted residual Sum of squares.

N	P ₂ O ₅	K ₂ O	Zn	Yield
40	30	25	20	11.28
40	30	25	60	8.44
40	30	75	20	13.29
40	90	25	20	7.71
120	30	25	20	8.94
40	30	75	60	10.9
40	90	25	60	11.85
120	30	25	60	11.03
120	30	75	20	8.26
120	90	25	20	7.87
40	90	75	20	12.08
40	90	75	60	11.06
120	30	75	60	7.98
120	90	75	60	10.43
120	90	75	20	9.78
120	90	75	60	12.59
160	60	50	40	8.57
0	60	50	40	9.38
80	120	50	40	9.47
80	0	50	40	7.71
80	60	100	40	8.89
80	60	0	40	9.18
80	60	50	80	10.79
80	60	50	0	8.11
80	60	50	40	10.14
80	60	50	40	10.22
80	60	50	40	10.53
80	60	50	40	9.5
80	60	50	40	11.53
80	60	50	40	11.02

Procedure:

Prepare a data file.

```
/* yield at different levels of several factors */
```

```
title 'yield with factors N P K Zn';
```

```
data dose;
```

```
input n p k Zn y ; label y = "yield" ;
```



```

cards;
.....
.....
.....
;
*Use PROC RSREG.
ods graphics on;
proc rsreg data=dose plots(unpack)=surface(3d);
model y= n p k Zn/ nocode lackfit press;
run;
ods graphics off; *If we do not want surface plots, then we may
proc rsreg;
model y= n p k Zn/ nocode lackfit press;
Ridge min max;
run;

```

Exercise 3.15: Fit a second order response surface design to the following data. Take replications as covariate.

Fertilizer1	Fertilizer2	X ₁	X ₂	Yields(lb/plot)	
				Replication I	Replication II
50	15	-1	-1	7.52	8.12
120	15	+1	-1	12.37	11.84
50	25	-1	+1	13.55	12.35
120	25	+1	+1	16.48	15.32
35	20	$-\sqrt{2}$	0	8.63	9.44
134	20	$+\sqrt{2}$	0	14.22	12.57
85	13	0	$-\sqrt{2}$	7.90	7.33
85	27	0	$+\sqrt{2}$	16.49	17.40
85	20	0	0	15.73	17.00

Procedure:

```

Prepare a data file.
/* yield at different levels of several factors */
title 'yield with factors x1 x2';
data respcov;
input fert1 fert2 x1 x2 yield ;
cards;
.....
.....
.....
;
/*Use PROC RSREG.*/
ODS Graphics on;
proc rsreg plots(unpack)=surface(3d);
model yield = rep fert1 fert2/ covar=1 nocode lackfit ;
Ridge min max;
run;

```

ods graphics off;

Exercise 3.16: Following data is related to the length(in cm) of the ear-head of a wheat variety 9.3, 18.8, 10.7, 11.5, 8.2, 9.7, 10.3, 8.6, 11.3, 10.7, 11.2, 9.0, 9.8, 9.3, 10.3, 10.1 9.6, 10.4. Test the data that the median length of ear-head is 9.9 cm.

Procedure:

This may be tested using any of the three tests for location available in Proc Univariate viz. Student's *t* test, the sign test, and the Wilcoxon signed rank test. All three tests produce a test statistic for the null hypothesis that the mean or median is equal to a given value μ_0 against the two-sided alternative that the mean or median is not equal to μ_0 . By default, PROC UNIVARIATE sets the value of μ_0 to zero. You can use the MU0= option in the PROC UNIVARIATE statement to specify the value of μ_0 . If the data is from a normal population, then we can infer using t-test otherwise non-parametric tests sign test, and the Wilcoxon signed rank test may be used for drawing inferences.

Procedure:

```

data npsign;
input length;
cards;
9.3
18.8
10.7
11.5
 8.2
 9.7
10.3
 8.6
11.3
10.7
11.2
 9.0
 9.8
 9.3
10.3
10.0
10.1
 9.6
10.4
;
PROC UNIVARIATE DATA=npsign MU0=9.9;
VAR length;
HISTOGRAM / NOPLOT ;
      RUN;
QUIT;

```

Exercise 3.17: An experiment was conducted with 21 animals to determine if the four different feeds have the same distribution of Weight gains on experimental animals. The feeds 1, 3 and 4 were given to 5 randomly selected animals and feed 2 was given to 6 randomly selected animals. The data obtained is presented in the following table.

Feeds	Weight gains (kg)					
1	3.35	3.8	3.55	3.36	3.81	
2	3.79	4.1	4.11	3.95	4.25	4.4
3	4	4.5	4.51	4.75	5	
4	3.57	3.82	4.09	3.96	3.82	

Procedure:

```
data np;
```

```
input feed wt;
```

```
datalines;
```

```
1 3.35
```

```
1 3.80
```

```
1 3.55
```

```
1 3.36
```

```
1 3.81
```

```
2 3.79
```

```
2 4.10
```

```
2 4.11
```

```
2 3.95
```

```
2 4.25
```

```
2 4.40
```

```
3 4.00
```

```
3 4.50
```

```
3 4.51
```

```
3 4.75
```

```
3 5.00
```

```
4 3.57
```

```
4 3.82
```

```
4 4.09
```

```
4 3.96
```

```
4 3.82
```

```
;
```

```
PROC NPAR1WAY DATA=np WILCOXON; /*for performing Kruskal-Walis test*/;
```

```
VAR wt;
```

```
CLASS feed;
```

```
RUN;
```

Exercise 3.18: Finney (1971) gave a data representing the effect of a series of doses of carotene (an insecticide) when sprayed on *Macrosiphoniella sanborni* (some obscure insects). The Table below contains the concentration, the number of insects tested at each dose, the proportion dying and the probit transformation (probit+5) of each of the observed proportions.

Concentration (mg/l)	No. of insects (n)	No. of affected (r)	%kill (P)	Log concentration (x)	Empirical probit
10.2	50	44	88	1.01	6.18
7.7	49	42	86	0.89	6.08
5.1	46	24	52	0.71	5.05
3.8	48	16	33	0.58	4.56
2.6	50	6	12	0.41	3.82
0	49	0	0	-	-

Perform the probit analysis on the above data.

Procedure

```
data probit;
input con n r;
datalines;
10.2 50 44
7.7 49 42
5.1 46 24
3.8 48 16
2.6 50 6
0 49 0
```

```
;
```

```
ods html;
```

```
Proc Probit log10 ;
```

```
Model r/n=con/lackfit inversecl;
```

```
title ('output of probit analysis');
```

```
run;
```

```
ods html close;
```

Model Information	
Data Set	WORK.PROBIT
Events Variable	r
Trials Variable	n
Number of Observations	5
Number of Events	132
Number of Trials	243
Name of Distribution	Normal
Log Likelihood	-120.0516414

Model Information	
Number of Observations Read	6
Number of Observations Used	5
Number of Events	132
Number of Trials	243

Algorithm converged.

Goodness-of-Fit Tests			
Statistic	Value	DF	Pr > ChiSq
Pearson Chi-Square	1.7289	3	0.6305
L.R. Chi-Square	1.7390	3	0.6283

Response-Covariate Profile	
Response Levels	2
Number of Covariate Values	5

Since the chi-square is small ($p > 0.1000$), fiducial limits will be calculated using a t value of 1.96

Type III Analysis of Effects			
Wald			
Effect	DF	Chi-Square	Pr > ChiSq
Log10(con)	1	77.5920	<.0001

Analysis of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-2.8875	0.3501	-3.5737	-2.2012	68.01	<.0001
Log10(con)	1	4.2132	0.4783	3.2757	5.1507	77.59	<.0001

Probit Model in Terms of Tolerance Distribution	
MU	SIGMA
0.68533786	0.23734947

Estimated Covariance Matrix for Tolerance Parameters		
	MU	SIGMA
MU	0.000488	-0.000063
SIGMA	-0.000063	0.000726

Probit Analysis on Log10(con)			
Probability	Log10(con)	95% Fiducial Limits	
0.01	0.13318	-0.03783	0.24452
0.02	0.19788	0.04453	0.29830
0.03	0.23893	0.09668	0.33253
0.04	0.26981	0.13584	0.35834
0.05	0.29493	0.16764	0.37940
0.06	0.31631	0.19466	0.39737
0.07	0.33506	0.21832	0.41316
0.08	0.35184	0.23946	0.42733
0.09	0.36711	0.25866	0.44026
0.10	0.38116	0.27631	0.45218
0.15	0.43934	0.34898	0.50192
0.20	0.48558	0.40618	0.54202

Probit Analysis on Log10(con)			
Probability	Log10(con)	95% Fiducial Limits	
0.25	0.52525	0.45467	0.57700
0.30	0.56087	0.49759	0.60904
0.35	0.59388	0.53666	0.63942
0.40	0.62521	0.57295	0.66905
0.45	0.65551	0.60716	0.69861
0.50	0.68534	0.63983	0.72870
0.55	0.71516	0.67142	0.75986
0.60	0.74547	0.70240	0.79265
0.65	0.77679	0.73330	0.82766
0.70	0.80980	0.76480	0.86563
0.75	0.84543	0.79777	0.90761
0.80	0.88510	0.83352	0.95533
0.85	0.93133	0.87427	1.01188
0.90	0.98951	0.92456	1.08401
0.91	1.00357	0.93658	1.10155
0.92	1.01883	0.94960	1.12065
0.93	1.03562	0.96387	1.14170
0.94	1.05436	0.97976	1.16526
0.95	1.07574	0.99783	1.19218
0.96	1.10086	1.01898	1.22388
0.97	1.13174	1.04490	1.26294
0.98	1.17279	1.07924	1.31498
0.99	1.23750	1.13315	1.39721

Probit Analysis on con			
Probability	con	95% Fiducial Limits	
0.01	1.35888	0.91657	1.75599
0.02	1.57718	1.10799	1.98745
0.03	1.73353	1.24935	2.15043
0.04	1.86129	1.36724	2.28215
0.05	1.97212	1.47110	2.39553
0.06	2.07163	1.56554	2.49671
0.07	2.16302	1.65317	2.58917
0.08	2.24825	1.73565	2.67506
0.09	2.32868	1.81410	2.75586
0.10	2.40526	1.88932	2.83257
0.15	2.75005	2.23349	3.17629
0.20	3.05900	2.54788	3.48353
0.25	3.35157	2.84884	3.77571
0.30	3.63808	3.14478	4.06477
0.35	3.92538	3.44084	4.35935
0.40	4.21897	3.74068	4.66710
0.45	4.52389	4.04724	4.99582
0.50	4.84549	4.36343	5.35423
0.55	5.18995	4.69265	5.75260
0.60	5.56506	5.03963	6.20374
0.65	5.98127	5.41132	6.72450
0.70	6.45363	5.81830	7.33883
0.75	7.00531	6.27722	8.08377
0.80	7.67532	6.81590	9.02252
0.85	8.53758	7.48633	10.27723
0.90	9.76143	8.40534	12.13411

Probit Analysis on con			
Probability	con	95% Fiducial Limits	
0.91	10.08243	8.64132	12.63428
0.92	10.44313	8.90434	13.20233
0.93	10.85466	9.20181	13.85792
0.94	11.33346	9.54469	14.63036
0.95	11.90537	9.95006	15.56609
0.96	12.61427	10.44674	16.74479
0.97	13.54388	11.08927	18.32046
0.98	14.88655	12.00168	20.65263
0.99	17.27807	13.58779	24.95808

Interpretation: The goodness-of-fit tests (p-values = 0.6305, 0.6283) suggest that the distribution and the model fits the data adequately. In this case, the fitting is done on normal equivalent deviate only without adding 5. Therefore, log LD50 or lof ED50 corresponds to the value of Probit=0. Log LD50 is obtained as 0.685338. Therefore, the stress level at which the 50% of the insects will be killed is $(10^{0.685338} = 4.845 \text{ mg/l})$. Similarly the stress level at which 65% of the insects will be killed is $(10^{0.776793} = 5.981 \text{ mg/l})$. Although both values are given in the table above.

33.7 Discussion

We have initiated a link “Analysis of Data” at Design Resources Server (www.iasri.res.in/design) to provide steps of analysis of data generated from designed experiments by using statistical packages like SAS, SPSS, MINITAB, and SYSTAT, MS-EXCEL etc. For details and live examples one may refer to the link Analysis of data at <http://www.iasri.res.in/design/Analysis%20of%20data/Analysis%20of%20Data.html>.

How to see SAS/STAT Examples?

One can learn from the examples available at

<http://support.sas.com/rnd/app/examples/STATexamples.html>

How to use HELP?

Help → SAS help and Documentation → Contents → Learning to use SAS → Sample SAS Programs → SAS/STAT ...

Strengthening Statistical Computing for NARS

NAIP Consortium on Strengthening Statistical Computing for NARS (www.iasri.res.in/sscnars) targets at providing

- research guidance in statistical computing and computational statistics and creating sound and healthy statistical computing environment
- Providing advanced, versatile, and innovative and state-of the art high end statistical packages to enable them to draw meaningful and valid inferences from their research.

The efforts also involve designing of intelligent algorithms for implementing statistical techniques particularly for analysing massive data sets, simulation, bootstrap, etc.

The objectives of the consortium are:

- To strengthen the high end statistical computing environment for the scientists in NARS;
- To organize customized training programmes and also to develop training modules and manuals for the trainers at various hubs; and
- To sensitize the scientists in NARS with the statistical computing capabilities available for enhancing their computing and research analytics skills.

This consortium has provided the platform for closer interactions among all NARS organizations.

Capacity Building

For capacity building of researchers in the usage of high end statistical computing facility and statistical techniques,

- **209** trainers have been trained through 30 working days training programmes;
- **2166** researchers have been trained through 104 training programmes of one week duration each in the usage.

The capacity building efforts have paved the way for publishing research papers in the high impact factor journals.

33.8 Indian NARS Statistical Computing Portal

For providing service oriented computing, developed and established Indian NARS Statistical Computing portal, which is available to NARS users through IP authentication at <http://stat.iasri.res.in/sscnarsportal>. Any researcher from Indian NARS may obtain User name and password from Nodal Officers of their respective NARS organizations, list available at www.iasri.res.in/sscnars. It is a paradigm of computing techniques that operate on software-as-a-service). There is no need of installation of statistical package at client side. Following 24 different modules of analysis of data are available on this portal, which have been classified into four broad categories as

Basic Statistics

- Descriptive Statistics
- Univariate Distribution Fitting
- Test of Significance based on t-test
- Test of Significance based on Chi-square test
- Correlation Analysis
- Regression Analysis

Designs of Experiments

- Completely randomized designs
- Block Designs (includes both complete and incomplete block designs)
- Combined Block Designs
- Augmented Block Designs
- Resolvable Block Designs
- Nested Block Designs
- Row-Column Designs

SAS FOR STATISTICAL PROCEDURES

- Cross Over Designs
- Split Plot Designs
- Split-Split-Plot Designs
- Split Factorial (main A, sub B × C) designs
- Split Factorial (main A×B, sub C×D) designs
- Strip Plot Designs
- Response Surface Designs

Multivariate Analysis

- Principal Component Analysis
- Linear Discriminant Analysis

Statistical Genetics

- Estimation of Heritability from half- sib data
- Estimation of variance-Covariance matrix from Block Designs

The above modules can be used by uploading *.xlsx, *.csv and *.txt files and results can be saved as *.RTF or *.pdf files. This has helped them in analyzing their data in an efficient manner without losing any time.

Method	Data Type	Create Plots	Simple Statistics
<input type="checkbox"/> Covariance	<input checked="" type="checkbox"/> CENTERED DATA	<input type="checkbox"/> None	<input type="checkbox"/> Yes
<input type="checkbox"/> Correlation	<input type="checkbox"/> ORIGINAL DATA	<input type="checkbox"/> Histogram	<input type="checkbox"/> Mean
<input type="checkbox"/> PCA Scores		<input type="checkbox"/> Plot	<input type="checkbox"/> Score

Eigenvalues	Ph1st	Ph2d	Ph3c	Ph4d
ph	0.899749	0.19884	0.10263	0.00088
ph	0.891578	0.12541	0.04280	0.99900
ngl	0.814672	0.30262	0.37011	0.05479
ylt	0.818892	0.39577	0.92790	0.04505

Requirements of Excel Files during analysis over Indian NARS Statistical Computing Portal

1. Excel file must have the .xls, .xlsx, .csv or .txt extensions

2. This system will only consider the first sheet of the excel file which has name appearing first in lexicographic order. It will not analyze the data which lies in subsequent sheets in excel file.
3. Do not put period (.) or Zero (0) to display missing values in the treatment. It will not consider as missing. Please leave the missing observations as blank cells.
4. If you are getting some wrong analysis then kindly check your excel file. Go to First Column, first cell and then press Ctrl+Shift+End. It will select all the filled rows and columns. If it selects some missing rows and columns then kindly delete those rows and columns otherwise it will give wrong analysis result.
5. Do not use special characters in the variable/column names. Also variable names should not start with spaces.
6. Do not use any formatting to the Excel sheet including formats or expressions to the cell values. It should be data value.
7. If the First row cells has been merged then it will not detect as Column/Variable names.
8. If any rows or columns are hidden then it will be displayed during the analysis.

Basic Statistics

9. **Descriptive Statistics:** The data file should contain at least one quantitative analysis variable.
10. **Univariate Distribution Fitting:** The data file should contain at least one quantitative numeric variable.
11. **Test of Significance based on t-distribution:** The data file should contain at least one quantitative variable name and one classificatory variable.
12. **Chi-Square Test:** The data file should contain at least one categorical variable and weights or frequency counts variable if frequencies are entered in a separate column. Data may also have classificatory in it.
13. **Correlation:** The data file should contain at least two quantitative variables.
14. **Regression Analysis:** The data file should contain at least one Dependent and one Independent variable.

Design of Experiments

15. **Unblock Design:** Prepare a data file containing one variable to describe the Treatment details and at least one response/ dependent variable in the experimental data to be analyzed. Also, the treatment details may be coded or may have actual names (i.e. data values, for variable describing treatment column may be in numeric or character). The maximum length of treatment value is 20 characters. The variables can be entered in any order.
16. **Block Design:** Prepare a data file containing two variables to describe the block and treatment details. There should be at least one response/ dependent variable in the experimental data to be analyzed. Also, the block/treatment details may be coded or may have actual names (i.e. data values, for variables describing block and treatment column may be in numeric or character). The maximum length of treatment value is 20 character. The variables can be entered in any order. (These conditions are applicable to other similar experimental designs also)
17. **Combined Block Design:** The data file should contain three variables to describe Environment, Block, Treatment variables and at least one Dependent variable.

18. **Augmented Block Design:** The data file should contain two variables to describe Block & Treatment variables and at least one Dependent variable. At present, Portal supports only numeric treatment and block variables for augmented designs. An augmented block design involves two sets of treatments known as check or control and test treatments. The treatments should be numbered in such a fashion that the check or control treatments are numbered first followed by test treatments. For example, if there are 4 control treatments and 8 test treatments, then the control treatments are renumbered as 1, 2, 3, 4 and tests are renumbered as 5, 6, 7, 8, 9, 10, 11, 12.
19. **Resolvable Block Design:** The data file should contain three variables to describe the Replication, Block, Treatment variables and at least one Dependent/ response variable.
20. **Nested Block Design:** The data file should contain three variables to describe Block, SubBlock, Treatment variables and at least one Dependent variable.
21. **Row Column Design:** The data file should contain three variables to describe Row, Column, Treatment variables and at least one Dependent variable.
22. **Crossover Design:** Create a data file with at least 5 variables, one for units, one for periods, one treatments, one for residual, and one for the dependent or analysis variable. For performing analysis using the portal, please rearrange the data in the following order: animal numbers as units; periods can be coded as 1, 2, 3, and so on, treatments as alphabets or numbers (coding could be done as follows: for every first period the number one has assigned (fixed) and for other periods code 1 to 3 are given according to the treatment received by the unit in the previous period) and residual effect as residual. It may, however, be noted that one can retain the same name or can code in any other fashion. A carry-over or residual term has the special property as a factor, or class variate, of having no level in the first period because the treatment in the first period is not affected by any residual or carry over effect of any treatment. When we consider the residual or carryover effect in practice the fact that carry-over or residual effects will be adjusted for period effects (by default all effects are adjusted for all others in these analysis). As a consequence, any level can be assigned to the residual variate in the first period, provided the same level is always used. An adjustment for periods then removes this part of the residual term. (For details a reference may made to Jones, B. and Kenward, M.G. 2003. Design and Analysis of Cross Over Trials. Chapman and Hall/CRC. New York . Pp: 212)
23. **Split Plot Design:** The data file should contain three variables to describe Replication, Main Plot, Sub Plot variables and at least one Dependent variable.
24. **Split Split Plot Design:** The data file should contain four variables to describe Replication, Main Plot, Sub Plot, and Sub-Sub Plot Treatment variables and at least one Dependent variable.
25. **Split Factorial (Main A, Sub B×C) Plot Design** The data file should contain four variables to describe Replication, Main Plot, Sub Plot(1){levels of factor 1 in sub plot} , and Sub Plot(2)){levels of factor 21 in sub plot} Treatment variables and at least one Dependent variable.

26. **Split Factorial (Main A×B, Sub C×D) Plot Design:** Create a data file with at least 6 variables, one for block or replication, one for main plot- treatment factor 1, one main plot- treatment factor 2, one for subplot- treatment factor 1, one for subplot- treatment factor 2 and at least one for the dependent or analysis variable. If the data on more than one dependent variable is collected in the same experiment, the data on all variables may be entered in additional columns. One may give actual levels used for different factors applied in main plot-treatment factor 1, main plot- treatment factor 2, subplot- treatment factor 1 and subplot-treatment factor 2. Please remember that there should not be any space between a single data value. Main plot- treatment factor 1, main plot- treatment factor 2, subplot- treatment factor 1, subplot- treatment factor 2 treatments and block numbers may be coded as 1, 2, 3 and so on. One can have character values also.
27. **Strip Plot Design:** The data file should contain at least 4 variables to describe Replication, Horizontal Strip, Vertical Strip variables and at least one Dependent variable.
28. **Response Surface Design:** The data file should contain at least one treatment factor variable and at least one dependent variable

Multivariate Analysis

29. **Principal Component Analysis:** The data file should contain at least one quantitative analysis variable.
30. **Discriminant Analysis:** The data file should contain at least one quantitative analysis variable and a classificatory variable.

Statistical Genetics

31. **Genetic Variance Covariance:** Create a data file with at least 4 variables, one for blocking variable, one for treatments and at least two analysis variable.
32. **Heritability Estimation from Half-Sib Data:** The data file should contain at least one quantitative analysis variable and a classificatory variable.

Other IP Authenticated Services

Following can also be accessed through IP authenticated networks:

- Web Report Studio: <http://stat.iasri.res.in/sscnarswebreportstudio>
- BI DashBoard: <http://stat.iasri.res.in/sscnarsbidashboard>
- Web OLAP Viewer: <http://sas.iasri.res.in:8080/sscnarswebolapviewer>
- E-Miner 6.1: <http://sas.iasri.res.in:6401/AnalyticsPlatform>
- E-Miner 7.1: <http://stat.iasri.res.in/SASEnterpriseMinerJWS/Status>

Accessing SAS E-Miner through URL (IP Authenticated Services)

For Accessing E-miner 6.1 and 7.1 through URLs, following ports should be open

Server	Ports
1) Metadata server	8561
2) Object spawner	8581
3) Table Server	2171
4) Remote Server	5091

5) SAS App. Olap Server	5451
6) SAS Deployment Tester Server	10021
7) Analytics Platform Server	6411
8) Framework Server	22031

However, if you are accessing only E-miner 6.1, then following port need not be opened.

