उच्च संकाय प्रशिक्षण केंद्र
Centre for Advanced Faculty Training

संदर्भ पुस्तिका-I
Reference Manual-I

कृषि सर्वेक्षण के डेटा विश्लेषण की आधुनिक तकनीकें

(कृषि शिक्षा विभाग, भा.कृ.अनु.प. द्वारा प्रायोजित)

MODERN DATA ANALYTICS TECHNIQUES FOR AGRICULTURAL SURVEYS

(Sponsored by Agricultural Education Division, ICAR)

फरवरी 11- मार्च 03, 2025

February 11- March 03, 2025

पाठ्यक्रम समन्वयक : कौस्तब आदित्य
पाठ्यक्रम सह-समन्वयक : अंकुर बिश्वास
पाठ्यक्रम सह-समन्वयक : पंकज दास

Course Coordinator : Kaustav Aditya
Course Co-Coordinator : Ankur Biswas
Course Co-Coordinator : Pankaj Das

प्रतिदर्श सर्वेक्षण प्रभाग
भा.कृ.अनु.प. - भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली-110012

DIVISION OF SAMPLE SURVEYS
ICAR-INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE
LIBRARY AVENUE, PUSA, NEW DELHI-110012
https://iasri.icar.gov.in
2025

प्राक्कथन

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान (भा.कृ.सां.अ.सं.) ने 1930 में तत्कालीन इंपीरियल कृषि अनुसंधान परिषद में एक सांख्यिकी अनुभाग के रूप में अपनी यात्रा शुरू की और सांख्यिकी विज्ञान (सांख्यिकी, संगणक अनुप्रयोग और जैव सूचना विज्ञान) के क्षेत्र में अनुसंधान, शिक्षा और प्रशिक्षण आयोजित करने के लिए प्रासंगिकता के एक प्रमुख संस्थान के रूप में विकसित हुआ है। संस्थान मुख्य रूप से मौजूदा ज्ञान में अंतराल को पाटने के लिए कृषि सांख्यिकी और सूचना विज्ञान में अनुसंधान करने के लिए जिम्मेदार है। संस्थान, सांख्यिकी विज्ञान को, सूचना विज्ञान के साथ मिश्रित करके, कृषि विज्ञान में उनके विवेकपूर्ण सम्मिश्रण का उपयोग कर रहा है, ताकि कृषि के नए उभरते क्षेत्रों की चुनौतियों का सामना कर गुणवत्तापूर्ण कृषि अनुसंधान को बढ़ाया जा सके, और साक्ष्य आधारित नीति निर्णय लिए जा सके। संस्थान ग्रेजुएट स्कूल, भा.कृ.अनु.प.–भारतीय कृषि अनुसंधान संस्थान, नई दिल्ली के सहयोग से कृषि सांख्यिकी, संगणक अनुप्रयोग और जैव सूचना विज्ञान में एम.एस.सी. और पीएच.डी. डिग्री कार्यक्रम भी संचालित करता है। संस्थान राष्ट्रीय कृषि अनुसंधान एवं शिक्षा प्रणाली (एनएआरईएस) को सुदृढ़ बनाने के लिए सलाहकारी एवं परामर्श सेवाएं प्रदान करता है तथा राष्ट्रीय एवं अंतर्राष्ट्रीय संगठनों के लिए प्रायोजित अनुसंधान एवं परामर्श सेवाएं प्रदान करता है। राष्ट्रीय कृषि सांख्यिकी प्रणाली (एनएएसएस) को सुदृढ़ बनाने में पद्धतिगत सहायता भी प्रदान की जाती है। संस्थान एनएआरईएस के लिए मजबूत कृषि ज्ञान प्रबंधन प्रणालियों और कृत्रिम बुद्धिमत्ता आधारित अनुप्रयोगों के विकास में भी अग्रणी भूमिका निभा रहा है।

सांख्यिकीय रूप से मान्य एवं अर्थपूर्ण निष्कर्ष, जो अनुसंधान कार्यक्रमों से प्राप्त होते हैं, उच्च–गुणवत्ता वाले अनुसंधान का आधार बनते हैं एवं नीतिगत योजना, विशेष रूप से विकास संबंधी पहल एवं कार्यक्रम कार्यान्वयन के लिए अत्यंत महत्वपूर्ण होते हैं। इसलिए, डेटा संग्रह एवं विश्लेषण के लिए मजबूत सांख्यिकीय विधियों का उपयोग करना आवश्यक है। संस्थान द्वारा प्रदान किए जाने वाले प्रशिक्षण कार्यक्रम शोधकर्ताओं एवं योजनाकारों को नवीनतम सांख्यिकीय तकनीकों से परिचित कराने में अमूल्य सिद्ध होते हैं।

संस्थान कृषि सांख्यिकी एवं कंप्यूटर अनुप्रयोग में उन्नत संकाय प्रशिक्षण केंद्र (सीएएफटी) भी है। 11 फरवरी से 3 मार्च 2025 के दौरान आयोजित "कृषि सर्वेक्षण के डेटा विश्लेषण की आधुनिक तकनीकें" नामक यह प्रशिक्षण कार्यक्रम कृषि शिक्षा विभाग, भा.कृ.अनु.प. द्वारा प्रायोजित है। इस प्रशिक्षण कार्यक्रम को संकाय सदस्यों/वैज्ञानिकों को विभिन्न प्रतिदर्श पद्धतियों की जानकारी प्रदान करने के उद्देश्य से तैयार किया गया है, जिसमें नवीनतम विकास एवं सर्वेक्षण डेटा विश्लेषण के लिए सॉफ्टवेयर पैकेजों के उपयोग पर विशेष ध्यान दिया गया है। साथ ही, भारत में कृषि एवं बागवानी सांख्यिकी के संग्रह प्रणाली की भी जानकारी दी जाएगी। इसके अतिरिक्त, फसल उत्पादन अनुमान/पूर्वानुमान के लिए एआई/एमएल, भा.कृ.अनु.प. के दृष्टिकोण से डिजिटल कृषि से संबंधित कुछ महत्वपूर्ण विषयों आदि को भी शामिल किया गया है। प्रशिक्षण कार्यक्रम व्यावहारिक रूप से उन्मुख है जिसमें अनुभव एवं क्षेत्र के दौरे पर जोर दिया गया है।

इस प्रशिक्षण कार्यक्रम के संकाय में वैज्ञानिकों एवं प्रतिष्ठित सांख्यिकीविदों को शामिल किया गया है, जो प्रतिदर्श सर्वेक्षण एवं संबंधित क्षेत्रों में व्यापक अनुभव रखते हैं। इस प्रशिक्षण कार्यक्रम की शुरुआत से पहले प्रकाशित एवं वितरित की जाने वाली संदर्भ प्रशिक्षण पुस्तिका प्रतिभागियों के कार्य कौशल को समृद्ध करने में बहुमूल्य ज्ञान प्रदान करेगी। यह आशा की जाती है कि इस प्रशिक्षण कार्यक्रम से प्राप्त अनुभव प्रतिभागियों को अपने कार्यस्थल पर इस ज्ञान का प्रभावी रूप से उपयोग करने में सक्षम बनाएगा। मैं इस संदर्भ प्रशिक्षण पुस्तिका को समय पर प्रस्तुत करने के लिए प्रभागाध्यक्ष (प्रतिदर्श सर्वेक्षण) एवं पाठ्यक्रम समन्वयकों को बधाई देता हूँ।

(राजेन्द्र प्रसाद)
निदेशक, भा.कृ.अनु.प.–भा.कृ.सां.अ.सं.

नई दिल्ली
11 फरवरी, 2025

FOREWORD

The ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI) began its journey as a Statistical Section in 1930 in the then Imperial Council of Agricultural Research. Over the years, it has evolved into a premier institute dedicated to research, education, and training in Statistical Sciences (Statistics, Computer Applications and Bioinformatics). ICAR-IASRI has played a key role in advancing research in Agricultural Statistics and Informatics, addressing knowledge gaps in these fields. The Institute offers M.Sc. and Ph.D. programs on Agricultural Statistics, Computer Applications and Agricultural Bioinformatics in collaboration with the Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi. Additionally, ICAR-IASRI provides customized and sponsored training courses at both national and international levels, aiming to be a center of excellence in Human Resource Development. The Institute also offers advisory and consultancy services to strengthen the National Agricultural Research and Education System (NARES), conducts sponsored research for national and international organizations, and supports the development of a robust Agricultural Knowledge Management System for NARES.

Statistically valid and meaningful inferences derived from research programmes form the basis of high-quality research and are crucial for policy planning, particularly in developmental initiatives and programme implementation. Consequently, it is vital to employ robust statistical methodologies for data collection and analysis. The training programs offered by the Institute are invaluable in familiarizing researchers and planners with the latest advancements in statistical techniques.

The Institute is also a Centre of Advanced Faculty Training in Agricultural Statistics and Computer Application. This training programme entitled "**Modern Data Analytics Techniques for Agricultural Surveys**" organized during February 11-March 3, 2025 is sponsored by Agricultural Education Division, ICAR. 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 the system of collection of agricultural and horticultural statistics in India. In addition, some important topics related to AI/ML for crop yield estimation/forecasting in India, Digital agriculture - ICAR perspective etc. have also been included. The training programme is practical oriented with emphasis on hands on experience and field visits.

The faculty of this training programme comprises of scientists and eminent statisticians with vast experiences 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 Head, Division of Sample Surveys and Course Coordinators for bringing out this valuable document in time.

11/2/2025

(Rajender Parsad)
Director, ICAR-IASRI

New Delhi
11 February, 2025

vkeq[k

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकीय अनुसंधान संस्थान (भा.कृ.अनु.प.–भा.कृ.सां.अ.सं.) कृषि सांख्यिकी, कृषि जैवसूचना एवं संगणक अनुप्रयोग के क्षेत्र में अनुसंधान को बढावा देने तथा संचालित करने के लिये एक प्रमुख संस्थान है। यह संस्थान, भारतीय कृषि अनुसंधान परिषद् (भा.कृ.अनु.प.) के मानव संसाधन विकास कार्यक्रम के तत्वाधान में कृषि सांख्यिकी एवं संगणक अनुप्रयोग में उच्च संकाय प्रशिक्षण केन्द्र के रूप में भी कार्यरत है। कृषि फसलों, बागवानी फसलों, पशुधन एवं मत्स्य पालन के मामलों में विभिन्न प्राचलों के आकलन से सम्बन्धित प्रतिदर्श सर्वेक्षणों सहित कृषि सांख्यिकीय के विभिन्न क्षेत्रों में मौलिक एवं व्यवहारिक, दोनों प्रकार के अनुसंधान किये जा रहे हैं। प्रतिदर्श सर्वेक्षण प्रभाग के वैज्ञानिक प्रतिदर्श सर्वेक्षण के विभिन्न पहलुओं जैसे जटिल सर्वेक्षणों की अभिकल्पना और विश्लेषण, सर्वेक्षण आँकड़ों के विश्लेषण हेतु सॉफ्टवेयर विकसित करना, बूटस्ट्रैप विचरण आकलन तकनीक, अंशांकन और मॉडल अंशांकन अनुमानक, लघु क्षेत्र आकलन, रैंक सेट प्रतिचयन, अनुकूली क्लस्टर प्रतिचयन, एकाधिक फ्रेम सर्वेक्षण, स्थानिक नमूनाकरण एवं आकलन, भौगोलिक रूप से भारित प्रतिगमन आधारित आकलन, कृषि सर्वेक्षणों में भौगोलिक सूचना प्रणाली एवं सुदूर संवेदी तकनीकों का अनुप्रयोग इत्यादि के अनुसंधान में लगे हुए हैं। यह प्रभाग प्रतिदर्श सर्वेक्षण के क्षेत्र में कई अनुप्रयुक्त अनुसंधान गतिविधियों के लिए संयुक्त राष्ट्र के खाद्य एवं कृषि संगठन (एफएओ) के साथ अंतरराष्ट्रीय सहयोग में भी प्रवृत्त है।

''कृषि सर्वेक्षण के डेटा विश्लेषण की आधुनिक तकनीकें'' नामक इस प्रशिक्षण कार्यक्रम का मुख्य उद्देश्य कृषि विज्ञान के विभिन्न विषयों से सम्बन्धित प्रतिभागियों को प्रतिचयन की विभिन्न तकनीकों एवं आकलन विधियों, प्रतिदर्श सर्वेक्षणों में नवीनतम विकास एवं प्रतिदर्श आँकड़ों के विश्लेषण में प्रयोग होने वाले सॉफ्टवेयर पैकेज जैसे MS-Excel, R, SAS, Python एवं SPSS के प्रयोग, कृषि सर्वेक्षणों में भौगोलिक सूचना प्रणाली एवं सुदूर संवेदी तकनीकों का अनुप्रयोग इत्यादि की जानकारी प्रदान करना है। सैद्धांतिक के अपेक्षा व्यावहारिक पहलुओं पर अधिक जोर दिया गया है। प्रतिभागियों के उपयोग के लिए संदर्भ पुस्तिका सरल रूप में प्रस्तुत की गयी है।

हम संस्थान एवं अतिथि संकाय के सभी संकाय सदस्यों का धन्यवाद करते है जिन्होंने इस कार्यक्रम को सार्थक एवं सफल बनाने में अपना बहुमुल्य समय लगा कर सहयोग दिया है। हम प्रशिक्षण कार्यक्रम के आयोजन के लिए विभिन्न व्यवस्थाएं करने के लिए शामिल विभिन्न समितियों के अध्यक्षों एवं सदस्यों के भी आभारी हैं। उनके अथक प्रयासो से इस संदर्भ पुस्तिका को समय से तैयार करने में मदद मिली है। हम इस प्रशिक्षण कार्यक्रम में प्रतिभागियों को नामित करने के लिए भारतीय कृषि अनुसंधान परिषद् के विभिन्न संस्थानों, राज्य कृषि विश्वविद्यालयों आदि के आभारी हैं। इस प्रशिक्षण कार्यक्रम के आयोजन का दायित्व हमें सौंपने के लिए हम भारतीय कृषि अनुसंधान परिषद् के शिक्षा प्रभाग के आभारी हैं। हम डॉ. राजेंद्र प्रसाद, निदेशक, भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान एवं डॉ. तौकीर अहमद, प्रभागाध्यक्ष, प्रतिदर्श सर्वेक्षण प्रभाग का इस कार्यक्रम में मार्गदर्शन एवं निरंतर सहयोगएवं प्रशिक्षण कार्यक्रम को सुचारू संचालन के लिए सभी आवश्यक सुविधाएं उपलब्ध कराने के लिए आभारी हैं। अंत में, हम उन सभी का आभार प्रकट करते हैं जिन्होंने, इस संदर्भ पुस्तिका को तैयार करने में सहयोग दिया है।

ub fnYyh **y[kdx.k**
11 Qjojh] 2025

PREFACE

The ICAR-Indian Agricultural Statistics Research Institute, New Delhi is a premier Institute for promoting and conducting research in the field of Agricultural Statistics, Agricultural Bioinformatics and Computer Applications. 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 are 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. Scientists of the Division of Sample Surveys are engaged in research on various aspects of sample surveys like design and analysis of complex surveys, application of statistical softwares for survey data analysis, bootstrap variance estimation techniques, calibration and model calibration estimators, small area estimation, ranked set sampling, adaptive cluster sampling, multiple frame surveys, spatial sampling and estimation, geographically weighted regression based estimation approaches, application of GIS and remote sensing techniques in agricultural surveys etc. The division is also engaged in international collaborations with Food and Agriculture Organization of the United Nations (FAO) for several applied research activities in the field of sample surveys.

The broader objective of this training programme on "**Modern Data Analytics Techniques for Agricultural Surveys**" is to provide exposure to the participants belonging to different disciplines of agricultural sciences in proper understanding of various sampling techniques and estimation procedures, some recent developments in sample surveys, use of software packages for survey data analysis like MS-Excel, R, SAS, Python and SPSS, application of remote sensing and GIS techniques in agricultural surveys etc. More emphasis is given on the applied aspects rather than theoretical. The reference manual is presented in a simplified and comprehensive manner for better usage by the participants.

We take this opportunity to thank all the faculty members from the institute and the guest faculties who have devoted their valuable time and energy in making this training program successful. Their sincere efforts helped in bringing out this lecture manual on time. We are also thankful to the Chairman and members of various committees involved in successful organization of this training programme. We are also thankful to various ICAR Institutes, State Agricultural Universities etc. for nominating participants to this training programme. We are indebted to the Agricultural Education Division of ICAR for entrusting the responsibility of organizing this training programme. We are also grateful to Dr. Rajender Parsad, Director, ICAR-IASRI and Dr. Tauqueer Ahmad, Head, Division of Sample Surveys for their guidance and continuous support in this training programme and providing all the necessary facilities for smooth conduct of this training programme. In the end, we are thankful to one and all who helped in preparing this reference manual.

New Delhi **Authors**
February 11, 2025

विषय-सूची

CONTENTS

BASIC STATISTICAL METHODS

Bharti

ICAR-Indian Agricultural Statistics Research Institute, New Delhi

1. Introduction

The word "Statistics" is derived from the Latin word "Status," the Italian "statista," or the German "statistik," all of which refer to a 'political state.' Over time, various authors have defined statistics in different ways. Generally, the term "Statistics" is used in two contexts: (i) in its plural form, it refers to quantitative data, and (ii) in its singular form, it denotes the science of collecting, classifying, analyzing, and interpreting data.

Statistics as Numerical data – definitions

- Statistics are numerical statement of facts in any department of enquity placed in relation to each other -Bowley
- By statistics we mean quantitative data affected to a marked extent by multiplicity of causes -Yule and Kendall

Statistics as statistical Methods- definitions

- Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data -Croxton and cowden
- Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of population of natural phenomenon -Kendall

2. Scope of Statistics in Agriculture

Statistics plays a crucial role in collecting, classifying, tabulating, analyzing, and interpreting agricultural data, enabling informed decision-making. Some key areas where statistics plays a significant role in agriculture include:

- **Collection of Agricultural Statistics**: Statistics is essential for gathering data on various aspects of agriculture, including crop production, livestock statistics, fishery statistics etc. This data helps in creating accurate records and forms the foundation for informed decision-making.
- **Analyzing Experimental Data**: Agricultural research often involves controlled experiments to test different farming practices, crop varieties, or fertilizers. Statistical techniques are used to analyze experimental data, ensuring that the results are reliable and meaningful. This helps in identifying the most effective methods for improving crop productivity and sustainability.
- **Prediction of Yields**: Statistical models, particularly time series analysis and forecasting techniques, are used to predict future crop yields based on historical data and environmental factors like weather patterns, soil health, and irrigation methods. Accurate predictions help farmers plan better and mitigate risks related to crop failure or low yields.
- **Crop Yield Estimation**: Estimating crop yield is vital for agricultural planning, policy formulation, and ensuring food security. Statistics plays a key role in designing and analyzing surveys and field experiments to estimate the potential yield of crops in different regions, helping policymakers allocate resources effectively.

- **Agricultural Policy and Economics**: Governments and policy-makers use statistical data to formulate agricultural policies, subsidies, and programs that support farmers. It also helps in evaluating the economic performance of the agricultural sector.

3. **Limitation of Statistics:**
   Although statistics is a powerful tool, it does have several limitations, including:

- **Statistical studies are true only on an average:** Statistical methods often deal with general trends, averages, and patterns observed across large datasets. The results of a statistical study typically reflect the central tendency or overall trend of a population, rather than the behavior of every individual case. This means that while the findings may be true for the population as a whole, they might not always apply accurately to each specific instance within the population. For example, an average crop yield prediction based on historical data might be accurate overall, but it could vary for particular farms due to local conditions such as soil quality, weather, or farming practices.
- **It is associated with some amount of error:** All statistical methods come with a degree of uncertainty or error. These errors can arise from various sources, including measurement inaccuracies, sampling errors, or even flaws in the assumptions made when conducting the analysis. For instance, if a sample is used to estimate the characteristics of a larger population, the sample might not perfectly represent that population, introducing sampling error. Even with precise data, random fluctuations or uncontrollable factors might influence results.
- **It is liable to be misused:** Statistics can be powerful in uncovering trends, making predictions, and guiding decisions, but they are also vulnerable to misuse. This occurs when data is manipulated or selectively presented to support a specific agenda, or when statistical methods are applied incorrectly. For example, presenting only favorable data while ignoring conflicting results, or using misleading charts and graphs, can create a false impression of the findings. Even the way questions are framed or the sample is chosen can skew results.
- **It does not study qualitative phenomena:** Statistical analysis is inherently quantitative, meaning it focuses on numerical data and measurable aspects of a phenomenon. While it excels at summarizing large volumes of numerical information and identifying patterns, it is not equipped to directly study qualitative phenomena such as emotions, behaviors, or the subjective experiences of individuals. For example, while statistics can quantify the average income of farmers in a region, it cannot capture the underlying reasons why farmers feel stressed or motivated, nor can it assess the cultural and social aspects of farming that may influence decision-making.

4. Types of Data and Measurement Scales

Types of Data

In statistics, data is broadly classified into two main types:

- **Qualitative (Categorical) Data**: This type of data represents categories or labels and cannot be measured numerically. It is further divided into:
  - **Nominal Data**: Data that represents categories without any order. (e.g., gender, eye color, marital status)
  - **Ordinal Data**: Data that has a meaningful order but the difference between values is not consistent. (e.g., education level, customer satisfaction ratings)

- **Quantitative (Numerical) Data:** This type of data consists of numbers and can be measured. It is further divided into:
  - **Discrete Data:** Data that takes only specific values and cannot be broken down further (countable). (e.g., number of students, number of cars)
  - **Continuous Data:** Data that can take any value within a given range (measurable). (e.g., height, weight)

Each type of data requires different statistical methods for analysis, interpretation, and visualization.

Measurement Scales: There are four main types of measurement scales:

- **Nominal Scale** (Qualitative, No Order): Used for labeling or categorizing data without any order; No numerical significance or ranking. Example: gender (male, female), eye color (blue, green, brown), blood type (A, B, AB, O).
- **Ordinal Scale** (Qualitative, Ordered): Data is categorized with a meaningful order, but the differences between values are not equal or measurable. Example: education level, satisfaction ratings (satisfied, neutral, dissatisfied).
- **Interval Scale** (Quantitative, No True Zero): Numeric data where differences between values are meaningful, but there is no true zero. Example: Temperature
- **Ratio Scale** (Quantitative, True Zero): Similar to the interval scale but with a true zero point, allowing meaningful ratios. Example: height, weight, age, income, distance, time.

5. Descriptive Statistics

Descriptive statistics deals with summarizing and organizing data to make it easier to understand.

5.1 Measures of Central Tendency: It is a single value within the range of data which is used to represent all values of the series. The objective of this is to get a single value which represent all values of the series and to facilitate the comparison between two or more than two series.

Measures of Central Tendency:

- Arithmetic mean (or Mean)
- Median
- Mode
- Geometric Mean
- Harmonic Mean

- **Arithmetic mean (Mean):** It is sum of observations divided by total number of observations. Suppose $X_1, X_2, \ldots, X_n$ be n observations, then mean can be calculated by

$$\text{Mean } \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Median:** Median is that value which divide the given set of data in two equal parts. The procedure for finding median includes arrangement of the given set of data either in increasing or decreasing order. There are two cases.

  Case I: Number of observation is odd: If number of observations 'n' is odd, then median is $\frac{n+1}{2}$th term.

  Case II: Number of observation is even: Let number of observations 'n' is even. There will be two middle terms i.e $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$ th term, then median is mean of these two middle terms

- **Mode:** Mode is that value which is have maximum frequency in given data set.

- **Geometric mean (GM):** Suppose $X_1, X_2, \ldots, X_n$ be n observations, then geometric mean is nth root of product of n observations (GM = $\sqrt[n]{X_1 . X_2 . X_3 \ldots X_n}$)

- **Harmonic Mean (HM):** Suppose $X_1, X_2, \ldots, X_n$ be n observations, then harmonic mean is reciprocal of arithmetic mean of reciprocal of observations i.e. HM = $\frac{n}{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+\cdots+\frac{1}{X_n}}$

5.2 **Measure of dispersion:** It measures the variability or scatterdness of mass of figures in the given data set from its average. Measure of Dispersion are:

- Range
- Quartile deviation
- Mean deviation
- Standard deviation

- **Range:** It is difference between maximum and minimum value.

- **Quartile deviation (QD):** It is half of inter-quartile range. Mathematically QD = $\frac{Q_3 - Q_1}{2}$, where Q_3 is 3rd quartile and Q_1 is 1st quartile. Quartile is the value which divide given data set in four equal parts.

- **Mean deviation (MD):** It is arithmetic mean of absolute value of deviation of observation from its average. Suppose $X_1, X_2, \ldots, X_n$ be n observations, then MD = $\frac{1}{n}$ $\sum |X - A|$, where A may be mean, median or mode.

- **Standard deviation (SD) :** It is positive square root of arithmetic mean of square of deviation of observations from mean. It is denoted by σ.

- **Variance:** It is arithmetic mean of square of deviation of observations from mean . It is denoted by σ^2.

- **Coefficient of variation (CV):** When two or more than two series differ in their unit and we want to compare these series, then suitable measure is coefficient of variation. It

measure the percentage variation in the mean, where standard deviation is considered as total variation.

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

The series having smaller value of CV is more consistent or having less variability than the series having high value of CV.

5.3 **Skewness:** It measures the lack of symmetry in the shape of distribution. It is of two types

- **Positively skewed:** In positively skewed distribution more frequency is towards right hand side of the distribution. Value of mean is more as compared to mode.
- **Negatively Skewed:** In negatively skewed distribution more frequency is towards left hand side of distribution. Value of mode is more as compared to mean.

Symmetrical distribution: A distribution is symmetrical if value of mean, median and mode are same. Curve is bell shaped and symmetric at mean.

5.4 **Kurtosis:** It measure the flatness or peakness of distribution. Mathematically,

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

If $\beta_2 = 3$, distribution is mesokurtic i.e. distribution is neither more flat nor more peaked
$\beta_2 > 3$, distribution is leptokurtic i.e. distribution is more peaked
$\beta_2 < 3$, distribution is platykurtic i.e distribution is more flatter.

6. **Graphical and Diagrammatic representation of data:** Graphical and diagrammatic representations help present data visually, making it easier to understand, analyze, and interpret. Instead of looking at large tables of numbers, graphs and charts provide a clear, summarized view of the data. Graphical Representation includes histogram, frequency curve, frequency polygon etc. diagram includes one dimensional diagram (simple bar diagram, multiple bar diagram etc.), two dimentional disgram (squares, circular diagram) etc.

7. **Correlation and Regression:** Correlation and regression are both statistical methods used to analyze the relationship between two or more variables, but they serve different purposes.
   - **Correlation:** The measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1
     - **r = 1**: Perfect positive correlation (as one variable increases, the other increases in exact proportion).
     - **r = -1**: Perfect negative correlation (as one variable increases, the other decreases in exact proportion).
     - **r = 0**: No linear correlation (the variables are unrelated in a linear fashion).
     - **0 < r < 1**: Positive correlation (as one variable increases, the other tends to increase).
     - **-1 < r < 0**: Negative correlation (as one variable increases, the other tends to decrease).

- **Regression:** Regression is used to model the relationship between a dependent (response) variable and one or more independent (predictor) variables. Unlike correlation, regression involves predicting the value of the dependent variable based on the independent variables.

8. **Hypothesis testing:** Hypothesis testing is a statistical method used to make inferences or draw conclusions about a population based on a sample of data. It allows researchers to test an assumption (hypothesis) about a population parameter. Steps involved in hypothesis testing are:
   - **State the Hypotheses:**
     - **Null Hypothesis (H_0):** The assumption that there is no difference
     - **Alternative Hypothesis (H_1 or Ha):** The statement that contradicts the null hypothesis, indicating there is an effect or difference.
   - **Choose the Significance Level (α):** This is typically set at 0.05, which means there is a 5% chance of rejecting the null hypothesis when it is actually true.
   - **Select the Appropriate Test:** Depending on the type of data and hypothesis, different tests may be used, such as:
     - **t-test** (for comparing two means)
     - **Chi-square test** (for categorical data)
     - **ANOVA** (for comparing means across more than two groups)
     - **z-test** (for large sample sizes with known population variance)
   - **Collect and Analyze Data:** Gather the sample data and perform the selected statistical test.
   - **Calculate the Test Statistic:** Depending on the test, calculate the statistic (e.g., t, z, F, etc.), which will measure how far the sample statistic is from the population parameter.
   - **Find the p-value:** The **p-value** indicates the probability of obtaining the observed results (or more extreme) given that the null hypothesis is true. If the p-value is less than the significance level (α), reject the null hypothesis.
   - **Make a Decision:**
     - If the **p-value < α**: Reject the null hypothesis (H_0) in favor of the alternative hypothesis (H_1).
     - If the **p-value ≥ α**: Fail to reject the null hypothesis.
   - **Conclusion:** Based on the decision, conclude whether or not there is enough evidence to support the alternative hypothesis.
9. **Sampling Methods:** Sampling methods refer to the techniques used to select a representative subset of individuals from a larger population to give conclusion about the whole target population. Sampling methods can be broadly classified into two categories: probability sampling and non-probability sampling. These categories differ in how the sample is selected and the degree of randomness or bias involved.
   - **Probability Sampling:** In probability sampling, every individual or unit in the population has a known, non-zero chance of being selected. These methods are

considered more statistically rigorous because they reduce bias and allow for generalizations about the population. It includes:

- Simple Random Sampling
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling
- Multistage Sampling

- **Non-Probability Sampling:** In non-probability sampling, not all individuals have a known or equal chance of being selected. This introduces the possibility of bias, making it harder to generalize findings to the entire population. It includes:
  - Convenience Sampling
  - Judgmental (Purposive) Sampling
  - Snowball Sampling
  - Quota Sampling

10. **Conclusion:** Basic statistical methods are crucial for analyzing and interpreting data, offering essential tools for understanding and making decisions based on information. Descriptive statistics summarize data patterns, and inferential statistics inform decision-making by drawing conclusions from samples. A solid understanding of hypothesis testing, correlation, and regression further strengthens analytical skills, enabling deeper insights across various fields. This chapter serves as the foundation for more advanced statistical techniques, empowering data-driven decision-making with greater reliability and effectiveness.

Reference:

Gupta, S. C., and V. K. Kapoor. *Fundamentals of Mathematical Statistics*. Sultan Chand & Sons, 2002.

HYPOTHESIS TESTING: PARAMETRIC AND NON-PARAMETRIC METHODS

Med Ram Verma

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

In situations when values of parameters are not known then we estimate the values of parameters from the sample data. There are two situations (i) when sample value is exactly same as parametric value then we accept this value without any hitch (ii) when the sample value is not same as population value we cannot accept the sample value as parametric value. In this situation there are some procedures or methods that enable us to decide whether to accept or reject the hypothetical value on the basis of sample values. Such procedure is known as testing of hypothesis.

2. Essential Terminology and Concepts

- **Statistical Hypothesis**: This is a statement about the distribution of one or more random variables. If the statistical hypothesis completely specifies the distribution, it is called a simple statistical hypothesis, if it does not, it is called composite statistical hypothesis.
- **Null Hypothesis**: This is the hypothesis which is tested under the assumption that it is true. It is a statistical statement about the parameter that is tested for its possible rejection under the assumption that it is true. It is denoted by H_0.
- **Alternative Hypothesis:** The alternative hypothesis provides an alternate to null hypothesis. A hypothesis used in testing of hypothesis that is contrary to the null hypothesis is known as the alternative hypothesis and denoted by H_1.
- **Statistical Test:** A test of statistical hypothesis is a rule when the experimental sample values obtained leads to a decision to accept or reject the null hypothesis under consideration.
- **Errors in Testing of Hypothesis:**

  Type - I Error: In a hypothesis testing type I error occurs when the null hypothesis is rejected when in fact it is true. The probability of Type-I Error is denoted by α.

  Type - II Error: In a hypothesis testing a type II error occurs when the null hypothesis H_0 is accepted when in fact it is false. The probability of Type-II error is denoted by β.

- **Level of significance**: It is the probability of rejecting null hypothesis (H_0) when it is true. It is denoted by α.

  $$P(\text{Reject } H_0 / H_0 \text{ is true}) = \alpha$$

  Usually, the significance level is chosen to is either 5% or 1% .

- **Power of the Test**: The power of a test is the probability of not committing Type –II error. It is the ability of the test to reject the null hypothesis when it is actually false. It is calculated by subtracting the probability of Type - II error from 1.

  Power of test = 1 – P (Type–II error) = 1-β

3. **Steps in Procedure of Hypothesis Testing:** The usual process of hypothesis testing consists following steps:
   - State the null hypothesis H_0 and the alternative hypothesis H_1 about the populationparameter
   - Choose the level of significance
   - Choose the appropriate test statistic
   - Compute the value of test statistic
   - Decide critical value and critical region
   - Decision about the acceptance or rejection of the null hypothesis
   - Write the conclusion of test in simple language

Tests of significance

1. Large sample tests (n>30)
2. Small sample tests (n<30)

Since small samples tests are mostly used in actual practice. So we will discuss small sample tests and their applications.

4. Parametric Tests

4.1 t-test: When the sample size is small then the distribution of the variable is not normal. In that case variable follows a distribution which is known as t-distribution. This test was given by W.S. Gosset in 1908 who wrote under the nickname 'Student'. This test is based on the following assumptions.

Assumptions

(i) The parent population from which the sample is drawn is normal

(ii) Sample observations are independent

(iii) Population variance is unknown

(iv) Sample size is small (n<30).

Applications of t-test

1. Testing the significance of population mean
2. Testing the equality of two population means
3. Paired t test
4. Testing the significance of correlation coefficient
5. Testing the significance of regression coefficient

Testing the significance of population mean (Student t-test)

When the sample size is small and we want to test the hypothesis whether the sample has been drawn from a population with specified population mean value. Then we use the t-test. The null and alternate hypotheses under this test are given below.

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

Test statistics is given by

$$t_{cal} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean based on n observations. s is the sample standard deviation which is calculated by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The t_{cal} statistic follows the t- distribution with n-1 degrees of freedom.

Testing the equality of two population means (Independent Samples t-test)

This test is also known as Fisher's test. This test is used to test the hypothesis that the population means of the two independent samples are same. In other words, this test is used to test the hypothesis whether the two samples have been drawn from the same populations. Suppose we have two independent samples X ($X_1 \ldots X_{n_1}$) and Y ($Y_1 \ldots Y_{n_2}$) drawn from the normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. This test is based on the assumption that population variances for the both the samples are same. The Null and alternate hypothesis are given below.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Then the test statistic is given by

$$t_{cal} = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$$

$$s_1^{\,2} = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2 \quad \text{and} \quad s_2^{\,2} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$$

The t_{cal} statistic follows t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

4.2 Paired t test: This analysis is performed in case of two matched/ paired samples to determine whether two means are equal or not. We can use paired t-test when there is a natural pairing of observations in the samples, such as when individuals/animals in the group are tested twice, before and after experiment.

This test is used to test the hypothesis of equality of means of two dependent samples**.**

$H_0 : \bar{d} = 0$

$H_1 : \bar{d} \neq 0$

Then the test statistic is given by

$$t_{cal} = \frac{\bar{d}}{s_d / \sqrt{n-1}}$$

where $\bar{d} = \frac{\sum d_i}{n}$ and $s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$

The test statistic t_{cal} follows t- distribution with n-1 degrees of freedom.

4.3 Testing the Significance of Correlation Coefficient

Suppose, r is sample correlation coefficient in a sample of n pairs of observations from bivariate normal population and here we want to test

$H_0 : \rho = 0$. There is no correlation between two variables

$H_1 : \rho \neq 0$ Correlation coefficient is significant

The t-statistic is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The t statistic follows the t- distribution with n-2 degrees of freedom.

4.4 Testing the significance of Regression Coefficient

For testing the significance of regression coefficient null and alternative hypothesis are given below

$H_0 : b_1 = 0$ i.e. the slope coefficient is equal to zero

$H_1 : b_1 \neq 0$ i.e. the slope coefficient is not equal to zero.

The t- test statistic is given by

$$t = \frac{b}{s_b}$$

where b is the slope coefficient of the regression line and s_b is the standard error of the slope.

The standard error of the slope is $s_b = \frac{b}{r}\frac{\sqrt{1-r^2}}{\sqrt{n-2}}$

The t statistic follows the t- distribution with n-2 degrees of freedom.

4.5 F Test (Variance Ratio Test)

The F-test was first discovered by Professor R.A. Fisher and later developed by G.W. Snedecor. This test is also known as F- test (in the honor of Fisher). Since the F test is based on the ratios of two variances so this is also called as Variance Ratio Test and is based on Snedecor's F distribution.

When testing the equality of two normal population variances the F statistic used is the ratio of two sample variances and is as follow:

$$F = \frac{s_x^2}{s_y^2}$$

where $s_x^2 = \frac{\sum(x-\bar{x})^2}{n_1 - 1}$ and $s_y^2 = \frac{\sum(y-\bar{y})^2}{n_2 - 1}$

Note: Put the higher value of estimates of sample variance in numerator and smaller value of estimated variance in denominator.

Applications of F- Test

1. To test equality of two population variances.
2. To test the equality of several population means (ANOVA).
3. To test the significance of multiple correlation coefficient.
4. To test the significance of sample correlation ratio.
5. To test the linearity of the regression.

Chi-Square Test (χ^2) Test

The chi-square test is an important statistical test used in data analysis. This test was developed by Prof. Karl Pearson in 1900.

Uses of Chi-Square (χ^2) Test

1. To test the significance of population variance (σ_0^2).
2. To test the goodness of fit of the data.
3. To test the independence of attributes.
4. To test the linkages in backcross.
5. To test equality or homogeneity of more than two (several) population proportions.
6. To test the homogeneity of several population variances (Bartlett's test).
7. To test equality of several population correlation coefficients.

5. Non-parametric Methods

Parametric tests are used when the information about the distribution of the population parameter is known. For example, sample has been drawn from the normal population in case of t-test. But when there is no or few information available about the distribution of the population parameters, non-parametric tests are used. However, non-parametric tests make no or fewer assumptions about the distribution of data. Hence non-parametric tests are called distribution free tests.

5.1 Sign Test: This test is non parametric counterpart of the student t test for single sample. This is used to test the hypothesis that the median (η) of a population has specific value say, η_0.

Null hypothesis $H_0: \eta = \eta_0$

Alternative hypothesis $H_1: \eta \neq \eta_0$

Suppose X_1, X_2, ……, X_n is a random sample of size n taken from a population with median η_0. Now we will perform the sign test in following manner:

1. Subtract the given value of population median η_0 from each value of sample and give plus (+) sign if deviation is positive and give minus (-) sign if deviation is negative.
2. Give zero if deviation is zero.
3. Here we consider only the signs not magnitude of ranks.
4. Drop out the zero signs.

Test Statistic: The test statistic is X defined as the smaller of X+ and X- which are the sums of the positive and negative ranks of the difference scores, respectively.

When the sample size is small **(** $n \leq 25$ **)** then

Test statistic X is given by

X= Minimum (Positive signs, Negative signs)

Decision: If calculated value of test statistic X is less than the table value (critical value) of test at given level of significance, then accept the null hypothesis and otherwise reject it.

when the sample size is large ($n > 25$) then

Test statistic is:

$$Z = \frac{(X + 0.5) - \frac{n}{2}}{\frac{1}{2}\sqrt{n}} \sim N(0,1)$$

Decision: If calculated value of Z is less than the table value of z at given level of significance, then accept the null hypothesis, and otherwise reject it.

5.2 Wilcoxon Signed Ranked Test: The Wilcoxon Signed Ranked test is a nonparametric counterpart of the Paired t-test. The test is used to compare two dependent or matched or paired samples with ordinal continuous data. This test is used to test the hypothesis that the median (η) of a population has specific value say, η_0.

Null hypothesis (H_0): Median difference is zero

Alternative hypothesis (H_1): Median difference is not zero

Suppose there are X_1, X_2, ……, X_n sample of size n taken from a population with median η_0. To calculate the Test statistics W subtract the given value of population median η_0 from each value of sample and give plus (+) sign if deviation is positive, minus (-) sign if deviation is negative and zero if deviation is zero. Drop out the zero signs from the counting of statistic. Here we consider both signs and magnitude of deviations.

Test Statistic: The test statistic is W, defined as the smaller of W+ and W- which are the sums of the positive and negative ranks of the difference scores, respectively.

When Sample Size is Small ($n \leq 30$)

Test statistic W is given by

W = Minimum (no. of plus signs, no. of minus signs)

Decision: If calculated value of test statistic W is less than the table value (critical value) of test at given level of significance, then accept the null hypothesis, and otherwise reject it.

When Sample Size is large ($n > 30$)

Test statistic is given by

$$Z = \frac{W - \mu_W}{\sigma_W}$$

where $\mu_W = \frac{n(n+1)}{4}$ and $\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

W = Smaller of (no. of plus signs, no. of minus signs

n = Total sample size

Decision: If calculated value of Z is less than the table value of Z at given level of significance, then accept the null hypothesis, and otherwise reject it. Reject H_0 if W $\leq$ critical value from table.

5.3 Mann-Whitney -Wilcoxon U Test

The Mann-Whitney-Wilcoxon U test is a nonparametric alternative to independent samples t test. This test is used to compare two different independent groups or conditions or treatments, when the dependent variable is either ordinal or continuous under the assumption that values are not normally distributed. The only assumption is that variable(s) is (are) continuous and sample are random and not less than 10. Here we test whether the two populations differ through determining if there are differences in medians between two groups.

There are following steps: This test uses the sum of the ranks of each sample.

1. Set up hypotheses: H_0: The two populations are equal
   H_1: The two populations are not equal.
2. Combine the observations both samples (X-values) and (Y-values) in to a single sample.
3. Arrange the observations of combined sample in ascending order of magnitude.
4. Assign the ranks to ordered data values.
5. Compute the value of U test statistic
   U = Smaller of (U_x and U_y)

$$U_X = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_X \qquad \text{and} \qquad U_Y = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_Y$$

Where, n_1 = No. of observations in first sample (X-values)

n_2 = No. of observations in second sample (Y-values)

R_x = Sum of ranks of observations of first sample

R_y = Sum of ranks of observations of second sample

6. Reject the null hypothesis if critical value at chosen value of α (tabulated value) of U is less than calculated value of U, otherwise accept it.

For a Large Sample: If sample size is larger than 20, $U \sim N(\mu_u, \sigma_u)$ then use following normal test: $Z = \frac{U - \mu_u}{\sigma_u}$

Where, Mean (μ_u) = $\frac{n_1 n_2}{2}$ and Standard Deviation (σ_u) = $\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1}{12}}$

5.4 Run Test for Testing the Randomness

Run test is a non-parametric statistical method that examines whether a given data set has been randomly selected from a specific distribution. Run test used to test whether the distribution functions $F(x)$ and $G(y)$ of two continuous random variables X and Y are equal or not.

The runs test analyses the occurrence of similar events that are separated by other events that are different from them. Thus this tests the randomness (independency) of data. A run is defined as a sequence of similar or like events, items or symbols that is preceded by and followed by an event, item or symbol of a different type, or by none at all.

A run has two characteristics, number of runs and the length of run.

Procedure:

1. Assume that the data available for the analysis can categorize into two mutually exclusive types.
2. Determine the total sample size (n), number of observation of each type (n_1 = the number of observation of one type and n_2 = the number of observations of the other type).
3. State the null and alternate hypothesis

   H_0: Pattern of occurrence of items is random

   H_1: The pattern of occurrence is not random

4. Calculate the test statistic (r)
   r = total number of runs
5. Determine the critical value from the statistical table using n_1 and n_2.
6. Take a decision about acceptance or rejection of null hypothesis.

   If calculated value of test statistic is less than the value of lower critical limits or more than value of upper critical limit, reject the null hypothesis otherwise accept.

Run Test for Large Sample

Suppose there are n_1 elements of one type and n_2 of the other, where $n_1 \geq n_2$ and n_1 is large enough (approximately $n_1 > 20$). Suppose further there are *r* runs.

$$Z = \frac{r - \mu_r}{\sigma_r}$$

where: *r* is the number of runs, μ_r is the expected number of runs and

σ_r is the standard deviation of the number of runs.

Then based on the null hypothesis H_0 that the order is random, *r* has an approximately normal distribution $N(\mu,\sigma)$.

The values of μ_r and σ_r are computed as follows

$$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1 \qquad \sigma_r = \sqrt{\frac{2n_1n_2\left(2n_1n_2 - n_1 - n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)}}$$

Run test is a statistical test that is used to know the randomness in data.

5.5 Kruskal -Wallis Test: This test is analogue of the ANOVA for one way classified data. It is used to compare more than two independent groups with ordinal data with respect to their identicalness (distributions of populations). This test compares medians among the k comparison groups (k > 2) and is sometimes described as One Way ANOVA with the ranked data. It is an extension of the Mann-Whitney test to situations where more than two levels/populations are involved.

Assumptions

1. The population need not be normal.
2. There are minimum three independent groups or samples having minimum five observations in each sample.

H_0 : The medians of k populations are equal

H_1 : At least medians of two populations are not equal

The procedure for the test involves pooling the observations from the k samples into one combined sample and then ranking the observations in ascending order from lowest to highest, i.e., from 1 to N. Where N is the total sample size, $N = n_1 + n_2 + n_3 + ...n_k$., k is the number of comparison groups, R_j is the sum of the ranks in the i-th group or sample and n_i is the sample size in the i-th sample, then the test statistic H is given by:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(\frac{R_i^2}{n_i} \right) - 3(N+1)$$

or $$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} + ... + \frac{R_k^2}{n_k} \right) - 3(N+1)$$

Test statistics is approximately distributed as chi-square variate with (k -1) d.f.

Decision Rule: Reject the null hypothesis H_0 if calculated value of H is more or equal to tabulated value of Chi -Square for α level of significance at (k-1) degrees of freedom otherwise accept it.

5.6 Friedman Test: The Friedman test is the non-parametric alternative to the ANOVA with repeated measures. It is used to determine differences among the three or more measurements from the same group of subjects are significantly different from each other or not, when the dependent variable is measured on ordinal scale. Here assumption is that data is random and continuous that has violated the assumptions of normality.

Procedure to Conduct Friedman Test

1. Rank each row (block) together and independently of the other rows. When there are ties, give the average ranks of the observations.
2. Find the sum of the ranks for each column (treatments) and then sum of the squared rank total.
3. Compute the Friedman test statistic

$$F = \frac{12}{rk(k+1)} \sum_{j=1}^{k} R_j^2 - 3r(k+1)$$

where r = no. of rows (blocks) and k = no. of treatments (groups or columns)

Test statistics is approximately distributed as chi-square variate with (k -1) degrees of freedom.

4. Determine critical value from chi-square distribution table with k-1 degrees of freedom.
5. Decision and conclusion: Accept the null hypothesis if calculated value is less than tabulated value of Chi-Square for α level of significance at (k-1) degrees of freedom otherwise reject it.

5.7 Kolmogorov- Smirnov Test: The Kolmogorov–Smirnov test is a nonparametric goodness-of-fit test. This test is based on empirical distribution and compare empirical distribution of a sample of any value of x with the Cumulative Distribution Function (CDF) of the population. The empirical distribution function $S_n(x)$ is the fraction of sample values that are equal to or less than *x.* This test can also be used to test whether a sample comes from a population that is normally distributed. It makes better use of available data than Chi-square test and compares an empirical distribution with cumulative distribution function for a variable. In this the null hypothesis assumes that there is no difference between the observed and theoretical distribution.

H_0: Data follow a specified distribution, $F(x) = F_0(x)$

H_1: Data don't follow a specified distribution, $F(x) \neq F_0(x)$

The K-S test statistic measures the largest distance between the empirical distribution function $S_n(x)$ and the hypothesized or theoretical distribution function $F_0(x)$.

The value of test statistic 'D' is calculated as:

$$D = \underset{\text{overall x}}{\text{Max}} \left| F_0(x) - S_n(x) \right|$$

where, $S_n(x)$ is empirical distribution of the observed data

$S_n(x)$ =k/n =(No. of Observations ≤ X)/(Total No. of Observations).

$F_0(x)$ = CDF of the population

The critical value of D is found from the K-S table values for one sample test.

Decision Criteria: If calculated value is less than tabulated value, $D \leq D_{n,\alpha}$ accept null hypothesis otherwise reject.

5.8 Spearman's Rank Correlation (r_s): Spearman's rank correlation coefficient can be defined as a special case of Pearson's correlation coefficient applied to ranked or sorted variables. Unlike Pearson, Spearman's correlation is not restricted to linear relationships. Instead, it measures monotonic association (only strictly increasing or decreasing but not mixed) between two variables and relies on the rank order of values. In other words, rather than comparing means and variances, Spearman's coefficient looks at the relative order of values for each variable. This makes it appropriate to use with both continuous and discrete data. The formula for computing Spearman's Rank Correlation Coefficient (r_s) is given below.

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

where, d = Difference of ranks
n = Number of observations

The value of Spearman's rank correlation coefficient lies between -1 and +1.

CORRELATION AND LINEAR REGRESSION ANALYSIS

Pankaj Das

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction:

Correlation and regression analysis are fundamental statistical techniques used to measure relationships between variables. These methods are widely applied in various fields such as economics, finance, biology, and social sciences to understand and quantify dependencies among different factors.

Correlation analysis helps in determining the strength and direction of a relationship between two variables, indicating whether they move together or in opposite directions. However, it does not imply causation. On the other hand, regression analysis not only assesses the relationship but also provides a mathematical model to predict one variable based on another. This makes regression an essential tool in predictive analytics and decision-making processes.

By understanding correlation and regression, analysts can make informed conclusions about data trends, develop predictive models, and enhance decision-making strategies in both academic research and practical applications.

Correlation Analysis

Correlation is a statistical technique to ascertain the association or relationship between two or more variables. Correlation analysis is a statistical technique to study the degree and direction of relationship between two or more variables. A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.

Uses of correlations:

1. Correlation analysis helps inn deriving precisely the degree and the direction of such relationship.

2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.

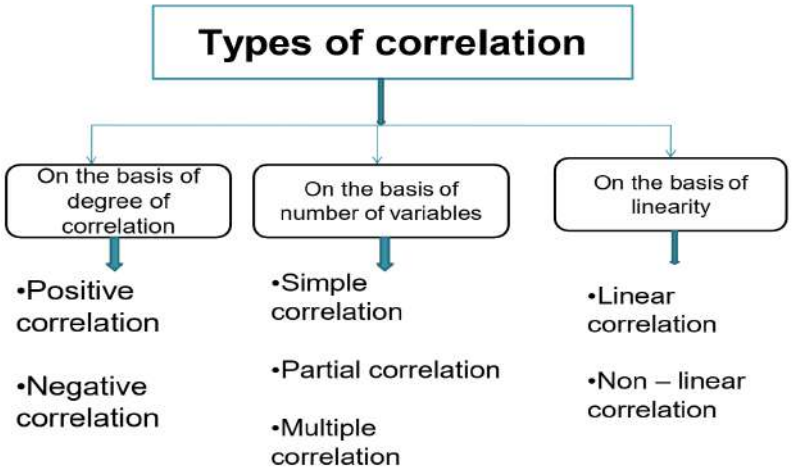
3. Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing farces may become effective.

4. Economic theory and business studies show relationships between variables like price and quantity demanded advertising expenditure and sales promotion measures etc.

5. The measure of coefficient of correlation is a relative measure of change.

Types of Correlation:

Correlation is described or classified in several different ways. Three of the most important are:



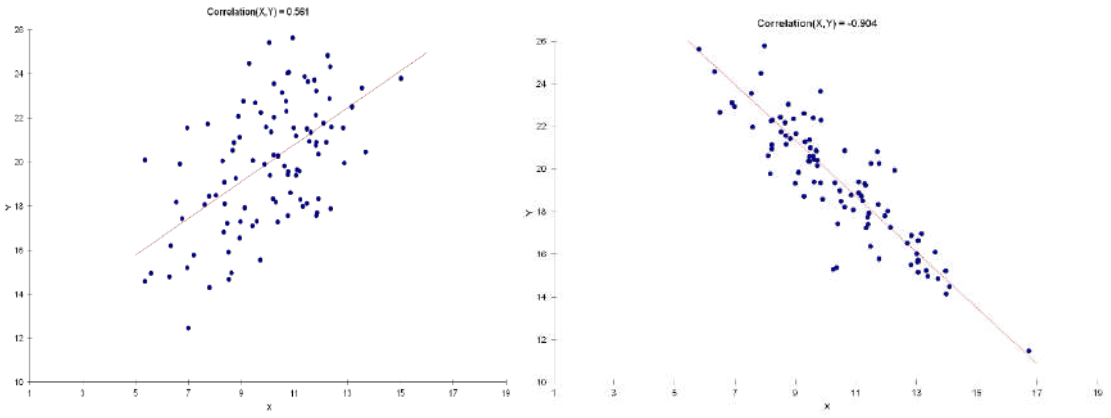
I. Correlation on the basis of degree: Whether correlation is positive (direct) or negative (in-versa) would depend upon the direction of change of the variable.

Positive Correlation: If both the variables vary in the same direction, correlation is said to be positive. It means if one variable is increasing, the other on an average is also increasing or if one variable is decreasing, the other on an average is also deceasing, then the correlation is said to be positive correlation. For example, the correlation between heights and weights of a group of persons is a positive correlation.

Height (cm) : X	158	160	163	166	168	171	174	176
Weight (kg) : Y	60	62	64	65	67	69	71	72

Negative Correlation: If both the variables vary in opposite direction, the correlation is said to be negative. If means if one variable increases, but the other variable decreases or if one variable decreases, but the other variable increases, then the correlation is said to be negative correlation. For example, the correlation between the price of a product and its demand is a negative correlation.

Price of Product (Rs. Per Unit) : X	6	5	4	3	2	1
Demand (In Units) : Y	75	120	175	250	215	400



a. Positive correlation b. negative correlation

Zero Correlation: Actually it is not a type of correlation but still it is called as zero or no correlation. When we don't find any relationship between the variables then, it is said to be zero correlation. It means a change in value of one variable doesn't influence or change the value of other variable. For example, the correlation between weight of person and intelligence is a zero or no correlation.

II. Correlation on the basis of number of variables: The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation: When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks secured by student and the attendance of student in class, it is a problem of simple correlation.

Partial Correlation: In case of partial correlation one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in above example of relationship between student marks and attendance, the other variable influencing such as effective teaching of teacher, use of teaching aid like computer, smart board etc. are assumed to be constant.

III. Correlation on the basis of linearity: Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

Linear Correlation: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear. If such variables are plotted on a graph paper all the plotted points would fall on a straight line. For example: If it is assumed that, to produce one unit of finished product we need 10 units of raw materials, then subsequently to produce 2 units of finished product we need double of the one unit.

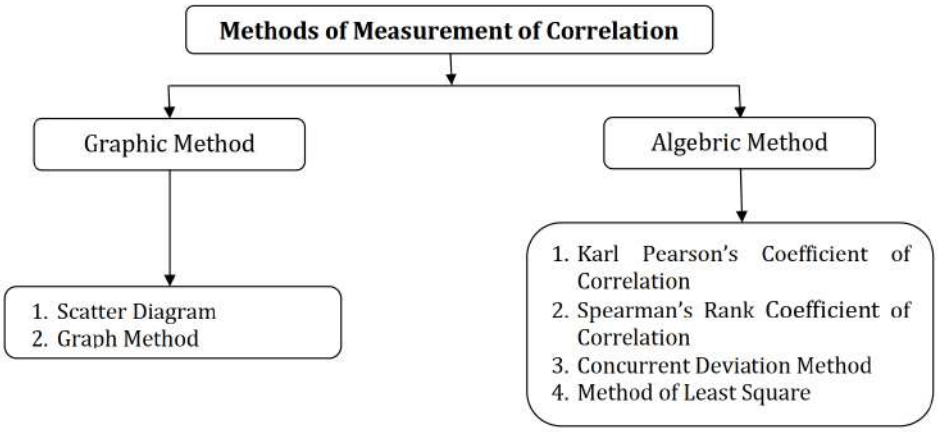
Raw material : X	10	20	30	40	50	60
Finished Product : Y	2	4	6	8	10	12

Non-linear Correlation: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear. If such variables are plotted on a graph, the points would fall on a curve and not on a straight line. For example, if we double the amount of advertisement expenditure, then sales volume would not necessarily be doubled.

Advertisement Expenses : X	10	20	30	40	50	60
Sales Volume : Y	2	4	6	8	10	12

Methods of measurement of correlation:

Quantification of the relationship between variables is very essential to take the benefit of study of correlation. For this, we find there are various methods of measurement of correlation, which can be represented as given below:



1. Scatter plot:

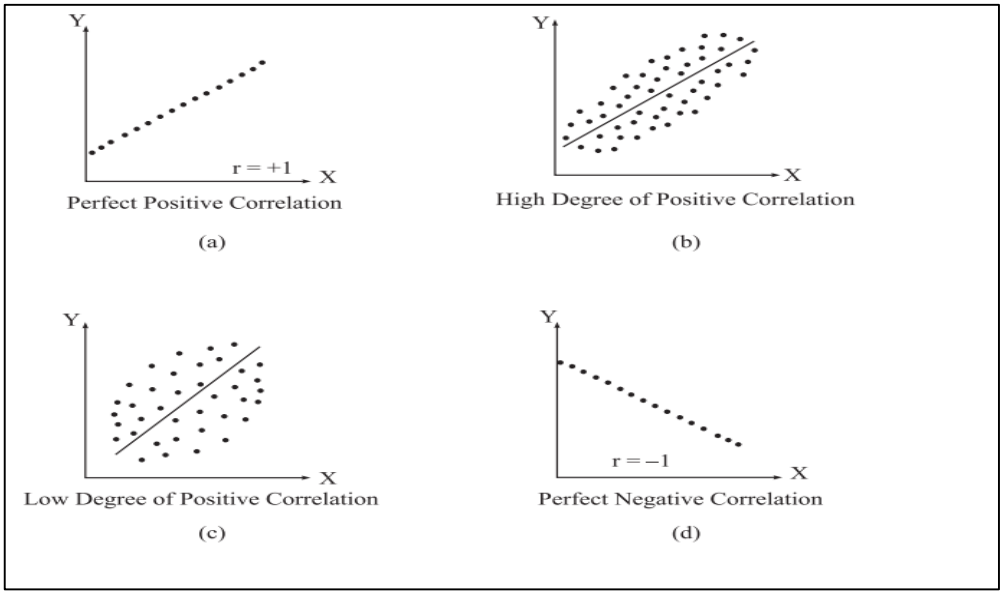
Scatter Plots (also called scatter diagrams) are used to graphically investigate the possible relationship between two variables without calculating any numerical value. In this method, the values of the two variables are plotted on a graph paper. One is taken along the horizontal (X-axis) and the other along the vertical (Y-axis). By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'Scatter Plot'. The manner in which these points are scattered, suggest the degree and the direction of correlation. The degree of correlation is denoted by 'r' and its direction is given by the signs positive and negative.

A scatter diagram reveals whether the movements in one series are associated with those in the other series.

• Perfect Positive Correlation: In this case, the points will form on a straight line falling from the lower left hand corner to the upper right hand corner.

• Perfect Negative Correlation: In this case, the points will form on a straight line rising from the upper left hand corner to the lower right hand corner.

• High Degree of Positive Correlation: In this case, the plotted points fall in a narrow band, wherein points show a rising tendency from the lower left hand corner to the upper right hand corner.

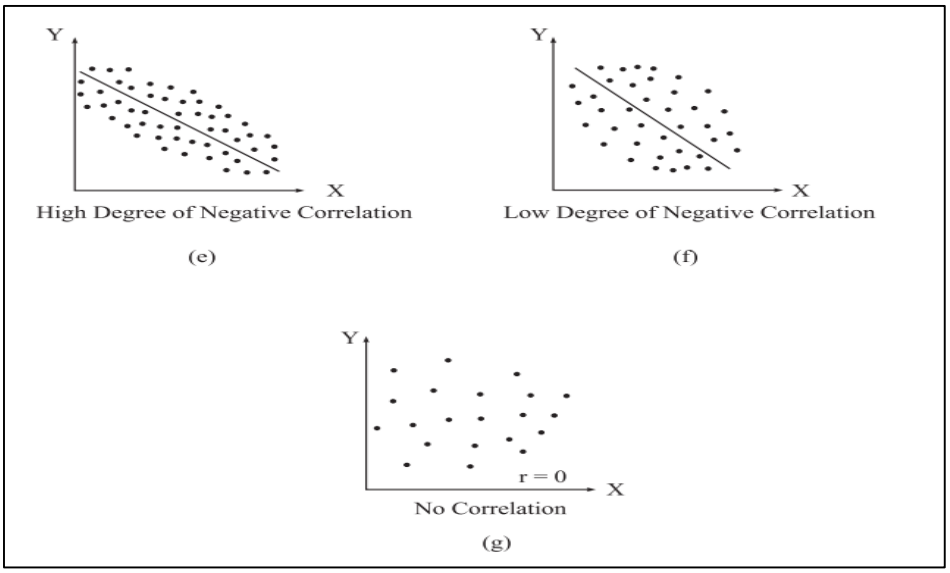


•High Degree of Negative Correlation: In this case, the plotted points fall in a narrow band, wherein points show a declining tendency from upper left hand corner to the lower right hand corner.

• Low Degree of Positive Correlation: If the points are widely scattered over the diagrams, wherein points are rising from the left hand corner to the upper right hand corner.

• Low Degree of Negative Correlation: If the points are widely scattered over the diagrams, wherein points are declining from the upper left hand corner to the lower right hand corner.

• Zero (No) Correlation: When plotted points are scattered over the graph haphazardly, then it indicates that there is no correlation or zero correlation between two variables.



2. **Karl Pearson's coefficient of correlation:**

Karl Pearson's method of calculating coefficient of correlation is based on the covariance of the two variables in a series. This method is widely used in practice and the coefficient of correlation is denoted by the symbol "r". If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship of correlation.

$$r = \frac{Covariance\ (x,y)}{S.D.(x)S.D.(y)}$$

$$r = \frac{Cov(X, Y)}{\sigma_x\ \sigma_y}$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad \text{where } X = x - \bar{x},\ Y = y - \bar{y}$$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{(Y - \bar{Y})^2}}$$

Where, $\bar{X}$ = mean of X variable
$\bar{Y}$ = mean of Y variable

$$r = \frac{\sum f(dx)(dy) - \frac{\sum fdx)(\sum fdy)}{N}}{\sqrt{\sum(fdx)\wedge 2 - \frac{(\sum fdx)\wedge 2}{N}}\sqrt{\sum(dy)\wedge 2 - \frac{(\sum fdy)\wedge 2}{N}}} \qquad d_x = X - A \qquad d_y = Y - A$$

Illustration 01: From following information find the correlation coefficient between advertisement expenses and sales volume using Karl Pearson's coefficient of correlation method.

Firm	1	2	3	4	5	6	7	8	9	10
Advertisement Exp. (Rs. In Lakhs)	11	13	14	16	16	15	15	14	13	13
Sales Volume (Rs. In Lakhs)	50	50	55	60	65	65	65	60	60	50

Solution: Let us assume that advertisement expenses are variable X and sales volume are variable Y. Calculation of Karl Pearson's coefficient of correlation

Firm	X	Y	$x = X - \dot{X}$	x^2	$y = Y - \dot{Y}$	y^2	xy
1	11	50	-3	9	-8	64	24
2	13	50	-1	1	-8	64	8
3	14	55	0	0	-3	9	0
4	16	60	2	4	2	4	4
5	16	65	2	4	7	49	14
6	15	65	1	1	7	49	7
7	15	65	1	1	7	49	7
8	14	60	0	0	2	4	0
9	13	60	-1	1	2	4	-2
10	13	50	-1	1	-8	64	8
	140	**580**		**22**		**360**	**70**
	ΣX	**ΣY**		**Σx2**		**Σy2**	**Σxy**

$$\dot{X} = \frac{\Sigma X}{n} = \frac{140}{10} = 14 \qquad \dot{Y} = \frac{\Sigma Y}{n} = \frac{580}{10} = 58$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad \text{where} \quad X = x - \bar{x}, \; Y = y - \bar{y}$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x2 \, \Sigma y2}} = \frac{70}{\sqrt{22 * 360}} = \frac{70}{88.9944} = \underline{0.7866}$$

Interpretation: From the above calculation it is very clear that there is high degree of **positive correlation** i.e. r = 0.7866, between the two variables. i.e. Increase in advertisement expenses leads to increased sales volume.

3. Spearman's Rank Coefficient of Correlation:

When quantification of variables becomes difficult such beauty of female, leadership ability, knowledge of person etc, then this method of rank correlation is useful which was developed by British psychologist Charles Edward Spearman in 1904. In this method ranks are allotted to each element either in ascending or descending order. The correlation coefficient between these allotted two series of ranks is popularly called as "Spearman's Rank Correlation" and denoted by "R".

To find out correlation under this method, the following formula is used.

$R = 1 - \frac{6\sum D^2}{N^3 - N}$ where D =Difference of the ranks between paired items in two series.

N = Number of pairs of ranks

In case of tie in ranks or equal ranks:

In some cases it may be possible that it becomes necessary to assign same rank to two or more elements or individual or entries. In such situation, it is customary to give each individual or entry an average rank. For example, if two individuals are ranked equal to 5th place, then both of them are allotted with common rank (5+6)/2 = 5.5 and if three are ranked in 5th place, then they are given the rank of (5+6+7)/3 = 6. It means where two or more individuals are to be ranked equal, the rank assigned for the purpose of calculating coefficient of correlation is the average of the ranks with these individual or items or entries would have got had they differed slightly with each other.

Where equal ranks are assigned to some entries, an adjustment factor is to be added to the value of 6∑D2 in the above formula for calculating the rank coefficient correlation. This adjustment factor is to be added for every repetition of rank.

$$\text{Adjustment factor } = \frac{1}{12}(m_1^3 - m_1)$$

where, m = number of items whose rank are common. For example, if a particular rank repeated two times then m=2 and if it repeats three times then m= 3 and so on.

Hence the above formula can be re-written as follows:

$$\mathbf{R = 1 - \frac{6 * [\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{N^3 - N}}$$

Illustration 02:

Find out spearman's coefficient of correlation between the two kinds of assessment of graduate students' performance in a college.

Name of students	A	B	C	D	E	F	G	H	I
Internal Exam	51	68	73	46	50	65	47	38	60
External Exam	49	72	74	44	58	66	50	30	35

Solution:

Calculation of Spearman's Rank Coefficient of Correlation

Name	**Internal Exam**	**Ranks (R1)**	**External Exam**	**Ranks (R2)**	**D = R1 - R2**	**D^2**
A	51	5	49	6	-1	1
B	68	2	72	2	0	0
C	73	1	74	1	0	0
D	46	8	44	7	1	1
E	50	6	58	4	2	4
F	65	3	66	3	0	0
G	47	7	50	5	2	4
H	36	9	30	9	0	0
I	60	4	35	8	-4	16
					$\sum D^2$ =	**26**

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6*26}{9^3 - 9} = 1 - \frac{156}{729 - 9} = 1 - \frac{156}{720} = 1 - 0.2167 = \underline{0.7833}$$

Interpretation: From the above calculation it is very clear that there **is high degree of positive correlation** i.e. **R = 0.7833**, between two exams. It means there is a high degree of positive correlation between the internal exam and external exam of the students.

Properties of Coefficient of Correlation:

1. The coefficient of correlation always lies between – 1 to +1, symbolically it can be written as $-1 \leq r \leq 1$.

2. The coefficient of correlation is independent of change of origin and scale.

3. The coefficient of correlation is a pure number and is independent of the units of measurement. It means if X represent say height in inches and Y represent say weights in kgs, then the correlation coefficient will be neither in inches nor in kgs but only a pure number.

4. The coefficient of correlation is the geometric mean of two regression coefficient, symbolically $r = \sqrt{b_{xy} * b_{yx}}$

5. If X and Y are independent variables then coefficient of correlation is zero.

Regression analysis:

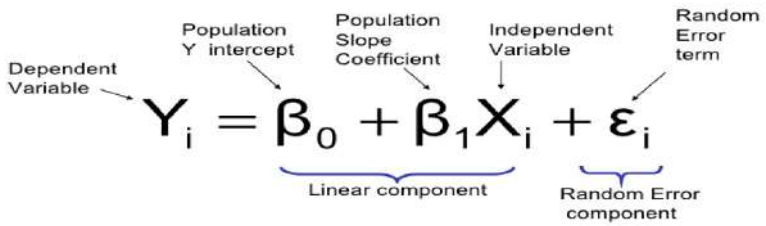
A study of measuring the relationship between associated variables, wherein one variable is dependent on another independent variable, called as Regression. It is developed by Sir Francis Galton in 1877 to measure the relationship of height between parents and their children.

Regression analysis is a statistical tool to study the nature and extent of functional relationship between two or more variables and to estimate (or predict) the unknown values of dependent variable from the known values of independent variable.

The variable that forms the basis for predicting another variable is known as the Independent Variable and the variable that is predicted is known as dependent variable. For example, if we know that two variables price (X) and demand (Y) are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy.

Components in regression analysis:

A typical regression model looks like



Independent/Explanatory/ Regressor /Predictor: The variable which influences the value or is used for prediction. It is denoted as "x".