Framework Server	22031
------------------	-------

Steps for accessing SAS Enterprise Miner 6.1 and SAS Enterprise Miner 7.1 separately

SAS Enterprise Miner 6.1

Pre-requisite:

- JRE 1.5 Update 15
- If Firewall and proxy has been implemented then kindly open following ports:

Server	Ports
1) Metadata server	8561
2) Object spawner	8581
3) Table Server	2171
4) Remote Server	5091
5) SAS App. OLAP Server	5451
6) SAS Deployment Tester Server	10021
7) Analytics Platform Server	6411

Steps to be followed:

- If you have installed multiple Java Runtime Environment then
Go to Control Panel → Java → Java tab → View → Keep check on JRE 1.5.0_15 and Uncheck all others
- Check the entry of the **sas.iasri.res.in** in the host file, if not then open host file **C:\Windows\System32\drivers\etc** and edit the host file by entering the IP as shown below or specify the internal/external IP given by IASRI. Internal IP is to be specified only at IASRI, New Delhi. All other NARS organizations should specify external IP only which is: 203.197.217.209 sas.iasri.res.in sas as shown below

```

hosts - Notepad
File Edit Format View Help
# Copyright (c) 1993-2009 Microsoft Corp.
#
# This is a sample HOSTS file used by Microsoft TCP/IP for Windows.
#
# This file contains the mappings of IP addresses to host names. Each
# entry should be kept on an individual line. The IP address should
# be placed in the first column followed by the corresponding host name.
# The IP address and the host name should be separated by at least one
# space.
#
# Additionally, comments (such as these) may be inserted on individual
# lines or following the machine name denoted by a '#' symbol.
#
# For example:
#
#       102.54.94.97       rhino.acme.com       # source server
#       38.25.63.10      x.acme.com        # x client host
#
# localhost name resolution is handled within DNS itself.
#   127.0.0.1       localhost       L0174INA2.apac.sas.com  L0174INA2.in.sas.com  L0174INA2
#   ::1            localhost
127.0.0.1 www.presentation-3d.com
#to access Eminer 6.1 internally
10.10.10.35   sas.iasri.res.in      SAS
#to access Eminer 6.1 Externally
203.197.217.209 sas.iasri.res.in      sas

```

- Now Go to URL: <http://sas.iasri.res.in:6401/AnalyticsPlatform>
- Click on Launch and then Run

SAS Enterprise Miner 7.1

Pre-requisite:

- JRE 1.6 Update 16 or higher
- If Firewall and/or proxy has been implemented then kindly open the following ports:

Server	Ports
1) Metadata server	8561
2) Object spawner	8581
3) Framework Server	22031
4) Remote Server	5091
5) SAS App. Olap Server	5451
6) SAS Deployment Tester Server	10021

Steps to be followed:

- If you have installed multiple Java Runtime Environment then
Go to Control Panel → Java → Java tab → View → Keep check on JRE 1.6.0_16 or higher available version and Uncheck all other
- Check the entry of the **stat.iasri.res.in** in the host file, if not then open host file **C:\Windows\System32\drivers\etc** and edit the host file by entering the IP as shown below or specify the internal/external IP given by IASRI, New Delhi. Internal IP is to be specified only at IASRI, New Delhi. All other NARS organizations should specify external IP only which is: 14.139.56.156 stat.iasri.res.in stat (earlier 203.197.217.221 stat.iasri.res.in stat) as shown below stat.iasri.res.in stat as shown below

```

hosts - Notepad
File Edit Format View Help
# Copyright (c) 1993-2009 Microsoft Corp.
#
# This is a sample HOSTS file used by Microsoft TCP/IP for Windows.
#
# This file contains the mappings of IP addresses to host names. Each
# entry should be kept on an individual line. The IP address should
# be placed in the first column followed by the corresponding host name.
# The IP address and the host name should be separated by at least one
# space.
#
# Additionally, comments (such as these) may be inserted on individual
# lines or following the machine name denoted by a '#' symbol.
#
# For example:
#
#       102.54.94.97       rhino.acme.com       # source server
#       38.25.63.10      x.acme.com          # x client host
#
# localhost name resolution is handled within DNS itself.
#       127.0.0.1        localhost           L0174INA2.apac.sas.com  L0174INA2.in.sas.com  L0174INA2
#       ::1             localhost
127.0.0.1 www.presentation-3d.com
#to access Eminer 7.1 internally
10.10.10.21  stat.iasri.res.in  stat
#to access Eminer 7.1 Externally
203.197.217.221 stat.iasri.res.in  stat

```

- Now Go to URL: <http://stat.iasri.res.in/SASEnterpriseMinerJWS/Status>
- Click on Launch and then Run

Please note: You cannot run both E-Miner 6.1 and E-Miner 7.1 together. If you want to run JMP 6.1 then JAVA 1.5.0_15 should be available and for running JMP 7.1, JAVA version 1.6 onwards should be available on your system.

Indian NARS Statistical Computing Portal and other IP authenticated services are best viewed in **Internet Explorer 6 to 8 and Firefox 2.0.0.11 and 3.0.6**

Macros Developed

Macros have been developed for some commonly used statistical analysis and made available at Project Website www.iasri.res.in/sscnars. Following macros have been developed:

1. Analysis of data from Augmented Block designs
<http://www.iasri.res.in/sscnars/augblkdsgn.aspx>
2. Analysis of data from Split Factorial (main A, Sub B × C) designs
<http://www.iasri.res.in/sscnars/spltfctdsgn.aspx>
3. Analysis of data from Split Factorial (Main A×B, Sub C) designs
<http://www.iasri.res.in/sscnars/spltfctdsngm2s1.aspx>
4. Analysis of data from Split Factorial (main A×B, Sub C × D) designs
<http://www.iasri.res.in/sscnars/spltfactm2s2.aspx>
5. Analysis of data from Split Split Plot designs
<http://www.iasri.res.in/sscnars/spltpltdsgn.aspx>
6. Analysis of data from Strip Plot designs
<http://www.iasri.res.in/sscnars/StripPlot.aspx>
7. Analysis of data from Strip-Split Plot designs
<http://www.iasri.res.in/sscnars/stripssplit.aspx>

8. Econometric Analysis ((diversity indices, instability index, compound growth rate, Garret scoring technique and Demand analysis using LA-AIDS model) and available at <http://www.iasri.res.in/sscnars/ecoanalysis.aspx>
9. Estimation of heritability along with its standard error from half sib data
<http://www.iasri.res.in/sscnars/heritability.aspx>
10. Generation of Polycross designs
<http://www.iasri.res.in/sscnars/polycrossdesign.aspx>
11. Generation of TFNBCB designs
<http://www.iasri.res.in/sscnars/TFNBCBdesigns.aspx>

How to see updated version of reference manual?

Reference manual is updated regularly and updated version may be downloaded from <http://www.iasri.res.in/sscnars/contentmain.htm>

How to Renew License Files for SAS 9.2M2?

1. Go to <http://stat.iasri.res.in/sscnarsportal/public>
2. Click on SAS License Downloads 2011-12. It will redirect to New Page. It will start the Download of the SAS_Licenses11-12.zip. If it does not start automatically, then it would show Yellow Bar below the URL bar. Click on the Yellow Bar and Select Download File. Dialog box showing Open/Save/Cancel would appear. Click on Save and Browse the desired Location for saving the file.
3. Click on Portal Page link which is on top of the Page to go back to the main page.
4. Click on How to apply License Files?. Again it will redirect to the New Page and will start the Download Renew_the_licenses_for_SAS92_JMP8_JMPGenomics4.doc If it does not start automatically, then it would show Yellow Bar below the URL bar. Click on the Yellow Bar and Select Download File. Dialog box showing Open/Save/Cancel would appear. Click on Save and Browse the desired Location for saving the file.

You can also follow the following links for renewal of SAS Licenses:

<http://support.sas.com/kb/31/187.html>

Following link is only for Windows 7 and Windows Vista:

<http://support.sas.com/kb/31/290.html>

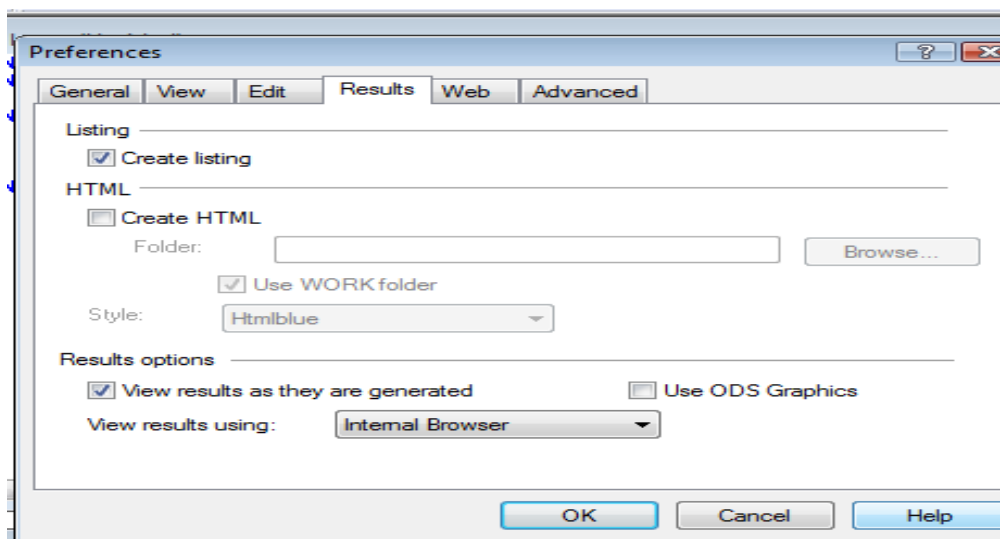
SAS 9.3

In SAS 9.3, the default destination in the SAS windowing environment is HTML, and ODS Graphics is enabled by default. These new defaults have several advantages. Graphs are integrated with tables, and all output is displayed in the same HTML file using a new style. This new style, HTML Blue, is an all-color style that is designed to integrate tables and modern statistical graphics. The default settings in the Results tab are as follows:

- The Create listing check box is not selected, so LISTING output is not created.
- The Create HTML check box is selected, so HTML output is created.
- The Use WORK folder check box is selected, so both HTML and graph image files are saved in the WORK folder (and not your current directory).
- The default style, HTMLBlue, is selected from the Style drop-down list.

- The Use ODS Graphics check box is selected, so ODS Graphics is enabled.
- Internal browser is selected so results are viewed in an internal SAS browser

We can view and modify the default settings by selecting **Tools→Options→Preferences→Result Tab** from the menu at the top of the SAS window usually known as TOPR pronounced "topper". Snap shot is as under.



- To get SAS listing instead of HTML, Select check box Create listing option and deselect Create HTML check box.
- Once HTML checkbox is deselected "Use work folder " get deselected automatically.
- Select View results as they are generated , if ODS Graphics is not required as default output. In many cases, graphs are an integral part of a data analysis. If we do not need graphics, ODS Graphics should be disabled, which will improve the performance of our program in terms of time and memory. One can disable and re-enable ODS Graphics in our SAS programs with the ODS GRAPHICS OFF and ODS GRAPHICS ON statements.

References

Littel, R.C., Freund, R.J. and Spector, P.C. (1991). *SAS System for Linear Models, Third Edition*. SAS Institute Inc.

Searle, S.R. (1971). *Linear Models*. John Wiley & Sons, New York.

Searle, S.R., Casella, G and McCulloch, C.E. (1992). *Analysis of Variance Components*. John Wiley & Sons, New York.

www.sas.com

www.support.sas.com

www.iasri.res.in/design

www.iasri.res.in/sscnars

<http://stat.iasri.res.in/sscnarsportal>

LARGE SCALE SURVEY DATA ANALYSIS USING SAS

Anil Rai

Indian Agricultural Statistics Research Institute, New Delhi 110012

34.1 Introduction

Multistage sampling has been found to be very useful in practice and this procedure is being commonly used in large-scale surveys. This sampling procedure is a compromise between cluster sampling and direct sampling of units. Further, this design is more flexible as it permits the use of different selection procedures at different stages. It may also be mentioned that multi-stage sampling may be the only choice in a number of practical situations where a satisfactory sampling frame of ultimate-stage units is not readily available and the cost of obtaining such a frame is large and time consuming. SAS has procedure for selection of samples using various designs i.e. SURVEYSELECT. Users can use SURVEYMEANS procedure for estimation of various important statistics. Further, in case of complex survey data analysis users can perform regression analysis (SURVEYREG) and categorical data analysis (SURVEYFREQ) when data has been selected through a complex sampling design. Details of SURVEYMEANS and brief description about rest of the procedure are given below from SAS User's Guide.

34.2 Sample Selection Procedures:

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or can sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of selection. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population.

To select a sample with PROC SURVEYSELECT, you input a SAS data set that contains the sampling frame or list of units from which the sample is to be selected. You also specify the selection method, the desired sample size or sampling rate, and other selection parameters. The SURVEYSELECT procedure selects the sample, producing an output data set that contains the selected units, their selection probabilities, and sampling weights. When you are selecting a sample in multiple stages, you invoke the procedure separately for each stage of selection, inputting the frame and selection parameters for each current stage.

The SURVEYSELECT procedure provides methods for both equal probability sampling and probability proportional to size (PPS) sampling. In equal probability sampling, each unit in the sampling frame, or in a stratum, has the same probability of being selected for the sample. In PPS sampling, a unit's selection probability is proportional to its size measure. For details on probability sampling methods, refer to Lohr (1999), Kish (1965, 1987), Kalton (1983), and Cochran (1977).

The SURVEYSELECT procedure provides the following equal probability sampling methods:

- simple random sampling
- unrestricted random sampling (with replacement)
- systematic random sampling
- sequential random sampling

This procedure also provides the following probability proportional to size (PPS) methods:

- PPS sampling without replacement
- PPS sampling with replacement
- PPS systematic sampling
- PPS algorithms for selecting two units per stratum
- sequential PPS sampling with minimum replacement

The procedure uses fast, efficient algorithms for these sample selection methods. Thus, it performs well even for large input data sets or sampling frames, which may occur in practice for large-scale sample surveys.

The SURVEYSELECT procedure can perform stratified sampling, selecting samples independently within the specified strata, or non-overlapping subgroups of the survey population. Stratification controls the distribution of the sample size in the strata. It is widely used in practice toward meeting a variety of survey objectives. For example, with stratification you can ensure adequate sample sizes for subgroups of interest, including small subgroups, or you can use stratification toward improving the precision of the overall estimates. When you are using a systematic or sequential selection method, the SURVEYSELECT procedure also can sort by control variables within strata for the additional control of implicit stratification.

The SURVEYSELECT procedure provides replicated sampling, where the total sample is composed of a set of replicates, each selected in the same way. You can use replicated sampling to study variable non-sampling errors, such as variability in the results obtained by different interviewers. You can also use replication to compute standard errors for the combined sample estimates.

34.3 Estimation Procedures:

PROC SURVEYMEANS uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or primary sampling units (PSUs), in the sample design, the procedure estimates variance from the variation among PSUs. When the design is

stratified, the procedure pools stratum variance estimates to compute the overall variance estimate.

PROC SURVEYMEANS uses the Output Delivery System (ODS) to place results in output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality.

Statistical Computations

The SURVEYMEANS procedure uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or PSUs, in the sample design, the procedure estimates variance from the variation among PSUs. When the design is stratified, the procedure pools stratum variance estimates to compute the overall variance estimate. For t tests of the estimates, the degrees of freedom equal the number of clusters minus the number of strata in the sample design.

For a multistage sample design, the variance estimation method depends only on the first stage of the sample design. So, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling. This variance estimation method assumes that the first-stage sampling fraction is small, or the first-stage sample is drawn with replacement, as it often is in practice.

Quite often in complex surveys, respondents have unequal weights, which reflect unequal selection probabilities and adjustments for non-response. In such surveys, the appropriate sampling weights must be used to obtain valid estimates for the study population.

Definition and Notation

For a stratified clustered sample design, together with the sampling weights, the sample can be represented by an $n \times (P+1)$ matrix

$$\begin{aligned} (\mathbf{w}, \mathbf{Y}) &= (w_{hij}, \mathbf{y}_{hij}) \\ &= \left(w_{hij}, y_{hij}^{(1)}, y_{hij}^{(2)}, \dots, y_{hij}^{(P)} \right) \end{aligned}$$

where

- $h = 1, 2, \dots, H$ is the stratum number, with a total of H strata
- $i = 1, 2, \dots, n_h$ is the cluster number within stratum h , with a total of n_h clusters
- $j = 1, 2, \dots, m_{hi}$ is the unit number within cluster i of stratum h , with a total of m_{hi} units
- $p = 1, 2, \dots, P$ is the analysis variable number, with a total of P variables
- $n = \sum_{h=1}^H \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample

- w_{hij} denotes the sampling weight for observation j in cluster i of stratum h
- $\mathbf{y}_{hij} = (y_{hij}^{(1)}, y_{hij}^{(2)}, \dots, y_{hij}^{(P)})$ are the observed values of the analysis variables for observation j in cluster i of stratum h , including both the values of numerical variables and the values of indicator variables for levels of categorical variables.

For a categorical variable C , let l denote the number of levels of C , and denote the level values as c_1, c_2, \dots, c_l . Then there are l indicator variables associated with these levels. That is, for level $C=c_k$ ($k = 1, 2, \dots, l$), a $y^{(q)}(q \in \{1, 2, \dots, P\})$ contains the values of the indicator variable for the category $C=c_k$, with the value of observation j in cluster i of stratum h :

$$y_{hij}^{(q)} = I_{\{C=c_k\}}(h, i, j) = \begin{cases} 1 & \text{if } C_{hij} = c_k \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the total number of analysis variables, P , is the total number of numerical variables plus the total number of levels of all categorical variables.

Also, f_h denotes the sampling rate for stratum h . You can use the TOTAL= option or the RATE= option to input population totals or sampling rates. If you input stratum totals, PROC SURVEYMEANS computes f_h as the ratio of the stratum sample size to the stratum total. If you input stratum sampling rates, PROC SURVEYMEANS uses these values directly for f_h . If you do not specify the TOTAL= option or the RATE= option, then the procedure assumes that the stratum sampling rates f_h are negligible, and a finite population correction is not used when computing variances.

This notation is also applicable to other sample designs. For example, for a sample design without stratification, you can let $H=1$; for a sample design without clusters, you can let $m_{hi}=1$ for every h and i .

Mean

When you specify the keyword MEAN, the procedure computes the estimate of the mean (mean per element) from the survey data. Also, the procedure computes the mean by default if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement.

PROC SURVEYMEANS computes the estimate of the mean as

$$\hat{Y} = \left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \right) / w_{\dots}$$

where

$$w_{...} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

is the sum of the weights over all observations in the sample.

Variance and Standard Error of the Mean

When you specify the keyword STDERR, the procedure computes the standard error of the mean. Also, the procedure computes the standard error by default if you specify the keyword MEAN, or if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement. The keyword VAR requests the variance of the mean.

PROC SURVEYMEANS uses the Taylor series expansion theory to estimate the variance of the mean \hat{Y} . The procedure computes the estimated variance as

$$\widehat{V}(\hat{Y}) = \sum_{h=1}^H \widehat{V}_h(\hat{Y})$$

where if $n_h > 1$,

$$\widehat{V}_h(\hat{Y}) = \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2$$

$$e_{hi.} = \left(\sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - \hat{Y}) \right) / w_{...}$$

$$\bar{e}_{h..} = \left(\sum_{i=1}^{n_h} e_{hi.} \right) / n_h$$

and if $n_h = 1$,

$$\widehat{V}_h(\hat{Y}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \dots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

The standard error of the mean is the square root of the estimated variance.

$$\text{StdErr}(\hat{Y}) = \sqrt{\widehat{V}(\hat{Y})}$$

Ratio

When you use a RATIO statement, the procedure produces statistics requested by the statistics-keywords in the PROC SURVEYMEANS statement.

Suppose that you want to calculate the ratio of variable Y over variable X . Let x_{hij} be the value of variable X for the j th member in cluster i in the h th stratum.

The ratio of Y over X is

$$\hat{R} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} x_{hij}}$$

PROC SURVEYMEANS uses the Taylor series expansion method to estimate the variance of the ratio \hat{R} as

$$\widehat{V}(\hat{R}) = \sum_{h=1}^H \widehat{V}_h(\hat{R})$$

where if $n_h > 1$,

$$\widehat{V}_h(\hat{R}) = \frac{n_h(1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (g_{hi\cdot} - \bar{g}_{h\cdot\cdot})^2$$

$$g_{hi\cdot} = \frac{\sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - x_{hij} \hat{R})}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} x_{hij}}$$

$$\bar{g}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} g_{hi\cdot} \right) / n_h$$

and if $n_h = 1$,

$$\widehat{V}_h(\hat{R}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \dots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

The standard error of the ratio is the square root of the estimated variance.

$$\text{StdErr}(\hat{R}) = \sqrt{\widehat{V}(\hat{R})}$$

t Test for the Mean

If you specify the keyword T, PROC SURVEYMEANS computes the t -value for testing that the population mean equals zero, $H_0 : \bar{Y} = 0$. The test statistic equals

$$t(\hat{Y}) = \hat{Y} / \text{StdErr}(\hat{Y})$$

The two-sided p -value for this test is

$$\text{Prob}(|T| > |t(\hat{Y})|)$$

where T is a random variable with the t distribution with df degrees of freedom.

PROC SURVEYMEANS calculates the degrees of freedom for the t test as the number of clusters minus the number of strata. If there are no clusters, then df equals the number of observations minus the number of strata. If the design is not stratified, then df equals the number of clusters minus one. The procedure displays df for the t test if you specify the keyword DF in the PROC SURVEYMEANS statement.

If missing values or missing weights are present in your data, the number of strata, the number of observations, and the number of clusters are counted based on the observations in non-empty strata.

Confidence Limits for the Mean

If you specify the keyword CLM, the procedure computes two-sided confidence limits for the mean. Also, the procedure includes the confidence limits by default if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement.

The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\hat{Y} \pm \text{StdErr}(\hat{Y}) \cdot t_{df, \alpha/2}$$

where \hat{Y} is the estimate of the mean, $\text{StdErr}(\hat{Y})$ is the standard error of the mean, and $t_{df, \alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the t distribution

If you specify the keyword UCLM, the procedure computes the one-sided upper $100(1 - \alpha)$ confidence limit for the mean:

$$\hat{Y} + \text{StdErr}(\hat{Y}) \cdot t_{df, \alpha}$$

If you specify the keyword LCLM, the procedure computes the one-sided lower $100(1 - \alpha)$ confidence limit for the mean:

$$\hat{Y} - \text{StdErr}(\hat{Y}) \cdot t_{df, \alpha}$$

Coefficient of Variation

If you specify the keyword CV, PROC SURVEYMEANS computes the coefficient of variation, which is the ratio of the standard error of the mean to the estimated mean.

$$cv(\bar{Y}) = \text{StdErr}(\hat{Y}) / \hat{Y}$$

If you specify the keyword CVSUM, PROC SURVEYMEANS computes the coefficient of variation for the estimated total, which is the ratio of the standard deviation of the sum to the estimated total.

$$cv(Y) = \text{Std}(\hat{Y}) / \hat{Y}$$

Proportions

If you specify the keyword MEAN for a categorical variable, PROC SURVEYMEANS estimates the proportion, or relative frequency, for each level of the categorical variable. If you do not specify any statistic-keywords in the PROC SURVEYMEANS statement, the procedure estimates the proportions for levels of the categorical variables, together with their standard errors and confidence limits.

The procedure estimates the proportion in level c_k for variable C as

$$\hat{p} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}^{(q)}}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}}$$

where $y_{hij}^{(q)}$ is the value of the indicator function for level $C=c_k$, and $y_{hij}^{(q)}$ equals 1 if the observed value of variable C equals c_k , and $y_{hij}^{(q)}$ equals 0 otherwise. Since the proportion estimator is actually an estimator of the mean for an indicator variable, the procedure computes its variance and standard error according to the method of survey mean

Total

If you specify the keyword SUM, the procedure computes the estimate of the population total from the survey data. The estimate of the total is the weighted sum over the sample.

$$\hat{Y} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

For a categorical variable level, \hat{Y} estimates its total frequency in the population.

Variance and Standard Deviation of the Total

When you specify the keyword STD or the keyword SUM, the procedure estimates the standard deviation of the total. The keyword VARSUM requests the variance of the total.

PROC SURVEYMEANS estimates the variance of the total as

$$\hat{V}(\hat{Y}) = \sum_{h=1}^H \hat{V}_h(\hat{Y})$$

where if $n_h > 1$,

$$\widehat{V}_h(\widehat{Y}) = \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (y_{hi\cdot} - \bar{y}_{h\cdot\cdot})^2$$

$$y_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

$$\bar{y}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} y_{hi\cdot} \right) / n_h$$

and if $n_h=1$,

$$\widehat{V}_h(\widehat{Y}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \dots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

The standard deviation of the total equals

$$\text{Std}(\widehat{Y}) = \sqrt{\widehat{V}(\widehat{Y})}$$

Confidence Limits of a Total

If you specify the keyword CLSUM, the procedure computes confidence limits for the total. The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\widehat{Y} \pm \text{Std}(\widehat{Y}) \cdot t_{df, \alpha/2}$$

where \widehat{Y} is the estimate of the total, $\text{Std}(\widehat{Y})$ is the estimated standard deviation, and $t_{df, \alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the t distribution with df

If you specify the keyword UCLSUM, the procedure computes the one-sided upper $100(1 - \alpha)$ confidence limit for the sum:

$$\widehat{Y} + \text{Std}(\widehat{Y}) \cdot t_{df, \alpha}$$

If you specify the keyword LCLSUM, the procedure computes the one-sided lower $100(1 - \alpha)$ confidence limit for the sum:

$$\widehat{Y} - \text{Std}(\widehat{Y}) \cdot t_{df, \alpha}$$

Domain Statistics

When you use a DOMAIN statement to request a domain analysis, the procedure computes the requested statistics for each domain.

For a domain D , let I_D be the corresponding indicator variable:

$$I_D(h, i, j) = \begin{cases} 1 & \text{if observation } (h, i, j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

Let

$$z_{hij} = y_{hij} I_D(h, i, j) = \begin{cases} y_{hij} & \text{if observation } (h, i, j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

The requested statistics for variable y in domain D are computed based on the values of z .

Domain Mean The estimated mean of y in the domain D is

$$\widehat{Y}_D = \left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij} z_{hij} \right) / v_{\dots}$$

where

$$v_{hij} = w_{hij} I_D(h, i, j) = \begin{cases} w_{hij} & \text{if observation } (h, i, j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

$$v_{\dots} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}$$

The variance of \widehat{Y}_D is estimated by

$$\widehat{V}(\widehat{Y}_D) = \sum_{h=1}^H \widehat{V}_h(\widehat{Y}_D)$$

where if $n_h > 1$,

$$\begin{aligned}\widehat{V}_h(\widehat{Y}_D) &= \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (r_{hi\cdot} - \bar{r}_{h\cdot\cdot})^2 \\ r_{hi\cdot} &= \left(\sum_{j=1}^{m_{hi}} v_{hij} (z_{hij} - \widehat{Y}_D) \right) / v_{\dots} \\ \bar{r}_{h\cdot\cdot} &= \left(\sum_{i=1}^{n_h} r_{hi\cdot} \right) / n_h\end{aligned}$$

and if $n_h=1$,

$$\widehat{V}_h(\widehat{Y}_D) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \dots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

Domain Total The estimated total in domain D is

$$\widehat{Y}_D = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij} z_{hij}$$

and its estimated variance is

$$\widehat{V}(\widehat{Y}_D) = \sum_{h=1}^H \widehat{V}_h(\widehat{Y}_D)$$

where if $n_h > 1$,

$$\begin{aligned}\widehat{V}_h(\widehat{Y}_D) &= \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (z_{hi\cdot} - \bar{z}_{h\cdot\cdot})^2 \\ z_{hi\cdot} &= \sum_{j=1}^{m_{hi}} v_{hij} z_{hij} \\ \bar{z}_{h\cdot\cdot} &= \left(\sum_{i=1}^{n_h} z_{hi\cdot} \right) / n_h\end{aligned}$$

and if $n_h=1$,

$$\widehat{V}_h(\widehat{Y}_D) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \dots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

Degrees of Freedom For domain analysis, PROC SURVEYMEANS computes the degrees of freedom for t tests as the number of clusters in the non-empty strata minus the number of non-empty strata. When the sample design has no clusters, the degrees of freedom equal the number of observations in non-empty strata minus the number of non-empty strata. Missing values and missing weights can result in empty strata. In domain

analysis, an empty stratum can also occur when the stratum contains no observations in the specified domain. If no observations in a whole stratum belong to a domain, then this stratum is called an empty stratum for that domain.

Survey Data Analysis

Specification of Population Totals and Sampling Rates

If your analysis includes a finite population correction (*fpc*), you can input either the sampling rate or the population total using the RATE= option or the TOTAL= option. (You cannot specify both of these options in the same PROC SURVEYMEANS statement.) If you do not specify one of these options, the procedure does not use the *fpc* when computing variance estimates. For fairly small sampling fractions, it is appropriate to ignore this correction. If your design has multiple stages of selection and you are specifying the RATE= option, you should input the first-stage sampling rate, which is the ratio of the number of PSUs in the sample to the total number of PSUs in the study population. If you are specifying the TOTAL= option for a multistage design, you should input the total number of PSUs in the study population.

For a non-stratified sample design, or for a stratified sample design with the same sampling rate or the same population total in all strata, you should use the RATE=*value* option or the TOTAL=*value* option. If your sample design is stratified with different sampling rates or population totals in the strata, then you can use the RATE= *SAS-data-set* option or the TOTAL= *SAS-data-set* option to name a SAS data set that contains the stratum sampling rates or totals. This data set is called a *secondary data set*, as opposed to the *primary data set* that you specify with the DATA= option.

The secondary data set must contain all the stratification variables listed in the STRATA statement and all the variables in the BY statement. If there are formats associated with the STRATA variables and the BY variables, then the formats must be consistent in the primary and the secondary data sets. If you specify the TOTAL=*SAS-data-set* option, the secondary data set must have a variable named `_TOTAL_` that contains the stratum population totals. Or if you specify the RATE=*SAS-data-set* option, the secondary data set must have a variable named `_RATE_` that contains the stratum sampling rates. If the secondary data set contains more than one observation for any one stratum, then the procedure uses the first value of `_TOTAL_` or `_RATE_` for that stratum and ignores the rest.

The *value* in the RATE= option or the values of `_RATE_` in the secondary data set must be nonnegative numbers. You can specify *value* as a number between 0 and 1. Or you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

If you specify the `TOTAL=value` option, *value* must not be less than the sample size. If you provide stratum population totals in a secondary data set, these values must not be less than the corresponding stratum sample sizes.

Primary Sampling Units (PSUs)

When you have clusters, or primary sampling units (PSUs), in your sample design, the procedure estimates variance from the variation among PSUs. You can use the `CLUSTER` statement to identify the first stage clusters in your design. `PROC SURVEYMEANS` assumes that each cluster represents a PSU in the sample and that each observation is an element of a PSU. If you do not specify a `CLUSTER` statement, the procedure treats each observation as a PSU.

Domain Analysis

It is common practice to compute statistics for subpopulations, or domains, in addition to computing statistics for the entire study population. Analysis for domains using the entire sample is called *domain analysis* (subgroup analysis, subpopulation analysis, sub-domain analysis). The formation of these subpopulations of interest may be unrelated to the sample design. Therefore, the sample sizes for the subpopulations may actually be random variables.

In order to incorporate this variability into the variance estimation, you should use a `DOMAIN` statement. Note that using a `BY` statement provides completely separate analyses of the `BY` groups. It does not provide a statistically valid subpopulation or domain analysis, where the total number of units in the subpopulation is not known with certainty. For more detailed information about domain analysis, refer to Kish (1965).

PROC SURVEYMEANS Statement

PROC SURVEYMEANS < options > < statistic-keywords > ;

The `PROC SURVEYMEANS` statement invokes the procedure. In this statement, you identify the data set to be analyzed and specify sample design information. The `DATA=` option names the input data set to be analyzed. If your analysis includes a finite population correction factor, you can input either the sampling rate or the population total using the `RATE=` or `TOTAL=` option. If your design is stratified, with different sampling rates or totals for different strata, then you can input these stratum rates or totals in a SAS data set containing the stratification variables.

In the `PROC SURVEYMEANS` statement, you also can use statistic-keywords to specify statistics for the procedure to compute. Available statistics include the population mean and population total, together with their variance estimates and confidence limits. You can also request data set summary information and sample design information.

You can specify the following options in the `PROC SURVEYMEANS` statement:

ALPHA= α : sets the confidence level for confidence limits. The value of the ALPHA= option must be between 0 and 1, and the default value is 0.05. A confidence level of α produces $100(1 - \alpha)$ % confidence limits. The default of ALPHA=0.05 produces 95% confidence limits.

DATA=*SAS-data-set*: specifies the SAS data set to be analyzed by PROC SURVEYMEANS. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

MISSING: requests that the procedure treat missing values as a valid category for all categorical variables, which include categorical analysis variables, strata variables, cluster variables, and domain variables.

ORDER=**DATA** | **FORMATTED** | **INTERNAL**: specifies the order in which the values of the categorical variables are to be reported. The following shows how PROC SURVEYMEANS interprets values of the ORDER= option:

DATA

orders values according to their order in the input data set.

FORMATTED

orders values by their formatted values. This order is operating environment dependent. By default, the order is ascending.

INTERNAL

orders values by their unformatted values, which yields the same order that the SORT procedure does. This order is operating environment dependent. By default, ORDER=FORMATTED.

The ORDER= option applies to all the categorical variables. When the default ORDER=FORMATTED is in effect for numeric variables for which you have supplied no explicit format, the levels are ordered by their internal values.

RATE=*value* | *SAS-data-set*

R=*value* | *SAS-data-set*

specifies the sampling rate as a nonnegative *value*, or names an input data set that contains the stratum sampling rates. The procedure uses this information to compute a finite population correction for variance estimation. If your sample design has multiple stages, you should specify the *first-stage sampling rate*, which is the ratio of the number of PSUs selected to the total number of PSUs in the population.

For a nonstratified sample design, or for a stratified sample design with the same sampling rate in all strata, you should specify a nonnegative *value* for the RATE= option. If your design is stratified with different sampling rates in the strata, then you should name a SAS data set that contains the stratification variables and the sampling rates. The sampling rate *value* must be a nonnegative number. You can specify *value*

as a number between 0 and 1. Or you can specify *value* in percentage form as a number between 1 and 100, and PROC SURVEYMEANS will convert that number to a proportion. The procedure treats the value 1 as 100%, and not the percentage form 1%.

If you do not specify the **TOTAL=** option or the **RATE=** option, then the variance estimation does not include a finite population correction. You cannot specify both the **TOTAL=** option and the **RATE=** option.

TOTAL=*value* *SAS-data-set*

N=*value* *SAS-data-set*

specifies the total number of primary sampling units (PSUs) in the study population as a positive *value*, or names an input data set that contains the stratum population totals. The procedure uses this information to compute a finite population correction for variance estimation.

For a nonstratified sample design, or for a stratified sample design with the same population total in all strata, you should specify a positive *value* for the **TOTAL=** option. If your sample design is stratified with different population totals in the strata, then you should name a SAS data set that contains the stratification variables and the population totals. If you do not specify the **TOTAL=** option or the **RATE=** option, then the variance estimation does not include a finite population correction. You cannot specify both the **TOTAL=** option and the **RATE=** option.

statistic-keywords : specifies the statistics for the procedure to compute. If you do not specify any statistic-keywords, PROC SURVEYMEANS computes the NOBS, MEAN, STDERR, and CLM statistics by default.

The statistics produced depend on the type of the analysis variable. If you name a numeric variable in the CLASS statement, then the procedure analyzes that variable as a categorical variable. The procedure always analyzes character variables as categorical. PROC SURVEYMEANS computes MIN, MAX, and RANGE for numeric variables but not for categorical variables. For numeric variables, the keyword MEAN produces the mean, but for categorical variables it produces the proportion in each category or level. Also for categorical variables, the keyword NOBS produces the number of observations for each variable level, and the keyword NMISS produces the number of missing observations for each level. If you request the keyword NCLUSTER for a categorical variable, PROC SURVEYMEANS displays for each level the number of clusters with observations in that level. PROC SURVEYMEANS computes SUMWGT in the same way for both categorical and numeric variables, as the sum of the weights over all non-missing observations.

PROC SURVEYMEANS performs uni-variate analysis, analyzing each variable separately. Thus the number of non-missing and missing observations may not be the same for all analysis variables. If you use the keyword **RATIO** without the keyword

MEAN, the keyword MEAN is implied.

Other available statistics computed for a ratio are N, NCLU, SUMWGT, RATIO, STDERR, DF, T, PROBT, and CLM, as listed below. If no statistics are requested, the procedure will compute the ratio and its standard error by default for a RATIO statement.

The valid statistic-keywords are as follows:

ALL : all statistics listed

CLM: $100(1 - \alpha)$ % two-sided confidence limits for the MEAN, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

CLSUM : $100(1 - \alpha)$ % two-sided confidence limits for the SUM, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

CV: coefficient of variation for MEAN

CVSUM :coefficient of variation for SUM

DF :degrees of freedom for the *t* test

LCLM : $100(1 - \alpha)$ % one-sided lower confidence limit of the MEAN, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

LCLMSUM : $100(1 - \alpha)$ % one-sided lower confidence limit of the SUM, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

MAX :maximum value

MEAN:mean for a numeric variable, or the proportion in each category for a categorical variable

MIN :minimum value

NCLUSTER :number of clusters

NMISS: number of missing observations

NOBS:number of nonmissing observations

RANGE :range, MAX-MIN

RATIO :ratio of means or proportions

STD :standard deviation of the SUM. When you request SUM, the procedure computes STD by default.

STDERR :standard error of the MEAN or RATIO. When you request MEAN or RATIO, the procedure computes STDERR by default.

SUM :weighted sum, $\sum w_i y_i$, or estimated population total when the appropriate sampling weights are used

SUMWGT :sum of the weights, $\sum w_i$

T :*t*-value and its corresponding *p*-value with DF degrees of freedom for

$H_0 : \theta = 0$: where θ is the population mean or the population ratio

UCLM : $100(1 - \alpha)$ % one-sided upper confidence limit of the MEAN, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

UCLMSUM : $100(1 - \alpha)$ % one-sided upper confidence limit of the SUM, where α is determined by the ALPHA= option, and the default is $\alpha = 0.05$

VAR :variance of the MEAN or RATIO

VARSUM :variance of the SUM

BY Statement

BY *variables* ;

You can specify a BY statement with PROC SURVEYMEANS to obtain separate analyses on observations in groups defined by the BY variables.

Note that using a BY statement provides completely separate analyses of the BY groups. It does not provide a statistically valid subpopulation or domain analysis, where the total number of units in the subpopulation is not known with certainty. You should use the DOMAIN statement to obtain domain analysis.

When a BY statement appears, the procedure expects the input data sets to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If you specify more than one BY statement, the procedure uses only the latest BY statement and ignores any previous ones.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Use the BY statement options NOTSORTED or DESCENDING in the BY statement. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure.

CLASS Statement

CLASS | **CLASSES** *variables* ;

The CLASS statement names variables to be analyzed as categorical variables. For categorical variables, PROC SURVEYMEANS estimates the proportion in each category or level, instead of the overall mean. PROC SURVEYMEANS always analyzes character variables as categorical. If you want categorical analysis for a numeric variable, you must include that variable in the CLASS statement.

The CLASS *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the CLASS variables determine the categorical variable levels. Thus, you can use formats to group values into levels. You can use multiple CLASS statements to specify categorical variables.

When you specify class variables, you may use the SAS system option SUMSIZE= to limit (or to specify) the amount of memory that is available for data analysis.

CLUSTER Statement

CLUSTER | **CLUSTERS** *variables* ;

The CLUSTER statement names variables that identify the clusters in a clustered sample design. The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata.

If your sample design has clustering at multiple stages, you should identify only the first-stage clusters, or primary sampling units (PSUs), in the CLUSTER statement. The CLUSTER *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the CLUSTER variables determine the CLUSTER variable levels. Thus, you can use formats to group values into levels. You can use multiple CLUSTER statements to specify cluster variables. The procedure uses variables from all CLUSTER statements to create clusters.

DOMAIN Statement

DOMAIN | **SUBGROUP** *variables* < *variable*variable*
*variable*variable*variable ...* > ;

The DOMAIN statement requests analysis for subpopulations, or domains, in addition to analysis for the entire study population. The DOMAIN statement names the variables that identify domains, which are called domain variables.

It is common practice to compute statistics for domains. The formation of these domains may be unrelated to the sample design. Therefore, the sample sizes for the domains are random variables. In order to incorporate this variability into the variance estimation, you should use a DOMAIN statement.

Note that a DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently. A domain variable can be either character or numeric. However, the procedure treats domain variables as categorical variables. If a variable appears by itself in a DOMAIN statement, each level of this variable determines a domain in the study population. If two or more variables are joined by asterisks (*), then every possible combination of levels of the variables determines a domain. The procedure performs a descriptive analysis within each domain defined by the domain variables.

The formatted values of the domain variables determine the categorical variable levels. Thus, you can use formats to group values into levels.

RATIO Statement

RATIO < *'label'* > *variables / variables* ;

The RATIO statement requests ratio analysis for means or proportions of analysis variables. A ratio statement names the variables whose means will be used as numerators or denominators in a ratio. Variables appearing before the slash (/), called *numerator variables*, are used for numerators. Variables appearing after the slash (/), called *denominator variables*, are used for denominators. These *variables* can be any number of analysis variables, either continuous or categorical, in the input data set.

You can optionally specify a label for each RATIO statement to identify the ratios in the output. Labels must be enclosed in single quotes.

If a RATIO statement does not have any numerator variable or denominator variable specified, the RATIO statement will be ignored.

A numerator or denominator variable must be an analysis variable. That is, if there is a VAR statement, then a numerator or denominator variable must appear in the VAR statement. If there is no VAR statement, a numerator or denominator variable must be on the default analysis variable list. If a numerator or denominator variable is not an analysis variable, it is ignored.

The computation of ratios depends on whether the numerator and denominator variables are continuous or categorical. For continuous variables, ratios are calculated with the mean of the variables. If a continuous variable appears as both a numerator and a denominator variable, the ratio of this variable itself is ignored. For categorical variables, ratios are calculated with the proportions for the categories of a categorical variable. If a categorical variable appears as both a numerator and a denominator variable, then the ratios of the proportions for all categories are computed, except the ratio of each category with itself. You may have more than one RATIO statement. Each RATIO statement produces ratios independently using its own numerator and denominator variables. Each RATIO statement also produces its own ratio analysis table.

Available statistics for a ratio are

- N, number of observations used to compute the ratio
- NCLU, number of clusters
- SUMWGT, sum of weights
- RATIO, ratio
- STDERR, standard error of ratio
- VAR, variance of ratio
- T, *t*-value of ratio
- PROBT, *p*-value of *t*
- DF, degrees of freedom of *t*
- CLM, two-sided confidence limits of ratio
- UCLM, one-sided upper confidence limit of ratio
- LCLM, one-sided lower confidence limit of ratio

The procedure will calculate these statistics based on the statistic-keywords which you specified in the PROC statement. If a statistic-keyword is not appropriate for RATIO statement, that statistic-keyword is ignored. If no valid statistics are requested for a RATIO statement, the procedure will compute the ratio and its standard error by default.

Note that ratios within a domain are currently not available.

When calculating the means or proportions for the numerator and denominator variables in a ratio, an observation is excluded if it has a missing value in either the continuous numerator variable or the denominator variable. An observation with missing values is also excluded for the categorical numerator or denominator variables, unless the MISSING option is used.

STRATA Statement

```
STRATA | STRATUM variables < / option > ;
```

The STRATA statement names variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample. If your sample design has stratification at multiple stages, you should identify only the first-stage strata in the STRATA statement. The STRATA *variables* are one or more variables in the DATA= input data set. These variables can be either character or numeric. The formatted values of the STRATA variables determine the levels. Thus, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *SAS Procedures Guide*.

You can specify the following option in the STRATA statement after a slash (/):

LIST: displays a "Stratum Information" table, which includes values of the STRATA variables and sampling rates for each stratum. This table also provides the number of observations and number of clusters for each stratum and analysis variable.

VAR Statement

```
VAR variables ;
```

The VAR statement names the variables to be analyzed. If you want a categorical analysis for a numeric variable, you must also name that variable in the CLASS statement. For categorical variables, PROC SURVEYMEANS estimates the proportion in each category or level, instead of the overall mean. Character variables are always analyzed as categorical variables. If you do not specify a VAR statement, then PROC SURVEYMEANS analyzes all variables in the DATA= input data set, except those named in the BY, CLUSTER, STRATA, and WEIGHT statements.

WEIGHT Statement

WEIGHT | **WGT** *variable* ;

The WEIGHT statement names the variable that contains the sampling weights. This variable must be numeric. If you do not specify a WEIGHT statement, PROC SURVEYMEANS assigns all observations a weight of 1. Sampling weights must be positive numbers. If an observation has a weight that is non-positive or missing, then the procedure omits that observation from the analysis. If you specify more than one WEIGHT statement, the procedure uses only the first WEIGHT statement and ignores the rest.

34.4 Regression Analysis of Survey Data:

The SURVEYREG procedure performs regression analysis for sample survey data. This procedure can handle complex survey sample designs, including designs with stratification, clustering, and unequal weighting. The procedure fits linear models for survey data and computes regression coefficients and their variance-covariance matrix. The procedure also provides significance tests for the model effects and for any specified estimable linear functions of the model parameters. Using the regression model, the procedure can compute predicted values for the sample survey data.

PROC SURVEYREG computes the regression coefficient estimators by generalized least-squares estimation using element-wise regression. The procedure assumes that the regression coefficients are the same across strata and primary sampling units (PSUs). To estimate the variance-covariance matrix for the regression coefficients, PROC SURVEYREG uses the Taylor expansion theory for estimating sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimator for this approximation to estimate the variance of the estimator itself.

PROC SURVEYREG uses the ODS (Output Delivery System) to place results in output data sets. This is a departure from older SAS procedures that provide OUTPUT statements for similar functionality.

34.5 Categorical Data Analysis of Survey Sampling:

The SURVEYFREQ procedure produces one-way to n-way frequency and cross-tabulation tables from sample survey data. These tables include estimates of population totals and proportions, and the corresponding standard errors. PROC SURVEYFREQ computes the variance estimates based on the sample design used to obtain the survey data. The design can be a complex multistage survey design with stratification, clustering, and unequal weighting. PROC SURVEYFREQ also provides design-based tests of independence and association between variables.

PROC SURVEYFREQ uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method is appropriate for all designs where the first-stage sample is selected with replacement, or where the first-stage

sampling fraction is small, as it often is in practice. The Taylor expansion method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters or primary sampling units (PSUs) in the sample design, the procedure estimates variance from the variation among PSUs. When the design is stratified, the procedure combines stratum variance estimates to compute the overall variance estimate

Example: This survey aims to assess the harvest and post-harvest losses of major crops and commodities i.e. 46, produced in India. The survey was conducted in 14 agro-climatic zones of the country. Only quantitative losses will be assessed in the survey. The survey was conducted in 120 districts by all 33 AICRP on PHT Centers throughout the country. Food grains, oilseeds, pulses, plantation crops, vegetables, fruits, livestock and aquaculture produce were covered in this survey. The data was collected for estimating the losses at farm level (Harvest/picking, collection, threshing and other unit of operations), other channels at producer level and various channels at market level (Trader, processing unit, etc.).

1. Farm level: Two blocks was taken randomly from each selected district. Then five villages were taken randomly from each block. A random sample of ten farmers were selected from each village for data collection by enquiry at farm level. For data collection by actual observation, two farmers from the list of already 10 selected farmers of each village were taken randomly.

2. Storage at producer level: Same sample size of farmers (as taken for data collection at farm level) was taken for data collection by enquiry and observation at this level.

3. Storage at market level: Two units of each channel such as traders, processing unit, packaging units etc for each crop/commodity was taken randomly from the list of the units prepared after complete enumeration of units for each channel of each selected district. If a particular channel is not available in the selected district then nearby districts for data collection by enquiry/ actual observation was considered.

Day2	Month2	Year2	Period	Aczon	Centre	State	District	Block	Village	Wt_blo	Wt_vill	Wt_tot	Serial	H_ho
14	JUN	2006	3	East c	ANGR/ANDH	GUNTUR	TADEFCHIRR	28.5	1.6	45.6	2	DON		
14	JUN	2006	3	East c	ANGR/ANDH	GUNTUR	TADEFCHIRR	28.5	1.6	45.6	2	DON		
07	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	2	ARUI		
07	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	2	ARUI		
09	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	5	ARUI		
09	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	5	ARUI		
07	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	3	BOD		
07	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	3	BOD		
07	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	4	CHA		
07	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	4	CHA		
05	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	1	KON		
05	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	1	KON		
08	JUN	2006	6	East c	ANGR/ANDH	GUNTUR	TADEFKUNCH	28.5	1.6	45.6	7	TENI		
11	JUN	2006	5	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	5	ALLA		
10	JUN	2006	5	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	3	ALLA		
12	JUN	2006	5	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	8	ANN		
12	JUN	2006	5	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	8	ANN		
12	JUN	2006	5	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	7	ANN		
12	JUN	2006	5	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	10	AVU		
10	JUN	2006	3	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	1	BHEI		
10	JUN	2006	5	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	4	BOM		
10	JUN	2006	3	East c	ANGR/ANDH	GUNTUR	TADEFMELLE	28.5	1.6	45.6	2	DON		

Program Code:

```

ODS html;
Title 'ESTIMATION OF PH_LOSS (BY ENQUIRY) AT FARM LEVEL BY OPERATIONS
Using FPC (Guntur}';

proc surveymeans
DATA=WORK.Test ALL RATIO;
CLUSTERS BLOCK;
STRATA DISTRICT;
WEIGHT WT_Total;
RATIO Qty_lost/Qty_handled;
DOMAIN cropl*OPERATION;
VAR Qty_handled Qty_lost;
run;

```


RESULT:

Crop1	operation	Variable	Label	N	N Miss	Minimum	Maximum	Range	Clus
BANANA	HARVESTING/ PICKING	Qty_handled	Qty_handled	30	0	1100.000000	8400.000000	7300.000000	
		Qty_lost	Qty_lost	30	0	0	1200.000000	1200.000000	
	TRANSPORTATION	Qty_handled	Qty_handled	11	0	1050.000000	5300.000000	4250.000000	
		Qty_lost	Qty_lost	11	0	0	0	0	
BLACKGRAM	HARVESTING/ PICKING	Qty_handled	Qty_handled	10	0	300.000000	1600.000000	1300.000000	
		Qty_lost	Qty_lost	10	0	0	35.000000	35.000000	
	PACKING	Qty_handled	Qty_handled	9	0	290.000000	865.000000	575.000000	
		Qty_lost	Qty_lost	9	0	0.500000	2.000000	1.500000	
	THRESHING/ DEHUSKING	Qty_handled	Qty_handled	9	0	290.000000	1560.000000	1270.000000	
		Qty_lost	Qty_lost	9	0	0	25.000000	25.000000	
	TRANSPORTATION	Qty_handled	Qty_handled	9	0	289.000000	863.000000	574.000000	
		Qty_lost	Qty_lost	9	0	0	0.500000	0.500000	

SPSS: AN OVERVIEW

Seema Jaggi

Indian Agricultural Statistics Research Institute, New Delhi-110012

35.1 Introduction

The abbreviation SPSS stands for **Statistical Package for the Social Sciences** and is a comprehensive system for analyzing data. This package of programs is available for both personal and mainframe (or multi-user) computers. SPSS package consists of a set of software tools for data entry, data management, statistical analysis and presentation. SPSS integrates complex data and file management, statistical analysis and reporting functions. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and complex statistical analyses.

35.2 FEATURES OF SPSS

- (i) It is easy to learn and use
- (ii) It includes a full range of data management system and editing tools
- (iii) It provides in-depth statistical capabilities
- (iv) It offers complete plotting, reporting and presentation features.

SPSS makes statistical analysis accessible for the casual user and convenient for the experienced user. The data editor offers a simple and efficient spreadsheet-like facility for entering data and browsing the working data file. To invoke SPSS in the windows environment, select the appropriate **SPSS** icon. There are a number of different types of windows in SPSS.

Data Editor: This window displays the contents of the data file. One can create new data files or modify existing ones. The Data Editor window opens automatically when one starts an SPSS session. One can have only one data file open at a time. This editor provides two views of the data.