Dependent/regressed/ Explained variable: The variable whose value is influenced or is to be predicted. It is denoted as "y".

In linear regression equation contains one independent variable, one constant and coefficient (also known as weight) and we are trying to predict dependent variable

Uses of Regression Analysis:

1. It provides estimates of values of the dependent variables from values of independent variables.

2. It is used to obtain a measure of the error involved in using the regression line as a basis for estimation.

3. With the help of regression analysis, we can obtain a measure of degree of association or correlation that exists between the two variables.

4. It is highly valuable tool in economies and business research, since most of the problems of the economic analysis are based on cause and effect relationship.

Regression Lines and Regression Equation:

Regression lines and regression equations are used synonymously. Regression equations are algebraic expression of the regression lines. Let us consider two variables: X & Y. If y depends on x, then the result comes in the form of simple regression. If we take the case of two variable X and Y, we shall have two regression lines as the regression line of X on Y and regression line of Y on X. The regression line of Y on X gives the most probable value of Y for given value of X and the regression line of X on Y given the most probable value of X for given value of Y. Thus, we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression line will coincide, i.e. we will have one line. If the variables are independent, r is zero and the lines of regression are at right angles i.e. parallel to X axis and Y axis.

Therefore, with the help of simple linear regression model we have the following two regression lines

1. Regression line of Y on X: This line gives the probable value of Y (Dependent variable) for any given value of X (Independent variable).

Regression line of Y on X	:	$Y - \dot{Y} = b_{yx}\,(X - \dot{X})$
OR	:	$Y = a + bX$

2. Regression line of X on Y: This line gives the probable value of X (Dependent variable) for any given value of Y (Independent variable).

Regression line of X on Y	:	$X - \dot{X} = b_{xy}\,(Y - \dot{Y})$
OR	:	$X = a + bY$

In the above two regression lines or regression equations, there are two regression parameters, which are "a" and "b". Here "a" is unknown constant and "b" which is also denoted as "b_{yx}" *or* "b_{xy}", is also another unknown constant popularly called as regression coefficient. Hence, these "a" and "b" are two unknown constants (fixed numerical values) which determine the position of the line completely. If the value of either or both of them is changed, another line is determined. The parameter "a" determines the level of the fitted line (i.e. the distance of the line directly above or

below the origin). The parameter "b" determines the slope of the line (i.e. the change in Y for unit change in X).

If the values of constants "a" and "b" are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of least squares. With the little algebra and differential calculus, it can be shown that the following two normal equations, if solved simultaneously, will yield the values of the parameters "a" and "b".

Two normal equations:

X on Y	Y on X
$\sum X = Na + b\sum Y$ $\sum XY = a\sum Y + b\sum Y^2$	$\sum Y = Na + b\sum X$ $\sum XY = a\sum X + b\sum X^2$

This above method is popularly known as direct method, which becomes quite cumbersome when the values of X and Y are large. This work can be simplified if instead of dealing with actual values of X and Y, we take the deviations of X and Y series from their respective means.

In that case:

Regression equation Y on X:

$Y = a + bX$ will change to $\left(Y - \dot{Y}\right) = b_{yx}\left(X - \dot{X}\right)$

Regression equation X on Y:

$X = a + bY$ will change to $\left(X - \dot{X}\right) = b_{xy}\left(Y - \dot{Y}\right)$

In this new form of regression equation, we need to compute only one parameter i.e. "b". This "b" which is also denoted either "b_{yx}" *or* "b_{xy}" which is called as regression coefficient.

Regression Coefficient: The quantity "b" in the regression equation is called as the regression
coefficient or slope coefficient. Since there are two regression equations, therefore, we have two regression coefficients.

1. Regression Coefficient X on Y, symbolically written as "bxy"
2. Regression Coefficient Y on X, symbolically written as "byx

Different formula's used to compute regression coefficients:

Method	Regression Coefficient X on Y	Regression Coefficient Y on X
Using the correlation coefficient (r) and standard deviation (σ)	$b_{xy} = r\frac{\sigma x}{\sigma y}$	$b_{yx} = r\frac{\sigma y}{\sigma x}$
Direct Method: Using sum of X and Y	$b_{xy} = \frac{N\sum XY - \sum X\sum Y}{N\sum Y^2 - (\sum Y)^2}$	$b_{yx} = \frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - (\sum X)^2}$
When deviations are taken from arithmetic mean	$b_{xy} = \frac{\sum xy}{\sum y^2}$ where $x = X - \dot{X}$ and $y = Y - \dot{Y}$	$b_{yx} = \frac{\sum xy}{\sum x^2}$ where $x = X - \dot{X}$ and $y = Y - \dot{Y}$

Assumptions of Linear Regression

- Linearity: The relationship between X and Y is linear.
- Independence: Observations are independent.
- Homoscedasticity: Constant variance of residuals.
- Normality: Residuals are normally distributed.

Properties of Regression Coefficients:

1. The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically $r = \sqrt{b_{xy} * b_{yx}}$
2. If one of the regression coefficients is greater than unity, the other must be less than unity, since the value of the coefficient of correlation cannot exceed unity. For example, if $b_{xy} = 1.2$ and $b_{yx} = 1.4$ "r" would be = $\sqrt{1.2 * 1.4} = 1.29$, which is not possible.
3. Both the regression coefficient will have the same sign. i.e. they will be either positive or negative. In other words, it is not possible that one of the regression coefficients are having minus sign and the other plus sign.
4. The coefficient of correlation will have the same sign as that of regression coefficient, i.e. if regression coefficient has a negative sign, "r" will also have negative sign and if the regression coefficient has a positive sign, "r" would also be positive. For example, if $b_{xy} = -0.2$ and $b_{yx} = -0.8$ then $r = -\sqrt{0.2 * 0.8} = -0.4$
5. The average value of the two regression coefficient would be greater than the value of coefficient of correlation. In symbol $(b_{xy} + b_{yx}) / 2 > r$. For example, if $b_{xy} = 0.8$ and $b_{yx} = 0.4$ then average of the two values = (0.8 + 0.4) / 2 = 0.6 and the value of r = r = $\sqrt{0.8 * 0.4} = 0.566$ which less than 0.6.
6. Regression coefficients are independent of change of origin but not scale.

Illustration 03:

After investigation it has been found the demand for automobiles in a city depends mainly, if not entirely, upon the number of families residing in that city. Below are the given figures for the sales of automobiles in the five cities for the year 2019 and the number of families residing in those cities

City	No. of Families (in lakhs): X	Sale of automobiles (in '000): Y
Belagavi	70	25.2
Bangalore	75	28.6
Hubli	80	30.2
Kalaburagi	60	22.3
Mangalore	90	35.4

Fit a linear regression equation of Y on X by the least square method and estimate the sales for the year 2020 for the city Belagavi which is estimated to have 100 lakh families assuming that the same relationship holds true.

Solution:

Calculation of Regression Equation

City	**X**	**Y**	**X^2**	**XY**

Belagavi	70	25.2	4900	1764
Bangalore	75	28.6	5625	2145
Hubli	80	30.2	6400	2416
Kalaburagi	60	22.3	3600	1338
Mangalore	90	35.4	8100	3186
	375	**141.7**	**28,625**	**10,849**
	$\sum X$	**$\sum Y$**	**$\sum X^2$**	**$\sum XY$**

Regression equation of Y on X: $Y = a + bX$

The two normal equations are:

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values in above normal equations, we get

$$141.7 = 5a + 375*b \quad \text{..... (i)}$$

$$10849 = 375*a + 28625*b \quad \text{..... (ii)}$$

Let us solve these equations (i) and (ii) by simultaneous equation method. Multiply equation (i) by 75 we get 10627.5 = 375a + 28125*b

Now rewriting these equations:

10627.5	=	375a	+	28125b
10849	=	375a	+	28625b
(-)		(-)		(-)
-221.5	=			-500b

Therefore, now we have -221.5 = -500*b, this can be rewritten as 500*b = 221.5

Now, b = 0.443

Substituting the value of b in equation (i), we get,

141.7	=	5a	+	(375 * 0.443)
141.7	=	5a	+	166.125
5a	=	141.7	-	166.125
5a	=	-24.425		
a	=	-24.425/5		
a	=	-4.885		

Thus we got the values of a = -4.885 and b = 0.443

Hence, the required regression equation of Y on X:

$$Y = a + bX \Rightarrow Y = -4.885 + 0.443X$$

Estimated sales of automobiles (Y) in city Belagavi for the year 2020, where number of families (X) are 100(in lakhs):

$$Y = -4.885 + 0.443X$$

$$Y = -4.885 + (0.443 * 100)$$

$$Y = -4.885 + 44.3$$

$$Y = 39.415 (‘000)$$

Means sales of automobiles would be 39,415 when number of families are 100,00,000.

Types of regression models:

Linear regression model: A linear regression model is used to depict a relationship between variables which are proportional to each other. Meaning, the dependent variable increases/decreases with the independent variable. In the graphical representation, it has a straight linear line plotted between the variables. Even if the points are not exactly in a straight line (which is always the case) we can still see a pattern and make sense out of it. Example: As the age of a person increases, the level of glucose in their body increases as well.

Multiple regression model: A multiple regression model is used when there is more than one independent variable affecting a dependent variable. While predicting the outcome variable, it is important to measure how each of the independent variables moves in their environment and how their changes will affect the output or target variable. Example: Chances of a student failing their test can be dependent on various input variables like hard work, family issues, health issues, etc.

Non-linear regression model: In the non-linear regression model, the graph doesn't show a linear progression. Depending on how the response variable reacts to the input variable, the line will rise or fall showing the height or depth of the effect of the response variable. To know that a non-linear regression model is the best fit for your scenario, make sure you look into your variables and their patterns. If you see that the response variable is showing not so constant output to the input variable, you can choose to use a non-linear model for your problem. Example: A patient's response to treatment can be good or bad depending on their body tendency and willpower.

Distinction between Correlation and Regression

Sl No	Correlation	Regression
1	It measures the degree and direction of relationship between the variables.	It measures the nature and extent of average relationship between two or more variables in terms of the original units of the data
2	It is a relative measure showing association between the variables.	It is an absolute measure of relationship.
3	Correlation Coefficient is independent of change of both origin and scale.	Regression Coefficient is independent of change of origin but not scale.
4	Correlation Coefficient is independent of units of measurement.	Regression Coefficient is not independent of units of measurement.
5	Expression of the relationship between the variables ranges from –1 to +1.	Expression of the relationship between the variables may be in any of the forms like: $Y = a + bX$ $Y = a + bX + cX2$
	It is not a forecasting device.	It is a forecasting device which can be used to predict the value of dependent variable from the given value.

	There may be zero correlation such as weight of wife and income of husband.	There is nothing like zero regression independent variable.

Conclusion: In correlation analysis, when we are keen to know whether two variables under study are associated or correlated and if correlated what is the strength of correlation. The best measure of correlation is proved by Karl Pearson's Coefficient of Correlation. However, one severe limitation of this method is that it is applicable only in case of a linear relationship between two variables. If two variables say X and Y are independent or not correlated, then the result of correlation coefficient is zero.

Correlation coefficient measuring a linear relationship between the two variables indicates the amount of variation one variable accounted for by the other variable. A better measure for this purpose is provided by the square of the correlation coefficient, known as "coefficient of determination". This can be interpreted as the ratio between the explained variance to total variance:

$r^2 = \frac{\text{Explained variance}}{\text{Total variance}}$ Similarly, Coefficient of non-determination = $(1 - r^2)$.

Regression analysis is concerned with establishing a functional relationship between two variables and using this relationship for making future projection. This can be applied, unlike correlation for any type of relationship linear as well as curvilinear. The two lines of regression coincide i.e. become identical when $r = -1$ or $+1$ in other words, there is a perfect negative or positive correlation between the two variables under discussion if r = 0, then regression lines are perpendicular to each other.

References

- Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). *Introduction to Linear Regression Analysis*.
- Gujarati, D.N. (2011). *Basic Econometrics*.
- Freedman, D.A. (2009). *Statistical Models: Theory and Practice*.

DIAGNOSTIC AND RESIDUAL MEASURES IN REGRESSION ANALYSIS

Achal Lama and K N Singh

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012

1. Overview and Definition

Regression analysis is a fundamental statistical technique used to model and analyze the relationships between dependent and independent variables. It is widely used in various fields such as economics, finance, machine learning, and social sciences to predict outcomes and understand relationships between variables.

Regression analysis helps in:

1. Understanding the relationship between one or more independent variables and a dependent variable.
2. Predicting the dependent variable based on the values of independent variables.
3. Identifying trends and patterns in data.
4. Evaluating the strength of relationships between variables.
5. Making data-driven decisions.

There are several types of regression analyses, including:

Linear Regression – Establishes a linear relationship between independent and dependent variables. The simplest form of regression analysis involves a single independent variable and is expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y is the dependent variable,
- x is the independent variable,
- β_0 is the y -intercept,
- β_1 is the slope coefficient,
- ϵ is the error term.

Multiple Linear Regression

Extends simple linear regression to multiple independent variables, expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where $x_1, x_2, \ldots x_n$ are independent variables and $\beta_0, \beta_1, \ldots, \beta_n$ are corresponding coefficients.

Key Assumptions of Linear Regression

1. **Linearity** – The relationship between dependent and independent variables is linear.
2. **Independence** – Observations are independent of each other.

3. **Homoscedasticity** – The variance of residuals is constant across all levels of independent variables.
4. **Normality** – Residuals should be normally distributed.
5. **No Multicollinearity** – Independent variables should not be highly correlated with each other.

Model Evaluation Metrics

To assess the performance of a regression model, the following metrics are commonly used:

Mean Squared Error (MSE): $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

R-squared (R^2) Score: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$

Where $\hat{y}_i$ is the predicted value and $\bar{y}_i$ is the mean of observed values.

2. Residual Diagnostics in Regression Analysis

Residual diagnostics in regression analysis is a crucial process used to evaluate whether a regression model meets its underlying assumptions and accurately describes the relationship between the dependent and independent variables. Residuals, defined as the differences between observed and predicted values, play a significant role in assessing model adequacy. The residual for an individual observation is given by:

$$e_i = y_i - \hat{y}_i$$

where:

- e_i is the residual (error term),
- y_i is the actual observed value, and
- $\hat{y}_i$ is the predicted value from the regression model.

Residual diagnostics involve analyzing these residuals using various techniques to identify issues such as non-linearity, heteroscedasticity (non-constant variance), autocorrelation, and violations of normality. If residuals exhibit systematic patterns, it suggests that the model may not be properly specified, and improvements or transformations might be necessary.

Importance of Residual Diagnostics

Residual diagnostics are essential for validating a regression model and ensuring its reliability for prediction and inference. Some key reasons why residual diagnostics are important include:

- **Checking Linearity Assumption**
  Regression models, especially linear regression, assume a linear relationship between independent and dependent variables. If residuals show a systematic pattern (e.g., a curve), it indicates a potential non-linear relationship, suggesting that polynomial terms or transformations may be required.
- **Detecting Heteroscedasticity (Variance Issues)**
  Homoscedasticity (constant variance of residuals) is a key assumption in regression. If residual variance changes with predicted values (e.g., forming a funnel shape), it

suggests heteroscedasticity, which can lead to inefficient and biased parameter estimates.

- **Ensuring Normality of Residuals**
  Many inferential techniques (e.g., hypothesis tests and confidence intervals) rely on the assumption that residuals follow a normal distribution. Skewed or heavy-tailed residual distributions indicate potential violations, affecting statistical inference.
- **Identifying Autocorrelation**
  In time-series and sequential data, residuals should not be correlated with one another. The presence of autocorrelation (systematic patterns in residuals over time) violates this assumption, requiring correction through techniques like differencing or autoregressive modelling.
- **Detecting Outliers and Influential Observations**
  Outliers and highly influential data points can disproportionately affect model estimates. Residual analysis helps in identifying these points, allowing analysts to decide whether to remove or adjust them.
- **Improving Model Fit and Predictive Accuracy**
  A well-diagnosed regression model ensures robust and reliable predictions. Addressing residual issues can lead to a better-fitting model, reducing errors and improving generalizability to new data.

Residual diagnostics act as a quality check in regression analysis, ensuring that the model is statistically sound and interpretable. Without proper residual analysis, incorrect conclusions may be drawn, leading to poor decision-making based on flawed models.

3. Important Residual Diagnostic Statistics in Regression Analysis

Residual diagnostics involve several statistical measures that help assess the validity of a regression model. These statistics help detect issues such as non-linearity, heteroscedasticity, autocorrelation, and influential data points. Below are some key residual diagnostic statistics:

- **Mean of Residuals:** The residuals should have a mean close to zero. A nonzero mean suggests model misspecification, such as missing explanatory variables or incorrect functional form.

$$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$$

where $\varepsilon_i = y_i - \widehat{y_i}$ are the residuals.

- **Standard Deviation of Residuals:** Measures the spread of residuals around their mean (ideally zero). High residual variability indicates poor model fit.

$$\sigma_\varepsilon = \sqrt{\frac{\sum(\varepsilon_i - \overline{\varepsilon}_i)}{n-1}}$$

- **Residual Sum of Squares (RSS):** Represents the total variation in residuals. Lower RSS suggests a better-fitting model.

$$RSS = \sum_{i=1}^{n} (\varepsilon_i)^2$$

- **Durbin-Watson (DW) Statistic:** The **Durbin-Watson (DW) statistic** is a crucial test used in regression analysis to detect **autocorrelation** (serial correlation) in the residuals. Autocorrelation occurs when residuals from one observation are correlated with residuals from another, which violates the assumption of independence in regression models, particularly in time-series data.

Understanding Autocorrelation

- **Positive Autocorrelation** ($0 < DW < 2$): Residuals follow a pattern where an increase in one residual is likely to be followed by another increase. This suggests that the model may be missing important lagged effects.

- **Negative Autocorrelation** ($2 < DW < 4$): Residuals alternate in sign, meaning an increase is often followed by a decrease, and vice versa.

- **No Autocorrelation** ($DW \approx 2$): Residuals are independent, meaning no systematic pattern exists.

Autocorrelation is problematic because it can:

- Lead to **inefficient** and **biased** standard errors.
- Affect hypothesis testing, making p-values unreliable.
- Reduce the predictive power of the model.

The Durbin-Watson statistic is calculated as:

$$DW = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}(\varepsilon_i)^2}$$

where: ε_i are the residuals from the regression model, n is the number of observations.

The numerator measures the squared differences between consecutive residuals, while the denominator accounts for the total squared residuals.

Interpretation of Durbin-Watson Statistic

DW Value	Interpretation
DW≈2	No autocorrelation (ideal condition).
0<DW<2	Positive autocorrelation (common in time-series data).
2<DW<4	Negative autocorrelation (rare but possible).
DW≈0	Strong positive autocorrelation (problematic).
DW≈4	Strong negative autocorrelation (problematic).

A **DW value close to 2** indicates that the residuals are randomly distributed and that there is no significant autocorrelation. Values deviating significantly from 2 suggest issues that need correction.

Suppose we run a regression model and obtain residuals. If we calculate the **Durbin-Watson statistic** and get **DW = 1.1**, this suggests **positive autocorrelation**. To address

this, we might introduce lagged variables or apply the Cochrane-Orcutt correction to improve the model's reliability

How to Handle Autocorrelation

If the Durbin-Watson statistic suggests autocorrelation, you can address it using the following approaches:

- **Add Lagged Variables**: Include past values of dependent or independent variables to account for time-related dependencies.
- **Use Generalized Least Squares (GLS)**: Adjusts the standard regression model to handle correlated errors.
- **Apply the Cochrane-Orcutt Method**: Iteratively estimates and corrects for autocorrelation.
- **Difference the Data**: Subtract the previous observation from the current observation to remove systematic patterns.
- **Use Time-Series Models**: If the data is time-dependent, models like ARIMA (AutoRegressive Integrated Moving Average) can better capture autocorrelation.

The **Durbin-Watson test** is an essential diagnostic tool in regression analysis, particularly for time-series data. It ensures that residuals remain independent, preserving the statistical validity of hypothesis testing and model predictions. If autocorrelation is detected, corrective measures should be applied to improve the regression model's accuracy and efficiency.

- **Breusch-Pagan Test for Heteroscedasticity**

The **Breusch-Pagan (BP) test** is a statistical test used to detect **heteroscedasticity** in a regression model. Heteroscedasticity occurs when the variance of the residuals (errors) is not constant across different levels of the independent variables. This violates a key assumption of **Ordinary Least Squares (OLS) regression**, which assumes that residuals have a constant variance (homoscedasticity).

Heteroscedasticity

In an **ideal regression model**, the residuals should be **homoscedastic**, meaning they exhibit a uniform spread across all levels of the independent variables. **Heteroscedasticity** means that the spread of residuals **changes systematically**, which can lead to inefficient estimates and unreliable hypothesis tests.

Common Causes of Heteroscedasticity:

- **Omitted Variables**: If relevant variables are missing, the model may fail to capture all systematic variations.
- **Non-Linear Relationships**: If the relationship between variables is not linear but is modeled as linear, residual variance may increase.
- **Data with Large Differences in Scale**: If some observations have much higher values than others, residual variance might not be constant.
- **Time-Series Data Issues**: Financial and economic data often show increasing variance over time.

Effects of Heteroscedasticity:

- The **coefficient estimates remain unbiased**, but their **standard errors are incorrect**, leading to misleading statistical inferences.
- Confidence intervals and hypothesis tests (e.g., t-tests and F-tests) become unreliable.
- Prediction accuracy decreases, particularly for extreme values.

- **Breusch-Pagan Test: The Method**

The **Breusch-Pagan test** examines whether the residual variance depends on the independent variables. It does this by testing if the squared residuals are systematically related to one or more independent variables.

Step 1 Fit the Original Regression Model

First, estimate the standard OLS regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where ε are the residuals.

Step 2 Compute the Squared Residuals

Calculate the squared residuals from the regression model:

$$\hat{\varepsilon}_i^{\,2} = (y_i - \hat{y}_i)^2$$

If heteroscedasticity is present, these squared residuals will systematically increase or decrease with the independent variables.

Step 3: Regress the Squared Residuals on Independent Variables

Run the following auxiliary regression:

$$\hat{\varepsilon}_i^{\,2} = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_n x_n + u$$

If the independent variables significantly explain the variance in residuals, heteroscedasticity is likely present.

Step 4: Compute the Breusch-Pagan Test Statistic

The test statistic is given by:

$$BP = n \times R^2$$

where:

- n is the number of observations,
- R^2 is the coefficient of determination from the auxiliary regression.

Step 5: Compare Against the Chi-Square Distribution

The BP statistic follows a **Chi-Square (χ^2) distribution** with degrees of freedom equal to the number of independent variables in the auxiliary regression.

- **Null Hypothesis (H_0)**: The residuals are homoscedastic (constant variance).
- **Alternative Hypothesis (H_a)**: The residuals exhibit heteroscedasticity (non-constant variance).

If the test statistic is **significant (p-value < 0.05)**, we reject $\boldsymbol{H_0}$, indicating that heteroscedasticity is present.

Interpretation of the Breusch-Pagan Test

Test Outcome	**Interpretation**	**Implication**
p>0.05 (Fail to reject H_0)	No evidence of heteroscedasticity	OLS standard errors are reliable.
p<0.05 (Reject H_0)	Significant heteroscedasticity present	OLS standard errors may be incorrect, and adjustments are needed.

How to Correct Heteroscedasticity

If heteroscedasticity is detected, there are several ways to address it:

- Use Robust Standard Errors (Heteroscedasticity-Consistent Errors)

  Use **White's Robust Standard Errors** or **Huber-White standard errors** to adjust for heteroscedasticity.

  These adjustments ensure valid hypothesis tests even when variance is not constant.

- Transform the Dependent Variable

  Apply a **log transformation** or **square root transformation** to stabilize variance:

$$y^* = \log(y)$$

- Weighted Least Squares (WLS)

  Assign weights to observations based on the inverse of the residual variance.

  Helps in giving more weight to low-variance observations.

- Re-specify the Model

  If a non-linear relationship exists, adding polynomial terms or interaction terms might improve the fit.

- Use Generalized Least Squares (GLS)

  GLS modifies OLS to account for heteroscedasticity by transforming variables before estimation.

The **Breusch-Pagan test** is a fundamental diagnostic tool in regression analysis, helping to detect heteroscedasticity. Since heteroscedasticity leads to inefficient and biased standard errors, failing to correct it can result in unreliable hypothesis testing and inaccurate predictions. If heteroscedasticity is found, analysts should consider robust standard errors, transformations, or alternative estimation methods such as WLS or GLS to ensure model reliability.

- **Shapiro-Wilk Test (for Normality of Residuals)**

Checks if residuals follow a normal distribution, which is essential for valid hypothesis testing in regression.

A small p-value (< 0.05) indicates non-normal residuals

- **Q-Q Plot (Quantile-Quantile Plot)**
  - A graphical diagnostic tool comparing residual quantiles to a normal distribution.
  - Deviations from the 45-degree line indicate departures from normality.

Residual diagnostic statistics are essential for verifying regression model assumptions and improving model performance. Evaluating residual behavior helps in detecting misspecifications, improving predictive accuracy, and ensuring reliable statistical inference.

MULTIVARIATE ANALYSIS TECHNIQUES

Rajender Parsad

ICAR-Indian Agricultural Statistics Research Institute, New Delhi - 110 012

1. Introduction

The researchers in biological, physical and social sciences frequently collect measurements on several variables. Generally the data is analyzed by taking one variable at a time. The inferences drawn by analyzing the data for each of the variables may be misleading. This can best be explained from the story of the six blind persons, who tried to describe an elephant after each one touching and feeling a part of it. All of us know that they came out with six different versions of what an elephant was like, each version being partially correct but none was near to reality. Therefore, the data on several variables should be analyzed using multivariate analytical techniques.

Various statistical methods for describing and analyzing these multivariate data sets are Hotelling T^2; Multivariate analysis of variance (MANOVA), Discriminant Analysis, Principal Component Analysis, Factor Analysis, Canonical Correlation Analysis, Cluster Analysis, etc. In this talk, we present an overview of the multivariate analytical techniques.

1. Testing of mean vector - One Sample Case

This is useful for the situations where the data on the different variables are collected and it is required to test whether the sample mean vectors is equal to a specified mean vector. To be specific: Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a random sample of size n is drawn from the population with p-dimensional mean vector $\boldsymbol{\mu}_0$ and based on this sample we want to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

If variance covariance matrix Σ is known or the sample is large, χ^2 test is used.

$$\chi^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

with p degrees of freedom where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j$ is the sample mean calculated from the sample, p is number of variable in the study.

If Σ is not known and sample size is small. Hotelling T^2 is used.

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{s}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j$, $\mathbf{s} = \frac{1}{n-1}\sum_{j=1}^{n} (\mathbf{x_j} - \bar{\mathbf{x}})(\mathbf{x_j} - \bar{\mathbf{x}})'$.

$$\frac{(n-p)}{(n-1)p}T^2 \approx F_{p,\,n-p}.$$

Example 1: {Example 5.2 in Johnson and Wichern, 2002}. Perspiration from 20 healthy females was analyzed. Three components, X_1= sweat rate, X_2 = sodium content and X_3 = potassium content were measured and the results are presented in table 1.

Table 1: Sweat Data

Individual	X_1 **(sweat rate)**	X_2 **(sodium content)**	X_3 **(potassium content)**
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Test the hypothesis, $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0}$ given $\boldsymbol{\mu_0} = \begin{bmatrix} 4 & 50 & 10 \end{bmatrix}$ against $H_1 : \mu \neq \mu_0$. From Table 1, we can calculate

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j = \begin{bmatrix} 4.640 \\ 45.400 \\ 9.965 \end{bmatrix},\ \mathbf{s} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{x_j} - \bar{\mathbf{x}}\right)\left(\mathbf{x_j} - \bar{\mathbf{x}}\right)' = \begin{bmatrix} 2.879368 & 10.01 & -1.80905 \\ 10.01 & 199.7884 & -5.64 \\ -1.80905 & -5.64 & 3.627658 \end{bmatrix}$$

and the observed T^2 value is

$$= n(\bar{\mathbf{x}} - \mu_{\mathbf{0}})' \mathbf{s}^{-1} (\bar{\mathbf{x}} - \mu_{\mathbf{0}})$$

$$= 20 \begin{bmatrix} 4.640-4 & 45.400-50 & 9.965-10 \end{bmatrix} \begin{bmatrix} 2.879368 & 10.01 & -1.80905 \\ 10.01 & 199.7884 & -5.64 \\ -1.80905 & -5.64 & 3.627658 \end{bmatrix}^{-1} \begin{bmatrix} 4.640-4 \\ 45.400-50 \\ 9.965-10 \end{bmatrix}$$

$$= 20 \begin{bmatrix} 0.640 & -4.600 & -0.035 \end{bmatrix} \begin{bmatrix} 0.467705 \\ -0.04199 \\ 0.158308 \end{bmatrix} = 9.738774$$

Comparing the observed T^2 =9.738774 with the critical value $\frac{(n-1)p}{(n-p)} F_{p,n-p(\alpha)} = 3.353 \times 3.20 = 10.73$ we see that T^2 = *9.74* < *10.73*, and consequently we accept H_0.

2. Testing of mean vectors - Two Sample Case

Consider that we have two independent random samples of sizes n_1 and n_2 with mean vectors $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ and sample dispersion matrices $\mathbf{s}_1$ and $\mathbf{s}_2$ respectively and want to test the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{against} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

$\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are mean vectors of populations from which samples are drawn. If population dispersion matrices are unknown but same, we use

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \mathbf{s}_{pooled}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

where $\mathbf{s}_{pooled} = \frac{(n_1 - 1)\mathbf{s}_1 + (n_2 - 1)\mathbf{s}_2}{n_1 + n_2 - 2}$.

T^2 is distributed as $\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}$

Example 2: {Example 6.4 in Johnson and Wichern, 2002}. Samples of sizes $n_1 = 45$ and $n_2 = 55$ were taken of homeowners with and without air conditioning respectively. Two measurements of electrical usage (is kilowatt-hours) were considered. The first is a measure of total on-peak consumption ($\mathbf{x}_1$) during July and the second is a measure of total off-peak consumption during July. Test weather there is a difference in electrical consumption between those with air conditioning and those without.

The summary statistics given are

$$\mathbf{x}_1 = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix}$$

$$\mathbf{s}_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix}, \quad \mathbf{s}_2 = \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}$$

$$n_1 = 45, n_2 = 55$$

Here the null hypothesis is $H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2}$ and alternate hypothesis is $H_1 : \boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}$. To test the difference, first we calculate

$$\mathbf{s}_{pooled} = \frac{(n_1 - 1)\mathbf{s}_1 + (n_2 - 1)\mathbf{s}_2}{n_1 + n_2 - 2}$$

$$= \begin{bmatrix} 10963.7 & 21505.5 \\ 21505.5 & 63661.3 \end{bmatrix}.$$

Now $$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\mathbf{x}_1 - \mathbf{x}_2)' \mathbf{s}_{pooled}^{-1}(\mathbf{x}_1 - \mathbf{x}_2)$$

$$= 16.22066$$

Comparing the observed T^2 with the critical value

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1(\alpha)} = \frac{98(2)}{97} F_{2.97(0.05)} = 6.26 .$$

We see that the observed $T^2 = 16.22066 > 6.26$, we reject the null hypothesis and conclude that there is a difference in electrical consumption between those with air conditioning and those without.

Note:

(i) For this testing, Mahalnobis D^2 can also be used which is a linear function of T^2

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{s}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$= \frac{n_1 n_2}{n_1 + n_2} T^2$$

(ii) If $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, the above test cannot be used. For large sample size or dispersion matrices known, χ^2 test can be used. However, test for small sample sizes and dispersion matrices not known to be equal is beyond the scope of discussion. Readers may go through the references given at the end.

Steps to carry out the Analysis: **Testing Mean Vector (s) (Using MS-EXCEL)**

We to use the inbuilt Functions of MS-EXCEL like *Average*: Mean; *VAR*: Variance and *COVAR*n/(n-1)*: Covariance. Correlation can be obtained using the function *CORREL*.

Matrix Inverse

Mark the area for the resultant matrix → Formula bar → =minverse (mark range of original matrix) → press control + shift + enter

Matrix multiplication

Mark the area for the resultant matrix → Formula bar → =mmult (mark range of first matrix, mark range of second matrix) → press control + shift + enter

Using the matrix multiplication and matrix inversion one can easily calculate Hotelling's T^2.

3. Multivariate Analysis of Variance (MANOVA)

One way Classified Data

Consider that the random samples from each of g (say) populations using are arranged as

Population 1: $\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$

Population 2: $\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$

:

:

:

Population g: $\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gn_g}$

Multivariate analysis of variance is used first to investigate whether the populations mean vectors are the same and, if not, which mean components differ significantly. MANOVA is carried out under the following two assumptions: 1. Dispersion matrices of various populations are same. 2. Each population is multivariate normal. One-way Classified MANOVA Table for testing the equality of g-population mean Vectors is given below:

Source of variation	Degrees of freedom	SSP matrix
Population or treatment	g-1	$\mathbf{T} = \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$
Residual (error)	$\sum_{i=1}^{g} n_i - g$	$\mathbf{R} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$
Total	$\sum_{i=1}^{g} n_i - 1$	$\mathbf{T} + \mathbf{R} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})'$

We reject the null hypothesis of equal mean vectors if the ratio of generalized variance (*Wilk's lambda* statistic) $\Lambda^* = \dfrac{|\mathbf{R}|}{|\mathbf{T}+\mathbf{R}|}$ is too small. The distribution of Λ^* in different cases are as below.

$p = 1 \quad g \geq 2 \qquad \left(\dfrac{\sum n_i - g}{g-1}\right)\left(\dfrac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1,(\sum n_i - g)}(\alpha)$

$p = 2 \quad g \geq 2 \qquad \left(\dfrac{\sum n_i - g - 1}{g-1}\right)\left(\dfrac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(\sum n_i - g - 1)}(\alpha)$

$p \geq 1 \quad g = 2 \qquad \left(\dfrac{\sum n_i - p - 1}{p}\right)\left(\dfrac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p,(\sum n_i - p - 1)}(\alpha)$

$p \geq 1 \quad g = 3 \qquad \left(\dfrac{\sum n_i - p - 2}{p}\right)\left(\dfrac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(\sum n_i - p - 2)}(\alpha)$

and for other cases $-\left(n - 1 - \dfrac{(p+g)}{2}\right) \text{In } \Lambda^* \sim \chi^2_{p(g-1)}(\alpha)$ (approximate).

Example 3: {Example 6.8 in Johnson and Wichern, 2002}. Consider the following independent samples:

	R1	R2	R3	Total
Population 1	$\begin{bmatrix}9\\3\end{bmatrix}$	$\begin{bmatrix}6\\2\end{bmatrix}$	$\begin{bmatrix}9\\7\end{bmatrix}$	$\begin{bmatrix}24\\12\end{bmatrix}$
Population 2	$\begin{bmatrix}0\\4\end{bmatrix}$	$\begin{bmatrix}2\\0\end{bmatrix}$		$\begin{bmatrix}2\\4\end{bmatrix}$
Population 3	$\begin{bmatrix}3\\8\end{bmatrix}$	$\begin{bmatrix}1\\9\end{bmatrix}$	$\begin{bmatrix}2\\7\end{bmatrix}$	$\begin{bmatrix}6\\24\end{bmatrix}$
Grand Total				$\begin{bmatrix}32\\40\end{bmatrix}$

Due to variable 1

Sum of squares (Population) $= \dfrac{24^2}{3} + \dfrac{2^2}{2} + \dfrac{6^2}{3} - \dfrac{38^2}{8} = 78$

Sum of squares (Total) $= 9^2 + 6^2 + \ldots + 2^2 - \dfrac{38^2}{8} = 88$

Sum of squares (Residual) $= 88 - 78 = 10$ (by subtraction)

Due to variable 2

$$\text{Sum of squares (Population)} = \frac{12^2}{3} + \frac{4^2}{2} + \frac{24^2}{3} - \frac{40^2}{8} = 48$$

$$\text{Sum of squares (Total)} = 3^2 + 2^2 + ... + 7^2 - \frac{40^2}{8} = 72$$

Sum of squares (Residual) $= 72 - 48 = 24$ (by subtraction)

Due to variable 1 and 2

$$\text{Sum of cross products (Population)} = \frac{24 \times 12}{3} + \frac{2 \times 4}{2} + \frac{6 \times 24}{3} - \frac{32 \times 40}{8} = -12$$

$$\text{Sum of cross products (Total)} = 9 \times 3 + 6 \times 2 + ... + 2 \times 7 - \frac{32 \times 40}{8} = -11$$

Sum of cross products (Residual) $= -11 - (-12) = 1$

MANOVA

Source of Variation	**Degrees of freedom**	**SSP matrix**
Population	$3 - 1 = 2$	$\begin{bmatrix} 78 & -12 \\ -12 & 48 \end{bmatrix}$
Residual (error)	$3 + 2 + 3 - 3 = 5$	$\begin{bmatrix} 10 & 1 \\ 1 & 24 \end{bmatrix}$
Total	$3 + 2 + 3 - 1 = 7$	$\begin{bmatrix} 88 & -11 \\ -11 & 72 \end{bmatrix}$

To test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$. We use *Wilk's lambda* statistic

$$\Lambda^* = \frac{|\mathbf{R}|}{|\mathbf{T} + \mathbf{R}|} = \frac{\left|\begin{bmatrix} 10 & 1 \\ 1 & 24 \end{bmatrix}\right|}{\left|\begin{bmatrix} 88 & -11 \\ -11 & 72 \end{bmatrix}\right|} = \frac{10(24) - (1)^2}{88(72) - (-11)^2} = \frac{239}{6215} = 0.0385$$

Since $p = 2$ (variables) and $g = 3$ (populations), we use the following $\left(\frac{\sum n_i - p - 2}{p}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) = \left(\frac{8 - 3 - 1}{3 - 1}\right)\left(\frac{1 - \sqrt{0.0385}}{\sqrt{0.0385}}\right) = 8.19$ with a percentage point of an F-distribution having $n_1 = 4$ & $n_2 = 8$ d.f. Since $8.19 > F_{4,8}(0.01) = 7.01$, we reject the null hypothesis at 1% level of significance and conclude that there exists treatment differences. The pairwise comparisons can be done using the contrast analysis.

Remark: One complication of multivariate analysis that does not arise in the univariate case is due to the ranks of the matrices. The rank of **R** should not be smaller than p or in other words error degrees of freedom s should be greater than or equal to p ($s \geq p$).

For performing MANOVA using SAS, the following procedures/statements may be used.

PROC ANOVA and PROC GLM can be used to perform analysis of variance even for more than one dependent variables. PROC ANOVA performs the analysis of variance for balanced data whereas PROC GLM can analyze both balanced and unbalanced data. As ANOVA takes into account the special features of a balanced data, 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 Variable;

TEST H=effects E= effect M = equations/options;

REPEATED factor - name levels / options;

BY variables;

RUN;

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).

PROC GLM for analysis of variance is similar to PROC ANOVA. The statements listed for PROC ANOVA are also used for PROC GLM. 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;

However, if the MODEL statement includes more than one dependent variable, additional multivariate statistics can be requested with the MANOVA statement.

When a MANOVA statement appears before the first RUN statement, GLM or ANOVA enters a multivariate mode with respect to the handling of missing values; observations with missing independent or dependent variables are excluded from the analysis. If you want to use this mode of handling missing values and do not need any multivariate analysis, specify the MANOVA option in the PROC GLM statement.

If both the CONTRAST and MANOVA statements are to be used, the MANOVA statement must appear after the CONTRAST statement. The basic syntax of MANOVA statement is

MANOVA;

MANOVA < H=effects | INTERCEPT | _ALL_ ><E=effect></options>;

MANOVA < H=effects | INTERCEPT | _ALL_><E=effect>

<M=equation,...,equation |(row-or-matrix,...,row-or-matrix)>

<MNAMES=names><PREFIX=name></options>;

The terms given in the MANOVA statement are specified as follows:

H=effects | INTERCEPT | _ALL_ : specifies effects in the preceding model to use as hypothesis matrices. For each H matrix (the SSCP matrix associated with that effects), the H=specification prints the characteristic roots and vectors of $\mathbf{E}^{-1}\mathbf{H}$ (where **E** is the matrix associated with the error effects), Hotelling-Lawley trace, Pillai's trace, Wilks' criterion, and Roy's maximum root criterion with approximate F statistic. Use the keyword INTERCEPT to print tests for the intercept. To print tests for all effects listed in the MODEL statement, use the keyword _ALL_ in place of a list of effects.

E=effect : specifies the error effect. If we omit the E=specification, GLM uses the error SSCP (residual) matrix from the analysis.

<M=equation, ..., equation |(row-or-matrix,...,row-or-matrix)> : specifies a transformation matrix for the dependent variables listed in the MODEL statement. The equations in the M=specification are of the form

$$C_1{}^{*}\text{dependent-variable} \pm C_2{}^{*}\text{dependent-variable\_} \pm C_n{}^{*}\text{dependent-variable}$$

where the C_i values are coefficients for the various dependent-variables. If the value of a given C_i is 1, it may be omitted; in other words, 1*Y is the same as Y. Equations should involve two or more dependent variables. Alternatively, we can input the transformation matrix directly by entering the elements of the matrix with commas separating the rows, and parentheses surrounding the matrix. When this alternate form of input is used, the number of elements in each row must equal the number of dependent variables. Although

these combinations actually represent the columns of the M matrix, they are printed by rows.

When we include an M=specification, the analysis requested in the MANOVA statement is carried out for the variables defined by the equations in the specification, not the original dependent variables. If M=is omitted, the analysis is performed for the original dependent variables in the MODEL statement.

If an M=specification is included without either the MNAMES= or PREFIX=option, the variables are labelled by default as MVAR1, MVAR2, and so on.

MNAMES= names: provides names for the variables defined by the equations in the M=specification. Names in the list correspond to the M=equations or the rows of the M matrix (as it is entered).

PREFIX = name : is an alternative means of identifying the transformed variables defined by the M=specification. For example, if you specify PREFIX = DIFF, the transformed variables are labelled DIFF1, DIFF2, and so on.

The following options can be used in the MANOVA statement

CANONICAL : Prints a canonical analysis of the **H** and **E** matrices (transformed by the M matrix, if specified) instead of the default printout of characteristic roots and vectors.

ETYPE=n : specifies the type(1,2,3, or 4) of the **E** matrix. By default, the procedure uses an ETYPE=value corresponding to the highest type (largest n) used in the analysis.

HTYPE =n : specifies the type (1,2,3, or 4) of the H matrix.

ORTH : requests that the transformation matrix in the M=specification of the MANOVA statement be orthonormalized by rows before the analysis.

PRINTE : prints the **E** matrix. If the **E** matrix is the error SSCP (residual) matrix from the analysis, the partial correlations of the dependent variables given the independent variables are also printed. For example, the statement

```
manova / printe;
```

prints the error SSCP matrix and the partial correlation matrix computed from the error SSCP matrix.

PRINTH : prints the **H** matrix (the SSCP matrix) associated with each effect specified by the **H**=specification.

SUMMARY: produces analysis-of-variance tables for each dependent variable. When no **M** matrix is specified, a table is printed for each original dependent variable from the MODEL statement; with an **M** matrix other than the identity, a table is printed for each transformed variable defined by the **M** matrix.

Various ways of using a MANOVA statement are given as follows:

```
proc glm;
```

```
class a b;

model y1-y5=a b(a);

manova h=a e=b(a) / printh printe htype=1 etype=1;

manova h=b(a) / printe;

manova h=a e=b(a) m=y1-y2, y2-y3, y3-y4, y4-y5  prefix=diff;

manova h=a e=b(a) m=(1  -1  0  0  0,
                     0   1 -1  0  0,
                     0   0  1 -1  0,
                     0   0  0  1 -1) prefix=diff;
run;
```

Since this MODEL statement requests no options for type of sums of squares, GLM uses Type I and Type III. The first MANOVA statement specifies A as the hypothesis effect and B(A) as the error effect. As a result of PRINTH, the procedure prints the **H** matrix associated with the A effect; and, as a result of PRINTE, the procedure prints the **E** matrix associated with the B(A) effect. HTYPE=1 specifies a Type I **H** matrix, and ETYPE =1 specifies a Type I **E** matrix.

The second MANOVA statement specifies B(A) as the hypothesis effect. Since no error effect is specified, GLM uses the error SSCP matrix from the analysis as the **E** matrix. The PRINTE option prints this **E** matrix. Since the **E** matrix is the error SSCP matrix from the analysis, the partial correlation matrix computed from this matrix is also printed.

The third MANOVA statement requests the same analysis as the first MANOVA statement, but the analysis is carried out for variables transformed to be successive differences between the original dependent variables. PREFIX=DIFF labels the transformed variables as DIFF1, DIFF2, DIFF3, and DIFF4.

Finally, the fourth MANOVA statement has the identical effect as the third, but it uses an alternative form of the M=specification. Instead of specifying a set of equations, the fourth MANOVA statement specifies rows of a matrix of coefficients for the five dependent variables.

SPSS: To obtain MANOVA, from the menus choose Analyze → General Linear Models… → Multivariate…→ Select at least two dependent variables→ Optionally, one can specify Fixed Factor(s), Covariate(s), and WLS Weight.

4. Principal Component Analysis

The purpose of principal component analysis is to derive a small number of linear combinations (principal components) of a set of variables that retain as much information

in the original variables as possible. Often a small number of principal components can be used in place of the original variables for plotting, regression, clustering and so on. Principal component analysis can also be viewed as a technique to remove multicollinearity in the data. In this technique, we transform the original set of variables to a new set of uncorrelated random variables. These new variables are linear combinations of the originals variables and are derived in decreasing order of importance so that the first principal component accounts for as much as possible of the variation in the original data. Let $x_1, x_2, x_3, ..., x_p$ are variables under study, then first principal component may be defined as

$$z_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$

such that variance of z_1 is as large as possible subject to the condition that

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$$

This constraint is introduced because if this is not done, then $Var(z_1)$ can be increased simply by multiplying any $a_{1j}'s$ by a constant factor. The second principal component is defined as

$$z_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p$$

such that $Var(z_2)$ is as large as possible next to $Var(z_1)$ subject to the constraint that

$a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$ and $Cov(z_1, z_2) = 0$ and so on.

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial *p* variables in subsequent analysis, thus reducing the effective dimensionality of the problem. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result. However, Principal Components Analysis is more of a mean to an end rather than end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique, which does not require user to specify the statistical model or assumption about distribution of original variates. It may also be mentioned that principal components are artificial variables and often it is not possible to assign physical meaning to them. Further, since Principal Components Analysis transforms original set of variables to new set of uncorrelated variables. It is worth stressing that if the original variables are uncorrelated, then there is no point in carrying out the Principal Components Analysis. It is important to note here that the principal components depend on the scale of measurement. Conventional way of getting rid of this problem is to use the standardized variables with unit variances.

Example 4: Let us consider the following data on average minimum temperature (x_1), average relative humidity at 8 hrs. (x_2), average relative humidity at 14 hrs. (x_3) and total

rainfall in cm. (x_4) pertaining to Raipur district from 1970 to 1986 for kharif season from 21st May to 7th Oct.

	x_1	x_2	x_3	x_4
	25.0	86	66	186.49
	24.9	84	66	124.34
	25.4	77	55	98.79
	24.4	82	62	118.88
	22.9	79	53	71.88
	7.7	86	60	111.96
	25.1	82	58	99.74
	24.9	83	63	115.20
	24.9	82	63	100.16
	24.9	78	56	62.38
	24.3	85	67	154.40
	24.6	79	61	112.71
	24.3	81	58	79.63
	24.6	81	61	125.59
	24.1	85	64	99.87
	24.5	84	63	143.56
	24.0	81	61	114.97
Mean	23.56	82.06	61.00	112.97
S.D.	4.13	2.75	3.97	30.06

with the variance co-variance matrix.

$$\sum = \begin{bmatrix} 17.02 & -4.12 & 1.54 & 5.14 \\ & 7.56 & 8.50 & 54.82 \\ & & 15.75 & 92.95 \\ & & & 903.87 \end{bmatrix}$$

Find the eigenvalues and eigenvectors of the above matrix. Arrange the eigenvalues in decreasing order. Let the eigenvalues in decreasing order and corresponding eigenvectors are

$\lambda_1 = 916.902 \quad \mathbf{a_1} = (0.006, 0.061, 0.103, 0.993)$

$\lambda_2 = 18.375 \quad \mathbf{a_2} = (0.955, -0.296, 0.011, 0.012)$

$\lambda_3 = 7.87 \quad \mathbf{a_3} = (0.141, 0.485, 0.855, -0.119)$

$\lambda_4 = 1.056 \quad \mathbf{a_4} = (0.260, 0.820, -0.509, 0.001)$

The principal components for this data will be

$$z_1 = 0.006x_1 + 0.061x_2 + 0.103x_3 + 0.993x_4$$
$$z_2 = 0.955x_1 - 0.296x_2 + 0.011x_3 + 0.012x_4$$
$$z_3 = 0.141x_1 + 0.485x_2 + 0.855x_3 - 0.119x_4$$
$$z_4 = 0.26x_1 + 0.82x_2 - 0.509x_3 + 0.001x_4$$

The variance of principal components will be eigenvalues i.e.

$$Var(z_1) = 916.902,\ Var(z_2) = 18.375,\ Var(z_3) = 7.87,\ Var(z_4) = 1.056$$

The total variation explained by principal components is

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 916.902 + 18.375 + 7.87 + 1.056 = 944.20$$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.

The proportion of total variation accounted for by the principal components is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{916.902}{944.203} = 0.97$$

Continuing, the first two components account for a proportion

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{935.277}{944.203} = 0.99 \text{ of the total variance.}$$

Hence, in further analysis, the first or first two principal components z_1 and z_2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of x_i's in the equations of z_i's. For above data, the first two principal components for first observation i.e. for year 1970 can be worked out as

$$z_1 = 0.006 \times 25.0 + 0.061 \times 86 + 0.103 \times 66 + 0.993 \times 186.49 = 197.380$$
$$z_2 = 0.955 \times 25.0 - 0.296 \times 86 + 0.011 \times 66 + 0.012 \times 186.49 = 1.383$$

Similarly for the year 1971

$$z_1 = 0.006 \times 24.9 + 0.061 \times 84 + 0.103 \times 66 + 0.993 \times 124.34 = 135.54$$
$$z_2 = 0.955 \times 24.9 - 0.296 \times 84 + 0.011 \times 66 + 0.012 \times 124.34 = 1.134$$

Thus the whole data with four variables can be converted to a new data set with two principal components.

Example 5: Consider the same data as given in Example 1. The variance-covariance matrix was given as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.879368 & 10.01 & -1.80905 \\ 10.01 & 199.7884 & -5.64 \\ -1.80905 & -5.64 & 3.627658 \end{bmatrix}$$

Now find the eigenvalues and eigenvectors of the above matrix. Arrange the eigenvalues in decreasing order. Let the eigenvalues in decreasing order and corresponding eigenvectors are

$\lambda_1 = 200.462 \quad \mathbf{a_1} = (0.0508, 0.9983, -0.0291)$

$\lambda_2 = 4.532 \quad \mathbf{a_2} = (-0.5737, 0.0530, \quad 0.8173)$

$\lambda_3 = 1.301 \quad \mathbf{a_3} = (0.8175, -0.0249, 0.5754)$

The principal components for this data are

$$z_1 = 0.0508x_1 + 0.9983x_2 - 0.0291x_3$$
$$z_2 = -0.5737x_1 + 0.0530x_2 + 0.8173x_3$$
$$z_3 = 0.8175x_1 - 0.0249x_2 + 0.5754x_3$$

The variance of principal components will be eigenvalues i.e.

$Var(z_1) = 200.462, \ Var(z_2) = 4.532, \ Var(z_3) = 1.301$

The total variation explained by principal components is

$\lambda_1 + \lambda_2 + \lambda_3 = 200.462 + 4.532 + 1.301 = 206.295$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.

The proportion of total variation accounted for by the principal components is

$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{200.462}{206.295} = 0.9717$ of the total variance.

Continuing, the first two components account for a proportion

$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{204.994}{206.295} = 0.9937$ of the total variance.

Hence, in further analysis, the first or first two principal components z_1 and z_2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of x_i's in the equations of z_i's. For above data, the first two principal components for first observation i.e. for first individual is

$$z_1 = 0.0508 \times 3.7 + 0.9983 \times 48.5 - 0.0291 \times 9.3$$
$$z_2 = -0.5737 \times 3.7 + 0.0530 \times 48.5 + 0.8173 \times 9.3$$

Similarly principal component scores for other individuals can be obtained. Thus the whole data with three variables can be converted to a new data set with two principal components.

Following steps of SAS may be used for performing the principal component analysis.

The PROC PRINCOMP can be used for performing principal component analysis. Raw data, a correlation matrix, a covariance matrix or a sum of squares and cross products (SSCP) matrix can be used as input data. The data sets containing eigenvalues, eigenvectors, and standardized or unstandardized principal component scores can be created as output. The basic syntax of PROC PRINCOMP is as follows:

PROC PRINCOMP Cov <options>;

BY variables;

FREQ Variable;

PARTIAL Variables;

VAR Variables;

WEIGHT Variable;

RUN;

The PROC PRINCOMP and RUN are must. However, the VAR statement listing the numeric variables to be analysed is usually used alongwith PROC PRINCOMP statement. If the DATA= data set is TYPE=SSCP, the default set of variables does not include intercept. Therefore, INTERCEPT may also be included in the VAR statement. The following options are available in PROC PRINCOMP.

A. DATA SETS SPECIFICATION

1. DATA= SAS-data-set : names the SAS data set to be analysed. This data set can be ordinary data set or a TYPE = CORR, COV, FACTOR, UCORR or UCOV data set.

2. OUT = SAS-data-set : creates an output data set containing original data alongwith principal component scores.

3. OUTSTAT-SAS-data-set : creates an output data set containing means, standard deviations, number of observations, correlations or covariances, eigenvalues and eigenvectors.

B. ANALYTICAL DETAILS SPECIFICATION

1. COV: computes the principal components from the covariance matrix. The default option is computation of principal components using a correlation matrix.

2. N=: the non-negative integer equal to the number of principal components to be computed.

3. NOINT : omits the intercept from the model

4. PREFIX= name: specifies a prefix for naming the principal components. The default option is PRIN1, PRIN2, ...

5. STANDARD (STD): standardizes the principal component scores to unit variance from the variance equal to corresponding eigenvalue.

6. VARDEF=DF | N | WDF | WEIGHT: specifies the divisor (error degree of freedom | number of observations | sun of weights | sum of weights-1) in calculating variances and standard deviations. The default option is DF.

Besides these options NOPRINT option suppresses the output. The other statements in PROC PRINCOMP are:

By variables: obtains the separate analysis on observations in groups defined by variables.

FREQ statement: It names a variable that provides frequencies of each observation in the data set. Specifically, if n is the value of the FREQ variable for a given observation, then that observation is used 'n' times.

PARTIAL Statement: used to analyze for a partial correlation or covariance matrix.

VAR statement: Lists the numeric variables to be analysed.

WEIGHT Statement: If we want to use relative weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a weight statement. This is often done when the variance associated with each observation is different and the values of the weight variable are proportional to reciprocals of the variances. The observation is used in the analysis only if the value of the WEIGHT statement variable is non-missing and greater than zero.

The other closely related procedures with PROC PRINCOMP are

PROC PRINQUAL: It performs a principal component analysis of a qualitative data.

PROC CORRESP: performs correspondence analysis, which is a weighted principal component analysis of contingency tables.

For detailed steps for performing principal component analysis using SAS and SPSS, a reference may be made to link "Analysis of Data" at Design Resources Server. SAS and SPSS codes can be obtained from http://www.iasri.res.in/design/Analysis of data/principal_component.html

5. Canonical Correlation Analysis

Canonical correlation is a technique for analyzing the relationship between two sets of variables. Each set can contain several variables. Simple and multiple correlation are

special cases of canonical correlation in which one or both sets contain a single variable. This analysis actually focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in the second set. The idea is first to determine the pair of linear combinations having the largest correlation. Next we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair. This process continues until the number of pairs of canonical variables equals the number of variables in the smaller group. The pairs of linear combinations are called the **canonical variables** and their correlations are called **canonical correlations**. The canonical correlations measure the strength of association between the two sets of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pair of canonical variables.

The PROC CANCORR procedure tests a series of hypotheses that each canonical correlation and all smaller correlations are zero in population using an F-approximation. At least one of the two sets of the variables should have an approximate multivariate normal distribution. PROC CANCORR can also perform partial canonical correlation, a multivariate generalization of ordinary partial correlation. Most commonly used parametric statistical methods, ranging from t-tests to multivariate analysis of covariance are special cases of partial canonical correlations.

6. Discriminant Analysis

The term discriminant analysis refers to several types of analysis viz. classificatory discriminant analysis (used to classify observations into two or more known groups on the basis of one or more quantitative variables), Canonical discriminant analysis (a dimension reduction technique related to principal components and canonical correlation), Stepwise discriminant analysis (a variable selection technique i.e. to try to find a subset of quantitative variables that best reveals differences among the classes).

For classificatory discriminant analysis, Fisher's Discriminant function is generally used. It is described in the sequel.

Fisher's idea was to transform the multivariate observations **x** to univariate observations *y* such the *y*'s derived from the populations π_1 and π_2 were separated as much as possible. Fisher's approach assumes that the populations are normal and also assumes the population covariance matrices are equal because a pooled estimate of common covariance matrix is used.

A fixed linear combination of the **x**'s takes the values $y_{11}, y_{12}, ..., y_{1n_1}$ for the observations from the first population and the values $y_{21}, y_{22}, ..., y_{2n_2}$ for the observations from the second population. The separation of these two sets of univariate *y*'s is assessed in terms of the differences between $\bar{y}_1$ and $\bar{y}_2$ expressed in standard deviation units. That is,

$$\text{separartion} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \text{ where } s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the variance. The objective is to select the linear combination of the **x** to achieve maximum separation of the sample means $\bar{y}_1$ and $\bar{y}_2$.

Result: The linear combination $y = \hat{\mathbf{l}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \mathbf{x}$ maximizes the ratio

$$\frac{(\text{Squared distance between sample mean of y})}{(\text{Sample variance of y})} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{(\hat{\mathbf{l}}\bar{\mathbf{x}}_1 - \hat{\mathbf{l}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{l}}' S_{pooled} \hat{\mathbf{l}}}$$

$$= \frac{(\hat{\mathbf{l}}'\mathbf{d})^2}{\hat{\mathbf{l}}' S_{pooled} \hat{\mathbf{l}}'}$$

over all possible coefficient vectors $\hat{\mathbf{l}}'$ where $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. The maximum of the above ratio is $\mathbf{D}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, the Mahalanobis distance.

Fisher's solution to the separation problem can also be used to classify new observations. An allocation rule is as follows.

Allocate $\mathbf{x}_0$ to π_1 if $y_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \mathbf{x}_0 \geq \hat{\mathbf{m}} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$

and to π_2 if $y_0 < \hat{\mathbf{m}}$

If we assume the populations π_1 and π_2 are multivariate normal with a common covariance matrix, the a test of $\mathbf{H}_0 : \mu_1 = \mu_2$ versus $\mathbf{H}_1 : \mu_1 \neq \mu_2$ is accomplished by referring

$$\frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} \left(\frac{n_1 n_2}{n_1 + n_2} \right) \mathbf{D}^2$$

to an F-distribution with $\nu_1 = p$ and $\nu_2 = n_1 + n_2 - p - 1$ degrees of freedom. If $\mathbf{H}_0$ is rejected, we can conclude the separation between the two populations is significant.

Following procedure statements of SAS that can be used for above discriminant analyses.

PROC DISCRIM : Classificatory discriminant analysis

PROC CANDISC : Cannonical discriminant analysis

PROC STEPDISC : Stepwise discriminant analysis.

SPSS: To Obtain a Discriminant Analysis, from the menus choose: Analyze → Classify →Discriminant... → Select an integer-valued grouping variable and click Define Range to specify the categories of interest → Select the independent, or predictor, variables. (If the grouping variable does not have integer values, Automatic Recode on the Transform menu will create one that does.

Example 6: {Example 11.3 in Johnson and Wichern, 2002}. To construct a procedure for detecting potential hemophilia 'A' carriers, blood samples were analyzed for two groups of women and measurements on two variables, $x_1 = \log_{10}(AHF\ activity)$ and $x_2 = \log_{10}(AHF-like\ antigens)$ recorded. The first group of $n_1 = 30$ women were selected from a population who do not carry hemophilia gene (normal group). The second group of $n_2 = 22$ women were selected from known hemophilia 'A' carriers (obligatory group). The mean vectors and sample covariance matrix are given as

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} \text{ and } \mathbf{S}_{pooled}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Now the linear discriminant function is

$$y_0 = \hat{\mathbf{l}}'\mathbf{x}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x}_0$$

$$= \begin{bmatrix} 0.2418 & -0.0652 \end{bmatrix} \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 37.61\, x_1 - 28.92\, x_2$$

Moreover

$$\bar{y}_1 = \hat{\mathbf{l}}'\bar{\mathbf{x}}_1 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} = 0.88$$

$$\bar{y}_2 = \hat{\mathbf{l}}'\bar{\mathbf{x}}_2 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.2483 \\ -0.0262 \end{bmatrix} = -10.10$$

and the mid-point between these means is

$$\hat{\mathbf{m}} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = -4.61$$

Now to classify a women who may be a hemophilia 'A' carrier with $x_1 = -0.210$ and $x_2 = -0.044$.

We calculate: $y_0 = \hat{\mathbf{l}}'\mathbf{x}_0 = 37.61\,x_1 - 28.92\,x_2 = -6.62$. Since $y_0 < \hat{\mathbf{m}}$ we classify the women in π_2 population, *i.e*., to obligatory carrier group.

7. Factor Analysis

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying but unobservable random quantities called *factors*. A frequent source of confusion in the field of factor analysis is the term factor. It sometimes refers to a hypothetical, unobservable variable as in the phrase common factor. In this sense, factor analysis must be distinguished from component analysis since a component is an observable linear combination. Factor is also used in the sense of matrix factor, in that one matrix is a factor of second matrix if the first matrix multiplied by its transpose equals the second matrix. In this sense, factor analysis refers to all methods of data analysis using matrix factors, including component analysis and common factor analysis. A common factor is an unobservable hypothetical variable that contributes to that variance of at least two of the observed variables. The unqualified term " factor" often refers to a common factor. A unique factor is an unobservable hypothetical variable that contributes to the variance of only one of the observed variables. The model for common factor analysis posits one unique factor for each observed variable. The PROC FACTOR can be used for several types of common factor and component analysis. Both orthogonal and oblique rotations are available. We can compute scoring coefficients by the regression method. All major statistics computed by PROC FACTOR can also be saved in an output DATA SET. The PROC FACTOR can be invoked by the following statements:

PROC FACTOR <options>;

VAR variables;

PRIORS Communalities;

PARTIAL Variables;

FREQ Variable;

WEIGHT Variable;

BY variables;

RUN;

Usually only the VAR statement is needed in addition to the PROC FACTOR statement. The some of the important options available with PROC FACTOR are:

METHOD=NAME : specifies the method of extracting factors. The default option is METHOD = PRINCIPAL, which yields principal component analysis if no PRIORS is used or if PRIORS = ONE is specified; if a PRIORS = value other than one is specified, a principal factor anlaysis is performed.

METHOD= PRINT : yields iterated principal factor analysis.

METHOD=ML : performs maximum- likelihood factor analysis.

METHOD = ALPHA : produced alpha factor analysis.

METHOD =ULS: produced unweighted least squares factor analysis.

NFACTORS=n | NFACT=n | N=n specifies the maximum number of factors to be extracted.

PRIORS =name: (ASMC | INPUT | MAX | ONE | RANDOM | SMC) : specifies a method for computing prior communality estimates

ROTATE=name: gives the rotation method. The default is ROTATE=NONE. FACTOR performs the following orthogonal rotation methods:

- EQUAMAX
- ORTHOMAX
- QUARTIMAX
- PARSIMAX
- VARIMAX

After the initial factor extraction, the common factors are uncorrelated with each other. If the factors are rotated by an orthogonal transformation, the rotated factors are uncorrelated. If the factors are rotated by an oblique transformation, the rotated factors become correlated. Oblique rotations often produce more useful patterns than do orthogonal rotations. However, a consequence of correlated factors is that there is no single unambiguous measure of the importance of a factor in explaining a variable. Thus, for oblique rotations, the pattern matrix doesn't provide all the necessary information for interpreting the factors.

SPSS: To Perform Factor Analysis. From the menus choose: Analyze → Data Reduction → Factor... → Select the variables for the factor analysis.

To understand the role of Factor Analysis, consider the following examples

Example 7: What underlying attitudes lead people to respond to the questions on a political survey as they do? Examining the correlations among the survey items reveals that there is significant overlap among various subgroups of items--questions about taxes tend to correlate with each other, questions about military issues correlate with each other, and so on. With factor analysis, you can investigate the number of underlying factors and, in many cases, you can identify what the factors represent conceptually. Additionally, you can compute factor scores for each respondent, which can then be used in subsequent analyses. For example, you might build a logistic regression model to predict voting behavior based on factor scores.

Example 8: A manufacturer of fabricating parts is interested in identifying the determinants of a successful salesperson. The manufacturer has on file the information

shown in the following table. He is wondering whether he could reduce these seven variables to two or three factors, for a meaningful appreciation of the problem.

Data Matrix for Factor Analysis of seven variables (14 sales people)

Sales Person	Height (x_1)	Weight (x_2)	Education (x_3)	Age (x_4)	No. of Children (x_5)	Size of Household (x_6)	IQ (x_7)
1	67	155	12	27	0	2	102
2	69	175	11	35	3	6	92
3	71	170	14	32	1	3	111
4	70	160	16	25	0	1	115
5	72	180	12	30	2	4	108
6	69	170	11	41	3	5	90
7	74	195	13	36	1	2	114
8	68	160	16	32	1	3	118
9	70	175	12	45	4	6	121
10	71	180	13	24	0	2	92
11	66	145	10	39	2	4	100
12	75	210	16	26	0	1	109
13	70	160	12	31	0	3	102
14	71	175	13	43	3	5	112

Can we now collapse the seven variables into three factors? Intuition might suggest the presence of three primary factors: maturity revealed in age/children/size of household, physical size as shown by height and weight, and intelligence or training as revealed by education and IQ.

The sales people data have been analyzed by the SAS program. This program accepts data in the original units, automatically transforming them into standard scores. The three factors derived from the sales people data by principal component analysis (SAS program) are presented below:

Three-factor results with seven variables

Variable	Sales People Characteristics			Communality
	Factor I	Factor II	Factor III	
Height	0.59038	0.72170	-0.30331	0.96140 (sumsq I,II and III)
Weight	0.45256	0.75932	-0.44273	0.97738
Education	0.80252	0.18513	0.42631	0.86006
Age	-0.86689	0.41116	0.18733	0.95564
No. of Children	-0.84930	0.49247	0.05883	0.96730
Size of Household	-0.92582	0.30007	-0.01953	0.94756
IQ	0.28761	0.46696	0.80524	0.94918
Sum of squares	3.61007	1.85136	1.15709	
Variance summarized	0.51572	0.26448	0.16530	Average=0.94550

Factor Loadings

The coefficients in the factor equations are called "factor loadings". They appear above in each factor column, corresponding to each variable. The equations are:

$$\mathbf{F}_1 = 0.59038x_1 + 0.45256x_2 + 0.80252x_3 - 0.86689x_4 - 0.84930x_5 - 0.92582x_6 + 0.28761x_7$$
$$\mathbf{F}_2 = 0.72170x_1 + 0.75932x_2 + 0.18513x_3 + 0.41116x_4 + 0.49247x_5 + 0.30007x_6 + 0.46696x_7$$
$$\mathbf{F}_3 = -0.30331x_1 - 0.44273x_2 + 0.80252x_3 + 0.18733x_4 + 0.58830x_5 - 0.01953x_6 + 0.80524x_7$$

The factor loadings depict the relative importance of each variable with respect to a particular factor. In all the three equations, education (x_3) and IQ (x_7) have got positive loading factor indicating that they are variables of importance in determining the success of sales person.

Variance summarized

Factor analysis employs the criterion of maximum reduction of variance - variance found in the initial set of variables. Each factor contributes to reduction. In our example Factor I accounts for 51.6% of the total variance. Factor II for 26.4% and Factor III for 16.5%. Together the three factors "explain" almost 95% of the variance.

Communality

In the ideal solution the factors derived will explain 100% of the variance in each of the original variables, "Communality" measures the percentage of the variance in the original variables that is captured by the combinations of factors in the solution. Thus communality is computed for each of the original variables. Each variables communality might be thought of as showing the extent to which it is revealed by the system of factors. In our example the communality is over 85% for every variable. Thus the three factors seem to capture the underlying dimensions involved in these variables.

There is yet another analysis called varimax rotation, after we get the initial results. This could be employed if needed by the analyst. We do not intend to dwell on this and those who want to go into this aspect can use SAS program for varimax rotation.