- **Data view:** Displays the actual data values or defined value labels.
- **Variable view:** Displays variable definition information, including defined variable and value labels, data type, etc.

With the Data Editor, one can modify data values in the Data view in many ways like change data values; cut, copy and paste data values; add and delete cases; add and delete variables, change the order of variables.

Viewer: All statistical results, tables, and charts are displayed in the Viewer. The output can be edited and saved for later use. A Viewer window opens automatically the first time you run a procedure that generates output.

Draft Viewer: The output can be displayed as a simple text in this window.

Syntax Editor: One can paste the dialog box choices into a syntax window, where the selections appear in the form of command syntax. One can then edit the command syntax to utilize special features of SPSS not available through dialog boxes. These commands can be saved in a file for use in subsequent SPSS sessions.

Pivot Table Editor: Output is displayed in pivot tables that can be modified in many ways with this editor. One can edit text, swap data in rows and columns, create multidimensional tables, and selectively hide and show results.

Text Output Editor: Text output not displayed in pivot tables can be modified with the Text Output Editor. One can edit the output and change font characteristics (type, style, colour and size).

Chart Editor: High-resolution charts and plots can be modified in chart windows. One can change the colours, select different type of fonts and sizes etc.

Many of the tasks that are to be performed with SPSS start with **menu** selections. Each window has its own menu bar with menu selections appropriate for that window type. The various procedures under SPSS are

File Edit View Data Transform Analyze Graphs Utilities Windows Help

Analyze and Graphs menus are available on all windows, making it easy to generate new output without having to switch windows. Most menu selections open dialog boxes. One can use dialog boxes to select variables and options for analysis. Since most procedures provide a great deal of flexibility, not all of the possible choices can be contained in a single dialog box. The main dialog box usually contains the minimum information required to run a procedure. Additional specifications are made in subdialog boxes. All these above-mentioned options have further suboptions. To see what applications there are, we simply move the cursor to a particular option and press, when a drop-down menu will appear. To cancel a drop-down menu, place the cursor anywhere outside the option and press the left button.

The three dots after an option term (...) on a drop-down menu, such as **Define Variable...**option in Data option, signifies that a dialog box will appear when this option is chosen. To cancel a dialog box, select the **Cancel** button in the dialog box. A right-facing arrowhead after an option term indicates that a further submenu will appear to the right of the drop-down menu. An option with neither of these signs means that there are no further drop-down menus to select. There are five standard command pushbuttons in most dialog boxes.

OK: Runs the procedure. After the variables and additional specifications are selected, click OK to run the procedure.

Paste: Generates command syntax from the dialog box selections and pastes the syntax into a syntax window.

Reset: Deselects any variables in the selected variable list and resets all specifications in the dialog box.

Cancel: Cancels any changes in the dialog box settings since the last time it was opened and closes the dialog box.

Help: Contains information about the current dialog box.

Entering and Editing data

The easiest way of entering data in SPSS is to type it directly into the matrix of columns and numbered rows in the **Data Editor** window. The columns represent variables and the rows represent cases. The variables can be defined in the variable view. Variable name must be no longer than eight characters and the name must begin with a letter.

Saving data

To be able to retrieve a file, we need to save it and give it a name. The default extension name for saving files is **sav**. Thus, we could call our data file see .sav.

To save this file on a floppy disk, we carry out the following sequence:

→**File** →**Save As...** [opens **Save Data As** dialog box]

→box under **Drives:** →drive [e.g. **a**] from options listed

→box under **File Name:**, delete the asterisk and type file

stem name [e.g. see] →**OK**

The output file can also be printed and saved. The extension name for output file is **spo**.

Retrieving a saved file

To retrieve this file at a later stage when it is no longer the current file, use the following procedure:

→**File**→**Open**→**Data...**[opens the **Open Data File** dialog box]

→box under **Drives:** →drive [e.g. **a**]from options listed

→box under **File Name:** →file name [e.g. **see.sav**] → **OK**

35.3 BASIC STEPS IN DATA ANALYSIS

- **Get your data into SPSS:** You can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter your data directly in the Data Editor.
- **Select a procedure:** Select a procedure from the menus to calculate statistics or to create a chart.

- **Select the variables for the analysis:** The variables in the data file are displayed in a dialog box for the procedure.
- **Run the procedure:** Results are displayed in the Viewer.

35.4 STATISTICAL PROCEDURES

After entering the data set in **Data Editor** or reading an ASCII data file, we are now ready to analyze it. The **Analyze** option has the following sub options:

Reports

Descriptive Statistics

Custom Tables

Compare Means

General Linear Model (GLM)

Correlate

Regression

Loglinear

Classify

Data Reduction

Scale

Non Parametric Tests

Time Series

Survival

Multiple Response

DESCRIPTIVE STATISTICS: This submenu provides techniques for summarising data with statistics, charts, and reports. The various sub-sub menus under this are as follows:

Frequencies provides information about the relative frequency of the occurrence of each category of a variable. This can be used it to obtain summary statistics that describe the typical value and the spread of the observations. To compute summary statistics for each of several groups of cases, Means procedure or the Explore procedure can be used.

Descriptives is used to calculate statistics that summarize the values of a variable like the measures of central tendency, measures of dispersion, skewness, kurtosis etc.

Explore produces and displays summary statistics for all cases or separately for groups of cases. Boxplots, stem-and leaf plots, histograms, tests of normality, robust estimates of location, frequency tables and other descriptive statistics and plots can also be obtained.

Crosstabs is used to count the number of cases that have different combinations of values of two or more variables, and to calculate summary statistics and tests. The variables you use to form the categories within which the counts are obtained should have a limited number of distinct values.

List Cases displays the values of variables for cases in the data file.

Report Summaries in Rows produces reports in which different summary statistics are laid out in rows. Case listings are also available from this command, with or without summary statistics.

Report Summaries in Columns produces reports in which different summary statistics are laid out in separate columns.

Custom Tables submenu provides attractive, flexible displays of frequency counts, percentages and other statistics.

COMPARE MEANS: This submenu provides techniques for testing differences among two or more means for both independent and related samples.

Means computes summary statistics for a variable when the cases are subdivided into groups based on their values for other variables.

Independent Sample t test is used if two unrelated samples come from populations with the same mean. The observations should be from two unrelated groups, and for testing, the mean must be an appropriate summary measure for the variable to be compared in the two groups. For more than two independent groups, the **One-way ANOVA** option could be used.

Paired Sample t test is used to compare the means of the same subjects in two conditions or at two points in time i.e. to compare subjects who had been matched to be similar in certain respects and then to test if two related samples come from populations with the same mean. The related, or paired, samples often result from an experiment in which the same person is observed before and after an intervention. If the distribution of the differences of the values between the members of a pair is markedly non-normal you should consider one of the non-parametric tests.

One-Way ANOVA is used to test that several independent groups come from populations with the same mean. To see which groups are significantly different from each other, multiple comparison procedures can be used through **Post Hoc Multiple Comparison option** which consist of the options like **Least-significant difference, Duncan's multiple range test, Scheffe**, etc. The contrast analysis can also be performed in order to compare the different groups or treatments by using the **Contrast** option. The data obtained using completely randomized design can be analyzed through this option.

GENERAL LINEAR MODEL: This submenu provides techniques for testing univariate and multivariate Analysis-of-Variance models, including repeated measures. The **Univariate** suboption could be used to analyze the experimental designs like Completely randomized design, Randomized block design, Latin square design, Designs for factorial experiments, etc.

The covariance analysis can also be performed and alternate methods for partitioning sums of squares can be selected.

If only some of the interactions of a particular order are to be included, the **Custom** procedure should be used. If there is only one factor then One-Way ANOVA procedure should be used.

Multivariate analyses analysis-of-variance and analysis-of-covariance designs when you have two or more correlated dependent variables.

Multivariate analysis of variance is used to test hypotheses about the relationship between a set of interrelated dependent variables and one or more factor or grouping variables. For example, you can test whether verbal and mathematical test scores are related to instructional method used, sex of the subject, and the interaction of method and sex.

This procedure should be used only if there are several dependent variables that are related to each other. For a single dependent variable or unrelated dependent variables, the Univariate ANOVA procedures can be adopted. If the same dependent variable is measured on several occasions for each subject, the Repeated Measures procedure is to be used.

Repeated Measures is used to test hypotheses about the means of a dependent variable when the same dependent variable is measured on more than one occasion for each subject.

Subjects can also be classified into mutually exclusive groups, such as males or females, or type of job held. Then you can test hypotheses about the effects of the between-subject variables and the within-subject variables, as well as their interactions.

CORRELATE: This submenu provides measures of association for two or more variables measured at the interval level.

Bivariate calculates matrices of Pearson product-moment correlations, and of Kendall and Spearman non-parametric correlations, with significance levels and optional univariate statistics.

The **correlation coefficient** is used to quantify the strength of the linear relationship between two variables.

The **Pearson correlation coefficient** should be used only for data measured at the interval or ratio level. Spearman and Kendall correlation coefficients are non-parametric measures which are particularly useful when the data contain outliers or

when the distribution of the variables is markedly non-normal. Both the Spearman and Kendall coefficients are based on assigning ranks to the variables.

Partial calculates **partial correlation coefficients** that describe the relationship between two variables, while adjusting for the effects of one or more additional variables.

If the values of a dependent variable from a set of independent variables are to be predicted then the Linear Regression procedure may be used. If there are no control variables then the Bivariate Correlations procedure can be adopted. Nominal variables should not be used in the partial correlation procedure.

REGRESSION: This submenu provides a variety of regression techniques, including linear, logistic, nonlinear, weighted, and two-stage least-squares regression.

Linear is used to examine the relationship between a dependent variable and a set of independent variables. If the dependent variable is dichotomous, then the logistic regression procedure should be used. If the dependent variable is censored, such as survival time after surgery, use the Life Tables, Kaplan-Meier, or proportional hazards procedure.

Logistic estimates regression models in which the dependent variable is dichotomous.

If the dependent variable has more than two categories, use the Discriminant procedure to identify variables which are useful for assigning the cases to the various groups. If the dependent variable is continuous, use the Linear Regression procedure to predict the values of the dependent variable from a set of independent variables.

Probit performs probit analysis that is used to measure the relationship between a response proportion and the strength of a stimulus.

For example, the probit procedure can be used to examine the relationship between the proportion of plants dying and the strength of the pesticide applied or to examine the relationship between the proportion of people buying a product and the magnitude of the incentive offered. The Probit procedure should be used only if the response is dichotomous-buy/not buy, alive/dead-and several groups of subjects are exposed to different levels of some stimulus. For each stimulus level, the data must contain counts of the totals exposed and the totals responding.

If the response variable is dichotomous but you do not have groups of subjects with the same values for the independent variables you should use the Logistic Regression procedure.

Nonlinear estimates nonlinear regression models, including models in which parameters are constrained.

The nonlinear regression procedure can be used if one knows the equation whose parameters are to be estimated, and the equation cannot be written as the sum of parameters times some function of the independent variables. In nonlinear regression the parameter estimates are obtained iteratively.

If the function is linear, or can be transformed to a linear function, then the Linear Regression procedure should be used.

The **Loglinear** submenu provides general and hierarchical log-linear analysis and logit analysis.

CLASSIFY: This submenu provides cluster and discriminant analysis.

K-means Cluster performs cluster analysis using an algorithm that can handle large numbers of cases, but that requires you to specify the number of clusters.

The goal of cluster analysis is to identify relatively homogeneous groups of cases based on selected characteristics.

If the number of clusters to be formed is not known, then Hierarchical Cluster procedure can be used. If the observations are in known groups and one wants to predict group membership based on a set of independent variables, then the Discriminant procedure can be used.

Hierarchical Cluster combines cases into clusters hierarchically, using a memory-intensive algorithm that allows you to examine many different solutions easily.

Discriminant is used to classify cases into one of several known groups on the basis of various characteristics. To use the Discriminant procedure the dependent variable must have a limited number of distinct categories. Independent variables that are nominal must be recoded to dummy or contrast variables.

If the dependent variable has two categories, Logistic Regression can be used. If the dependent variable is continuous one may use Linear Regression.

DATA REDUCTION: This submenu provides factor analysis, correspondence analysis, and optimal scaling.

Factor is used to identify factors that explain the correlations among a set of variables. Factor analysis is often used to summarize a large number of variables with a smaller number of derived variables, called factors.

Distances compute many different measures of similarity, dissimilarity or distance. Many different measures can be used to quantify how much alike or how different two cases or variables are. Similarity measures are constructed so that large values indicate much similarity and small values indicate little similarity. Dissimilarity measures estimate the distance or unlikeness of two cases. A large dissimilarity value tells that two cases or variables are far apart. In order to decide which similarity or dissimilarity measure to use, one must consider the characteristics of the data. Special measures are available for interval data, frequency counts, and binary data. If the cases are to be classified into groups based on similarity or dissimilarity measures, one of the Cluster procedures should be used.

The **Conjoint** submenu provides for the generation and analysis of conjoint designs.

SCALE: This submenu provides reliability analysis and multidimensional scaling.

NONPARAMETRIC TESTS: This submenu provides non-parametric tests for one sample, or for two and more paired or independent samples.

Chi-Square is used to test hypotheses about the relative proportion of cases falling into several mutually exclusive groups. For example, if one wants to test the hypotheses that people are equally likely to buy six different brands of cereals, one can count the number buying each of the six brands. Based on the six observed counts Chi-Square procedure could be used to test the hypothesis that all six cereals are equally likely to be bought. The expected proportions in each of the categories don't have to be equal. The hypothetical proportions to be tested should be specified.

Binomial is used to test the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values. For example, to test that the probability of an item on the assembly line is defective is one out of ten ($p=0.1$), take a sample of 300 items and record whether each is defective or not. Then use the binomial procedure to test the hypothesis of interest.

Runs is used to test whether the two values of a dichotomous variable occur in a random sequence. The run test is appropriate only when the order of cases in the data file is meaningful.

1-Sample Kolmogorov-Smirnov is used to compare the observed frequencies of the values of an ordinal variable, such as rated quality of work, against some specified theoretical distribution. It determines the statistical significance of the largest difference between them. In SPSS, the theoretical distribution can be **Normal, Uniform or Poisson**.

Alternative tests for normality are available in the Explore procedure, in the Summarize submenu. The P-P and Q-Q plots in the Graphs menu can also be used to examine the assumption of normality.

2-Independent Samples is used to compare the distribution of a variable between two non-related groups. Only limited assumptions are needed about the distributions from which the sample are selected. The Mann-Whitney U test is an alternative to the two sample t-test. The actual values of the data are replaced by ranks. The Kolmogorov-Smirnov test is based on the differences between the observed cumulative distributions of the two groups. The Wald-Woflowitz run tests sorts the data values from smallest to largest and then performs a run test on the group numbers. The Moses Test of Extreme Reaction is used to test for differences in range between two groups.

K-Independent Samples is used to compare the distribution of a variable between two or more groups. Only limited assumptions are needed about the distributions from which the samples are selected. The Kruskal-Wallis test is an alternative to one-way analysis of variance, with the actual values of the data replaced by ranks. The Median test counts the number of cases in each group that are above and below the combined median, and then performs a chi-square test.

2 Related Samples is used to compare the distribution of two related variables. Only limited assumptions are needed about the distributions from which the samples are

selected. The Wilcoxon and Sign tests are non-parametric alternative to the paired samples t-test. The Wilcoxon test is more powerful than the Sign test.

McNemar's test is used to determine changes in proportions for related samples. It is often used for "before and after" experimental designs when the dependent variable is dichotomous. For example, the effect of a campaign speech can be tested by analyzing the number of people whose preference for a candidate changed based on the speech. Using McNemar's test you analyze the changes to see if change in both directions is equally likely.

K Related Samples is used to compare the distribution of two or more related variables. Only limited assumptions are needed about the distributions from which the samples are selected. The Friedman test is a non-parametric alternative to a single-factor repeated measures analysis of variance. You can use it when the same measurement is obtained on several occasions for a subject. For example, the Friedman test can be used to compare consumer satisfaction of 5 products when each person is asked to rate each of the products on a scale.

Cochran's Q test can be used to test whether several dichotomous variables have the same mean. For example, if instead of asking each subject to rate their satisfaction with five products, you asked them for a yes/no response about each, you could use Cochran's test to test the hypothesis that all five products have the same proportion of satisfied users.

Kendall's W measures the agreement among raters. Each of your cases corresponds to a rater, each of the selected variables is an item being rated. For example, if you ask a sample of customers to rank 7 ice-cream flavors from least to most liked, you can use Kendall's W to see how closely the customers agree in their ratings.

The **Time series** submenu provides exponential smoothing, autocorrelated regression, ARIMA, X11 ARIMA, seasonal decomposition, spectral analysis, and related techniques.

The **Survival** submenu provides techniques for analyzing the time for some terminal event to occur, including Kaplan-Meier analysis and Cox regression.

Multiple response submenu provides facilities to define and analyze multiple-response or multiple-dichotomy sets.

Weight Estimation estimates a linear regression model with differential weights representing the precision of observations. This command is in the Professional Statistics option.

If the variance of the dependent variable is not constant for all of the values of the independent variable, weights that are inversely proportional to the variance of the dependent variable can be incorporated into the analysis. This results in a better solution.

The Weight Estimation procedure can also be used to estimate the weights when the variance of the dependent variable is related to the values of an independent variable. If you know the weights for each case you can use the linear regression procedure to

obtain a weighted least squares solution. The linear regression procedure provides a large number of diagnostic statistics that help you evaluate how well the model fits your data.

2-Stage Least Squares performs two-stage least squares regression for models in which the error term is related to the predictors. This command is in the Professional Statistics option.

For example, if you want to model the demand for a product as a function of price, advertising expenses, cost of the materials, and some economic indicators, you may find that the error term of the model is correlated with one or more of the independent variables. Two-stage least squares allows you to estimate such a model.

Correspondence Analysis analyzes correspondence tables (such as cross-tabulations) to best measure the distances between categories or between variables. This command is in the Categories option.

Homogeneity Analysis is an optimal scaling procedure analogous in some ways to factor analysis, but capable of analyzing categorical or ordinal variables. The technique is also known as multiple correspondence analysis. This command is in the Categories option.

Nonlinear Components performs nonlinear principal-components analysis to try to reduce the dimensionality of a set of variables. This command is in the Categories option.

OVERALS performs nonlinear canonical correlation analysis to determine how similar sets of variables are to one another. This command is in the Categories option.

35.5 TRANSFORM

Compute calculates the values for either a new or an existing variable, for all cases or for cases satisfying a logical criterion.

Random Number Seed sets the seed used by the pseudo-random number generator to a specific value, so that you can reproduce a sequence of pseudo-random numbers.

Count creates a variable that counts the occurrences of the same value(s) in a list of variables for each case.

Recode into Same Variables reassigns the values of existing variables or collapses ranges of existing values into new values.

Recode into Different Variables reassigns the values of existing variables to new variables or collapses ranges of existing values into new variables.

Rank Cases creates new variables containing ranks, normal scores, or similar ranking scores for numeric variables.

Automatic Recode reassigns the values of existing variables to consecutive integers in new variables.

Create Time Series creates a time-series variable as a function of an existing series, for example, lagged or leading values, differences, cumulative sums. This command is in the Trends option.

Replace Missing Values substitutes non-missing values for missing values, using the series mean or one of several time-series functions. This command is in the Trends option.

Run Pending Transforms executes transformation commands that are pending due to the Transformation Options setting in the Preferences dialog.

35.6 UTILITIES

Command Index takes you to the dialog box for a command if you know its name in the SPSS command language.

Fonts let you choose a font, style, and size for SPSS Data Editor, output, and syntax windows.

Variable Information displays the Variables window, which shows information about the variables in your working data file, and allows you to scroll the data editor to a specific variable, or copy variable names to the designated syntax window.

File Information displays information about the working data file in the output window.

Output Page Titles lets you specify a title and subtitle for output from SPSS. They appear in the page header, if it is displayed. (Preferences in the Edit menu controls the page header.)

Define Sets defines sets of variables for use in other dialog boxes.

Use Sets lets you select which defined sets of variables should appear in the source-variable lists of other dialog boxes.

Grid Lines turns grid lines on and off in the Data Editor window. This command is available when the Data Editor is active.

Value Labels turns on and off the display of Value Labels (instead of actual values) in the Data Editor window. When Value Labels are displayed you can edit data with a pop-up menu of labels. This command is available when the Data Editor is active.

Auto New Case turns on and off the automatic creation of new cases by cursor movement below the last case in the Data Editor window. This command is available when the Data Editor is active.

Designate Window designates the active window to receive output from SPSS commands (if it is an output window); or to receive commands pasted from dialog boxes (if it is a syntax window). You can also designate a window by clicking the ! button on its icon bar. This command is available when an output or syntax window is active.

35.7 GRAPHS

Bar generates a simple, clustered, or stacked bar chart of the data.

Line generates a simple or multiple line chart of the data.

Area generates a simple or stacked area chart of the data.

Pie generates a simple pie chart or a composite bar chart from the data.

High-Low plots pairs or triples of values, for example high, low, and closing prices.

Pareto creates Pareto charts, bar charts with a line superimposed showing the cumulative sum.

Control produces the most commonly-used process-control charts.

Boxplot generates boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.

Scatter generates a simple or overlay scatterplot, a scatterplot matrix, or a 3-D scatterplot from the data.

Histogram generates a histogram showing the distribution of an individual variable.

Normal P-P plots the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.

Normal Q-Q plots the quantiles of a variable's distribution against the quantiles of the normal distribution.

Sequence produces a plot of one or more variables by order in the file, suitable for examining time-series data.

Time Series: Autocorrelations calculates and plots the autocorrelation function (ACF) and partial autocorrelation function of one or more series to any specified number of lags, displaying the Box-Ljung statistic at each lag to test the overall hypothesis that the ACF is zero at all lags.

Time Series: Cross-correlations calculate and plots the cross-correlation function of two or more series for positive, negative, and zero lags.

Time Series: Spectral calculates and plots univariate or bivariate periodograms and spectral density functions, which express variation in a time series (or covariation in two time series) as the sum of a series of sinusoidal components. It can optionally save various components of the frequency analysis as new series.

ANALYSIS OF SURVEY DATA USING SPSS

U.C. Sud

Indian Agricultural Statistics Research Institute, New Delhi-110012

36.1 INTRODUCTION

SPSS version 13.0 has many additional features over the version 12.0. One notable addition in version 13.0 is the SPSS complex samples. We cover some of the features of SPSS complex samples using data provided in the SPSS package.

There are many new features in SPSS 13.0.

36.2 SAMPLES PROCEDURES

An inherent assumption of analytical procedures in traditional software packages is that the observations in a data file represent a simple random sample from the population of interest. This assumption is untenable for an increasing number of companies and researchers who find it both cost-effective and convenient to obtain samples in a more structured way.

The SPSS Complex Samples option allows one to select a sample according to a complex design and incorporate the design specifications into the data analysis, thus ensuring that your results are valid.

36.3 PROPERTIES OF COMPLEX SAMPLES

A complex sample can differ from a simple random sample in many ways. In a simple random sample, individual sampling units are selected at random with equal probability and without replacement (WOR) directly from the entire population. By contrast, a given complex sample can have some or all of the following features:

36.4 STRATIFICATION

Stratified sampling involves selecting samples independently within non-overlapping subgroups of the population, or strata. For example, strata may be socioeconomic groups, job categories, age groups, or ethnic groups. With stratification, one can ensure adequate sample sizes for subgroups of interest, improve the precision of overall estimates, and use different sampling methods from stratum to stratum.

36.5 CLUSTERING

Cluster sampling involves the selection of groups of sampling units, or clusters. For example, clusters may be schools, hospitals, or geographical areas, and sampling units may be students, patients, or citizens. Clustering is common in multistage designs and area (geographic) samples.

36.6 MULTIPLE STAGES

In multistage sampling, one selects a first-stage sample based on clusters. Then it creates a second-stage sample by drawing subsamples from the selected clusters. If the second-stage sample is based on sub-clusters, one can then add a third stage to the sample. For example, in the first stage of a survey, a sample of cities could be drawn. Then, from the selected cities, households could be sampled.

Finally, from the selected households, individuals could be polled. The Sampling and Analysis Preparation wizards allow you to specify three stages in a design.

36.7 NON RANDOM SAMPLING

When selection at random is difficult to obtain, units can be sampled systematically (at a fixed interval) or sequentially.

36.8 UNEQUAL SELECTION PROBABILITIES

When sampling clusters that contain unequal numbers of units, one can use probability-proportional-to-size (PPS) sampling to make a cluster's selection probability equal to the proportion of units it contains. PPS sampling can also use more general weighting schemes to select units.

36.9 UNRESTRICTED SAMPLING

Unrestricted sampling selects units with replacement (WR). Thus, an individual unit can be selected for the sample more than once.

36.10 SAMPLING WEIGHTS

Sampling weights are automatically computed while drawing a complex sample and ideally correspond to the "frequency" that each sampling unit represents in the target population. Therefore, the sum of the weights over the sample should estimate the population size. Complex Samples analysis procedures require sampling weights in order to properly analyze a complex sample. Note that these weights should be used entirely within the Complex Samples option and should not be used with other analytical procedures via the Weight Cases procedure, which treats weights as case replications.

36.11 USAGE OF COMPLEX SAMPLES PROCEDURES

The usage of Complex Samples procedures depends on the particular needs. The primary types of users are those who: Plan and carry out surveys according to complex designs, possibly analyzing the sample later. The primary tool for surveyors is the Sampling Wizard. Analyze sample data files previously obtained according to complex designs.