8. Cluster Analysis

The basic aim of the cluster analysis is to find “natural” or “real” groupings, if any, of a set of individuals (or objects or points or units or whatever). This set of individuals may form a complete population or be a sample from a larger population. More formally, cluster analysis aims to allocate a set of individuals to a set of mutually exclusive, exhaustive groups such that individuals within a group are similar to one another while individuals in different groups are dissimilar. This set of groups is called partition or dissection. Cluster analysis can also be used for summarizing the data rather than finding natural or real groupings. Grouping or clustering is distinct from the classification methods in the sense that the classification pertains to a known number of groups, and the operational objective is to assign new observations to one of these groups. Cluster analysis is a more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities). Some of these distance criteria are:

Euclidean distance: This is probably the most commonly chosen type of distance. It is the geometric distance in the multidimensional space and is computed as:

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p}(x_i - y_i)^2\right]^{1/2} = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

where $\mathbf{x},\mathbf{y}$ are the p-dimensional vectors of observations.

Note that Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers). However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed. For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyses may be very different. Generally, it is good practice to transform the dimensions so they have similar scales.

Squared Euclidean distance: This measure is used in order to place progressively greater weight on objects that are further apart. This distance is square of the Euclidean distance.

Statistical distance: The statistical distance between the two p-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ is $d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{s}^{-1}(\mathbf{x}-\mathbf{y})}$, where $\mathbf{s}$ is the sample variance-covariance matrix.

Many more distance measures are available in literature. For details, a reference may be made to Romesburg (1984).

Several types of clusters are possible using various PROC statements:

- Disjoint cluster place each object in one and only one cluster. (PROC FASTCLUS, PROC VARCLUS).
- Hierarchical clusters are organised so that one cluster may be entirely contained within another cluster, but no other kind of overlap between clusters is allowed. (PROC CLUSTER, PROC VARCLUS).
- Overlapping clusters can be constrained to limit the number of objects that belongs simultaneously to two clusters. (PROC OVERCLUS)
- Fuzzy clusters are defined by a probabilities or grade of membership of each object in each cluster. Fuzzy clusters can be disjoint, hierarchical or overlapping.

SPSS: To Obtain a Hierarchical Cluster Analysis, from the menus choose: Analyze → Classify → Hierarchical Cluster... → For clustering cases, select at least one numeric variable, For clustering variables, select at least three numeric variables → Optionally, one can select an identification variable to label cases.

References

Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*. 5th Edition, Pearson Education Inc., New Delhi.

Romesburg, H.C. (1984). *Cluster Analysis for Researchers*. Lifetime Learning Publications, California.

Some E-learning Resources

kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/**PCA-tutorial**.pdf

www.cs.princeton.edu/picasso/mats/**PCA-Tutorial**-Intuition_jp.pdf -

en.wikipedia.org/wiki/Principal_components_analysis

AGRICULTURAL STATISTICS SYSTEM IN INDIA

Tauqueer Ahmad

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012

1. Introduction

India is primarily an agriculture-based country and its economy largely depends upon agriculture. Presently, contribution of agriculture about one third of the national GDP and provides employment to over seventy percent of Indian population in agriculture and allied activities. Therefore, our country's development largely depends upon the development of agriculture. The agricultural production information is very important for planning and allocation of resources to different sectors of agriculture. Agricultural statistics in India have a long tradition. Artha Shastra of Kautilya makes a mention of their collection as a part of the administrative system. During the Moghul period also some basic agricultural statistics were collected to meet the needs of revenue administration. Ain-e-Akbari is most important document that throws great light on the manner in which statistics were collected during the moghul period. After the Moghul period British rule started Ryotwari System, introduced during 18th Century to collect land revenue. In 1866, the British Government initiated collection of agricultural statistics mainly as a by product of revenue administration and these reflected the then primary interest of the Government in the collection of land revenue. Subsequently, the emphasis shifted to crop forecasts designed primarily to serve the British trade interests. On a representation made by a leading firm of Liverpool, trading in wheat, the preparation of wheat forecast was taken up in 1884 and the land utilization statistics are available in the country since 1884. By 1900, oilseeds, rice, cotton, jute indigo and sugarcane had also been added to the list of forecast crops.

The system of agricultural statistics generates valuable statistics on a vast number of parameters. Some of the very important statistics are land-use statistics and area under principal crops through the Timely Reporting Scheme (TRS) and also on complete enumeration basis, yield estimates through the General Crop Estimation Surveys (GCES), the scheme of Cost of Cultivation Studies (CCS), checks the reliability of TRS and yield estimates through the scheme of Improvement of Crop Statistics (ICS) , cost of production estimates, agricultural wages, irrigation statistics, conduct of Agricultural Census and Livestock Census on quinquennial basis. The system of agricultural statistics also generates data on livestock products through the scheme of Integrated Sample Survey (ISS), collects wholesale and retail prices, conducts market intelligence and observes rainfall and weather conditions.

2. Crop Area Statistics

The country can be divided into four broad categories with respect to collection of area statistics namely (i) Temporarily settled states, (ii) Permanently settled states and (iii) Other regions.

(i) Temporarily Settled States

The system of temporarily settlements was introduced in our country in 1892, with a view to fix land revenue for a period, which was subject to change at the time of the next settlement. Ordinarily, the interval between two settlements was 25 to 30 years. In order

to determine the land revenue and to make estimates of production forecast detail statistics are to be collected about land revenue, land value etc. In temporarily settled areas the information on crop area statistics are collected by the village accountant or Patwari and are recorded in a register which is popularly known in northern India as Khasra. The village accountant has been called by different names in different parts of the country such as Karnam in South, Telatti in Maharashtra, Karamchari in Bihar, lekhpal in Uttar Pradesh etc. This category covers around 86% of total reported area of the country.

The crop area statistics collected by village accountant are on the basis of complete enumeration called girdawari. The village accountant is to visit each and every field of the village in each crop season and record the information such as area under different crops/land use categories and its status in standard forms called Khasra register. The work of village accountant is supervised by immediate superior officer known by the name of Quanungo in northern India. Most of the geographical areas of temporarily settled states are cadastrally surveyed and detailed maps are available in tehsil and district Headquarters. The statistics obtained by different village accountants are aggregated to get the crop area statistics at higher administrative units such as blocks, tehsils, district, states etc. This system of data collection is being followed in 18 states namely Andhra Pradesh, Assam (excluding hill districts), Bihar, Chattisgarh, Goa, Gujarat, Haryana, Himachal Pradesh, Jammu & Kashmir, Jharkhand, Karnataka, Madhya Pradesh, Maharashtra, Punjab, Rajasthan, Tamil Nadu, Uttaranchal and Uttar Pradesh and 5 union territories Chandigarh, Dadra and Nagar Haveli, Daman & Due, Delhi and Puduchery.

(ii) Permanently Settled States

There are three states namely Kerala, Orissa and West Bengal which come under category of permanently settled states. In case of these states land revenue was permanently fixed and question of revision ordinarily did not arise. In these states there is no system of recording details of area statistics as there is no permanent revenue staff for a village like village accountant as in the case of temporarily settled area. Initially there was no uniform system of collecting area statistics in these regions. The police Chaukidar or village headman was usually providing the statistics on the basis of guess work which were quite unreliable. In order to improve the quality of these statistics in the permanently settled states, presently, the area statistics in these states are collected by specially appointed field staff under the scheme known as “Establishment of Agency for Reporting Agricultural Statistics (EARAS)” which was initiated in 1968-69. In the States covered by EARAS, the complete enumeration of all fields (survey numbers) i.e. girdawari is conducted every year in a random sample of 20% villages of the States, which are selected in such a way that during a period of 5 years, the entire state is covered. This category covers around 9% of total reported area of the country.

(iii) Other Regions

The remaining eight states in North Eastern regions namely Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura and two other union territories namely Andaman and Nicobar Islands and Lakshadweep do not have a proper reporting system, though states of Tripura and Sikkim (except some minor pockets) are cadastrally surveyed. In these regions, compilation of area statistics is based on conventional methods in which estimates are reported by village choukidars on the basis of personal assessment. This category covers around 5% of total reported area of the country.

3. Timely Reporting Scheme (TRS)

In order to reduce the time-lag between the sowing and availability of estimates of area and harvesting of crops and availability of estimates of production, a Centrally sponsored scheme for Timely Reporting of Estimates of Area and Production of Principal Crops (TRS), was initiated by the Ministry of Agriculture and Irrigation in the year 1968-70. The basic objective of TRS is to reduce time lag for making available area statistics of major crops in addition to providing the sample frame of selection of crop growing fields for crop cutting experiments in permanently settled states. Under the scheme, the villages in each stratum (tehsil/revenue inspector circle/patwari circle etc.) are divided into 5 independent non-overlapping sets, each comprising one fifth of the total number of villages. One set of randomly selected village is chosen for crop inspection on priority basis immediately after the sowing in each season are completed, but in advance of the period prescribed in the land records manuals for such crop inspection. The village crop area statements are submitted to higher authorities in stipulated date to estimate crop area statistics in advance for major crops. These estimates are further used for crop forecasting purposes. The sampled villages under TRS are selected from temporary area in such a way that the entire temporarily settled parts of the country are covered over a period of five years.

The TRS provides for recording the area under irrigation as well as area under high yielding varieties in the selected villages. Besides ensuring accuracy and timeliness of the enumeration of the area under crops, statistical staff under the scheme is required to inspect the fieldwork of crop cutting experiments and ensure timely dispatch of the returns. This scheme has been taken up in a phased manner in different States beginning with Uttar Pradesh and Maharashtra.

4. Establishment of an Agency for Reporting of Agricultural Statistics Scheme (EARAS)

In the states of Kerala, Orissa and West Bengal a scheme similar to TRS was introduced with same objectives of obtaining area estimates based on 20% sample for use of both by Center and States. Here also, it was envisaged that complete enumeration of fields for area figures would be available for all villages over a period of five years as in case of TRS.

5. System of Data Collection for Area Estimation

In the states where land record are maintained (temporary settled) the village accountant is in-charge of a village or a group of villages for carrying out field to field crop inspection in each crop season for an agricultural year to record the crop area and land utilization statistics. He is supposed to record the crop details related to area and land utilization in Khasra register. After the completion of entries for each survey number of the village, an abstract of area sown under different crops “Jinswar statement” is prepared and sent to next higher official in the revenue hierarchy. At the end of each agricultural year a land utilization area statistics are compiled and abstract is sent to related higher official. The crop wise and land utilization wise area statistics obtained from different villages are aggregated at the revenue circle, tehsil and district levels. The district wise area statistics are sent to State Agricultural Statistics Authority (SASA), which is generally Director of Statistical Bureau or the Director of Agriculture or the Director of Land Records. The state level aggregation is done by SASA and forwarded to Directorate

of Economics and Statistics (DES), Ministry of Agriculture & Cooperation, Govt. of India, which is the nodal agency for releasing the state level and the all India level estimates.

In order to improve the timeliness and quality of Agricultural Statistics, Ministry of Agriculture and Cooperation, Govt. of India introduced TRS and EARAS. Area enumeration under TRS has to be completed on priority basis in a random sample of 20% of the villages during each crop season in a state. EARAS was introduced as a sequel to TRS in the non-land record states namely Kerala, Orissa and West Bengal. This scheme provides for setting up whole time agency to cover 20% of villages every year so that all the villages of a state are covered in 5 years. In the sample villages under this scheme, the crop area is to be reported on the basis of complete enumeration.

6. System of Data Collection for Yield Estimation

The sampling design generally adopted for the Crop Estimation Surveys is one of Stratified Multi-Stage Random Sampling with tehsils/taluks/revenue inspector circles/blocks/anchals, etc. as strata and revenue village within a stratum as first stage unit of sampling, survey number/field within each selected village as sampling unit at the second stage and experimental plot of a specified shape and size as the ultimate unit of sampling. In each selected primary unit generally two survey numbers/fields growing the experimental crop are selected for conducting crop-cutting experiments. However, in Dadra and Nagar Haveli three fields are selected instead of two.

Generally, 80-120 experiments are conducted for a crop in a major district where a district is considered as major for a given crop if the area under the crop in the district exceeds 80,000 hectares or lies between 40,000 and 80,000 hectares but exceeds the average area per district in the State. Otherwise, district is considered a minor for a given crop. Experiments in minor districts are so adjusted that the precision of the estimates is fairly high and the workload on the field staff is manageable. On an average, about 44 or 46 experiments are planned in a minor district. The number of experiments allotted to a district is distributed among the strata within the district roughly in proportion to the area under the crop in the stratum. Generally, the crop cutting is done in a plot of size 5m x 5m size for most of the crops in most of the states. However, in UP the shape of the plot is of an equilateral triangle of size 10 meters and in West Bengal a circular plot of radius 1.745 meters is taken for crop cut.

The average yield is obtained after harvesting, threshing, weighing and recording the weight of the produce from the selected plots. In a sub-sample of experiments further processing of the harvested produce is done to determine the percentage recovery of dried grains or the marketable grain of the produce depending on the nature of the crop.

In the case of three non-land record states i.e. Kerala, Orissa and West Bengal both area and yield are estimated on the basis of sample surveys. The crop cutting experiments are planned in a sub-sample of the primary units selected for the purpose of area enumeration. The general procedure of selecting sampling units remains same at different stages as in that of other states. However, some special features of these states need to be mentioned specifically.

In Kerala block/city corporation or municipalities with an area of 10 sq. km. and above are treated as separate stratum. Municipalities with an area of less than sq. km. are merged with adjoining blocks and treated as a single stratum. These blocks are divided into a number of Investigator Zones depending on the area of a block, nature of land, etc.

City Corporation area is divided into three Investigator Zones. Each municipality with an area more than 10 sq. km. is treated as a single Investigator Zone. The number of crop cutting experiments conducted in each Investigator Zone is six per season for paddy, three each for Coconut and Banana and two each for Tapioca, Arecanut, Cashewnut, Pepper, Plantain and Jackfruit in an agricultural year. In a municipal area having separate Investigator Zone, 10 crop cutting experiments are conducted in respect of paddy per season and 5 for coconut per year. For City Corporation areas, six experiments for paddy per season and five for coconut per year in one Investigator Zone are conducted.

7. System for Crop Forecasting

The advance estimates of crop area and production are released with respect to principal food and non-food crops (food grain, oilseeds, sugarcane, fibres etc.), which covers nearly 87% of agricultural output. Four forecasts are issued, first in middle of September, the second in January, the third towards the end of March and fourth by the end of May.

The advance estimates released in September are related to Kharif crops, which is mostly based on reports submitted by states based on visual observation of the field officials. The second forecast which covers both Kharif and Rabi and released in January by taking into account additional information obtained from various sources including agricultural inputs, incidence of pests and diseases, weekly reports from state government regarding area coverage, conditions of standing crops etc. Presently estimates obtained through Remote Sensing are also considered at this stage. The third forecast, which is made in March, the estimates of Kharif and Rabi seasons are revised based on information received from sources such as Market Intelligence Units, Meteorological Department and the Crop Weather Watch Group (CWWG). The forecast made by the end of May is based on actual figures supplied by State Agricultural Statistics Authorities (SASAs) using yield estimates obtained through GCES. In addition to these four forecasts, the DES, MOA provides final estimates in December. The fully revised estimates are obtained in the next crop year in the following December in which all delayed information are incorporated and all India crop statistics are released.

The Mahalanobis National Crop Forecasting Centre (MNCFC) was setup by Ministry of Agriculture with the objective of examining existing mechanism of making forecasts and developing more objective technique. However, the MNCFC need to strengthen the crop forecasting system of the country by incorporating more objective techniques and models based on sound statistical techniques.

8. Co-ordination of Data Collection

The Field Operation Division of NSSO has the overall responsibility of assisting the States in developing suitable techniques for obtaining reliable and timely estimates, providing technical guidance and ensuring adoption of uniform concepts, definitions and procedures in the Crop Estimation Survey (CES) in the States. It reviews the design, plan, details of implementation and the results of the surveys and, associates itself in the conduct of training camps organized for the States field staff and participates in the primary field work of exercising technical supervision.

9. Supervision of Data Collection

Supervision of fieldwork is an essential part of any large-scale sample survey for ensuring quality of data collected. A three-fold approach is adopted in the States for supervision of crop cutting experiments planned under Crop Estimation Surveys. This includes:

a. Supervision by the statistical staff of State Agricultural Statistics Authorities (SASAs),

b. Supervision by the Departmental staff i.e. by the supervisory officers of the Departments whose workers are responsible for the conduct of crop cutting experiments in the field and

c. Supervision by the Technical personnel of the FOD of National Sample Survey Office.

In the States of Goa, Orissa, West Bengal and the UT of Pondicherry where the field work was conducted only by the staff of Statistics Department, the supervision was done by the Statistical staff only whereas in the case of Bihar, Himachal Pradesh, Union Territories of Dadra & Nagar Haveli and Daman & Diu, though there are other primary field agencies, the supervision was done by the State statistical staff only. Though supervision of the conduct of Crop Cutting Experiments in various states was in vogue since inception of Agricultural Statistical Wing from the year 1973-74 (Rabi) onwards, NSSO personnel are participating in the supervision by conducting sample check on crop cutting experiments in the post-harvest stages in a pre-assigned sample under the Scheme for Improvement of Crop Statistics (ICS) in 20 states and 2 Union Territories. Under this Scheme, State statistical staff also undertakes similar sample checks on a matching basis.

10. Applications of Remote Sensing and GIS Technology

In India, Indian Council of Agricultural Research (ICAR) and Indian Space Research Organization (ISRO) jointly conducted the first multi-spectral air born study for identification of root-wilt disease in coconut in 1969.

The country level studies related to applications of remote sensing technologies were initiated after launch of IRS-IA satellite. Crop Acreage and Production Estimation (CAPE) was one of the important projects in this direction for estimation of crop area under wheat, rice, cotton, ground nut, sorghum & mustered. Apart from these national level projects, numbers of small studies have been carried out to develop methodologies for application of satellite data in various fields of agricultural and rural development by Department of Space. Some of these studies are by Dadhwal *et al.* (1985, 1991), etc. Several methodological studies related to estimation of crop area and production have been carried out at Indian Agricultural Statistics Research Institute (IASRI), New Delhi. Singh *et al.* (1992) used satellite data for stratification of crop area for the general crop estimation surveys and obtained more precise estimator of crop yield. Singh *et al.* (1999) also developed small area estimator of crop yield. Singh *et al.* (2002) used satellite data and the farmers eye estimate for developing a reliable crop yield model. Application of remote sensing and GIS technology for estimation of land use statistics using spatial models has been explored by Rai *et al.* (2004). Now, a project entitled "Forecasting Agricultural Output Using Space, Agro-metrology and Land-based Observations" (FASAL) is undertaken under National Crop Forecasting Center (NCFC) of Ministry of Agriculture, to meet the requirements of timely nation wide and multi- crop reliable fore cast. A project has also been taken up jointly by IASRI, New Delhi, Space Application

Center (SAC), Ahmedabad and North-Eastern Space Application Centre (NESAC), Shillong with the support of Directorate of Economic and Statistic of Meghalaya State to explore the possibility of estimation of area and production of field crops by integration of remote sensing technology, GIS and field survey. The result of all these studies are very encouraging and indicates that in future remote sensing and GIS has a great potential tool to improve the quality of area and production statistics of the country.

11. Improvement of Crop Statistics (ICS)

In addition to the crop area estimates developed by the state government the National Sample Survey (NSS) use to develop area estimates based on sample surveys during its regular rounds of surveys. Estimates were obtained for the whole country and also for certain population zones. There used to be significant differences between two series of data on crop area statistics. In order to probe into these high differences a technical committee on crop statistics was set up in 1963. The committee favored inter alia the estimates based on complete enumeration. As a consequence, the NSS discontinued its land utilization surveys and also crop cutting experiments in 1970-71 under household surveys. Thereafter, the NSSO introduced the ICS scheme in 1973-74 with an objective of improving the quality of statistics through joint efforts of centre and state authorities. Currently the scheme is in operation in 20 states and two Union Territories of Delhi and Pondicherry. In this scheme an independent agency (NSSO) carries out the supervision and physical verification of girdawari in a sub-sample of four clusters of five survey numbers in each of the TRS sample villages. An assessment is made for extent of discrepancies between the entries of supervisor and girdawari completed by village accountant for each of the selected survey numbers in the sample. The supervisors for checking possible errors of aggregations also scrutinize the crop abstract of the village, which is prepared by patwaries. The permanently settled states are also covered under this scheme where a sub-sample of EARAS sample villages (survey number) is scrutinized following the same methodology as adopted for temporarily settled area. Generally, a total of 10,000 sample villages are covered by the ICS out of which 8,500 are in the temporarily settled states and 1,500 in the permanently settled states.

National Sample Survey Office (NSSO) is mainly responsible for planning and operations of ICS by employing full time field staff for supervision. The responsibility of field supervision is shared by designated state agencies which are responsible for carrying out the field supervision in approximately half of the sampled villages.

Major Issues Emerging from the ICS Scheme

1. The crop statements submitted by patwari are many times based on incomplete girdawaries.
2. The village crop statements are not submitted in time and there are large percentages of non-response.
3. The entries in the girdawaries are not correct at least for one third of survey numbers.
4. Recording area under mixed crop is a major source of errors as it is not uniform across the states.
5. Sometimes there is uncertainty of recording area under crop as area sown or area harvested. This leads to inaccurate estimation of area, if area sown is recorded as area under crop and there is no germination as expected.

6. Area sown more than once is also responsible for some confusion about statistics of area under various crops.
7. Inclusion of field ridges, bunds in measurements also result in accuracy, which may be higher in some of the cases.
8. Due to introduction of new technology/varieties number of short duration crops are grown and also, there is shift in cropping pattern towards value added crops which are not reflected properly in girdawari.
9. It has been observed that field staff approved by the State Government do not strictly adhere to the prescribed procedures and thereby the survey estimates are subject to a variety of non-sampling errors.
10. The errors are introduced mainly due to wrong selection of fields and duration of selected experimental plots. The use of defective instruments such as proper weighing machine introduces considerable amount of measurement errors.
11. The state departments of revenue and agriculture, which are responsible for carrying out the survey, keeps these programmes on low priority and there is inadequate higher level of supervision and control of field operations. The "High Level Coordination Committee (HLCC) on Agricultural Statistics" in the states has little impact in improving the quality of data.
12. In order to meet the requirements of getting estimates at block/village panchayat levels especially for crop insurance purposes some of the State increased the number of crop cutting experiments considerably. This imposes an enormous burden on the field agency, increases considerably the non-sampling errors, which results in further deterioration of quality of data collected through GCES. There is possibility of under estimation of yield rates in case of crop insurance due to local pressure from insured farmers where interest lies in depressing the crop yield.
13. It has been a matter of great debate in the past as production statistics obtained by different sources/agencies are quite different. The problem is especially significant in case of cash crops like cotton, oilseeds etc.
14. Inadequate training is provided to the field staff for conducting the crop cutting experiments.
15. Another important factor, which has bearing on the quality of production data is, the late time schedule fixed for certain crops in Kharif season in some states. In this case crop-cutting experiments are to be conducted before completion of the season due to early harvesting. Such situations have been arising in respect of Kharif crops like maize, jowar, bajra, groundnut, cotton, soybean etc. in States like Gujarat, Haryana, Karnataka and M.P. Due to early harvesting of these crops, area under crop is generally under reported and hence production too.
16. There is strong need to develop suitable forecasting models which integrate information from different sources on parameters related to crop production such as crop conditions, agro meteorology, water availability etc.
17. No multi-dimensional models exists in which the information generated from different sources can be integrated.
18. The flows of information from different generating agencies are not time bound and appropriate.

19. The DES, MOA is loosing confidence of users group due to frequent changes in production figures specially most of the time differences in the forecasted estimates are huge. These differences create lot of confusion and doubt among users

20. The present technique is mostly subjective and is not based on sound statistical technique.

References

Basu, D. (1969). Role of sufficiency and likelihood principles in sample survey theory. *Sankhya*, *31, 441-454*.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Apart I**.** *In Foundations of Statistical Inference, Holt, Rinehast and Winston, Toronto, 203-242*.

Bowley, A.L. (1906). Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science. *J. Roy. Statist. Soc., 69, 548-557*

Bowley, A.L. (1926). Measurement of the precision attained in Sampling. *Bull. Int. Statist. Inst.*, *22* Livre I.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Austral. J. Statist., 5,* 3-5.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley**.**

Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes.' *J. Amer. Statist. Assoc., 37,* 199-212.

Cochran, W.G. (1953). *Sampling Techniques*, 1st ed., 2nd ed. (1963). 3rd ed., (1977). New York: Wiley.

Cochran, W.G. and Watson, D.J. (1936). An experiment on observer's bias in the selection of shoot-heights. *Empire J. Exp. Agriculture, 4,* 69-76.

Datenius, T. (1962). Recent advances in sample survey theory and methods. *Ann. Math. Statist. 33,* 325-349.

Deming, W.E. (1950). *Some Theory of Sampling*. New York**.** Wiley

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *J. Roy. Statist. Soc. B, 17,* 269-278.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. John Wiley and Sons, New York, Vols. I and II.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc., 47*, 663-685.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin No. 304.*

Kiaer, A.N. (1895-6). Observations of experiences concernant des denombrements representatifs. *Bull. Int. Statist. Inst., 9*, Liv. 2, 176-183.

McCarthy, P.J. (1966). Pseudo-replication: an approach to the analysis of data from complex surveys. *Washington: NCHS Series 2,* No.14.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *J. Ind. Soc. Agril. Statist., 3,* 169-174.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc., 97,* 555- 606.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc., 33*, 101-116.

Rao, C.R., (1985): Evaluation of data collection censuses. S*ample Surveys and Design of Experiments.*

Rao, J.N.K. (1963). On two systems of unequal probability sampling without replacement. *Ann. Inst. Statist. Maths*. 15, 67-72.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables**.** *J. Amer. Statist. Assoc., 76*, 221-230.

Royall, R.M. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc., 63,* 1269-1279.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika, 57*, 377-387.

Verma, V., Scott, C. and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *J. Roy. Statist. Soc.,* A143, 431-473.

Sukhatme, P.V. (1959). Major developments in the theory and application of sampling during the last twenty five years. *Estadistica, 17*, 62-679.

ELEMENTARY CONCEPTS OF SAMPLE SURVEYS AND SIMPLE RANDON SAMPLING

Ankur Biswas

ICAR- Indian Agricultural Statistics Research Institute, New Delhi -110012

1. Introduction

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conservation and controversy require numerical data for their resolution. The data collected and analyzed in an objective manner and presented suitably serve as a basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include government, industry, business, research institutions, public organizations and international agencies and organizations. To discharge its various responsibilities, the government needs variety of information regarding different sectors of economy, trade, industrial production, health and mortality, population, livestock, agriculture, forestry, environment and available resources. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. For instance, the information on cost of living for different categories of people, living in various parts of the country is of importance in shaping its policies in respect of wages and price levels. Data on agricultural production are of immense use to the state for planning to feed the nation. In case of industry and business, the information is to be collected on labour, cost and quality of production, stock and demand and supply positions for proper planning of production levels and sales campaigns.

The purpose of a statistical survey is to obtain information about populations. By 'population' we mean, a group of units defined according to the objective(s) of a survey. Thus, the population may comprise of all the fields under a specified crop as in area and yield surveys, or all the agricultural holdings above a specified size as in agricultural surveys. Of course, the population may also refer to persons either of the whole population of a country or a particular sector thereof. The information that we seek about the population is normally the total number of units, aggregate values of the various characteristics, averages of these characteristics per unit, proportions of units possessing specified attributes, etc.

2. Complete enumeration

One way of obtaining the required information at regional and country level is to collect the data for each and every unit (person, household, field, factory, shop etc. as the case may be) belonging to the population which is the aggregate of all units of a given type under consideration and this procedure of obtaining information is termed as complete enumeration. The effort, money and time required for the carrying out complete enumeration to obtain the different types of data will, generally, be extremely large. However, if the information is required for each and every unit in the domain of study, a complete enumeration is clearly necessary. Examples of such situations are preparation of "voter list" for election purposes and recruitment of personnel in an establishment, etc. But there are many situations, where only summary figures are required for the domain of study as a whole or for group of units.

3. Need for sampling

An effective alternative to a complete enumeration can be sample survey where only some of the units selected in a suitable manner from the population are surveyed and an inference is drawn about the population on the basis of observations made on the selected units. It can be easily seen that compared to sample survey, a complete enumeration is time-consuming, expensive, has less scope in the sense of restricted subject coverage and is subject to greater coverage, observational and tabulation errors. In certain investigations, it may be essential to use specialized equipment or highly trained field staff for data collection making it almost impossible to carry out such investigations. It is of interest to note that if a sample survey is carried out according to certain specified statistical principles, it is possible not only to estimate the value of the characteristic of the population as a whole on the basis of the sample data, but also to get a valid estimate of the sampling error of the estimate. There are various steps involved in the planning and execution of the sample survey. One of the principal steps in a sample survey relates to methods of data collection.

4. Types of data

The collection of required information depends on the nature, object, and scope of study on the one hand and availability of financial resources, time, and man power on the other. The statistical data are of two types: (i) primary data, and (ii) secondary data. The data collected by the Investigator from the original source are called *primary data*. If the required data had already been collected by some agencies or individuals and are now available in the published or unpublished records, these are known as *secondary data.*

Thus, the primary data when used by some other Investigator/Agency become secondary data. There could be large number of publications presenting secondary data. Some of the important ones are given below:

- Official publications of the Federal, State, and Local Governments.
- Reports of Committees and Commissions.
- Publications and reports of business organizations, trade associations, and chambers of commerce.
- Data released by magazines, journals, and newspapers.
- Publications of different international organizations like United Nations Organization, World Bank, International Monetary Fund, United Nations Conference on Trade and Development, International Labor Organization, Food and Agricultural Organization, etc.

Caution must be exercised in using secondary data as they may contain errors of transcription from the primary source.

5. Need for a sample

Collection of information on every unit in the population for the characteristics of interest is known as **complete enumeration or census**. The money and time required for carrying out a census will generally be large, and there are many situations where with limited means complete enumeration is not possible. There are also instances where it is not

feasible to enumerate all units due to their perishable nature. In all such cases, the Investigator has no alternative except resorting to a sample survey.

The number of units (not necessarily distinct) included in the sample is known as the **sample size** and is usually denoted by '**n**', whereas the number of units in the population is called **population size** and is denoted by '**N**'. The ratio **n/N** is termed as **sampling fraction.**

There are certain advantages of a sample survey over complete enumeration, which are as follows:

a) Greater Speed

The time taken for collecting and analyzing the data for a sample is much less than that for a complete enumeration. Often, we come across situations where the information is to be collected within a specified period. In such cases, where time available is short or the population is large, sampling is the only alternative.

b) Greater Accuracy

A census usually involves a huge and unwieldy organization and, therefore, many types of errors may creep in. Sometimes, it may not be possible to control these errors adequately. In sample surveys, the volume of work is considerably reduced. On account of this, the services of better trained and efficient staff can be obtained without much difficulty. This will help in producing more accurate results than those for complete enumeration.

c) Greater scope

There can be investigations where highly trained investigators or sophisticated equipment are needed. In the event of limited availability of trained investigators and sophisticated equipment, the census investigation may become difficult to carry out. Furthermore, since data are obtained by observing limited number of items, their detailed investigation, if necessary, is also possible. Thus, the investigations that are based on samples have more scope.

d) Reduced Cost

Because of lesser number of units in the sample in comparison to the population, considerable time, money, and energy are saved in observing the sample units in relation to the situation where all units in the population are to be covered.

e) More detailed Information

As the number of units in a sample are much less than those in census, detailed information, therefore, can be obtained on more number of variables. However, in complete enumeration, such an effort becomes comparatively difficult.

From the above, it may be seen that the sample survey is more economical, provides more accurate information, and has greater scope in subject coverage as compared to a complete enumeration. It may, however, be pointed out here that sampling errors are present in the results of the sample surveys. This is due to the fact that only a part of the whole population is surveyed. On the other hand, non-sampling errors are likely to be more in case of a census study than these are in a sample survey.

6. Methods of data collection

The different methods of data collection are:

i. Physical observation or measurement
ii. Personal interview
iii. Mail enquiry
iv. Telephonic enquiry
v. Web-based enquiry
vi. Method of Registration
vii. Transcription from records

The first six methods relate to the collection of primary data from the units/ respondents directly, while the last one relates to the extraction of secondary data, collected earlier generally by one or more of the first six methods. These methods have their respective merits and therefore sufficient thought should be given in selection of an appropriate method(s) of data collection in any survey. The choice of the method of data collection should be arrived at after careful consideration of accuracy, practicability and cost from among the alternative methods.

i. Physical observations or measurement

Data collection by physical observation or measurement consists in physically examining the units/respondents and recording data as a result of personal judgment or using a measuring instrument by the investigator. For instance, in a crop cutting experiment for estimating the yield of a crop, the plot is demarcated, the crop in the selected plot is harvested and the produce is weighted to estimate the produce per unit area. Data obtained by this method are likely to be more accurate, but may often prove expensive.

ii. Personal interview

The method of personal interview consists in contacting the respondents and collecting statistical data by questioning. In this case, the investigator can clearly explain to the respondents the objectives of the survey and the exact nature of the data requirements and persuade them to give the required information, thus reducing the possibility of non-response arising from non-cooperation, indifference etc. Further, this method is most suitable for collecting data on conceptually difficult items from respondents. However, this method depends heavily on the availability of well trained interviewer.

iii. Mail enquiry

In a mail enquiry, data are collected by obtaining questionnaires filled in by the respondents, the questionnaires being sent and collected back through an agency such as the postal department. This method is likely to cost much less as compared to the above methods. However, the response may not always be satisfactory depending upon the cooperation of the respondents, the type of questionnaire and the design of the questionnaire. In developing countries where a large proportion of the population is illiterate, the method of mailed questionnaire may not even be feasible.

iv. Telephonic enquiry

In telephonic enquiry, data are collected by questioning the respondents. This method provides an opportunity of two-way communication and thus can reduce the possibility of item non-response. However, this method can be used only for those surveys in which all units of target populations have telephone otherwise it will cause bias in the results.

v. Web-based enquiry

The increasing popularity and wide availability of World Wide Web technologies provide a new mode of data collection. In web-based enquiry, data are collected by obtaining questionnaires filled in by the respondents, the questionnaires being posted on the net. One important advantage of using computer technology in data collection is to minimize the loss of data owing to incomplete or incorrectly completed data sets by using Client side validation. In an era of information superhighway, this method is one of the fastest means of data collection. However, in developing countries where a large proportion of the population does not have access to Internet, the method of web-based enquiry may not serve the purpose for most of the surveys. Various Internet sites are using this method for opinion poll on certain issues.

vi. Method of registration

In the registration method, the respondents are required to register the required information at specified place. The vital statistics registration system followed in many countries provides an illustration of the registration method. The main difficulty with this method, as in the case of the mail enquiry, is the possibility of non-response due to indifference, reluctance, etc. on the part of informants to visit the place of registration and supply the required data.

vii. Transcription from records

The method of transcription from records is used when the data needed for a specific purpose are already available in registers maintained in one or more places, making it no more necessary to collect them directly from the original units at much cost and effort. The method consists in compiling the required information from the registers for the concerned units. This method is extensively used since a good deal of government and business statistics are collected as by-product of routine administrative operations.

7. Various concepts and definitions

i. Element:

An element is a unit about which we require information. For example, a field growing a particular crop is an element for collecting information on the yield of a crop.

ii. Population

The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

iii. Sampling unit

Elementary units or group of such units which besides being clearly defined, identifiable and observable, are convenient for the purpose of sampling are called sampling units. For

instance, in a family budget enquiry, usually a family is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit.

iv. Sampling frame

A list of all the sampling units belonging to the population to be studied with their identification particulars or a map showing the boundaries of the sampling units is known as sampling frame. Examples of a frame are a list of farms and a list of suitable area segments like villages in India or counties in the United States. The frame should be up to date and free from errors of omission and duplication of sampling units.

v. Random sample

One or more sampling units selected from a population according to some specified procedures are said to constitute a sample. The sample will be considered as random or probability sample, if its selection is governed by ascertainable laws of chance. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection. For example, if a population consists of the N sampling units $U_1,U_2,\ldots,U_i,\ldots,U_N$ then, we may select a sample of n units by selecting them unit by unit with equal probability for every unit at each draw with or without replacing the sampling units selected in the previous draws.

vi. Non-random sample

A sample selected by a non-random process is termed as non-random sample. A non-random sample, which is drawn using certain amount of judgment with a view to get a representative sample, is termed as judgment or purposive sample. In purposive sampling units are selected by considering the available auxiliary information more or less subjectively with a view to ensuring a reflection of the population in the sample. This type of sampling is seldom used in large-scale surveys mainly because it is not generally possible to get strictly valid estimates of the population parameters under consideration and of their sampling errors due to the risk of bias in subjective selection and the lack of information on the probabilities of selection of the units.

vii. Population parameters

Suppose a finite population consists of the N units $U_1,U_2,\ldots,U_N$ and let Y_i be the value of the variable y, the characteristic under study, for the i^{th} unit U_i, (i=1,2,…,N). For instance, the unit may be a farm and the characteristic under study may be the area under a particular crop. Any function of the values of all the population units (or of all the observations constituting a population) is known as a population parameter or simply a parameter. Some of the important parameters usually required to be estimated in surveys are population total $Y = \sum_{i=1}^{N} Y_i$ and population mean $\overline{Y} = \sum_{i=1}^{N} Y_i / N$.

viii. Statistic, estimator and estimate

Suppose, a sample of n units is selected from a population of N units, according to some probability scheme and let, the sample observations be denoted by $y_1, y_2,\ldots, y_n$. Any function of these values which is free from unknown population parameters is called a statistic.

An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample and the samples are selected with specified probabilities. The particular value, which the estimator takes for a given sample, is known as an estimate.

ix. Sampling and non-sampling error

The error arises due to drawing inferences about the population on the basis of observations on a part (sample) of it, is termed sampling error. The sampling error is non-existent in a complete enumeration survey since the whole population is surveyed.

The errors other than sampling errors such as those arising through non-response, in-completeness and inaccuracy of response are termed non-sampling errors and are likely to be more wide-spread and important in a complete enumeration survey than in a sample survey. Non-sampling errors arise due to various causes right from the beginning stage when the survey is planned and designed to the final stage when the data are processed and analyzed.The sampling error usually decreases with increase in sample size (number of units selected in the sample) while the non-sampling error is likely to increase with increase in sample size.

As regards the non-sampling error, it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former.

8. Simple Random Sampling

Simple random sampling (SRS) can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure.

Simple random sampling is a method of selecting n units out of the N such that every one of the $\binom{N}{n}$ distinct samples has an equal chance of being drawn. In practice a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. Since a number that has been drawn is removed from the population for all subsequent draws, this method is also called random sampling without replacement. In case of a random sampling with replacement, at any draw all N members of the population are given an equal chance of being drawn, no matter how often they have already been drawn. The with-replacement assumption simplifies the estimation under complex sampling designs and is often adopted, although in practice sampling is usually carried out under a without replacement type scheme. Obviously, the difference between with replacement and without replacement sampling becomes less important when the population size is large and the sample size is noticeably smaller than it.

8.1 Procedure of selecting a random sample

Since probability sampling theory is based on the assumption of random sampling, the technique of random sampling is of basic significance. Some of the procedures used for selecting a random sample are as follows:

i) Lottery method

ii) Use of random number tables

i) Lottery Method:

Each unit in the population may be associated with a chit/ticket such that each sampling unit has its identification mark from 1 to N. All the chits/tickets are placed in a container, drum or metallic spherical device, in which a thorough mixing is possible before each draw. Chits/tickets may be drawn one by one and may be continued until a sample of the required size is obtained. When the size of population is large, this procedure of numbering units on chits/tickets and selecting one after reshuffling becomes cumbersome. In practice, it may be too difficult to achieve a thorough shuffling. Human bias and prejudice may also creep in this method.

ii) Use of Random Number Tables:

A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern where each position is filled with one of these digits. A Table of random numbers is so constructed that all numbers 0, 1, 2,…, 9 appear independent of each other. Some random number tables in common use are:

- Tippett's random number Tables
- Fisher and Yates Tables
- Kendall and Smith Tables
- A million random digits Table

Random number tables are the tables of digits *0, 1, 2,...,*9 each digit having an equal chance of selection at any draw. In *1927*, Tippett produced *41,600* digits (from *0* to *9*) arranged in sets of *4* in several columns and spread over *26* pages. This was followed by another publication by two great pioneering statisticians, Sir R.A. Fisher and Frank Yates, which contained *15,000* digits formed by listing the *15 - 19*th digits in some *20* figure logarithm tables. Rand Corporation (1955) published tables containing 1 million digits. Kendall and Smith (1938) published tables with 100,000 digits.

A practical method of selecting a random sample is to choose units one-by-one with the help of a Table of random numbers. By considering two-digit numbers, we can obtain numbers from 00 to 99, all having the same frequency. Similarly, three or more digit numbers may be obtained by combining three or more rows or columns of these Tables. The simplest way of selecting a sample of the required size is to select a random number from 1 to N and then taking the unit bearing that number. This procedure involves a number of rejections since all numbers greater than N appearing in the Table are not considered for selection. The procedure of selection of sample through the use of random numbers is, therefore, modified and some of these modified procedures are:

a) Remainder Approach

b) Quotient Approach

a) Remainder Approach:

Let N be an r-digit number and let its r-digit highest multiple be N'. A random number k is chosen from 1 to N' and the unit with serial number equal to the remainder obtained on dividing k by N is selected, *i.e.* the selected number is reduced mod (N). If the remainder

is zero, the last unit is selected. As an illustration, let N = 123, then highest three-digit multiple of 123 is 984. For selecting a unit, one random number from 001 to 984 has to be selected. Let the random number selected be 287. Dividing 287 by 123 gives the remainder as 41. Hence, the unit with serial number 41 is selected in the sample. Suppose that another random number selected is 245. Dividing 245 by 123 leaves 122 as remainder. So the unit bearing the serial number 122 is selected. Similarly, if the random number selected is 369, then dividing 369 by 123 leaves remainder as 0. So the unit bearing serial number 123 is selected in the sample.

b) Quotient Approach:

Let N be an r-digit number and let its r-digit highest multiple be N^* such that $N^*/N = d$. A random number k is chosen from 0 to $\left(N^* - 1\right)$. Dividing k by d, the quotient q is obtained and the unit bearing the serial number (q - 1) is selected in the sample. The selected number is reduced mod (N). For example, if q – 1 = -1, then unit bearing serial number N – 1 is selected and if q – 1 = 0, then unit bearing serial number N is selected. As an illustration, let N = 16 and hence $N^* = 96$ and d = 96/16 = 6. Let the two-digit random number chosen is 65 which lies between 0 and 95. Dividing 65 by 6, the quotient is 10 and hence the unit bearing serial number (10 - 1) = 9 is selected in the sample. Further, if the random number selected is 4, then the quotient is 4/6 = 0, and q –1= -1. The unit selected is 15. Similarly, if the random number selected is 9, then the quotient is 9/6 = 1, and q – 1 = 0. The unit selected is 16.

8.2 Estimation of Population Total

Let Y be the character of interest and $Y_1, Y_2, \cdots, Y_i, \cdots, Y_N$ be the values of the character from *N* units of the population. Further, let $y_1, y_2, \cdots, y_i, \cdots, y_n$ be the sample of size n selected by simple random sampling without replacement. For the total $Y = \sum_{i=1}^{N} Y_i$ we have an estimator

$$\hat{Y} = N\sum_{i=1}^{n} y_i / n = N\bar{y}_n$$

i.e., the sample mean $\bar{y}_n$ multiplied by the population size N.

The estimator can be expressed as

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i = \left(N/n\right)\sum_{i=1}^{n} y_i \text{ , where } w_i = N/n.$$

The constant N/n is the sampling weight and is the inverse of the sampling fraction n/N.

Alternatively, an estimator for the population total can be written by first defining the inclusion probability of a population element. Under SRS, the inclusion probability of a population element i is $\pi_i = n/N$, same or constant for every population element. Based on the inclusion probabilities, an estimator of the total can be expressed as a more general Horvitz-Thompson-type estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} \frac{1}{\Pi_i} y_i = \frac{N}{n} \sum_{i=1}^{n} y_i .$$

In this case, the estimator $\hat{Y}$ and $\hat{Y}_{HT}$ obviously coincide, because the inclusion probabilities $\pi_i = n/N$ are equal for each i. The Horvitz-Thompson-type estimator is often used for example, with probability-proportional to size sampling where inclusion probabilities vary. The estimator has the statistical property of unbiasedness in relation to the sampling design. Variance of the estimator $\hat{Y}$ of the population total under SRSWOR is given by

$$V_{SRS}(\hat{Y}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N}(Y_i-\bar{Y})^2/(N-1)$$

where $\bar{Y} = \sum_{i=1}^{N} Y_i / N$ is the population mean and $S^2 = \sum_{i=1}^{N}(Y_i-\bar{Y})^2/(N-1)$ is the population mean square.

An unbiased estimator of variance of the estimator $\hat{Y}$ of the total, $V_{SRS}(\hat{Y})$, under SRSWOR is given by

$$\hat{V}_{SRS}(\hat{Y}) = N^2\left(1-\frac{n}{N}\right)\sum_{i=1}^{n}(y_i-\bar{y}_n)^2/n(n-1)$$
$$= N^2\left(1-\frac{n}{N}\right)s^2/n$$

where $\bar{y}_n = \sum_{i=1}^{n} y_i / n$ is the sample mean and s^2 is an unbiased estimator of the population mean square S^2.

A similar approach applies when sampling is with replacement. In this case, there are N^n possible samples. The unbiased estimator of population total, sampling variance of the estimator and estimator of the sampling variance are given as

$$\hat{Y} = N\sum_{i=1}^{n} y_i / n = N\bar{y}_n \ , \quad V(\hat{Y}) = N^2\frac{\sigma^2}{n} \quad \text{and} \quad \hat{V}(\hat{Y}) = N^2\frac{s^2}{n}$$

where $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y}_N)^2$ is the population variance and s^2 is the sample mean square.

Consider all possible samples of size N which can be drawn from a given population. For a without replacement sampling scheme, there will be in all $\binom{N}{n}$ possible samples. For each sample, one can compute a statistic, such as the mean, standard deviation etc., which will vary from sample to sample. In this manner, one can obtain a distribution of the statistic which is called its sampling distribution.

From the above, it is clear that under Simple Random Sampling With Replacement (SRSWR),

i) the sample mean $(\bar{y}_n)$ is unbiased for the population mean $(\bar{Y}_N)$

ii) sample mean square (s^2) is unbiased for the population variance (σ^2)

iii) $V(\bar{y}_n) = \frac{\sigma^2}{n}$.

Like-wise, under Simple Random Sampling Without Replacement **(**SRSWOR),

i) the sample mean $(\bar{y}_n)$ is unbiased for the population mean $(\bar{Y}_N)$,

ii) sample mean square (s^2) is unbiased for the population mean square (S^2), and

iii) $V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$

8.3 Example

The data given below pertains to the average yield of wheat crop (in quintals) pertaining to 108 Villages in a Block of a District:

Village Sl. Nos.	**Yield (in quintals)**									
1-10	20	21	32	41	55	22	64	42	28	35
11-20	25	25	24	32	75	28	29	38	19	19
21-30	16	28	30	29	29	19	37	34	31	35
31-40	29	19	27	42	39	11	26	21	45	61
41-50	16	29	32	40	63	30	21	35	28	18
51-60	24	32	23	8	35	27	35	25	29	29
61-70	25	31	38	31	43	21	36	30	37	47
71-80	15	19	32	19	50	10	27	36	28	43
81-90	28	25	31	6	4	22	24	39	71	44
91-100	24	34	18	28	10	70	20	32	42	47
101-108	16	28	30	29	29	19	37	34		

a) Select a random sample of size 10 by simple random sampling without replacement (SRSWOR) and estimate the average yield along with its standard error on the basis of selected sample units.

b) Set up 95% confidence interval for the population mean.

SOLUTION:

As the population size N=108 is a three digit number, so for selecting a simple random sample of size n=10, we shall select three-digit random numbers from the Random Number Table (from 000 to 972, which is the highest multiple of 108 up to 999) as follows:

Random Number	**Sampling Unit Sl. No. (Remainder of Random Number/108)**	**Yield (q)**
120	12	25
572	32	19
649	01	20
211	103	30
327	03	32
673	25	29
153	45	63
317	101	16
586	46	30
943	79	28

- ✓ Estimate of Population Average yield = $\hat{\bar{Y}}_N = \bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{292}{10} = 29.2\,\text{q}$
- ✓ Estimate of population total is $= \hat{Y} = N x \bar{y}_n = 108 x 29.2 = 3153.6 q$
- ✓ The estimate of standard error of

$$\hat{Y} = S\hat{E}\left(\hat{Y}\right) = S\hat{E}\left(N.\bar{y}_n\right) = N.S\hat{E}\left(\bar{y}_n\right)$$

where $S\hat{E}\left(\bar{y}_n\right) = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}\, s$ and

$$s^2 = \frac{1}{n-1}\sum\left(y_i - \bar{y}_n\right)^2 = \frac{1}{10-1} \times 1533.6 = 170.4\ q^2$$

- ✓ So, $s = \sqrt{170.4} = 13.0537 q$.
- ✓ Hence,

$$S\hat{E}\left(\bar{y}_n\right) = \sqrt{\left(\frac{1}{10} - \frac{1}{108}\right)} x 13.0537 = 0.3012 x 13.0537 = 3.9322$$

- ✓ The 95% confidence interval for population mean is given by
  - $\bar{y}_n \pm t_{0.05/(10-1=9)df} \times S\hat{E}\left(\bar{y}_n\right) = 29.2 \pm 2.262 \times 3.9322 = 29.2 \pm 8.89$.

So, the 95% confidence interval for population mean is (29.2-8.89 to 29.2+8.89) i.e. (20.31, 38.09). It can be seen clearly that the population mean $\overline{Y}_N = \frac{3320}{108} = 30.74\,q$ is contained in this confidence interval. It may be mentioned here that out of total number of possible samples i.e. ${}^{108}C_{10}$, the population mean will be contained in such like confidence intervals corresponding to 95% of the total number of samples.

9. Conclusion

Simple random sampling and probability proportional size designs are most important uni-stage design. In most of the practical situations, complex sampling designs are utilized on the basis of these uni-stage sampling desisns. Stratified random sampling, multistage sampling, multiphase sampling, etc. are some examples of these complex designs.

References

1. Cochran, W.G. (1977). *Sampling techniques*. Wiley Eastern Ltd.
2. Des Raj, (1968). *Sampling theory*. Tata-Mcgraw-Hill Publishing Company Ltd.
3. Hansen, M. H. and Hurwitz, W. H. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, 14, 333-362.
4. Hansen, M. H., Hurwitz, W. H. and Madow, W. G. (1993). *Sample survey methods and theory*. **Vol. 1 and Vol. 2**, John Wiley & Sons, Inc.
5. Murthy, M. N. (1977). *Sampling theory and methods*. Statistical Publishing Society.
6. Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Ashok, C. (1984). *Sampling theory of surveys with applications*. Indian Society of Agricultural Statistics.

PLANNING AND EXECUTION OF SAMPLE SURVEYS

Prachi Misra Sahoo

ICAR- Indian Agricultural Statistics Research Institute, New Delhi -110012

1. Introduction

Sample surveys are widely used as a cost effective instrument of data collection and for making valid inferences about population parameters. Most of the steps involved while planning a sample survey are common to those for a complete enumeration. Three major stages of a survey are planning, data collection and tabulation of data. Some of the important aspects requiring attention at the planning stage are as follows:

1. formulation of data requirements - objectives of the survey
2. ad-hoc or repetitive survey
3. method of data collection
4. questionnaire versus schedules
5. survey, reference and reporting periods
6. problems of sampling frames
7. choice of sampling design
8. planning of pilot survey
9. field work
10. processing of data, and
11. preparation of report.
12. The different aspects listed above are inter-dependent.

(i) Formulation of Data Requirements

The users i.e. the persons or organizations requiring the statistical information are expected to formulate the objectives of the survey. The user's formulation of data requirements is not likely to be adequately precise from the statistical point of view. It is for the survey statistician to give a clear formulation of the objectives of the survey and to check up whether his formulation faithfully reflects the requirements of the users. The survey statistician's formulation of data requirements should include the following:

i. a clear statement of the desired information in statistical terms
ii. specification of the domain of study
iii. the form in which the data should be tabulated
iv. the accuracy aimed at in the final results and
v. cost of survey

Besides, these aspects, one may accommodate some additional items of information, directly or indirectly related to the objectives of the survey, which would provide checks on the accuracy of data or assist in interpreting the results.

(ii) Survey: Ad-hoc or Repetitive

An ad-hoc survey is one which is conducted without any intention of or provision for repeating it, whereas a repetitive survey is one, in which data are collected periodically for the same, partially replaced or freshly selected sample units. If the aim is to study only the current situation, the survey can be an ad-hoc one. But when changes or trends in some characteristics over time are of interest, it is necessary to carry out the survey repetitively.

(iii) Methods of Collecting Primary Data

There are varieties of methods that may be used to collect information. The method to be followed has to be decided keeping in view the cost involved and the precision aimed at. The methods usually adopted for collecting primary data are:

- Physical observation or measurement
- Direct Personal interview
- Mail enquiry
- Telephonic enquiry
- Web-based enquiry
- Method of registration
- Transcription from records
- Personal Digital Assistant (PDAs)

(a) Direct Personal Interview

The method of personal interview is widely used in social and economic surveys. In these surveys, the investigator personally contacts the respondents and can obtain the required data fairly accurately. The interviewer asks the questions pertaining to the objective(s) of survey and the information, so obtained, is recorded on a schedule prepared for the purpose. This method is mostly suitable for collecting data on conceptually difficult items from respondents. Under this method, the response rate is usually good and the information is more reliable and correct. However, more expenses and time is required to contact the respondents.

(b) Questionnaires sent through Mail

In this method, also known as mail inquiry, the investigator prepares a questionnaire and sends it by mail to the respondents. The respondents are requested to complete the questionnaires and return them to the investigator by a specified date. The method is suitable where respondents are spread over a wide area. Though the method is less expensive, normally it has a poor response rate. Usually, the response rate in mail surveys has been found to be about 40 per cent. The other problem with this method is that it can be adopted only where the respondents are literate and can understand the questions. They should also be able to send back their responses in writing. The success of the method depends on the skill with which the questionnaire is drafted, and the extent to which willing cooperation of the respondents is secured. For rural areas, this method has got its obvious limitation and is seldom used.

(c) Interview by Enumerators

This method involves the appointment of enumerators by the surveying agency. Enumerators go to the respondents, ask them the questions contained in the schedule, and then fill up the responses in the schedule themselves. For example, this method is used in collecting information during population census. For success of this method, the enumerators should be given proper training for soliciting co-operation of the respondents. The enumerators should be asked to carry with them their identity cards, so that the respondents are satisfied of their authenticity. They should also be instructed to be patient, polite, and tactful. This method can be usefully employed where the respondents to be covered are illiterate.

Telephone Interview

In case the respondents in the population to be covered can be approached by phone, their responses to various questions, included in the schedule, can be obtained over phone. If long distance calls are not involved and only local calls are to be made, this mode of collecting data may also prove quite economical. It is, however, desirable that interviews conducted over the phone are kept short so as to maintain the interest of the respondent.

Web-based Enquiry

Data collected by obtaining questionnaires posted on the net.

- Minimizes loss of data owing to incomplete or incorrectly completed data sets by using Client side validation.
- One of the fastest means of data collection.
- However, in developing countries where a large proportion of the population does not have access to Internet, the method of web-based enquiry may not serve the purpose for most of the surveys.
- Various Internet sites are using this method for opinion poll on certain issues.

Method of Registration

- ❖ The respondents are required to register the required information at specified places.
- ❖ The vital statistics registration system followed in many countries provide an illustration of the registration method.
- ❖ The main difficulty with this method, as in the case of the mail enquiry, is the possibility of non-response due to indifference, reluctance, etc. on the part of informants to visit the place of registration and supply the required data.

Transcription from records

- Used when the data needed for a specific purpose are already available in registers maintained in one or more places, making it no more necessary to collect them directly from the original units at much cost and effort.
- The method consists in compiling the required information from the registers for the concerned units.
- Extensively used since a good deal of government and business statistics are collected as by-product of routine administrative operations.

Personal Digital Assistant (PDAs)

- A personal digital assistant (PDA), also known as a handheld PC / palmtop computer, or personal data assistant, is a mobile device that functions as a personal information manager.
- Nearly all current PDAs have the ability to connect to the Internet.
- A PDA has an electronic visual display, enabling it to include a web browser, all current models also have audio capabilities enabling use as a portable media player, and also enabling most of them to be used as mobile phones. Most PDAs can access the Internet, intranets or extranets via Wi-Fi or Wireless Wide Area Networks.

Advantages of using Personal Digital Assistant (PDAs)

- Lightweight and easy to take anywhere.
- Online data transfer
- Online data supervision
- Online data scrutiny
- Reduce the time lag in data collection, scrutiny and entry
- Keep track of the enumerator

(iv) Questionnaire vs. Schedule

In the questionnaire approach, the informants or respondents are asked pre-specified questions and their replies to these questions are recorded by themselves or by investigators. In this case, the investigator is not supposed to influence the respondents. This approach is widely used in main enquiries. In the schedule approach, the exact form of the questions to be asked are not given and the task of questioning and soliciting information is left to the investigator, who backed by the training and instructions has to use his ingenuity in explaining the concepts and definitions to the informant for obtaining reliable information.

While planning a survey, preparation of questionnaire or schedules with suitable instructions needs to be given careful consideration. Respondent's bias and Investigator's bias are likely to be different in the two methods. Simple, unambiguous suitable wordings as well as proper sequence of questions are some considerations which contribute substantially towards reducing the respondents' bias. Proper training, skill of the Investigators, suitable instructions and motivation of investigators contribute towards reducing Investigator's bias.

(v) Survey, Reference and Reporting Periods

Another aspect requiring special attention is the determination of survey period, reference period and reporting periods.

i. **Survey Period**: The time period during which the required data is collected.

ii. **Reference Period**: The time period to which the collective data for all the units should refer.

iii. **Reporting Period**: The time period for which the required statistical information is collected for a unit at a time (reporting period is a part or whole of the reference period).

The reporting period should be decided after conducting suitable studies to examine recall errors and other non-sampling errors. For items of information subject to seasonal fluctuations, it is desirable to have one complete year as the survey and reference period, the data being collected every month or season with suitable reporting periods for the same or different sets of sample units.

(vi) Sampling Frames

A sampling frame is a list of all the items in your population. **It's a complete list of all the units one wants to study.** One of the most important practical problems in conducting sample surveys is that lists that can be used for selecting the samples are generally incomplete or out of date. Therefore, sample surveys can produce seriously biased estimates of the population parameters. Updating a list is a difficult and very expensive operation that has partially become easier due to the recent advances in managing databases. In any case, the single most important and expensive factor to be considered for updating a list is the data collection effort.

Types of Sampling Frames

There are many types of frames though the most common one is the list frame. Besides this there are area frames, Multi-Stage Frames, Frame for Series of Surveys. Each frames have its own advantages and disadvantages some of which are listed below.

- List Frame
- Area Frame
- Multi-Stage Frame
- Frame for Series of Surveys

Imperfectness in Sampling Frames

The Sampling frame is the key stone around which the sample is selected and the enumeration procedure must be determined. The nature & details of the frame become the basis for the choice of appropriate sampling design. Imperfection in frame arises due to two main reasons (i) Deviations in Coverage (ii) Deviation of content

Deviations in Coverage is mainly due to:

- Reporting units belonging to the target population are not included in the sampled population
- Reporting units belonging to the target population are contained in the sampled population more than once.
- Reporting units contained in the sampled population do not belong to the target population

Deviation of content is mainly due to following reasons:

- The frame provides incorrect auxiliary information on reporting unit.
- Auxiliary information for some of the reporting units is lacking in the frame.

Errors occurring in Sampling frame affecting the accuracy of estimates

- Incompleteness
- Non- Coverage

- Over – Coverage or Duplication
- Presence of Superfluous Units
- Incorrect or Inaccurate Information
- Non / Incomplete Auxiliary Information
- Non-Response (Out of scope)

Construction of sampling frames

The construction of sampling frame is to be done taking utmost care. First one should be clear regarding the choice of frame units and various parameters related to the units. Once this is done, the frame is developed and after the development of frame, its validation is required. These three major points to be considered during construction of sampling frames are mentioned below:

Choice of frame units

- Cost consideration in establishing and maintaining
- Availability of type of information for frame units
- Stability of frame units over time
- Time needed to construct frame

Development of frame

Construction of database including maps for area frames

Validation of frame

- Coverage achieved
- Quality of information

Maintenance and Updation of sampling frames

- Removing duplicates
- Removing 'deaths', such as
    - Closed establishments
    - Burned down or demolished housing units
- Incorporating 'births', such as
    - New establishments
    - New housing units in enumeration areas
- Updating auxiliary information
- To reflect population changes so it continues to be 'representative'
    - Prepare new listings of households in sample clusters
    - Periodical update of entire frame to account for post-censual high-growth areas

(vii) Choice of Sampling Design

The choice of a suitable sampling design for a given survey situation is one of the most important step in the process of planning sample surveys. The principle generally adopted in the choice of a design is either reduction of overall cost for a pre-specified permissible error or reduction of margin of error of the estimates for given fixed cost. Generally, a stratified uni-stage or multi-stage design is adopted for large scale surveys. For efficient planning, various auxiliary information, which is normally available, is utilized at various stages e.g. the area under particular crop as available for previous years is normally used for size stratification of villages. If the information is available for each and every unit of the population and there is wide variability in the information then it may be used for selecting the sample through probability proportional to size methods The choice of sample units, method of selecting sample and determination of sample size are some of the important aspects in the choice of proper sample design.

(viii) Pilot Surveys

Where some prior information about the nature of population under study, and the operational and cost aspects of data collection and analysis is not available from past surveys. It is desirable to design and carry out a pilot survey. It will be useful for

i. testing out provisional schedules and related instructions,

ii. evolving suitable procedure for field and tabulation work, and

iii. training field and tabulation staff.

(ix) Field Work

While planning the field work of the survey, a careful consideration is needed regarding choice of the field agency. For ad-hoc surveys, one may plan for ad-hoc staff but if survey is going to be a regular activity, the field agency should also be on a regular basis. Normally for regular surveys, the available field agency is utilized. A regular plan of work by the enumerators along with proper supervision is an important consideration for getting a good quality of data.

(x) Processing of Survey Data

The analysis of data collected in a survey has broadly two facets:

i. tabulation and summary of data and

ii. subject analysis.

The first task, which is of primary importance, is the reduction of collected data into meaningful tables. The tables should be presented along with the background information such as the objective(s) of the survey, the sampling design adopted, method used for data collection and tabulation, and margin of error applicable to the results. These margins of error provide the idea about the precision of estimates.

Subject analysis to be taken up after preparing summary tables, should include cross tabulation of data by the meaningful, geographical, economy, demographic or other breakdowns to study their relationship and trends among various characteristics. This is a detailed technical analysis and is likely to be time consuming. Hence, this part should not be tied up with the first part as otherwise the publication of the survey results might get delayed.

(xi) Preparation of Report

Although there are no set guidelines for presentation of results and preparation of report, however some points, which serve as guidelines in the preparation of sample survey reports, are given below:

i. Introduction & review of literature
ii. Objective(s)
iii. Scope
iv. Subject coverage
v. Method of data collection
vi. Survey references and recording
vii. Sampling design and estimation procedure
viii. Tabulation procedure
ix. Presentation of results
x. Activity of results
xi. Cost structure of the survey
xii. Agency for conducting the survey
xiii. References

2. Questionnaire Designing

Questionnaires and schedules are forms for recording the information as envisaged under the survey. Designing of these is one of the most important aspects of the survey. The words 'questionnaire' and 'schedule' as per the current practice are generally used synonymously. However, a technical distinction is sometimes made. The term questionnaire applies to forms distributed through mails or given to informants to be filled in, by and large, without the assistance or supervision of the interviewer, while a schedule is the form carried and filled in by the investigator or filled in his presence.

The question as to whether the questionnaire or schedule approach is to be used in a survey for collecting the required information needs consideration. In the former approach the respondents are asked pre-specified questions and their replies to these questions are recorded by themselves or by the investigators. This approach presumes that the respondents are capable of understanding and answering the questions, since in this case the investigator is not supposed to influence the responses in any way by his interpretation of the terms used in the form. This method is widely used in mail inquiries. In the schedule approach, the exact form of the questions to be asked are not given and the task of questioning and eliciting information is left to the investigator, who backed by his training, experience and instructions has to use his ingenuity in explaining the concepts and definitions to the informants for obtaining reliable information. Detailed instructions are, however, given to the investigator about concepts, definitions and procedures to be used in collecting data for the survey. In various socio-economic surveys, the method of collecting data after meeting the respondents and obtaining information of various characters by inquiry is commonly used.

From the above, it may appear that the schedule approach is subject to more investigator bias than the questionnaire approach, as there is added scope in it for the investigator to

influence the responses of the informants. This will not be so, if well-trained and skilled investigators are employed for the purpose. On the other hand, the respondent bias may be substantial in questionnaire approach, if the survey items are complicated and involve conceptual difficulties. In such a situation, it would be desirable to train investigators for explaining the terms involved rather than to burden the respondent with elaborate instructions and clarifications. As the cost of questionnaire approach is generally less than that of schedule approach, a decision as to which of the two methods should be followed in a particular survey needs to be arrived at after carefully examining the possible effects of investigator and respondent biases and the cost involved.

Designing of schedules/questionnaires with suitable instructions needs to be given careful consideration in planning a survey as utility of the results of the survey depends to a large extent on this. The framing of schedules or items should be done in a simple, unambiguous, interesting and tactful manner and they should be so worded as not to influence the answers of the respondents. The sequence of items is equally important. Those likely to help the investigator in establishing a good rapport with the respondents should be put first and item relating to a particular aspect of the survey should come together in a schedule/questionnaire. As far as possible the items should be such that the answers can be recorded in numbers or specific codes.

To reduce the non-sampling errors arising from ambiguous definitions and misunderstanding of the questions by investigators/respondents, it is desirable to give some typical examples, detailed explanatory notes and instructions for the items of information included in the schedule/questionnaire. Clarification of doubts raised by the investigators is to be done in such a manner that there is uniformity in the procedures followed by different investigators.

From what has been discussed above, it will appear that there are several considerations, which have to be kept in mind while designing the schedules. It is difficult to list out all of them. There may be some which are specific to a particular survey and may require special consideration. In the following paragraphs the main important considerations, which should be borne in mind while designing the schedule/questionnaire, are given.

2.1 Three Kinds of Schedule Items

The information included in the schedule may be classified under the following three headings:

2.1.1 Identification Information

This ensures that the schedule will not be misplaced or mixed-up, lost or duplicated; that the information on it pertains to the particular sample case, and the interviewer and respondent can be identified e.g. year, season, crop, name of the district, block, village, name of cultivator and his father's name etc. are entered against identification particulars.

2.1.2 Social Background or Census Type Factual Data

This information about respondent provides the variables by which the survey data are to be classified and also the basis for evaluating the sample viz. cultivator's total holding and holding size group, category namely, SC, ST, or General, monthly income, total number of family members, tenancy status, educational qualifications etc.

2.1.3 Questions on the Subject of the Survey

These questions may be directed towards obtaining more or less objective facts or towards revealing attitudes and opinions on matters of current interest.

2.2 Considerations to be borne in Mind while Designing Schedules/Questionnaires

The first step in designing a schedule/questionnaire is to define the problem to be tackled by the survey and hence to decide on what questions to be asked. The temptation is always to cover too much, to ask everything that might turn out to be interesting. This must be resisted. Lengthy questionnaires are as demoralizing for the interviewer as for the respondent, and the questionnaire should be no longer than is absolutely necessary for the purpose.

2.3 Agency which will Make the Entries in the Schedules

If a highly trained investigator is to ask the questions and enter the replies, the form should be different from the one drawn for informant to fill out himself since the interviewer can be instructed regarding details which will ensure uniform definitions, entries and interpretations.

The terminology and questions should be adapted to the type of people who will give the information. For example, a questionnaire addressed to specialist familiar with the subject matter of the survey can be much more technical than the one directed to a cross-section of the population. In designing schedules that are to be filled up by farmers, housewives, employers etc. The level of education should be taken into consideration.

2.4 Physical Appearance of the Schedule and Cooperation Received for the Survey

In surveys by mail, there is no doubt that an attractive looking questionnaire is a selling point for cooperation. Consequently, an unattractive one may cause the recipient to put it aside or even throw it. The fact that the form looks 'short', however, often contributes to securing individual's consent to be interviewed. Informants will tolerate a short interruption of only to get rid of the interviewer, but they may flatly refuse to answer a long list of questions.

2.5 How are the Questions to be worded?

The choice of the language used in expressing a question is of the greatest importance. It is too often presumed that the respondents must be aware of the concepts and definitions used in the questionnaire since these are obvious to the survey team. If the terminology is ambiguous, the respondents will have to use their own judgment and different persons will judge differently. This causes confusion and errors. Ambiguity arises with double barreled questions, such as, the following question to a public transport "Do you like travelling on trains and buses"? Respondent liking one and disliking other would be in a dilemma in answering this question. Clearly it needs to be divided into two questions.

2.5.1 Use simple words which are familiar to all potential informants

The basic principle in good question wording is to use the simplest words that will convey the exact meaning. Meaning of the questions becomes clear when the words used are well known and mean the same thing to everyone. The question 'Do you operate land?' used in agricultural surveys is poor. It is not clear whether the person is an owner cultivator or a tenant cultivator.

2.5.2 Make the Questions as Concise as Possible

A question that contains long dependent or conditional clauses may confuse the informants. In trying to comprehend the question as a whole he may over-look or forget clause and hence his answer may be wrong. However, in opinion or attitude survey, it may be important to have the complete question printed on the schedule.

2.5.3 Formulate the Question to Yield Exactly the Information Desired

The question should be self-explanatory. If the questions call for an answer in terms of units, these units must be clearly defined. Suppose we want to ask the cultivator the seed rate used for a specific crop. We should clearly mention whether the seed rate is to be reported in kg/ha or in kg for the entire field.

2.5.4 Avoid Multiple Meaning Questions

Unless each question covers only one point, there will be confusion as to which one, the answer applies to. Such items should be formulated as two or more questions so that separate answer can be secured.

2.5.5 Avoid Ambiguous Questions

A question which means different things to different people is ambiguous. The best course is to pre-test the questions through a pilot survey and thus detect ambiguities e.g. in a survey on consumption of milk and ghee, suppose the question is about the quantity of milk and ghee consumed during the month. It should be clearly mentioned whether it is during last calendar month or one month prior to Investigator's visit.

2.5.6 Avoid Leading Questions

A leading question is one which, by its content, structure or wording leads the respondent in the direction of a certain answer. In other words, all questions which produce biased answers may be regarded as leading questions. Such questions should be avoided.

2.5.7 Keep to a Minimum the Amount of Writing Required on the Schedule

When feasible, use symbols for the replies. Explain these symbols somewhere on the schedule. If the possible responses can be foreseen by pre-testing, the questions can be answered as Yes or No, by writing a number, by putting a cross, by putting a symbol or by encircling the correct answer.

2.5.8 Include a Few Questions that will serve as Checks on the Accuracy and Consistency of the Questions as a Whole

Two questions that bring out the same facts though worded differently and placed in different sections of the schedule, serve to check the internal consistency of the replies, e.g., in a socio-economic survey, suppose we are asking the total holding of the farmer. It would be better if we include the area owned, leased-in and leased-out separately in some block of the proforma. This serves as a check to tally the total holding size.

2.6 Handbook of Instructions for the Field Staff

It would be desirable to prepare a comprehensive handbook of instructions explaining concepts and definitions of various items for filling in the questionnaire/schedule under the survey and a copy of the same should be supplied to each Field Investigator.

2.7 Sequence of Questions

Careful consideration should be given to the problem of the order in which questions should appear. In order to guard against confusion and misunderstanding, questions should be arranged logically, one question leading to the next. Specific questions should always follow general questions. The opening question should be very interesting; this will ensure that the respondents cooperate in parting with the desired information for the survey. Questions which might embarrass the respondents should be placed towards the middle or end of the questionnaires. Questions with an emotional tinge may be interspersed between items which elicit more neutral reactions.

3. Conclusion

The problem of designing of questionnaires/schedules is not an easy task. Even if one follows all the accepted principles, there usually remains a choice of several question forms, each of which seems satisfactory. Every surveyor tries to phrase his questions in simple, everyday language, to avoid vagueness and ambiguity and to use neutral wording. His difficulty lies in judging whether, with any particular question, he has succeeded in these aims. He may appreciate perfectly that leading questions are to be avoided but how can he know which words will be 'leading' with the particular question, survey and population that confront him, perhaps for the first time?

The answer to this question lies in detailed pre-tests and pilot studies, more than anything else, they are the essence of a good questionnaire. However the experienced questionnaire designer is, any attempt to shortcut these preparatory stages will seriously jeopardize the quality of the questionnaire; past experience is a considerable asset, but in a fresh survey, there are always new aspects which may perhaps not be immediately recognized, but which exist and must be investigated through pre-tests and pilot studies.

References

1. Cochran, W.G. (1977). Sampling Techniques. Third Edition. John Wiley and Sons.
2. Des Raj (1968). Sampling Theory. TATA McGRAW-HILL Publishing Co. Ltd.
3. Des Raj and Chandok, P. (1998): Sample Survey Theory. Narosa Publishing House.
4. Murthy, M.N. (1977). Sampling Theory and Methods. Statistical Publishing Society, Calcutta.
5. Singh, D. and Chaudhary, F.S. (1986). Theory and Analysis of Sample Survey Designs. Wiley Eastern Limited.
6. Singh, D., Singh, P. and Kumar, P. (1978). Handbook of Sampling Methods. I.A.S.R.I., New Delhi.
7. Singh, R. and Mangat, N.S. (1996). Elements of Survey Sampling, Kluwer Academic Publishers.
8. Sukhatme, P.V. and Sukhatme, B.V. (1970). Sampling Theory of Surveys with Applications. Second Edition. Iowa State University Press, USA.
9. Sukhatme, P. V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). Sampling Theory of Surveys with Applications. Third Revised Edition, Iowa State University Press, USA.

STRATIFIED AND MULTISTAGE SAMPLING IN AGRICULTURAL SURVEYS

Kaustav Aditya

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Stratified Sampling

1.1 Introduction

The basic idea in stratified random sampling is to divide a heterogeneous population into sub-populations, usually known as strata, each of which is internally homogeneous in which case a precise estimate of any stratum mean can be obtained based on a small sample from that stratum and by combining such estimates, a precise estimate for the whole population can be obtained. Stratified sampling provides a better cross section of the population than the procedure of simple random sampling. It may also simplify the organization of the field work. Geographical proximity is sometimes taken as the basis of stratification. The assumption here is that geographically contiguous areas are often more alike than areas that are far apart. Administrative convenience may also dictate the basis on which the stratification is made. For example, the staff already available in each range of a forest division may have to supervise the survey in the area under their jurisdiction. Thus, compact geographical regions may form the strata. A fairly effective method of stratification is to conduct a quick reconnaissance survey of the area or pool the information already at hand and stratify the forest area according to forest types, stand density, site quality etc. If the characteristic under study is known to be correlated with a supplementary variable for which actual data or at least good estimates are available for the units in the population, the stratification may be done using the information on the supplementary variable. For instance, the volume estimates obtained at a previous inventory of the forest area may be used for stratification of the population.

In stratified sampling, the variance of the estimator consists of only the 'within strata' variation. Thus the larger the number of strata into which a population is divided, the higher, in general, the precision, since it is likely that, in this case, the units within a stratum will be more homogeneous. For estimating the variance within strata, there should be a minimum of 2 units in each stratum. The larger the number of strata the higher will, in general, be the cost of enumeration. So, depending on administrative convenience, cost of the survey and variability of the characteristic under study in the area, a decision on the number of strata will have to be arrived at.

1.2 Allocation and Selection of the Sample within Strata

Assume that the population is divided into k strata of $N_1, N_2, \ldots, N_k$ units respectively, and that a sample of n units is to be drawn from the population. The problem of allocation concerns the choice of the sample sizes in the respective strata, *i.e.*, how many units should be taken from each stratum such that the total sample is n.

Other things being equal, a larger sample may be taken from a stratum with a larger variance so that the variance of the estimates of strata means gets reduced. The application of the

above principle requires advance estimates of the variation within each stratum. These may be available from a previous survey or may be based on pilot surveys of a restricted nature. Thus, if this information is available, the sampling fraction in each stratum may be taken proportional to the standard deviation of each stratum.

In case the cost per unit of conducting the survey in each stratum is known and is varying from stratum to stratum an efficient method of allocation for minimum cost will be to take large samples from the stratum where sampling is cheaper and variability is higher. To apply this procedure one needs information on variability and cost of observation per unit in the different strata.

Where information regarding the relative variances within strata and cost of operations are not available, the allocation in the different strata may be made in proportion to the number of units in them or the total area of each stratum. This method is usually known as 'proportional allocation'.

For the selection of units within strata, In general, any method which is based on a probability selection of units can be adopted. But the selection should be independent in each stratum. If independent random samples are taken from each stratum, the sampling procedure will be known as 'stratified random sampling'. Other modes of selection of sampling such as systematic sampling can also be adopted within the different strata.

Stratification, if properly done as explained in the previous sections, will usually give lower variance for the estimated population total or mean than a simple random sample of the same size. However, a stratified sample taken without due care and planning may not be better than a simple random sample.

2. Multistage Sampling

2.1 Introduction

Cluster sampling is a sampling procedure in which clusters are considered as sampling units and all the elements of the selected clusters are enumerated. One of the main considerations of adopting cluster sampling is the reduction of travel cost because of the nearness of elements in the clusters. However, this method restricts the spread of the sample over population which results generally in increasing the variance of the estimator. In order to increase the efficiency of the estimator with the given cost it is natural to think of further sampling the clusters and selecting more number of clusters so as to increase the spread of the sample over population. This type of sampling which consists of first selecting clusters and then selecting a specified number of elements from each selected cluster is known as sub- sampling or two stage sampling, since the units are selected in two stages. In such sampling designs, clusters are generally termed as first stage units (fsu's) or primary stage units (psu's) and the elements within clusters or ultimate observational units are termed as second stage units (ssu's) or ultimate stage units (usu's). It may be noted that this procedure can be easily generalized to give rise to multistage sampling, where the sampling units at each stage are clusters of units of the next stage and the ultimate observational units are selected in stages, sampling at each stage being done from each of the sampling units or clusters selected in the previous stage. This procedure, being a compromise between uni-stage or direct sampling of units and cluster sampling, can be expected to be (i) more efficient than uni-stage sampling and less efficient than cluster sampling from considerations

of operational convenience and cost, and (ii) less efficient than uni-stage sampling and more efficient than cluster sampling from the view point of sampling variability, when the sample size in terms of number of ultimate units is fixed.

It may be mentioned that multistage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is prohibitive or where the cost of locating and physically identifying the usu's is considerable. For instance, for conducting a socio-economic survey in a region, where generally household is taken as the usu, a complete and up-to-date list of all the households in the region may not be available, whereas a list of villages and urban blocks which are group of households may be readily available. In such a case, a sample of villages or urban blocks may be selected first and then a sample of households may be drawn from each selected village and urban block after making a complete list of households. It may happen that even a list of villages is not available, but only a list of all tehsils (group of villages) is available. In this case a sample of households may be selected in three stages by selecting first a sample of tehsils, then a sample of villages from each selected tehsil after making a list of all the villages in the tehsil and finally a sample of households from each selected village after listing all the households in it. Since the selection is done in three stages, this procedure is termed as three stage sampling. Here, tehsils are taken as first stage units (fsu's), villages as second stage units (ssu's) and households as third or ultimate stage units (tsu's).

One of the advantages of this type of sampling is that at the first stage the frame of fsu's is required which is generally easily available and at the second stage the frame of ssu's is required for the selected fsu's only and so on. Moreover, this method allows the use of different selection procedures in different stages. It is because of these considerations that multistage sampling is used in most of the large scale surveys. It has been found to be very useful in practice. It is noteworthy that Prof. P.C. Mahalanobis used this sampling procedure in crop surveys carried out in Bengal during the period 1937-1941, and he had termed this procedure as nested sampling. Cochran (1939) and Hansen and Hurwitz (1943) have considered the use of this procedure in agricultural and population surveys respectively. Lahiri (1954) discussed the use of multistage sampling in the Indian Sample Survey.

2.2 Two Stage Sampling With Equal Probabilities, Equal First Stage Units

2.2.1 Estimation of population mean

Let the population under study consists of NM elements grouped into N first stage units, each first stage unit containing M second stage units.

Let us denote

Y_{ij} = the value of the characteristic under study for the j-th second stage unit of the i-th first stage unit, j = 1,2,…,M; i= 1,2,…,N

$\bar{Y}_{i.} = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}$, population mean of i-th fsu,

$\bar{Y}_{..} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij} = \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_{i.}$, the population mean.

Further, let a sample of size nm is selected by first selecting n fsu's from N fsu's by simple random sampling without replacement (srswor) and then selecting m ssu's from M ssu's by srswor from each of the selected fsu's. Let us denote

$\bar{y}_{im} = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$, sample mean based on m selected ssu's from the i-th fsu,

$\bar{y}_{nm} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{im}$, the sample mean based on all nm units in the sample.

Clearly, $\bar{y}_{nm}$ is an unbiased estimator of $\bar{Y}_{..}$ with its variance given by

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 \qquad \text{.....(3)}$$

where

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2 \quad \text{and} \quad \bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

The estimator of $V(\bar{y}_{nm})$ is given by

$$\hat{V}(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{s}_w^2 \qquad \text{.....(4)}$$

where

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_{im} - \bar{y}_{nm}\right)^2 \quad \text{and} \quad \bar{s}_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_{im}\right)^2$$

It is observed that the variance of sample mean ($\hat{\bar{Y}}$) in two stage sampling consists of two components, the first representing the contribution arising from sampling of first stage units and the second arising from sub-sampling within the selected first stage units. We note the following two cases:

Case (i) n = N, corresponds to stratified sampling with N first stage units as strata and m units drawn from each stratum.

Case (ii) m = M, corresponds to cluster sampling.

References

1. Cochran, W.G., (1939). The use of analysis of variance in enumeration by sampling; *Jour. Amer. Statist. Assoc.*, **34**, 492-510.
2. Cochron, W.G., (1977). Sampling techniques; Wiley Eastern Ltd.
3. Des Raj, (1968). Sampling theory; Tata-Mcgraw-Hill Publishing Company Ltd.
4. Murthy, M.N., (1977). Sampling theory and methods; Statistical Publishing Society
5. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984). Sampling theory of surveys with applications; Indian Society of Agricultural Statistics.

PROBABILITY PROPORTIONAL TO SIZE (PPS) SAMPLING

Ankur Biswas and Raju Kumar

ICAR-Indian Agricultural Statistics Research Institute New Delhi -110012

1. Introduction

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conservation and controversy require numerical data for their resolution. The data collected and analyzed in an objective manner and presented suitably serve as a basis for taking policy decisions in different fields of daily life.The important users of statistical data, among others, include government, industry, business, research institutions, public organizations and international agencies and organizations. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. Data on agricultural production are of immense use to the state for planning to feed the nation.

In sampling theory if the auxiliary information, related to the character under study, is available on all the population units, then it may be advantageous to make use of this additional information in survey sampling. One way of using this additional information is in the sample selection with unequal probabilities of selection of units. The knowledge of auxiliary information may also be exploited at the stratification and estimation stage.

2. Sampling with Varying Probability

Under certain circumstances, selection of units with unequal probabilities provides more efficient estimators than equal probability sampling, and this type of sampling is known as unequal or varying probability sampling. In the most commonly used varying probability sampling scheme, the units are selected with probability proportional to a given measure of size (PPS) where the size measure is the value of an auxiliary variable x related to the characteristic y under study and this sampling scheme is termed as probability proportional to size sampling. For instance, the number of persons in some previous period may be taken as a measure of the size in sampling area units for a survey of socio-economic characters, which are likely to be related to population. Similarly, in estimating crop characteristics the geographical area or cultivated area for a previous period, if available, may be considered as a measure of size, or in an industrial survey, the number of workers may be taken as the size of an industrial establishment.

Since a large unit, that is, a unit with a large value for the study variable y, contributes more to the population total than smaller units, it is natural to expect that a scheme of selection which gives more chance of inclusion in a sample to larger units than to smaller units would provide estimators more efficient than equal probability sampling. Such a scheme is provided by PPS sampling, size being the value of an auxiliary variable x directly related to y. It may appear that such a selection procedure would give biased estimators as the larger units are over-represented and the smaller units are under-represented in the sample. This would be so, if the sample means is used as an estimator of population mean. Instead, if the sample observations are suitably weighted at the estimation stage taking into consideration their probabilities of selection, it is possible to obtain unbiased estimators. Mahalanobis (1938) has referred to this procedure in the

context of sampling plots for a crop survey and this procedure has been discussed in detail by Hansen and Hurwitz (1943).

There is a basic difference between simple random sampling and pps sampling procedures. In simple random sampling, the probability of drawing any specified unit at any given draw is the same, while in pps sampling, it differs from draw to draw. The theory of pps sampling is consequently more complex than that of simple random sampling. In PPS sampling, the units may be selected with or without replacement. We shall discuss the theory appropriate to pps sampling with replacement (PPS wr) and pps sampling without replacement (PPS wor) in following sections.

3. Sample selection procedures under PPS sampling with replacement

i) Cumulative Total Method

To draw a sample of size n from a population of size N with probability proportional to size and with replacement, we proceed as follows:

If X_i is an integer proportional to the size of the i-th unit, i = 1,2,…N, we form successive totals X_1 , X_1+X_2, $X_1+X_2+X_3$, …, $\sum_{i=1}^{N} X_i$. Draw a random number R not exceeding $\sum_{i=1}^{N} X_i$ from a table of random numbers. If $X_1+X_2+\ldots+X_{i-1} < R \leq X_1+X_2+\ldots+X_i$, the i-th unit is selected. Repeat the procedure n times to get a sample of size n.

The main disadvantage of this method is that it involves cumulation of the sizes and writing down of the cumulative totals, which is time consuming and costly when N is large. For instance, if this method is used for selecting a sample of factories with probability proportional to the number of workers from the population of about 45,000 factories in India or for selecting a PPS sample of farms or fields with area as the size from a large number of such units, the selection operation becomes prohibitively costly. A procedure which avoids the need for calculating cumulative totals for each unit, is considered in the next sub-section. Of course, the work of cumulation is simple when the population is small.

ii) Lahiri's Method

Lahiri suggested a method of PPS selection in the year 1951, which does not require cumulation of sizes at all. In this approach a pair of random numbers, say (i,j) is selected such that $1 \leq i \leq N$ and $1 \leq j \leq M$, where M is the maximum of the sizes of the N units in the population. If $j \leq X_i$, the i-th unit is selected, otherwise it is rejected and another pair of random numbers is chosen. For selecting a sample of n units with probability proportional to size and with replacement, the procedure is to be repeated till n units are selected. It can be seen that the method leads to the required probability of selection.

Example 1. A village has 10 holdings consisting of 50,30,45,25,40,26,24,35, 28 and 27 fields, respectively. Select a sample of four holdings with the replacement method and with probability proportional to the number of fields in the holding. The first step in the selection of holdings is to form cumulative totals as shown below.

S.N. of holdings	Size (Xí)	Cumulative Size	Numbers associated
1	50	50	1-50
2	30	80	51-80
3	45	125	81-125
4	25	150	126-150
5	40	190	151-190
6	26	216	191-216
7	44	260	217-260
8	35	295	261-295
9	28	323	296-323
10	27	350	324-350

To select a holding, a random number not exceeding 350 is drawn with the help of a random number table. Suppose the random number selected is 123. It can be seen from the cumulative totals that the number is associated with the group 81-125, i.e. the 3rdholding is selected corresponding to the random number 123. Similarly, 3 more random numbers need to be selected. Suppose, these numbers are 346, 165 and 094. Then the holdings selected corresponding to these random numbers are the 10th, 5thand 3rd, respectively. Hence, a sample of 4 holdings selected with probability proportional to size will contain the 3rd, 10th, 5th, and 3rd Holdings.

Here, N= 10, m= 50. For selection by Lahiri's method, first, we have to select a random number which is not greater than 10 and a second random number which is not greater than 50. Referring to the random number table, the pair is (10,13). Hence , the 10th unit is selected in the sample. Similarly, choosing other pairs , we can have (4, 26), (5,35), (7,26). The pair (4,26) is rejected as 26 is greater than the size value (25) and so another pair is drawn which turns out to be (8, 16). Hence, the sample will consist of the holdings with serial numbers 10, 5, 7, and 8.

4. Estimation of the population mean ($\bar{Y}$)

Let y be the characteristic under study and denote by Y_i , the y-value for the i-th unit in the population, i=1,2,…,N. P_i be the probability of selecting the i-th unit in the population.Obviously, $\sum_{i=1}^{N} P_i = 1$. We shall now consider the problem of estimating the population mean $\bar{Y}$ based on a sample of n units selected with probabilities P_i and with replacement.

Let, $Z_i = \frac{Y_i}{N P_i}$, i = 1,2,…,N.

As an estimator of $\overline{Y}$, consider

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{NP_i}$$

$\overline{z}$ is an unbiased estimator of $\overline{Y}$ as $E(\overline{z}) = \overline{Y}$.

Further, since the units are selected with replacement,

$$V(\overline{z}) = \frac{\sigma_z^2}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{N} P_i \left(\frac{Y_i}{N\,P_i} - \overline{Y} \right)^2$$

Estimate of variance is given by

$$\hat{V}(\overline{z}) = \frac{s_z^2}{n} \text{ as } E(s_z^2) = \sigma_z^2 \text{ where } s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n} (z_i - \overline{z})^2 .$$

Example 2. A sample survey was conducted to study the yield of wheat in Haryana. These dataset is available in Singh and Chaudhary (1986). A sample of 20 farms from a total of 100 was taken, with probability proportional to the area under wheat crop, with replacement method. The total area under wheat crop (X) was 484.5 hectares. The area under crop (x) and yield (y) were noted in hectares and quintals per hectare, respectively.The sample selected by the cumulative methods was

Area under Crop	4.8	4.1	1.3	5.2	6.9	6.0	2.0	6.3	5.2	4.2
Yield of Crop	22	19	6	25	54	43	4	40	28	29
Area under Crop	4.8	5.9	5.8	5.8	5.1	4.7	5.6	5.2	4.0	4.6
Yield of Crop	22	39	39	44	30	27	34	31	18	31

Here, N= 100, n = 20, X = 484.5

The estimate of average yield/farm is given by

$$\hat{\overline{Y}}_{PPS} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{NP_i} = \frac{X}{nN}\sum_{i=1}^{n} \frac{y_i}{x_i} = 484.5 \times 1205930/100 \times 20 = 29.11.$$

The estimate of ariance of the estimator is

$$\hat{V}\left(\hat{\overline{Y}}_{PPS}\right) = \frac{1}{20 \times 19}\left[\frac{(484.5)^2 \times 728.142}{100 \times 100} - 20(29.11)^2 \right] = 2.00$$

The standard error of $\hat{\overline{Y}}_{PPS} = \sqrt{2.00} = 1.41.$

The estimate of variance of the estimate based on simple random sample on assumption of the PPS sample is given by

$$\hat{V}\left(\hat{\bar{Y}}_{SRS}\right)=\frac{100}{20}\left[\frac{484.4\times 4156.05}{20\times(100)^2}-\frac{1}{100}\times\left\{(29.11)^2-2.00\right\}\right]=7.64$$

Hence, the percentage gain in precision due to PPS sampling as compared to SRSWR, is given by

$$\text{Percentage gain in precision} = \frac{\hat{V}\left(\hat{\bar{Y}}_{SRS}\right)-\hat{V}\left(\hat{\bar{Y}}_{PPS}\right)}{\hat{V}\left(\hat{\bar{Y}}_{PPS}\right)}\times 100$$

$$= \frac{7.64-2.00}{2.00}\times 100 = 282\,.$$

5. PPS sampling without replacement

It is generally observed that sampling without replacement provides a more efficient estimator than sampling with replacement, since the effective sample size is more in the former than in the latter. Considerable development has taken place in the field of sampling with varying probabilities without replacement since 1950. But most of the suggested procedures, estimators and variance estimators are rather complicated and hence these are not commonly used in practice, especially in large-scale sample surveys with a small sampling fraction, since in such cases the efficiencies of sampling with and without replacement are not likely to differ much. However, it may be worthwhile to use these procedures of selection and estimation, if the sampling fraction is moderately large, as in that case the gain in efficiency in sampling without replacement is likely to be substantial.

5. Conclusion

Simple random sampling and probability proportional size designs are most important uni-stage design. In most of the practical situations, complex sampling designs are utilized on the basis of these uni-stage sampling desisns. Stratified random sampling, multistage sampling, multiphase sampling, etc. are some examples of these complex designs.

References

Cochran, W.G. (1977). *Sampling Techniques*. Wiley Eastern Ltd.

Des Raj. (1968). *Sampling Theory*. Tata-Mcgraw-Hill Publishing Company Ltd.

Hansen, M.H. and Hurwitz, W.H. (1943). On the theory of sampling from finite populations.*Ann. Math. Statist.*, 14, 333-362.

Hansen, M.H., Hurwitz, W.H. and Madow, W.G. (1993). *Sample Survey Methods and Theory*. Vol. 1 and Vol. 2, John Wiley & Sons, Inc.

Murthy, M.N. (1977). *Sampling Theory and Methods*. Statistical Publishing Society.

Singh, D. and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. Wiley Eastern Ltd.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Indian Society of Agricultural Statistics.

RATIO AND REGRESSION METHODS OF ESTIMATION IN SAMPLE SURVEYS

Kaustav Aditya and Deepak Singh

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

In sampling theory the auxiliary information is being utilized in following ways:

- Utilization of information at pre-selection stage i.e. for stratifying the population.
- Utilization of information at selection stage i.e. in selecting the units with probabilities proportional to some suitable measure of size (size being based on some auxiliary variables).
- Utilization of information at estimation stage i.e. in formulation of the ratio-type, regression, difference and product estimators etc.
- Auxiliary information may also be utilized in mixed ways.

Usually the information available is in the form that:

- The values of the auxiliary character(s) are known in advance for each and every sampling unit of the population.
- The population total(s) or mean(s) of auxiliary character(s) are known in advance.
- If it is desired to stratify the population according to the values of some variate x, their frequency distribution must be known.

The use of auxiliary information at estimation stage in the formation of ratio-type and regression estimators and sampling scheme providing unbiased regression estimator has been discussed in the following sections.

In sample surveys, many a time the characteristic y under study is closely related to an auxiliary characteristic x, and data on x are either readily available or can be easily collected for all the units in the population. In such situations, it is customary to consider estimators of population mean $\overline{Y}_N$ of survey variable y that use the data on x and are more efficient than the estimators which use data on the characteristic y alone. The fact that the data on the auxiliary variable can be used even at a later stage after selecting the sample, encourages such procedures. Two types of these commonly used methods are as follows:

- the ratio-type method of estimation
- the regression method of estimation

2. Ratio-Type Method of Estimation

Let a sample of size n be drawn by SRSWOR (Simple random sampling without replacement) from a population of size N. Denote by

y_i = the value of the characteristic under study for the i^{th} unit of the population,

x_i = the value of the auxiliary characteristic on the i^{th} unit of the population,

Y = the total of the y values in the population,

X = the total of the x values in the population,

$r_i = \frac{y_i}{x_i}$, the ratio of y to x for the i^{th} unit,

$\bar{r}_N = \frac{1}{N}\sum_{i=1}^{N} r_i$, the simple arithmetic mean of the ratio for all the units in the population,

$\bar{r}_n = \frac{1}{n}\sum_{i=1}^{n} r_i$, the simple arithmetic mean of the ratios for all the units in the sample,

$R_N = \frac{\bar{Y}_N}{\bar{X}_N} = \frac{Y}{X}$, the ratio of the population mean of y to the population mean of x, and

$R_n = \frac{\bar{y}_n}{\bar{x}_n} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$, the corresponding ratio for the sample.

With this, an estimator of the population mean $\bar{Y}_N$ is given by

$$\bar{y}_R = R_n \bar{X}_N = \frac{\bar{y}_n}{\bar{x}_n}\bar{X}_N .$$

This estimator is known as the ratio-type estimator and pre-supposes the knowledge of $\bar{X}_N$. Here, R_n provide an estimator of the population ratio R_N. For example, if y is the number of bullocks on a holding and x its area in acres, the ratio R_n is an estimator of the number of bullocks per acre of holding in the population. The product of R_n with $\bar{X}_N$, the average size of a holding in acres would provide an estimator of $\bar{Y}_N$, the average number of bullocks per holding in the population.

2.1 Expected Value of the Ratio Estimator

Note that R_n is a biased estimator of R_N and the bias in R_n is given by

Bias in $R_n = \frac{-Cov(R_n, \bar{x}_n)}{\bar{x}_N}$.

Expected value of the ratio estimator to the first approximation is given by

$$E_1(\bar{y}_R) = \bar{y}_N\left[1 + (\frac{N-n}{Nn})(C_x^2 - \rho C_y C_x)\right],$$

where, $C_x = \frac{S_x}{\bar{X}_N}$, $C_y = \frac{S_y}{\bar{Y}_N}$ and ρ = population correlation coefficient between x and y.

It may be noted here that the bias to the first approximation vanishes when the regression of y on x is a straight line passing through the origin.

2.2 Variance of the Ratio Estimator

The variance of the ratio estimator to a first approximation is given by

$$V_1(R_n) = R_N^2 \left(\frac{N-n}{Nn}\right)(C_y^2 + C_x^2 - 2\rho\, C_y C_x),$$

and the variance of the ratio estimator of population mean to a first approximation is given by

$$V_1(\bar{y}_R) = \frac{N-n}{Nn}\left(S_y^2 + R_N^2 S_x^2 - 2R_N S_{yx}\right).$$

2.3 Estimator of the Variance of the Ratio Estimator

A consistent estimator of the relative variance of a ratio estimator is given by

$$\hat{V}_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn}\left[\frac{s_y^2}{\bar{y}_n^2} + \frac{s_x^2}{\bar{x}_n^2} - \frac{2s_{yx}}{\bar{y}_n \bar{x}_n}\right]$$

and the estimator of variance of the ratio estimator of population mean to a first approximation is given by

$$\hat{V}_1(\bar{y}_R) = \frac{N-n}{Nn}\left[s_y^2 + R_n^2 s_x^2 - 2R_n s_{yx}\right]$$

where s_y^2, s_x^2 and s_{yx} are the corresponding sample values.

2.4 Efficiency of the Ratio Estimator

In large samples, the ratio estimator will be more efficient than the corresponding sample estimator based on the simple arithmetic mean if

$$\rho \frac{C_y}{C_x} > \tfrac{1}{2} \quad \text{or} \quad \rho > \frac{1}{2}\frac{C_x}{C_y}.$$

If $C_x = C_y$, as may be expected, for example, when y and x denote values of the same variate, in two consecutive periods, ρ will be larger than one-half in order that the ratio estimator may be more efficient than the one based on the simple arithmetic mean.

3. Ratio Estimator in Stratified Sampling

Let there be K stratum in the population. Let N_t denotes the number of units in the t^{th} stratum and n_t the size of the sample to be selected there from, so that

$$\sum_{t=1}^{K} N_t = N \quad \text{and} \quad \sum_{t=1}^{K} n_t = n.$$

Denote by R_{n_t} the estimate of the population ratio $R_{N_t} = \bar{Y}_{N_t} / \bar{X}_{N_t}$ and by $\bar{y}_{Rt}$ the ratio estimate of the population mean $\bar{Y}_{N_t}$ for the t^{th} stratum. Then clearly, the ratio estimator of the population mean $\bar{Y}_N = \sum_{i=1} \frac{N_t}{N} \bar{Y}_{N_t}$ has been discussed in the next section.

3.1 Separate Ratio Estimator $(\bar{y}_{Rs})$

$$\bar{y}_{Rs} = \sum_{t=1}^{K} \frac{N_t}{N} \bar{y}_{R_t} = \sum_{t=1}^{K} p_t \bar{y}_{R_t} \text{ , where } p_t = \frac{N_t}{N} \quad (t = 1,...,K).$$

This is a biased but consistent estimator of population mean $\bar{Y}_N$. The bias to the first approximation is given by

$$\text{Bias in } (\bar{y}_{Rs}) = E_1(\bar{y}_{Rs}) - \bar{Y}_N = \sum_{t=1}^{K} p_t \bar{Y}_{N_t} \left(\frac{N_t - n_t}{N_t n_t}\right)(C_{tx}^2 - \rho_t C_{tx} C_{ty}),$$

where $C_{tx} = \frac{S_{tx}}{\bar{X}_{N_t}}$ and $C_{ty} = \frac{S_{ty}}{\bar{Y}_{N_t}}$. The variance of $\bar{y}_{R_S}$ to a first approximation is given by

$$V_1(\bar{y}_{Rs}) = \sum_{t=1}^{K} p_t^2 \left(\frac{1}{n_t} - \frac{1}{N_t}\right)\left(S_{ty}^2 + R_{N_t}^2 S_{tx}^2 - 2R_{N_t} S_{txy}\right),$$

$$V_1(\bar{y}_{Rs}) = \frac{1}{N} \sum_{t=1}^{K} p_t \left(\frac{N_t - n_t}{n_t}\right)(S_{ty}^2 + R_{N_t}^2 S_{tx}^2 - 2R_{N_t} S_{txy}),$$

$$V_1(\bar{y}_{Rs}) = \frac{1}{N} \sum_{t=1}^{K} p_t \left(\frac{N_t - n_t}{n_t}\right)(S_{ty}^2 + R_{N_t}^2 S_{tx}^2 - 2R_{N_t} \rho_t S_{tx} S_{ty}).$$

The above formula is based on the assumption that n_t is large. A consistent estimator of $V_1(\bar{y}_{Rs})$ is given by

$$\hat{V}_1(\bar{y}_{Rs}) = \frac{1}{N} \sum_{t=1}^{K} p_t \left(\frac{N_t - n_t}{n_t}\right)(s_{ty}^2 + R_{n_t}^2 s_{tx}^2 - 2R_{n_t} s_{tyx}).$$

In practice, the assumption that n_t is large is not always true. To get over this difficulty, a combined ratio estimator has been suggested as below:

3.2 Combined Ratio Estimator $(\bar{y}_{Rc})$

$$\bar{y}_{Rc} = \frac{\sum_{t=1}^{K} p_t \bar{y}_{n_t}}{\sum_{t=1}^{K} p_t \bar{x}_{n_t}} \bar{X}_N.$$

This is again a biased estimator, however, it is a consistent estimator. The relative bias to the first approximation is given by

$$\text{Relative Bias in } (\bar{y}_{Rc}) = ((E_1(\bar{y}_{Rc}) - \bar{Y}_N)/\bar{Y}_N = \sum_{t=1}^{K} p_t^2 \left(\frac{N_t - n_t}{N_t n_t}\right)(C_{tx}^2 - \rho_t C_{tx} C_{ty}).$$

The variance of $\bar{y}_{Rc}$ to a first approximation is given by

$$V_1(\bar{y}_{Rc}) = \frac{1}{N} \sum_{t=1}^{K} p_t \frac{N_t - n_t}{n_t} (S_{ty}^2 + R_N^2 S_{tx}^2 - 2R_N \rho_t S_{ty} S_{tx}),$$

and an estimator of the variance is given by

$$\hat{V}_1(\bar{y}_{Rc}) = \frac{1}{N}\sum_{t=1}^{K} p_t \frac{N_t - n_t}{n_t}(s_{ty}^2 + R_n^2 s_{tx}^2 - 2R_n \, s_{tyx}),$$

where, $R_{nt} = \dfrac{\bar{y}_{n_t}}{\bar{x}_{n_t}}$ and $R_n = \dfrac{\sum_{t=1}^{K} p_t \bar{y}_{n_t}}{\sum_{t=1}^{K} p_t \bar{x}_{n_t}}$

4. Regression Method of Estimation

We have seen that the ratio estimate provides on efficient estimate of the population mean if the regression of *y*, the variable under study, on *x*, the auxiliary variable is linear and the regression line passes through the origin. It happens frequently that even though the regression of *y* on *x* is linear, the regression line does not pass through the origin. Under such conditions, it is more appropriate to use the regression method of estimation rather than ratio method of estimation.

4.1 Simple Regression Estimate

Since the regression coefficient β is generally not known, the usual practice is to use estimate

$$\hat{\beta} = \frac{s_{xy}}{s_x^2},$$

where $s_{xy} = \dfrac{1}{n-1}\sum^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)$ and $s_x^2 = \dfrac{1}{n-1}\sum^{n} (x_i - \bar{x}_n)^2$ giving the simple regression estimate,

$$\bar{y}_{lr} = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n).$$

Note: The general form of the estimator is $\hat{\bar{Y}} = \bar{y} + k(\bar{X}_N - \bar{x}_n)$.

(i) If $k = \hat{\beta}$, then $\hat{\bar{Y}} = \bar{y}_n + \hat{\beta}(\bar{X}_N - \bar{x}_n)$ i.e. $\hat{\bar{Y}}$ is regression estimator

(ii) If $k = \dfrac{\bar{y}}{\bar{x}}$ then $\hat{\bar{Y}} = \bar{y}_n + \dfrac{\bar{y}_n}{\bar{x}_n}\left(\bar{X}_N - \bar{x}_n\right) = \dfrac{\bar{y}_n}{\bar{x}_n}\bar{X}_N$ i.e. $\hat{\bar{Y}}$ is a ratio estimator.

4.2 Expected Value of the Simple Regression Estimator

$$E(\bar{y}_{lr}) = \bar{y}_N - Cov(\hat{\beta}, \bar{x}_n)$$

showing that the simple regression estimate is biased by an amount $-Cov(\hat{\beta}, \bar{x}_n)$.

4.3 Variance of the Simple Regression Estimate

To a first approximation,

$$V(\bar{y}_{lr}) \cong (\frac{1}{n} - \frac{1}{N}) S_y^2 (1 - \rho^2)$$

where ρ is the correlation coefficient between y and x in the population.

4.4 Estimator of the Variance

$$\hat{V}(\bar{y}_{lr}) = (\frac{1}{\mathrm{n}} - \frac{1}{N})\, \mathrm{s}_{\mathrm{y}}^2\, (1 - r^2)$$

where r = $\frac{s_{xy}}{s_x s_y}$ is the sample correlation coefficient.

5. Regression Estimators in Stratified Sampling

At first, we shall consider two difference estimates, namely

(i) Separate difference estimator

(ii) Combined difference estimate

5.1 Separate Regression Estimate

When β_i, s are not known in case of separate difference estimator, we estimate these from the sample and in that case the estimator is known as separate regression estimator.

$$\bar{y}_{lrs} = \sum_{i=1}^{K} p_i \left[\bar{y}_{n_i} + \hat{\beta}_i (\bar{x}_{N_i} - \bar{x}_{n_i})\right] \quad \text{where} \quad \hat{\beta}_i = \frac{s_{ixy}}{s_{ix}^2}$$

This estimator is biased and the variance of the estimator to the first approximation, is given by

$$\mathrm{V}(\bar{y}_{lrs}) \cong \sum_{i=1}^{K} p_i^2 (\frac{1}{n_i} - \frac{1}{N_i}) S_{iy}^2 (1 - \rho_i^2)$$

where ρ_i is the correlation coefficient between y and x for the i-th stratum and

$$\hat{V}(\bar{y}_{lrs}) = \sum_{i=1}^{K} p_i^2 (\frac{1}{n_i} - \frac{1}{N_i})(s_{iy}^2 + \hat{\beta}_i^2 s_{ix}^2 - 2\hat{\beta}_i s_{ixy})$$

5.2 Combined Regression Estimator

When the pooled regression coefficient β is not known then we replace it by $\hat{\beta}$ and get the combined regression estimator,

$$\bar{y}_{lrc} = \sum_{i=1}^{K} p_i \bar{y}_{n_i} + \hat{\beta}(\bar{X}_N - \sum_{i=1}^{K} p_i \bar{x}_{n_i}),$$

where $\hat{\beta} = \dfrac{\sum_{i=1}^{K} p_i^2 (\frac{1}{n_i} - \frac{1}{N_i}) s_{ixy}}{\sum_{i=1}^{K} p_i^2 (\frac{1}{n_i} - \frac{1}{N_i}) s_{ix}^2}$.

The variance of the estimator along with its estimator, to the first approximation are given by

$$V(\bar{y}_{lrc}) \cong \sum_{i=1}^{K} p_i^2 (\frac{1}{n_i} - \frac{1}{N_i})(S_{iy}^2 + \beta^2 S_{ix}^2 - 2\beta S_{ixy}),$$

and $\hat{V}(\bar{y}_{lrc}) = \sum_{i=1}^{K} p_i^2 (\frac{1}{n_i} - \frac{1}{N_i})(s_{iy}^2 + \hat{\beta}^2 s_{ix}^2 - 2\hat{\beta} s_{ixy})$.

6. Practical Examples

Let $y_i\ (i = 1,..., N)$ be the variate under study, and $x_i\ (i = 1,..., N)$ be the auxiliary variate. Let N be the population size out of which a sample of size n is drawn. Let X_N be the population total of the auxiliary variate.

STEP-I: Calculate: $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i^2$, $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i y_i$.

STEP-II: Calculate:

$$s_y^2 = \frac{1}{(n-1)}\left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]$$

$$s_x^2 = \frac{1}{(n-1)}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$$

$$s_{xy} = \frac{1}{(n-1)}\left[\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}\right]$$

$$b = \frac{s_{xy}}{s_x^2} \qquad r = \frac{s_{xy}}{s_x . s_y}$$

$$\bar{y}_n = \frac{1}{n}\sum y_i \qquad \bar{x}_n = \frac{1}{n}\sum x_i$$

$$R_n = \frac{\bar{y}_n}{\bar{x}_n} \qquad \bar{X} = \frac{X_N}{N}$$

STEP-III: Calculate:

(a) Ratio estimate .

$$\bar{y}_R = \frac{\bar{y}_n}{\bar{x}_n}\ \bar{X}_N .$$

Estimate of its variance

$$\hat{V}(\bar{y}_R) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[s_y^2 + R_n^2 s_x^2 - 2R_n s_{xy}\right].$$

(b) Regression estimate $(\bar{y}_{lr})$

$$\bar{y}_{lr} = \bar{y}_n + b\left(\bar{X}_N - \bar{x}_n\right).$$

Estimate of its variance

$$\hat{V}(\bar{y}_{lr})) = \left(\frac{1}{n}-\frac{1}{N}\right)\left[s_y^2 + b^2 s_x^2 - 2bs_{xy}\right] = \left(\frac{1}{n}-\frac{1}{N}\right)\left(1-r^2\right)s_y^2$$

(c) Simple Mean estimate .

$\bar{y}_{srs} = \bar{y}_n$.

Estimate of its variance .

$$\hat{V}(\bar{y}_{SRS}) = \left(\frac{1}{n}-\frac{1}{N}\right)s_y^2 .$$

STEP-IV: Calculate Estimate of Relative Efficiency

(a) Estimate of Relative Efficiency of Ratio estimate over Simple Mean estimate

$$= \frac{\hat{V}(\bar{y}_{SRS})}{\hat{V}(\bar{y}_R)} \text{ x } 100$$

(b) Estimate of Relative Efficiency of Regression estimate over Simple Mean estimate

$$= \frac{\hat{V}(\bar{y}_{SRS})}{\hat{V}(\bar{y}_{lr})} \text{ x}100$$

(c) Estimate of Relative Efficiency of Regression estimate over Ratio estimate

$$= \frac{\hat{V}(\bar{y}_R)}{\hat{V}(\bar{y}_{lr})} \text{ x } 100$$

Note: Estimate of Standard Error (SE) of the estimate can be worked out by taking square root of the corresponding value of the estimate of the variance.

Practical Exercise 1

A sample survey for the study of yield and cultivation practices of guava was conducted in Allahabad district. Out of a total of 146 guava growing villages in Phulpur-Saran tehsil, 13 villages were selected by method of simple random sampling. The Table below presents total number of guava trees and area under guava orchards for the selecte 13 villages. It is also given that the total area under guava orchards of 146 villages is 354.78 acres.

Using area under guava orchards as auxiliary variate, estimate the total number of guava trees in the tehsil along with its standard error, by using

(i) Ratio method of estimation, and

(ii) Regression method of estimation.

(iii) Discuss the efficiency of these estimates with the one which does not make use of the information on the auxiliary variate.

Sl. No. of Village	**Total number of guava trees** (y_i)	**Area under guava orchards (in acres)** (x_i)
1.	492	4.80
2.	1008	5.99
3.	714	4.27
4.	1265	8.43
5.	1889	14.39
6.	784	6.53
7.	294	1.88
8.	798	6.35
9.	780	6.58
10.	619	9.18
11.	403	2.00
12.	467	2.20
13.	197	1.00

SOLUTION:

$\sum_{i=1}^{n} y_i = 9710,$ $\sum_{i=1}^{n} x_i = 73.60,$ $\sum_{i=1}^{n} y_i^2 = 9685234,$ $\sum_{i=1}^{n} x_i^2 = 579.20,$

$\sum_{i=1}^{n} x_i y_i = 72879.72$, $s_y^2 = 202717.60,$ $s_x^2 = 13.54,$ $s_{xy} = 1492.18,$ $b = 110.19$

$r = 0.90,$ $\bar{y}_n = 746.92,$ $\bar{x}_n = 5.66,$ $R_n = 131.93,$ $\bar{X}_N = 2.43,$

$\bar{y}_R = 320.59$ $\hat{V}(\bar{y}_R) = 3132.35$ (Estimate of Standard Error = 55.97)

$\bar{y}_{lr} = 390.85$ $\hat{V}(\bar{y}_{lr}) = 2683.74$ (Estimate of Standard Error = 51.80)

$\bar{y}_n = 746.92$ $\hat{V}(\bar{y}_n) = 14205.18$ (Estimate of Standard Error = 119.19)

(a)	Estimate of Relative Efficiency of Ratio estimate over Simple Mean estimate	453.50
(b)	Estimate of Relative Efficiency of Regression estimate over Simple Mean estimate	529.31
(c)	Estimate of Relative Efficiency of Regression estimate over Ratio estimate	116.72

Practical Exercise 2

A sample survey was conducted for studying milk yield, feeding and management practices of cattle and buffaloes in the eastern districts of U.P. The whole of the eastern districts of U.P. were divided into four Zones (strata). The Table below present total number of milch cows in 17 randomly selected villages of Zone-I as enumerated in winter season and as per Livestock Census.

Sl. No. of Village	Number of Milch Cows	
	Winter Season (y_i)	Livestock Census (x_i)
1.	29	41
2.	44	44
3.	25	27
4.	38	53
5.	37	17
6.	27	40
7.	63	53
8.	53	46
9.	64	89
10.	30	37
11.	53	70
12.	25	15
13.	16	30
14.	15	18
15.	12	22
16.	12	13
17.	23	66

Estimate the number of milch cows per village with its standard error for the rural area of Zone-I in winter season by using (i) Ratio method of estimation, and (ii) Regression method of estimation. It is given that total number of milch cows in Zone-I as per Livestock Census was 10,87,004 and number of villages in Zone-I was 22,654. Also compare the efficiency of these estimates with Simple Mean estimate.

SOLUTION:

$$\sum_{i=1}^{n} y_i = 566, \quad \sum_{i=1}^{n} x_i = 681, \quad \sum_{i=1}^{n} y_i^2 = 23450, \quad \sum_{i=1}^{n} x_i^2 = 34617, \quad \sum_{i=1}^{n} x_i y_i = 26879$$

$s_y^2 = 287.85$, $s_x^2 = 458.56$, $s_{xy} = 262.86$, $b = 0.57$, $r = 0.72$

$\bar{y}_n = 33.29$, $\bar{x}_n = 40.06$, $R_n = 0.83$, $\bar{X}_N = 47.98$

$\hat{\bar{y}}_R = 39.88$ $\hat{V}(\hat{\bar{y}}_R) = 9.86$ $SE(\bar{y}_R) = 3.14$ (Estimate of Standard Error = 3.14)

$\hat{\bar{y}}_{lr} = 37.84$ $\hat{V}(\hat{\bar{y}}_{lr}) = 8.06$ (Estimate of Standard Error = 2.84)

$\hat{\bar{y}}_n = 33.29$ $\hat{V}(\hat{\bar{y}}_n) = 16.92$ (Estimate of Standard Error = 4.11)

(a)	Estimate of Relative Efficiency of Ratio estimate over Simple Mean estimate	171.67
(b)	Estimate of Relative Efficiency of Regression estimate over Simple Mean estimate	209.85
(c)	Estimate of Relative Efficiency of Regression estimate over Ratio estimate	122.24

Practical Exercise 3

A pilot sample survey for estimating the extent of cultivation and production of fresh fruits was conducted in three districts of Uttar Pradesh State during the agricultural year 1976-77. The following data were collected

Stratum Number	Total number of villages (N_m)	Total area under orchards (ha.) (X_m)	Number of villages in Sample (n_m)	Area under orchards (ha.) (x_m)			Total number of trees (y_m)		
1	985	11253	6	10.63	9.90	1.45	747	719	78
				3.38	5.17	10.35	201	311	448
2	2196	25115	8	14.66	2.61	4.35	580	103	316
				9.87	2.42	5.60	739	196	235
				4.70	36.75		212	1646	
3	1020	18870	11	11.60	5.29	7.94	488	227	374
				7.29	8.00	1.20	491	499	50
				11.50	1.70	2.01	455	47	879
				7.96	23.15		115	115	

Estimate the total number of trees in the three districts by different methods and compare their precision.

SOLUTION

The calculations have been shown in the Table given below:

Stratum	W_m	$\left(\frac{1}{n_m}-\frac{1}{N_m}\right)$	$\bar{x}_m$	$\bar{y}_m$	$\hat{R}_m$	$W_m \bar{x}_m$	$W_m \bar{y}_m$	$s^2_{x_m}$	$s^2_{y_m}$	s_{xy_m}
1	0.2345	0.16598	6.81	417.33	61.28	1.60	97.66	16.03	74778.80	1008.75
2	0.5227	0.12454	10.07	503.38	49.99	5.26	263.12	129.64	259107.98	5643.81
3	0.2428	0.08902	7.97	340.00	42.66	1.94	82.55	38.39	65885.60	1403.69

$$W_m = N_m / \sum N_m \quad , \quad \hat{R}_m = \bar{y}_m / \bar{x}_m$$

(A) Ratio Estimators

(i) Separate Ratio Estimate (y_{Rs})

$$y_{Rs} = \sum_{m=1}^{K} \hat{R}_m X_m = 2750077$$

Estimate of its variance $\hat{V}(y_{Rs})$

$$\hat{V}(y_{Rs}) = \sum N_m^2 \left(\frac{1}{n_m}-\frac{1}{N_m}\right)\left(s^2_{y_m} + \hat{R}_m^2 . s^2_{x_m} - 2.\hat{R}_m . s_{xy_m}\right) = 2441137855.48$$

(ii) Combine Ratio Estimate (y_{Rc})

$$y_{Rc} = \frac{\sum W_m \bar{y}_m}{\sum W_m \bar{x}_m} X = (2783995)$$

Estimate of its variance $\hat{V}(y_{Rc})$

$$\hat{V}(y_{Rc}) = \sum N_m^2 \left(\frac{1}{n_m} - \frac{1}{N_m} \right) \left(s_{y_m}^2 + \hat{R}.s_{x_m}^2 - 2.\hat{R}.s_{xy_m} \right)$$

where $\hat{\mathrm{R}} = \sum W_m \bar{y}_m \Big/ \sum W_m \bar{x}_m$

(iii) Efficiency of Separate Ratio Estimate (y_{Rs}) **over the Combined Ratio Estimate** (y_{Rc})

Estimate of Relative Precision Efficiency (R.P.)= $\frac{\hat{V}(y_{Rc})}{\hat{V}(y_{Rs})}$ x 100 (246.58%)

(B) Regression estimators

(i) Separate Regression Estimate (y_{ls})

$$y_{ls} = \sum_m^K N_m \left[\bar{y}_m + b_m \left(\bar{X}_m - \bar{x}_m \right) \right] = 2672911$$

Estimate of its variance $\hat{V}(y_{ls})$

$$\hat{V}(y_{ls}) = \sum_m^K N_m^2 \left(\frac{1}{n_m} - \frac{1}{N_m} \right) \left(s_{y_m}^2 - b_m^2 . s_{x_m}^2 \right) = 1870633332$$

(ii) Combine Regression Estimate (y_{lc})

$$y_{lc} = \mathrm{N} \left[\bar{y}_{st} + b_c \left(\bar{X} - \bar{x}_{st} \right) \right] \text{ where } b_c = \frac{\sum_m^K \sum_j^{n_m} \left(y_{mj} - \bar{y}_m \right) \left(x_{mj} - \bar{x}_m \right)}{\sum_m^K \sum_j^{n_m} \left(x_{mj} - \bar{x}_m \right)^2} = 2643949$$

$$\bar{y}_{st} = \sum_m^K N_m \bar{y}_m \quad \text{and} \quad \bar{x}_{st} = \sum_m^K N_m \bar{x}_m$$

Estimate of its variance $\hat{V}(y_{lc})$

$$\hat{V}(y_{lc}) = \sum_m^K \frac{W_m^2 (1 - f_m)}{n_m (n_m - 1)} \sum_j^{n_m} \left[\left(y_{mj} - \bar{y}_m \right) - b_c \left(x_{mj} - \bar{x}_m \right) \right]^2 = 2020917640 \quad \text{where } f_m = \frac{n_m}{N_m}$$

a) Estimate of Efficiency of Separate Regression Estimate (y_{ls}) **over the Separate Ratio Estimate** (y_{Rs}) **is given by**

Relative Precision (R.P.) = $\frac{\hat{V}(y_{Rs})}{\hat{V}(y_{ls})}$. 100 = 130.50%

b) Estimate of Efficiency of Combine Regression Estimate (y_{lc}) over the Combined Ratio Estimate (y_{Rc}) is given by

Relative Precision (R.P.) = $\frac{\hat{V}(y_{Rc})}{\hat{V}(y_{lc})} \cdot 100 = 297.86\%$

c) Estimate of Efficiency of Separate Regression Estimate (y_{ls}) over the Combined Regression Estimate (y_{lc}) is given by

Relative Precision (R.P.) = $\frac{\hat{V}(y_{lc})}{\hat{V}(y_{ls})} \cdot 100 = 108.03\%$

References

Cochran, William G. (1977). *Sampling Techniques*. Third Edition. John Wiley and Sons.

Des Raj (1968). *Sampling Theory*. TATA McGRAW-HILL Publishing Co. Ltd.

Des Raj and Promod Chandok (1998). *Sample Survey Theory*. Narosa Publishing House.

Murthy, M.N. (1977). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Singh, Daroga and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. Wiley Eastern Limited.

Singh, Daroga, Singh, Padam and Pranesh Kumar (1978). *Handbook of Sampling Methods*. I.A.S.R.I., New Delhi.

Singh Ravindra and Mangat N.S. (1996). *Elements of Survey Sampling*. Kluwer Academic Publishers.

Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Application*. Second Edition. Iowa State University Press, USA.

Sukhatme, P. V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Third Revised Edition, Iowa State University Press, USA.

SAMPLE SIZE DETERMINATION WITH HANDS ON EXERCISE

Raju Kumar

ICAR- Indian Agricultural Statistics Research Institute, New Delhi -110012

1. Introduction

In survey studies, once data are collected, the most important objective of a statistical analysis is to draw inferences about the population using sample information. "How big a sample is required?" is one of the most frequently asked questions by the investigators. If the sample size is not taken properly, conclusions drawn from the investigation may not reflect the real situation for the whole population. Obtaining a representative sample size remains critical to survey researchers because of its implication for cost, time and precision of the sample estimate. The survey statistician has to be careful while choosing the sample size, because too large a sample implies waste of resources (time and cost), and too small a sample reduces the utility of the result inferences. However, the difficulty of obtaining a good estimate of population variance coupled with insufficient skills in sampling theory impede the researchers' ability to obtain an optimum sample in survey research. Use of efficient sampling plan enable an optimum utilization of budgetary resources to provide the best estimators (Highest efficiency) of the population parameters. As is well known, efficiency of an estimator is measured by inverse of mean square error (or variance in case of unbiased estimators). Therefore, target is to minimize both cost as well as variance, simultaneously. But, unfortunately, it is not possible as there is a trade-off between cost of survey and variance of estimator. With an increase in the sample size, increases the cost of the survey while the variance decreases, that means improvement in efficiency. Hence, it is very important to determine appropriate sample size, to maintain a balance which is reasonable with respect to cost as well as efficiency. Survey sampling theory provides a framework within which the problem of determining sample size may be tackled reasonably, see for example, Cochran (1977), Singh and Chaudhury (1985), Field (2005) etc.

One of the key challenges that researchers face in survey research is the determination of appropriate sample size which is representative of the population under study. This is to ensure that findings generalized from the sample drawn back to the population are with limits of random error. It is impossible to make accurate inferences about the population when a test sample does not truly represent the population from which it is drawn due to sample bias. Statistical inference regarding population characteristics using the sample data generally adopts one of the two methods, namely, the estimation of population parameters or testing of a hypothesis. The process of obtaining an estimate of the unknown value of a parameter by a statistic is known as estimation. There are two types of estimations viz. point estimation and interval estimation. Again, when we draw inference about parameter from statistic, some kind of error arises. The error which arises due to only a sample being used to estimate the population parameters is termed as sampling error or sampling fluctuations. Whatever may be the degree of cautiousness in selecting sample, there will always be a difference between the parameter and its corresponding estimate. A sample with the smallest sampling error will always be considered a good representative of the population. Bigger samples have lesser sampling errors. When the sample survey becomes the census survey, the sampling error becomes zero. On the other hand, smaller samples may be easier to manage and have less non-sampling error. Handling of bigger samples is more expensive than smaller ones. The

non-sampling error increases with the increase in sample size. The computation of the appropriate sample size is generally considered to be one of the most important steps in statistical study. But, it is observed that in most of the studies this particular step has been overlooked. The sample size computation must be done appropriately because if the sample size is not appropriate for a particular study then the inference drawn from the sample will not be authentic and it might lead to some wrong conclusions.

2. Criteria of determining sample size

a) **Level of Precision:**

Sample size is to be determined according to some pre assigned 'degree of precision' or permissible margin of error between the estimated value and the population value. In other words, the level of precision may be also termed as sampling error. According to W.G. Cochran (1977), precision desired may be made by giving the amount of errors that are willing to tolerate in the sample estimates. It depends on the amount of risk a researcher is willing to accept while using the data to make decisions. It is often expressed in percentage. Thus, if a researcher finds that 60% of farmers in the sample have adopted a recommended practice with a precision rate of ±5%, then he or she can conclude that between 55% and 65% of farmers in the population have adopted the practice. High level of precision requires larger sample sizes and higher cost to achieve those samples.

b) **Confidence level desired**:

The confidence or risk level is based on ideas encompassed under the Central Limit Theorem. The key idea encompassed in the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value. Furthermore, the values obtained by these samples are distributed normally about the true value, with some samples having a higher value and some obtaining a lower score than the true population value. In a normal distribution, approximately 95% of the sample values are within two standard deviations of the true population value (e.g., mean). While calculating the sample size, the desired confidence level is specified by the z value. The z-value is a point along the abscissa of the standard normal distribution. For example, 1.96 and 2.58 for 95% and 99% confidence level. In other words, this means that, if a 95% confidence level is selected, 95 out of 100 samples will have the true population value within the range of precision specified earlier. There is always a chance that the sample you obtain does not represent the true population value.

c) Degree of variability:

The degree of variability in the attributes being measured refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to be, to obtain a given level of precision. For less variable (more homogeneous) population, smaller sample sizes works nicely. Note that a proportion of 50% indicates a greater level of variability than that of 20% or 80%. This is because 20% and 80% indicate that a large majority do not or do, respectively, have the attribute of interest. Because a proportion of 0.5 indicates the maximum variability in a population, it is often used in determining a more conservative sample size.

3. Principal Steps in Determination of Sample Size

- Choice of desired confidence level depending upon tolerance limit (depends on type of study and resources available).
- Some equation should be found that connects n with the desired precision of the sample.
- This equation solution depends upon one or more, parameters, (contain unknown properties of the population), must be estimated or achieved from prior knowledge.
- In case of sample size determination many a time multiple characteristics are available. The desired degree of precision is prescribed for each characteristic can produce conflicting values of n. Hence, some method must be developed to reconcile these values.

Finally, the calculated *n* must be appraised with the practical scenario available i.e. is n consistent with the resources (cost, labour, time and material required) available to take the sample. Many a time, situation arises to reduce the n drastically. Decision has to make either to go with a much smaller sample size, that reducing precision, or to wait until required resources. Regarding the choice of a level for tolerable margin of error and the confidence level, the user normally has only a vague idea and these measures are mainly subjective and depend largely on the judgment of the user regarding the importance, applicability and vulnerability of the results.

4. Determining the sample size

Regarding the sample sizes in case of simple random sampling, the cases for qualitative and quantitative data are presented below:

Taro Yamane (1967) formula:

$$n = \frac{N}{1 + N(e)^2}$$

where n = Desired sample size

N = Population of the study

e = precision of sampling error (0.05)

Qualitative data:

Cochran's formula for when the population is infinite

Cochran (1977) proposed a formula to calculate the sample size based on the sample required to estimate a proportion with an approximate $(1-z)\times 100\%$ confidence level. The units are classified into two classes, C and C' Some margin of error d in the estimated proportion p of units in class C has been agreed on, and there is a small risk α, probability that the actual error is larger than d; i.e., $\Pr(|p-P|>d)=\alpha.$ Simple random sampling is assumed, and p (proportion of the population having the characteristic) is taken as normally distributed. Hence, the standard deviation is $\sigma_p = \sqrt{\left(\frac{N-n}{N-1}\right)\frac{P(1-P)}{n}}.$

Hence the formula that connects n with the desired degree of precision is

$d = z\sqrt{\left(\frac{N-n}{N-1}\right)\frac{P(1-P)}{n}}$, where z is the abscissa of the normal curve that cuts off an area of at the tails or selected critical value of desired confidence level. Solving for *n*, get the formula $n_r = \frac{z^2P(1-P)}{d^2}\Big/\left[1+\frac{1}{N}\left(\frac{z^2P(1-P)}{d^2}-1\right)\right]$ where n_r = required sample size. For practical use, an advance estimate *p* of *P* is substituted in this formula. If *N* is large, a first approximation is $n_r = \frac{z^2p(1-p)}{d^2} = \frac{p(1-p)}{V}$, where, $V = \frac{p(1-p)}{n_r}$ is desired variance of the sample proportion. The proportion of the population (p) may be known from prior research or other sources; if it is unknown use p = 0.5 which assumes maximum heterogeneity (i.e. a 50/50 split). For example, suppose we want to calculate a sample size of a large population whose degree of variability is not known. Assuming the maximum variability, which is equal to 50% (p =0.5) and taking 95% confidence level with ±5% precision, the calculation for required sample size will be as follows:

p = 0.5 and hence *q* =1-0.5 = 0.5; *d*= 0.05; *z* =1.96.

$n_r = \frac{z^2pq}{d^2} = \frac{(1.96)^2\times 0.5\times 0.5}{(0.05)^2} = 384.16 = 384$, So the minimum sample size would be 384.

Again, taking 99% confidence level with ±5% precision, the calculation for required sample size will be as follows: *p* = 0.5 and hence *q* =1-0.5 = 0.5; *e* = 0.05; *z* =2.58.

So, $n_r = \frac{(2.58)^2\times 0.5\times 0.5}{(0.05)^2} = 665.64 = 666$

Table1: Each cell represents required sample size for different confidence level and precision.

Confidence level	**Desired margin of error**		
	d=0.03	d=0.05	d=0.1
95%	1067	384	96
99%	1849	666	166

Cochran's formula for when the population is finite

First calculate n_r using the previous section formula. Then find n_r/N value, if it is negligible (approx. less than 5 percent), n_r is a final calculated sample size. If not, the use the formula $n_0 = \frac{n_r}{1+\frac{(n_r-1)}{N}} \cong \frac{n_r}{1+n_r/N}$, n_o is the ultimate sample size. Now, suppose *N* =13191, n_r=666 at 99% confidence level with margin of error equal to (0.05). $n_r/N \approx 5\%$

not negligible, hence sample size is $n_0 = \frac{666}{1+\frac{(666-1)}{13191}} = 634.03 = 634.$ But, if the sample size is calculated at 95% confidence level with margin of error equal to (0.05), the sample size become 384 i.e. $n_r/N < 5\%,$ which does not need correction formula. So, in this case the representative sample size for our study is 384.

Quantitative data: Cochran's formula for calculating sample size when the population is infinite

A drawback with this formula is the need to know the population standard deviation. This may be known from prior research; if a good estimate is unavailable the formula will not be reliable.

Further, sample size determination using relative error in the estimated population total or mean as

$$\Pr\left(\left|\frac{\bar{y}-\bar{Y}}{\bar{Y}}\right| > r\right) = \Pr\left(\left|\frac{N\bar{y}-N\bar{Y}}{N\bar{Y}}\right| > r\right) = \Pr\left(\left|\bar{y}-\bar{Y}\right| > r\bar{Y}\right) = \alpha$$

Where, *r* be the relative error in the estimated population total or mean, $\bar{y}$ be the sample mean estimates, $\bar{Y}$ be population mean and α is a small probability. Standard error of $\bar{y}$ is $\sigma_{\bar{y}} = \sqrt{\frac{(N-n)}{N}\frac{S}{n}}.$ Hence, $r\bar{Y} = z\sigma_{\bar{y}} = z\sqrt{\frac{(N-n)}{N}\frac{S}{n}}$ and Solving for n gives $n_r = \left(\frac{zS}{r\bar{Y}}\right)^2 \Big/ \left[1+\frac{1}{N}\left(\frac{zS}{r\bar{Y}}\right)^2\right] = \frac{1}{(CV)^2}\left(\frac{S}{\bar{Y}}\right)^2.$

Note that the population characteristic on which n_r depends is its coefficient of variation (CV) and $S/\bar{Y}$.

If the interest is to control the absolute error instead of relative error formula specified as $n_r = \frac{z^2\sigma^2}{d^2}$ Where n_r= required sample size, σ = the population standard deviation and *d* = the degree of precision required. For example, If investigation is done on the average (mean) level of employee satisfaction in Government offices and want to know the required sample size. You decide on a 95% confidence level. Prior studies have reported a standard deviation (σ) of 1.5 so you decide to use the same figure in your estimate. Satisfaction will be measured on a 10-point scale and you set a margin of error of ±0.25 units. To determine the minimum sample size you then apply the formula: $n_r = \frac{(1.96)^2 \times (1.5)^2}{0.25^2} \approx 144.$ So your minimum sample size would be 144. When population size is finite i.e. sample represents a significant (e.g. over 5%) proportion of the population, a finite population correction factor can be applied. This will reduce the sample size required. The formula for this is: $n_0 = \frac{n_r}{1+\frac{(n_r-1)}{N}}$, where n_0= the adjusted sample size, n_r= the original required sample size and *N* = population size. For example, calculated sample size (n_r= 144) for the employee satisfaction survey in the previous example, you decide to apply a finite population correction factor because the total

number of employees is only 650 (N = 650). To determine the adjusted sample size you apply the following formula.

$n_a = \frac{n_r}{1+\frac{(n_r-1)}{N}} = \frac{144}{1+\frac{(144-1)}{650}} = \frac{144}{1.22} = 118.$ So your adjusted minimum sample size would be 118.

Sometimes the specification error to be tolerated is only given in terms of desired per cent S.E. of the estimator e.g. the estimate is desired with a maximum of say 5 % S.E. In such cases, n is obtained from the corresponding formulae. In simple random sampling, if the desired % S.E. is d, then n is given by $n_0 = \frac{1}{\left(\frac{1}{N}+\frac{d^2}{C^2}\right)}$.

Illustration of determination of sample size with practical example (Cochran, 1977)

The example is "An Anthropologist is preparing to study the inhabitants of some island. Among other things, he wishes to estimate the percentage of inhabitants belonging to blood group O. Co-operation has been secured so that it is feasible to take a simple random sample. How large should the sample be?" This is just a typical example. In fact, in almost all the sampling investigations, one has to face such problems. An answer to the question is not straight forward. First of all, one must be very clear about the objective of the study. Or at least, the user must know to what use their results are going to be put, so that he should be able to answer as to what is the margin of error he is going to tolerate in the results. In the above example, the Anthropologist should be able to answer as to how accurately does he wish to know the percentage of people with blood group O? In this case he is reported to be content with a 5% margin in the sense that if the sample shows 43% to have blood group O, the percentage for the whole island is sure to be between 38 and 48. Since a random sampling procedure has been used, every sample has got some chance of selection and the possibility of getting the estimates lying outside the above specified range cannot be ruled out. Aware of this fact, the Anthropologist is prepared to take a 1 in 20 chance of getting an unlucky sample with the estimate lying outside the above margin.

With the above information, ignoring fpc and assuming that the sample proportion p is assumed to be normally distributed, a rough estimate of n may be obtained. In technical terms, p is to lie in the range $(P \pm 5)$, except for a 1 in 20 chance. Since p is assumed to be normally distributed about the population proportion P, it will lie in the range $(P \pm 2\sigma_p)$ apart from a 1 in 20 chance (in 95% cases). Further, since the standard error of p is approximately given by $\sigma_p \cong \sqrt{\frac{PQ}{n}}$,where Q=1-P. Using confidence interval formula, approximately get $2\sigma_p = 5$. Hence $n = \frac{4PQ}{25}$. At this point a difficulty appears that is common to all problems in the estimation of sample size. A formula for n has been obtained, but n depends on some property of the population that is to be sampled. Here, it is the quantity P that we would like to measure. They therefore ask the anthropologist if he can give us some idea of the likely value of P. He replied that from previous data on other ethnic groups, and from his speculations about the racial history of this island, he will be surprised if P lies outside the range 30 to 60%. This information is sufficient to

provide a usable answer. For any value of P between 30 and 60, the product P.Q lies between 2100 and a maximum of 2500 at P = 50. The corresponding n lies between 336 and 400. To be on the safe side, 400 is taken as the initial estimate of n.