Before using the Complex Samples analysis procedures, one may need to use the Analysis Preparation Wizard. Regardless of which type of user one may be, one needs to supply design information to Complex Samples procedures. This information is stored in a **plan file** for easy reuse.

36.12 PLAN FILES

A plan file contains complex sample specifications. There are two types of plan files:

Sampling Plan The specifications given in the Sampling Wizard define a sample design that is used to draw a complex sample. The sampling plan file contains those specifications. The sampling plan file also contains a default analysis plan that uses estimation methods suitable for the specified sample design.

Analysis Plan This plan file contains information needed by Complex Samples analysis procedures to properly compute variance estimates for a complex sample. The plan includes the sample structure, estimation methods for each stage, and references to required variables, such as sample weights. The Analysis Preparation Wizard allows you to create and edit analysis plans.

There are several advantages to saving your specifications in a plan file, including:

A surveyor can specify the first stage of a multistage sampling plan and draw first-stage units now, collect information on sampling units for the second stage, and then modify the sampling plan to include the second stage.

An analyst who doesn't have access to the sampling plan file can specify an analysis plan and refer to that plan from each Complex Samples analysis procedure. A designer of large-scale public use samples can publish the sampling plan file, which simplifies the instructions for analysts and avoids the need for each analyst to specify his or her own analysis plans.

36.13 SAMPLING WIZARD: SAMPLING METHOD

This step allows one to specify how to select cases from the working data file.

Method Controls in this group are used to choose a selection method. Some sampling types allow one to choose whether to sample with replacement (WR) or without replacement (WOR). See the type descriptions for more information. Note that some probability-proportional-to-size (PPS) types are available only when clusters have been defined and that all PPS types are available only in the first stage of a design. Moreover, WR methods are available only in the last stage of a design.

- **Simple Random Sampling** Units are selected with equal probability. They can be **selected** with or without replacement.
- **Simple Systematic** Units are selected at a fixed interval throughout the sampling frame (or strata, if they have been specified) and extracted without replacement. A randomly selected unit within the first interval is chosen as the starting point.
- **Simple Sequential** Units are selected sequentially with equal probability and without replacement.

- **PPS** This is a first-stage method that selects units at random with probability proportional to size. Any units can be selected with replacement; only clusters can be sampled without replacement.
- **PPS Systematic** This is a first-stage method that systematically selects units with probability proportional to size. They are selected without replacement.
- **PPS Sequential** This is a first-stage method that sequentially selects units with probability proportional to cluster size and without replacement.
- **PPS Brewer** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Murthy** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Sampford** This is a first-stage method that selects more than two clusters from each stratum with probability proportional to cluster size and without replacement. It is an extension of Brewer's method. A cluster variable must be specified to use this method.
- **Use WR estimation for analysis.** By default, an estimation method is specified in the plan file that is consistent with the selected sampling method. This allows one to use with-replacement estimation even if the sampling method implies WOR estimation. This option is available only in stage 1.

36.14 MEASURE OF SIZE (MOS)

If a PPS method is selected, one must specify a measure of size that defines the size of each unit. These sizes can be explicitly defined in a variable or they can be computed from the data. Optionally, one can set lower and upper bounds on the MOS, overriding any values found in the MOS variable or computed from the data. These options are available only in stage 1.

Preparing a Complex Sample for Analysis

The Analysis Preparation Wizard guides you through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, one should have a sample drawn according to a complex design.

Creating a new plan is most useful when one does not have access to the sampling plan file used to draw the sample (recall that the sampling plan contains a default analysis plan). If one does have access to the sampling plan file used to draw the sample, one can use the default analysis plan contained in the sampling plan file or override the default analysis specifications and save ones changes to a new file.

Steps for Drawing the Sample and Analysis of Sampled Data

- START - SPSS 13.0 for windows

OR

- START - All Programs - SPSS for Windows - SPSS 13.0 for windows

Prepare a file from which data to be sampled in **SPSS Data Editor**

File - Open - Data (employee data.sav)

36.15 TO DRAW SAMPLES

ANALYZE

COMPLEX SAMPLES

Select a Sample

1. Sampling Wizard

Design a **Sample Plan File** (emp.csplan)/

Edit a sample design/

Draw a sample

Next

2. Sampling Wizard - Stage1 - **design variables** - Stratify By - Clusters -Input sample weight

Next

3. Sampling Wizard - **Sampling Method** – Type -WOR/WR

Next

4. Sampling Wizard-**Sample Size** -Units - Counts/Proportions – Value - Unequal values for Strata

Next

5. Sampling Wizard-**Output Variables**- Population size/Sample proportion/
Sample size/Sample weight

Next

6. Sampling Wizard-**Plan Summary** – Summary -Add stage 2

Next

7. Sampling Wizard-**Draw a sample** - Selection options - Do you want draw sample Yes/No - What type of seed value do you want to use-A random chosen number/ Custom value

Next

8. Sampling Wizard-Draw sample: **Out files** - Working data file/ External file (emp.sav)

Next

9. Sampling Wizard -Completing the sampling wizard - Save the design to a plan file and draw the sample - Finish

10. Output- **SPSS Viewer**

36.16 TO ANALYZING SAMPLE DATA

Open a selected sample file in **SPSS Data Editor**

File - Open – Data - SPSS data document (emp.sav)

ANALYZE

COMPLEX SAMPLES

Descriptives-

1. Complex Samples Plan for Descriptives analysis Wizard -
File – Browse - Plan file name (emp.csplan)
Continue
2. Complex Samples Descriptives Wizard – Measures - Sub Population - ok
3. Output- **SPSS Viewer**

OVERVIEW OF R SOFTWARE

Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi-110012

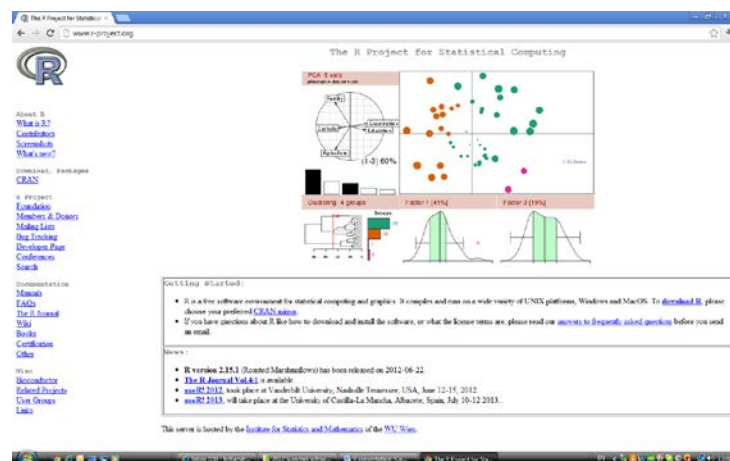
Email: hchandra@iasri.res.in

37.1 INTRODUCTION

R is a free software environment for statistical computing and graphics. It is almost perfectly compatible with S-plus. The only thing you need to do is download the software from the internet and use an editor to write your program (e.g. Notepad). It contains most standard methods of statistics as well as lot of less commonly used methods and can be used for programming and to construct your own functions. It is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. It is available for download from <http://www.r-project.org/>. The primary purpose of this chapter is to introduce R software for beginners.

37.2 TO DOWNLOAD R SOFTWARE

In any web browser (e.g. Microsoft Internet Explorer), go to: <http://www.r-project.org>. You will see a page like

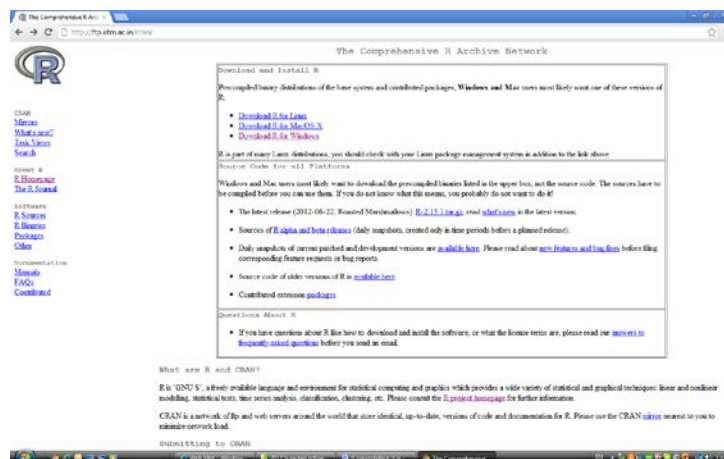


Downloads: CRAN (On your left hand side you will see CRAN).

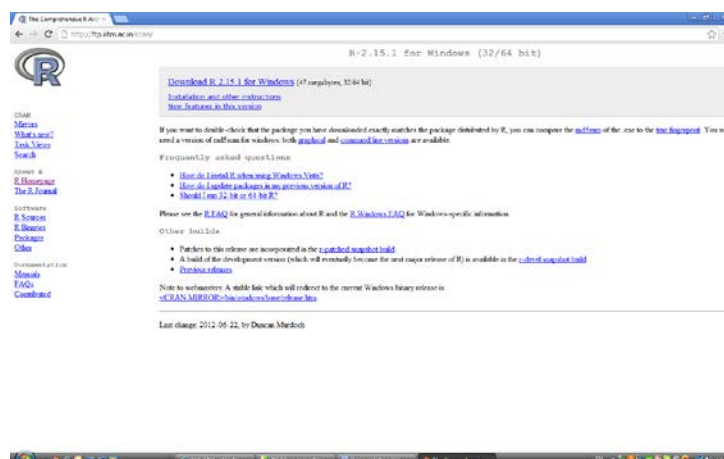
OVERVIEW OF R SOFTWARE



- Set your Mirror: Anyone in the **India** or **any other country** is fine.
- On your right hand side you will see Download R for Windows. Click there
- Click on [base](#)



- Click on [R-3.0.1.exe](#) and save it to your hard disc.



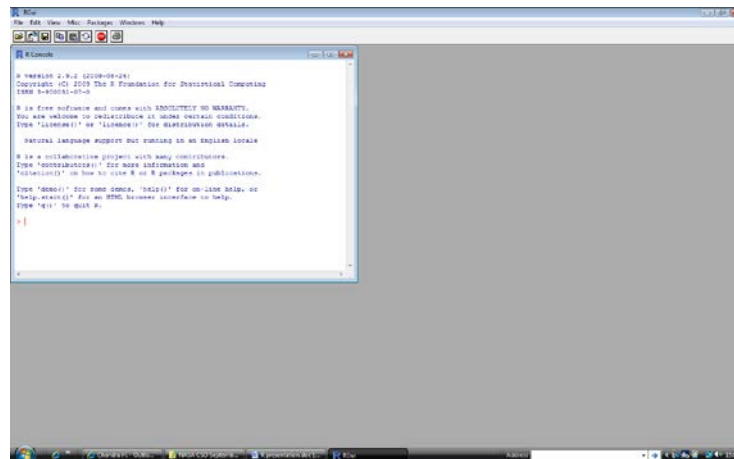
- This is the latest available version of the software. It is an '.exe' file, which you can save in your hard disc. By double clicking on the name of this file, R is automatically installed. All you need to do is follow the installation process.

37.3 TO OPEN R SOFTWARE

The installation process automatically creates a shortcut for R. Double click this icon to open the R environment. R will open up with the appearance of a standard Windows implementation (i.e. various windows and pull-down menus). Note that R is an interpreted language and processes commands on a line by line basis. Consequently it is necessary to hit **ENTER** after typing in (or pasting) a line of R code in order to get R to implement it.

37.4 TO RUN R PROGRAM CODE

The main active window within the R environment is the **R Console**. This is a line editor and output viewer combined into one window. Here at the command **prompt** (the symbol `>`), we can enter R commands which run instantly upon pressing the carriage return key. This sign (`>`) is called prompt, since it prompts the user to write something, see below.



```

R Console
R version 2.10.1 (2009-08-08)
Copyright (C) 2009 The R Foundation for Statistical Computing
1000 N-10000-10-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.search()' for an HTML browser interface to help.
Type 'q()' to quit R.

>

```

We can also run blocks of code which we have copied into the paste buffer from another source. In this session we shall use the Windows-supplied editor Notepad to display and edit our R program code. If we were to write some R of code, then simply copy it from the editor and paste it into the R Console, then the code would run in real time.

37.5 TO OPEN THE EDITOR

Here we are using the Windows-supplied editor **Notepad** to display and edit our R program code, although any general-purpose editor will suffice. Open Notepad by going to the Start button and clicking on:

Start > All Programs > Accessories > Notepad

Having opened Notepad, open the file, for example, **Intro_to_R.txt** (containing the program code, assume that it is copied in `C: / derive`) by selecting the following option from the pull-down menu:

File > Open

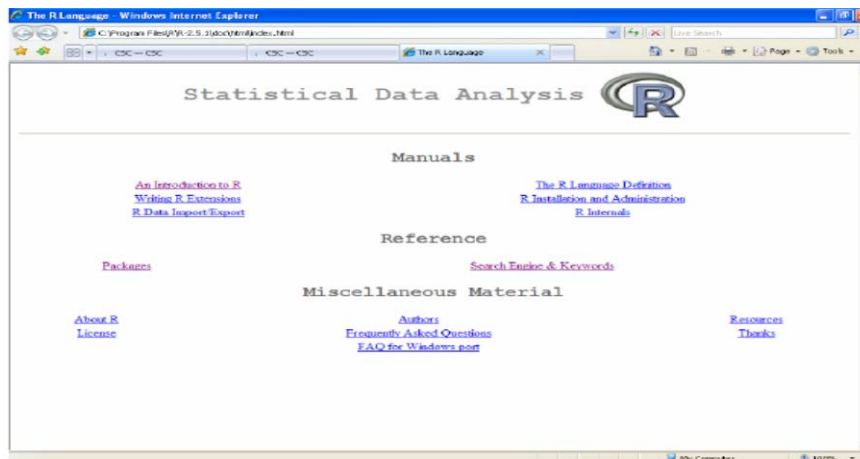
Click on the down-arrow at the top of the “Open” dialog box and change the selection to “Look in” C:\. You should now see the filename **Intro_to_R.txt** among a list of files. Double-click on the filename to open it.

A COUPLE OF OTHER USEFUL THINGS

- Please remember that R is **case-sensitive** so we need to be consistent in our use of lower and upper case letters, both for commands and for objects.
- When the program has finished, we should see the **red** command prompt (`>`) pop up in the R Console window. This indicates that control is returned to the user, so that you can now type more R commands if you wish.
- A **comment** in R code begins with a hash symbol (`#`). Whole lines may be commented or just the tail-end of a line. Examples are:

37.6 HELP

Html-help can be invoked from the Help-menu. From the opening webpage, you can access manuals, frequently asked questions, references to help for individual packages, and most importantly, Search Engine. Help is the best place to find out new functions, and get descriptions on how to use them.



37.7 GETTING STARTED WITH R

Commands in R are given at the command prompt.

Simple calculations, vectors and graphics

To begin with, we’ll use R as a calculator. Try the following commands:

`2+7`

`2/(3+5)`

`sqrt(9)+5^2`

`sin(pi/2)-log(exp(1))`

Help about a specific command can be had by writing a question mark before the command, for instance:

?log

As an alternative, help can be used; in this case, help (log). The help files are a great resource and you will soon find yourself using them frequently.

Comments can be written using the #-symbol as follows:

```
2+3          # The answer should be 5
```

Vectors and matrices

Vectors and matrices are of great importance in many numerical problems. To create a vector named mydata and assign the values 7, -2, 5 to it, we write as follows:

```
mydata <- c(7,-2,5)
```

The symbol <- (or alternatively use =) should be read as “assigns”. The command c can be interpreted (by you, the user) as column or combine. The second element of the vector can be referred to by the command

```
mydata[2]
```

and elements between 2 and 3 (i.e. elements 2 and 3) by

```
mydata[2:3]
```

Vectors can be manipulated, for instance by adding a constant to all elements, as follows.

```
myconst <- 100; mydata + myconst
```

Using the semicolon allows us to write multiple commands on a single line.

A vector x consisting of the integers between 1 and 10; 1, 2, . . . , 10; can be created by writing

```
x <- c(1:10)
```

Vectors with sequences of numbers with particular increments can be created with the seq command:

```
mydata1 <- seq(0,10,2)      # integers between 0 and 10, with increment 2
```

Read x and y

```
x<- c(2,3,1,5,4,6,5,7,6,8)
```

```
y<- c(10, 12, 14, 13, 34, 23, 12, 34, 25, 43)
```

Read two vectors

```
weight<- c(60, 72, 57,90)
```

```
height<-c(1.75, 1.80, 1.65, 1.90)
```

```
bmi<- weight/height^2
```

```
# Compute body mass index
```

(BMI)

Functions on vectors

```
length(x)           #To compute length of data in x. [1] 10
sum(x)              #To compute sum of data in x. [1] 47
sum(x^2)           [1] 265
mean(x)            #To compute mean of data in x. [1] 4.7
mean(y)            [1] 22
var(x)             #To compute variance of x. [1] 4.9
sqrt(var(x))       # To compute standard deviation of x. [1] 2.213594
sum((x-mean(x))^2) [1] 44.1
sqrt(var(x))/mean(x)*100 #To compute coefficient of variation
```

To compute summary features of data in x

```
summary(x)
Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
1.00  3.25    5.00    4.70  6.00    8.00
```

To compute summary features of data in x^2

```
summary(x^2)
Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
1.00 10.75    25.00    26.50 36.00    64.00
```

Some calculations

```
sum(weight)
mean(weight) or sum(weight)/ length(weight)
Denote by  $\bar{x}$  = mean(weight) then
sqrt(sum((weight-  $\bar{x}$ )^2)/ length(weight))
sd(weight)
cor(x,y)           #To compute correlation coefficient between x and y.
var(x,y)           #To compute covariance between x and y.
```

Slightly more complicated example ...

The rule of thumb is that the BMI for a normal weight individual should be between 20 and 25, and we want to know if our data deviate systematically from that.

- We can use a one sample t test to assess whether the 6 persons' BMI can be assumed to have mean 22.5 given that they come from a normal distribution.
- We can use function t.test

- Although you might not be knowing about t test but example is just to give some indication of what real statistical output look like

t test (see ? t.test)

```
t.test (bmi, mu=22.5)
```

One Sample t-test

data: bmi

t = -0.5093, df = 3, p-value = 0.6456

alternative hypothesis: true mean is not equal to 22.5

95 percent confidence interval: 18.29842 25.54231

sample estimates:

mean of x

21.92036

If mu is not given then t.test would use default mu=0

The p value is not small, indicating that it is not at all unlikely to get data like those observed if the mean were in fact 22.5

Packages

The base distribution already comes with some high priority add on packages, for example, boot, nlme, stats, grid, foreign, MASS, spatial etc. The packages included as default in base distribution implement standard statistical functionality, for example, linear models, classical tests, a huge collection of high level plotting functions etc. Packages not included in the base distribution can be installed directly from R prompt.

Classical Tests

To load the library of classical tests statistics available with R software use
library(stats)

#To get results of t-test for comparing population means of x and y when variances are not equal.

```
t.test(x,y)
```

To get results for usual t-test when variances are equal. If T is replaced by F then it is equal to t.test(x, y)

```
t.test(x,y,var.equal=T)
```

```
?t.test
```

```
library(stats)
```

```
x<- c(2,3,1,5,4,6,5,7,6,8)
```

```
y <- c(10, 12, 14, 13, 34, 23, 12, 34, 25, 43)
```

```
mean(x)
```

```
mean(y)
```

OVERVIEW OF R SOFTWARE

```
var(x,y)
cor(x,y)
t.test(x)
t.test(x,y)
t.test(x,y,var.equal=T)
var.test(x,y)          #To compare variances of x and y.
```

The commands **rbind** and **cbind** can be used to merge row or column vectors to matrices. Try the following:

```
x <- c(1,2,3)
y <- c(4,5,6)
A = cbind(x,y)
B = rbind(x,y)
C = t(B)
```

The last command gives the matrix transpose of B. Now type A, B or C to see what the different matrices look like.

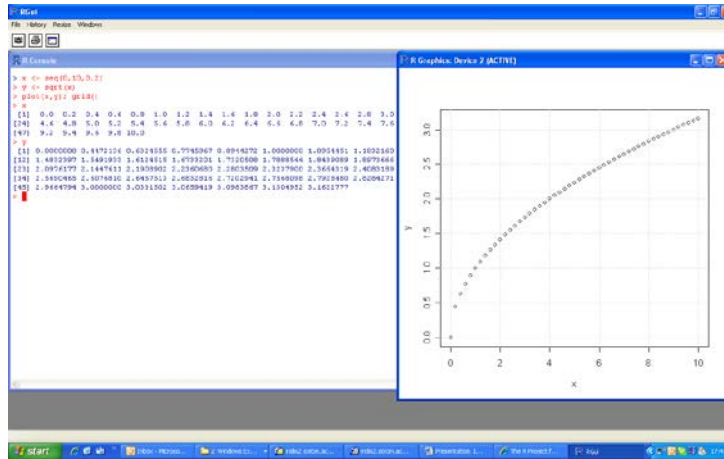
Simple graphics

Graphics- one of the most important aspects of presentation and analysis of data is generation of proper graphics

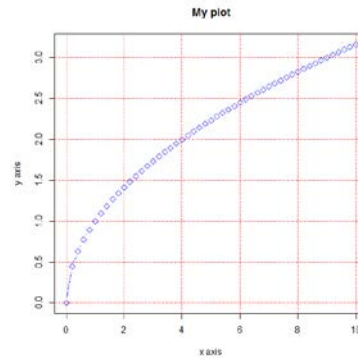
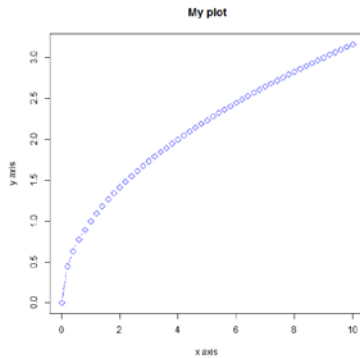
- Graphic features of a data can be viewed very effectively using R-software
- Graphs of functions can be drawn by constructing suitable vectors and using the plot command.
- **plot**: both 1D and 2D plots (see ?plot)

Scatter plots: are useful for studying dependencies between variables. Try writing

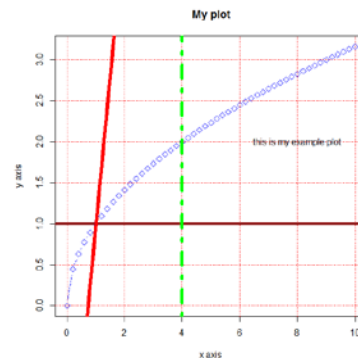
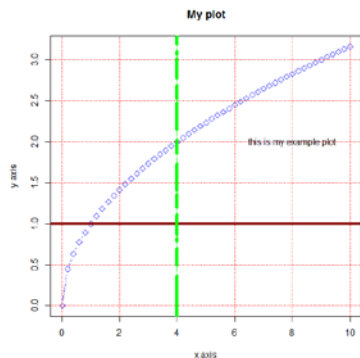
- Using the **plot command**.
x <- seq(0,10,0.2)
y <- sqrt(x)
plot(x,y); grid()
- As one might guess, the last command adds a grid to the plot.



```
plot(x,y,type="b",col="blue",lwd=1,lty=4,pch=5, main="My plot", xlab="x axis",
ylab="y axis")
grid(col="red")
```



```
plot(x,y,type="b",col="blue",lwd=1,lty=4,pch=5, main="My plot", xlab="x axis",
ylab="y axis")
grid(col="red")
text(8,2,"this is my example plot")
abline(h=1,v=4, col=c("darkred","green"), lty=c(1,4),lwd=c(4,6))
reg.lm=lm(x~y)
abline(reg.lm, col="red",lwd=6) #To add the regression line
```



Save graphics by choosing File -> Save as

Barplot

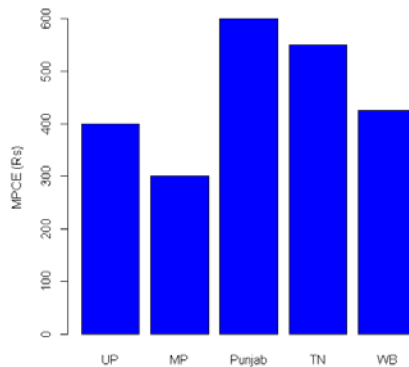
```
x1 <- c(400, 300, 600, 550, 425)
```

Suppose data in x1 are average MPCE of some states whose names are to be assigned against their value. Following commands are required:

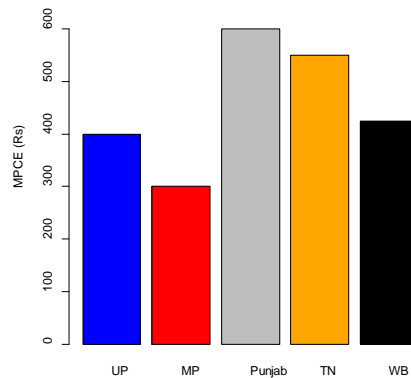
```
names(x1) <- c("UP", "MP", "Punjab", "TN", "WB")
```

To assign names of states. Double quotation mark “ ” means that names are characters not numeric.

```
barplot(x1, names=names(x1), ylab="MPCE (Rs)", col="blue")
```



```
barplot(x1, names=names(x1), ylab="MPCE (Rs)", col = c("blue", "red", "gray", "orange", "black"))
```



?barplot

Histograms

A histogram can be used to study the distribution of continuous data. Unless they are explicitly stated, R chooses the numbers of classes and class width when the command **hist** is used.