In the hypothetical blood groups example, we had d = 0.05, p = 0.5, α= 0.05, z = 2 .

Thus,

$$n_0 = \frac{4 \times 0.5 \times 0.5}{0.0025} = 400 .$$

Let us assume that there are only 3200 people on the island. The *fpc* is needed, and we find

$$n = \frac{n_0}{1 + (n_0 - 1)/N} = \frac{40}{1 + 399/3200} = 356 .$$

The formula for $n_0$holds also if *d, p* and *q* are all expressed as percentages instead of proportions. Since the product *pq* Increases as *p* moves towards 1/2, or 50%, a conservative estimate of n is obtained by choosing for *p* the value nearest to 1/2 in the range in which *p* is thought likely to lie. If *p* seems likely to lie between 5 and 9 %, for instance, we assume 9 % for the estimation of *n*.

5. Overall Sample Size

If a pilot survey is undertaken for testing questions and survey procedure before the main survey is launched, it may be possible to estimate roughly the parameters (For example, population mean, variance) required for the determination of sample size for the various items of interest. However, in practice pilot survey may not be always possible since, it requires advance budget prior to the main survey. Thus the determination of the sample size in most cases may have to be done in advance either by making reasonable guesses of the different parameters or prior information from different sources (For example, administrative or census sources).

In stratified multi-stage random sampling design is used, information is required not merely on the population mean and standard deviation (SD), but also its components of variance between primary stage sampling units (PSUs) and within PSUs. In such circumstances statistician have to proceed by finding sample sizes in stages, required for a simple random sample (SRS) and to make adjustments to the overall sample size the design effects of multi-stage sampling are taken into consideration

Generally, the level of precision desired of an estimate is expressed as a percentage of itself or, strictly speaking of the population parameter. Let *Y* be the characteristic under study, $\bar{Y}$ be the population mean and $\bar{y}$ be the sample mean. The required sampling precision is prescribed as a percentage of $\bar{y}$. For example, sampling precision of $\bar{y}$ should be $\alpha\%$ of $\bar{y}$ i.e., the population average should lie in between $\bar{y} - \frac{\alpha \times \bar{y}}{100}$ and $\bar{y} + \frac{\alpha \times \bar{y}}{100}$.Taking 95% confidence interval (*CI*), the margin of error $2SE(\bar{y})$ equalizes to $\frac{\alpha \times \bar{y}}{100}$ and that produce, $\frac{SE(\bar{y})}{\bar{y}} \times 100 = \%$ Relative $SE = \frac{\alpha}{2}$. Now, Putting the value of $SE(\bar{y})$ under simple random sampling (SRS) ultimately produce,

$n = 40000 \times \frac{(CV)^2}{\alpha^2}$, where, $CV = \text{Coefficient of variation} = \frac{\text{Stndard Deviation}\,(\sigma)}{\text{Population Mean}\,(\bar{Y})}$. For example, If, α=5, then $n = 40000 \times \frac{(CV)^2}{25}$ .Thus to sample sizes determination require the value of CV which can be estimated using prior information from different data sources. Coefficient of variation, generally being a stable quantity can also be approximated using some related information. For example, average household income can be taken to be equal to per capita income x average household size and 1/6 of the known range of household income can be taken as an approximation of the SD on the assumption of a normal distribution. However, a better assumption is that of log-normal distribution, i.e. instead of assuming y to be distributed normally, assume that $\log_e y$ is distributed normally.

Let us assume that $\log_e y$ is distributed normally with mean 'a' and standard deviation 'b', then Mean= $e^{a+\frac{b^2}{2}}$, variance= $e^{2a+b^2}e^{b^2-1}$, $(\text{CV})^2 = e^{b^2-1}$, median= e^a, mode= e^{a-b^2}. Hence, $\text{Mean}/\text{Median} = e^{b^2/2}$ and $\text{Mean}/\text{Mode} = e^{3b^2/2}$.

Thus if mean is approximated and the median or mode is roughly guessed, it is possible to calculate the value of 'b' and therefore, approximate C.V. It may also be noted that an error in the estimate of mode affects the estimate of 'b' less than the same relative error in the estimate of the median. Also it may be perhaps less difficult to make a good guess of the mode than median. Further, log normal distribution has the property that the proportion of population with values less than or equal to the mean is given by P(b/2), where P(t) is the area to the left of 't' of a standard normal probability density function. Thus if we guess the proportion of households whose income is less than or equal to the average, it is possible to obtain the value of 'b' by referring to the corresponding proportion in the standard normal distribution tables and thus arrive at an estimate of CV. Now discussion will be made regarding a simple method that does not depend upon the estimation of either the population mean or the standard deviation, but assumes log-normal distribution. Take the case of estimation of household income. Based on empirical data collected from a large number of countries, it is reported in a technical study on Household Income and Expenditure Surveys (Published in 1989 by Statistics Office of the United Nations under National Household Survey Capability Programme) that nearly two-thirds of the population in the case of distribution of income or similar economic variables lie below the average value. Thus with the property mentioned in above gives $CV = 1.0492 \cong 1$. It is not claimed that the observation mentioned above is universal. If the proportion between the average is different, CV will be different as given below:

Table 2: Table 5 (pp180-187) of the UN publication mentioned above gives the value of CV and the proportion of households below the average for some 54 countries.

Percentage of population below average	C.V.
55	0.2554
60	0.5409
65	0.9005
70	1.4157
75	2.2739
80	4.0000

It would be seen that as the percentage of population below the average changes, the value of CV changes very fast. Thus, rather than working with the assumed proportion of two-thirds, one may prefer to err on the safe side and take the proportion as 70% and use $CV = 1.4157 \cong 1.42$ or $(\mathrm{CV})^2 = 2$. Table 2 shows that only few cases CV exceeded 1.42. Thus the assumption of $(\mathrm{CV})^2 = 2$ is not unrealistic. If $(\mathrm{CV})^2 = 2$, then find n = 3200 and if $(\mathrm{CV})^2 = 1$, as it happens in most cases, get n = 1600.

In the above calculations, finite population correction (fpc) factor is ignored, i.e. the population size is very large as compared to sample size. As a working rule, when $n/N < 5\%$ fpc can be ignored.

However, if fpc cannot be ignored, then the sample size n' for a simple random sample will be $n' = \dfrac{n}{1 + n/N}$, where n is the sample size for the case when fpc can be ignored. In large scale sample surveys, generally one uses a two-stage random sampling design. A two-stage design is generally less efficient than SRS of the same ultimate size and to achieve the same level of precision as in a SRS, a larger number of ultimate stage sampling units has to be surveyed. This is called the design effect and the extent of the upward adjustment to the sample size depends on the degree of similarity of second stage units within a PSU, which is measured by intra-class correlation coefficient. As a good working rule, one can take the value of 2 for the design effect as indicated in the Handbook of Household Surveys (Revised Edition), Studies in Methods, Series F.No.31, 1984 of the United Nations. Using the value 2 for the design effect, find n = 6400, if $(\mathrm{CV})^2 = 2$ and n =3200, if $(\mathrm{CV})^2 = 1$. Here again the sample size required would be less if appropriate stratification is used at various stages.

6. Sample Size for Domains

Estimates are generally required not only at the national level but also for certain domains such as geographical regions, rural and urban areas. One has then to work out the sample size for each domain and add them to arrive at the national sample size. We shall assume that domain-wise estimates are required with the same precision of 5%. Further, for sake of simplicity, we will deal with the case of two domains of study. For sake of illustration, let us consider a Household Income and Expenditure Survey (HIES) and rural and urban areas may be taken as the two domains of study. Further, let us assume that 80% of households (hh) are rural and 20% of the hh are urban. Suppose further that the average household income for the urban area is twice the national average. With the 80:20 ratios between rural and urban hh, it means that the average household income in the rural areas, is 75% of the national average. Let $(\mathrm{CV})_{rural}$ and $(\mathrm{CV})_{urban}$ be the CV for rural and urban areas respectively. If $(\mathrm{CV})^2 = 2$, it can be shown that

$$0.45(\mathrm{CV})^2_{rural} + 0.80(\mathrm{CV})^2_{urban} = 1.75. \qquad (*)$$

If it assume that $(\mathrm{CV})_{rural} = (\mathrm{CV})_{urban}$, then find $(\mathrm{CV})^2_{rural} = (\mathrm{CV})^2_{urban} = 1.40$. Thus to obtain the same relative precision of 5% (i.e. relative SE of 2.5%), the sample size for each of two sectors will be $1.4 \times 1600 \times 2(\text{design weight}) = 4480 hh$. Hence, National sample size = 8960 *hh*. Thus the national sample size is increased by 40% from 6400 to 8960. If we do not have that many resources, we can divide the national sample size of 6400 equally to rural and urban areas. The effect of allocation of 3200 instead of 4480 hh

would be that, instead of 5% we shall have 5.9% precision (2.95% relative SE). It is likely that $(CV)_{rural} < (CV)_{urban}$.

Table 3: Sample size required with different $(CV)_{urban}$ and $(CV)_{rural}$ but satisfying the equation (*).

$(CV)^2_{rural}$	$(CV)^2_{urban}$	Rural	Urban	Total
1.0	1.625000	3200	5200	8400
1.1	1.56875	3520	5020	8540
1.2	1.515250	3840	4840	8680
1.3	1.45625	4150	4660	8820
1.4	1.40000	4480	4480	8960
1.5	1.34375	4800	4300	9100

References:

Cochran, W.G. (1977): *Sampling Techniques*. Third Edition. John Wiley and Sons.

Des Raj (1968): *Sampling Theory*. TATA McGRAW-HILL Publishing Co. Ltd.

Des Raj and Chandok, P. (1998): *Sample Survey Theory*. Narosa Publishing House.

Field, A.P. (2005). *Discovering statistics using SPSS*. London : Sage.

Murthy, M.N. (1977): *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Singh, D. and Chaudhary, F.S. (1986): *Theory and Analysis of Sample Survey Designs*. Wiley Eastern Limited.

Singh, D., Singh, P. and Kumar, P. (1978): *Handbook of Sampling Methods*. I.A.S.R.I., New Delhi.

Singh, R. and Mangat, N.S. (1996): *Elements of Survey Sampling*, Kluwer Academic Publishers.

Sukhatme, P.V. and Sukhatme, B.V. (1970): *Sampling Theory of Surveys with Application*. Second Edition. Iowa State University Press, USA

Sukhatme, P. V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984): *Sampling Theory of Surveys with Applications*. Third Revised Edition, Iowa State University Press, USA.

Tyagi, K.K., An overview of various sampling schemes and determination of sample sizes, Indian Agricultural statistics Research Institute.

Yamane, Taro. 1967. *Statistics, An Introductory Analysis*, 2nd Ed., New York: Harper and Row.

Chapter 9 Management Research: Applying the Principles © 2015 Susan Rose, Nigel Spinks & Ana Isabel Canhoto.

MULTI-PHASE SAMPLING AND SUCCESSIVE SAMPLING IN SAMPLE SURVEYS

Kaustav Aditya

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

The procedure called double sampling or two-phase sampling is typically employed in the following situation. There exists a procedure, relatively cheap to implement, that produces a vector of observations denoted by x. The vector x is correlated with the characteristics of interest, where the vector of interest is denoted by y. It is very expensive to make determinations on y. In the most popular form of two-phase sampling, a relatively large sample is selected and x determined on this sample. This sample is called the firstphasesample or phase *I* sample. Determinations for the vector y are madeon a subsample of the original sample. The subsample is called the secondphasesample orphase 2 sample. In the form originally suggested by Neyman(1938), the original sample was stratified on the basis of **x** and the stratifiedestimator for **y** constructed using the estimated stratum sizes estimated withthe phase 1 sample. We first describe this particular, and important, case oftwo-phase sampling. We simplify the discussion by considering scalar y. Double sampling can be used both with ratio or regression estimation technique and stratified sampling for better precession.

The general procedure for both double sampling with the ratio estimator and for double sampling with the regression estimator is identical. Contrary to double sampling for stratification where a categorical variable is observed in the first phase, it is usually metric variables that serve as ancillary variables when double sampling with the ratio or regression estimator is being used. In the first phase, a sample of size 'n' is taken to estimate the mean or total of the auxilliary variable X. The sample taken is usually large because measurement of X is cheap, fast and easy. In the second phase, a sample is selected on which both target and ancillary variable are observed; from these pairs of observations, a relationship between the two variables can be established, either a ratio or a regression. The second phase sample is usually small because the observation of Y is usually more expensive, difficult and time consuming. Then, the observations from the first phase are used to estimate the total and mean of the target variable for the entire area of interest.

In both approaches, dependent or independent phases are possible and the corresponding estimators need to be used. It is interesting to note, that double sampling is also interesting in context of Sampling with partial replacement (SPR) that is a very efficient technique to estimate changes.

Notations

N	Total number of samples in the entire area of interest;
n'	Number of samples in the first phase;
n	Number of samples in the second phase;

$\overline{y}_{mdr}$ Estimated mean of target variable *Y* from the ratio estimator for entire area;

$\overline{y}_{mdreg}$ Estimated mean of target variable *Y* from regression estimator for entire area;

$\overline{x}'$ Estimated mean of ancillary variable *X* in the first phase:

$\overline{x}$ Estimated mean of ancillary variable *X* in the second phase;

$\overline{y}$ Estimated mean of target variable *Y* in the second phase;

y_i i-th Observed value of target variable *Y*;

r Estimated ratio of the ratio estimator

b Estimated slope coefficient of regression estimator;

s^2_y Estimated variance of the target variable *Y*;

$s^{2\prime}_x$ Estimated variance of ancillary variable *X* in the first phase;

s_{xy} Estimated covariance of *Y* and *X* in the second phase;

$\hat{\rho}$ Estimated coefficient of correlation of *Y* and *X*.

For the *ratio estimator*, the mean of the target variable is estimated as,

$$\overline{y}_{mdr} = \frac{\overline{y}}{\overline{x}}\overline{x}' = r\overline{x}'$$

with an estimated variance of the estimated mean as,

$$\hat{V}(\overline{y}_{mdr}) = \frac{s_y^2 + r^2 s_x'^2 - 2rs_{xy}}{n} + \frac{2rs_{xy} - r^2 s_x'^2}{n'} - \frac{s_y^2}{N}$$

And for the *regression estimator*, the mean is estimated as,

$$\overline{y}_{mdreg} = \overline{y} + b(\overline{x}' - \overline{x})$$

with an estimated variance of the estimated mean as,

$$\hat{V}(\overline{y}_{mdreg}) = \frac{s_y^2}{n}\left(1 - \frac{n' - n}{n'}\hat{\rho}^2\right)$$

Examples:

1. Aerial photographs or satellite images are used to measure the ancillary variable, for example percentage crown cover. In the second phase, field plots are selected to measure the target variable such as volume or biomass per ha and the ancillary variables. Thus, a regression can be established which allows to predict the target variable once the ancillary variable is known. In many cases, this regression, however, is not very strong so that the overall precision that can be achieved is moderate. One of the main issues and source of errors in this example is the accuracy of co-registration between remote sensing imagery and [sample plot|field plots].
2. This example is on the estimation of leaf area of a tree, as, for example, needed to determine the leaf area index. Here, leaf area is difficult to measure; it is much easier

to observe leaf weight. Therefore, a regression is established in the second phase that allows predicting leaf area from leaf weight; a sample of leaves is taken in the second phase sample of which both leaf area and leaf weight are determined. In order to apply this regression, the mean (or total) leaf weight needs to be determined: for this purpose, a large sample is taken in the first phase. In this example, a major issue is the sampling frame for the first phase sample, that needs to be carefully defined (or a sampling technique is applied that does not require the a-priori definition of the sampling frame such asrandomized branch sampling).

2. Sampling on Successive Occasions or Successive Sampling

2.1 Introduction

Surveys often gets repeated on many occasions (over years or seasons) for estimating same characteristics at different points of time. The information collected on previous occasion can be used to study the change or the total value over occasion for the character and also in addition to study the average value for the most recent occasion. For example in milk yield survey one may be interested in estimating the

1. Average milk yield for the current season,

2. The change in milk yield for two different season and

3. Total milk production for the year.

The successive method of sampling consists of selecting sample units on different occasions such that some units are common with samples selected on previous occasions. If sampling on successive occasions is done according to a specific rule, with partial replacement of sampling units, it is known as successive sampling. The method of successive sampling was developed by Jessen (1942) and extended by Patterson (1950) and by Tikkiwal (1950, 53, 56, 64, 65, 67) and also Eckler (1955). Singh and Kathuria (1969) investigated the application of this sampling technique in the agricultural field. Hansen *et al*. (1955) and Rao and Graham (1964) have discussed rotation designs for successive sampling. Singh and Singh (1965), Singh (1968), Singh and Kathuria (1969) have extended successive sampling for many other sampling designs.

Generally, the main objective of successive surveys is to estimate the change with a view to study the effects of the forces acting upon the population. For this, it is better to retain the same sample from occasion to occasion. For populations where the basic objective is to study the overall average or the total, it is better to select a fresh sample for every occasion. If the objective is to estimate the average value for the most recent occasion, the retention of a part of the sample over occasions provides efficient estimates as compared to other alternatives. One important question arises in the context of devising efficient sampling strategies for repetitive surveys is whether the same sample is to be surveyed on all occasions, or fresh samples are to be chosen on each of the occasions; in what manner the composition of the sample is changed from occasion to occasion.

The answer depends on, apart from field difficulties, the specific problems of estimation at hand. For instance if the aim is to estimate only the difference between the item mean on the current ($\bar{y}$) and on the previous ($\bar{x}$) occasion, then the sample on both the occasion would

give rise to a better estimate than the independent samples since the variance of the estimate in the former case viz,

$$V(\bar{y}-\bar{x}) = V(\bar{y}) + V(\bar{x}) - 2COV(\bar{y}, \bar{x}) < V(\bar{y}) + V(\bar{x}),$$

as y and x are highly correlated so that $\mathbf{Cov}(\bar{y}, \bar{x}) > 0$.

On the contrary, for estimating the average of the means the latter would be better than the former in that

$$V(\bar{y}+\bar{x}) = \mathbf{V}(\bar{y}) + \mathbf{V}(\bar{x}) + \mathbf{2Cov}(\bar{y}, \bar{x}) > \mathbf{V}(\bar{y}) + \mathbf{V}(\bar{x}),$$

But, if the difference between the means and also their average are to be estimated simultaneously, clearly neither of this alternatives are desirable ,hence arises the idea of retaining a part (say S_c) of the previous sample (say S_1) and supplement it by a set (sayS_f) of fresh units on the current occasion, and the data retaining to x on , x and y on, and y on1*ScSfS* build up the optimum estimator of $\bar{Y}$ so that it ,together with the estimate of $\bar{X}$, would give rise to efficient result for difference between $\bar{Y}$ and $\bar{X}$,and also their average .The question then would be that big or small the set of common units or fresh units, should be for the surveys on the current occasion,how should these samples be chosen and what procedure be employed for working out estimates. The entire question is interrelated and depends ultimately on the regression of y on x. It is known that regression of y on x is linear with significant intercepts then we may choose fromby SRS without replacement and then employ regression estimator, or when the intercept is not significant the sample may be chosen by SRS and ratio estimator be employed.

2.2 Sampling on Two Successive Occasions

It is assumed that the survey population remains unaltered from occasion to occasion. For the purpose of generality, let the sample size for the first occasion be n_1 and for second occasion be, $n_2=n_{12}+n_{22}$, where n_{12} is the number of common units between the 1st and the 2nd occasion and n_{22} units to be drawn afresh on the second occasion. The data obtained on current (i.e. 2nd in this case) occasion would be denoted by y and that on the previous occasion (i.e. 1st in this case) by x. Now the sampling procedure consists of the following steps:

1. From the given survey population choose a sample $S_1$of size n_1 units by SRS without replacement for survey on the first occasion.
2. On the second occasion choose a set $S_c$ofn_{12} units from the sample taken at step(1a) either by SRS or PPS sampling depending on the situation at hand and supplement it to another set S_f of n_{22} units taken independently from the unsurveyed $N- n_1$ units of the population by SRS without replacement so that the total sample S_2on the second occasion comprises $n_2$units. Now S_1 acts as a preliminary sample.
3. The unbiased estimator of $\bar{Y}$ based on y and x values of S_c and x values of S_1 would be given as,

$$t_c = \frac{1}{n_{12}} \sum_{j=1}^{n_{12}} \frac{y_j}{p_j}, p_j = \frac{x_j}{\sum_{j=1}^{n_{12}} x_j}$$

with variance,

$$\mathrm{V}(t_c) = \frac{S_y^2}{n_1} + \frac{\sum_{j=1}^{N} P_j \left[\frac{y_j}{nP_j} - \overline{Y} \right]^2}{n_{12}} - \frac{S_y^2}{N}$$

Also in view of selection of $S_j$as noted in the step (2), the unbiased estimator of $\overline{Y}$ *is*,

$\overline{y}_f = \frac{1}{n_{22}} \sum_{j=1}^{n_{22}} y_j$ with variance as,

$$V\left(\overline{y}_f\right) = \left(\frac{1}{n_{22}} - \frac{1}{N} \right) S_y^2$$

Further, t_c and $\overline{y}_f$ are correlated so,

$$\mathrm{COV}(t_c, \overline{y}_f) = -\frac{1}{N} S_y^2$$

So in this sampling on two successive occasion, the best minimum variance combination of t_c and $\overline{y}_f$ will be,

$$\overline{\mathrm{y}}_{ss} = at_c + (1-a)\overline{y}_f, \text{where}, a = \frac{V_f}{V_c + V_f}$$

with variance,

$$V(\overline{\mathrm{y}}_{ss}) = \frac{V_c V_f}{V_c + V_f} + \mathrm{COV}(t_c, \overline{y}_f);$$

$$\text{where}, V_f = V(\overline{y}_f) - \mathrm{COV}(t_c, \overline{y}_f) \text{ and } V_c = V(t_c) - \mathrm{COV}(t_c, \overline{y}_f)$$

3. Application of Successive Sampling in Agriculture for Estimating the Incidence of Pest and Diseases on the Field Crops

This work is done by T.P. Abraham, R.K. Khosla and O.P. Kathuria from Institute of Agricultural Research Statistics, New Delhi-12, in the year 1969.

Surveys to estimate the incidence of pest and diseases on field crops have to be generally repeated due to large variation in the incidence of pest and diseases from year to year. It is therefore interesting to examine the partial replacement of units in such repeat surveys especially when taking some of the sampling units common from one year to another is operationally convenient. In particular, we examine how far partial matching of sampling units is helpful in obtaining a better estimate of,

1. The incidence in the second year of the survey.
2. The changes in occurrence from one year to other,
3. Overall mean incidence over the two year.

For this a survey was conducted in Cuttack district of Orissa on major pest of rice (i.e. stem borer and gallfly) and major disease was *Helminthosporium oryzae*, in each of the fields periodical observation was on various pest and diseases are taken at an interval about a month and up to and including harvest the first observation is taken after a month of planting. Now in each of the plots the no of the plots the number of the plants are recorded. The number of dead hearts due to stem borer is recorded and silver shoot by gall fly is also recorded. In case of helminthosporium disease, some plants are selected and the leaves with maximum infection are chosen and the intensity of the infection was noted in comparison with the standard chart given by Central Rice Research Institute, Cuttack. Also the manifestation by those pests also noted field wise average percentage of incidence of pest and diseases was worked out.

The estimates in change in incidence of stem borer and gallfly was taken mainly on kharif and rabi season and it is found after applying the method s of sampling on different occasion it is seen that the incidence of stem borer and gallfly in the months of March and October during rabi and kharif seasons respectively is more than any other months. it is also found that the incidence of those pest is much more in kharif than in rabi season. So we can see how this sampling scheme can be used in agricultural experiments.

References

Eckler, A.R. (1955). Rotation Sampling. *A.M.S.*, **26**: 664-685.

Jessen, R.J. 91942). Statistical Investigation of a sample survey for obtaining farm facts. *Iowa Agri. Expt. Station Res. Bull*. No. 304.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *JASA*, **33**,101-116.

Parzen, E. (1959). *Statistical Inference on Time Series boy Hilbert Space Methods, I.* Technical report No. 23: Department of Statistics, Stanford University.

Parzen, E. (1961). An approach to time series analysis. *Annals of Mathematical Statistics*, **32**: 951-989.

Rao, C.R. (1952). Some theorems on Minimum Variance Unbiased Estimation. *Sankhya, Sr. A*, **12**: 27-42.

Rao, J.N.K. and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions, *Jour Amer Assoc*., **59**, 492-509.

Tikkwal, B.D. (1951). *Theory of Successive Sampling*. Unpublished Thesis for Diploma, I.C.A.R., New Delhi.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. Charles Griffin & Company LTD., London.

NON-SAMPLING ERRORS IN SAMPLE SURVEYS

Kaustav Aditya

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

The reliability of the estimates from a survey depends on the errors that are affecting the survey. Groves (1989, Chapter 1) gives an excellent review of the potential sources of survey errors. Total survey error is sum of sampling error and non-sampling error. The former is as a result of selecting a sample instead of canvassing the whole population, while the latter is mainly due to adopting wrong procedures in the system of data collection and/or processing. In other words, sampling errors arise solely as a result of drawing a probability sample rather than conducting a complete enumeration. Non-sampling errors, on the other hand, are mainly associated to data collection and processing procedures. The quality of a sample estimator of a population parameter is therefore a function of total survey error, comprising both sampling and non-sampling errors. Both sampling and non-sampling errors need to be controlled and reduced to a level at which their presence does not defeat or obliterate the usefulness of the final sample results. This chapter will focus of non-sampling error in surveys.

2. Definition, Concept and Source of Non-Sampling Errors

Non-sampling error is an error in sample estimates which cannot be attributed to sampling fluctuations. Non-sampling errors may arise from many different sources such as defects in the frame, faulty demarcation of sample units, defects in the selection of sample units, mistakes in the collection of data due to personal variations or misunderstanding or bias or negligence or dishonesty on the part of the investigator or of the interviewer, mistakes at the stage of the processing of the data, etc. It may also arise from poorly designed survey questionnaires, improper sample allocation and selection procedures, and/or errors in estimation methodology. These errors are unpredictable and not easily controlled. Unlike in the control of sampling error this error may increase with increases in sample size. If not properly controlled non-sampling error can be more damaging than sampling error. It is noteworthy that increasing the sample size will not reduce this type of error.

These errors are caused by the mistakes in data processing. It includes:

- Over coverage: Inclusion of data from outside of the population.
- Under coverage: Sampling frame does not include elements in the population.
- Measurement error: The respondents misunderstand the question.
- Processing error: Mistakes in data coding.
- Non-response: errors because some selected units could not be contacted or refused to provide the information

Acquisition errors arise from the recording of incorrect responses, due to:

- incorrect measurements being taken because of faulty equipment,
- mistakes made during transcription from primary sources,

- inaccurate recording of data due to misinterpretation of terms, or
- inaccurate responses to questions concerning sensitive issues

Note that non-sampling errors can be generally defined as any source of bias or error in the estimation of a population characteristic in which the uncertainty about the resulting estimate is NOT due to the fact that we're sampling. We can think of them as errors for which increasing the sample size will not aid us in our estimation.

3. Types of Non-Sampling Errors

Brieumer and Lyberg (2003) identify five components of non sampling error, namely specification, frame, non-response, measurement and processing error. We may add that estimation error is another error, which should be considered. However, non-response and measurement errors are two main non-sampling errors that we generally talk. These types of error are briefly discussed below.

i. Specification Error

This occurs when the concept implied by the question is different from the underlying construct that should be measured. A simple question such as how many children does a person have can be subject to different interpretations in some cultures. In households with extended family member's biological children may not be distinguished from children of brothers or sisters living in the same household. In a disability survey, a general question asking people whether or not they have a disability can be subject to different interpretations depending on the severity of the impairment or the respondent's perception of disability. People with minor disabilities may perceive themselves to have no disability. Unless the right screening and filter questions are included in the questionnaire, the answers may not fully bring out the total number of people with disabilities.

ii. Coverage or Frame Error

In most area surveys primary sampling units comprise clusters of geographic units generally called enumeration areas (EAs). It is not uncommon that the demarcation of EAs is not properly carried out during census mapping. Thus households may be omitted or duplicated in the second stage frame. Frame imperfections can bias the estimates in the following ways: If units are not represented in the frame but should have been part of the frame, these results in zero probability of selection for those units omitted from the frame. On the other hand if some units are duplicated, this results in over coverage with such units having larger probabilities of selection. Errors associated with the frame may, therefore, result in both over coverage and under coverage. Non-coverage denotes failure to include some sample units of a defined survey population in the sampling frame. Because such units have zero probability of selection, they are effectively excluded from the survey results.

It is important to note that we are not referring here to deliberate and explicit exclusion of sections of a larger population from survey population. Survey objectives and practical difficulties determine such deliberate exclusions. For example attitudinal surveys on marriage may exclude persons under the minimum legal age for marriage. Residents of institutions are often excluded because of practical survey difficulties. Areas in a country infested with landmines may be excluded from a household survey to safeguard the safety of field workers. When computing non-coverage rates, members of the group deliberately and explicitly excluded should not be counted either in the survey population or under

non-coverage. In this regard defining the survey population should be part of the clearly stated essential survey conditions. Non-coverage is often associated with problems of incomplete frames. Examples are to omissions in preparing the frame but also missed units, implying omissions due to faulty execution of survey procedures. Thus non-coverage refers to the negative errors resulting from failure to include elements that would, under normal circumstances, belong in the sample. Positive errors of over coverage also occur due to inclusion in the sample of elements that do not belong there.

The term gross coverage error refers to the sum of the absolute values of non-coverage and over coverage error rates. The net non-coverage refers to the excess of non-coverage over coverage. It is, therefore, their algebraic sum. The net coverage measures the gross coverage only if over coverage is absent. Most household surveys in developing countries suffer mainly from under coverage errors. Most survey research practitioners agree that in most social surveys non-coverage is a much more common problem than over coverage. Corrections and weighting for non-coverage are much more difficult than for non-responses, because coverage rates cannot be obtained from the sample itself, but only from outside sources.

The non-coverage errors may be caused by the use of faulty frames of sampling units. If the frames are not updated or old frames are used as a device to save time or money, it may lead to serious bias. For example, in a household survey if an old list of housing units is not updated from the time of its original preparation say 10 years prior the current survey, newly added housing units in the selected enumeration area will not be part of the second stage frame of housing units. Similarly, some disbanded housing units will remain in the frame as blanks. In such a situation, there may be both omission of units belonging to the population and inclusion of units not belonging to the population.

At times there is also failure to locate or visit some units in the sample. This is a problem with area sampling units in which the enumerator must identify and list the households according to some definition. This problem arises also from use of incomplete lists. Some times weather or poor transportation facilitates make it impossible to reach certain units during the designated period of the survey. Survey results can, therefore, be distorted if the extent of non-coverage differs between geographical regions, sub groups, the population such as sex, age groups, ethnic and socio-economic classes. In general good frames should provide a list of sampling units from which a sample can be selected and sufficient information on the basis of which the sample units can be uniquely identified in the field.

Non-coverage errors differ from non-response. The latter, results from failure to obtain observations on some sample units, due to refusals, failure to locate addresses or find respondents at home and losses of questionnaires. The extent of non-response can be measured from the sample results by comparing the selected sample with that achieved. By contrast the extent of non-coverage can only be estimated by some kind of check external to the survey operation.

Sample selection and implementation errors

This strictly refers to losses and distortions within then sampling frame. Example, the wrong application of the selection procedures and selection probabilities. One glaring example is the inappropriate substitution of the selected units by others especially when systematic sampling is used in the field.

Reducing coverage error

The most effective way to reduce coverage error is to improve the frame by excluding erroneous units and duplicates and updating the frame through filed work to identify units missing from the frame. It is also important to undertake a good mapping exercise during the preparatory stages of a population and housing census. However, the frame prepared during the census should be updated periodically. It is also imperative to put in place procedures that will ensure the coverage of all selected sample units.

iii. Non-response errors

Non-response is error due to not all selected elements yield their information (i.e., failure to measure some of the sample units), which usually means that the population of interest is not the population from which the sample is drawn. It is a problem usually associated with surveys or interviews – any situation in which the human element is involved. People can and will refuse information for a wide variety of reasons – they could be busy, uninterested, suspicious of the surveyor's intentions, afraid they won't be anonymous, or simply uncooperative. The problem with non-response is that it changes our sampling frame – if some elements will not give us their information, then effectively we are sampling from the population of potential responders, not the population of interest. For example, let:

N = total population size, and μ = population mean

N_1 = total potential responders, and μ_1 = population mean of responders

N_2 = total potential non-responders, and

μ_2 = population mean of non-responders

Suppose we conduct a simple random sampling (SRS) from this population, with estimation via the usual sample mean (which is unbiased under SRS when all unit respond). Is the sample mean unbiased when there is non-response? No, because all of our data is drawn from the population of responders, and thus we are really estimating is μ_1, not μ. Let y denote the variable of interest. The bias in this case can be shown to be $(N_2 / N)(\bar{y}_1 - \bar{y}_2)$, where $\bar{y}_1$ and $\bar{y}_2$ are the averages of y for responders and non-responders respectively. We can think of this situation as a stratified sample where the population is broken into two strata, and we only have data from one stratum. Remember that the simple estimator used on data from a stratified sample is biased for μ - the same thing applies here.

Notice that if $\mu_1 = \mu_2$, in other words, if the populations of responders and non-responders are the same, then $\mu_1 = \mu$, and we are out of the woods – we can do everything in the same manner as we have all along. Evaluating whether or not the responders and non-responders are the same involves making an assumption, and that assumption is more or less reasonable depending on each specific situation. So what if we can't reasonably assume that the groups of responders and non-responders are similar, or if we prefer not to let our analysis ride on a subjective assessment? There are some alternatives.

In most cases non-response is not evenly spread across the sample units but is heavily concentrated among subgroups. As a result of differential non-response, the distribution of the achieved sample across the subgroups will deviate from that of the selected sample. This deviation is likely to give rise to non-response bias if the survey variables are also related to the subgroups.

The most obvious method of reducing non-response bias is to convert non-responders into responders. Recall the equation for non-response bias: $(N_2 / N)(\overline{y}_1 - \overline{y}_2)$. One way to reduce the absolute value of this quantity is to reduce N_2 / N, i.e., reduce the proportion of non-responders in the population. The ways to do this are numerous. Here is a medium-sized list, with short discussions of pros and cons. Some are specific, some are general, some are practical and some are psychological. They appear in no particular order.

Ways to Convert Non-responders into responders

i. If you are conducting a telephone or face-to-face interview, make sure you call/visit at times when the person to be interviewed is likely to be home.

ii. If you intend to send a mail survey, confirm that the people you wish to survey still live at the address you have on file. If a particular individual does not respond, you may want to send a representative to the address to find out if they are there, or perhaps to find out to where they have moved. If you want to sample whoever is currently living in the address you've selected, label the envelope, for example, "Mr. and Mrs. XYZ or current resident."

iii. For mailed surveys in particular, studies have shown that using attractive, high quality, official-looking envelopes and letterhead can improve response significantly. Include a carefully typed cover letter explaining your intentions, and guaranteeing their confidentiality. Get a big-wig from your company or organization to sign it (personally, if possible). Always send materials through first-class mail, and include a return envelope with first-class postage.

iv. Keep surveys and interviews as short as possible. As a general rule, the more questions you ask, the less likely you are to get accurate (or any) information.

v. Use the guilt angle whenever possible (but do it implicitly, don't beg). What I mean by this is simply to increase the amount and quality of personal contact with your population. Psychologically speaking, for most people it's easy to throw away a mailed survey, considerably harder to hang-up on an interviewer, and harder yet to walk away. Therefore, choose a face-to-face interview over a phone interview, and choose a phone interview over a mailed survey, whenever it is practical to do so.

vi. Publicizing or advertising your survey often helps with non-response. This lets people know they are not the only one being surveyed and helps with credibility. Use endorsements by celebrities, important individuals, or respected institutions if you are able.

vii. Offer an incentive. Money is by far the best, because it has the most universal appeal. Be careful when using other incentives, because you do not want to elicit responses from some specific subgroup of the population who happens to want or like what you're offering. Whether to offer the incentive up-front or upon return of the survey is basically a toss up in terms of effectiveness – but the former will be considerably more expensive.

In addition to the above, there is one more method that requires a bit more attention, called 'double sampling.' At the core, it is really just a two-stage sample. In the first stage, try to elicit responses through a cheap and easy method, such as a mailed survey. In the second stage, go after a random sample of the non-responders from stage 1 with the

big guns – telephone or face-to-face interviewing. This is a fairly well studied method, with suggested estimators.

Non-response rate

The non-response rate can be accurately measured if accounts are kept of all eligible elements that fall into the sample. Response rate for a survey is defined as the ratio of the number of questionnaires completed for sample units to the total number of sample units.

Reporting of non-response is good practice in surveys. Non-response can be due to respondents not being -at-home, refusing to participate in the survey, being incapacitated to answer questions and to lost schedules/ questionnaires. All categories of non-response refer to eligible respondents and should exclude ineligibles.

There are two types of non-responses: unit non-response and item non-response. Unit non-response implies that no information is obtained from certain sample units. This may be because respondents refuse to participate in the survey when contacted or they cannot be contacted. Item non-response refers to a situation where for some units the information collected is incomplete. Item non-response is therefore, evidenced by gaps in the data records for responding sample units. Reasons may be due to refusals, omissions by enumerators and incapacity.

The magnitude of unit (total) non-response, among other reasons, is indicative of the general receptivity, complexity, organisation and management of the survey. The extent of item non-response is indicative of the complexity, clarity and acceptability of particular items sought in a questionnaire and the quality of the interviewer work in handling those items.

Non-response errors can introduce bias in the survey results especially in situations in which the non-responding units are not representative of those that responded. Non-response increases both the sampling error, by decreasing the sample size, and non-sampling errors.

The basics assumption in the previous sections dealing with basic theory of sampling is that the probability of the sample unit being available for interview is one. In practice non-response occurs with varying degrees in different surveys. In general, follow ups can increase the number of responses.

In summary the types of non respondents include:

1. Not-at-homes: prospective respondents who may not be at home when enumerators visit their households.
2. Refusals: respondents who refuse to give information for whatever reasons.
3. Not identifiable respondents.

Causes of non-response

Respondents to provide information can cause non-response error, they are being not at home or by sample units not being accessible. This introduces errors in the survey results because sample units excluded may have different characteristics from the sample units for which information was collected. Refusal by a prospective respondent to take part in a survey may be influenced by many factors, among them, lack of motivation, shortage of time, sensitivities of the study to certain questions, etc. Groves and Couper (1995) suggest a number of causes of refusals, which include social context of the study,

characteristics of the respondent, survey design (including respondent burden), interviewer characteristics and the interaction between interviewer and respondent.

Errors arise from the exclusion of some of the units in the sample. This may not be a serious problem if the characteristics of the non-responding units are similar to those of the responding units, serve for large sampling errors. But such similarity is not common in practice.

With specific reference to item non-response, questions in the survey may be perceived by the respondent as being embarrassing, sensitive or/and irrelevant to the stated objective. The enumerator may skip a question or ignore recording an answer. In addition, a response may be rejected during editing. Non-response cannot be completely eliminated in practice, however it can be minimized by persuasion through repeated visits or other methods.

Reducing non-response

A number of procedures can be used in survey design in an attempt to reduce the number of refusals. For example in face-to-face interviews, interviewers are supposed to be carefully trained in strategies to avoid refusals, and they are to return to conduct an interview at the convenience of the respondent. The objectives and value of the surveys should generally and carefully be explained to respondents so that they can appreciate and cooperate. Assurance of confidentiality can help to alleviate fear respondents may have about the use of their responses for purposes other than those stipulated for the survey. The following are some of the steps that can be undertaken to reduce non-response on household surveys:

Good frames

In many developing countries there are problems of locating sample units. This results in some form of non-response error. In such cases it would be helpful to have good frames of both area units and housing listings, to facilitate easy identification of all respondents. In addition, the workloads of enumeration staff should be manageable within the allotted time frame for the survey. This enables them to reach all sample units within the assigned cluster or enumeration area. During listing of households, for example, enough auxiliary information should be collected to facilitate distinction and easy location of the sample unit. Whenever, possible enumerators should know the area they work in very well and should preferably be stationed in the assigned work areas.

Interview training, selection and supervision

In personal interview surveys, the enumerator can play an important role in maximising response from respondents. The way interviewers introduce themselves, what they say about the survey, the identity they carry, and the courtesy they show to respondents matter. In most household surveys the enumerator is the only link between the survey organisation and respondent. It is for this reason that enumerators and their supervisors should be carefully selected, well trained and motivated. Close supervision of enumerator's work and feedback on achieved response rate is of paramount importance.

Follow up of non-responding units

There should be follow up of non-respondents or make all effort to collect information from a sub-sample of the units who did not respond in the first place. This can be treated as a different stratum, from the responding stratum, in which better enumerators or supervisors may be assigned to interview respondents. The extent of refusals will depend on the subject matter of the survey (sensitive subjects are prone to high refusals), length

of and complexity of the questionnaire and skills of the survey team. The not-at-home respondents should be followed up. Depending on the resources and duration of the survey in face-to-face interviews at least four callbacks are recommended. These should be made during different days and different times of the day (villages give example of farming period).

iv. Measurement Errors

These errors arise from the fact that what is observed or measured departs from the actual values of sample units. These errors centre on the sustentative content of the survey such as definition of survey objectives, their transformation into usable questions, and the obtaining, recording, coding and processing of responses. These errors concern the accuracy of measurement at the level of individual units.

For example at the initial stage wrong or misleading definitions and concepts on frame construction and questionnaire design lead to incomplete coverage and varied interpretations by different enumerators leading to inaccuracies in the collected data.

Inadequate instructions to field staff are another source of error. For some surveys instructions are vague and unclear leaving enumerators to use their own judgement in carrying out fieldwork. At times sample units in the population lack precise definition, thereby resulting in defective and unsatisfactory frames. The enumerators themselves can be a source of error. At times the information on items for all units may be wrong, this is mainly due to inadequate training of field workers. Depending on the type and nature of enquiry or information collected, these errors may be assigned to respondents or enumerators or both. At times there may be interaction between the two, which may contribute to inflating such errors. Likewise, the measurement device or technique may be defective and may cause observational errors. Reasons for such errors are:

- Inadequate supervision of enumerators.
- Inadequately trained and experienced field staff.
- Problems involved in data collection and other type of errors on the part of respondents.

Non-sampling errors occur because procedures of observation or data collection are not perfect and their contribution to the total error of the survey may be substantially large thereby affecting the survey results adversely. At times respondents may introduce errors because of the following reasons:

- Failure to understand the question.
- Careless and incorrect answers from respondent due to, for example, lack of adequate understanding of the objective(s) of the survey. The respondent may not give sufficient time to think over the questions.
- Respondents answering questions even when they do not know the correct answer.
- Deliberate inclination to give wrong answers, for example, in surveys dealing with sensitive issues, such as income and stigmatised diseases.
- Memory lapses if there is along reference period, a case in point is the collection information on non-durable commodities in expenditure surveys.

The cumulative effect of various errors from different sources may be considerable since errors from different sources may not cancel. The net effect of such errors can be a large bias.

v. Processing Errors

Processing errors comprise:

- Editing errors.
- Coding errors.
- Data entry errors.
- Programming errors etc.

The above errors arise during the data processing stage. For example in coding open ended answers related to economic characteristics, coders may deviate from the laid out procedures in coding manuals, and therefore assign wrong codes to occupations. In addition, the weighting procedures may be wrongly applied during the processing stage, etc.

vi. Errors of estimation

These arise in the process of extrapolation of results from the observed sample units to the entire target population. These include errors of coverage, sample selection and implementation, non-response, as well as sampling variability and estimation bias. This group of errors centres on the process of sample design, implementation and estimation. Biases of the estimating procedure may either be deliberate, due to the uses of a biased estimation procedure or it may be due to inadvertent use of wrong formula.

Bias and variable error

The main types of survey errors are generally divided into two main kinds:

- Survey biases due to definitions, measurement and responses.
- Sampling variable errors.

However, we should also take note that there are sampling biases and variable non-sampling errors. Bias refers to systematic errors that affect any sample taken under a specified survey design with the same constant error. Ordinarily, sampling errors account for most of the variable errors of a survey, and biases arise mainly from non-sampling sources. In this connection, bias arises from the flaws in the basic survey design and procedures. While variable error occurs because of the failure to consistently apply survey designs and procedures. A widely accepted model combines the variable error and the bias into total error, which is a sum of variable error, and bias.

The mean square error (MSE) for an estimate is equal to the variance plus the squared bias (MSE = Variance + Squared bias). If for arguments sake the bias were zero, the MSE would therefore be the variance of the estimate. In most cases bias is not zero. As earlier indicated measuring bias in surveys may not be easy, partly because its computation requires the knowledge of the true population value which in most cases is not a practical proposition.

In practice non-sampling errors can decompose into variable component and systematic errors. According to Biemer and Lyberg (2003) there are two types of non-sampling error, namely systematic and variable error, the latter are generally non compensating errors and therefore tend to agree (in most cases, mostly in the same direction e.g.

positive), while the latter are compensating errors that tend to disagree (cancelling each other).

Variable component

The variable component of an error arises from chance (random) factors affecting different samples and repetition of the survey. In the case of the measurement process we can imagine that the whole range of procedures from enumerator selection, data collection to data processing can be repeated using the same specified procedures, under the same given conditions, and independently without one repetition affecting another. The results of repetitions are affected by random factors, as well as systematic factors, which arise from conditions under which repetitions are undertaken and affect the results of the repetition the same way. When the variable errors (VE) are caused only by sampling errors, VE squared equals sampling variance. The deviation of the average survey value from the true population value is the bias. Both variable errors and biases can arise either from sampling or non-sampling operations. The variable error will measure the divergence of the estimator from its expected value and it comprises both sampling variance and non-sampling variance. The difference of the expected value of the estimator from its true value is total bias and comprises both sampling bias and non-sampling bias.

Systematic error

This occurs when there is a tendency either to consistently underreport or over report in a survey. For example in some societies where there are no birth certificates, there is a tendency among men to exaggerate. This will result in systematic bias of the average age in the male population, producing a higher average than what the true average age should be. Variable errors can be assessed on the basis of appropriately designed comparisons between repetitions (replications) of survey operation under the same conditions. Reduction in variable errors depends on doing more of something e.g. larger sample size, more interviewers etc. on the other hand bias can be reduced only by improving survey procedures by doing something more, e.g. additional quality control measures at various stages of the survey operation.

Sampling bias

Sampling biases may arise from inadequate or faulty conduct of the specified probability sample or from faulty methods of estimation of the universe values. The former includes defects in frames, wrong selection procedures, and partial or incomplete enumeration of selected units. In general, biases are difficult to measure, that is why we emphasize their rigorous control. Their assessment can only be done by comparing the survey results with external reliable data sources. On the other hand variable error can be assessed through comparisons between sub-divisions of the sample or repetition of the survey under the same conditions. Bias can be reduced by improving survey procedures. As earlier stated biases can be negative or positive.

In summary, bias arises from factors, which are a part of essential conditions and affect all repetitions in more or less the same way. Biases arise from shortcomings in the basic survey design and procedures. In general, biases are harder to measure and can only be assessed on the basis of comparison with more reliable sources outside the normal survey or with information obtained by using improved procedures. Some sources of error appear mainly in the form of bias, among them coverage, non-response, and sample selection. On the other hand errors in coding and data entry may appear largely as variable error. Although both systematic and variable error reduces accuracy, bias is more

damaging in estimates such as population means, proportions and totals. These linear estimates are sums of observations in the sample. It should be noted that variable non-sampling errors like sampling errors could be reduced by increasing the sample size. For nonlinear estimates such as correlation coefficients, standard errors and regression estimates both variable and systematic error can lead to serious bias (Biemer and Lyberg, 2003).

Precision and accuracy

These terms are widely used to separate the effects of bias. Precision generally refers to small variable errors; at times it denotes only the inverse of the sampling variance, i.e. it excludes bias. Accuracy refers to small total errors and includes the effect of bias. A precise design must have small variable errors while an accurate design must be precise and have zero or small bias. A survey design is still precise if it has a large bias but with small variable errors. Such a design is however, not accurate. Note that reliability refers mainly to precision of measurements whereas validity to lack of bias in the measurements.

4. Assessing Non-Sampling Errors

Consistency check

In designing the survey instruments (questionnaires), special care has to be taken to include certain items of information that will serve as a check on the quality of the data to be collected. If the additional items of information are easy to obtain, they may be canvassed for all units covered in the survey, otherwise, they may be canvassed only for a sub-sample of units. For example, in a post census enumeration survey (PES), where the de jure method is followed it may be helpful to also collect information on de facto basis, so that it will be possible to work out the number of persons temporarily present and the number of persons temporarily absent. A comparison of these two figures will give an idea of the quality of data. Similarly, inclusion of items leading to certain relatively stable ratios such as sex ratios may be useful in assessing the quality of survey data.

Sample check/verification

One way of assessing and controlling non-sampling errors in surveys is to independently duplicate the work at the different stages of operation with a view to facilitating the detection and rectification of errors. For practical reasons the duplicate checking can only be carried out on a sample of the work by using a smaller group of well- trained and experienced staff. If the sample is properly designed and if the checking operation is efficiently carried out, it would be possible, not only to detect the presence of non-sampling errors, but also to get an idea of their magnitude. If it were possible to completely check the survey work, the quality of the final results could be considerably improved. With the sample check, rectification work can only be carried out on the sample checked. This difficulty can be overcome by dividing the output at different stages of the survey, e.g. filled in schedules, coded schedules, computation sheets, etc., into lots and checking samples from each lot. In this case, when the error rate in a particular lot is more than the specified level, the whole lot may check and corrected for the errors, thereby improving the quality of the final results.

Post-survey checks

An important sample check, which may be used to assess non-sampling errors consists of selecting a sub-sample, or a sample in the case of a census, and re-enumerating it by using

better trained and more experienced staff than those employed for the main investigation. For this approach to be effective, it is necessary to ensure that;

- The re-enumeration is taken up immediately after the main survey to avoid any possible recall error.
- Steps are taken to minimize the conditioning effect that the main survey may have on the work of the post survey check.

Usually the check-survey is designed to facilitate the assessment of both coverage and content errors. For this purpose, it is first desirable to re-enumerate all the units in the sample at the high stages, e.g. EAs and villages, with the view of detecting coverage errors and then to resurvey only a sample of ultimate units ensuring proper representation for different parts of the population which have special significance from the point of view of non-sampling errors. A special advantage of the check-survey is that it facilitates a unitary check, which consists first, of matching the data obtained in the two enumerations for the units covered by the check-sample and then analyzing the observed individual differences. When discrepancies are found, efforts are made to identify the cause of their presence and gain insight into the nature and types of non-sampling errors. If the unitary check is a problem due to time and financial constraints, an alternative but less effective procedure called aggregate check, may be used. This method consists in comparing estimates of parameters given by check-survey data with those from the main survey. The aggregate check gives only an idea of net error, which is the resultant of positive and negative errors. The unitary check provides information on both net and gross error.

In post survey check, the same concepts and definitions, as those used in the original survey should be followed.

Quality control techniques

There is ample scope for applying statistical quality control techniques to survey work because of the large scale and repetitive nature of the operations involved in such work. Control charts and acceptance-sampling techniques could be used in assessing the quality of data and improving the reliability of the final results in large-scale surveys. Just for illustration, work of each data entry clerk could be checked 100 percent for an initial period of time, but if the error rate falls below a specified level, only a sample of the work may be verified.

Study of recall errors

Response errors, as earlier mentioned in this chapter, arise due to various factors such as:

- The attitude of the respondent towards the survey.
- Method of interview.
- Skill of the enumerator.
- Recall error.

Of these, recall error needs particular attention as it presents special problems often beyond the control of the respondent. It depends on the length of reporting period and on the interval between the reporting period and the date of the survey. The latter may be taken care of by choosing for the reporting period a suitable interval preceding the date of survey or as near a period as possible. One way of studying recall error is to collect and analyse data relating to more than one reporting period in a sample or sub-sample of units

covered in a survey. The main problem with this approach is the effect of certain amount of conditioning effect possibly due to the data reported for one reporting period influencing those reported for the other period. To avoid the conditioning effect, data for the different periods under consideration may be collected from different sample units. Note tha t large samples are necessary for this comparison. Another approach is to collect some additional information, which will permit estimates for different reporting periods to be obtained. For example in a demographic survey one may collect not only age of respondent, but also date month and year of birth. The discrepancy will reveal any recall error that may be present in the reported age.

Interpenetrating sub-sampling

This method involves drawing from the overall sample two or more sub-samples, which should be selected in an identical manner and each capable of providing a valid, estimate of the population parameter. This technique helps in providing an appraisal of the quality of the information, as the interpenetrating sub-samples can be used to secure information on non-sampling errors such as differences arising from differential enumerator bias, different methods of eliciting information, etc. After the sub-samples have been surveyed by different groups of enumerators and processed by different teams of workers at the tabulation stage, a comparison of the estimates based on sub-samples provides a broad check on the quality of the survey results. For example, in comparing the estimates based on four sub-samples surveyed and processed by different groups of survey personnel, if three estimates are close to each other and the other estimate differs widely from them despite the sample size being large enough, then normally one would suspect the quality of work in the discrepant sub-sample.

5. Conclusion

Non-sampling errors should be given due attention in household sample surveys because they can cause huge biases in the survey results if not controlled. In most surveys very little attention is given to the control of such errors at the expense of producing results that may be unreliable. The best way to control non-sampling errors is to follow the right procedures of all survey activities from planning, sample selection up to the analysis of results.

References

Banda, J.P. (2003). Nonsampling errors in surveys. UNITED NATIONS SECRETARIAT ESA/STAT/AC.93/7 Statistics Division.

Biemer, P.P. and Lyberg, L. E. (2003). *Introduction to Survey Quality*. Wiley Series in Survey Methodology, Wiley, Hoboken.

Biemer, P. P. (editors) (1991). *Measurement Errors in Surveys.* Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Groves, R. (1989). *Survey Errors and Survey Costs*. Wiley New York.

Groves, R.M. and Couper, M. P. (1995). Theoretical motivation for Post-Survey Nonresponse Adjustment in Household Surveys. *Journal of Official Statistics*. **11(1)**, 93- 106.

Kalton, G. and Heeringa, S. (2003). *Leslie Kish: Selected papers*, Wiley Series in Survey Methodology, Hoboken.

Kish, L. (1965). *Survey Sampling*, Wiley, New York.

Murthy, M.N. (1967). *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

Raj, D. (1972*). The Design of Sample Surveys*, McGraw-Hill Book Company, New York.

CROP CUTTING EXPERIMENTS TECHNIQUE FOR CROP YIELD ESTIMATION

Tauqueer Ahmad

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1.0 Introduction

The estimation of crop production is based on acreage under the particular crop and its average yield per hectare. Thus, the data on crop area and crop yield is a basic and timely requirement for estimation of prop production. In India, crop area is compiled on the basis of complete enumeration while the crop yield is estimated on the basis of sample survey approach. The estimates of yield rates are obtained using Crop Cutting Experiment (CCE) approach. The crops may be sown in rows (one direction, two directions) and without rows (broadcasting). Keeping in view the proper representation of each plant sown either in rows or otherwise, the three different methods are recommended for demarcation of CCE plot. The measurement of length and breadth of the field, determination of random number pair for marking of CCE plot is to be done atleast one month before the harvest of crop and marking of CCE plot may be done on the date of harvesting or before as per situation.

2.0 Demarcation of CCE plot when crop is sown without rows (broadcasting)

The crops like wheat, barley, mustard, gram, lentil, peas, greengram, blackgram, redgram maize, jowar, bajra, etc sown through either broadcasting or in compact rows without maintaining plant to plant distance within the row. Method of marking CCE plot is as under:

2.1 Determination of random number pair for random step for length and breadth

Two random numbers, one for length and the other for breadth have to be selected with the help of random number table. A column number of the random number is assigned to the primary worker for selecting these two random numbers. Steps in the length as well as breadth in a CCE plot have to be deducted separately from length and breadth of the selected field to ensure the whole CCE plot gets accommodate in the selected field. Suppose the shape of CCE plot is square of side 5 meter. (7 steps = 5 meter approximately).

Example:

Length of the selected field (in steps)	120
Steps in the length of CCE plot	007
Length (in steps) of the selected field **minus** number of steps in the length of CCE plot	113
Breadth of the selected field (in steps)	70
Steps in breadth of CCE plot	07
Breadth (in steps) of the selected field **minus** steps in the breadth of CCE plot	63

Let column number 1 of random number table is assigned to the primary worker. A random number which is less than or equal to 113 is to be selected for length. Since 113 comprises of three digits, therefore, by referring column number one of three digits random number table, random number 058 appeared first which is less than 113. Therefore, random number 058 is selected as random step for length. The second random number is to be selected for breadth. It should be less than or equal to 63. Since, 63 comprises of two digits, therefore, by referring column number one of two digit random number table, random number 51 appeared first which is less than 63. Hence, random number 51 is selected random step for breadth. Random number pair (58, 51) is selected for locating the south-west corner of the CCE plot in the selected field. If the assigned column of random number table is exhausted during the process of selection of random numbers, the next column on the right hand side will have to be referred. If the whole or part of the CCE plot goes beyond the boundary of the selected field owing to irregular shape of the selected field, the random number pair should be rejected and a new random number pair should be selected till whole CCE plot accommodates within the field.

2.2 Marking of CCE plot

2.2.1 Marking of south-west corner of the CCE plot

The selected random number for length as a random step is 58. Therefore, starting from the south-west corner of the selected field, measure 58 steps along the length of the selected field and the point where reached, measure 51 steps perpendicular to the length and parallel to breadth of the selected field because 51 is the random number selected as random step for breadth. Thus, the point "A" where reached by measuring 51 steps, is the south-west corner of the CCE plot (Figure-6.5.1.2.1). The point "A" is also called as the key point or first corner of the CCE plot. Fix a peg at the key point of the CCE plot.

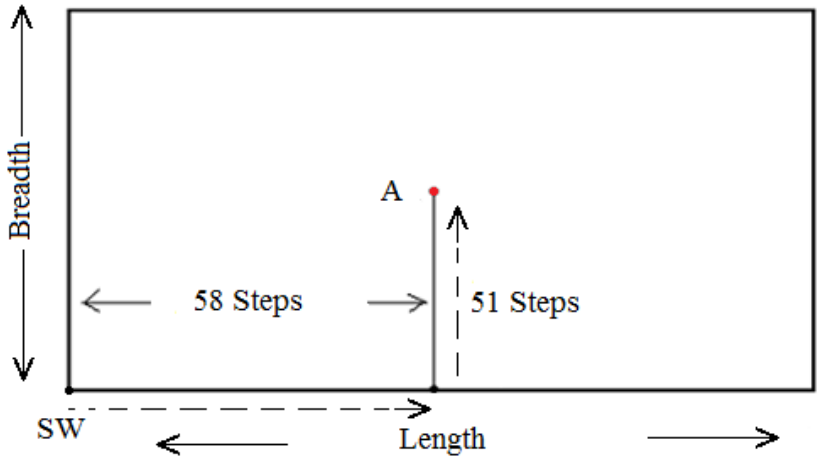


Figure-6.5.1.2.1: South-west corner of the CCE plot (Step-1)

2.2.2 Marking of second corner of the CCE plot

We measured five meter along the length of the selected field from corner "A" and the second point where we reached which is 5 meter away from corner "A" is the second corner "B" of the CCE plot. Fix a peg at corner "B" (Figure-6.5.1.2.2). The line joining "A" and "B" point is the base of the CCE plot.

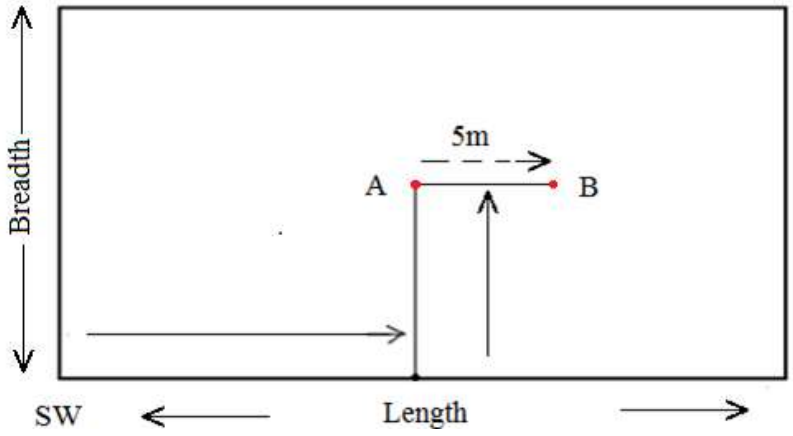


Figure-6.5.1.2.2: Second corner of the CCE plot (Step-2)

2.2.3 Marking of third corner of the CCE plot

Third and fourth corner of the CCE plot is to be marked with the help of right angle triangle method. To mark the third corner, let first person stands at corner "A" by holding the measuring tape at 0 meter mark and second person must has to stands at corner "B" holding at 12.07 (7.07 diagonal + 5.0 one side) meter mark on the same measuring tape. The third person holding at 7.07 [sq rt (5^2 + 5^2)] meter mark on the measuring tape should stretch the measuring tape in the direction of breadth of the selected field, the point where he reached is the third corner "C" of the CCE plot. The third corner is 7.07 meter (diagonal) away from corner "A" and 5 meter from corner "B". Fix a peg at corner "C" (Figure-6.5.1.2.3).

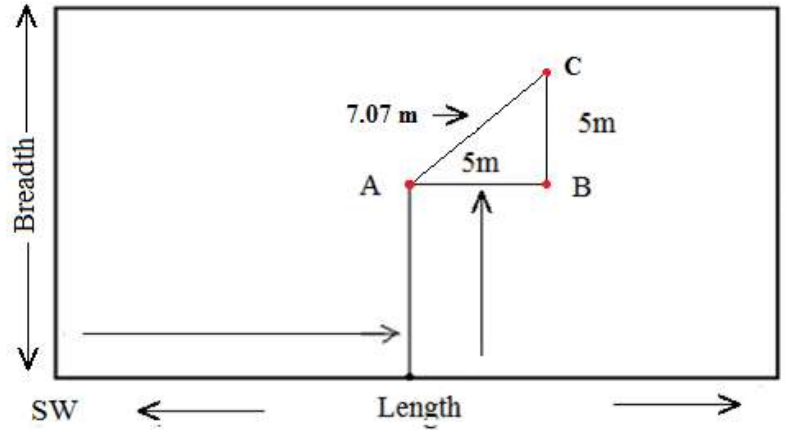


Figure-6.5.1.2.3: Third corner of the CCE plot (Step-3)

2.2.4 Marking of fourth corner of the CCE plot

For locating the fourth corner of the CCE plot, the third person should change the holding position on the measuring tape as 5.0 meter mark away from corner "A" and 7.07 meter away from corner "B". He should stretch the measuring tape in the direction of breadth of the field and where he reached is the fourth corner "D" of the CCE plot. Fix a peg at corner "D" (Figure-6.5.1.2.4).

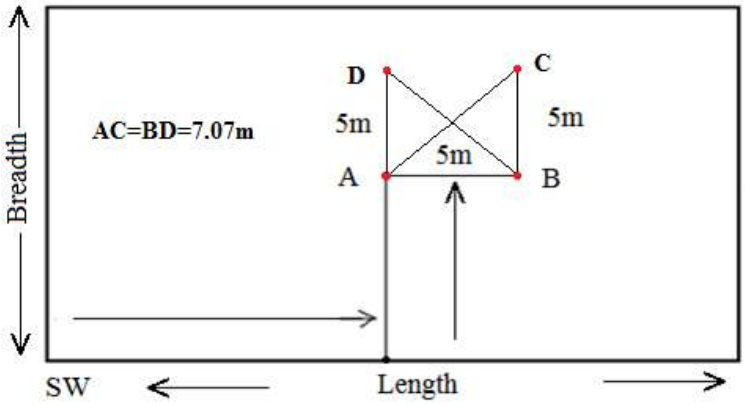


Figure-6.5.1.2.4: Fourth corner of the CCE plot (Step-4)

2.2.5 CCE plot

A, B, C and D is the four corners of the CCE plot. We have to check the distance 5 meter between the each corner A & B, B & C, C & D and A & D. The distance 7.07 meter between each diagonal AC and BD should also be checked (Figure- 6.5.1.2.5).

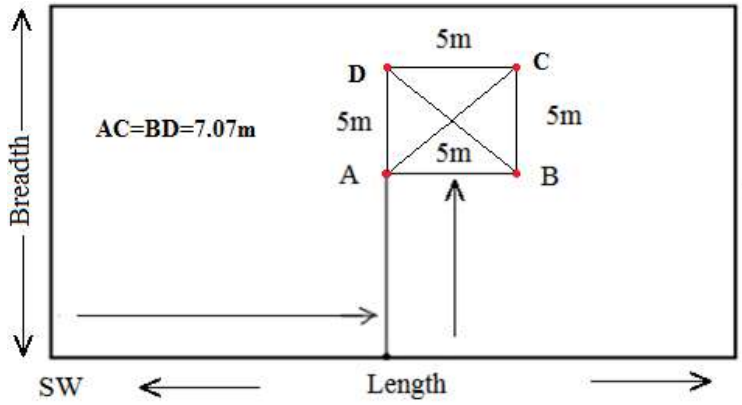


Figure-6.5.1.2.5: CCE plot (Step-5)

CCE plot

3.0 Demarcation of CCE plot when crop is sown in distinct rows in one direction

The crops like potato, redgram, sugarcane, caster, cotton, etc. are sown in rows without maintaining plant to plant distance within the row. Procedure of demarcation of the CCE plot is as under:

3.1 Enumeration of rows

Rows are to be enumerated starting from the south-west corner of the selected field. Conventionally, this side may be considered as breadth of the selected field (Figure-6.5.2.1).

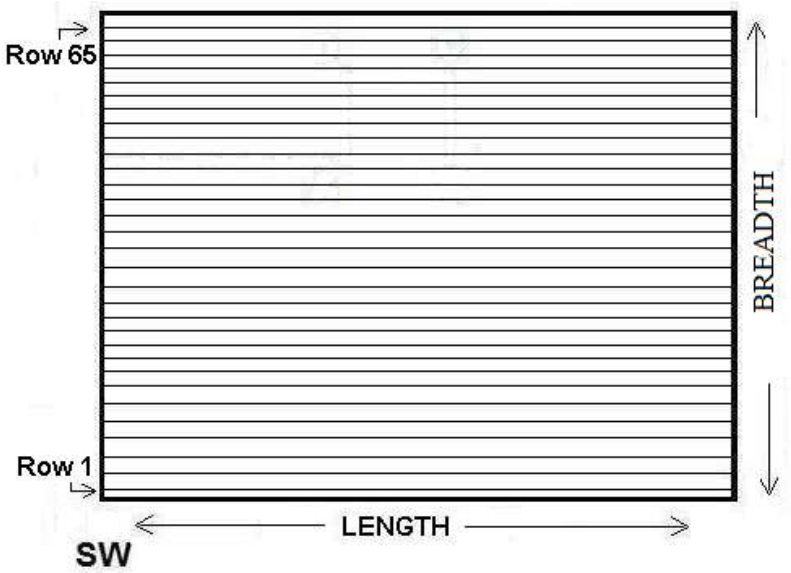


Figure-6.5.2.1: Enumeration of rows

3.2 Measurement of length of longest row

Measure the length (in normal steps) of longest row of the selected field (Figure-6.5.2.2).

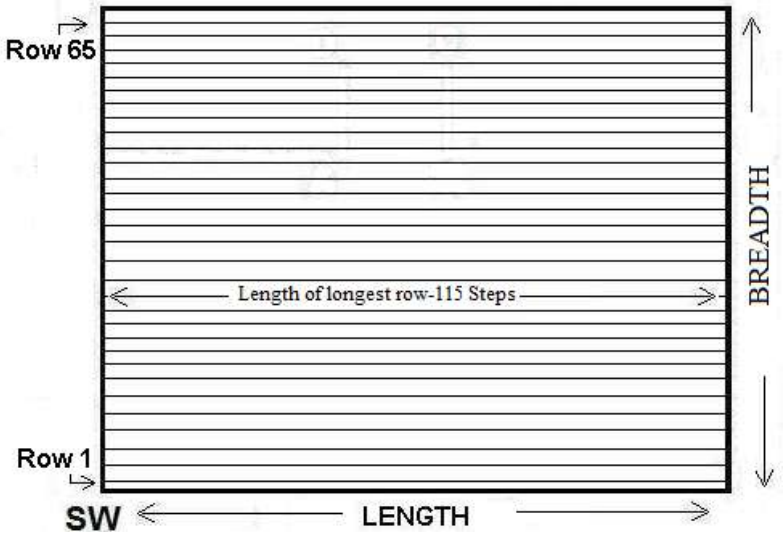


Figure-6.5.2.2: Length of longest row

3.3 Average number of rows in the breadth of CCE plot

Average number of rows in the specified breadth of CCE plot is to be workout. Therefore, observations of rows in the specified breadth of the CCE plot (5 meters or 10 meters in plains and 2 meters in hills) have to be taken at three randomly selected places in the selected field. The observations may be taken in the starting, middle and end point of the breadth of the selected field.

3.4 Determination of random number for random row

The average number of rows is to be deducted from the total number of rows in the selected field and add one. Deduction of average rows is necessary for ensuring that the CCE plot must be within the selected field. Addition of one is compulsory for inclusion of last row of the selected field in the CCE plot. A random number less than or equal to the number obtained after deducting average number of rows workout in the breadth of CCE and adding one should be selected using assigned column of random number table.

Example:

Let, total number of rows in the selected field is 65 and average number of rows workout in the specified breadth (5 meter) of CEE plot is 6. The deduction and addition of rows for selection of random number for random row is as under.

Total number of rows in the selected field	65
Average number of rows in 5 meter breadth of CCE plot	6
(Number of rows in the selected field **minus** Average number of rows in 5 meter breadth of CCE plot) + One	60

The number obtained after deducting 6 and adding one in total number of rows in the selected field is 60. Since 60 is the two digit number, therefore, using assigned column number **one** of two digited random number table, random number 22 is appeared first which is less than 60. Thus random number 22 is selected for identifying the random row. The random row (22) will be the first row of the CCE plot.

3.5 Determination of random number for length

Number of steps in specified length of CCE plot has to be deducted from the length (step) of longest row to ensure the whole CCE plot gets accommodate in the selected field. A random number which is less than or equal to the length obtained after deducting steps in specified length of CCE plot from the length (step) of longest row is to be selected using assigned column number of the random number table.

Example:

Let, the length of longest row is 115 steps and length of CCE plot is five meter. There are seven steps in five meter. The calculation the deduction of steps for selection of random number for random row is as under.

Length of longest row in the selected field (in steps)	115
Steps in the length of CCE plot	7
Length (in steps) of longest row **minus** steps in length of CCE plot	108

The number as obtained after deduction seven steps is 108 which is three digited number, therefore, using allotted column number **one** of three-digit random number table, select a random number less than or equal to 108. The number 10 is appeared first which is less than or equal to 108, therefore, random number 10 is considered as random step.

The selected random number pair is (22, 10) for locating south-west corner of the CCE plot.

3.6 Marking of the CCE plot

3.6.1 Marking of south-west corner of the CCE plot

Starting from first row from south west corner of the selected field, count the rows up to row number 22 i.e. random row. From the starting point of random row along its length moving between the inter-space of selected random row (22) and its preceding row (21) by measuring random steps (10) where we reached is the south-west corner "A" of the CCE plot (Figure-6.5.2.6.1). This may also be called as first corner or key point of the CCE plot. Fix a peg "A" at this point "A" in between the inter-space of the selected row (22) and its preceding row (21).

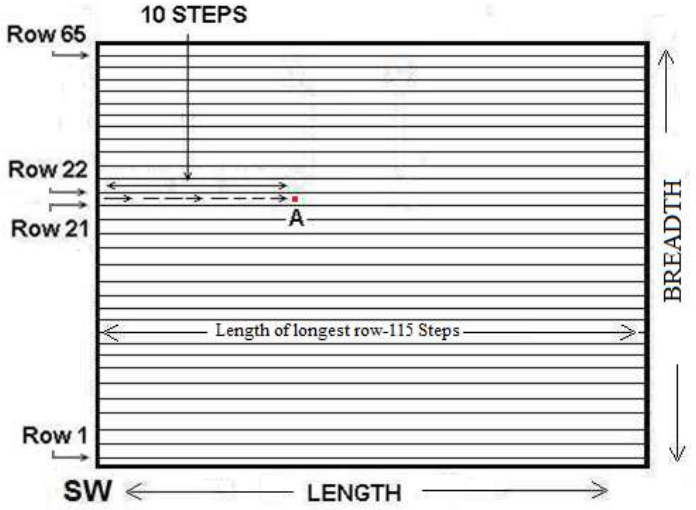


Figure-6.5.2.6.1: SW of the CCE plot (Step-1)

3.6.2 Marking of second corner of the CCE plot

Measure 5 meter meters (as per length of CCE plot) from the key point (First corner of CCE plot) moving in between random row (22) and its preceding row (21) toward the length of row and fix second peg "B" at other corner. It is the second corner of the CCE plot (Figure-6.5.2.6.2).

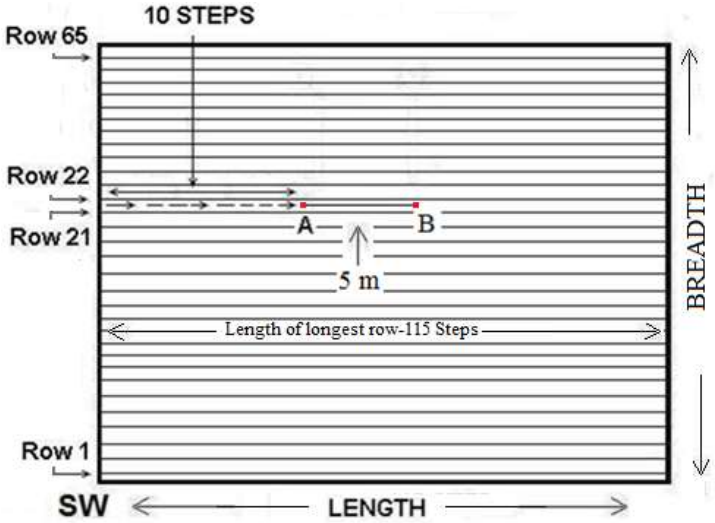


Figure-6.5.2.6.2: Second corner of the CCE plot (Step-2)

3.6.3 Marking of third corner of the CCE plot

Moving perpendicular direction to the random row 22 (second corner "B") towards inner side of the selected field start counting random row (22) as first row of the CCE plot and end the counting at sixth row (as average number of rows is six). Fix third peg "C" in between the inter-space of last row (i.e. 6^{th}) to be included in the CCE plot and its succeeding row. The sixth row of the CCE plot is 27^{th} row of the selected field and 28^{th} is the succeeding row (Figure-6.5.2.6.3).

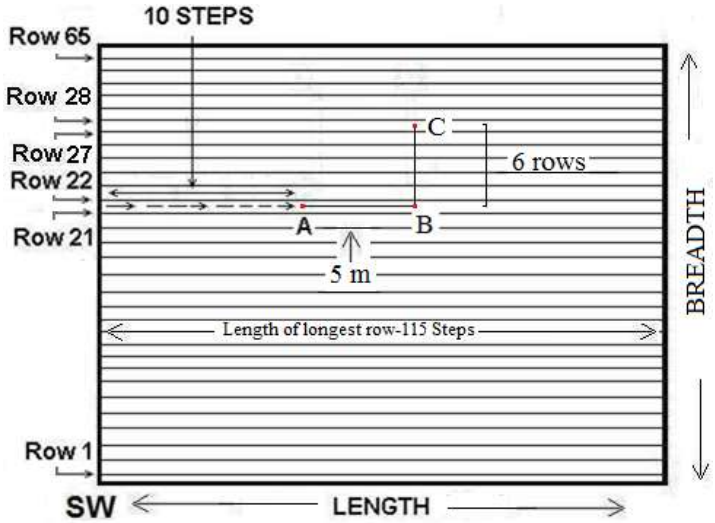


Figure-6.5.2.6.3: Third corner of the CCE plot (Step-3)

3.6.4 Marking of fourth corner of the CCE plot

Start moving from third corner "C" parallel to line "B"-"A" in the direction of Key point "A" in between last row i.e. sixth row of CCE plot or 27^{th} and its succeeding row 28^{th} of selected field, measure a length 5 meter (as par the length of CCE plot). The point where we reached is the fourth corner of the CCE plot. Fix fourth peg "D" at this point (Figure-6.5.2.6.4).

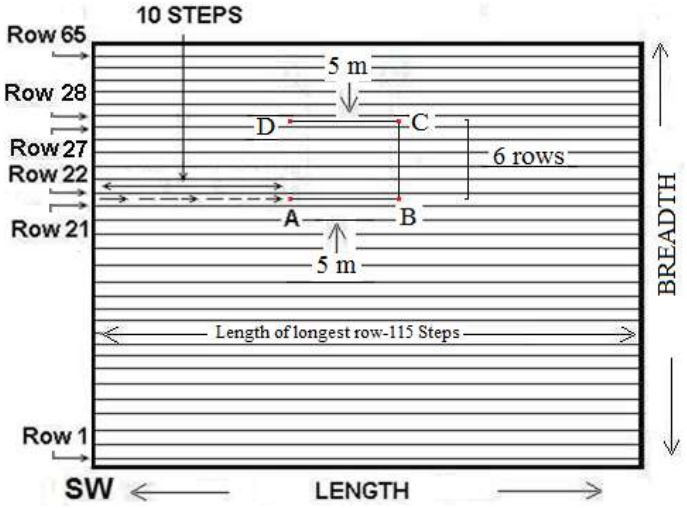


Figure-6.5.2.6.4: Fourth corner of the CCE plot (Step-4)

3.6.5 CCE plot

"A", "B", "C" and "D" is the four corner of the CCE plot. The number of rows in the breadth side between "B" and "C" corner should be equal to the number of rows between "A" and "D" corner (i.e. six). The distance between "A" "B" and "C" "D" corner should be equal to the length of CCE plot i.e. 5 meter. The distance of all the sides and diagonal should be measured and recorded (Figure-6.5.2.6.5).

If the CCE plot does not fall wholly within the selected field due to irregular shape of the field, reject the random number pair and select a new random number pair for making CCE plot.

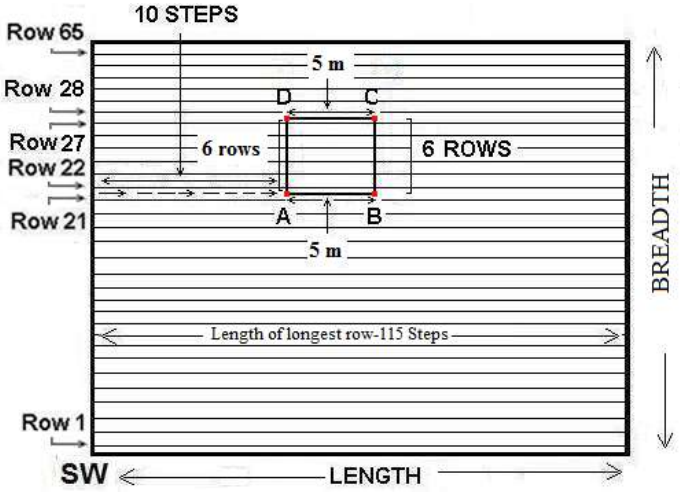


Figure-6.5.2.6.5: CCE plot (Step-5)

4.0 Demarcation of CCE plot when crop is sown in lines in two directions

The crop like tobacco is sown in both the directions in lines. The procedure for making CCE plot is slightly differ from the procedure of making the CCE plot when crop is sown in one direction in line. Procedure of demarcation of the CCE plot is as under:

4.1 Enumeration of rows

Rows are to be enumerated in both the directions i.e. length and breadth from the south-west corner of the selected field (Figure-6.5.3.1). Suppose 108 rows are in length side and 65 are in breadth side of the selected field.

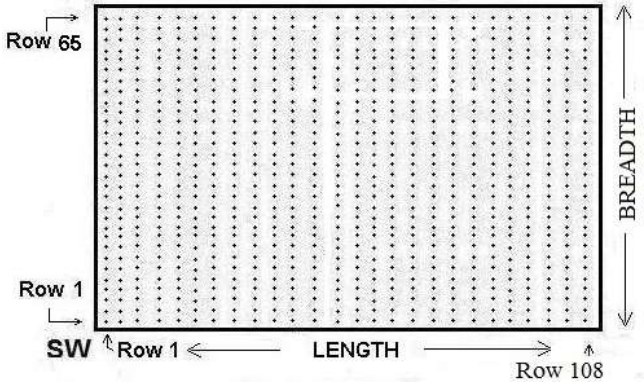


Figure-6.5.3.1: Enumeration of rows

4.2 Average number of rows

Average number of rows in the specified length and breadth of CCE plot is to be workout. Therefore, observations of rows in the specified length and breadth of the CCE plot (5 x 5 meters or 10 x10 meters or 10 x 5 meter in plains and 10 x 2 meters in hills) have to taken at three randomly selected places in the selected field. The observations may be taken in the starting, middle and end point of the length and breadth of the selected field.

4.3 Determination of random number for random row in the direction of length

Average number of rows in specified length may be deducted from the total number of rows in longer side i.e. length and add one for inclusion of last row in the CCE plot. Deduction of average number of rows is essential for ensuring that the whole CCE plot gets accommodate in the selected field.

Example:

Let, the shape of CCE plot is square having 5 meter length of each side. Suppose, average number of rows in 5 meter length are 6 and total number of rows are 108 in the direction of length, the deduction and addition of rows for selection of random number for random row is as under.

Total number of rows in the direction of length of selected field	108
Average rows in 5 meter length of CCE plot	6
(Total number of rows **minus** average rows in the length of CCE plot) + one	103

The number 103 is obtained after deducting average number of rows from total number of rows and addition of one. Hence 103 is three digited number, therefore, using assigned column number one of three digited random number table, a random number less than or equal to 103 is to be selected. The random number 48 is appeared, hence, it is considered as selected random number for random row for length side.

4.4 Determination of random number for random row in the direction of breadth

Average number of rows in specified breadth may be deducted from the total number of rows in shorter side i.e. breadth and add one for inclusion of last row in the CCE plot. Deduction of average number of rows is essential for ensuring that the whole CCE plot gets accommodate in the selected field.

Example:

Let, the shape of CCE plot is square having 5 meter length of each side. Suppose, average number of rows in 5 meter breadth is 8 and total number of rows are 65 in the direction of breadth, the deduction and addition of rows for selection of random number for random row is as under.

Total number of rows in the direction of breadth of the selected field	65
Average rows in 5 meter breadth of CCE plot	8
(Total number of rows **minus** average rows in the breadth of CCE plot) + one	58

The number 58 is obtained after deducting average number of rows from total number of rows and addition of one. Hence 58 is two digited number, therefore, using assigned

column number one of three digited random number table, a random number less than or equal to 58 is to be selected. The random number 22 is appeared first, hence, it is considered as selected random number for random row for breadth side.

The random number pair is (48, 22) for locating south-west corner of CCE plot.

4.5 Marking of the CCE plot

4.5.1 Marking of south-west corner of the CCE plot

Starting from south-west corner of the selected field, move towards the direction of length of the selected field by counting from row number one and stop at row number 48 which is selected as random number for row in the direction of length. From this point, move between the inter-space of selected random row (48) and its preceding row (47) towards the direction of breadth and perpendicular to the length of the selected field by counting from row number one and stop at row number 22 which is the selected as random number for row in the direction of breadth. Fix first peg "A" between the interspace of random row (22) selected for breadth and the preceding row (21). The point "A" is the south-west corner (key point) or first corner of the CCE plot (Figure-6.5.3.5.1).

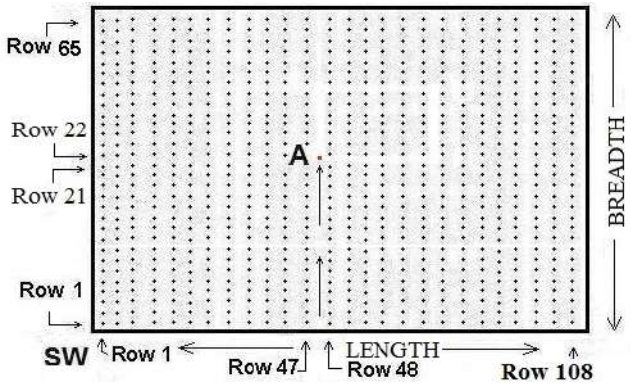


Figure-6.5.3.5.1: South-West corner of the CCE Plot

4.5.2 Marking of second corner of the CCE plot

From the key point "A" move in between the interspace of selected random row (22) and preceding row (21) by counting the average number of rows in the length of CCE plot (i.e. 6) in the direction of length of the selected field and stop at row number 6^{th} which is to be included in CCE plot. Row number 48^{th} and 53^{rd} of the selected field is the first and 6^{th} (last) row, respectively, of the CCE plot. Fix second peg "B" between the interspace of last row (i.e. 6^{th} of CCCE plot or 53^{rd} row of the selected field) and its succeeding row number 54^{th} row of the selected field (Figure-6.5.3.5.2).

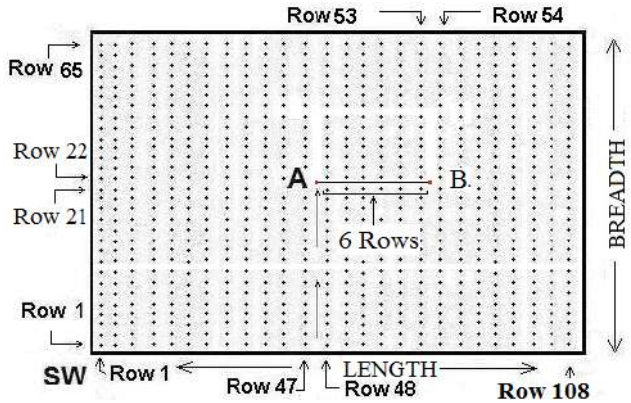


Figure-6.5.3.5.2: Second Corner of the CCE Plot

4.5.3 Marking of third corner of the CCE plot

From second corner "B" proceed along the breadth of the selected field between the interspace of last row of CCE (i.e. 6th row of CCE or 53rd of selected field) and its succeeding row (i.e. 54th) by counting the average number of rows workout in 5 meter breadth (i.e. 8) and stop at last row (8th) to be included in CCE plot (or row number 29th of selected field). Row number 22nd is the first row while 29th is 8th (last) row of the CCE plot in the direction of breadth. Fix third peg at "C" point between the interspace of last row (8th) of CCE plot (or row number 29 of selected field) and its succeeding row (row number 30). This is the third corner of CCE plot (Figure-6.5.3.5.3).

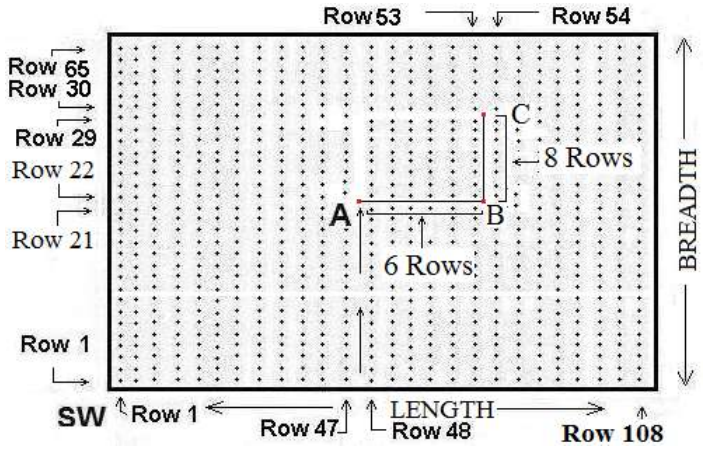


Figure-6.5.3.5.3: Third corner of the CCE plot

4.5.4 Fourth corner of the CCE plot

Proceed from third corner "C" along the interspace of 8th (last) row (i.e. row number 29 of selected field) of CCE plot and its succeeding row (i.e. row number 30 of selected field) parallel to line "A" - "B" and towards south-west corner of the CCE plot by counting average number of rows in the length of CCE plot (i.e. 6). We reached back between the interspace of selected random row 48 and preceding row number 47 in the direction of length. This is the fourth corner of CCE plot. Fix the fourth peg "D" at this point (Figure-6.5.3.5.4).

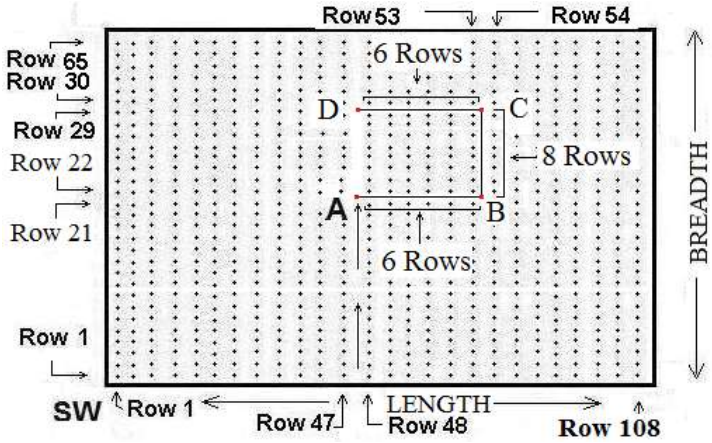


Figure-6.5.3.5.4: Fourth corner of the CCE plot

4.5.5 CCE plot

"A" "B" "C" "D" is the four corner of the CCE plot. The number of rows between A and B corner should be equal to the number of rows between C and D corner (i.e. 6) in length

side while in the breadth side rows between B and C corner should be equal to the number of rows between D and A (i.e. 8). The distance of all the sides and diagonal should be measured and recorded (Figure-6.5.3.6.5).

If the CCE plot does not fall wholly within the selected field due to irregular shape of the field, reject the random number pair and select a new random number pair for making CCE plot.

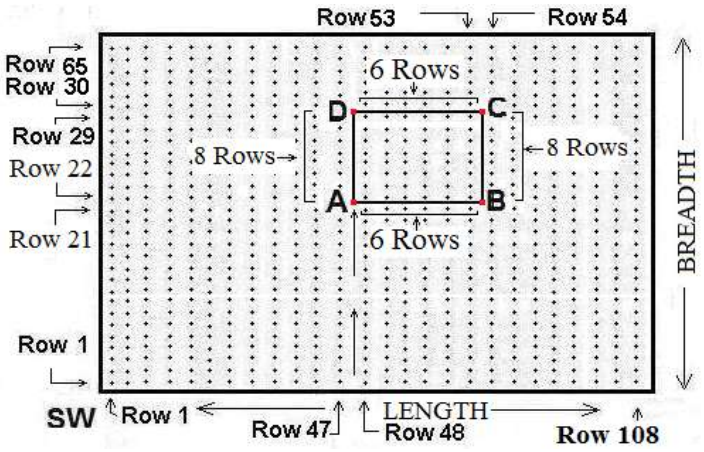


Figure-6.5.3.5.5: CCE plot

5.0 Harvesting of the crop of CCE plot

The boundary of the CCE plot should be demarcated a rope/string. The length of rope/string should not be increase on stretching. A well stretch rope/string should be tied around the tall and straight pegs firmly fixed on the ground and lowered gradually to the ground level for demarcating the boundary of the CCE plot. The decision for harvesting the plants of CCE plot is based on the position of roots. The plants on the boundary line of the CCE plot will be harvested only if the roots are more than half inside the CCE plot. All plants within the CCE plot have to be harvested and gathered carefully. The harvested pants should be bundled with coloured rope, marked properly with full identification particulars by permanent marker on tag and transported to proper place for drying/threshing/rotting/crushing operation etc. No plant and ear head should be fallen during harvesting, bundling and transporting. Weight of each bundle should be taken to the nearest possible weighing unit by a perfect weighing balance / machine.

5.0 Harvesting of the CCE crop

6.0 Threshing of the crop of CCE plot

The harvested CCE crop should be spread on a piece of hessian cloth for drying in sun light. After proper drying the crop, it should be threshed carefully as per the usual method.

6.0 Threshing the crop of CCE plot

7.0 Winnowing and cleaning of threshed crop of CCE plot

Grains from straw should be separated by winnowing with the help of wind, winnowing fan and other cleaning tool. The produce should be free from seed of other crops, weed seed, dust particles, stone, husk etc.

7.0 Winnowing and cleaning

8.0 Weighing of the wet produce (Wet weight)

Weight of clean produce should be taken just after its winnowing/ cleaning. At this time most of the crops have excess moisture, therefore, this weight is called as wet weight.

Weight should be taken to the nearest possible weighing unit by a perfect weighing balance / machine. After weighing, the produce should be returned to the farmer.

8.0 Weighing of the wet produce

9.0 Drying of produce (Driage experiment)

Driage experiments are necessary to obtain final estimates of yield in terms of dry produce. Therefore, driage experiments are conducted, if the produce has excess moisture. Sample of recommended quantity of the produce has to be taken in cloth bag and kept for drying in the sun light as per local practice. Driage experiments for different crops are to be conducted by the district statistical officer and selected out of the CCE supervised by the district level officers at the district level. The driage experiments are conducted in respect of 15 per cent of the experiments planned for the specific crops or subject to a minimum four experiments per crop.

Generally, one kilogram sample of harvested produce should be taken at random for drying by the District Statistical Supervisor. If, the produce obtained from the CCE plot is less than one kilogram, the entire produce is to be taken for drying. In the case of sugarcane, the final produce is expressed in terms of cane only while in the case of cotton, the final produce is expressed in terms of lint. The cotton (Kapas) is converted into lint by using ginning percentage (kapas to lint) which is obtained from the ginning factories.

In case of jute and similar crop, the labelled bundles should be left in the field or any proper place for drying the leaves for one or two days. After that the bundles should be dipped under the water in the pond or pit for 10 to 15 days for rotting as per local practice. The date for extracting the fibre may be fixed with the consultation of the farmer. The fibre should be extracted, washed, cleaned properly and kept for drying in the sun light as per local practice. When the fibre dried properly, the weight of dry fibre should be taken to the nearest possible weighing unit up to ten gram.

10.0 Weighing of dry produce (Dry weight)

Weight of dry produce should be taken to the nearest possible weighing unit by a perfect weighing balance / machine after proper sun drying. This weight is called as dry weight of the produce. The produce should be returned to the farmer.

REFERENCES

Ahmad, T., Sahoo, P.M., Singh, M. and Biswas, A. (2024). Crop Cutting Experiment Techniques for Determination of Yield Rates of Field Crops. Monograph. ICAR-Indian Agricultural Statistics Research Institute. Monograph No.: I.A.S.R.I./M-01/2024. http://krishi.icar.gov.in/jspui/handle/123456789/83961

Mahalanobis, P.C. (1945). A Report on Bihar Crop Survey, 1943-44, Sankhya, 7, 29-118.

Panse, V.G. and Sukhatme, P.V. (1967). Statistical Method for Agricultural Workers, ICAR Publication.

Panse, V.G. (1946 b). Plot size in Yield Surveys on Cotton, Curr. `Sci. 15, 218-19.

Panse, V.G. (1947). Plot size in Yield Surveys, Nature, 15, 159, 820.

Raut, K.C. and Singh, D. (1976). Methods of Collection of Agricultural Statistics in India, IASRI publication.

Sukhatme, P.V. (1946 a). Bias in the Use of Small Size Plots in Sample Surveys for Yield, Curr. Sci. 15, 119-20.

Sukhatme, P.V. (1946 b). Bias in the Use of Small Size Plots in Sample Surveys for Yield, Nature, 15, 7, 630.

Sukhatme, P.V. (1947 a). The Problem of Plot Size in Large-scale Yield Surveys, Jour. Amer. Stat. Assoc., 42, 297-310.

Sukhatme, P.V. and Panse, V.G. (1951). Crop surveys in India-II. *Jour. Ind. Soc. Agril. Statist*. Vol.III:(2), 95-168.

ADAPTIVE CLUSTER SAMPLING WITH APPLICATIONS

Ankur Biswas and Raju Kumar

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Consider a survey of a rare and endangered bird species in which observers record the number of individuals of the species seen or heard at sites or units within a study area. At many of the sites selected for observation, zero abundance may be observed. But wherever substantial abundance is encountered, observation of neighbouring sites is likely to reveal additional concentrations of individuals of the species. Similar patterns of clustering or patchiness are encountered with many other types of animals from whales to insects, with vegetation types from trees to lichens, and with mineral and fossil fuel resources. A related pattern is found in epidemiological studies of rare, contagious diseases. Whenever an infected individual is encountered, addition to the sample of closely associated individuals reveals a higher than expected incidence rate. In such situations, the field workers may feel the inclination to depart from the preselected sample plan and add nearby or associated units to the sample.

Adaptive cluster sampling refers to designs in which an initial set of units is selected by some probability sampling procedure, and, whenever the variable of interest of a selected unit satisfies a given criterion, additional units in the neighbourhood of that unit are added to the sample. Adaptive cluster sampling provides a means of taking advantage of clustering tendencies in a population, when the locations and shapes of the clusters cannot be predicted prior to the survey. Thompson (1990, 1991a, 1991b) described some designs in which, whenever the observed value of a selected unit satisfies a condition of interest, additional units are added to the sample from the neighbourhood of that unit. Many purposes may be served by such a design such as increasing the "yield" of interesting units. For such surveys, Birnbaum and Sirken (1965) obtained unbiased estimators of the Hansen and Hurwitz (1943) type, in which observations are divided by draw-by-draw selection probabilities, and of the Horvitz and Thompson (1952) type, in which observations are divided by inclusion probabilities.

The designs given by Thompson (1990) are related to network sampling in that selection of certain units may lead to observation of others. Because of the way the decisions to observe additional units depend adaptively on the observed values of the variable of interest, however, the selection and inclusion probabilities are not in general known for all units in the sample. Modifications must, therefore, be made in estimators of the Hansen-Hurwitz or Horvitz-Thompson types to obtain unbiased estimators.

2. Sampling Design

The basic idea of the adaptive cluster sampling design is illustrated in Figure1. Suppose that the interest lies in studying a particular weed that grows in strawberry fields. The weed is not particularly abundant, but serves as a host plant for a disease of strawberries. The purpose of the estimation of the total (and average) number of weeds in the field can be achieved using

adaptive cluster sampling. The field is divided using a grid system to produce 400 square contiguous sampling units. An initial random sample of 10 units is shown in Figure 1a. Whenever one or more of the objects is observed in a selected unit, the adjacent neighbouring units to the left, right, top and bottom are added to the sample. When this process is completed, the sample consists of 45 units, shown in Figure 1b. Neighbourhoods of units may be defined in many ways other than the spatial proximity system of this example.

In the designs considered here, the initial sample consists of a simple random sample of n_1 units, selected either with or without replacement. As in the usual finite population sampling situation, the population consists of N units with labels 1, 2,. . ., N and with associated variables of interest $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$. The sample s is a set or sequence of labels identifying the units selected for observation. The data consists of the observed y-values together with the associated unit labels. The object of interest is to estimate the population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ or total $N\mu$ of the y-values.

A sampling design is a function $p(s|\mathbf{y})$ assigning a probability to every possible sample s. In designs such as those described here, these selection probabilities depend on the population y-values. It is assumed that for every unit i in the population a **neighbourhood** A_i is defined, consisting of a collection of units including i. These neighbourhoods do not depend on the population y-values. In the spatial sampling example, the neighbourhood of each unit consists of a set of geographically nearest neighbours, but more elaborate neighbourhood patterns are also possible, including a larger contiguous set of units or a non-contiguous set such as a systematic

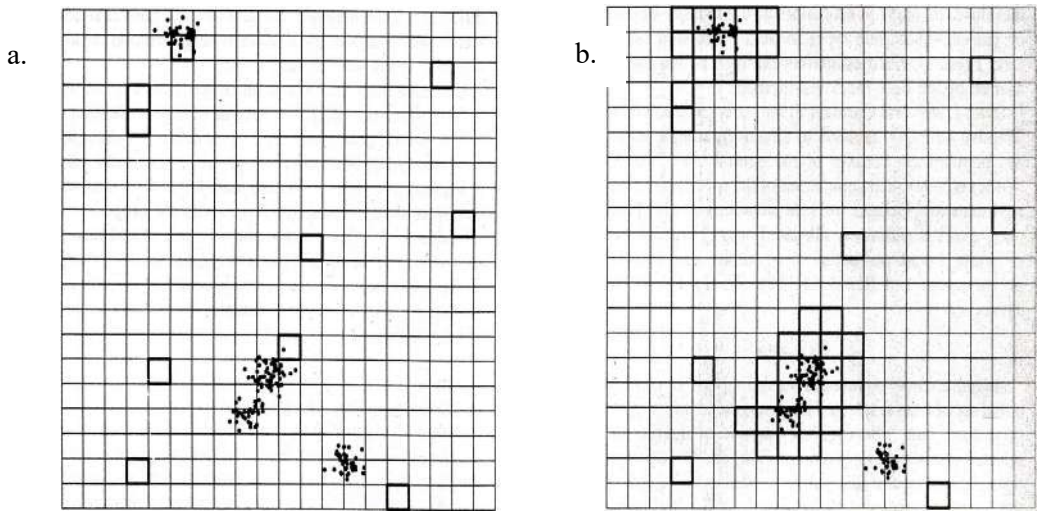


Figure 1. Adaptive cluster sampling to estimate the number of point-objects in a study region of 400 units. An initial random sample of 10 units is shown in (a). Adjacent neighbouring units are added to the sample whenever one or more of the objects of the population are observed in a selected unit. The resulting sample of 45 units is shown in (b).

Grid pattern around the initial unit. In other sampling situations, neighbourhoods may be defined by social or institutional relationships between units. The neighbourhood relation is symmetric: if unit *j* is in the neighbourhood of unit *i*, then unit *i* is in the neighbourhood of unit *j*.

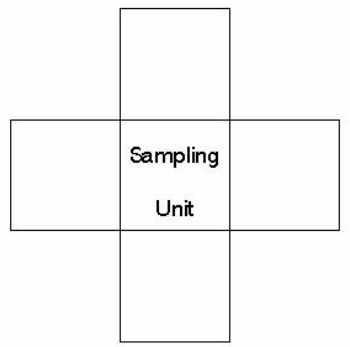


Figure 2. Neighbourhood for a sampling unit in the strawberry field study.

The **condition** for additional selection of neighbouring units is given by an interval or set *C* in the range of the variable of interest. The unit *i* is said to satisfy the condition if $y_i \in C$. In the examples, a unit satisfies the condition if the variable of interest y_i is greater than or equal to some constant *c*, that is, $C=\{ y : y \geq c \}$.

When a selected unit satisfies the condition, all units within its neighbourhood are added to the sample and observed, some of these units may in turn satisfy the condition and some may not. For any of these units that do satisfy the condition, the units in its neighbourhood are also included in the sample, and so on.

Consider the collection of all of the units that are observed under the design as a result of initial selection of unit *i*. Such a collection, which may consist of the union of several neighbourhoods, will be termed as **cluster** when it appears in a sample. Within such a cluster there is a sub-collection of units, termed as a **network**, with the property that selection of any unit within the network would lead to inclusion in the sample of every other unit in the network. In the example of Figure 1, inside either of the obvious clusters of units in the final sample, the sub-collection of units with one or more of the point-objects forms a network.

Any unit not satisfying the condition but in the neighbourhood of one that does is termed an **edge unit**. Although selection of any unit in the network will result in inclusion of all units in the network and all associated edge units, selection of an edge unit will not result in the inclusion of any other units. It is convenient to consider any unit not satisfying the condition a network of size one, so that, given the *y*-values, the population may be uniquely partitioned into networks.

When the initial sample of n_1 units is selected by simple random sampling without replacement, the n_1 units in the initial sample are distinct because of the without-replacement sampling, but the data may nevertheless contain repeat observations due to selection in the initial sample of more than one unit in a cluster. The unit *i* will be included in the sample either if any unit of the network to which it belongs (including itself) is selected as part of the initial sample or if any unit of a network of which unit *i* is an edge unit is selected. Let m_i

denote the number of units in the network to which unit i belongs, and let a_i denote the total number of units in networks of which unit i is an edge unit. Note that if unit i satisfies the criterion C then $a_i = 0$, whereas if unit i does not satisfy the condition then $m_i = 1$. The probability of selection of unit i on any one of the n_1 draws is $p_i = (m_i + a_i)/N$. The probability that unit i is included in the sample is

$$\alpha_i = 1 - \binom{N - m_i - a_i}{n_1} \Big/ \binom{N}{n_1} \tag{2.1}$$

When the initial simple is selected by simple random sampling with replacement, repeat observations in the data may occur due either to repeat selections in the initial sample or to initial selection of more than one unit in a cluster. With this design, the draw-by-draw selection probability is $p_i = (m_i + a_i)/N$ and the inclusion probability is

$$\alpha_i = 1 - (1 - p_i)^{n_1} \tag{2.2}$$

With either initial design, neither the draw-by-draw selection probability p_i nor the inclusion probability α_i can be determined from the data for all units in the sample, because some of the a_i may be unknown.

3. Estimators for Population Parameters

Classical estimators such as the sample mean $\bar{y}$, which is an unbiased estimator of the population mean under a non-adaptive design such as simple random sampling, or the mean of the cluster means $\bar{\bar{y}}$, which is unbiased under cluster sampling with selection probabilities proportional to cluster sizes, are biased when used with the adaptive designs described earlier. These biases are demonstrated later in example. In this section several estimators that are unbiased for the population mean under the adaptive designs are given.

The expected value of an estimator t is defined in the design sense, that is, $E[t] = \sum t_s . p(s \mid y)$, where t_s is the value of the estimate computed when sample s is selected, $p(s|y)$ is the design, and the summation is over all possible samples s. The sampling strategy i.e. the estimator together with the design, is design unbiased for the population mean if $E[t] = \frac{1}{N}\sum_{i=1}^{N} y_i$ for all population vectors $\boldsymbol{y}$.

3.1 The Initial Sample Mean

If the initial sample in the adaptive design is selected by simple random sampling, with or without replacement, the mean, $\bar{y}$ of the n_1 initial observations is an unbiased estimator of the population mean. This estimator ignores all observations in the sample other than those initially selected.

3.2 A Modified Hansen-Hurwitz Type of Estimator

For sampling designs in which n units are selected with replacement and the probability p_i of selecting unit i on any draw is known for all units, the Hansen-Hurwitz estimator, in which each y-value is divided by the associated selection probability and multiplied by the number of times the unit is selected, is an unbiased estimator of the population mean.

With the adaptive cluster sampling design, the selection probabilities are not known for every unit in the sample. An unbiased estimator can be formed by modifying the Hansen-Hurwitz estimator to make use of observations not satisfying the condition only when they are selected as part of the initial sample. Let Ψ_k denote the network that includes unit k, and let m_k be the number of units in that network. (Recall that a unit not satisfying the criterion is considered a network of size one.) Let $\bar{y}_k^*$ represent the average of the observations in the network that includes the k^{th} unit of the initial sample, that is,

$$\bar{y}_k^* = \frac{1}{m_k} \sum_{j \in \Psi_k} y_j .$$

The modified estimator is

$$t_{HH^*} = \frac{1}{n_1} \sum_{k=1}^{n_1} \bar{y}_k^* . \qquad \ldots(3.2.1)$$

The variance of t_{HH^*} is

$$Var(t_{HH^*}) = \left(\frac{1}{n_1} - \frac{1}{N} \right) \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_i^* - \mu)^2 , \qquad \ldots(3.2.2)$$

if the initial sample is selected without replacement and

$$Var(t_{HH^*}) = \frac{1}{n_1} \frac{1}{(N-1)} \sum_{i=1}^{N} (\bar{y}_i^* - \mu)^2 \qquad \ldots(3.2.3)$$

if the initial sample is selected with replacement.

An unbiased estimator of this variance is

$$\hat{V}ar(t_{HH^*}) = \left(\frac{1}{n_1} - \frac{1}{N} \right) \frac{1}{(n_1 - 1)} \sum_{k=1}^{n_1} (\bar{y}_k^* - t_{HH^*})^2 , \qquad \ldots(3.2.4)$$

if the initial sample is selected without replacement and

$$\hat{V}ar(t_{HH^*}) = \frac{1}{n_1} \frac{1}{(n_1 - 1)} \sum_{k=1}^{n_1} (\bar{y}_k^* - t_{HH^*})^2 \qquad \ldots(3.2.5)$$

if the initial sample is selected with replacement.

3.3 A Modified Horvitz-Thompson Type of Estimator

For sampling designs in which the probability α_i that unit *i* is included in the sample is known for every unit, the Horvitz-Thompson estimator, in which each y-value is divided by the associated inclusion probability, is an unbiased estimator of the population mean.

With the adaptive designs here, the inclusion probabilities are not known for all units included in the sample. An unbiased estimator can be formed by modifying the Horvitz-Thompson estimator to make use of observations not satisfying the condition only when they are included in the initial sample. Then the probability that a unit is used in the estimator can be computed, even though its actual probability of inclusion in the sample may be unknown. If the initial sample is selected by simple random sampling without replacement, define

$$\alpha_k^* = 1 - \binom{N - m_k}{n_1} \Big/ \binom{N}{n_1} \quad , \tag{3.3.1}$$

where m_k is the number of units in the network that includes unit k. If the initial selection is made with replacement, define $\alpha_k^* = 1 - (1 - m_k/N)^{n_1}$. For any unit not satisfying the condition, $m_k = 1$.

Let the indicator variable J_k be 0 if the k^{th} unit in the sample does not satisfy the condition and was not selected in the initial sample; otherwise, J_k=1. The modified estimator is

$$t_{HT*} = \frac{1}{N} \sum_{k=1}^{\nu} y_k J_k / \alpha_k^* , \tag{3.3.2}$$

where ν is the number of distinct units in the sample.

To obtain the variance of t_{HT*}, it will be most convenient to change notation to deal with the networks into which the population is partitioned, rather than individual units. Let ς denote the number of networks in the population and let Ψ_j be the set of units comprising the j^{th} network. Let m_j be the number of units in network *j*. The total of the y-values in network *j* will be denoted by $y_j = \sum_{i \in \psi_j} y_i$.

The probability α_i^* that the unit *i* is used in the estimator is the same for all units within a given network *j*; this common probability will be denoted by π_j. The probability π_{jh} that the initial sample contains at least one unit in each of the networks *j* and *h* is

$$\pi_{jh} = 1 - \left\{ \binom{N - m_j}{n_1} + \binom{N - m_h}{n_1} - \binom{N - m_j - m_h}{n_1} \right\} \Big/ \binom{N}{n_1}, \tag{3.3.3}$$

when the initial sample is selected without replacement and

$$\pi_{jh} = 1 - \left[\left\{1 - m_j/N\right\}^{n_1} + \left\{1 - m_h/N\right\}^{n_1} - \left\{1 - \left(m_j + m_h\right)/N\right\}^{n_1} \right] , \tag{3.3.4}$$

when the initial sample is selected with replacement.

With the convention that $\pi_{jj} = \pi_j$, the variance of the estimator t_{HT*} is

$$\mathrm{Var}(t_{HT*}) = \frac{1}{N^2}\sum_{j=1}^{\varsigma}\sum_{h=1}^{\varsigma} y_j y_h \left(\pi_{jh} - \pi_j\pi_h\right)\Big/\left(\pi_j\pi_h\right) \qquad (3.3.5)$$

An unbiased estimator of the variance of t_{HT*} is

$$\hat{\mathrm{V}}\mathrm{ar}(t_{HT*}) = \frac{1}{N^2}\sum_{k=1}^{\kappa}\sum_{m=1}^{\kappa} y_k y_m \left(\pi_{km} - \pi_k\pi_m\right)\Big/\left(\pi_k\pi_m\pi_{km}\right), \qquad (3.3.6)$$

where the summation is over the κ distinct networks represented in the initial sample.

3.4. A Small Example

In this section, the sampling strategies are applied to a very small population to shed light on the computations and properties of the adaptive strategies in relation to each other and to conventional strategies. The population consists of just five units, the y-values of which are {1, 0, 2, 10, 1000}. The neighbourhood of each unit includes all adjacent units. The condition is defined by $C = \{y : y \geq 5\}$. The initial sample size is $n_1 = 2$.

With the adaptive design in which the initial sample is selected by simple random sampling without replacement, there are ${}^5C_2 = 10$ possible samples, each having probability 1/10. The resulting observations and the values of each estimator are listed in Table 1.

In this population, the 4th and 5th units, with the y-values 10 and 1000, respectively, form a network, and the 3rd, 4th and 5th units, with y-values 2, 10 and 1000, respectively, form a cluster. In the fourth row of the table, the 1st and 5th units, with y-values 1 and 1000, were selected initially; since $1000 \geq 5$, the single neighbour of the 5th unit, having y-value 10, is added to the sample. Since y-value 10 also exceeds 5, the neighbouring unit with y-value 2 is also added to the sample.

Table 1. All possible outcomes of Adaptive Cluster Sampling for a population of five units with y-values 1, 0, 2, 10 and 1000 in which the neighbourhood of each unit consists of itself plus adjacent units.

Observations	$\bar{y}_1$	t_{HH*}	t_{HT*}	$\bar{y}$	$\bar{\bar{y}}$
1, 0	0.50	0.50	0.50	0.50	0.50
1, 2	1.50	1.50	1.50	1.50	1.50
1, 10; 2, 1000	5.50	253.00	289.07	253.25	169.67
1, 1000; 10, 2	500.50	253.00	289.07	253.25	169.67
0, 2	1.00	1.00	1.00	1.00	1.00

0, 10; 2, 1000	5.00	252.50	288.57	253.00	168.67
0, 1000; 10, 2	500.00	252.50	288.57	253.00	168.67
2, 10; 1000	6.00	253.50	289.57	337.33	337.33
2, 1000; 10	501.00	253.50	289.57	337.33	337.33
10, 1000; 2	505.00	505.00	288.57	337.33	337.33
Mean	202.6	202.6	202.6	202.75	169.17
Bias	0	0	0	0.15	-33.43
MSE	59615	22862	17418.4	18660	18086

The computations for the estimators are- $t_{HH*} = (1 + (10 + 1000) / 2) / 2 = 253$ and $t_{HT*} = (1/0.4 + 10/0.7 + 1000/0.7) / 5 = 289.07$, in which $\alpha_1^* = 1 - \binom{4}{2} \Big/ \binom{5}{2} = 0.4$ and $\alpha_2^* = \alpha_3^* = 1 - \binom{3}{2} \Big/ \binom{5}{2} = 0.7$. The classical estimator $\bar{y} = 253.25$ is obtained by averaging all four observations in the sample, and $\bar{\bar{y}} = (1 + (10 + 2 + 1000) / 3) / 2 = 169.67$.

The population mean is 202.6 and the population variance (defined with N-1 in the denominator) is 198718. From the Table 1 it is clear that the unbiased adaptive strategies indeed have mean 202.6 and the estimators $\bar{y}$ and $\bar{\bar{y}}$, used with the adaptive design, are biased.

From the variances and MSEs given in the last row of the Table 1, it is clear that for this population, the adaptive design with the estimator t_{HT*} has the lowest variance among the unbiased strategies and all of the adaptive strategies are more efficient than simple random sampling.

5. Conclusions

Adaptive cluster sampling appears to be an effective method for sampling from populations with rare events as well as aggregation tendencies in these rare events. Unbiased estimators can be obtained by modifying the estimators of the Hansen-Hurwitz or Horvitz-Thompson types in case of adaptive cluster sampling. As per the example shown here, the adaptive Horvitz-Thompson estimator t_{HT*} clearly outperformed its Hansen-Hurwitz counterpart t_{HH*} and all of the adaptive strategies are more efficient than simple random sampling.

References

Birnbaum, Z.W. and Sirken, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: three unbiased estimates. *Vital and Health Statistics*, Ser. **2**, No. **11**, Washington, D.C.: Govt. printing office.

Hansen, M.M. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, **14**:333-362.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* **47**:663-685.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, **85**:1050-1059.

Thompson, S.K. (1991a). Stratified adaptive cluster sampling. *Biometrika*, **78**:389-397.

Thompson, S.K. (1991b). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics*, **47**:1103-1115.

RANKED SET SAMPLING AND APPLICATIONS

Ankur Biswas

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

The method of Simple Random Sampling (SRS) is the most commonly used method of sampling. The reason lies in its simplicity in selection as well as mathematical derivation. The probability of selection of every sample in the method of SRS is equal. Further, the units are selected one by one and the probability of selection of every unit of population in the sample is same.

The selection of units in the sample following SRS is purely random. Thus, it may happen that all the units selected in the sample may belong to one type or representing some part of the population only. Thus, one may end up with a sample where certain parts of the population are over represented while some other parts are under represented. Or in other words, the selected sample may not be representative enough resulting in misleading inferences about the population under study. An improved sampling mechanism which is capable of producing representative samples is, therefore, very much a practical necessity.

In agricultural, environmental and ecological sampling one may encounter a situation where the exact measurement (or quantification) of a selected unit is either difficult or expensive in terms of time, money or labour, but where the ranking of a small set of selected units according to the character of interest can be done with reasonable success on the basis of visual inspection or any other rough method not requiring actual measurement.

Suppose the objective is to estimate the distribution of volume of trees in a forest. If the forest were believed to be homogenous, a simple random sample could be taken by choosing the nearest tree to each of a set of randomly selected coordinates across the region of the forest. If homogeneity were less believable, the forest could be grided and trees randomly selected from within each grid-cell. Natural forests, however, are not so conveniently arranged. Stratification, clustering and various other area sampling schemes could be considered in such a situation.

Characteristics of these sampling mechanisms are simple random sampling at the ultimate stage of sampling. Replacement of SRS in the ultimate smallest group by some other efficient sampling mechanism may lead to further increase in the precision of sample estimates.

In statistical settings where actual measurements of the sample observations are difficult or costly or time consuming or destructive etc. but acquisition and subsequent ranking of the potential sample data is relatively easy, improved methods of statistical inference can result from using Ranked Set Sampling (RSS) technique. In what follows, we describe the method of RSS.

Consider the example explained earlier. Select two trees randomly and make judgement with the help of eyes about the content of wood. Mark the tree having lesser wood content and discard the one having higher wood content. Next, select two more trees, make judgement

through eyes and mark the tree having higher wood content and discard the other one. Repeat the procedure of alternately selecting the tree having lesser wood content and the other having higher wood content 25 times. Thus, out of 100 randomly selected trees only 50 are retained. Out of these 50 trees, 25 are from a stratum of trees generally having lesser wood content and the other 25 are from a stratum of trees having higher wood content. These 50 kept trees constitute the Ranked Set Sample. The sample so selected is expected to contain trees of almost all the sizes. Thus, it is likely to provide a better representation of trees in the population as compared to the method of SRS.

In situations where visual inspection is not directly available, ranking can sometimes be done on the basis of a covariate that is more accessible requiring less costs than, but correlated with, the character of interest. Thus, if we are interested in the volumes of trees, we may use the ranking by diameter to approximate the ranking by volume. This procedure is called as ranking using concomitant variables. This was first discussed by Stokes (1977) and referred it as "ranked set sampling with concomitant variables".

2. Method of RSS

The RSS procedure with equal allocation involves randomly drawing m^2 units from a population with mean μ and a finite variance σ^2 and then randomly partitioning them into m equal-sized sets with set size m. The units are then ranked within each set with respect to other than variable of interest. Here, ranking of the units could be based on visual inspection, judgement, auxiliary information or by some other relatively inexpensive methods not requiring actual measurement of the variable of interest. The unit receiving the smallest rank is accurately quantified from the first set, the unit receiving the 2^{nd} smallest rank is accurately quantified from the 2^{nd} set, and so forth, until the unit with largest rank is accurately quantified from the m^{th} set. This constitutes one cycle. This procedure involves the measurement of m units out of the m^2 originally selected units. The entire cycle is replicated r times until altogether $n = mr$ observations have been quantified out of m^2r originally selected units. These n quantified units constitute the ranked set sample.

Example: Consider the set size $m = 3$ with $r = 4$ cycles. This situation is illustrated in figure 1, where each row denotes a judgment-ordered sample within a cycle, and the units selected for quantitative analysis are circled. Here, 36 units have been randomly selected in 4 cycles; however, only 12 units are actually measured to obtain the ranked set sample for quantitative analysis.

Cycles	**Rank**		
	1	2	3
1	Θ	.	.
	.	Θ	.
	.	.	Θ

2	Θ . .	. Θ .	. . Θ
3	Θ . .	. Θ .	. . Θ
4	Θ . .	. Θ .	. . Θ

Figure 1: A ranked set sample with set size $m = 3$ and no. of sampling cycles $r = 4$.

3. RSS Estimator and its Variance

Let us consider only one cycle first. Let, X_{11}, X_{12}, …, X_{1m}, X_{21}, X_{22}, …, X_{2m}, …, X_{m1}, X_{m2}, …, X_{mm} be independent random variables all having the same cumulative distribution function F(x). Also let $X_{i(1)}$, $X_{i(2)}$, …, $X_{i(m)}$ be the corresponding order statistics of X_{i1}, X_{i2}, …, X_{im} (for all i=1,2,…,m). Then $X_{1(1)}$, $X_{2(2)}$, …, $X_{i(i)}$, …, $X_{m(m)}$ is the ranked set sample, since $X_{i(i)}$ is the i^{th} order statistics in the i^{th} sample.

The values of X_{ij} for randomly drawn units can be arranged in the following diagram:

Set				
1	X_{11}	X_{12}	…	X_{1m}
2	X_{21}	X_{22}	…	X_{2m}
.				
.				
.				
m	X_{m1}	X_{m2}	…	X_{mm}

After ranking the units appear as shown below:

Set	Order statistics			
1	$X_{1(1)}$	$X_{1(2)}$	…	$X_{1(m)}$
2	$X_{2(1)}$	$X_{2(2)}$	…	$X_{2(m)}$
.				
.				
.				
m	$X_{m(1)}$	$X_{m(2)}$	…	$X_{m(m)}$

The quantified units are presented as given below:

Set				
1	$X_{1(1)}$	*	…	*
2	*	$X_{2(2)}$	…	*
.				
.				
m	*	*	…	$X_{m(m)}$

The mean of ranked set sample is denoted by $\bar{X}_{(m)}$ where,

$$\bar{X}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{i(i)}$$

For convenience, $X_{i(i)}$ can also be written as $X_{(i:m)}$ which denotes the i:mth order statistics from the population, and the parenthesis are used to surround the subscript to show that $X_{(i:m)}$ are independent unlike the usual i:mth order statistics denoted by $X_{i:m}$ which are positively correlated.

Now,

$$\bar{X}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(i:m)}$$

so that,

$$E[\bar{X}_{(m)}] = \frac{1}{m}\sum_{i=1}^{m} E[X_{(i:m)}] = \mu$$

This shows that $\bar{X}_{(m)}$ is an unbiased estimator of population mean, μ.

This estimator can be compared with the sample mean based on m iid quantifications based on usual order statistics. The latter can be written as,

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_{i:m}$$

where $X_{i:m}$ are the order statistics of the m quantifications. Since, as pointed out, the $X_{i:m}$ are positively correlated, it follows that $\bar{X}_{(m)}$ is more efficient than $\bar{X}$ for estimating μ. In essence, the RSS quantifications $X_{(i:m)}$ are more regularly spaced with less clustering than is simple random sample of size m.

When the whole process of drawing random sample is repeated r times, the i^{th} order statistics from i^{th} sample in j^{th} cycle will be denoted by $X_{(i:m)j}$, i=1,2,…,m and j=1,2,…,r. Here, these are not iid in general, but for a given value of i these are so with $E[X_{(i:m)j}]=\mu_{(i:m)}$ and $V[X_{(i:m)j}]=\sigma^2_{(i:m)}$ in the absence of ranking error. The estimator, $\hat{\mu}_{RSS}$, of population mean, μ, is defined as follows:

$$\hat{\mu}_{RSS} = \bar{X}_{(m)r} = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} X_{(i:m)j} \qquad \text{.....(1)}$$

Also if, $\hat{\mu}_{(i:m)} = \frac{1}{r}\sum_{j=1}^{r} X_{(i:m)j}$ then,

$$\hat{\mu}_{RSS} = \bar{X}_{(m)r} = \frac{1}{m}\sum_{i=1}^{m} \hat{\mu}_{(i:m)} .$$

Now,

$$E[\hat{\mu}_{RSS}] = E[\bar{X}_{(m)r}] = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} E[X_{(i:m)j}] = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} \mu_{(i:m)} = \mu$$

Hence, $\hat{\mu}_{RSS}$ is unbiased estimator of population mean, μ.

The variance of $\hat{\mu}_{RSS}$ is given by,

$$V[\hat{\mu}_{RSS}] = V[\bar{X}_{(m)r}] = \frac{1}{mr}\sum_{i=1}^{m} \frac{\sigma^2_{(i:m)}}{m} \qquad \text{…….(2)}$$

An equivalent expression of variance is given by

$$V[\hat{\mu}_{RSS}]=\frac{1}{mr}[\sigma^2-\frac{1}{m}\sum_{i=1}^{m}\{\mu_{(i:m)}-\mu\}^2] \qquad \ldots\ldots\ldots\ldots(3)$$

where σ^2 denotes the population variance.

Ranked Set Sampling works by creating an "artificially" stratified sample. RSS provides a more precise estimator of population mean than SRS and it is also more cost efficient in a given situation. This is due to the fact that RSS results in a sample in which units are more evenly spaced. Since the units in RSS are more evenly spaced than SRS, the variance of RSS estimates is expected to be less than SRS estimates.

4. Relative Precision of the RSS Estimator of Population Mean Relative to the SRS Estimator and its Estimator

The relative precision, (RP) of RSS estimator, $\hat{\mu}_{RSS}$, as compared with simple random sample (SRS) estimator, $\hat{\mu}_{SRS}$, with same sample size, n, is computed as follows:

$$RP=\frac{V(\hat{\mu}_{SRS})}{V(\hat{\mu}_{RSS})}$$

Here, SRS estimator, $\hat{\mu}_{SRS}$, is based on a random sample of n = mr observations and not a random sample of m^2r observations. This is because the cost of acquiring and ranking samples is not taken into account, but only the cost of quantification is considered. Therefore,

$$V(\hat{\mu}_{SRS})=\sigma^2/mr \qquad \ldots\ldots(4)$$

RP is given by,

$$RP=\frac{V(\hat{\mu}_{SRS})}{V(\hat{\mu}_{RSS})}=\frac{1}{1-\frac{1}{m}\sum_{i=1}^{m}(\frac{\tau_{(i)}}{\sigma})^2} \qquad ..(5)$$

where, $\tau_{(i)}=\mu_{(i:m)}-\mu$.

An equivalent and useful measure of RP are relative cost (RC) and relative savings (RS). These are defined as:

$$RC=1/RP \quad \text{and} \quad RS=1-RC$$

In this context, the relative savings (RS) is given by,

$$RS=\frac{1}{m}\sum_{i=1}^{m}(\frac{\tau_{(i)}}{\sigma})^2 \qquad \ldots\ldots(6)$$

Since this expression is positive, RSS is always more cost efficient than SRS with same number of observations.

McIntyre (1952) and Takahasi and Wakimoto (1968) showed that $1 \le RP \le \frac{m+1}{2}$ and so, $0 \le RS \le \frac{m-1}{m+1}$.

References

Hall, L.S. and Dell, T.R. (1966). Trial of ranked set sampling for forage yield. *Forest Science*, 12:22-26.

Kaur, A., Patil, G.P. and Taillie, C. (1997). Unequal allocation models for ranked set sampling with skew distributions. *Biometrics*, 53: 123-130.

McIntyre, G. (1952). A method of unbiased selective sampling, using rank sets. *Australian Journal of Agricultural Research*, 3:385-390.

Patil, G.P., Sinha, A.K. and Taillie, C. (1994). Ranked set sampling, *Handbook of Statistics,* 12 (G.P. Patil and C.R. Rao eds.):167-198, North Holland, Elsevier Science B.V., Amsterdam.

Sinha, A.K. (2005). On some recent developments in ranked set sampling. *Bulletin of Informatics and Cybernatics*, 37: 137-160.

Stokes, S.L. (1977). Ranked set sampling with concomitant variables. *Communication in Statistics-Theory and methods*, 6:1207-1211.

Takahasi, K. and Wakimoto, K. (1968). On unbiased estimators of the population mean based on the sample stratified by means of ordering. *Annals of Institute of Statistical Mathematics*, 20:1-31.

Multiple Frame Surveys and Application

Bharti

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Statistical tools can be applied to data sets to derive meaningful inferences, which are then utilized for various purposes. For instance, governments rely on these inferences to shape policies aimed at enhancing public welfare, while marketing firms analyze data from consumer surveys to refine their strategies and improve customer services. This data is typically gathered through sample surveys, which are conducted globally by both governmental and non-governmental organizations. In India, for example, the National Sample Survey Organization carries out such surveys. Sampling theory provides essential tools and methods for data collection, ensuring alignment with the intended objectives and the characteristics of the target population. Information can be gathered in two primary ways: through sample surveys or complete enumeration. Sample surveys focus on collecting data from a subset of the population, while a census involves gathering information from the entire population. Certain surveys, such as economic and agricultural surveys, are carried out on a regular basis, providing ongoing insights.

In sample surveys, sampling frame provide access to the elements of finite population of interest. The sampling frame refers to the list of all the units of the population to be surveyed. Each unit in the frame has specific identification details, ensuring that all elements can be accurately tracked and sampled. For instance, in a household-based survey, the sampling frame would include all households, with details such as the head of the household's name or the house address to ensure proper identification. In a study of crop yield, the sampling frame might consist of a list of all commercial farms in a region that grow a specific crop, with details such as farm location, farm size, and the crop being cultivated. This frame would allow researchers to survey a representative sample of farms, helping to obtain insights into the broader agricultural landscape in that region. In forestry, the sampling frame could be constructed by listing all the farms or forestry units in a specific region or a frame that lists all managed forest plots, identified by their plot number or geographic location, to study forest health, biodiversity, or timber production.

2. Sampling Frames

Sampling frames can be broadly categorized into two main types: list frames and area frames. List frame is the exhaustive list of units in the survey population (e.g. a list of all agricultural holdings, a list of farm operators involved in agricultural activities). On the other hand, an area frame is set of geographical unit which may be either points, transects or segments of land. For examples, segments with physical boundaries: a river, a sequence of mountain peaks, etc. Both types of frames are essential for ensuring a structured and effective sampling process in various surveys.

2.1 List Frame: A list frame is a comprehensive enumeration of individuals, households, institutions, or other units within a population that can be sampled. In agricultural statistics,

list frames consist of lists of farms and/or households, which are typically derived from agricultural or population censuses and/or administrative data. The ultimate sampling units are lists of names of holders or agricultural households (Global Strategy, 2015). Typically, the sampling unit from the list frame is a name of a farm operator, while the reporting unit is the holding operated by the name.

Advantages of List Frame:

- ✓ Easy to use
- ✓ Enable in-depth analysis of alternative sampling designs
- ✓ Typically more cost-effective than constructing area frames
- ✓ A key advantage of list frames is the availability of ancillary information for improving sampling designs and estimators

Disadvantages of List Frame:

- ✓ The relationship between frame units and target population units, as well as issues related to multiplicity, and their impact on the inferences drawn
- ✓ Imperfections in the list frame, such as under coverage or over coverage
- ✓ The necessity of maintaining and regularly updating the list frame

2.2 Area Frame: An area frame is a set of land elements, which may be either points or segments of land, that geographically cover a target population (e.g. agricultural land)**.** The sampling process can occur in one or more stages, involving the selection of land segments or points. Information is then gathered directly from these land elements through observations or measurements, as well as details about farming activities associated with the land, typically collected via interviews with the landholders.

Advantages of Area Frame:

- ✓ Ensures complete coverage of the target population
- ✓ Remains stable over extended periods with minimal maintenance costs
- ✓ Enables the adoption of efficient sampling designs based on the survey variables and type of area frame (AF)
- ✓ The ability to collect data through direct observation significantly reduces biases associated with the reliability of farmers' responses regarding cultivated areas or yields.
- ✓ When the same points or segments are surveyed annually, area frames facilitate monitoring of land conditions and inventorying of natural resources.
- ✓ Technological tools, such as aerial photographs, remote sensing, satellite images, GPS, and GIS, can enhance the creation and implementation of area frames in surveys.

Disadvantages of Area Frame:

- ✓ The initial cost of constructing an area frame can be high.

- ✓ Area sampling frames are often less effective for items not closely related to cultivated land use, such as specialty or rare crops. Sampling errors, compared to list frames, may be higher for rare items.
- ✓ The method is also less representative for small areas and crops typically grown on small farms, such as tobacco, vegetables, orchards, and vineyards.
- ✓ Poor road infrastructure and access limitations, making it hard to reach certain segments.
- ✓ Challenges in locating respondents and confirming the existence of households or farm headquarters.
- ✓ Practical issues like farmers living far from their holdings and determining whether a farm should be included in a segment when segments are close to one another.
- ✓ Variables such as livestock or large farms may present additional challenges when using area frames. Surveying farms with livestock that use common pastures, particularly for nomadic livestock, is difficult.
- ✓ Linking selected points to specific farmers can be problematic.

3. Multiple Frame Surveys

Sampling frame is a device which is used to obtain observational access to the elements of finite population of interest. In most of the surveys, it is assumed that sampling frame is complete and up to date, but in reality, sometimes, it is difficult for a single sampling frame to include the entire population of interest and also it is expensive. As a result, multiple frame surveys are becoming more common. In multiple frame surveys population parameters are estimated by using more than one frame which together covers the entire population (Hartley, 1962). Independent samples are selected respectively from each of the frames and information about the target population is gathered based on the combined sample. Hansen *et al.* (1953) first described about dual frame surveys. There are two major motivations behind the use of multiple frame sampling: first is to achieve a desired level of accuracy with reduced cost and second is to have a better coverage of the target population and hence simultaneously reduce the bias occurring due to coverage errors. In some situations, sampling frame may be complete, but sampling using it is quite expensive and on other hand, sampling other frames may be less expensive. For example, suppose in an agricultural survey on wheat in Haryana state, an area frame (e.g. satellite image based frame) may include all of the wheat-growing areas but selecting samples from this frame will increase the cost and complexity because there is already one existing system (Timely Reporting Scheme for enumeration of crop area) for collecting samples based on the list frame. A better option would be to combine random samples taken from both list frame and area frame for estimation of the cropped area with a higher precision (Das *et al.,* 2013). Therefore, it is more cost effective to select a sample of reduced size from the costly complete frame and supplement the sample by additional data taken from other cheaper frames. Multiple frame sampling methods in many agricultural surveys in different countries, combine area frame consisting of segments of land with identifiable physical boundaries that completely covers the entire population and a list frame consisting of the names and address of agricultural holdings which may not be complete. Even though the area frame is complete, but the cost

for building an area frame is high and the cost to reach a reporting unit is also greater than the cost associated to a list frame. Hartley (1962) first derived the basic theory for utilizing two frames in the estimation of population parameter. Later Saxena *et al.* (1984) proposed the estimator for population total for multiple frame surveys under two stage sampling design using domain estimation considering the frames as independent domains. Das *et al.* (2013) proposed a generalized estimator of the population mean under multistage sampling design framework for estimating the average yield of wheat in state of Haryana, India by using list frame of Crop Cutting Experiments (CCE) data collected under General Crop Estimation Surveys (GCES) and area frame from wide field sensor and linear imaging self scanner (LISS-III) data from the Indian Remote Sensing satellite. They proposed a Horvitz Thompson estimator of population mean under two stage sampling design.

Advantages of Multiple Frame Surveys:

- ✓ Combines the advantages of both area frames and list frames, while minimizing their limitations.
- ✓ Allows the easy and not expensive creation of lists of agricultural holdings only in the selected areas, instead of making it in the entire country
- ✓ Data collection becomes more affordable as sample units are concentrated in specific areas
- ✓ Variability can be controlled and measured effectively.
- ✓ Enables the study of specialty or rare products.

Disadvantages of Multiple Frame Surveys:

- ✓ Every holding in the population must appear in at least one frame.
- ✓ The overlap of sampling units between frames must be clearly identified to prevent duplication, as incorrect overlap could introduce bias in the estimation.
- ✓ Both list and area frames should be updated separately.
- ✓ The formulas used for estimation can be complex.

4. Summary:

This chapter explores the concept of multiple frame surveys, which combine area and list frames to optimize sampling in surveys. It explains the importance of sampling frames in obtaining data for various purposes, such as policy-making or marketing strategies. List frames involve comprehensive lists of population units, like households or farms, and are advantageous for their ease of use and cost-effectiveness. However, they have limitations like under coverage or over coverage. Area frames, on the other hand, consist of geographical units like land segments, providing full population coverage and reducing bias through direct observation. Yet, they come with high initial costs and challenges in accessing certain areas. Multiple frame surveys address these issues by integrating both frame types, thus improving cost-efficiency and reducing bias while providing more accurate data coverage. Though they offer several benefits, such as improved variability control and enabling the study of rare

products, they require careful management of overlaps between frames and complex estimation formulas.

References

Das, S.K. and Singh, R. (2013). A multiple frame approach to crop yield estimation from satellite-remotely sensed data. *International Journal of Remote Sensing*, 34(11): 3803-3819.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. Volume1. Wiley, New York.

Hartlay, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association,* 203-206.

Hartlay, H.O. (1974). Multiple frame methodology and selected applications. *Sankhya C*, 36: 99-118.

Saxena, B.C., Narain, P. and Srivastava, A.K. (1984). Multiple frame surveys in two stage sampling. *Sankhya Series B,* 46(1): 75-82.