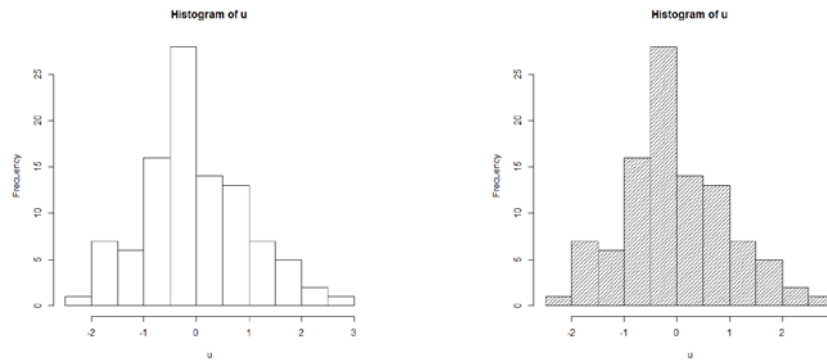
```
#generate 100 random numbers from standard normal distribution
```

```
u <- rnorm(100)
```

```
hist(u) #default histogram
```

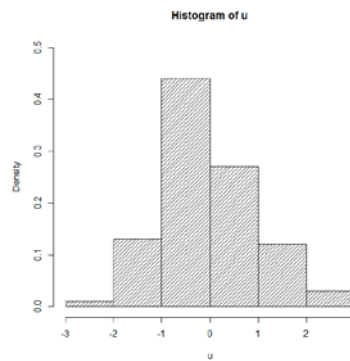
```
#with shading
```

```
hist(u, density=20)
```



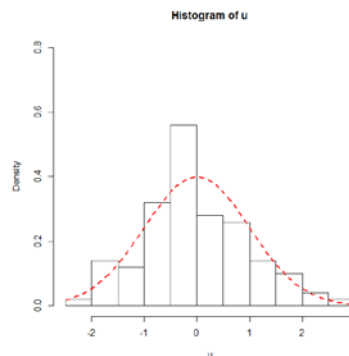
proportion, instead of frequency also specifying y-axis

```
hist(u, density=20, breaks=-3:3, ylim=c(0,.5), prob=TRUE)
```

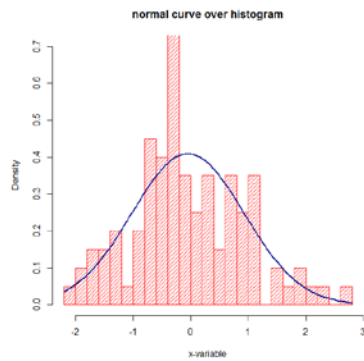


```
hist(u, freq=F, ylim = c(0,0.8))
curve(dnorm(x), col = 2, lty = 2, lwd = 2, add = TRUE)
```

The freq=F argument to hist ensures that the histogram is in terms of densities rather than absolute counts

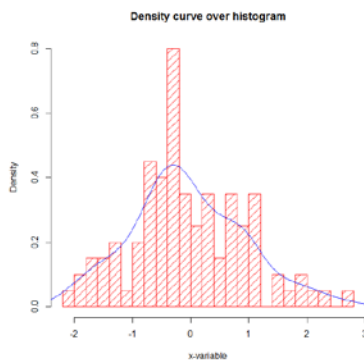


```
# overlay normal curve with x-lab and ylim
# colored normal curve
m<-mean(u) ;std<-sqrt(var(u))
hist(u, density=20, breaks=20, prob=TRUE, xlab="x-variable", col="red", ylim=c(0, 0.7), main="normal curve over histogram")
curve(dnorm(x, mean=m, sd=std), col="darkblue", lwd=2, add=TRUE)
```



```
hist(u, density=10, breaks=20, col="red", prob=TRUE, xlab="x-variable",
ylim=c(0,0.8),main="Density curve over histogram")
```

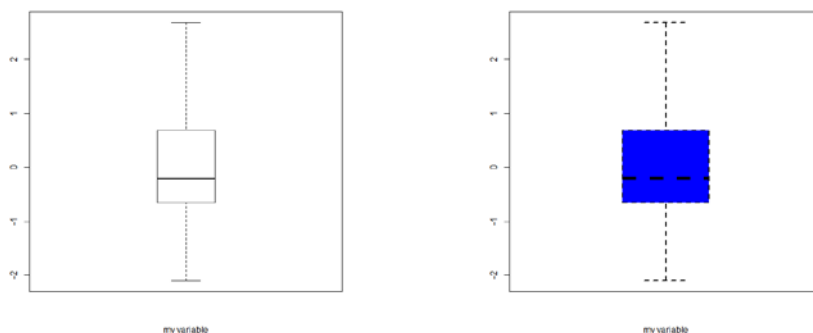
```
lines(density(u),col = "blue")
```



Boxplots

Boxplots are also a useful tool for studying data. It shows the median, quartiles and possible outliers. The R command is **boxplot**, which we use on the same variables as the histogram.

```
boxplot(u, xlab="my variable", boxwex=.4)
boxplot(u, xlab="my variable", boxwex=.6,col="blue", lty=2,lwd=2)
```

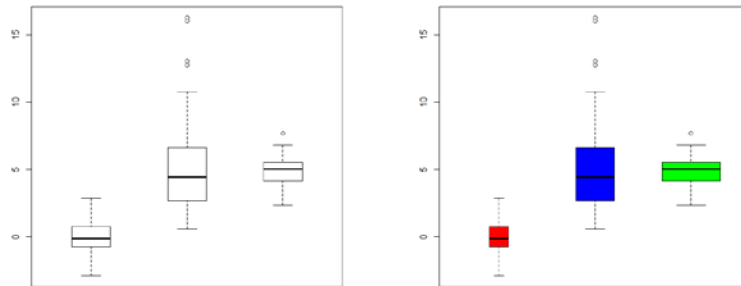


```
## we creat data: three variables
u1<- rnorm(100) # 100 random number from standard normal distribution
u2<- rchisq(100,5) # 100 random number from chisq distribution with mean 5
```

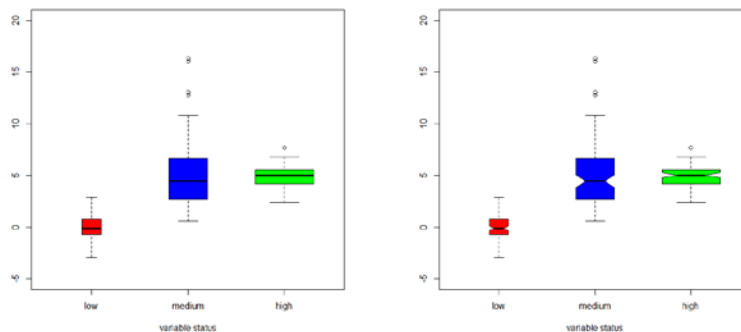


```
u3<- rnorm(100,5,1) # 100 random number from normal distribution with mean 5, sd
1
```

```
boxplot(u1,u2,u3, boxwex=.4)
boxplot(u1,u2,u3, boxwex=c(.2,.4,.6),col=c("red","blue","green"))
```



```
variablename<-c("low","medium", "high")
boxplot(u1,u2,u3,names=variablename,boxwex=c(.2,.4,.6), col=c("red","blue","green"),
ylim=c(-5, 20), xlab="variable status")
boxplot(u1,u2,u3,names=variablename,
boxwex=c(.2,.4,.6),col=c("red","blue","green"),ylim=c(-5, 20),xlab="variable status", notch =
TRUE)
```



```
## try
boxplot(u, xlab="my variable", pars = list(boxwex = 0.5, staplewex = .5, outwex = 0.5),plot =
F)
boxplot(u, xlab="my variable", pars = list(boxwex = 0.5, staplewex = .5, outwex = 0.5),plot =
T)
```

?boxplot

37.8 HANDLING DATA

Creating data frames

The command `data.frame` can be used to organize data of different kinds and to extract subsets of said data. Assume that we have data about three persons and that we store it as follows:

```
length <- c(180,175,190)
weight <- c(75,82,88);
name <- c("Anil","Ankit","Sunil")
friends <- data.frame(name,length,weight)
```

friends is now a data frame containing the data for the three persons. Data can easily be extracted:

```
my.names <- friends$name
length1 <- friends$length[1]
```

Reading data from files

It is common that data is stored in a text file and that we wish to import the data to R. We will study two cases; one with purely numerical data and one with numerical data with headers. In the file **coins.dat** data about the amount of silver in 27 silver coins from different epochs is stored. When we use the command

```
mynt1 <- read.table("coins.dat")
```

R creates a structure with two columns with headers V1 (amount of silver) and V2 (epoch). To see the properties of the tenth observations we write

```
mynt1[10,]          # [10,] means "everything on row 10"
```

If we only want to see the amount of silver in the tenth coin we write

```
mynt1$V1[10] # vector V1 (the first vector), row 10
or
mynt1[10,1] # row 10, column 1
```

Clearly, it's important to know which data is stored in which column. To make this clear one can give the columns headers when importing the data:

```
mynt2 <- read.table("coins.dat",col.names=c("Silver","Epoch"))
```

so that the strings Silver and Epoch will be used instead of V1 and V2. Let us know study the same data set, but with the data stored in a slightly different way. The file coins.txt has the same data stored, but with headers stored on the first row of the file. To import the data with headers we write

```
mynt2 <- read.table ("coins.txt", header=TRUE)
```

Try to access information about particular observations as before!

There are functions for importing data from, for instance, databases or Excel spreadsheets as well, but these are more advanced and not covered in this lecture. Usually, however, one can copy the data from the database or the spreadsheet to a **text file** and then import it, so that the **read.table** command can be used.

```
# clean out the workspace
rm(list=ls())
#List objects in workspace
ls()
```

```
#File path is relative to working directory
#Get or Set Working Directory
  getwd()
  setwd()
```

```
# e.g. setwd("C:/Documents and Settings/Myfiles")
```

Writing data from files

```
x <- matrix(1:20,ncol=5)    # generate data in matrix form
write(x, "C:/ xm.txt")
write(x, " C:/ xm.csv")
write(x, "C:/ xm.csv",sep=",")
write.table(x," C:/ xm.xls",sep="\t")
```

37.9 ANALYSIS OF A DATA SET

We will study a data set from the early 70's, with data about different cars (Cars data set). Load the data set by writing

```
data(mtcars)
```

You can read more about the data by looking at the help file:

```
?mtcars
```

mtcars	package:datasets	R Documentation
Motor Trend Car Road Tests		
Description:		
The data was extracted from the 1974 <i>Motor Trend</i> US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).		
Usage:		
mtcars		
Format:		
A data frame with 32 observations on 11 variables.		
[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (lb/1000)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors
Source:		
Henderson and Velleman (1981), Building multiple regression models interactively. <i>Biometrics</i> , *37*, 391-411.		
Examples:		
pairs(mtcars, main = "mtcars data")		
coplot(mpg ~ disp as.factor(cyl), data = mtcars, panel = panel.smooth, rows = 1)		

EXERCISE. Answer the following questions using the help file:

1. How many cars are included in the data set?
2. Which years are the models from?
3. What does the mpg value describe?

To see the entire data set, simply write

```
mtcars
```

EXERCISE. To get familiar with the data set, answer the following non-statistical questions.

1. Are there any cars that weigh more than 5000 (lb/1000)?
2. How many cylinder has the motor of the Volvo 142E?
3. Are there any cars with 5 forward gears? Do they have automatic or manual transmission?

Descriptive statistics

Data can be summarized using simple measures such as mean, median, standard deviation, maximum and minimum and so on. A summary of a few such measures for the mtcars data set is obtained by writing

```
summary(mtcars)
```

Measures can also be studied one at a time:

```
mean(mtcars$hp);          median(mtcars$hp);          quantile(mtcars$wt);
max(mtcars$mpg)
sd(mtcars$mpg)           # standard deviation
var(mtcars$mpg)          # variance
sd(mtcars$mpg)^2        # sd*sd=var?
```

The command `attach` is very useful when dealing with data frames. By writing `attach(mtcars)` the references to the variables in `mtcars` can be shortened; instead of the long references above we can write:

```
mean(hp); median(hp); quantile(wt); max(mpg)
```

The sequence of commands below plots two histograms in one window, the first being the histogram for `mpg` and the second for `wt`.

```
par(mfrow=c(1,2)); hist(mtcars$mpg); hist(mtcars$wt)
```

`par(mfrow=c(a,b))` gives a rows with `b` plots on each row. Using the parameters `freq` when calling `hist` we can plot a histogram with relative frequencies instead of frequencies. Such histograms can be viewed as estimates of the density function of the data. Read in the help file about what `hist(mtcars$mpg,freq=FALSE)` means and then see for yourself by typing the command.

```
boxplot(mtcars$mpg); x11(); boxplot(mtcars$wt)
```

The `x11` command opens a new window which the next figure will be plotted in.

```
plot(mtcars$wt,mtcars$mpg)
```

Does the slope of the cluster seem reasonable? The correlation (which measures linear dependence) can be calculated using the command `cor` (use to help file to see how). What is the correlation in this case? Does it agree with the slope?

```
cor(mtcars$wt,mtcars$mpg)
```

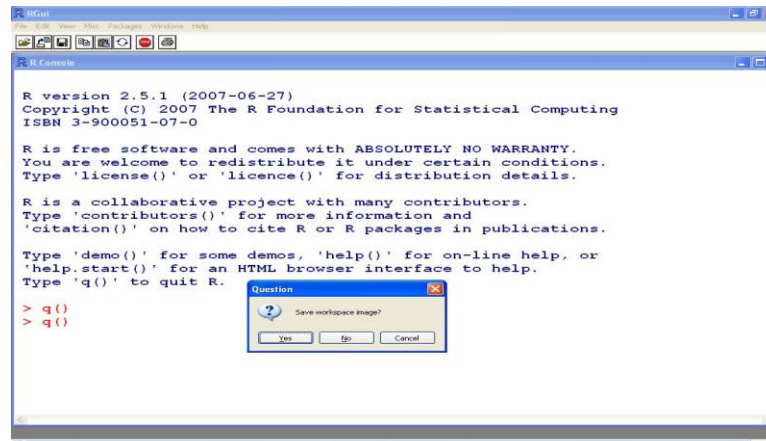
Linear regression

```
lm(mtcars$wt~mtcars$mpg)
```

Try to see help (lm)

37.10 QUITTING R

R can be closed with the command `q()`. After issuing the quit command, R asks whether to save the workspace or not:



It is usually a good idea to save the workspace, since this creates a special file that can be directly read into R, and one can commence working with the same datasets and results already generated without a need to start from the scratch again. Saved workspace is in a file called `.RData`, and all the commands given during the same R session are saved in a file called `.Rhistory`. To load the workspace into R again, one can simply double-click on the file `.Rdata`, and R should open automatically with all the data and results loaded. Note however that libraries are not loaded automatically, and these should be loaded (if needed) before commencing the work.

37.11 STRENGTHS AND WEAKNESSES OF R

Strengths

- free and open source, supported by a strong user community
- highly extensible and flexible
- implementation of modern statistical methods
- moderately flexible graphics with intelligent defaults

Weaknesses

- slow or impossible with large data sets
- non-standard programming paradigms

REFERENCES

- R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org>.
- www.r-project.org
- <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>
- <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/index.html>

ANALYSYS OF SURVEY DATA USING R

Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi-110012

Email: hchandra@iasri.res.in

38.1 Introduction

A sample survey is a process for collecting data on a sample of observations which are selected from the population of interest using a probability-based sample design. In sample surveys, certain methods are often used to improve the precision and control the costs of survey data collection. These methods introduce a complexity to the analysis, which must be accounted for in order to produce unbiased estimates and their associated levels of precision. This write up provides a brief introduction to the impact these design complexities have on the sampling variance and then summarizes survey function in **software R** to carry out analysis on sample survey data.

38.2 Complex Sample Designs

Statistical methods for estimating population parameters and their associated variances are based on assumptions about the characteristics and underlying distribution of the observations. Statistical methods in most general-purpose statistical software tacitly assume that the data meet certain assumptions. Among these assumptions are that the observations were selected independently and that each observation had the same probability of being selected. Data collected through surveys often have sampling schemes that deviate from these assumptions. For logistical reasons, samples are often clustered geographically to reduce costs of administering the survey, and it is not unusual to sample households, then subsample families and/or persons within selected households. In these situations, sample members are not selected independently, nor are their responses likely to be independently distributed.

In addition, a common survey sampling practice is to oversample certain population subgroups to ensure sufficient representation in the final sample to support separate analyses. This is particularly common for certain policy-

relevant subgroups, such as ethnic and racial minorities, the poor, the elderly, and the disabled. In this situation, sample members do not have equal probabilities of selection. Adjustments to sampling weights (the inverse of the probability of selection) to account for nonresponse, as well as other weighting adjustments (such as poststratification to known population totals), further exacerbate the disparity in the weights among sample members.

In brief, the complications in a complex survey sample result from following:

- **Stratification**- Dividing the population into relatively homogenous groups (strata) and sampling a predetermined number from each stratum will increase precision for a given sample size.
- **Clustering**- Dividing the population into groups and sampling from a random subset of these groups (eg geographical locations) will decrease precision for a given sample size but often increase precision for a given cost.
- **Unequal sampling**- Sampling small subpopulations more heavily will tend to increase precision relative to a simple random sample of the same size.
- **Finite population**- Sampling all of a population or stratum results in an estimate with no variability, and sampling a substantial fraction of a stratum results in decreased variability in comparison to a sample from an infinite population. I have described these in terms of their effect on the design of the survey.
- **Weighting** -When units are sampled with unequal probability it is necessary to give them correspondingly unequal weights in the analysis. The inverse-probability weighting has generally the same effect on point estimates as the more familiar inverse-variance weighting, but very different effects on standard errors.

Most standard statistical procedures in software packages commonly used for data analysis do not allow the analyst to take most of these properties of survey data into account unless specialized survey procedures are used. That is standard methods of statistical analysis assume that survey data arise from a *simple random sample* of the target population. Little attention is given to characteristics often associated with survey data, including missing data,

unequal probabilities of observation, stratified multistage sample designs, and measurement errors. Failure to do so can have an important impact on the results of all types of analysis, ranging from simple descriptive statistics to estimates of parameters of multivariate models.

38.3 Impact of Complex Sample Design on Sampling Variance

Because of these deviations from standard assumptions about sampling, such survey sample designs are often referred to as complex. While stratification in the sampling process can decrease the sampling variance, clustering and unequal selection probabilities generally increase the sampling variance associated with resulting estimates. Not accounting for the impact of the complex sample design can lead to an underestimate of the sampling variance associated with an estimate. So while standard software packages can generally produce an unbiased weighted survey estimate, it is quite possible to have an underestimate of the precision of such an estimate when using one of these packages to analyze survey data.

That is, analyzing a stratified sample as if it were a simple random sample will *overestimate* the standard errors, analyzing a cluster sample as if it were a simple random sample will usually *underestimate* the standard errors, as will analyzing an unequal probability sample as if it were a simple random sample.

The magnitude of this effect on the variance is commonly measured by what is known as the design effect. The design effect is the sampling variance of an estimate, accounting for the complex sample design, divided by the sampling variance of the same estimate, assuming a sample of equal size had been selected as a simple random sample. A design effect of unity indicates that the design had no impact on the variance of the estimate. A design effect greater than one indicates that the design has increased the variance, and a design effect less than one indicates that the design actually decreased the variance of the estimate. The design effect can be used to determine the effective sample size, simply by dividing the nominal sample size by the design effect. The effective sample size gives the number of observations that would yield an equivalent level of precision from an independent and identically distributed (iid) sample.

38.4 Software Packages R for Survey data analysis

Several packages are available to the public designed specifically for use with sample survey data. However, this lecture will discuss only **Software R** for

analyzing complex surveys. The survey functions for R were contributed by Thomas Lumley, Department of Biostatistics, University of Washington, USA.

Types of designs that can be accommodated

- Designs incorporating stratification, clustering, and possibly multistage sampling, allowing unequal sampling probabilities or weights.
- Simple two-phase designs
- Multiply-imputed data

Types of estimands and statistical analyses that can be done in R

- Mean, Totals, Quantiles, Variance, Tables, Ratios,
- Generalised linear models (e.g. linear regression, logistic regression etc.)
- Proportional hazards models
- Proportional odds and other cumulative link models
- Survival curves
- Post-stratification, raking, and calibration
- Tests of association in two-way tables

Restrictions on number of variables or observations: Only those due to limitations of available memory or disk capacity.

Variance estimation methods: Taylor series linearization and replication weighting.

Platforms on which the software can be run

- Intel computers with Windows 2000 or better
- Mac OS X 10.3 or later
- Linux
- Most Unix systems.

Pricing and terms: Free download. R is updated about twice per year and the survey package is updated as needed. For information on R see <http://www.r-project.org/>.

38.5 Implementation of survey package in R

First install survey package. The command **svydesign** in **library (survey)** is used for survey data analysis in R, described as below.

Svydesign (id=~1, strata=~stype, weights=~pw, data=apistrat, fpc=~fpc)

where different arguments of function **svydesign()** are

ids	Formula or data frame specifying cluster ids from largest level to smallest level, ~0 or ~1 is a formula for no clusters.
probs	Formula or data frame specifying cluster sampling probabilities
strata	Formula or vector specifying strata, use NULL for no strata
variables	Formula or data frame specifying the variables measured in the survey. If NULL, the data argument is used.
fpc	Finite population correction
weights	Formula or vector specifying sampling weights as an alternative to prob
data	Data frame to look up variables in the formula arguments
nest	If TRUE, relabel cluster ids to enforce nesting within strata
check.strata	If TRUE, check that clusters are nested in strata

The **svydesign** object combines a data frame and all the survey design information needed to analyse it. These objects are used by the survey modelling and summary functions. The **id** argument is **always required**, the strata, fpc, weights and probs arguments are **optional**. If these variables are specified they must **not** have any missing values.

By default, svydesign assumes that all PSUs, even those in different strata, have a unique value of the id variable. This allows some data errors to be detected. If your PSUs reuse the same identifiers across strata then set nest=TRUE.

The *finite population correction* (fpc) is used to reduce the variance when a substantial fraction of the total population of interest has been sampled. It may not be appropriate if the target of inference is the process generating the data rather than the statistics of a particular finite population.

The finite population correction can be specified either as the total population size in each stratum or as the fraction of the total population that has been

sampled. In either case the relevant population size is the sampling units. That is, sampling 100 units from a population stratum of size 500 can be specified as 500 or as $100/500=0.2$.

If population sizes are specified but not sampling probabilities or weights, the sampling probabilities will be computed from the population sizes assuming simple random sampling within strata.

For multistage sampling the `id` argument should specify a formula with the cluster identifiers at each stage. If subsequent stages are stratified strata should also be specified as a formula with stratum identifiers at each stage. The population size for each level of sampling should also be specified in `fpc`. If `fpc` is not specified then sampling is assumed to be with replacement at the top level and only the first stage of cluster is used in computing variances. If `fpc` is specified but for fewer stages than `id`, sampling is assumed to be complete for subsequent stages. The variance calculations for multistage sampling assume simple or stratified random sampling within clusters at each stage except possibly the last.

If the strata with one only PSU are not self-representing (or they are, but `svydesign` cannot tell based on `fpc`) then the handling of these strata for variance computation is determined by `options("survey.lonely.psu")`.

Example -Read the `api` data - Academic Performance Index (`api`) is computed for all California schools. The full population data in **`apipop`** are a data frame with 6194 observations on the 37 variables. Read **`apipop`** data available in `survey` package

```
data(api)           #This load the api population data apipop
dim(apipop)       # Shows the dimension of the data set
```

The details of 37 variables are

- | | | |
|-----|-----------------------|---|
| 1. | <code>cds</code> | |
| 24. | <code>acs.k3</code> | average class size years K-3 |
| 25. | <code>acs.46</code> | average class size years 4-6 |
| 26. | <code>acs.core</code> | Number of core academic courses |
| 27. | <code>pct.resp</code> | percent where parental education level is known |

28.	not.hsg	percent parents not high-school graduates
29.	hsg	percent parents who are high-school graduates
30.	some.col	percent parents with some college
31.	col.grad	percent parents with college degree
32.	grad.sch	percent parents with postgraduate education
33.	avg.ed	average parental education level
34.	full	percent fully qualified teachers
35.	emer	percent teachers with emergency qualifications
36.	enroll	number of students enrolled Unique identifier
2.	stype	Elementary/Middle/High School
3.	name	School name (15 characters)
4.	sname	School name (40 characters)
5.	snum	School number
6.	dname	District name
7.	dnum	District number
8.	cname	County name
9.	cnum	County number
10.	flag	reason for missing data
11.	pctest	percentage of students tested
12.	api00	API in 2000
13.	api99	API in 1999
14.	target	target for change in API
15.	growth	Change in API
16.	sch.wide	Met school-wide growth target?
17.	comp.imp	Met Comparable Improvement target
18.	both	Met both targets
19.	awards	Eligible for awards program
20.	meals	Percentage of students eligible for subsidized meals
21.	ell	`English Language Learners' (percent)
22.	yr.rnd	Year-round school
23.	mobility	percent of students for whom this is the first year at the school
37.	api.stu	number of students tested.

Type **summary(apiipop)** and see what you get?

The other data sets contain additional variables `pw` for sampling weights and `fpc` to compute finite population corrections to variance. **apiipop** is the entire population, **apiclus1** is a cluster sample of school districts, **apistrat** is a sample stratified by `stype`, and **apiclus2** is a two-stage cluster sample of schools within districts. The sampling weights in **apiclus1** are incorrect (the weight should be 757/15) but are as obtained from UCLA. Data were obtained from the survey sampling help pages of UCLA Academic Technology Services, at

http://www.ats.ucla.edu/stat/stata/Library/svy_survey.htm.