SMALL AREA ESTIMATION - AN OVERVIEW

Pradip Basak

Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal

1. Introduction

The concept of small area is pretty old although the name is somewhat new. Small areas and domains are synonymous. A part of the population is called a domain. Domains can be local areas often geographic areas for which separate estimates are planned. According to Purcell and Kish (1979, 1980), for major domains (0.1 of the population or more) reasonable sample-based estimates can be obtained with standard methods, from probability samples. Minor domains comprise, say, less than 0.1 or even 0.01 of the population, hence separate estimates may be imprecise, but these days they are increasingly computed with 'Small domain estimates. Mini domains range from 0.01 to 0.0001 of the population and for them censuses have been the traditional sources. But these days improved small area estimation techniques are used for building estimate for such domains. For building estimates for rare items, comprising less than 0.0001 of the population, sample surveys are useless requiring separate and distinct methods. This classification of domains provides a fairly good idea about the smallness of the 'small area'.

Small domain or area refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Examples of domains include a geographical region (e.g. a municipality, a census division, block, tehsil, gram panchayat etc.), a demographic group (e.g. age x sex), a demographic group within a geographic region. 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 considerable increase in the overall sample size which might exceed an already constrained budget and which could further lengthen the data processing time. The decision-making process is more effective when granular or disaggregated data are available because they may be utilised as the basis for developing policies, identifying suitable population groups for policy targets, and keeping track of a programme that has already been put into action.

2. Small Area Estimators

A survey population U consists of N distinct elements (or ultimate units) identified through the labels $j = 1, 2, \ldots, N$. A sample s is selected from U with probability $p(s)$, and the probability of including the j^{th} element in the sample is π_j. The design weight for each selected unit $j \in s$ is defined as $w_j=1/\pi_j$. Suppose U_i denotes a domain (or subpopulation) of interest and $s_i=s\cap U_i$ denotes the part of the sample s that falls in domain U_i, $\forall\ i=1, 2, \ldots, m$. The realized sample size of s_i is a random variable n_i , where $0\leq n_i \leq N_i$. Auxiliary data x will either be known at the element level x_j for $j \in s$ or for each small area i as totals $X_i=\sum_{j\epsilon U_i} x_j$ or means $\bar{X}_i=\frac{X_i}{N_i}$. Here, the problem is to estimate the domain total $Y_i=\sum_{j\epsilon U_i} y_j$ or means $\bar{Y}_i=\frac{Y_i}{N_i}$, where N_i, the number of elements in U_i , may or may not be known. Let us define y_{ij} to be y_j if j$\epsilon U_i$and 0 otherwise. An indicator variable a_{ij} is similarly defined it is equal to one if j$\epsilon U_i$and 0 otherwise. The domain total Y_i can then be written as

$Y_i=\sum_{j\epsilon U} y_{ij}=\sum_{j\epsilon U} y_j\, a_{ij}$. Small area estimation is categorized into two types of estimators: direct and indirect estimators.

2.1 Direct estimator

A direct estimator is one that uses values of the variable of interest, *y*, only from the sample units in the domain of interest.

Suppose a linear estimator based on sample weights $\{w_j; j\epsilon s\}$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 $U = \{1, \ldots., N\}$ of size *N*. Further, if $\pi_j = \sum_{j\epsilon s} p(s)$ 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 $w_j = Nn^{-1}$. Let us 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 such that $s = \bigcup_{i=1}^{m} s_i$ and $n = \sum_{i=1}^{m} n_i$. It may be noted that n_i is a random variable. Let y_j denotes the value of characteristic of interest y for j^{th} population unit in small area i. The population mean of y in the small area i is given by, $\hat{\bar{Y}}_i = N_i^{-1}\sum_{j\in U_i} y_j$. If the population size N_i of small areas i is unknown then the population mean of y in the small area i could be estimated using Hajek type estimator,

$$\hat{\bar{Y}}_i^{Hajek} = \left(\sum_{j\epsilon s_i} w_j\right)^{-1}\left(\sum_{j\epsilon s_i} w_j y_j\right),$$

or, if the population size N_i of the small areas i is known it leads to Horvitz-Thompson estimator,

$$\hat{\bar{Y}}_i^{HT} = N_i{}^{-1}\left(\sum_{j\epsilon s_i} w_j y_j\right).$$

The disadvantage of direct estimator is that irrespective of the form of direct estimator being used, it is easy to see that its variance can be large when the sample size n_i in i^{th} area is small.

Example:

Simple random sampling, with no auxiliary information, a direct estimator of the population mean of y, $\hat{\bar{Y}}_i = N_i^{-1}\sum_{j\in U_i} y_j$ for small area i is given by,

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

where, $\bar{y}_i = \frac{\sum_{s_i} w_j y_j}{\sum_{s_i} w_j} = \frac{\sum_{j\epsilon s_i} y_j}{n_i}$ is the sample mean of y in i^{th} small area and its variance is given by,

$$Var_p\left(\hat{\bar{Y}}_i\right) = \frac{(1-f_i)S_i^2}{n_i},$$

with $f_i = {}^{n_i}/_{N_i}$ and $S_i^2 = \left(N_i - 1\right)^{-1}\sum_{j=1}^{N_i}\left(y_j - \bar{Y}_i\right)^2$, $N_i \geq 2$. Here Var_p denotes the variance under the design-based approach. An unbiased estimator of S_i^2 is given by, $s_i^2 = \left(n_i - 1\right)^{-1}\sum_{j=1}^{n_i}\left(y_j - \bar{y}_i\right)^2$. Thus, an unbiased estimator for variance is given by

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

For unknown N_i, $f_i = {}^{n_i}/_{N_i}$ is replaced by $f = {}^{n}/_{N}$ and then estimator for variance is given by

$$v\left(\hat{\bar{Y}}_i\right) = \frac{(1-f)s_i^2}{n_i}.$$

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 in addition to survey variable *y*, values of *p*-auxiliary variables are also known. Consider $\mathbf{x}_{ij}$ is a $pX1$ vector of auxiliary variable **x** for the j^{th} unit in i^{th} small area. Then with known auxiliary information, a more efficient design based direct estimator of $\bar{Y}_i$ is the regression estimator defined as

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

where, $\hat{\boldsymbol{\beta}}_i = \dfrac{\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(y_{ij} - \bar{y}_i)}{\left(\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'\right)^{-1}}$ is the vector of estimated regression coefficients in i^{th} small area, $\bar{x}_i = n_i^{-1}\sum_{j\in s_i} x_j$ and $\bar{X}_i = N_i^{-1}\sum_{j\in U_i} x_j$ are the vectors of sample mean and population mean of *p* auxiliary variable in i^{th} small area respectively. The variance of the regression estimator is given by

$$Var_p\left(\hat{\bar{Y}}_i^{REG}\right) \approx \frac{(1-f_i)}{n_i} S_i^2(1-\rho_i^2) = (1-\rho_i^2)Var_p\left(\hat{\bar{Y}}_i\right)$$

where ρ_i is the multiple correlation between survey variable *y* and auxiliary variable vector **x** in area i.

The estimate of variance is provided by

$$v\left(\hat{\bar{Y}}_i^{REG} \mid n_i\right) = \frac{(1-f_i)}{n_i} s_i^2(1-\hat{\rho}_i^2).$$

By using auxiliary variables, the variance is reduced by the factor $(1-\rho_i^2)$ indicating that use of a good auxiliary information, in the sense of high correlation with survey variable *y*, increases the accuracy in small area estimation.

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.

2.2.1 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'.

Rao and Choudry (1995) suggested the use of a ratio synthetic estimator. Let us consider availability of a single auxiliary variable x. The ratio synthetic estimator for the population total of y in small area i is $\hat{T}_{y_i}^{SynR} = \hat{R}_i T_{x_i}$. It is assumed that in area i population ratio is $R_i = T_{y_i} / T_{x_i}$,

$T_{y_i} = \sum_{j=1}^{N} y_j$ and $T_{x_i} = \sum_{j=1}^{N} x_j$ 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, T_x are the values for the whole population and R_U is estimated by $\hat{R}_U = \bar{y} / \bar{x}$, where $\bar{y}$ and $\bar{x}$ are the overall sample means of y and x respectively. Here, a subscript of U is being used 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.

Now, in the case of model-based synthetic estimation, let us consider the regression model,

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

where, y_{ij} is the value of variable of interest for the $j^{th}(j = 1, \ldots, n_i)$ unit in the small area $i(1, \ldots, m)$, $\mathbf{x}_{ij}$ is the $pX1$ vector of auxiliary variables, $\boldsymbol{\beta}$ is a $pX1$ vector of regression coefficients and e_{ij} is error term often assumed to be normally distributed with mean zero and variance σ^2.

The regression synthetic estimator for the population mean of study variable y in small area i is defined as,

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

where $\hat{\boldsymbol{\beta}} = \dfrac{\sum_{i=}^{m}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(y_{ij}-\bar{y}_i)}{\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'\right)^{-1}}$ is the full sample estimate, i.e. calculated using data from entire areas and thus it is different from direct regression estimator. For the areas with no sample data, the model-based synthetic estimator for population mean of study variable y in small area i is defined as $\hat{\bar{Y}}_i^{MSyn} = \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}$. This will be very efficient when small area i does not exhibit strong individual effect with respect to the regression coefficient.

2.2.2 Composite estimator

As the sample size in a small area increase, 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

$$\hat{T}_{y_i}^{c} = \emptyset_i \hat{T}_{y_i}^{d} + (1-\emptyset_i)\hat{T}_{y_i}^{syn},$$

where, $\hat{T}_{y_i}^{d}$ and $\hat{T}_{y_i}^{syn}$ are the direct and synthetic estimator of population total of y in i^{th} small area respectively. Here, $\emptyset_i (0 \leq \emptyset_i \leq 1)$ is a suitably chosen weight.

The traditional indirect estimators such as synthetic and composite estimator have the advantage of being simple to implement. These 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 effect 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.

3. Mixed Models 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:

1. The area level mixed effect model (or Area level model) which uses area-specific auxiliary information and
2. 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.

3.1 Area level models

Area level models can be used when individual measurements for auxiliary variables are not available and the auxiliary information is available only at the area level. 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,

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}} = \boldsymbol{\theta}_{\boldsymbol{i}} + \boldsymbol{e}_{\boldsymbol{i}}\,; \boldsymbol{\theta}_{\boldsymbol{i}} = \boldsymbol{x}_{\boldsymbol{i}}'\,\boldsymbol{\beta} + \boldsymbol{u}_{\boldsymbol{i}}$$

where, $\tilde{\theta}_i$ denotes direct sample estimator (ex: sample mean $\bar{y}_i$) and e_i represents sampling error, assumed to have zero mean and known design variance $Var_D(e_i) = \sigma_{Di}^2$ and x_i represent the area level information. Here, θ_i is true population parameter.

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

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\iota}}=\boldsymbol{\gamma}_{\boldsymbol{i}}\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}} + (\mathbf{1}-\boldsymbol{\gamma}_{\boldsymbol{i}})\boldsymbol{x}_{\boldsymbol{i}}'\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = \boldsymbol{x}_{\boldsymbol{i}}'\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} + \boldsymbol{\gamma}_{\boldsymbol{i}}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{i}}'\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}\right),$$

where , $\gamma_i=\sigma_u^2/(\sigma_{Di}^2+\sigma_u^2)$ and

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = \left(\sum_{i=1}^{m}\frac{\bar{X}_i'\bar{X}_i}{\sigma_{Di}^2+\sigma_u^2}\right)^{-1}\left(\sum_{i=1}^{m}\frac{\bar{X}_i'\tilde{\theta}_i}{\sigma_{Di}^2+\sigma_u^2}\right).$$

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).

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\iota}}^{\,*}=\hat{\boldsymbol{\gamma}}_{\boldsymbol{\iota}}\,\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}} + (\mathbf{1}-\hat{\boldsymbol{\gamma}}_{\boldsymbol{i}})\boldsymbol{x}_{\boldsymbol{i}}'\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}^{*} = \boldsymbol{x}_{\boldsymbol{i}}'\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}^{*} + \hat{\boldsymbol{\gamma}}_{\boldsymbol{\iota}}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{i}}'\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}^{*}\right)$$

where, $\hat{\gamma}_\iota = \hat{\sigma}_u^2/(\hat{\sigma}_{Di}^2+\hat{\sigma}_u^2)$ and

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}^{*} = \left(\sum_{i=1}^{m}\frac{\bar{X}_i'\bar{X}_i}{\hat{\sigma}_{Di}^2+\hat{\sigma}_u^2}\right)^{-1}\left(\sum_{i=1}^{m}\frac{\bar{X}_i'\tilde{\theta}_i}{\hat{\sigma}_{Di}^2+\hat{\sigma}_u^2}\right).$$

3.2 Unit level models

A basic unit-level model assumes that the unit *y*-values, y_{ij} , associated with the j^{th} population unit $(j = 1, \dots, N_i)$ in the i^{th} area are related to unit-level covariates, y_{ij} for which the population mean vector $\bar{X}_i$ is known. If y is a continuous response (e.g. crop yield), we assume a one-fold nested error linear regression model

$$y_{ij} = \boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{\beta} + u_i + e_{ij}, j = 1, \dots, N_i; i = 1, \dots, m$$

where the random sample area effects u_i have mean 0 and common variance σ_u^2 and are independently distributed. Further, the u_i are independent of the residual errors e_{ij} which are assumed to be independently distributed with mean 0 and common variance σ_e^2 (Battese et al., 1988). If N_i is large, the population mean $\bar{Y}_i$ is approximately equal to $X_{ij}^T\beta + u_i$. The sample data $\{y_{ij}, x_{ij},\ j = 1, \dots, n_i; i = 1, \dots, m\}$ are assumed to obey the population model. This implies that sample selection bias is absent, which is satisfied by simple random sampling within areas. For more general sampling designs, the sample data will satisfy the assumption if the selection probabilities, p_{ij} depend only on the auxiliary variables in x_{ij}; for example, for probability proportional to size (PPS) sample, where size is used as an auxiliary variable in model. Non-probability samples obeying above model can also be used to estimate the mean $\bar{Y}_i$.

In vector form the model can be expressed as,

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

where, $\mathbf{y}_i = \left(y_{i1}, y_{i2,} \ldots y_{in_i}\right)'$, $\mathbf{X}_i = \left(x_{i1}, x_{i2,} \ldots x_{in_i}\right)'$ a $n_i Xp$ matrix and $\mathbf{e}_i = \left(e_{i1}, e_{i2,} \ldots e_{in_i}\right)'$. The variance-covariance matrix of $\mathbf{y}_i$ is $Var(\mathbf{y}_i) = \mathbf{V}_i = \sigma_e^2\mathbf{I}_{n_i} + \sigma_u^2\mathbf{1}_{n_i}\mathbf{1}'_{n_i}$.

Population mean of y in area i is $\bar{Y}_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + u_i + \bar{e}_i$, where $\bar{\mathbf{X}}_i^T = N_i^{-1}\sum_{j=1}^{N_i}\mathbf{x}_j$ is assumed to be known. For sufficiently large N_i, $\bar{e}_i \approx 0$ then mean of y in small area i is approximated by, $\mu_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + u_i$.

Example: Assume that $\hat{\bar{Y}}_{i,DIR} = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + v_i$, where $v_i = u_i + e_i$. Then the EBLUP estimate of $\bar{Y}_i$ is a composite estimate of the form

$$\hat{\bar{Y}}_{i,EBLUP} = \hat{\gamma}_i[\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T\boldsymbol{\beta}] + (1 - \hat{\gamma}_i)\bar{\mathbf{X}}_i^T\boldsymbol{\beta},\ \forall\ i = 1, \ldots, m \text{ or}$$

$$\hat{\bar{Y}}_{i,EBLUP} = \bar{\boldsymbol{X}}_i'\tilde{\boldsymbol{\beta}} + \gamma_i\left(\bar{y}_i - \bar{\boldsymbol{x}}_i'\tilde{\boldsymbol{\beta}}\right)$$

where, $\tilde{\boldsymbol{\beta}} = \frac{\sum_i \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{y}_i}{\sum_i \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i}$ is the BLUE of $\boldsymbol{\beta}$ and $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_i}$.

4. Applications

- Small area estimation techniques can be applied in Agriculture and allied sector, For example
  a. Estimation of crop yield at GP, tehsil level and Block level
  b. Estimation of Post harvest loses at district level
  c. Estimation of milk production at district level
  d. To estimate food insecurity at district level.
- Small area estimation techniques can also be applied in NSSO data
  a. To obtain district level estimates of amount of loan outstanding per household.
  b. To obtain district level poverty estimates
  c. To estimate income and unemployment rate at district level

5. Case Study

5.1 Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India- an application of small area estimation technique (Chandra *et. al.* 2011)

Objective: Obtaining district wise estimates of proportion of indebted farm households for different land-holding classes in the State of Uttar Pradesh.

Data: The variable of interest for which small area estimates are required is drawn from the Debt-Investment Survey 2002-03 of NSSO**.** The auxiliary (covariates) variables known for the population are drawn from the Population Census 2001 and the Agriculture Census 2003.

Methodology: The value of variable of interest y (which is the number of indebted household) in the area d is defined by y_d, y_{sd} *and* y_{rd} denotes sample and non-sample counts of indebted households in the area *d* respectively, and x_d denote the *k*-vector of the covariates for area d from the previous sources. The model linking the probabilities of success with the covariates is the logistic linear mixed model given as,

$$logit(\pi_d) = \mathbf{x}_d'\boldsymbol{\beta} + u_d,\ \text{d=1,2,…,D}$$

It is a special case of a generalized linear mixed model (GLMM) with logit link function and suitable for discrete, particularly binary variable.

Results: The direct estimates and model-based estimates of proportion of indebted farm households for the state of Uttar Pradesh for five different land-holding classes (Cat 0 = All land holdings, Cat 1= Marginal, Cat 2= Small, Cat 3= Semi medium, Cat 4= Medium, Cat 5= Large) is generated using SAE methodology and are presented in the following tables.

Table 1: Direct estimates and model-based estimates of proportion of indebted farm households for Cat 0

District	**Direct**	**Model based**	**District**	**Direct**	**Model based**	**District**	**Direct**	**Model based**
Saharanpur	0.6	0.6	Etawah	0.51	0.52	Ambedker Nr	0.48	0.5
Muzaffarnagar	0.61	0.6	Auraya	0.5	0.51	Sultanpur	0.51	0.5
Bijnor	0.55	0.56	Kheri	0.58	0.58	Bahraich	0.51	0.52
Moradabad	0.56	0.57	Sitapur	0.56	0.56	Srawasti	0.51	0.52
Rampur	0.55	0.57	Hardoi	0.46	0.5	S.Kabir Nr	0.61	0.54
J.B.P.Nr	0.52	0.52	Unnao	0.59	0.57	Kushi Nagar	0.58	0.55
Meerut	0.59	0.58	Lucknow	0.51	0.52	Balrampur	0.46	0.48
Baghpat	0.58	0.55	Raibarely	0.51	0.52	Gonda	0.54	0.53
Ghaziabad	0.55	0.55	Kanpur Dehat	0.47	0.5	Sidharth Nr	0.43	0.47
Bulad Shahar	0.56	0.56	Kanpur Nr	0.52	0.53	Basti	0.58	0.55
Aligarh	0.59	0.6	Fatehpur	0.59	0.54	Maharajganj	0.47	0.49
Mathura	0.55	0.56	Jalaun	0.63	0.58	Gorakhpur	0.52	0.52
Hathras	0.68	0.57	Jhanshi	0.55	0.54	Deoria	0.54	0.54
Agra	0.63	0.58	Lalitpur	0.71	0.57	Azamgarh	0.47	0.49
Firozabad	0.52	0.54	Hamirpur	0.52	0.52	Mau	0.58	0.54
Etah	0.52	0.55	Mahoba	0.48	0.52	Ballia	0.56	0.53
Farukhabad	0.52	0.54	Banda	0.55	0.53	Jaunpur	0.47	0.48
Mainpuri	0.57	0.55	Chitrakut	0.61	0.54	Ghazipur	0.43	0.47
Badaun	0.55	0.56	Pratapgarh	0.44	0.47	Chandauli	0.4	0.47
Bareilly	0.63	0.61	Kaushambi	0.57	0.53	Varanasi	0.51	0.51
Pilibhit	0.57	0.56	Allahabad	0.57	0.54	St. Ravidas Nr	0.48	0.51
Shahjahanpur	0.68	0.62	Barabanki	0.57	0.56	Mizapur	0.44	0.47
Kannauj	0.5	0.53	Faizabad	0.48	0.5	Shanbhadra	0.36	0.44

Table 2: Direct estimates and model-based estimates of proportion of indebted farm households for Cat 1

District	Direct	Model based	District	Direct	Model based	District	Direct	Model based
Saharanpur	0.59	0.58	Etawah	0.49	0.49	Ambedker Nr	0.47	0.48
Muzaffarnagar	0.58	0.57	Auraya	0.48	0.48	Sultanpur	0.47	0.47
Bijnor	0.51	0.52	Kheri	0.56	0.56	Bahraich	0.49	0.49
Moradabad	0.51	0.53	Sitapur	0.55	0.54	Srawasti	0.46	0.48
Rampur	0.52	0.53	Hardoi	0.43	0.46	S.Kabir Nr	0.6	0.52
J.B.P.Nr	0.47	0.48	Unnao	0.59	0.55	Kushi Nagar	0.56	0.53
Meerut	0.56	0.55	Lucknow	0.46	0.48	Balrampur	0.43	0.45
Baghpat	0.62	0.54	Raibarely	0.45	0.48	Gonda	0.56	0.54
Ghaziabad	0.5	0.51	Kanpur Dehat	0.43	0.46	Sidharth Nr	0.38	0.43
Bulad Shahar	0.56	0.55	Kanpur Nr	0.5	0.5	Basti	0.52	0.51
Aligarh	0.53	0.55	Fatehpur	0.58	0.53	Maharajganj	0.45	0.47
Mathura	0.48	0.51	Jalaun	0.53	0.53	Gorakhpur	0.52	0.51
Hathras	0.67	0.54	Jhanshi	0.46	0.49	Deoria	0.55	0.53
Agra	0.61	0.56	Lalitpur	0.63	0.51	Azamgarh	0.45	0.46
Firozabad	0.47	0.49	Hamirpur	0.53	0.5	Mau	0.52	0.5
Etah	0.5	0.52	Mahoba	0.28	0.46	Ballia	0.57	0.54
Farukhabad	0.43	0.48	Banda	0.52	0.5	Jaunpur	0.44	0.45
Mainpuri	0.49	0.49	Chitrakut	0.56	0.51	Ghazipur	0.43	0.45
Badaun	0.53	0.53	Pratapgarh	0.36	0.42	Chandauli	0.3	0.41
Bareilly	0.61	0.58	Kaushambi	0.59	0.53	Varanasi	0.5	0.49
Pilibhit	0.56	0.53	Allahabad	0.6	0.56	St. Ravidas Nr	0.51	0.49
Shahjahanpur	0.65	0.59	Barabanki	0.52	0.52	Mizapur	0.46	0.47
Kannauj	0.46	0.49	Faizabad	0.43	0.47	Shanbhadra	0.25	0.38

Table 3: Direct estimates and model-based estimates of proportion of indebted farm households for Cat 2

District	Direct	Model based	District	Direct	Model based	District	Direct	Model based
Saharanpur	0.5	0.65	Etawah	0.48	0.6	Ambedker Nr	0.41	0.55
Muzaffarnagar	0.67	0.63	Auraya	0.52	0.59	Sultanpur	0.55	0.51
Bijnor	0.64	0.63	Kheri	0.59	0.63	Bahraich	0.45	0.57
Moradabad	0.72	0.64	Sitapur	0.5	0.61	Srawasti	0.56	0.58
Rampur	0.69	0.64	Hardoi	0.47	0.59	S.Kabir Nr	0.7	0.56
J.B.P.Nr	0.57	0.59	Unnao	0.58	0.59	Kushi Nagar	0.58	0.5
Meerut	0.73	0.64	Lucknow	0.63	0.58	Balrampur	0.5	0.52
Baghpat	0.55	0.61	Raibarely	0.5	0.55	Gonda	0.55	0.57

Ghaziabad	0.68	0.62	Kanpur Dehat	0.55	0.6	Sidharth Nr	0.44	0.53
Bulad Shahar	0.53	0.6	Kanpur Nr	0.58	0.58	Basti	0.62	0.56
Aligarh	0.73	0.66	Fatehpur	0.69	0.52	Maharajganj	0.62	0.53
Mathura	0.66	0.63	Jalaun	0.5	0.61	Gorakhpur	0.53	0.54
Hathras	0.78	0.61	Jhanshi	0.64	0.59	Deoria	0.58	0.57
Agra	0.71	0.62	Lalitpur	1	0.59	Azamgarh	0.47	0.52
Firozabad	0.63	0.62	Hamirpur	0.63	0.58	Mau	0.74	0.57
Etah	0.63	0.65	Mahoba	0.45	0.58	Ballia	0.57	0.53
Farukhabad	0.75	0.62	Banda	0.59	0.55	Jaunpur	0.65	0.53
Mainpuri	0.72	0.61	Chitrakut	0.8	0.58	Ghazipur	0.43	0.53
Badaun	0.6	0.63	Pratapgarh	0.6	0.53	Chandauli	0.64	0.56
Bareilly	0.71	0.65	Kaushambi	0.54	0.52	Varanasi	0.45	0.56
Pilibhit	0.58	0.62	Allahabad	0.42	0.48	St. Ravidas Nr	0.5	0.59
Shahjahanpur	0.69	0.63	Barabanki	0.65	0.58	Mizapur	0.25	0.51
Kannauj	0.5	0.62	Faizabad	0.5	0.55	Shanbhadra	0.47	0.52

Table 4: Direct estimates and model-based estimates of proportion of indebted farm households for Cat 3

District	**Direct**	**Model based**	**District**	**Direct**	**Model based**	**District**	**Direct**	**Model based**
Saharanpur	0.62	0.66	Etawah	0.64	0.55	Ambedker Nr	0.53	0.54
Muzaffarnagar	0.69	0.66	Auraya	0.55	0.54	Sultanpur	0.63	0.53
Bijnor	0.5	0.61	Kheri	0.69	0.64	Bahraich	0.53	0.56
Moradabad	0.52	0.64	Sitapur	0.57	0.6	Srawasti	0.7	0.55
Rampur	0.5	0.63	Hardoi	0.56	0.56	S.Kabir Nr	0.4	0.53
J.B.P.Nr	0.56	0.54	Unnao	0.6	0.57	Kushi Nagar	0.84	0.53
Meerut	0.56	0.63	Lucknow	0.67	0.55	Balrampur	0.48	0.53
Baghpat	1	0.58	Raibarely	0.69	0.57	Gonda	0.26	0.55
Ghaziabad	0.59	0.6	Kanpur Dehat	0.52	0.57	Sidharth Nr	0.53	0.53
Bulad Shahar	0.55	0.59	Kanpur Nr	0.33	0.56	Basti	0.67	0.55
Aligarh	0.68	0.67	Fatehpur	0.41	0.52	Maharajganj	0.38	0.54
Mathura	0.82	0.63	Jalaun	0.75	0.61	Gorakhpur	0.41	0.55
Hathras	0.5	0.57	Jhanshi	0.52	0.57	Deoria	0.6	0.57
Agra	0.72	0.59	Lalitpur	0.71	0.56	Azamgarh	0.56	0.55
Firozabad	0.53	0.57	Hamirpur	0.29	0.55	Mau	0.67	0.54
Etah	0.56	0.64	Mahoba	0.78	0.56	Ballia	0.48	0.54
Farukhabad	0.44	0.58	Banda	0.47	0.54	Jaunpur	0.52	0.53
Mainpuri	0.86	0.56	Chitrakut	0.8	0.55	Ghazipur	0.48	0.54
Badaun	0.65	0.61	Pratapgarh	0.63	0.54	Chandauli	0.3	0.53
Bareilly	0.67	0.63	Kaushambi	0.44	0.52	Varanasi	0.46	0.53

Pilibhit	0.65	0.59	Allahabad	0.54	0.53	St. Ravidas Nr	0.33	0.54
Shahjahanpur	0.75	0.61	Barabanki	0.68	0.59	Mizapur	0.29	0.52
Kannauj	0.5	0.58	Faizabad	0.57	0.55	Shanbhadra	0.58	0.52

Table 5: Direct estimates and model-based estimates of proportion of indebted farm households for Cat 4

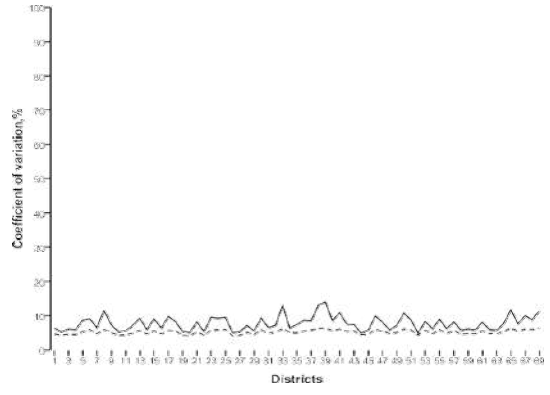
District	**Direct**	**Model based**	**District**	**Direct**	**Model based**	**District**	**Direct**	**Model based**
Saharanpur	0.83	0.69	Etawah	0.5	0.62	Ambedker Nr	0.6	0.63
Muzaffarnagar	0.71	0.68	Auraya	0.67	0.63	Sultanpur	0.69	0.63
Bijnor	0.6	0.65	Kheri	0.6	0.66	Bahraich	0.82	0.65
Moradabad	0.67	0.67	Sitapur	0.75	0.66	Srawasti	0.75	0.63
Rampur	0.5	0.66	Hardoi	0.55	0.63	S.Kabir Nr	0.75	0.62
J.B.P.Nr	0.75	0.63	Unnao	0.64	0.64	Kushi Nagar	0.54	0.61
Meerut	0.62	0.66	Lucknow	0	0.62	Balrampur	0.6	0.62
Baghpat	0.2	0.62	Raibarely	0.86	0.66	Gonda	0.63	0.63
Ghaziabad	0.75	0.65	Kanpur Dehat	0.83	0.65	Sidharth Nr	0.63	0.62
Bulad Shahar	0.63	0.65	Kanpur Nr	1	0.64	Basti	0.89	0.65
Aligarh	0.73	0.69	Fatehpur	0.67	0.62	Maharajganj	0.5	0.62
Mathura	0.55	0.65	Jalaun	0.93	0.68	Gorakhpur	0.6	0.63
Hathras	0.8	0.64	Jhanshi	0.78	0.65	Deoria	0.31	0.61
Agra	0.5	0.63	Lalitpur	0.57	0.63	Azamgarh	0.56	0.62
Firozabad	1	0.65	Hamirpur	0.82	0.65	Mau	0.67	0.63
Etah	0.33	0.65	Mahoba	0.46	0.61	Ballia	0.5	0.61
Farukhabad	0.83	0.65	Banda	0.77	0.64	Jaunpur	0.33	0.59
Mainpuri	0.63	0.63	Chitrakut	0.33	0.62	Ghazipur	0.5	0.62
Badaun	0.46	0.64	Pratapgarh	0.58	0.62	Chandauli	0.71	0.63
Bareilly	0.43	0.65	Kaushambi	0.67	0.62	Varanasi	1	0.63
Pilibhit	0.57	0.64	Allahabad	0.67	0.63	St. Ravidas Nr	0.33	0.62
Shahjahanpur	1	0.67	Barabanki	0.9	0.67	Mizapur	0.63	0.62
Kannauj	1	0.66	Faizabad	0.71	0.63	Shanbhadra	0.54	0.61

Table 6: Direct estimates and model-based estimates of proportion of indebted farm households for Cat 5

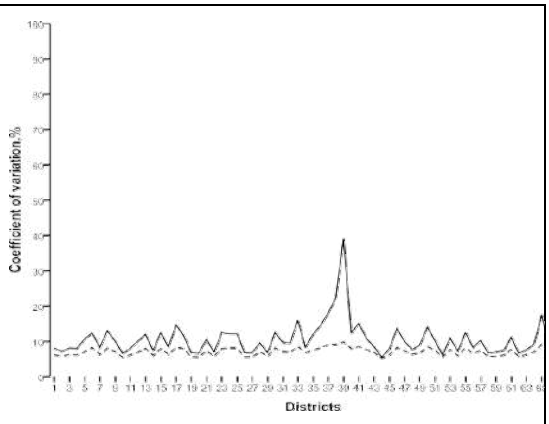
District	**Direct**	**Model based**	**District**	**Direct**	**Model based**	**District**	**Direct**	**Model based**
Saharanpur	0.33	0.62	Etawah	1	0.79	Ambedker Nr	0.8	0.64
Muzaffarnagar	0.75	0.55	Auraya	0.5	0.77	Sultanpur	0.4	0.55
Bijnor	0.75	0.69	Kheri	0.67	0.61	Bahraich	0.5	0.63

Moradabad	1	0.63	Sitapur	0.8	0.68	Srawasti		0.74
Rampur	1	0.7	Hardoi	0.67	0.73	S.Kabir Nr		0.71
J.B.P.Nr	1	0.79	Unnao	1	0.7	Kushi Nagar		0.5
Meerut	1	0.68	Lucknow	1	0.74	Balrampur	1	0.59
Baghpat	1	0.75	Raibarely	0.4	0.55	Gonda	0	0.66
Ghaziabad	0.5	0.73	Kanpur Dehat	0.6	0.73	Sidharth Nr	0.8	0.6
Bulad Shahar	0.63	0.66	Kanpur Nr		0.69	Basti	0.78	0.68
Aligarh	0.25	0.6	Fatehpur	0.75	0.57	Maharajganj	0.5	0.56
Mathura	0	0.65	Jalaun	1	0.68	Gorakhpur	1	0.56
Hathras		0.76	Jhanshi	0.75	0.7	Deoria	0.5	0.64
Agra		0.74	Lalitpur	1	0.76	Azamgarh	0.67	0.53
Firozabad		0.77	Hamirpur	0.25	0.7	Mau	0.67	0.7
Etah		0.67	Mahoba	0.8	0.72	Ballia	1	0.58
Farukhabad		0.77	Banda	1	0.64	Jaunpur	0.44	0.62
Mainpuri	1	0.79	Chitrakut	0.5	0.71	Ghazipur	0.33	0.57
Badaun	1	0.72	Pratapgarh	1	0.59	Chandauli	1	0.7
Bareilly	0.75	0.7	Kaushambi	0.5	0.6	Varanasi	0.5	0.73
Pilibhit	0	0.75	Allahabad	0.33	0.42	St. Ravidas Nr		0.78
Shahjahanpur	1	0.73	Barabanki	0.5	0.59	Mizapur	0.5	0.56
Kannauj	1	0.76	Faizabad	1	0.65	Shanbhadra	0.33	0.61

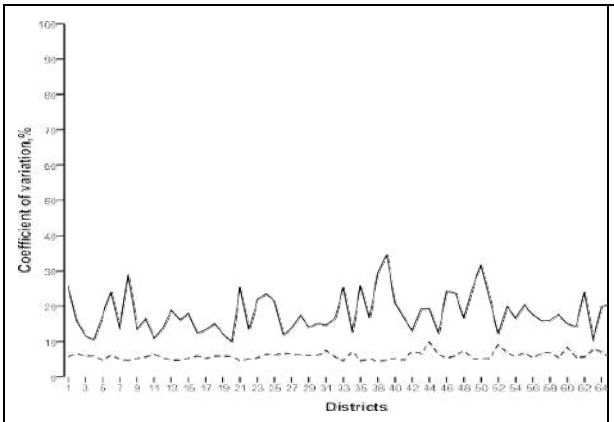
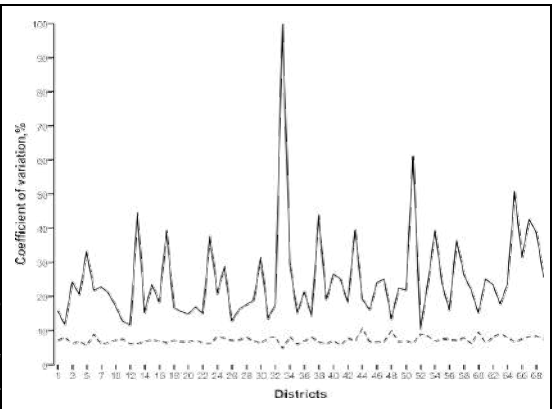
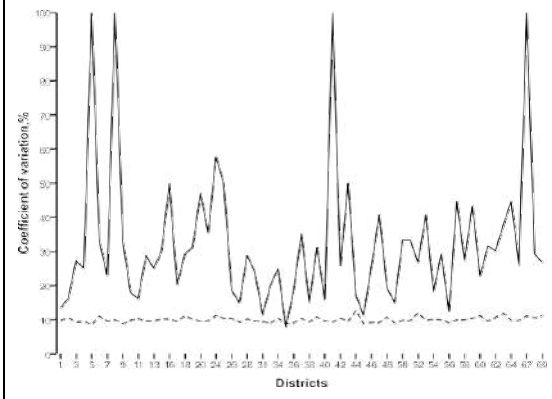
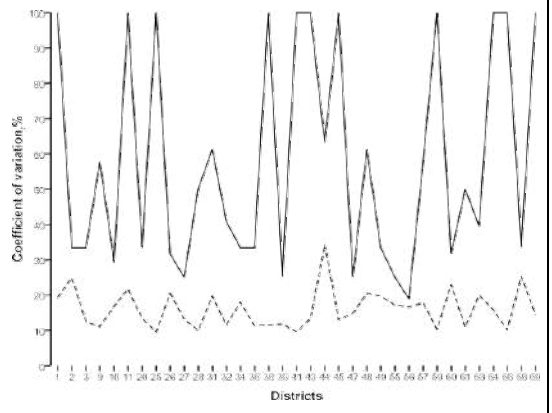
The graphs representing the district-wise coefficient of variation for direct (thick line) and model-based estimate (dotted line) for Cat0- Cat5 are given below:



Cat 0



Cat 1

	
Cat 2	Cat 3
	
Cat 4	Cat 5

Conclusion: The model-based district level estimates for different land holding classes have reasonably good precision as compared to direct estimates. The SAE method has also generated reliable estimates for the districts with zero or very small sample sizes.

5.2 EBLUP estimate of crop yield at sub-district level in Hisar district, Haryana, India using MODIS/Terra data (Muhammed et al., 2020)

Objective: Obtaining EBLUP estimate of crop yield at sub-district level in Hisar district, Haryana, India using MODIS/Terra data.

Data: MODIS/Terra Data Acquisition: MODIS/Terra is a satellite sensor that provides various spectral bands, including visible, near-infrared, and thermal infrared, suitable for monitoring vegetation health and crop conditions. Time-series data from MODIS/Terra were acquired for the study area in Hisar district, Haryana.

Crop Yield Sampling and Ground Truth Data: A systematic random sampling approach was used to collect ground truth data, including crop yield, from selected sub-districts within Hisar district. These data served as reference points for model calibration and validation.

Methodology: Spatial Analysis and Preprocessing: MODIS/Terra data were pre-processed to extract relevant vegetation indices (e.g., NDVI) and environmental variables (e.g., temperature, precipitation). The data were then spatially aligned with the ground truth points using geospatial techniques.

Area level random effect model was used to derive EBLUP estimate of crop yield by considering y_i denote the observed direct estimate of the unobservable population-level

quantity (e.g. average yield) Y_i of variable of interest y for area (or sub district) i. Let X_i be the known auxiliary variable, obtained from NDVI data set, related to the population mean Y_i. The area specific Fay and Herriot model takes into account both the spatial and temporal variations in crop growth conditions to generate accurate predictions. EBLUP incorporates the observed data, covariate information, and residual errors to estimate crop yield at unsampled locations.

Software used: ArcGIS and ENVI software's were used for processing and analysing geospatial imagery and Fay–Herriot model based EBLUP estimator was generated using SAE package in R software interface

Results: The authors generated Direct and EBLUP estimates for rice and wheat crop yield by using selected covariates along with CV. The estimated CVs for model-based estimates are much more precise than direct estimates which are given in below tables.

Rice yield									
Block	**Sample size**	**Direct estimate**	**CV(%)**	**EBLUP Estimate Yi~NDVI**	**CV(%)**	**RMSE**	**EBLUP Estimate Yi~iNDVI**	**CV(%)**	**RMSE**
Adampur	14	3420.91	14.84	3048.09	10.07	307.09	3818.48	9.62	367.38
Agroha	17	2849.72	10.18	3092.3	9	278.18	3018.37	7.76	234.37
Barwala	36	3068.55	8.89	3042.7	8.09	246.22	2834.2	7.8	221.02
Hisar 1	36	2662.48	9.03	2687.11	8.27	222.21	2511.55	8.78	220.62
Hisar 2	11	3451.85	5.58	3325.93	5.76	191.51	3325.06	6.09	202.47
Uklana	12	2538.8	13.51	2565.18	10.64	273.01	2493.6	9.51	237.07
Narnaund	31	1827.91	12.76	2070.24	10.77	222.91	2127.52	10.4	221.34
Hansi 1	36	2620.21	8.55	2577.83	8.33	214.86	2695.14	7.68	207.03
Hansi 2	22	2615.9	7.81	2534.32	8.06	204.24	2574.73	7.81	201.05
Wheat yield									
Block	**Sample size**	**Direct estimate**	**CV (%)**	**EBLUP Estimate Yi~NDVI**	**CV (%)**	**RMSE**	**EBLUP Estimate Yi~iNDVI**	**CV (%)**	**RMSE**
Adampur	24	4863.61	8.58	4697.84	6.69	314.11	4813.19	6.32	304.22
Agroha	23	4866.9	6.21	4941.07	6.34	313.04	4966.83	6.29	312.47
Barwala	38	5077	9.71	5037.23	3.81	191.76	5076.5	4.01	203.58
Hisar 1	46	4902.53	6.71	4975.01	5.74	285.51	4927.9	6.02	296.85
Hisar 2	39	4773.34	13.95	4505.52	8.49	382.66	4778.34	5.87	280.32
Uklana	12	5304.65	4.31	5167.33	8.19	423.21	5079.92	8.17	414.9
Narnaund	31	5497.33	8.81	5263.49	5.03	264.77	5247.68	6.05	317.29
Hansi 1	40	4825.95	9.26	4912.79	4.49	220.78	4926.1	4.61	227.29
Hansi 2	22	4138.58	12.58	5065.51	3.66	185.18	5169.64	4.79	247.5

Conclusion: The integration of remote sensing data and advanced statistical methods like EBLUP holds great potential for improving crop yield estimation and agricultural decision-making. This study contributes to the advancement of precision agriculture and offers insights into enhancing agricultural resource management and policy formulation. The combination of remote sensing technology and advanced statistical modelling has the potential to revolutionize crop yield estimation practices, leading to more sustainable and productive farming practices.

Reference:

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.

Cochran W G (1977) Sampling Techniques. John Wiley and Sons, New York.

Chandra H, Salvati N and Sud UC (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.

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.

Hidiroglou M (2007) Small-Area Estimation: Theory and Practice. *JSM proceedings,* survey research methods section.

Kott P (1989) Robust small domain estimation using random effects modelling. *Survey Methodology* **15**, 1-12.

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.

Muhammed P K, Kumar, Manoj, Airon, Anurag, Manjeet, (2020) EBLUP estimate of crop yield at sub-district level in Hisar district, Haryana, India using MODIS/Terra data. *Current Science*.

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.

APPLICATION OF SMALL AREA ESTIMATION USING LARGE-SCALE SURVEY DATA (CROP ESTIMATION SURVEY DATA, NSSO DATA ETC.)

Kaustav Aditya

ICAR- *Indian Agricultural Statistics Research Institute*, New Delhi-110012

Kaustav.aditya@icar.gov.in

1. Introduction

For planned development of a country, information on various aspects of economy is required to be collected on regular basis. The information can be collected through Census i.e. complete enumeration of the population under study. However, the conduct of Census is very time consuming, involves massive operations requiring huge resources, besides, being subject to large errors. Consequently, these can only be conducted after fairly long time gaps, which vary from country to country. In India, while the Population and Economic Censuses are conducted every 10 years, the Agricultural and Livestock Censuses are conducted every five years. For obtaining information during the intervening periods, large scale sample surveys are resorted to so that reliable, timely and adequate information on the parameters of interest from large populations can be provided. In India, National Sample Survey Organisation (NSSO) carries out country wide surveys on various socio-economic parameters related to the national economy such as follow up enterprise surveys of Economic Census, Annual Survey of Industries, supervision of Area enumeration and Crop Estimation surveys conducted by the state agencies so that appropriate data can be made available for policy planning and decision making on various issues of National importance. Similarly, the crop cutting experiments are organized by the Directorate of Economics and Statistics of various States for estimation of yield rates of various crops under the scheme of General Crop Estimation Surveys (GCES). While the sample sizes for the surveys conducted by the NSSO are fixed in such a manner that it is possible to get reasonably precise estimates at the State level, the sample size in the GCES are adequate to provide estimates at the District level. The state or regional level estimates generated by these survey are often masked the local level variation. But, the NSSO data cannot be used directly to produce reliable disaggregate level (e.g. district or further disaggregate level) estimates due to small sample sizes. The emphasis on micro-level planning reliable estimates of various parameters of interest are being demanded by the administrators and policy planners at the small area level. Due to the lack of robust and reliable estimates at lower level, proper planning, fund allocation and also monitoring of various plans is likely to suffer. In the survey literature, an area (or domain) is regarded as small if the area-specific (or domain - specific) sample is not large enough to support a direct survey estimator of adequate precision with unacceptably large coefficient of variation. A small area in the context of NSSO surveys may be a district while it may be a Community Development Block/ Gram Panchayat in case of GCES. In view of the demand for reliable statistics at the local level there is a burst of activity in the area of Small Area Estimation (SAE) technique. Newer techniques are increasingly being developed using tools of statistical inference and linear model. Simultaneously, attempts are also being made to apply these techniques so that precise estimates are available at the small/local area level. In many countries, SAE techniques are extensively used to produce the lower area level estimates, e.g. in United Kingdom the estimate of unemployment levels and rates for their Local Authority Districts (Ambler et al., 2001) and in United States the estimates of poor school-age

children at County level (Citro and Kalton, 2000). In India also, attempts have been made to use SAE techniques for various purposes (Sharma et al., 2004).

The growing demand for small area statistics in recent years has increased the popularity of SAE techniques. In this context model-based methods are widely used (Rao, 2003, chapter 2). The underlying idea is to use statistical models to link the variable of interest with auxiliary information to define the model-based estimator for small areas. Since the area-specific direct estimators do not provide adequate precision, for generating estimates for small areas it is necessary to employ model-based estimators that "borrow strength" from the related area. Small area model based techniques 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; these relate small area direct estimators to area-specific covariates (Fay and Herriot, 1979), and (ii) nested error unit level regression models, employed originally by Battese, Harter and Fuller (1988) for predicting areas under corn and soybean in 12 counties of the state of Iowa in the U.S.; these models relate the unit values of a study variable to unit-specific covariates.

The purpose of the study is to apply already available SAE technique. To achieve this we used NSS (2002-03 and 2004-05) and Agriculture census (1995-96) data to produce precise district level estimates. In particular, we employed an area level small area model to compute the empirical best linear unbiased predictor estimates and its mean squared error estimates because covariates, collected from Agriculture Census, are available at area level. Throughout this paper district and small area (or area) is used interchangeably

2. The Empirical Best Linear Unbiased Predictor for Small Areas

In the small area estimation method used here the covariates are collected from the Census which are available at District level. Here Districts are small area of interest. Widely used 'area level random effects model', is used because the auxiliary information is available only at the area level. This model was originally used by Fay and Herriot (1979) for the prediction of mean per capita income (PCI) in small geographical areas (less than 500 persons) within counties in USA, often referred to as Fay and Herriot model (hereafter FH model). In area level model there are two components:

(i) the direct survey estimate of the parameter based on the sampling design, expressed as

$$Y_d = y_d + e_d, \quad d = 1,...,D, \tag{1}$$

where D is total number of small areas that constitute our finite population, y_d are unobserved small area means (i.e., our parameter of interest), Y_d are observed direct survey estimators (the sample mean in our case) and the e_d's are independent sampling errors of survey estimate with $E(e_d / y_d) = 0$ and $V(e_d / y_d) = v_d$. The model (1) is a sampling model and v_d is a design-based sampling variance.

(ii) A linking model

$$y_d = \mathbf{z}_d^T \boldsymbol{\beta} + u_d, \quad d = 1, \ldots, D, \tag{2}$$

where $\mathbf{z}_d$ denotes p-vector of area (or District) level covariates, $\boldsymbol{\beta}$ is a p-vector of unknown fixed-effect coefficients and u_d is random effects (also called the model errors), assumed to be independent and identically distributed with $E(u_d) = 0$ and $V(u_d) = \sigma_u^2$.

Combining (1) and (2), we obtain the model

$$Y_d = \mathbf{z}_d^T\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D. \tag{3}$$

Clearly, the model (3) integrates a model dependent random effect u_d and a sampling error e_d with the two errors being independent. The model (3) is a special case of the linear mixed model. For known variance σ_u^2, assuming model (3) holds, the Best Linear Unbised Predictor (BLUP) for y_d (Henderson, 1963) is given by

$$\tilde{y}_d = \mathbf{z}_d^T\hat{\boldsymbol{\beta}}_{GLS} + \gamma_d(Y_d - \mathbf{z}_d^T\hat{\boldsymbol{\beta}}_{GLS}) = \gamma_d Y_d + (1-\gamma_d)\mathbf{z}_d^T\hat{\boldsymbol{\beta}}_{GLS} \tag{4}$$

where $\gamma_d = \sigma_u^2/(v_d + \sigma_u^2)$ and $\hat{\boldsymbol{\beta}}_{GLS} = \left(\sum_d (v_d + \sigma_u^2)^{-1}\mathbf{z}_d\mathbf{z}_d^T\right)^{-1}\left(\sum_d (v_d + \sigma_u^2)^{-1}\mathbf{z}_d Y_d\right)$ is the generalised least square estimate of $\boldsymbol{\beta}$. In practice, the variance σ_u^2 is usually unknown and they are replaced by sample estimates, $\hat{\sigma}_u^2$ (in equation (4) and $\hat{\boldsymbol{\beta}}_{GLS}$) yielding the corresponding empirical BLUP (EBLUP) denoted by $\hat{y}_d$. We note that the EBLUP $\hat{y}_d$ is a linear combination of a direct estimate Y_d and the model dependent regression synthetic estimate $\mathbf{z}_d^T\hat{\boldsymbol{\beta}}_{GLS}$, with weights given by γ_d. Here γ_d is called 'shrinkage factor' since it 'shrinks' the direct estimator towards the synthetic estimator $\mathbf{z}_d^T\hat{\boldsymbol{\beta}}_{GLS}$ (Rao, 2003, chapter 5).

Turning to mean squared error (MSE) estimation, if $\boldsymbol{\beta}$ and σ_u^2 are also known, the variance of the BLUP (4) is given as

$$Var[\tilde{y}_d(\sigma_u^2, \boldsymbol{\beta})] = \gamma_d v_d = g_{1d}$$

In practice, $\boldsymbol{\beta}$ and σ_u^2 are estimated from the sample data 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_{1d}. This estimator ignores the variability of $\hat{\sigma}_u^2$ and hence underestimates the true variance. Prasad and Rao (1990), extending the work of Kackar and Harville (1984) 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,

$$MSE[\hat{y}_d(\hat{\sigma}_u^2, \hat{\boldsymbol{\beta}}_{GLS})] = g_{1d} + g_{2d} + g_{3d} \tag{5}$$

where $g_{2d} = (1-\gamma_d)^2\mathbf{z}_d^T Var(\hat{\boldsymbol{\beta}}_{GLS})\mathbf{z}_d$ with $Var(\hat{\boldsymbol{\beta}}_{GLS}) = \left(\sum_d (v_d + \sigma_u^2)^{-1}\mathbf{z}_d\mathbf{z}_d^T\right)^{-1}$ is the excess in MSE due to estimation of $\boldsymbol{\beta}$ and $g_{3d} = [\sigma_{Di}^4/(\sigma_{Di}^2 + \sigma_u^2)^3] \times Var(\hat{\sigma}_u^2)$ is the excess in MSE due to estimation of σ_u^2. The neglected terms in the approximation are of order *o(1/D).* Building on the approximation, Prasad and Rao (1990) derive a MSE estimator of (5) with bias of order o(*1/D*) as,

$$mse[\hat{y}_d(\hat{\sigma}_u^2, \hat{\boldsymbol{\beta}}_{GLS})] = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2). \tag{6}$$

where $g_{kd}(\hat{\sigma}_u^2)$ is obtained from g_{kd} by substituting $\hat{\sigma}_u^2$ for σ_u^2, k=1,2,3. The MSE estimator (6) is robust with respect to departures from normality of the random area

effects u_d (but not the sampling errors e_d) (Lahiri and Rao, 1995). Here, standard error of the EBLUP is calculated as square root of MSE. Note that the leading term in (6) is $g_{1d} = \gamma_d v_d$ so for the small values of γ_d (i.e., the model variance σ_u^2 is small relative to the sampling variance v_d), $MSE[\hat{y}_d(\hat{\sigma}_u^2, \hat{\boldsymbol{\beta}}_{GLS})] << v_d = V_D(Y_d)$ illustrating the possible gains from using the model dependent estimator. Further, the availabilty of good auxiliary data is key to successful application of the small area technique since this provides a basis for good model fit. An excellent example of application of this method is in a study on Small Area Estimates of School-Age Children in Poverty in USA (Citro and Kalton, 2000).

3. Empirical study

The theory described in the previous section has been applied to develop district level estimates using the NSSO data. SAE techniques were applied to produce reliable small area estimates of incidence of poverty at district level in the State of Bihar in India by linking data from the existing Household Consumer Expenditure Survey data and the Population Census.

The incidence of poverty is defined as proportion of poor households, i.e. head count ratio (HCR). The HCR is poverty indicator or incidence measures the frequency of households under poverty line. Two types of variables are required for SAE analysis, the variable of interest and the auxiliary variables. In this study, the variable of interest for which small area estimates are required is drawn from the Household Consumer Expenditure Survey 2011-12 of NSSO for rural areas of the State of Bihar. The sampling design used in the NSSO data is stratified multi-stage random sampling with districts as strata, villages as first stage units and households as the second stage units. A total of 3312 households were surveyed from the 38 districts of the Bihar.

The district-wise sample size varied from minimum 64 to maximum 128 with average of 87 (Table 1). From Table 1, it is evident that district level sample sizes are very small with very low values of average sampling fraction as 0.00025. From Table 1, it is evident that district level sample sizes are very small with very low values of average sampling fraction as 0.00025. Therefore, it is difficult to produce reliable estimates and their standard errors at district level. The SAE is an obvious choice for such cases. The SAE technique is expected to provide reliable estimates for the districts having small sample data [8, 9 and 10]. The target variable used for the study is poor households. The poverty line has been used to identify whether given household is poor or not. A household having monthly per capita consumer expenditure below the state's poverty line (Rs 778) is categorized as poor household. The poverty line used in this study is same as those of year 2011-12, given by then planning commission, Govt of India.

Table 1. Distribution of district wise sample sizes (*n*), estimates of poverty incidence (Estimate) along 95 % confidence interval (95% CI) and percentage coefficient of variation (% CV) generated by direct survey estimate (DIR) and model based small area estimate (SAE estimate) for Bihar.

		DIR estimate				SAE estimate			
			95% CI				95% CI		
District	*N*	Estimate	Lower	Upper	% CV	Estimate	Lower	Upper	% CV
Pashchim Champaran	96	0.34	0.25	0.44	14.55	0.33	0.24	0.42	14.06
Purba Champaran	128	0.13	0.07	0.18	24.00	0.14	0.08	0.19	20.50
Sheohar	64	0.30	0.18	0.41	20.21	0.28	0.18	0.38	18.34
Sitamarhi	96	0.38	0.28	0.47	13.33	0.36	0.27	0.45	12.96
Madhubani	128	0.10	0.05	0.15	29.54	0.12	0.06	0.17	22.81

Supaul	64	0.05	-0.01	0.10	64.00	0.09	0.03	0.14	33.28
Araria	96	0.07	0.02	0.13	41.14	0.10	0.04	0.15	27.56
Kishanganj	64	0.09	0.02	0.17	42.67	0.12	0.05	0.19	29.11
Purnia	88	0.27	0.18	0.37	18.33	0.26	0.18	0.35	16.94
Katihar	88	0.18	0.10	0.26	22.00	0.19	0.11	0.26	20.94
Madhepura	64	0.00	0.00	0.00	0.00	0.06	0.02	0.10	37.27
Saharsa	64	0.08	0.01	0.14	38.40	0.11	0.04	0.17	31.00
Darbhanga	128	0.23	0.16	0.31	17.07	0.23	0.16	0.30	15.47
Muzaffarpur	128	0.23	0.16	0.31	17.07	0.23	0.16	0.29	15.33
Gopalganj	96	0.21	0.13	0.29	19.20	0.20	0.13	0.28	19.17
Siwan	96	0.29	0.20	0.38	17.14	0.28	0.20	0.37	15.40
Saran	128	0.16	0.09	0.22	19.20	0.16	0.10	0.22	18.75
Vaishali	96	0.09	0.04	0.15	32.00	0.12	0.06	0.18	25.64
Samastipur	128	0.17	0.11	0.24	17.45	0.18	0.11	0.24	17.97
Begusarai	96	0.06	0.01	0.11	32.00	0.09	0.04	0.14	29.73
Khagaria	64	0.09	0.02	0.17	42.67	0.12	0.05	0.19	28.87
Bhagalpur	96	0.18	0.10	0.25	22.59	0.18	0.11	0.25	20.14
Banka	64	0.22	0.12	0.32	22.86	0.22	0.13	0.31	21.42
Munger	64	0.27	0.16	0.38	22.59	0.25	0.15	0.35	20.08
Lakhisarai	64	0.16	0.07	0.25	32.00	0.16	0.08	0.23	25.64
Sheikhpura	64	0.17	0.08	0.27	29.09	0.17	0.09	0.25	24.96
Nalanda	96	0.29	0.20	0.38	17.14	0.28	0.19	0.36	15.85
Patna	96	0.30	0.21	0.39	16.55	0.29	0.21	0.38	14.98
Bhojpur	96	0.38	0.28	0.47	13.33	0.35	0.26	0.44	13.06
Buxar	64	0.34	0.23	0.46	17.45	0.31	0.20	0.41	17.43
Kaimur	64	0.23	0.13	0.34	21.33	0.22	0.13	0.31	21.03
Rohtas	96	0.33	0.24	0.43	15.00	0.31	0.22	0.40	14.47
Jehanabad	64	0.27	0.16	0.38	22.59	0.26	0.16	0.35	19.53
Aurangabad	64	0.19	0.09	0.28	26.67	0.19	0.11	0.28	23.17
Gaya	128	0.20	0.13	0.26	20.48	0.19	0.13	0.26	17.27
Nawada	64	0.16	0.07	0.25	32.00	0.16	0.08	0.24	25.45
Jamui	64	0.39	0.27	0.51	15.36	0.34	0.24	0.45	15.92
Arwal	64	0.20	0.10	0.30	24.62	0.19	0.10	0.28	23.79

The auxiliary variables used are collected drawn from the Population Census 2011. There are around 50 covariates available from Population Census 2011 to consider for small area modelling. However, we developed a composite score for selected group of variables using Principal Component Analysis (PCA). We first divided the selected number of variables in three groups and then considered PCA for these groups of variables. The first PCA (denoted by X_1) is based on gender-wise literacy rate and gender-wise worker population. The first principal component for first group of PCA (X_{11}) explained 52% of the variability in the dataset, while adding the second component (X_{12}) explained 100%. A second PCA (X_2) is based on following group of variables; gender-wise main worker, gender-wise main cultivator and gender-wise main agricultural labourers. The first principal component (X_{21}) for second group of PCA explained 67% of the variability in the dataset, while adding the second component (X_{22}) explained 94%. Finally, the third PCA (X_3) is derived from gender-wise marginal cultivator and gender-wise marginal agriculture labourers. The first principal component (X_{31}) for third group of PCA explained 52% of the variability in the dataset, while adding the second component (X_{32}) explained 77%. These six PCA scores (i.e. $X_{11}, X_{12}, X_{21}, X_{22}, X_{31}$ and X_{32}) developed from three group of variables were then used as auxiliary variables. We fitted a generalised linear model using direct survey estimates

of proportion of poor households as response variable and these six variables (i.e. $X_{11}, X_{12}, X_{21}, X_{22}, X_{31}$ and X_{32}) as covariates. The selected model with three significant covariates X_{11}, X_{21} and X_{31} have residual deviance and AIC values as 327.18 and 636.89 respectively. These three covariates X_{11}, X_{21} and X_{31} are used in SAE.

Let us assume a finite population *U* of size *N* and a sample *s* of size *n* is drawn from this population with a given survey design. We assume that this population is consist of *D* small areas or small domains (or simply areas or domains) $U_d (d = 1,...,D)$ such that $U = \bigcup_{d=1}^{D} U_d$ and $N = \sum_{d=1}^{D} N_d$. Throughout, we use a subscript *d* to index the quantities belonging to small area d $(d = 1,...,D)$, where D is the number of small areas (or areas) in the population. The subscript *s* and *r* are used for denoting the quantities related to the sample and non-sample parts of the population. So that n_d and N_d represent the sample and population (i.e., number of households in sample and population) sizes in district d, respectively. Let s_d denotes the part of sample from area *d* such that $s = \bigcup_{d=1}^{D} s_d$ and $n = \sum_{d=1}^{D} n_d$. Let y_{di} denotes the value of target variable of interest y for unit *i* in small area *d*. Let assume that the variable of interest *y* is binary and the target is the estimation of population counts $y_d = \sum_{i \in U_d} y_{di}$ or population proportions $P_d = N_d^{-1}\left(\sum_{i \in U_d} y_{di}\right)$ in area *d*. The direct estimator of proportion of poor household is $\hat{p}_d^{Direct} = \left(\sum_{i \in s_d} w_{di}\right)^{-1}\left(\sum_{i \in s_d} w_{di} y_{di}\right)$, where w_{di} is the survey weight associated with household *i* in area *d*. Assuming that joint inclusion $1/w_{di,d'j} = 0$ for $d \neq d'$ or $i \neq j$, the estimate of variance of $\hat{p}_d^{Direct}$ is $v(\hat{p}_d^{Direct}) = \left(\sum_{i \in s_d} w_{di}\right)^{-2}\left\{\sum_{i \in s_d} w_{di}(w_{di} - 1)(y_{di} - \hat{p}_d^{Direct})^2\right\}$. Let us denote by y_{sd} and y_{rd} the sample and non-sample counts of poor households in area (or district) *d*. The sample count y_{sd} has a Binomial distribution with parameters n_d and π_d, denoted by $y_{sd} \sim Bin(n_d, \pi_d)$, where π_d is the probability of a poor household in area *d*, often termed as the probability of a 'success'. Similarly, $y_{rd} \sim Bin(N_d - n_d, \pi_d)$. Further, y_{sd} and y_{rd} are assumed to be independent Binomial variables with π_d being a common success probability. Here we assume that only aggregated level data is available for the modelling. For example, from survey data y_{sd} and from secondary data sources (i.e. Census and Administrative records etc) $\mathbf{x}_d$ the *p*-vector of the covariates are available for area *d*. Following [4 and 6], the model linking the probabilities of success π_d with the covariates $\mathbf{x}_d$ is the logistic linear mixed model given by

$$logit(\pi_d) = \ln\left\{\frac{\pi_d}{1-\pi_d}\right\} = \eta_d = \mathbf{x}_d^T \boldsymbol{\beta} + u_d, \quad (1)$$

where $\boldsymbol{\beta}$ is the *p*-vector of regression coefficient often known as fixed effect parameters and u_d is the area-specific random effect that accounts for between area dissimilarity beyond that explained by the auxiliary variables included in the in the fixed part of the model. We assume that u_d's are independent and normally distributed with mean zero and variance φ. Under model (1),

$$\pi_d = \exp(\eta_d)\{1+\exp(\eta_d)\}^{-1} = \exp(\mathbf{x}_d^T\boldsymbol{\beta}+u_d)\{1+\exp(\mathbf{x}_d^T\boldsymbol{\beta}+u_d)\}^{-1} = \text{expit}(\mathbf{x}_d^T\boldsymbol{\beta}+u_d)$$

.

It is noteworthy that model (1) relates the area level proportions to area level covariates. This type of model is often referred to as 'area-level' model in SAE terminology. Area level model was originally proposed by Fay and Herriot [1979]. The Fay and Herriot method for SAE is based on area level linear mixed model and their approach is applicable to a continuous variable. The model (1) on the other hand is a special case of a generalized linear mixed model (GLMM) with logit link function and suitable binary variable. Here,

$y_{ds}|u_d \sim \text{Binomial}\left(n_d, \text{expit}(\mathbf{x}_d^T\boldsymbol{\beta}+u_d)\right)$ and $y_{dr}|u_d \sim \text{Binomial}\left(N_d-n_d, \text{expit}(\mathbf{x}_d^T\boldsymbol{\beta}+u_d)\right)$. This leads to $E\left(y_{sd}|u_d\right) = n_d\,\text{expit}(\mathbf{x}_d^T\boldsymbol{\beta}+u_d)$ and $E\left(y_{rd}|u_d\right) = \left(N_d-n_d\right)\text{expit}(\mathbf{x}_d^T\boldsymbol{\beta}+u_d)$. Collecting the area level models (1), we can write the model at population level as

$$g(\boldsymbol{\pi}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}. \quad (2)$$

Here $\boldsymbol{\pi} = (\pi_1,...,\pi_D)^T$, $\mathbf{X}=(\mathbf{x}_1^T,....,\mathbf{x}_D^T)^T$ is a $D\times p$ matrix, $\mathbf{Z}$ is a $D\times D$ diagonal matrix and $\mathbf{u} = (u_1,...,u_D)^T$ is a vector of $D\times 1$ of area random effects, which is normally distributed with mean zero and variance $\boldsymbol{\Omega} = \phi\mathbf{I}_D$. Here, $\mathbf{I}_D$ is a $D\times D$ diagonal matrix. Note that estimation of fixed effect parameters $\boldsymbol{\beta}$ and area specific random effects u_d's uses the data from all small areas. We used an iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ with restricted maximum likelihood (REML) estimation of ϕ to estimate these unknown parameters.

Let us write the total population counts, i.e. the total number of poor households in district *d* as $y_d = y_{sd}+y_{rd}$, where the first term y_{sd}, the sample count is known whereas the second term y_{rd}, the non-sample count, is unknown. Therefore, a plug-in empirical best predictor (EBP) of the population count in area *d* is

$$\hat{y}_i^{EBP} = y_{si}+\hat{E}\left(y_{rd}|u_d\right) = y_{si}+\left(N_d-n_d\right)\left[\text{expit}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\mathbf{Z}_i^T\hat{\mathbf{u}})\right], \quad (3)$$

where $\mathbf{Z}_i^T = \left(0,..,1,.,0\right)$ is $1\times D$ vector with 1 in position *i*-th. An estimate of proportion in area *d* is obtained as $\hat{p}_d^{EBP} = N_d^{-1}\hat{y}_d^{EBP}$**.** For area with zero sample sizes (i.e. non-sampled areas), the conventional approach for estimating area proportions or counts is synthetic estimation, based on a suitable GLMM fitted to the data from the sampled areas [4]. Under (1), for non-sampled areas, the synthetic type predictor of total population count

for area d is $\hat{y}_d^{SYN} = N_d \,\text{expit}(\mathbf{x}_{d,out}^T \hat{\boldsymbol{\beta}})$, where $\mathbf{x}_{d,out}$ denote the vector of covariates associated with non-sampled area *d*.

The mean squared error (MSE) estimates are computed to assess the reliability of estimates and also to construct the confidence interval for the small area estimates. Estimation of the MSE of the EBP (3) is followed from development reported in [4,6,7 and 11] and references therein. Let us denote by $\hat{\mathbf{V}}_{sd} = diag\left\{n_d \hat{p}_d^{EBP}(1-\hat{p}_d^{EBP})\right\}$ and $\hat{\mathbf{V}}_{rd} = diag\left\{(N_d - n_d)\hat{p}_d^{EBP}(1-\hat{p}_d^{EBP})\right\}$, the diagonal matrices defined by the corresponding variances of the sample and non-sample part respectively. Similarly, $\mathbf{A} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}$, $\mathbf{B} = \left\{diag(N_d^{-1})\right\}\left\{\hat{\mathbf{V}}_{rd}\mathbf{X} - \mathbf{A}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_{sd}\mathbf{X}\right\}$ and $\hat{\boldsymbol{\Sigma}} = \left(\phi^{-1}\mathbf{I}_D + \hat{\mathbf{V}}_{sd}\right)^{-1}$, where $\mathbf{I}_D$ is an identity matrix of order *D*. We further define $\hat{\mathbf{V}}_{(1)} = \left\{\mathbf{X}^T\hat{\mathbf{V}}_{sd}\mathbf{X} - \mathbf{X}^T\hat{\mathbf{V}}_{sd}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_{sd}\mathbf{X}\right\}^{-1}$ and $\hat{\mathbf{V}}_{(2)} = \hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}_{sd}\mathbf{X}\hat{\mathbf{V}}_{(1)}\mathbf{X}^T\hat{\mathbf{V}}_{sd}^T\hat{\boldsymbol{\Sigma}}$. With these notations, assuming model (1) holds, the MSE estimate of (3) is given by

$$mse(\hat{p}_d^{EBP}) = m_1(\hat{\phi}) + m_2(\hat{\phi}) + 2m_3(\hat{\phi}). \qquad (4)$$

The first two components m_1 and m_2 constitute the largest part of the overall MSE estimates in (4). These are the MSE of the best linear unbiased predictor type estimator when ϕ is known [11]. The third component m_3 is the variability due to the estimate of ϕ. The three components of (4) are defined as follows:

$m_1(\hat{\phi}) = \mathbf{A}\hat{\boldsymbol{\Sigma}}\mathbf{A}^T$, $m_2(\hat{\phi}) = \mathbf{B}\hat{\mathbf{V}}_{(1)}\mathbf{B}^T$, and $m_3(\hat{\phi}) = trace\left(\hat{\nabla}_i\hat{\boldsymbol{\Sigma}}^+\hat{\nabla}_j^T v(\hat{\phi})\right)$ with $\hat{\boldsymbol{\Sigma}}^+ = \hat{\mathbf{V}}_{sd} + \hat{\phi}\mathbf{I}_D\hat{\mathbf{V}}_{sd}\hat{\mathbf{V}}_{sd}^T$.