The API program and original data files are at <http://api.cde.ca.gov/>

```
# api00 is API in 2000
```

```
mean (apipop$api00)
```

```
[1] 664.7126
```

```
# enroll is number of students enrolled
```

```
sum (apipop$enroll, na.rm=TRUE)
```

```
[1] 3811472
```

Here na.rm=TRUE means –logical, Should missing values be removed?

Specifying a complex survey design – use function svydesign ()

[i] Stratified sample

Here we use data set apistrat, see dim(apistrat), c(apistrat[1,]), attach(apistrat) commands etc.

```
dstrat<- svydesign(id=~1,strata=~stype, weights=~pw, data=apistrat, fpc=~fpc)
```

```
summary(dstrat)
```

Stratified Independent Sampling design

```
svydesign(id = ~1, strata = ~stype, weights = ~pw, data = apistrat, fpc = ~fpc)
```

Probabilities:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02262	0.02262	0.03587	0.04014	0.05339	0.06623

Stratum Sizes:

	E	H	M
obs	100	50	50
design.PSU	100	50	50
actual.PSU	100	50	50

Population stratum sizes (PSUs):

E	M	H
4421	1018	755

Data variables:

```
[1] "cds" "stype" "name" "sname" "snum" "dname"
[7] "dnum" "cname" "cnum" "flag" "pctest" "api00"
[13] "api99" "target" "growth" "sch.wide" "comp.imp" "both"
[19] "awards" "meals" "ell" "yr.rnd" "mobility" "acs.k3"
[25] "acs.46" "acs.core" "pct.resp" "not.hsg" "hsg" "some.col"
[31] "col.grad" "grad.sch" "avg.ed" "full" "emer" "enroll"
[37] "api.stu" "pw" "fpc"
```

Some functions used to compute means, variances, ratios and totals for data from complex surveys are as follows.

svymean () and **svytotal ()** functions are use to extract mean and total estimate along with their standard error, specified as below.

```
svymean(x, design, na.rm=FALSE,deff=FALSE,...)
```

```
svytotal(x, design, na.rm=FALSE,deff=FALSE,...)
```

Arguments

x	A formula, vector or matrix
design	survey.design or svyrep.design object
na.rm	Should cases with missing values be dropped?
rho	parameter for Fay's variance estimator in a BRR design
return.replicates	Return the replicate means?
deff	Return the design effect
object	The result of one of the other survey summary functions
quietly	Don't warn when there is no design effect computed
estimate.only	Don't compute standard errors (useful when svyvar is used to estimate the design effect)
names	vector of character strings

Also see

```
Svyvar (x, design, na.rm=FALSE,...)
```

```
svyratio (x, design, na.rm=FALSE,...)
```

```
svyquantile (x, design, na.rm=FALSE,...)
```

svymean(~api00, dstrat)

	mean	SE
api00	662.29	9.4089

svymean(~api00, dstrat, deff=TRUE)

	mean	SE	DEff
api00	662.29	9.4089	1.2045

svytotal(~enroll, dstrat, na.rm=TRUE)

	total	SE
enroll	3687178	114642

#stratified sample, Now try these code for your self

```

dstrat<-svydesign(id=~1, strata=~stype, weights=~pw, data=apistrat,
fpc=~fpc)
summary(dstrat)
svymean(~api00, dstrat)
svyquantile(~api00, dstrat, c(.25,.5,.75))
svyvar(~api00, dstrat)
svytotal(~enroll, dstrat)
svyratio(~api.stu, ~enroll, dstrat)
# coefficients of variation
cv(svytotal(~enroll,dstrat))

```

[ii] One-stage cluster sample

```

dclus1<-svydesign(id=~dnum, weights=~pw, data=apiclus1, fpc=~fpc)
summary(dclus1)
svymean(~api00, dclus1, deff=TRUE)
svymean(~factor(stype),dclus1)
svymean(~interaction(stype, comp.imp), dclus1)
svyquantile(~api00, dclus1, c(.25,.5,.75))
svyvar(~api00, dclus1)
svytotal(~enroll, dclus1, deff=TRUE)
svyratio(~api.stu, ~enroll, dclus1)

```

summary(dclus1)

1 - level Cluster Sampling design

With (15) clusters.

svydesign(id = ~dnum, weights = ~pw, data = apiclus1, fpc = ~fpc)

Probabilities:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02954	0.02954	0.02954	0.02954	0.02954	0.02954

Population size (PSUs): 757

Data variables:

```

[1] "cds" "stype" "name" "sname" "snum" "dname"
[7] "dnum" "cname" "cnum" "flag" "pctest" "api00"
[13] "api99" "target" "growth" "sch.wide" "comp.imp" "both"
[19] "awards" "meals" "ell" "yr.rnd" "mobility" "acs.k3"
[25] "acs.46" "acs.core" "pct.resp" "not.hsg" "hsg" "some.col"
[31] "col.grad" "grad.sch" "avg.ed" "full" "emer" "enroll"
[37] "api.stu" "fpc" "pw"

```

svymean(~api00, dclus1)

	mean	SE
api00	644.17	23.542

svytotal(~enroll, dclus1, na.rm=TRUE)

	total	SE
enroll	3404940	932235

[iii] Two-stage cluster sample

**dclus2<-svydesign(id=~dnum+snum, fpc=~fpc1+fpc2,
data=apiclus2)
summary(dclus2)**

```

2 - level Cluster Sampling design
With (40, 126) clusters.
svydesign(id = ~dnum + snum, fpc = ~fpc1 + fpc2, data = apiclus2)
Probabilities:
      Min.      1st Qu.  Median      Mean      3rd Qu.  Max.
0.003669  0.037740 0.052840  0.042390  0.052840 0.052840

Population size (PSUs): 757
Data variables:
[1] "cds"  "stype" "name"  "sname" "snum"  "dname"
[7] "dnum"  "cname" "cnum"  "flag"  "pctest" "api00"
[13] "api99" "target" "growth" "sch.wide" "comp.imp" "both"
[19] "awards" "meals" "ell"    "yr.rnd" "mobility" "acs.k3"
[25] "acs.46" "acs.core" "pct.resp" "not.hsg" "hsg"    "some.col"
[31] "col.grad" "grad.sch" "avg.ed" "full"    "emer"   "enroll"
[37] "api.stu" "pw"      "fpc1"  "fpc2"

```

svymeans(~api00, dclus2)

	mean	SE
api00	670.81	30.099

svytotal(~enroll, dclus2, na.rm=TRUE)

	total	SE
enroll	2639273	799638

[iv] Two-stage `with replacement'

```
dclus2wr<-svydesign(id=~dnum+snum, weights=~pw,
data=apiclus2)
summary(dclus2wr)
```

2 - level Cluster Sampling design (with replacement)

With (40, 126) clusters.

```
svydesign(id = ~dnum + snum, weights = ~pw, data = apiclus2)
```

Probabilities:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003669	0.037740	0.052840	0.042390	0.052840	0.052840

Data variables:

```
[1] "cds" "stype" "name" "sname" "snum" "dname"
[7] "dnum" "cname" "cnum" "flag" "pcttest" "api00"
[13] "api99" "target" "growth" "sch.wide" "comp.imp" "both"
[19] "awards" "meals" "ell" "yr.rnd" "mobility" "acs.k3"
[25] "acs.46" "acs.core" "pct.resp" "not.hsg" "hsg" "some.col"
[31] "col.grad" "grad.sch" "avg.ed" "full" "emer" "enroll"
[37] "api.stu" "pw" "fpc1" "fpc2"
```

svymean(~api00, dclus2wr)

	mean	SE
api00	670.81	30.712

svytotal(~enroll, dclus2wr, na.rm=TRUE)

	total	SE
enroll	2639273	820261

Reference

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley Series in Survey Methodology

OVERVIEW OF SSDA 2.0 SOFTWARE

S.B. Lal

Indian Agricultural Statistics Research Institute, New Delhi-110012

39.1 Introduction

Specialised software is used when analysis required is descriptive or analytical for survey design that includes stratification and selection at different stages. But standard statistical software generally does not take into account four common characteristics of survey data (i) selection of units with unequal probability, (ii) clustering of observations, (iii) stratification and (iv) non-response and other adjustments. By ignoring these aspects, standard packages generally under estimate the variance of point estimate. Some of the software developed for survey data analysis for personal computer can be put into following categories: DOS based packages: These are PC CARP, CENVAR, CLUSTERS etc.

- (i) Standard packages: Generally costlier such as SUDAAN, STATA, WesVarPC etc.
- (ii) Open source software: Such as “R”. The full source code is free.
- (iii) Need of expertise to use: Such as SAS, it is costlier, vast, and very extensive but at the same time needs expertise to make use of it.

Most of these packages have more extensive features than only estimation for complex sample survey data, incorporating software for processing and managing survey data as well. Some of the distinguished features of the software that have been reviewed are discussed below:

SDAP: Priyanka (2003) developed window-based software, SDAP using Visual Basic 6.0 at IASRI as a part of her research work for M.Sc. (CA). This software is restricted to the analysis of data obtained through two stage-stratified sampling with 250 records only.

SSDA 1.0: Mahajan *et al.* (2008) developed an indigenous software for survey data analysis. It is a windows-based software developed using .NET programming language C# which has full GUI based features. The data handling has spreadsheet like feature and also importing from various file types such as excel, text and MS Access database. The results are shown using the crystal reports. It covers sampling schemes namely systematic, Simple Random Sampling (SRS), Probability Proportional to Size (PPS), stratified, cluster, two stage and stratified two stage. It also has most commonly used imputation methods for filling the missing observations. These are mean substitution, mean of neighboring units and zero substitution.

Sample: A sample is a subject chosen from a population for investigation. A random sample is one chosen by a method involving an unpredictable component.

Sampling Frame: Sampling frame is the set of individuals with the property that every single element can be identified and any member of this set can be included in our sample.

Probability sampling: A probability sampling method is any method of sampling that utilizes some form of *random selection*. In *probability sampling*, the sample is selected in such a way that each unit within the population or universe has a known chance of being selected. It is this concept of "known chance" that allows for the statistical projection of characteristics based on the *sample* to the *population*.

Summary Statistics: The purpose of this module is to provide the summary statistics of the sampled data. The summary statistics include the following measures:

- Mean, median and mode
- Minimum and maximum value
- 1st and 3rd quartile
- Standard deviation

Data Imputation: There are three types of imputations being carried in SSDA2. It takes care of stratum and stages up to three stage level. Any lower level design from stratified three stage sampling is also taken into consideration while imputing missing values using any of the following available methods. These are:

- (i) **Zero Imputation:** All the missing values are filled with value zero.
- (ii) **Mean Imputation:** The missing values are filled with the mean value of the sample/stratum as the case may be.
- (iii) **Mean of Neighboring Units:** This method fills the gaps by calculating mean of neighboring units. If it is first or last value within the available data then it is filled with the one of the available values. In case of more than one continuous missing value it goes for next available value for calculating mean.

Sample Selection: The sample selection procedure incorporated in the software includes standard methodology described in Cochran (2002). The selection methods which have been taken for inclusion in the software are as follows:

- Equal probability selection: Under this, the selection methods included are
 - Simple Random Sampling With Replacement (SRSWR)
 - Simple Random Sampling Without Replacement (SRSWOR)
 - Systematic Random Sampling
- Unequal probability selection: Under this method of sample selection Probability Proportional to Size (PPSWR) has been implemented. The selection method includes Cumulative Frequency method.

Sample Selection Procedure: Selecting a sample from population involves two tasks as follows:

- (i) How to select the elements?
- (ii) How to estimate the population characteristics – from the sampling units?

We employ some randomization process for sample selection so that there is no preferential treatment in selection which may introduce any selection bias.

Simple Random Sampling (SRS): It is a simplest sample design. Each element has an equal probability of being selected from a list of all population units (sample of n from N population). SRS are EPSEM samples:

- (i) Equal Probability of Selection Method
- (ii) Equal Probability of Selection [of Element] Method

Sample Selection with Systematic sampling: As we know that the systematic sampling is a statistical method involving the selection of elements from an ordered sampling frame. The most common form of systematic sampling is an equal-probability method, in which every k -th element in the frame is selected, where k , the sampling interval. Only the first unit is selected at random. To select a systematic sample of n units, the first unit is selected with a random start r from 1 to k sample, where $k=N/n$ sample intervals and after the selection of first sample, every k -th unit is included where $1 \leq r \leq k$. (n = sample size, and N = population size). Using this procedure each element in the population has a known and equal probability of selection. In case of the population is not evenly divisible, the decimal value is taken and rounded to the nearest integer. This gives an equal chance of every unit to be selected in the population.

Sample Selection with Probability Proportional to Size (PPS): On the other hand, in this technique it makes use of auxiliary information present on the frame that could make the design of the sample more efficient. Probability sampling requires that each member of the survey population have a chance of being included in the sample, but it does not require that this chance be the same for everyone. If there is information available on the frame about the size of each unit and those units vary in size, this information can be used in the sample selection in order to increase the efficiency. This is known as sampling with probability proportional to size (PPS). With this method, the bigger the size of the unit, the higher the chance it has of being included in the sample. For this method to bring increased efficiency, the measure of size needs to be accurate.

The auxiliary variable available on the frame is used for selecting the units by cumulative frequency method. A random number selected between the minimum and maximum value of the auxiliary variable. Looking at the cumulative frequencies, it is found out that which value is greater than the selected random value and less than the next cumulative frequency. The later cumulative frequency is the right choice and the corresponding unit is selected. The process is repeated for n number of times where n is the number of units to be selected from the population. This procedure of sample selection can be summaries in following steps:

- i. Write down cumulative totals for the sizes X_i , $i=1,2,\dots,N$.
- ii. Chose a random number r , such that $1 \leq r \leq X$
- iii. Select the i^{th} population unit if $T_{i-1} < r \leq T_r$ where $T_{i-1} = X_1 + X_2 + \dots + X_{i-1}$ and $T = T_{i-1} + X_i$.
- iv. For selecting a sample of n units with PPS with replacement, repeat the method n times.

Survey Weight Calculation

The algorithms used in the development of this software have been adopted from the PC-CARP. In particular, we use the same approach for the estimation as has been implemented in PC-CARP. The detailed descriptions are given as follows. SSDA2 is meant for analysis of data up to three stage stratified sampling. Taking this in view, the software has been developed. For this analysis, the basic input required for the program is the stratum and stages identification, weight and data vector for a set of observations. Analysis can be performed with data containing less identification too.

It is noteworthy that the calculation of survey weights is not an essential part of estimation if the same has been supplied by the user i.e. if survey weights are available in the data. Otherwise, user can calculate the survey weights but this will require some additional information. With same notations as illustrated above, algorithm of the survey weights calculation is described here. We denote Primary Stage Unit as PSU, Secondary Stage Unit as SSU and Third Stage Unit as TSU. The survey weight for unit i, j, k, s is obtained as

$$w_{ijks} = \left(\frac{1}{b_i p_{ij}} \right) \left(\frac{1}{v_{ij} p_{ijk}} \right) \left(\frac{1}{h_{ijk} p_{ijks}} \right)$$

Where,

L = Number of stratum

b_i = Selected number of PSU's in stratum $i = 1, 2, 3, \dots, L$

v_{ij} = Selected number of SSU's in PSU j and stratum i ($j = 1, 2, 3, \dots, b_i$)

h_{ijk} = Selected number of TSU's in SSU k , PSU j and stratum i , ($k = 1, 2, 3, \dots, v_{ij}$)

p_{ij} = Probability of selection of PSU j in stratum i

p_{ijk} = Probability of selection of SSU k in PSU j in stratum i

p_{ijks} = Probability of selection of TSU s in SSU k in PSU j in stratum i

Data Analysis Algorithm

We now denote Y_{ijks} as the observation for stratum i ($i=1, \dots, L$), PSUs j ($j=1, \dots, b_i$), SSU's k ($k=1, \dots, v_{ij}$) and TSU's s ($s=1, \dots, h_{ijk}$). Let w_{ijks} denote the survey weights associated with unit Y_{ijks} . The estimate of population total of Y is given by

$$\hat{Y} = \sum_{i=1}^L \sum_{j=1}^{b_i} \sum_{k=1}^{v_{ij}} \sum_{s=1}^{h_{ijk}} w_{ijks} Y_{ijks}$$

The estimate of variance of \hat{Y} is given by

$$\hat{V}ar(\hat{Y}) = \sum_{i=1}^L (n_i - 1)^{-1} n_i (1 - f_i) \sum_{j=1}^{b_i} (d_{ijk} - \bar{d}_{i..})^2 (d_{ijk} - \bar{d}_{i..}).$$

Where,

$$d_{ijks} = \sum_{k=1}^{v_{ij}} w_{ijks} Y_{ijks}$$

$$\bar{d}_{i..} = n_i^{-1} \sum_j \sum_k \sum_s w_{ijks} y_{ijks}$$

f_i = sampling fraction (or rate) for the i -th stratum,

n_i = number of sampling units in the i -th stratum, and

N_i = population size for stratum i .

Note that in case of SRSWR, the sampling fraction f_i is negligible. Therefore, in the variance expression given above $(1 - f_i)$ has been set as 1. As usual, the estimate of mean and its variance are calculated from the estimates of total and its variance.

$$\hat{\bar{Y}} = N^{-1} \left(\sum_{i=1}^L \sum_{j=1}^{b_i} \sum_{k=1}^{v_{ij}} \sum_{s=1}^{h_{ijk}} w_{ijks} Y_{ijks} \right), \text{ and}$$

$$\hat{V}ar(\hat{\bar{Y}}) = N^{-2} \hat{V}ar(\hat{Y})$$

Where N = population size

39.2 Describing SSDA2.0

SSDA2 Home Page:

The home page of SSDA 2.0 is available at <http://nabg.iasri.res.in/ssda2web>. It contains the links for various important sections on the left hand side of the browser window. The rightmost part of the window has a text boxes for entering login id and password. “New User Sign Up” is a hyperlink provided for crating user profile with login id and password. It also has links for “Forgot Password”, which is very useful when a user forgets his password. By clicking on this link, the login id entered by the user is searched in the profile database and if found, the email address in the user’s profile is picked up to send the password. Login button also checks for valid entry and verifies the password entered by the user with the password stored in the database.

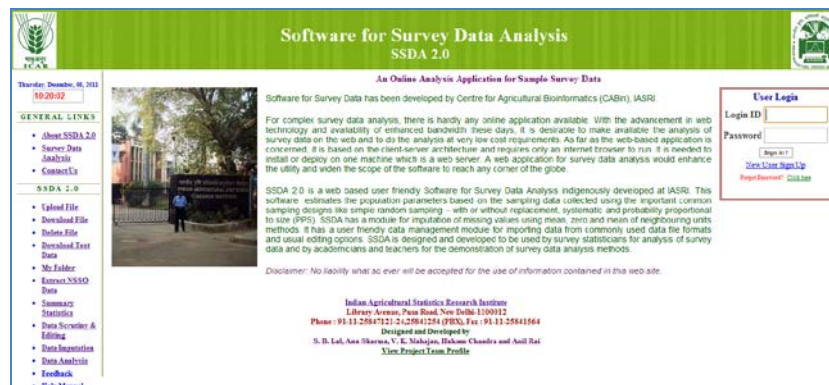


Fig. 1: Home Page

About SSDA 2.0: This link provides brief information about SSDA 2.0 software.

Survey Data Analysis: The survey data analysis concept has been presented in this section. People can have an idea about sampling methodology by clicking on this link.

User Login

Login ID

Password

[New User Sign Up](#)

Contact Us: This link shows contact information of the persons associated with this project. The detailed information can be also be viewed by clicking on their names. A user login is required to view the detailed profile.

Fig. 2: Login Window

SSDA2.0: This is the section where most of the links have been provided for carrying out the data analysis. The links provided under this heading includes file management (My folder, upload, download, delete a data file), extracting large sample survey data (includes data extraction program from collected data by 61st round of National Sample Survey Organization), data imputation for filling the gaps in the collected data, data analysis and results.

For carrying out the analysis process a user registration is mandatory. User registration can be done by clicking on the “New User Sign Up” link given on the home page. After registration, the user is verified by the webmaster and if approved, an email is sent to the user informing him about the approval.

To start using the software SSDA2.0, user logs in to the system as given below:

1. Enter login ID
2. Enter Password

Click on “Sign in” button. For new user, click on “New User Sign Up” (Fig. 2)

After login, you can start using SSDA2.0 for analyzing your survey data. The available functions displayed on left side of home page (Fig. 3) can be utilized for this.

How to Upload a File

1. Click on “Upload File” option given on the left side(Fig. 4). A window will appear which allows uploading data files with extensions “.txt” (text file), “.xls” and “.xlsx” (MS Excel File).

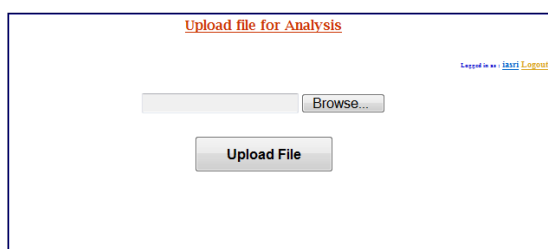


Fig. 3: File Upload

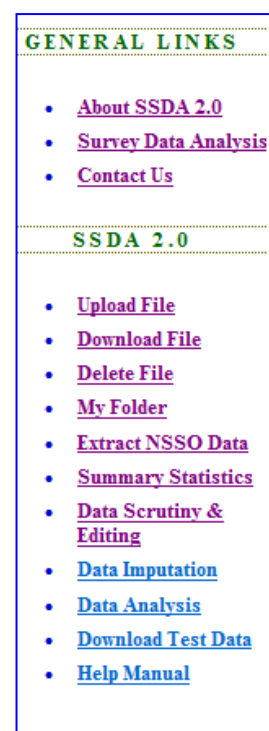


Fig. 4 General Links

2. Click on “Browse” button.
3. Select file.
4. Click on “Upload File” button.
5. Uploaded file will be saved in user’s folder of SSDA2 server.

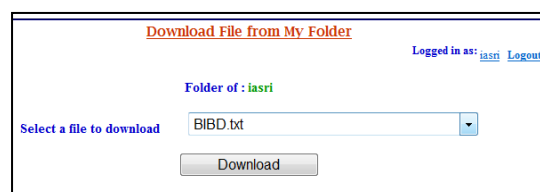


Fig. 5: Download File

How to Download a File

1. Click on “Download File” option given on left hand side of the browser.
2. A window will appear(Fig. 5)where a user can select the file for downloading.
3. Click on “Download” button.
4. Either open or save the file.

How to Delete a File

1. To delete a file from your folder, click on “Delete File” option given on left side bar.
2. Select the file (Fig. 6).
3. Click on “Confirm Delete” button.

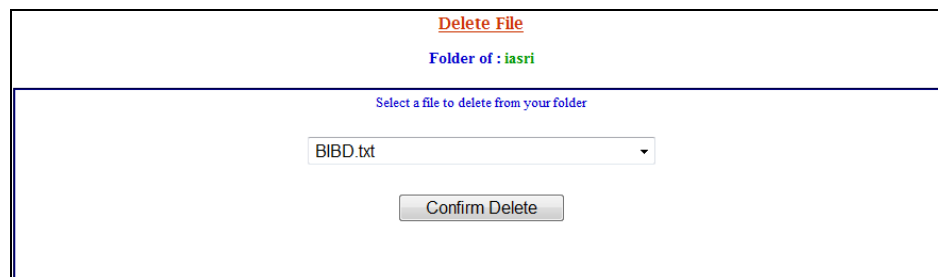


Fig. 6: Delete File

My Folder: “My Folder” shows all files saved in user’s folder (Fig. 7).

Welcome iasri My Folder Logged in as iasri Logout

Current Folder: iasri			
Name	Type	Size(KB)	Date Modified
BIBD.bt	Text Document	0	10/22/2011 12:28 PM
BIBD.xls	Excel file	13	10/22/2011 3:23 PM
Book1.xlsx	Excel file (xlsx)	30	11/13/2011 11:36 AM
clusteruneq_276.bt	Text Document	0	3/4/2008 3:04 PM
clusteruneq_276_copy.bt	Text Document	0	3/4/2008 3:05 PM
iasri (2).xls	Excel file	24	11/11/2011 2:44 PM
iasri.xls	Excel file	25	11/13/2011 6:46 PM
iasri46.xls	Excel file	22	11/7/2011 12:13 AM
iasri47.xls	Excel file	24	11/7/2011 1:02 AM
Result.xls	Excel file	15	11/16/2011 8:01 PM
StrOneStage.xls	Excel file	22	11/11/2011 2:50 PM
strTwoStage.xls	Excel file	23	11/11/2011 2:53 PM

Fig. 7: My Folder

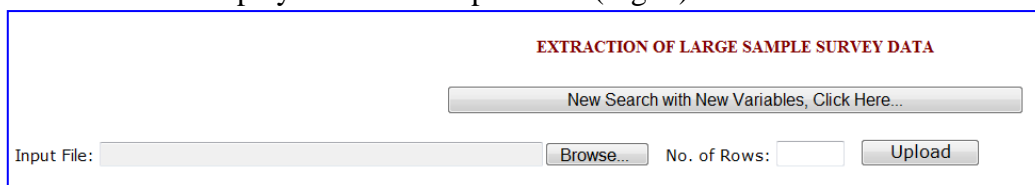
Extract NSSO Data

It is basically a web solution created in order to cater to the need of converting data collected by National Sample Survey Organization. This data is in ASCII format with meta data for defining entries in it. The data conversion needs to be done for doing the analysis. This link does the conversion of NSSO data to a format which can be utilized for doing the analysis available in SSDA2.0.

This web solution not only helps in reading up these files but also help in performing search in it.

To start with the web application steps are as follows:

1. First browse the text file on which you need to do the processing, specify the number of rows to be shown on the screen and click Upload. The uploaded file contents will be displayed in the area provided (Fig. 8).



EXTRACTION OF LARGE SAMPLE SURVEY DATA

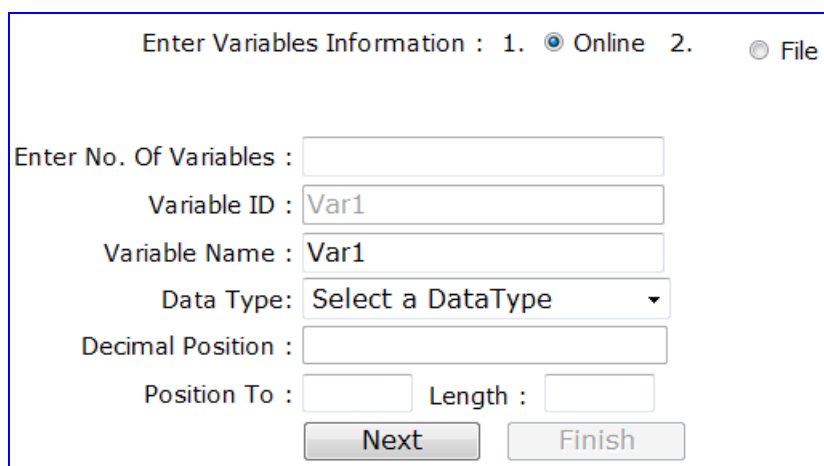
New Search with New Variables, Click Here...

Input File: Browse... No. of Rows: Upload

Fig. 8 Extraction of NSSO Data

2. Now to enter the variable information, select either online method or File Method.
 - a. Select the Online method to define variables as per requirement. (Fig. 9)
 - i. Enter the Number of variable you want to define.
 - ii. Var. ID – Default, it's Var1, Var2 etc. upto the number of variables defined.
 - iii. Variable Name - Default is Var1...but user can rename it.
 - iv. Datatype - Drop downlist of datatype as Long Integer, Long Double and Text.
 - v. Decimal Position - Default disabled, Enabled only when datatype chosen is Double.
 - vi. Position To - Starting position of the Variable in the data file.
 - vii. Length - Length of the variable in the data file.

Please click on “Next” button in case variable declaration number is greater than 1 otherwise click on “Finish”.



Enter Variables Information : 1. Online 2. File

Enter No. Of Variables :

Variable ID :

Variable Name :

Data Type:

Decimal Position :

Position To : Length :

Next Finish

Fig. 9: Data Extraction from NSSO

If you have the variables defined in the Excel sheet then select the File option with the specify file format that must contain the columns namely Variable ID, Variable Name, Starting Position and Length. (Fig. 10)

- i. Browse the Excel file of specified format from a location.
- ii. Select the Default data type for all the variables in the list.
- iii. Click on Show button to display the variables list in the grid same as in case from online submission. Any changes regarding Variable Name, Data type, Position, Length and Decimal value can be edited here. [The column names in the excel sheet must not contain comma(,)]

Fig. 10: Variable Information (Data Extraction from NSSO)

3. Click on Show Result button in order to display the Exact searched data in the area provided.

2428	1	98	0	1	0	1	1	5	3	9	1	212	2428	0	0	0	
003	2226	0															
	35580	61	010	1	1	201	01	01	01	1	1	2010	1	1	01	03	00000
	1	01	1	1	1	3	2	2	2	2	2	0	00604375		00628238		
004	35580	61	010	1	1	201	01	01	01	1	1	2010	1	1	01	04	00001
	1	1	5	7	2	0	6	0	0	2	0	0000	00060				
005	35580	61	010	1	1	201	01	01	01	1	1	2010	1	1	01	04	00002
	2	2	5	5	2	0	1	0	0	2	0	0000	00060				
006	35580	61	010	1	1	201	01	01	01	1	1	2010	1	1	01	04	00003
	3	1	3	0	2	0	6	0	0	2	0	0000	00060				
007	35580	61	010	1	1	201	01	01	01	1	1	2010	1	1	01	04	00004
	4	2	2	6	2	0	4	0	0	2	0	0000	00060				

Fig. 11: Extracted Data (Data Extraction from NSSO)

4. Finally click on the Upload into Excel to upload the searched data into the excel sheet.

Summary Statistics

1. On clicking “Summary Statistics”, folder containing uploaded files will appear on screen.
2. Select file.
3. Click on buttons shown as per requirement then click “Show Details”.
4. Click “Data Scrutiny & Editing” if required otherwise click “Summary Statistics”.

Overview of SSDA2.0 Software

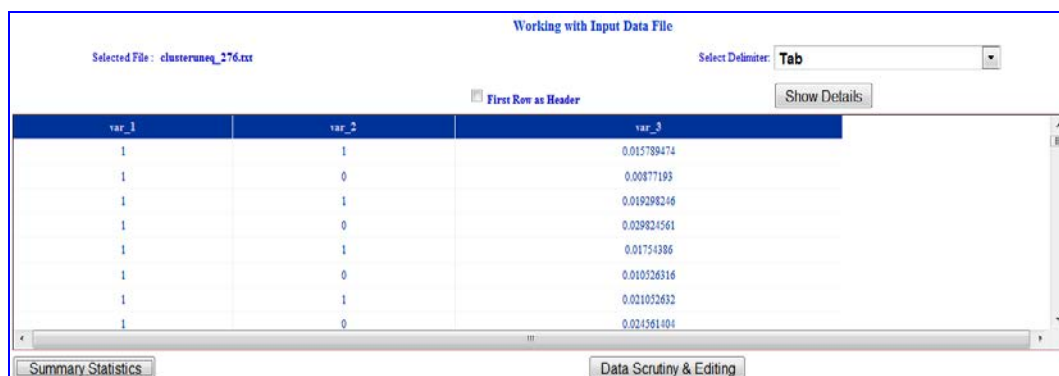


Fig. 12: Loading a Text File

5. Click on Download to MS-Excel to upload the results.(Fig. 13)

Fig. 13: Summary Statistics

Data Scrutiny and Editing

1. Click “Data Scrutiny and Editing” given in left side bar.
2. Select file.
3. Click Edit, make corrections, click Update (Fig. 14).
4. Then click on “Edit Records” button right side.

Fig. 14: Data Scrutiny and Editing

Data Imputation

Missing values, if any, are imputed using any of the three methods given below in this software. These are:

- (i) **Zero imputation:** All the missing values are filled with value zero.

- (ii) **Mean imputation:** The missing values are filled with the mean value within the same stratum or stages. The stratum and stages present in the data is supplied by the user of the software. In case the stratum is specified, the mean values are calculated from within the stratum. In case stratum is not present but stage(s) is(are) present, means are calculated within the stage(s). In case, both stratum and stage(s) are specified, the means are calculated within stage(s) of each stratum.
- (iii) **Mean of neighboring units:** This method fills the gaps by calculating mean of neighboring units. If it is first or last value within the available data then it is filled with the one of the available values. In case of more than one continuous missing value it goes for next available value for calculating mean. This method of imputation also takes care of stratum and stage(s) as described above.
 1. Click “Data Imputation” given in left side bar.
 2. Select “Stratum present”, “number of stages”, “which column” and Stage I, II, III according to the entry made in “number of stages”.(Fig. 15)
 3. Select “Method of Imputation”.
 4. Click “Yes” for “Sampling Parameter Selection Done” after selection is completed at steps 2 and 3.
 5. Select variables from the box showing list of variables under “Available Variables” box.
 6. Click Proceed (Imputation) button.

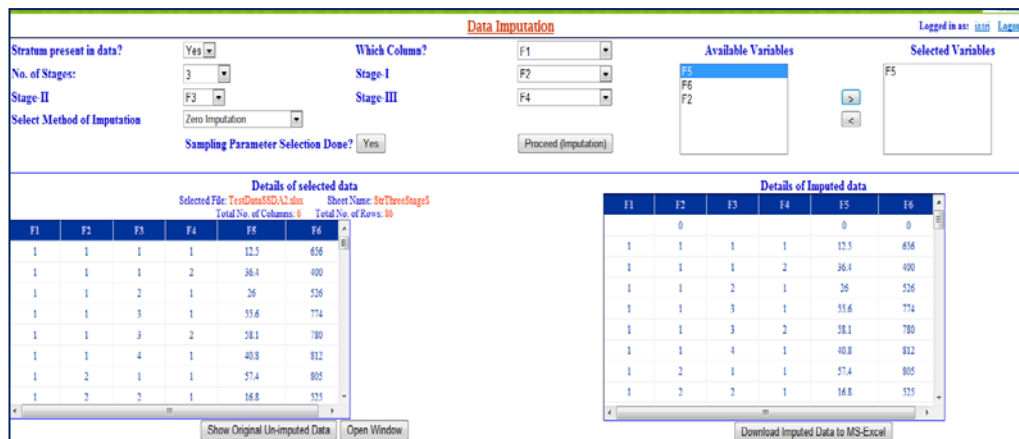


Fig. 15: Imputation in SSDA2

Sample Selection: The sample selection module helps a surveyor to select the sample from a population. The population area can have a number of strata with selections up to three stages. At every stage the surveyor needs to select representative samples. The module implements both equal and unequal probability selection methods. The selection with equal probability can have three methods of selection namely, Simple Random Sampling (SRS) – with replacement and without replacement and systematic sampling. The selection method in unequal probability is Probability Proportional to Size With Replacement (PPSWR). The Fig. 16 shows the screen snapshot of Sample Selection Module of SSDA2.

Sample Selection (Stage 1)
 Stratified Sampling? Yes

(i) No. of Stratum: 1 [Download Stratum Details Sheet](#)

(ii) File for Stratum Sizes: --Select File Name-- [Click to Upload](#)

(iv) Col. for Str-ID: []

(vi) Col. for Selected Unit: []

(iii) Sheet Name: []

(v) Col. for Total Size: []

[Read Data and Proceed](#)

Selection Method (Stage 1)
 Selection Probability: Equal Probability

(i) Method of Selection: SRSWR

(ii) Method of Selection: PPSWR Aux. Var Prob. Val

(iii) File for Prob./Aux Variable: [] [Click to Upload](#)

(iv) Sheet Name: []

(v) Col. for Aux/Prob. Values: []

Selection in Stratum

Selected Units in Stratum

[Show Selection -->](#)

[Proceed to Next Stage](#) [Finish](#)

Fig. 16: Sample Selection in SSDA2 (Stage 1)

Sample Selection (Stage 1)
 Stratified Sampling? Yes

(i) No. of Stratum: 1 [Download Stratum Details Sheet](#)

(ii) File for Stratum Sizes: iasriStrDetails.xls [Click to Upload](#)

(iv) Col. for Str-ID: StratumNum

(vi) Col. for Selected Unit: NumSelected

(iii) Sheet Name: Table1\$

(v) Col. for Total Size: StratumSize

[Read Data and Proceed](#)

Selection Method (Stage 1)
 Selection Probability: Equal Probability

(i) Method of Selection: SRSWR

(ii) Method of Selection: PPSWR Aux. Var Prob. Val

(iii) File for Prob./Aux Variable: [] [Click to Upload](#)

(iv) Sheet Name: []

(v) Col. for Aux/Prob. Values: []

Selection in Stratum

StratumNum	StratumSize	NumSelected
1	25	6
2	23	5

Selected Units in Stratum

StratumNum	Stage1
1	23
1	4
1	9
1	23
1	10
1	20

[Show Selection -->](#)

[Proceed to Next Stage](#) [Finish](#)

Fig. 17: Sample Selection in SSDA2 (Stage 2)

The sample selection module suggests the selection within the stratum, in case the sampling is to be done in more than one stratum. The total number of units and number of units to be selected within the stratum needs to be supplied by the user of the system. The available methods of selection are same in all the stages. The user can although select different selection methods at successive stages.

Sample Selection (Stage 3) Logged in as: [iasri](#) [Logout](#)

(i) File for Stage 1 Sizes: Insert Two Columns in this file as (i) Column containing Total No. of Units in the Stage 2, and (ii) Column Name containing No. of Units to be selected from each Stage 2

[Click to Upload](#) [Download Stage 2 Details Sheet](#)

(ii) Sheet Name: (iii) Col. for Str-ID:

(iv) Col. for Stage1-ID: (v) Col. for Stage2-ID:

(vi) Col. for Total Size: (vii) Col. for Selected Unit:

Selection Method (Stage3)

Selection Probability:

(i) Method of Selection:

(ii) Method of Selection: PPSWR Aux. Var Prob. Val (iii) File for Prob./Aux Variable:

(iv) Sheet Name: (v) Col. for Aux/Prob. Values:

[Click to Upload](#)

Selection in Stage 2

StratumNum	Stage1	Stage2	Total	Selected
1	5	3	4	2
1	5	4	4	2
1	5	3	3	1
1	8	4	4	2
1	8	7	5	2
1	10	3	3	1

Selected Units in Stage 2

StratumNum	Stage1	Stage2	Stage3
1	5	3	3
1	5	3	3
1	5	4	3
1	5	4	3
1	5	3	2
1	8	4	3

Fig. 18: Sample Selection in SSDA2 (Stage 3)

Software for Survey Data Analysis
SSDA 2.0

Sample Selection Report Logged in as: [iasri](#)

Selection in Stratum

StratumNum	Stage1
1	7
1	8
1	12
1	5
1	17
1	23
2	3
2	17
2	22
2	9
2	11

Selection in Stage 1

StratumNum	Stage1	Stage2
1	5	18
1	5	18
1	5	9
1	5	2
1	5	9
1	5	12
1	5	5
1	5	12
1	8	10
1	8	11

Fig. 19: Sample Selection in SSDA2 (Stage 3)

Data Analysis in SSDA2: SSDA performs data analysis for sampling designs for stratified data with up to three stages. It means it can perform analysis for one, two and three stages with or without stratified data. The collected data may be for equal or unequal probability. Simple Random Sampling with or without replacement (SRSWR or SRSWoR) and systematic sampling has been included under equal probability scheme. For unequal probability, Probability Proportional to Size With Replacement

(PPSWR) has been provided. The step wise procedure for carrying out the analysis is stated below:

1. After login go to “My Folder” and click on the input data file.
2. Click on “Data Analysis” link given in left side bar.
3. The analysis work is done in three steps as stated below:

Data Analysis (Step I)

1. If the input is a stratified data, Select “Yes” in the “Stratum present”.
2. The input data can have many stages. Select “number of stages” as “1”, “2” or “3” according to the input data you selected. Please select the column names of the stages accordingly.
3. Click “Yes” for “Sampling Parameter Selection Done” after selection is completed at step 2.
4. Select variables from the box showing list of variables under “Available Variables” box for which the analysis will be carried out.
5. Click on “Proceed” button.
6. In the Sample Selection Parameters box, samples selected at different stages can be viewed.
7. Before clicking on “Proceed to Next Step (ii)” button, user has to click on “Download to MS-Excel” button and to save the file.
8. Open the downloaded excel file. User has to add a new column to enter the figures regarding the population sizes of respective stratum or stages.
9. Click on “Proceed to Next Step (ii)” button. (Fig. 20)

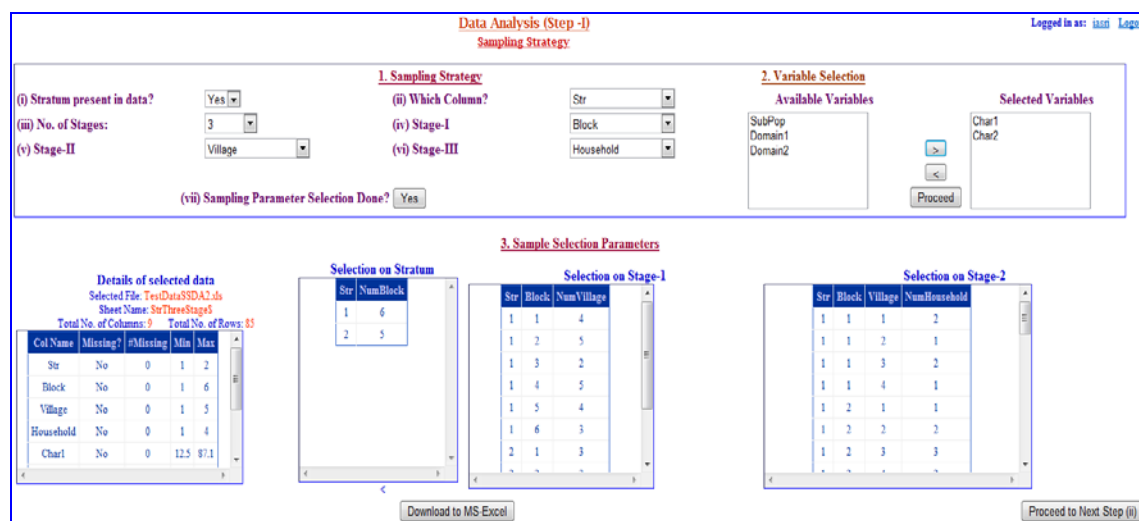


Fig. 20: Data Analysis in SSDA2 (Stage 1)

Data Analysis (Step II)

1. First bar displays the information regarding selection done by the user at Data Analysis (Step I).
2. Click on the fields regarding Population Sizes viz. select file name for population values, etc.

3. Click Probability Selection,

- i) If equal probability is selected, then click method of selection at different stages (Fig. 22).
- ii) If unequal probability is selected, then click on respective fields provided (Fig. 23).

4. Finally click on “Proceed for Analysis” button.

Data Analysis (Step -II) Logged in as: [iasri](#) [Logout](#)

Parameters Already Selected for Analysis

Selected File: [TestImputeDataSSDA2.xls](#) Sheet Name: [ThreeStage\\$](#) Total No. of Columns: 5 Total No. of Rows: 48

1. Sampling Strategy

(i) Stratum: F1 (iii) Stage-I: F2
 (ii) No. of Stages: 3 (iv) Stage-II: F3
 (vi) Stage-III: F4

2. Selected Variables

4. Other Required Parameters for Analysis

4.1 Population Sizes

(i) File Name for Population Values: --Select File Name-- [Click to Upload](#)
 (ii) Sheet Name for Stage-1:
 (iii) Column Name for Pop Size:
 (iv) Sheet Name for Stage-2:
 (v) Column Name for Pop Size:
 (vi) Sheet Name for Stage-3:
 (vii) Column Name for Pop Size:

4.2 Selection Probability

Equal Probability

(i) Method of Selection: Stage-1: SRSWR Stage-2: SRSWR Stage-3: SRSWR
 (ii) Method of Selection: PPSWR: Aux. Var Prob. Val
 (iii) File Name for Prob./Aux Variable: --Select File Name-- [Upload](#)
 (iv) Sheet Name for Stage-1:
 (v) Column Name for Prob. Values:
 (vi) Sheet Name for Stage-2:
 (vii) Column Name for Prob. Values:
 (viii) Sheet Name for Stage-3:
 (ix) Column Name for Prob. Values:

Fig. 21: Data Analysis in SSDA2 (Stage 1)

Data Analysis (Step -II) Logged in as: [iasri](#) [Log](#)

Parameters Already Selected for Analysis

Selected File: [TestDataSSDA2.xls](#) Sheet Name: [StrThreeStage\\$](#) Total No. of Columns: 9 Total No. of Rows: 85

1. Sampling Strategy

(i) Stratum: Str (iii) Stage-I: Block
 (ii) No. of Stages: 3 (iv) Stage-II: Village
 (vi) Stage-III: Household

2. Selected Variables

Char1
Char2

4. Other Required Parameters for Analysis

4.1 Population Sizes

(i) File Name for Population Values: [iasri.xls](#) [Click to Upload](#)
 (ii) Sheet Name for Stage-1: [Str\\$](#)
 (iii) Column Name for Pop Size: [TotalBlock](#)
 (iv) Sheet Name for Stage-2: [Block\\$](#)
 (v) Column Name for Pop Size: [TotalVillage](#)
 (vi) Sheet Name for Stage-3: [Village\\$](#)
 (vii) Column Name for Pop Size: [TotalHousehold](#)

4.2 Selection Probability

Unequal Probability

(i) Method of Selection: Stage-1: SRSWR Stage-2: SRSWR Stage-3: SRSWR
 (ii) Method of Selection: PPSWR: Aux. Var Prob. Val
 (iii) File Name for Prob./Aux Variable: [Book1.xls](#) [Upload](#)
 (iv) Sheet Name for Stage-1: [Sheet1\\$](#)
 (v) Column Name for Prob. Values: [AuxBlock](#)
 (vi) Sheet Name for Stage-2: [Sheet2\\$](#)
 (vii) Column Name for Prob. Values: [AuxVill](#)
 (viii) Sheet Name for Stage-3: [Sheet3\\$](#)
 (ix) Column Name for Prob. Values: [AuxHH](#)

Fig. 22: Data Analysis in SSDA2 (Stage 2)

Analysis Results

All the selections done by user at earlier stages of Data Analysis can be viewed. To correct, user can click on Back button and make selections.

Click any one or all the three radio button “Ratio Estimates”, Sub Population, Domain if required before clicking the Result button.

Overview of SSDA2.0 Software

Result Options

Ratio Estimates Sub Population Domain

Ratio Estimate
 Select Numerator : Char1 Select Denominator : Char2
 Mean (Ratio Method) Population Mean of Numerator : 23233
 Total (Ratio Method) Population Total of Numerator : 76767686

Sub-Population
 Sub-Pop Name : SubPop

Domain
 Select 1st Var : Domain1 Select 2nd Var : Domain2

ANALYSIS RESULTS [Click for Results](#)

Ratio Estimation **Ratio Mean : 1632.425985** **Ratio Total : 5393946.776885**

Sub Population Estimation

SubPop	Char1Total	Char1Mean	Char2Total	Char2Mean
1	9.707238	0.746711	168.646343	12.972796
2	5.605985	0.373732	90.898868	6.059925
3	4.822318	0.344451	52.760759	3.768626
4	1.637798	0.148891	21.419405	1.947219
5	3.344942	0.196761	34.073167	2.004304
6	1.712114	0.244388	17.585121	2.51216
7	1.479529	0.184941	17.52861	2.191076

[Save to Excel](#)

Domain Estimation

Domain1	Domain2	Char1Total	Char1Mean	Char2Total	Char2Mean
1	1	1.9691682	0.49229205	35.305098	8.8262745
1	2	3.5706172	0.71412344	59.886605	11.977321
1	3	5.7655031	0.72068789	109.171402	13.64642525
1	4	3.4293948	0.42867435	46.295108	5.7868885
1	5	3.2321737	0.29383397	35.390027	3.21727518
2	1	2.7166127	0.27166127	33.191946	3.3191946
2	2	1.770921	0.1770921	19.744676	1.9744676
2	3	2.6638891	0.19027779	28.81368	2.05812
2	4	3.191643	0.2127762	35.113731	2.3409154

[Save to Excel](#)

Fig. 23: Data Analysis in SSDA2 (Stage 3) – Part 1

[Save to Excel](#)

Domain Estimation

Domain1	Domain2	Char1Total	Char1Mean	Char2Total	Char2Mean
1	1	1.9691682	0.49229205	35.305098	8.8262745
1	2	3.5706172	0.71412344	59.886605	11.977321
1	3	5.7655031	0.72068789	109.171402	13.64642525
1	4	3.4293948	0.42867435	46.295108	5.7868885
1	5	3.2321737	0.29383397	35.390027	3.21727518
2	1	2.7166127	0.27166127	33.191946	3.3191946
2	2	1.770921	0.1770921	19.744676	1.9744676
2	3	2.6638891	0.19027779	28.81368	2.05812
2	4	3.191643	0.2127762	35.113731	2.3409154

[Save to Excel](#)

Total Estimate

VarName	Total	Mean
Char1	28.309924	42.28
Char2	431.222197	322.01

[Save to Excel](#)

Variance Estimate

Str	Count	SigmaW	SigmaWXChar1	MeanXChar1	SigmaWXChar2	MeanXChar2	VarianceChar1	VarianceChar2
1	47	0.502016	20.797378	93.891642	321.246866	90.067805	3027457.442005	4671963951.38839
2	38	0.167554	7.512544	227.499484	81.665407	233.954629	9827505.825453	317095295.53449

[Save to Excel](#)

Fig. 24: Data Analysis in SSDA2 (Stage 3) – Part 2

Finally click on “Click for Results” button and can save in MS-Excel files.

REFERENCES

Mahajan, V.K., Lal, S.B. and Sharma, Anu (2008). *Software for Survey Data Analysis*. Project Report. IASRI Publication.

Shahi, Priyanka (2003). Development of a Window Based Software for the Analysis of two stage Survey Data M.Sc. thesis, PG School , IARI.

Cochran, W. G. (2002). *Sampling Techniques*. John Wiley & Sons, Inc., New York.

P. C. CARP (1986, 1989). Users Manual Statistical Laboratory Iowa State University, Ames, Iowa. (Edited by Wayne Fuller, William Kannedy, Daniel Schnell GarySullivan,Heon Jin Park).

SAS | Business Intelligence Software and Predictive Analytics.

<http://www.sas.com/>

Sample Survey Design and Analysis: Overview

<http://www.sas.com/rnd/app/da/new/dasurvey.html>

Stata: Data Analysis and Statistical Software <http://www.stata.com>

WesVar - Software and Analysis of Data from Complex Samples

<http://www.westat.com/wesvar/>