Here $v(\hat{\phi})$ is the asymptotic covariance matrix of the estimates of variance components $\hat{\phi}$, which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\phi}$. Note that this also depends upon whether we are using ML or REML estimates for $\hat{\phi}$. We used REML estimates for $\hat{\phi}$, then $v(\hat{\phi}) = 2\left(\hat{\phi}^{-2}(D-2a_1) + \hat{\phi}^{-4}a_{11}\right)^{-1}$ with $a_1 = \hat{\phi}^{-1}\,trace(\hat{\mathbf{V}}_{(2)})$ and $a_{11} = trace(\hat{\mathbf{V}}_{(2)}\hat{\mathbf{V}}_{(2)})$. Let us write $\Delta = \mathbf{A}\hat{\boldsymbol{\Sigma}}$ and $\hat{\nabla}_i = \partial(\Delta_i)/\partial\phi\big|_{\phi=\hat{\phi}} = \partial(\boldsymbol{A}_i\hat{\boldsymbol{\Sigma}})/\partial\phi\big|_{\phi=\hat{\phi}}$, where $\boldsymbol{A}_i$ is the i^{th} row of the matrix $\mathbf{A}$.

In SAE application, generally two types of diagnostics measures are suggested and employed, the model diagnostics and the diagnostics for the small area estimates see [3 and 4]. The model diagnostics are applied to verify the assumptions of underlying model, i.e. how well working model is fitted to data. The random district specific effects in model (1) are assumed to have an independent and identical normal distribution with mean zero and fixed variance ϕ. If the model assumptions are satisfied then the district level residuals from model (1) are expected to be randomly (i.e., pattern less) distributed and not significantly different from the regression line *y=0.* Fig 1 shows histogram (left), normal probability plot of district level residuals (centre) and distribution of district level residuals (right). From Fig 1, it appears that district level residuals are randomly

distributed and the line of fit does not significantly differ from the line y=0. The histogram and q-q plot also confirm the normality assumption for random area effects. Therefore the model diagnostics are fully satisfied for the data.

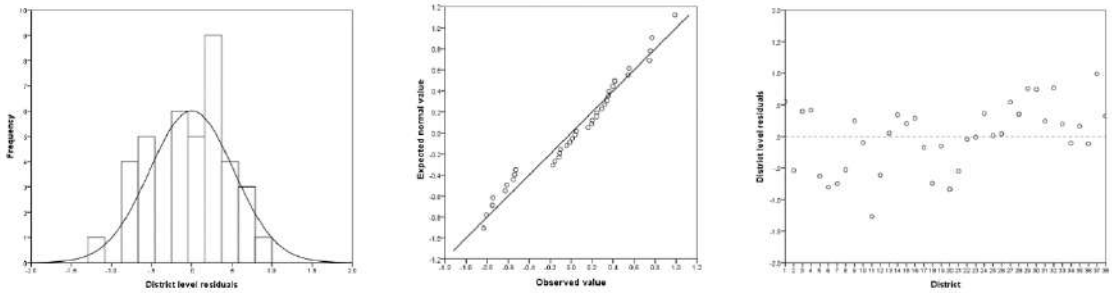


Fig 1. Histogram (left), Normal q-q plots (centre) and distribution of the district level residuals (right).

Second set of diagnostics is used for assessing the reliability and the validity of the small area estimates. Such diagnostics are suggested in [3]. Model-based small area estimates should be (a) consistent with unbiased direct survey estimates and (b) more precise than direct survey estimates. The values for the model-based small area estimates derived from the fitted model should be consistent with the unbiased direct survey estimates, wherever these are available, i.e. they should provide an approximation to the direct survey estimates that is consistent with these values being "close" to the expected values of the direct estimates. The model-based small area estimates should have mean squared errors significantly lower than the variances of corresponding direct survey estimates. For this purpose, we consider three commonly used measures namely the bias diagnostics, percent coefficient of variation (CV) and the 95 percent confidence intervals for small area estimates diagnostics.

The bias diagnostics is used to investigate if the small area estimates are less extreme when compared to the direct survey estimates, when it is available. In addition, if direct estimates are unbiased, their regression on the true values should be linear and correspond to the identity line. If small area estimates are close to the true values the regression of the direct estimates on the model-based estimates should be similar. We plotted direct estimates on y-axis and model based small area estimates on x-axis and we looked for divergence of regression line from $y= x$ and test for intercept = 0 and slope = 1. The bias scatter plot of the direct survey estimates against the model based small area estimates for EBP is given in Fig 2. The bias diagnostics plot in Fig 2 indicates that the small area estimates generated by EBP are less extreme when compared to the direct survey estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average. That is, the estimates of poverty incidence generated by EBP method lies along the line y= x for most of the districts which indicates that they are approximately design unbiased.

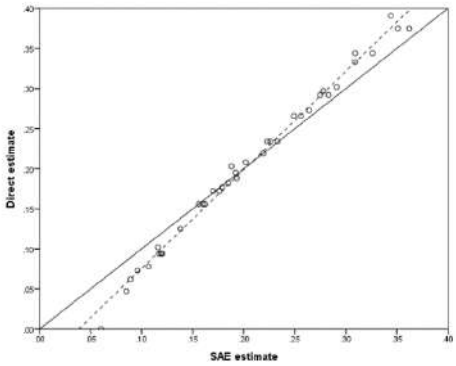


Fig 2. Bias diagnostics plots with $y = x$ line (solid line) and regression line (dotted line) model based small area estimate.

We also used Goodness of Fit (GoF) diagnostic. This diagnostic tests whether the direct and model-based estimates generated by EBP are statistically different. The null hypothesis is that the direct and model-based estimates are statistically equivalent. The alternative is that the direct and model-based estimates are statistically different. The GoF diagnostic is computed using the following Wald statistic for EBP estimate:

$$W = \sum_d \left\{ \frac{\left(\text{Direct estimate}_d - \text{EBP estimate}_d\right)^2}{\widehat{Var}(\text{Direct estimate}_d) + \widehat{MSE}(\text{EBP estimate}_d)} \right\}.$$

The value from the test statistic is compared against the value from a chi square distribution with D degrees of freedom. For our analysis, this is the chi square value with $D=38$ degrees of freedom which is 24.88 at 5% level of significance. For EBP, the value of Wald statistic is W=11.81. A smaller value (less than 24.88 in this case) indicates no statistically significant difference between model-based estimates generated by EBP and direct estimates. The diagnostic results clearly show that EBP estimates are consistent with direct estimates. We also examine the aggregation of direct and model based EBP estimate at state level. We computed state level incidence of poverty by aggregating the direct estimates as $\text{Direct estimate} = \sum_d (N_d \times \text{Direct estimate}_d) / \sum_d N_d$ and model based estimates as $\text{Model based estimate} = \sum_d (N_d \times \text{EBP estimate}_d) / \sum_d N_d$. The state level estimate of incidence of poverty by direct and EBP methods are 0.200 and 0.202 respectively. As one expects, model based estimate are aggregated well to state level direct estimate.

We use the percent CV to assess the comparative precision of model-based small area estimates (EBP) and direct survey estimates. The CVs show the sampling variability as a percentage of the estimate. Estimates with large CVs are considered unreliable (i.e. smaller is better). In general, there are no internationally accepted tables available that allow us to judge what is "too large". Different organization used different cut off for CV to release their estimate for the public use. For example, Office for National Statistics, United Kingdom has cut off CV value of 20% for acceptable estimates. The % CV of direct and EBP estimates are in given in Table 1. Fig 3 presents the district wise distribution of % CV for the model-based estimates and direct estimates. The results in Table 1 and district wise values in Fig 3 clearly show that direct estimates for small area poverty incidence are unstable with CV varies from 13.33% to 64 % with average of 24.69. The % CV of EBP ranges from 12.96 % to 37.27 % with average of 21.19%. In

Fig 3 and Table 1 we further notice that for direct estimates CVs are greater than 20% and 30% in 22 and 9 (out of 38) districts respectively. These results clearly reveal the model-based small area estimates generated by EBP are reliable. In contrast, the direct estimates are very unstable.

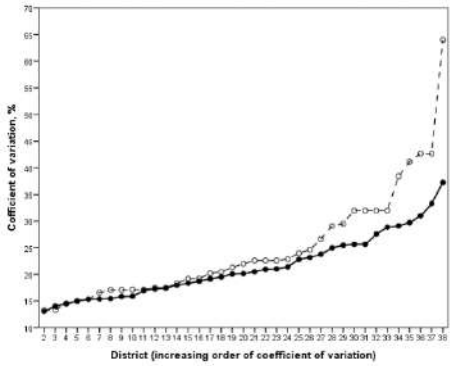


Fig 3. District-wise percentage coefficient of variation for the direct (dash line,°) and model based small area estimate (solid line, •).

The districts-wise 95 percent confidence intervals (95% Cis) of the EBP and the direct estimates are shown in Fig 4. It is important to note that the 95% CIs for the direct estimates are calculated assuming a simple random sample generated the simple proportions. This ignores the effects of differential weighting and clustering within districts that would further inflate the true standard errors of the direct estimates. The standard errors of the direct estimates are too large and therefore the estimates are unreliable. In Fig 4, we observe that 95% CIs for the direct estimates are wider than the 95% CIs for the EBP estimates. It indicates that the 95% CIs for the EBP estimates are more precise and contain both direct and EBP estimates of incidence of poverty.

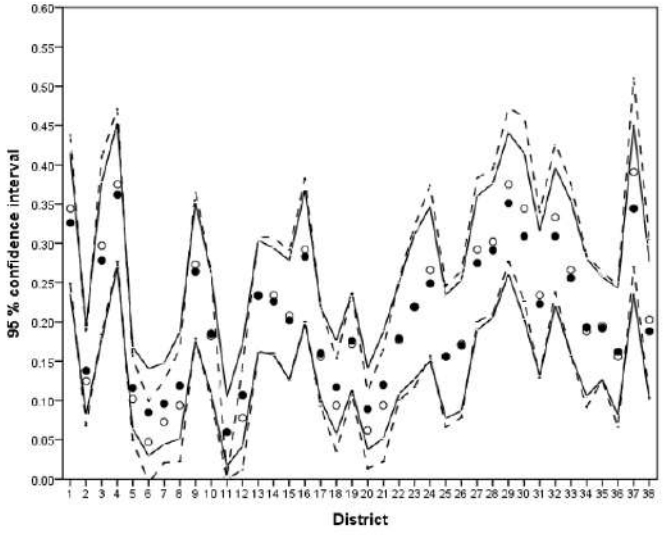


Fig 4. District-wise 95 percent confidence interval (lower and upper) for the direct (dash line) versus model based small area estimates (solid line). Direct estimate (dash line,°) and model based small area estimate (solid line, •).

The spatial mapping district-wise poverty incidence for the State of Bihar is shown in Fig 5. This map provides the spatial inequality in distribution of poverty incidence, i.e. the degree of inequality with respect to distribution of proportion of poor households in

different districts. This map is very useful in identifying the districts and regions with low and high level of poverty incidence in the state. The district-wise poverty incidence generated by EBP method in rural areas of Bihar ranges from 6 to 36% with average of 21%. From Fig 5, it can be seen that Madhepura (6 %) has lowest poverty incidence in the state. Supaul, Begusarai, Araria, Saharsa, Madhubani, Vaishali, Kishanganj, Khagaria and Purba Champaran have smallest poverty rate (9-14%) whereas Buxar, Rohtas, Pashchim Champaran, Jamui, Bhojpur and Sitamarhi have highest rate of poverty incidence (31-36%). This map clearly shows that districts bordering with eastern Uttar Pradesh have higher poverty incidence. The district level estimates as well as spatial maps of poverty rates are expected to provide invaluable information to policy-analysts and decision-makers for identifying the regions and districts requiring more attention in the State. This application and description of methodology can also be used a guideline for other application of SAE in different survey data as well as data from other countries.

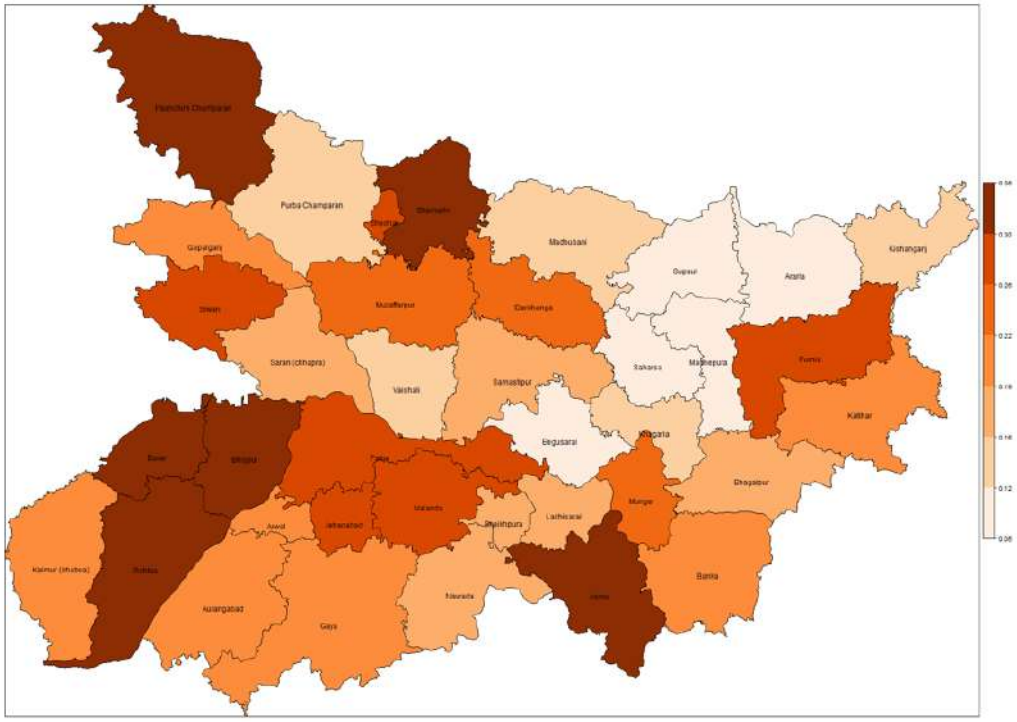


Fig 5. Poverty mapping generated for the state of Bihar in 2011-12.

Conclusions

Theory of SAE method for estimation of proportions for small areas is well developed, however, application in the area of agricultural or social sciences are not so popular. In developed countries like USA, UK, Australia etc., SAE has been initiated and included as a part of their objectives in the national statistical offices. Need of small area statistics has been felt in different agencies and organization in India, but, not much initiative has been taken place. In India, Censuses are usually limited as they tend to focus mainly on the basic socio-demographic and economic data and not available for every time period. On the other hand, country is fortunate to have regular NSSO survey for generating number of socio-economic indicators. The NSSO surveys are aimed to generate estimates at national and state level. They do not provide sub-state level statistics. The SAE can be used as cost effective and efficient approach for generating reliable micro level statistics from existing survey data and using auxiliary information

from different published sources. The results clearly indicates the advantage of using SAE technique to cope up the small sample size problem in producing the estimates or reliable confidence intervals. It is evident that model based SAE method brings gain in efficiency in district level estimates. Disaggregate level estimates of poverty incidence and poverty mapping are useful information for identifying the districts/regions with higher level of poverty rate. These information can be used by state government in allocation of budget in various government schemes.

References

Aditya K, Chandra H, Basak P, Kumari V and Das S. (2020). District Level Major Crop Yield Estimation using Reduced Number of Crop Cutting Experiments. *Indian Journal of Agricultural Sciences*, **90(6)**, 1185-89.

Anjoy P, Chandra H and **Aditya K**. (2020). Spatial Hierarchical Bayes Small Area Model for disaggregated Level Crop Acreage estimation. *Indian Journal of Agricultural Sciences*, **90(9),** 1780-85**.**

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M. and Wang, S. (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 Statisticians,* Meeting of the International Statistics Institute, 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 Statistical Association*, **95**, 1127-1142.

Chambers, R., Chandra, H. and Tzavidis, N. (2007). Small Area Estimation Course Notes. *Third International Conference on Establishment Surveys*, Montreal, Canada, June 18-21, 2007.

Chandra H, Sud UC and Salvati N. 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. 2001; 65(1): 1-8.

Citro, C. and Kalton, G. (2000). Small-Area Estimates of School-Age Children in Poverty. *Evaluation of Current Methodology* (National Research Council), Washington DC: Nat. Acad. Press.

Fay, R. E. and Herriot, R. A. (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.

H Chandra, K Aditya and U C Sud (2018). Localized estimates and spatial mapping of poverty incidence in the state of Bihar in India-An application of Small area estimation Techniques", PLOS One. doi.org/10.1371/journal.pone.0198502.

H Chandra, Kaustav Aditya*****, Swati Gupta, Saurav Guha and Bhanu Verma, (2020). Food and Nutrition in the Indo-Gangetic plain-a disaggregate Level analysis, *Current Science*, **119(11)**, 1783-1788.

Henderson, C.R. (1963). Selection Index and Expected Genetic Advance, in *Statistical genetics and Plant Breeding*, eds. W.D. Hanson and H.F. Robinson, Washington, DC: national Academic of Sciences- National Research Council, 141-163.

Johnson FA, Chandra H, Brown J and Padmadas S. Estimating district-level births attended by skilled attendants in ghana using demographic health survey and census data: an application of small area estimation technique. Journal Official Statistics. 2010; **26(2)**: 341-359.

Kackar, R.N. and Harville, D. A.(1984). Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models. *Journal of the American Statistical Association*, **79**, 853-862.

Lahiri, P. and Rao, J.N.K. (1995). Robust Estimation of Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association*, **90**, 758-766.

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**, 163-71.

Rao, J. N. K. (2003). *Small Area Estimation.* John Wiley & Sons, New York.

Särndal, C-E, Swensson, B, Wretman, J. Model Assisted Survey Sampling, Springer, New York; 1992.

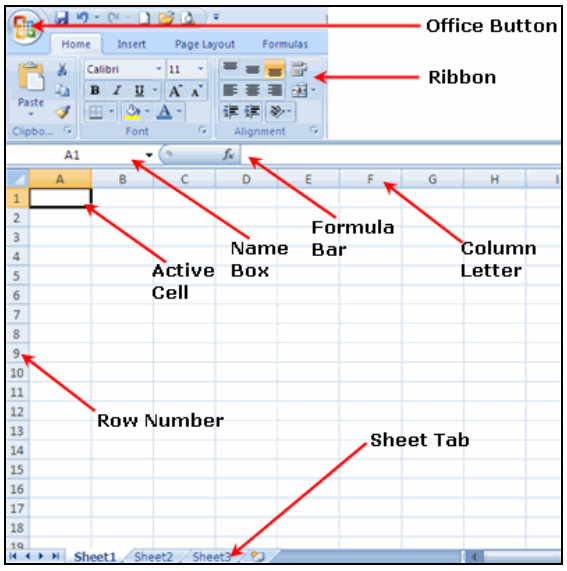
MS-EXCEL: STATISTICAL PROCEDURES

Cini Varghese

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012

Introduction:

Microsoft (MS) Excel ( ) is a powerful spreadsheet that is easy to use and allows you to store, manipulate, analyze, and visualize data. It also supports databases, graphic and presentation features. It is a powerful research tool and needs a minimum of teaching. Spreadsheets offer the potential to bring the real numerical work alive and make statistics enjoyable. But the main disadvantage is that some advanced statistical functions are not available and it takes a longer computing time as compared to other specialized software.



Data Entry in Spreadsheets

- Data entry should be started soon after data collection in the field
- The raw data collected should be entered directly into computer. Calculations (e.g. % dry matter) or conversions (e.g. kg/ha to t/ha) by hand will very likely result in errors and therefore require more data checking once the data are in MS-Excel. Calculations can be written in MS-Excel using formulae (e.g. sum of wood biomass and leaf biomass to give total biomass).

Data Checking

One can use calculations and conversions for data checking. For example, if the collected data is grain yield per plot it may be difficult to see whether the values are

reasonable. However, if these are converted to yield per hectare then one can compare the numbers with our scientific knowledge of grain yields. Simple formulae can be written to check for consistency in the data. For example, if tree height is measured 3 times in the year, a simple formula that subtracts 'tree height 1' from 'tree height 2'can be used to check the correctness of the data. The numbers in the resulting column should all be positive. We cannot have a shrinking tree! For new columns of calculated or converted data suitable header information (what the new column is, units and short name) at the top of the data should be included.

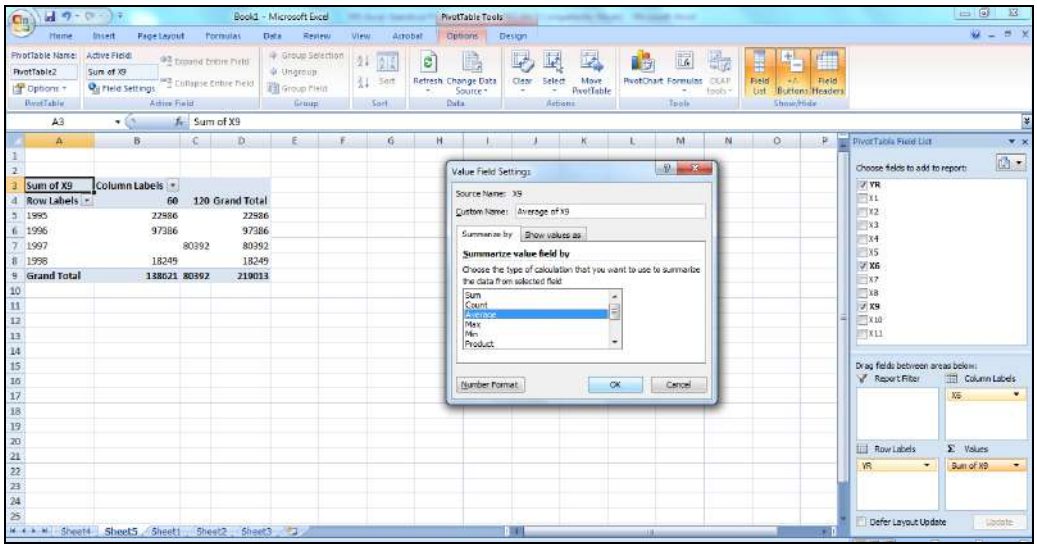
Missing Values

In MS-Excel the missing values are BLANK cells. It is useful to know this when calculating formulae and summaries of the data. For example, when calculating the average of a number of cells, if one cell is blank MS-Excel ignores this as an observation (i.e., the average is the sum/number of non-blank cells). But if the cell contains a '0' then this is included in the calculation (i.e., the average is the sum/no. of cells). In a column of 'number of fruit per plot', a missing value could signify zero (tree is there but no fruit), dead (tree was there but died so no fruit), lost (measurement was lost, illegible.) or not representative (tree had been browsed severely by goats). In this example, depending on the objectives of the trial, the scientist might choose to put a '0' in the cells of trees with no fruit and leave blank (but add comments) for the other 'missing values'.

Pivot Tables (to check consistency between replicates)

Variation between replicates is expected, but some level of consistency is also usual. We can use pivot tables to look at the data. A pivot table is an interactive worksheet table that quickly summarizes large amount of data using a format and calculation methods you choose. It is called pivot table because you can rotate its row and column heading around the core data area to give you different views of the source data. A pivot table provides an easy way for you to display and analyze summary information about data already created in MS-Excel or other application.

- Keep the cursor anywhere within the data range
- Choose “Insert” “Pivot Table” then “OK”
- From the “Pivot table Field List” drag and drop the respective fields under “Column Labels” , “Row Labels” and “Σ Values”
- Select “Value Field Settings” by clicking on the down arrow in “Σ Values” and choose the appropriate option and then click “OK”



Scatter Plots (to check consistency between variates)

We can often expect two measured variables to have a fairly consistent relationship with each other. For example, 'number of fruits' with 'weight of fruits' or Stover yield plotted against grain yield. To look for odd values we could plot one against the other in a scatter plot. Scatter plots are useful tools for helping to spot outliers. This option is available under "Insert" menu.

Line Plots (to examine changes over time)

Where measurements on a 'unit' are taken on several occasions over a period of time it may be possible to check that the changes are realistic. A check back at the problematic data which is not in the usual trend can be made. . This option is available under "Insert" menu.

Double Data Entry

One effective, although not always practical, way of checking for errors caused by data entry mistakes is double entry. The data are entered by two individuals onto separate sheets that have the same design structure. The sheets are then compared and any inconsistencies are checked with the original data. It is assumed that the two data entry operators will not make the same errors. There is no 'built-in' system for double entry in MS-Excel. However, there are some functions that can be used to compare the two copies. An example is the DELTA function that compares two values and returns a 1 if they are the same and a 0 otherwise. To use this function we would set up a third worksheet and input a formula into each cell that compares the two identical cells in the other two worksheets. The 0's on the third worksheet will therefore identify the contradictions between the two sets of data. This method can also be used to check survey data but for the process to work the records must be entered in exactly the same order in both sheets. If a section at the bottom of the third worksheet contains mostly 0's, this could indicate that you have omitted a record in one of the other sheets.

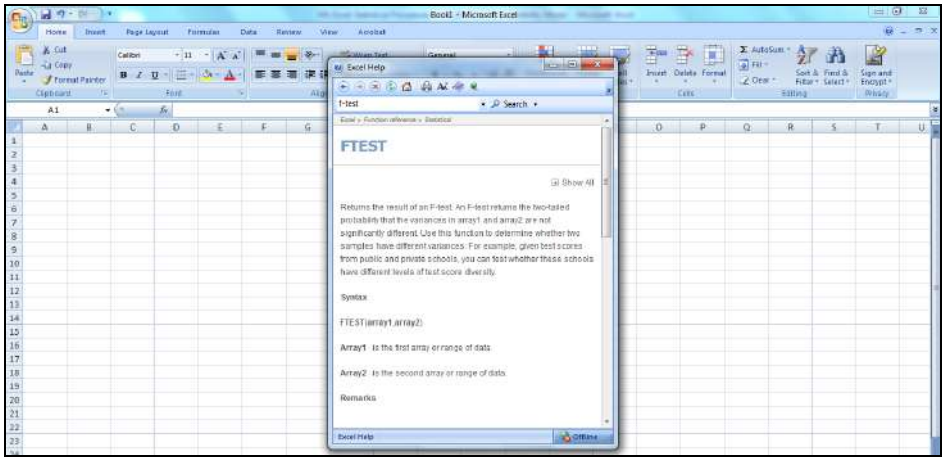
Preparing Data for Export to a Statistical Package

Statistical analysis of research data usually involves exporting the data into a statistical package such as GENSTAT, SAS or SPSS. These packages require you to give the MS-Excel cell range from which data are to be taken. In the latest editions of MS-Excel we can mark these ranges within MS-Excel and then transfer them directly into the statistical packages.

- Highlight the data you require including the column titles (the codes which have been used to label the factors and variables).
- Go to the Name Box, an empty white box at the top left of the spreadsheet. Click in this box and type a name for the highlighted range (e.g., Data). Press Enter.
- From now on, when you want to select your data to export go to the Name Box and select that name (e.g. Data). The relevant data will then be highlighted.

MS-Excel Help

If you get stuck on any aspect of MS-Excel then use the Help facility by clicking "F1" key. It contains extensive topics and by typing in a question you can extract the required information. See the snapshot below for an example:



FEATURES OF MS-EXCEL

Analytic Features

- The windows interface includes windows, pull down menus, dialog boxes and mouse support
- Repetitive tasks can be automated with MS-Excel. Easy to use macros and user defined functions
- Full featured graphing and charting facilities
- Supports on screen databases with querying, extracting and sorting functions

- Permits the user to add, edit, delete and find database records

Presentation Features

- Individual cells and chart text can be formatted to any font and font size
- Variations in font size, style and alignment control can be determined
- The user can add legends, text, pattern, scaling and symbols to charts.

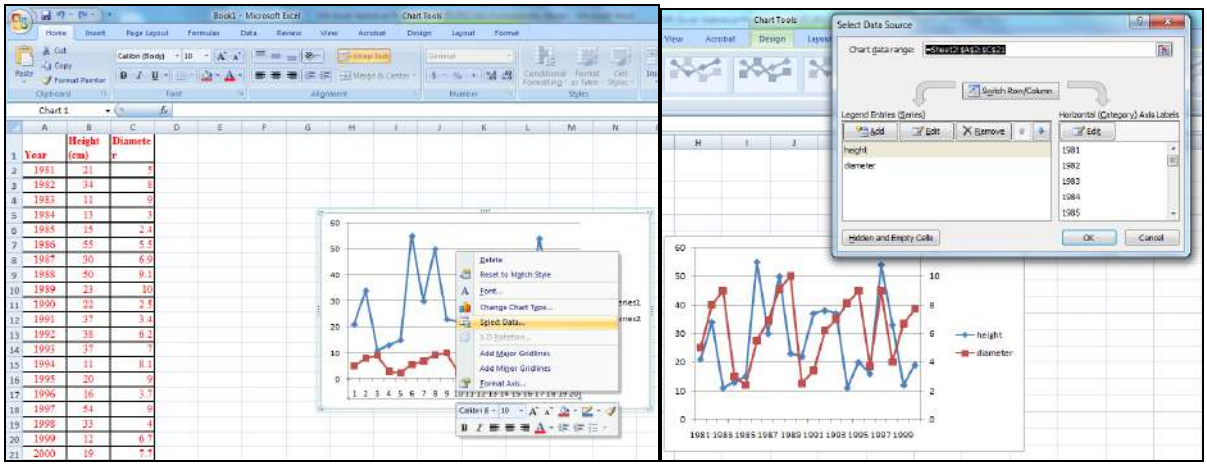
Charts and Graphs

A chart is a graphic representation of worksheet data. The dimension of a chart depends upon the range of the data selected. Charts are created on a worksheet or as a separate document that is saved with an extension xlsx. MS-Excel automatically scales the axes, creates columns categories and labels the columns. Values from worksheet cells or data points are displayed as bars, lines, columns, pie slices, or other shapes in the chart. Showing a data in a chart can make it clearer, interesting and easier to understand. Charts can also help the user to evaluate his/her data and make comparisons between different worksheet values.

Creating Line Chart

- Select relevant part of data
- Choose “Insert” “line”
- Select an appropriate option of line chart and click

Necessary changes in the chart can be done by clicking the right button of the mouse and choosing appropriate options.



Sorting and Filtering

MS-Excel makes it easy to organize, find and create report from data stored in a list.

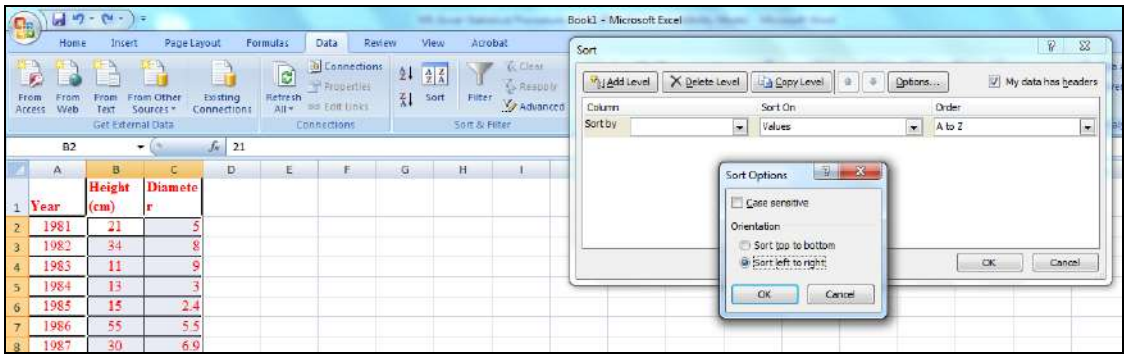
Sort: To organize data in a list alphabetically, numerically or chronologically.

(i) To sort entire list

- Select a single cell in the list
- Choose “data” “sort”

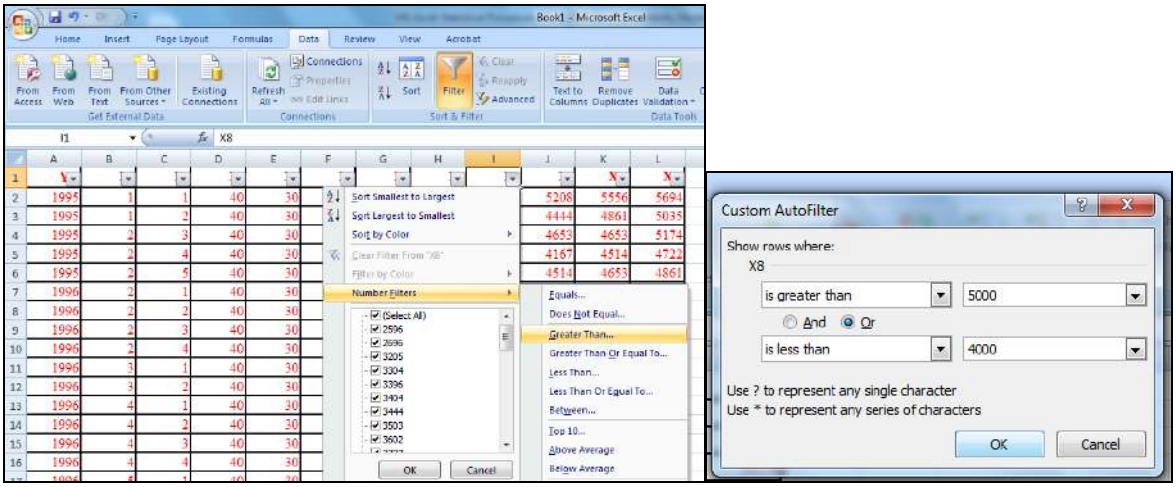
(ii) Sorting column from left to right

- Choose the “option” button in the sort dialog box
- In the sort option dialog box, select “sort left to right”
- Choose “OK”



Filter: To quickly find and work with a subset of your data without moving or sorting it.

- Choose “Data” and click on “Filter”
- MS-Excel place a drop down arrow directly on the column labels of the list
- Choose the column based on which the data has to be filtered. Clicking on the arrow displays a list of all the unique items in the column. Choose “Number Filter” option and define the required conditions.



STATISTICAL FUNCTIONS

Excel's statistical functions are quite powerful. In general, statistical functions take lists as arguments rather than single numerical values or text. A list could be a group of numbers separated by commas, such as (3,5,1,12,15,16), or a specified range of cells, such as (A1:A6), which is the equivalent of typing out the list (A1,A2,A3,A4,A5,A6). The function COUNT(list) counts the number of values in a list, ignoring empty or nonnumeric cells, whereas COUNTA(list) counts the number of values in the list that have any entry at all. MIN(list) returns a list's smallest value, whereas MAX(list) returns a list's largest value. The functions AVERAGE(list), MEDIAN(list), MODE(list), STDEV(list) all carry out the statistical operations you would expect (STDEV stands for standard deviation), when you pass a list of values as an argument.

Create a Formula

Formulas are equations that perform calculations on values in your worksheet. A formula starts with an equal sign (=). For example, the following formula multiplies 2 by 3 and then adds 5 to the result: =5+2*3. The following formulas contain operators and constants:

Example formula	What it does
=128+345	Adds 128 and 345
=5^2	Squares 5

- Click the cell in which you want to enter the formula.
- Type = (an equal sign).
- Enter the formula.
- Press ENTER.

Create a Formula that Contains References or Names: A1+23

The following formulas contain relative references and names of other cells. The cell that contains the formula is known as a dependent cell when its value depends on the values in other cells. For example, cell B2 is a dependent cell if it contains the formula =C2.

Example formula	What it does
=C2	Uses the value in the cell C2
=Sheet2!B2	Uses the value in cell B2 on Sheet2
=Asset-Liability	Subtracts a cell named Liability from a cell named Asset

- Click the cell in which the formula enter has to be entered.
- In the formula bar, type = (equal sign).
- To create a reference, select a cell, a range of cells, a location in another worksheet, or a location in another workbook. One can drag the border of the cell selection to move the selection, or drag the corner of the border to expand the selection.
- Press ENTER.

Create a Formula that Contains a Function: =AVERAGE(A1:B4)

The following formulas contain functions:

Example formula	What it does
=SUM(A:A)	Adds all numbers in column A
=AVERAGE(A1:B4)	Averages all numbers in the range

- Click the cell in which the formula enter has to be entered.
- To start the formula with the function, click “insert function” on the formula bar.
- Select the function.
- Enter the arguments. When the formula is completed, press ENTER.

Create a Formula with Nested Functions: =IF(AVERAGE(F2:F5)>50, SUM(G2:G5),0)

Nested functions use a function as one of the arguments of another function. The following formula sums a set of numbers (G2:G5) only if the average of another set of numbers (F2:F5) is greater than 50. Otherwise it returns 0.

STATISTICAL ANALYSIS TOOLS

Microsoft Excel provides a set of data analysis tools — called the Analysis ToolPak — that one can use to save steps when you develop complex statistical or engineering analyses. Provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Accessing the Data Analysis Tools: To access various tools included in the Analysis ToolPak click on “Data” menu, then click “Data Analysis” and select the appropriate analysis option. If the “Data Analysis” command is not available, we need to load the Analysis ToolPak “select and run the “Analysis ToolPack” from the “Add-Ins”.

Correlation

The “Correlation” analysis tool measures the relationship between two data sets that are scaled to be independent of the unit of measurement. It can be used to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive correlation), whether small values of one set are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (correlation near zero).

If the experimenter had measured two variables in a group of individuals, such as foot-length and height, he/she can calculate how closely the variables are correlated with each other. Select “Data”, “Data Analysis”. Scroll down the list, select “Correlation” and click OK. A new window will appear where the following information needs to be entered:

Input range. Highlight the two columns of data that are the paired values for the two variables. The cell range will automatically appear in the box. If column headings are included in this range, tick the Labels box.

Output range. Click in this box then select a region on the worksheet where the user want the data table displayed. It can be done by clicking on a single cell, which will become the top left cell of the table.

Click OK and a table will be displayed showing the correlation coefficient (r) for the data.

CORREL(array1, array2) also returns the correlation coefficient between two data sets.

Covariance

Covariance is a measure of the relationship between two ranges of data. The “covariance” tool can be used to determine whether two ranges of data move together, *i.e*., whether large values of one set are associated with large values of the other (positive covariance), whether small values of one set are associated with large values of the other (negative covariance), or whether values in both sets are unrelated (covariance near zero).

To return the covariance for individual data point pairs, use the COVAR worksheet function.

Regression

The “Regression” analysis tool performs linear regression analysis by using the "least squares" method to fit a line through a set of observations. You can analyze how a single dependent variable is affected by the values of one or more independent variables. For example, one can analyze how grain yield of barley is affected by factors like ears per plant, ear length (in cms), 100 grain weight (in gms) and number of grains per ear.

Descriptive Statistics

The "Descriptive Statistics" analysis tool generates a report of univariate statistics for data in the input range, which includes information about the central tendency and variability of the entered data.

Sampling

The "Sampling" analysis tool creates a sample from a population by treating the input range as a population. When the population is too large to process or chart, a representative sample can be used. One can also create a sample that contains only values from a particular part of a cycle if you believe that the input data is periodic. For example, if the input range contains quarterly sales figures, sampling with a periodic rate of four places values from the same quarter in the output range.

Random Number Generation

The "Random Number Generation" analysis tool fills a range with independent random numbers drawn from one of several distributions. We can characterize subjects in a population with a probability distribution. For example, you might use a normal distribution to characterize the population of individuals' heights.

ANOVA: Single Factor

"ANOVA: Single Factor" option can be used for analysis of one-way classified data or data obtained from a completely randomized design. In this option, the data is given either in rows or columns such that observations in a row or column belong to one treatment only. Accordingly, define the input data range. Then specify whether, treatments are in rows or columns. Give the identification of upper most left corner cell in output range and click OK. In output, we get replication number of treatments, treatment totals, treatment means and treatment variances. In the ANOVA table besides usual sum of squares, Mean Square, F-calculated and P-value, it also gives the F-value at the pre-defined level of significance.

ANOVA: Two Factors with Replication

This option can be used for analysis of two-way classified data with m-observations per cell or for analysis of data obtained from a factorial CRD with two factors with same or different levels with same replications.

ANOVA: Two Factors without Replication

This option can be utilized for the analysis of two-way classified data with single observation per cell or the data obtained from a randomized complete block design. Suppose that there

are ‘v’ treatments and ‘r’ replications and then prepare a v × r data sheet. Define it in input range, define alpha and output range.

t-Test: Two-Sample Assuming Equal Variances:

This analysis tool performs a two-sample student's t-test. This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic t-test. You can use t-tests to determine whether two sample means are equal. TTEST(array1,array2,tails,type) returns the probability associated with a student’s t test.

t-Test: Two-Sample Assuming Unequal Variances:

This t-test form assumes that the variances of both ranges of data are unequal; it is referred to as a heteroscedastic t-test. Use this test when the groups under study are distinct.

t-Test: Paired Two Sample For Means:

This analysis tool performs a paired two-sample student's t-test to determine whether a sample's means are distinct. This t-test form does not assume that the variances of both populations are equal. One can use this test when there is a natural pairing of observations in the samples, like a sample group is tested twice - before and after an experiment.

F-Test Two-Sample for Variances

The F-Test Two-Sample for Variances analysis tool performs a two-sample F-test to compare two population variances. For example, you can use an F-test to determine whether the time scores in a swimming meet have a difference in variance for samples from two teams. FTEST(array1, array2) returns the result of an F-test, the one tailed probability that the variances of Array1 and array 2 are not significantly different.

Transformation of Data

The validity of analysis of variance depends on certain important assumptions like normality of errors and random effects, independence of errors, homoscedasticity of errors and effects are additive. The analysis is likely to lead to faulty conclusions when some of these assumptions are violated. A very common case of violation is the assumption regarding the constancy of variance of errors. One of the alternatives in such cases is to go for a weighted analysis of variance wherein each observation is weighted by the inverse of its variance. For this, an estimate of the variance of each observation is to be obtained which may not be feasible always. Quite often, the data are subjected to certain scale transformations such that in the transformed scale, the constant variance assumption is realized. Some of such transformations can also correct for departures of observations from normality because unequal variance is many times related to the distribution of the variable also. Major aims of applying transformations are to bring data closer to normal distribution, to reduce relationship between mean and variance, to reduce the influence of outliers, to improve

linearity in regression, to reduce interaction effects, to reduce skewness and kurtosis. Certain methods are available for identifying the transformation needed for any particular data set but one may also resort to certain standard forms of transformations depending on the nature of the data. Most commonly used transformations in the analysis of experimental data are Arcsine, Logarithmic and Square root. These transformations of data can be carried out using the following options.

Arcsine (ASIN): In the case of proportions, derived from frequency data, the observed proportion p can be changed to a new form $\theta = \sin^{-1}(\sqrt{p})$. This type of transformation is known as angular or arcsine transformation. However, when nearly all values in the data lie between 0.3 and 0.7, there is no need for such transformation. It may be noted that the angular transformation is not applicable to proportion or percentage data which are not derived from counts. For example, percentage of marks, percentage of profit, percentage of protein in grains, oil content in seeds, etc., can not be subjected to angular transformation. The angular transformation is not good when the data contain 0 or 1 values for p. The transformation in such cases is improved by replacing 0 with (1/4n) and 1 with [1-(1/4n)], before taking angular values, where *n* is the number of observations based on which p is estimated for each group.

ASIN gives the arcsine of a number. The arcsine is the angle whose sine is number and this number must be from -1 to 1. The returned angle is given in radians in the range $-\pi/2$ to $\pi/2$. To express the arcsine in degrees, multiply the result by $180/\pi$. For this go to the CELL where the transformation is required and write =ASIN (Give Cell identification for which transformation to be done)* 180*7/22 and press ENTER. Then copy it for all observations.

Example: ASIN (0.5) equals 0.5236 ($\pi/6$ radians) and ASIN (0.5)* 180/PI equals 30 (degrees).

Logarithmic (LN): When the data are in whole numbers representing counts with a wide range, the variances of observations within each group are usually proportional to the squares of the group means. For data of this nature, logarithmic transformation is recommended. It squeezes the bigger values and stretches smaller values. A simple plot of group means against the group standard deviation will show linearity in such cases. A good example is data from an experiment involving various types of insecticides. For the effective insecticide, insect counts on the treated experimental unit may be small while for the ineffective ones, the counts may range from 100 to several thousands. When zeros are present in the data, it is advisable to add 1 to each observation before making the transformation. The log transformation is particularly effective in normalizing positively skewed distributions. It is also used to achieve additivity of effects in certain cases.

LN gives the natural logarithm of a positive number. Natural logarithms are based on the constant e (2.718281828845904). For this go the CELL where the transformation is required and write = LN(Give Cell Number for which transformation to be done) and press ENTER. Then copy it for all observations.

Example: LN(86) equals 4.454347, LN(2.7182818) equals 1, LN(EXP(3)) Equals 3 and EXP(LN(4)) equals 4. Further, EXP returns e raised to the power of a given number, LOG returns the logarithm of a number to a specified base and LOG 10 returns the base-10 logarithm of a number.

Square Root (SQRT): If the original observations are brought to square root scale by taking the square root of each observation, it is known as square root transformation. This is appropriate when the variance is proportional to the mean as discernible from a graph of group variances against group means. Linear relationship between mean and variance is commonly observed when the data are in the form of small whole numbers (*e.g.*, counts of wildlings per quadrat, weeds per plot, earthworms per square metre of soil, insects caught in traps, etc.). When the observed values fall within the range of 1 to 10 and especially when zeros are present, the transformation should be, $\sqrt{(y + 0.5)}$.

SQRT gives square root of a positive number. For this go to the CELL where the transformation is required and write = SQRT (Give Cell No. for which transformation to be done = 0.5) and press ENTER. Then copy it for all observations. However, if number is negative, SQRT return the #NUM ! error value.

Example: SQRT(16) equals 4, SQRT(-16) equals #NUM! and SQRT(ABS(-16)) equals 4.

Once the transformation has been made, the analysis is carried out with the transformed data and all the conclusions are drawn in the transformed scale. However, while presenting the results, the means and their standard errors are transformed back into original units. While transforming back into the original units, certain corrections have to be made for the means. In the case of log transformed data, if the mean value is $\bar{y}$, the mean value of the original units will be antilog ($\bar{y}$ + 1.15 $\bar{y}$) instead of antilog ($\bar{y}$). If the square root transformation had been used, then the mean in the original scale would be antilog $((\bar{y} + V(\bar{y}))^2$ instead of $(\bar{y})^2$ where $V(\bar{y})$ represents the variance of $\bar{y}$. No such correction is generally made in the case of angular transformation. The inverse transformation for angular transformation would be $p = (\sin q)^2$.

Sum(SUM): It gives the sum of all the numbers in the list of arguments. For this go to the CELL where the sum of observations is required and write = SUM (define data range for which the sum is required) and press ENTER. Instead of defining the data range, the exact numerical values to be added can also be given in the argument viz. SUM (Number1, number2,…), number1, number2,… are 1 to 30 arguments for which you want the sum.

Example: If cells A2:E2 contain 5, 15,30,40 and 50; SUM(A2:C2) equals 50, SUM(B2:E2,15) equals 150 and SUM(5,15) equals 20.

Some other related functions with this option are:

AVERAGE returns the average of its arguments, PRODUCT multiplies its arguments and SUMPRODUCT returns the sum of the products of corresponding array components.

Sum of Squares (SUMSQ): This gives the sum of the squares of the list of arguments. For this go to the CELL where the sum of squares of observations is required and write = SUMSQ (define data range for which the sum of squares is required) and press ENTER.

Example: If cells A2:E2 contain 5, 15, 30, 40 and 50; SUMSQ(A2:C2) equals 1150 and SUMSQ(3,4) equals 25.

Matrix Multiplication (MMULT): It gives the matrix product of two arrays, say array 1 and array 2. The result is an array with the same number of rows as array1, say a and the same number of columns as array2, say b. For getting this mark the a × b cells on the spread sheet. Write =MMULT (array 1, array 2) and press Control +Shift+ Enter. The number of columns in array1 must be the same as the number of rows in array2, and both arrays must contain only numbers. Array1 and array2 can be given as cell ranges, array constants, or references. If any cells are empty or contain text, or if the number of columns in array1 is different from the number of rows in array2, MMULT returns the ≠VALUE! error value.

Determinant of a Matrix (MDETERM): It gives the value of the determinant associated with the matrix. Write = MDETERM(array) and press Control + Shift + Enter.

Matrix Inverse (MINVERSE): It gives the inverse matrix for the non-singular matrix stored in a square array, say of order p. i.e., an array with equal number of rows and columns. For getting this mark the p × p cells on the spread sheet where the inverse of the array is required and write = MINVERSE(array) and press Control + Shift + Enter. Array can be given as a cell range, such as A1:C3; as an array constant, such as {1,2,3;4,5,6;7,8,8}; or as a name for either of these. If any cells in array are empty or contain text, MINVERSE returns the ≠VALUE! error value.

Example: MINVERSE ({4,-1;2,0}) equals {0,0.5;-1,2}and MINVERSE ({1,2,1;3,4,-1;0,2,0}) equals {0.25, 0.25,-0.75;0,0,0.5;0.75,-0.25,-0.25}.

Transpose (TRANSPOSE): For getting the transpose of an array mark the array and then select copy from the EDIT menu. Go to the left corner of the array where the transpose is required. Select the EDIT menu and then paste special and under paste special select the TRANSPOSE option.

EXERCISES ON MS-EXCEL

1. Table below contains values of pH and organic carbon content observed in soil samples collected from natural forest. Compute mean, median, standard deviation, range and skewness of the data.

2.

Soil pit	pH (x)	Organic carbon (%) (y)		Soil pit	pH (x)	Organic carbon (%) (y)
1	5.7	2.10		9	5.4	2.09
2	6.1	2.17		10	5.9	1.01
3	5.2	1.97		11	5.3	0.89
4	5.7	1.39		12	5.4	1.60
5	5.6	2.26		13	5.1	0.90
6	5.1	1.29		14	5.1	1.01
7	5.8	1.17		15	5.2	1.21
8	5.5	1.14				

3. Consider the following data on various characteristics of a crop:

pp	ph	ngl	yield
142	0.525	8.2	2.47
143	0.64	9.5	4.76
107	0.66	9.3	3.31
78	0.66	7.5	1.97
100	0.46	5.9	1.34
86.5	0.345	6.4	1.14
103.5	0.86	6.4	1.5
155.99	0.33	7.5	2.03
80.88	0.285	8.4	2.54
109.77	0.59	10.6	4.9
61.77	0.265	8.3	2.91
79.11	0.66	11.6	2.76
155.99	0.42	8.1	0.59
61.81	0.34	9.4	0.84
74.5	0.63	8.4	3.87
97	0.705	7.2	4.47
93.14	0.68	6.4	3.31
37.43	0.665	8.4	1.57
36.44	0.275	7.4	0.53
51	0.28	7.4	1.15
104	0.28	9.8	1.08
49	0.49	4.8	1.83
54.66	0.385	5.5	0.76
55.55	0.265	5	0.43

88.44	0.98	5	4.08
99.55	0.645	9.6	2.83
63.99	0.635	5.6	2.57
101.77	0.29	8.2	7.42
138.66	0.72	9.9	2.62
90.22	0.63	8.4	2

(i) Sort yield in ascending order and filter the data ph less than 0.3 or greater than 0.6 from the data.

(ii) Find the correlation coefficient and fit the multiple regression equation by taking yield as dependent variable.

4. Let **A**, **B** and **C** be three matrices as follows:

$$\mathbf{A}=\begin{bmatrix}2&4&6&1&9\\3&5&6&7&2\\8&3&9&1&5\\3&1&1&1&3\end{bmatrix} \quad \mathbf{B}=\begin{bmatrix}1&3\\5&7\\2&4\\1&9\\8&1\end{bmatrix} \quad \mathbf{C}=\begin{bmatrix}2&3&1&8&4\\3&6&7&8&8\\2&3&5&5&7\\2&3&6&6&1\\1&2&8&5&5\end{bmatrix}.$$

Find (i) **AB** (ii) $\mathbf{C}^{-1}$ (iii) $|\mathbf{A}|$ (iv) $\mathbf{A}^{T}$.

5. Draw line graph for the following data on a tree species:

Year	Height (cm)	Diameter
1981	21	5.0
1982	34	8.0
1983	11	9.0
1984	13	3.0
1985	15	2.4
1986	55	5.5
1987	30	6.9
1988	50	9.1
1989	23	10.0
1990	22	2.5
1991	37	3.4
1992	38	6.2
1993	37	7.0
1994	11	8.1
1995	20	9.0
1996	16	3.7
1997	54	9.0
1998	33	4.0
1999	12	6.7
2000	19	7.7

Also draw a bar diagram using the above data.

6. The table below lists plant height in cm of seedlings of rice belonging to the two varieties. Examine whether the two samples are coming from populations having equal variance, using F-test. Further, test whether the average height of the two groups are the same, using appropriate t-test.

Plot	Group I	Group II
1	23.0	8.5
2	17.4	9.6
3	17.0	7.7
4	20.5	10.1
5	22.7	9.7
6	24.0	13.2
7	22.5	10.3
8	22.7	9.1
9	19.4	10.5
10	18.8	7.4

7. Examine whether the average organic carbon content measured from two layers of a set of soil pits from a pasture are same using paired t-test from the data given below:

	Organic carbon (%)	
Soil pit	Layer 1 (x)	Layer 2 (y)
1	1.59	1.21
2	1.39	0.92
3	1.64	1.31
4	1.17	1.52
5	1.27	1.62
6	1.58	0.91
7	1.64	1.23
8	1.53	1.21
9	1.21	1.58
10	1.48	1.18

8. Mycelial growth in terms of diameter of the colony (mm) of *R. solani* isolates on PDA medium after 14 hours of incubation is given in the table below. Carry out the CRD analysis for the data. And draw your inferences.

R. solani isolates	Mycelial growth		
	Repl. 1	Repl. 2	Repl. 3
RS 1	29.0	28.0	29.0
RS 2	33.5	31.5	29.0
RS 3	26.5	30.0	
RS 4	48.5	46.5	49.0
RS 5	34.5	31.0	

9. Following is the data on mean yield in kg per plot of an experiment conducted to compare the performance of 8 treatments using a Randomized Complete Block design with 3 replications. Perform the analysis of variance.

Treatment (Provenance)	**Replication**		
	I	**II**	**III**
1	30.85	38.01	35.10
2	30.24	28.43	35.93
3	30.94	31.64	34.95
4	29.89	29.12	36.75
5	21.52	24.07	20.76
6	25.38	32.14	32.19
7	22.89	19.66	26.92
8	29.44	24.95	37.99

10. From the following data make a summary table for finding out the average of X_9 for various years and various levels of X_6 using pivot table and pivot chart report option of MS-Excel.

YR	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**
1995	1	1	40	30	0	60	40	4861	5208	5556	5694
1995	1	2	40	30	0	60	40	4167	4444	4861	5035
1995	2	3	40	30	0	60	40	4618	4653	4653	5174
1995	2	4	40	30	0	60	40	4028	4167	4514	4722
1995	2	5	40	30	0	60	40	4306	4514	4653	4861
1996	2	1	40	30	0	60	40	6000	5750	5499	6250
1996	2	2	40	30	0	60	40	5646	5000	5250	5444
1996	2	3	40	30	0	60	40	4799	5097	4896	5299
1996	2	4	40	30	0	60	40	5250	5299	4194	4847
1996	3	1	40	30	0	60	40	5139	5417	5764	5903
1996	3	2	40	30	0	60	40	5417	5694	6007	6111
1996	4	1	40	30	0	60	40	6300	7450	7750	8000
1996	4	2	40	30	0	60	40	6350	7850	7988	8200
1996	4	3	40	30	0	60	40	5750	6400	6600	6700
1996	4	4	40	30	0	60	40	6000	7250	7450	7681
1996	5	1	40	30	0	60	40	3396	4090	5056	5403
1996	5	2	40	30	0	60	40	5194	5000	6000	6500
1996	5	3	40	30	0	60	40	4299	4250	4750	5250
1996	6	1	40	30	0	60	40	4944	5194	5000	5097
1996	6	2	40	30	0	60	40	5395	5499	5499	5597
1996	6	3	40	30	0	60	40	3444	5646	5000	5000
1996	6	4	40	30	0	60	40	6250	6500	6646	6750
1997	1	1	120	30	30	120	60	5839	6248	6199	6335

1997	1	2	120	30	30	120	60	5590	5652	5702	5851
1997	2	1	120	30	30	120	60	4497	4794	4894	5205
1997	2	2	120	30	30	120	60	4696	5006	5304	5702
1997	2	3	120	30	30	120	60	4398	4596	4894	5304
1997	2	4	120	30	30	120	60	4497	5503	5702	6099
1997	3	1	120	30	30	120	60	4199	5602	5801	6000
1997	3	2	120	30	30	120	60	3404	3901	4199	4497
1997	3	3	120	30	30	120	60	3602	5404	5503	5801
1997	3	4	120	30	30	120	60	3602	4297	4497	4696
1997	4	1	120	30	30	120	60	3205	3801	4199	4894
1997	4	2	120	30	30	120	60	3801	4794	6099	6298
1997	4	3	120	30	30	120	60	3503	5205	6298	6795
1997	4	4	120	30	30	120	60	3205	4894	5503	6199
1997	5	1	120	30	30	120	60	4199	4099	4199	4297
1997	5	2	120	30	30	120	60	3304	3702	3602	3801
1997	5	3	120	30	30	120	60	2596	2894	3106	3205
1998	1	1	40	30	0	60	40	3727	3106	3404	3503
1998	1	2	40	30	0	60	40	4894	4348	4447	4534
1998	1	3	40	30	0	60	40	2696	2795	3056	3205
1998	2	2	40	30	0	60	40	5503	4298	4497	4795
1998	2	3	40	30	0	60	40	5006	3702	3702	3901

11. From the data given in problem 10, sort X_{10} in ascending order. Also, filter the data for $X_{11} < 4200$ or $X_{11} > 5000$.

HANDS-ON EXERCISE ON SAMPLING SCHEMES USING MS EXCEL

Bharti

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Sample surveys are a cost-effective method for data collection, allowing for accurate and reliable inferences about population parameters. These surveys involve selecting a representative subset of the population, from which conclusions about the entire target population can be drawn. Analyzing survey data is crucial for extracting meaningful insights from collected responses. Microsoft Excel offers a range of powerful tools for cleaning, organizing, and analyzing survey data efficiently. Key advantages of using Excel for survey analysis include:

- Easy data entry and organization
- Built-in statistical and analytical functions
- Visual representation through charts and graphs
- PivotTables for summarizing large datasets

2. Importing and Organizing Survey Data

- Importing Data
  - If survey data is in CSV or Excel format, open the file in Excel
  - If using Google Forms, download responses as a CSV file and open it in Excel

- Cleaning Data
  - Remove blank rows/columns
  - Ensure uniform responses (e.g., "Yes" vs. "yes")
  - Remove duplicates. Use Data > Remove Duplicates
  - Handle missing data

- Formatting Data for Analysis
  - Convert responses into numerical values where necessary (e.g., Yes = 1, No = 0)
  - Use Data Validation for consistency in data entry

3. Exploratory Data Analysis (EDA) of Survey Data

Basic statistical summaries help understand the central tendency and dispersion of survey data. Excel functions to compute these include:

3.1 Basic Statistical Functions:

- **Count**: =COUNT(cell range) for numerical data and =COUNTA(range) for categorical data
- **Minimum and Maximum**: =MIN(cell range), =MAX(cell range)
- **Mean (Average)**: =AVERAGE(cell range)
- **Median**: =MEDIAN(cell range)

- **Mode**: =MODE(cell range)
- **Standard Deviation**: =STDEV.P(range) or =STDEV.S(cell range)
- **Variance**: =VAR.P(cell range) or =VAR.S(cell range)

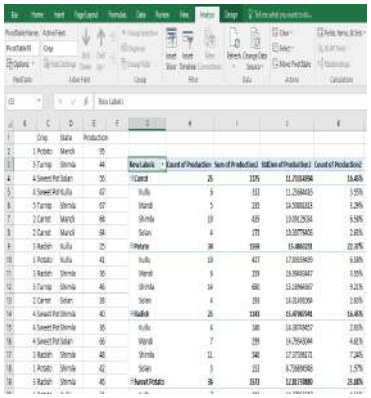
	Formula		
Sum	SUM(G2:G20)	COUNT	COUNT(G2:G20)
Count	COUNT(G2:G20)		COUNTA(G2:G20)
			COUNTBLANK(G2:G20)
			COUNTIF(G2:G20,">50")
Measure of Central Tendency			COUNTIFS(G2:G20,">50",G2:G20,"<70")
Mean	AVERAGE(G2:G20)		
Median	MEDIAN(G2:G20)	AVERAGE	AVERAGE(G2:G20)
Mode	MODE(G2:G20)		AVERAGEA(G2:G20)
GM	GEOMEAN(G2:G20)		AVERAGEIF(G2:G20,">45")
HM	HARMEAN(G2:G20)		AVERAGEIFS(G2:G20,F2:F20,"Onion")
Measure of Dispersion			
Standard Deviation	STDEV.P(G2:G20)		
Variance	VAR.P(G2:G20)		
Average Deviation	AVEDEV(G2:G20)		
Skewness	SKEW(G2:G20)		
Kurtosis	KURT(G2:G20)		

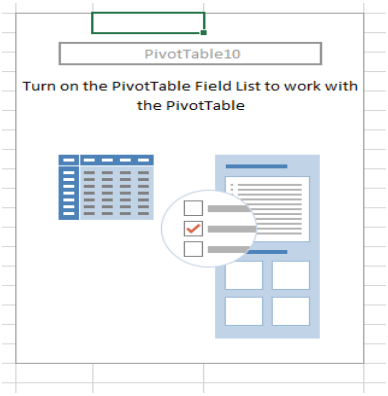
3.2 Data Visualization: Visualizing survey data helps identify trends and relationships. Some key chart types in Excel include:

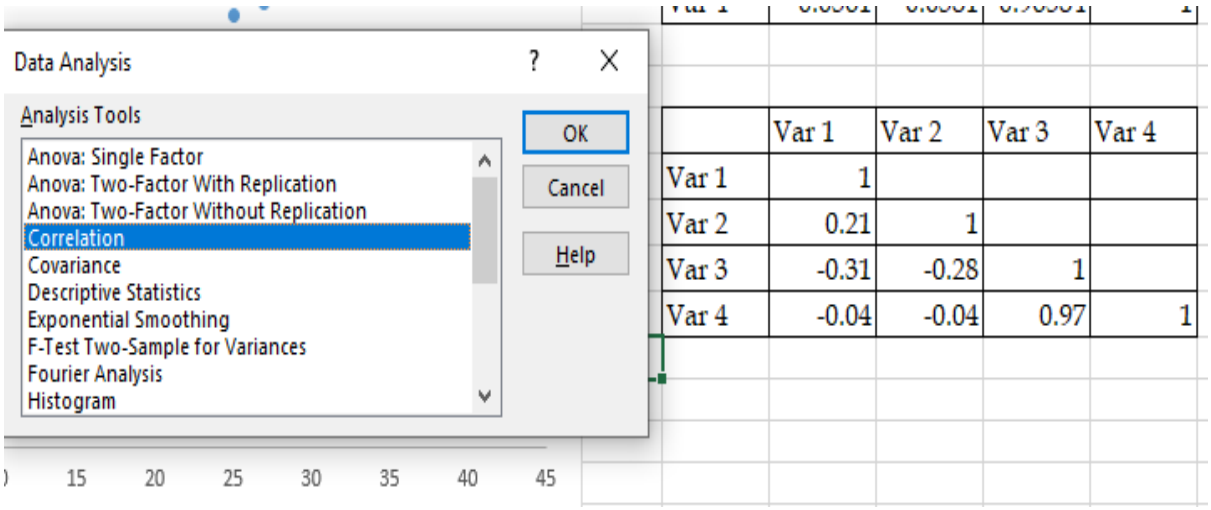
- **Bar Charts**: Useful for categorical data comparisons (Insert > Bar Chart)
- **Histograms**: Shows frequency distributions (Insert > Histogram)
- **Pie Charts**: Displays proportions (Insert > Pie Chart)
- **Box Plots**: Highlights outliers and spread (Insert > Box and Whisker)
- **Scatter Plots**: Shows relationships between numerical variables (Insert > Scatter Plot)
- **Pivot Charts**: Dynamic visual representation of aggregated data (Insert > PivotChart)

Pivot Tables for Survey Data Exploration: Pivot tables are one of the most powerful tools in MS Excel for summarizing and analyzing survey data. Steps to create a pivot table includes:

- Select the dataset and go to Insert > PivotTable.
- Choose whether to place the PivotTable in a new or existing worksheet.
- Drag fields into the Rows, Columns, Values, and Filters sections.
- Use Value Field Settings to apply functions like Count, Sum, Average, etc.
- Apply filters and sorting for deeper insights.







	Var 1	Var 2	Var 3	Var 4
Var 1	1			
Var 2	0.21	1		
Var 3	-0.31	-0.28	1	
Var 4	-0.04	-0.04	0.97	1

3.3 Identifying Trends and Patterns

- Use **Conditional Formatting** (Home > Conditional Formatting) to highlight trends.
- Apply **Filters and Sorting** to isolate specific groups.
- Utilize **Moving Averages** (=AVERAGE(cell range)) for trend analysis.
- Analyze **Correlations** using =CORREL(cell range1, cell range2) to measure relationships between numerical survey responses.

3.4 Detecting Outliers: Outliers can distort results and should be detected and handled appropriately.

- Use **Box Plots** to visually identify outliers.
- Apply the **Interquartile Range (IQR) method**:
  - Find Q1 (=QUARTILE.INC(range,1)) and Q3 (=QUARTILE.INC(range,3)).
  - Compute IQR: =Q3 - Q1.
  - Identify outliers as values below Q1 - 1.5*IQR or above Q3 + 1.5*IQR.
- Highlight outliers using **Conditional Formatting**.

3.5 **Descriptive Statistics Using Data Analysis ToolPak:** The Data Analysis ToolPak is an Excel add-in that provides a collection of analysis tools, including statistical, financial analysis, etc. This add-in simplifies tasks that would otherwise require complex formulas or external tools. It's especially useful for analysts, researchers, and students who need to conduct quick data processing, hypothesis testing, and advanced statistical analysis. Go to File > Options > Add-ins > Analysis ToolPak > Enable

Overview of the Data Analysis Tools: The ToolPak includes a variety of tools for:

- **Descriptive Statistics**: Measures of central tendency, dispersion, and distribution.
- **Regression Analysis**: Linear regression, multiple regression, and other statistical models.

- **ANOVA**: One-way and two-way analysis of variance.
- **Correlation**: Calculating the correlation coefficient between two variables.
- **t-Test**: Paired and two-sample tests for hypothesis testing.
- **Histograms**: Creating and analyzing frequency distributions.
- **F-Test**: Testing equality of variances between two datasets.
- **Moving Averages**: Smoothing out data to identify trends.
- **Sampling**: Creating random samples and conducting analysis on them.

Production	
Mean	46.54605
Standard Error	1.097378
Median	49
Mode	51
Standard Deviation	13.52938
Sample Variance	183.0442
Kurtosis	-1.24167
Skewness	-0.22964
Range	44
Minimum	23
Maximum	67
Sum	7075
Count	152
Largest(1)	67
Smallest(1)	23
Confidence Level(95.0%)	2.168198

4. **Sampling Schemes:** Sampling is the process of selecting a subset of individuals from a population to estimate characteristics of the entire population. Sampling schemes can be broadly categorized into two types, such as: Probability sampling (Simple Random Sampling, Systematic Sampling, Stratified Sampling, Cluster Sampling, Multi Stage Sampling etc.) and Non Probability Sampling (Quota Sampling, Judgement Sampling, Snow Ball Sampling etc.)

   4.1 Simple Random Sampling (SRS): In Simple Random Sampling (SRS), each element of the population has an equal chance of being selected. Steps involved to select simple random sample in MS Excel:

   - Create a list of items (or population). For example, assume a population of 100 people, numbered from 1 to 100 in column A.
   - In the next column (Column B), generate random numbers using function =RANDBETWEEN() function, which generates a random number between given range.
     - **Formula**: =RANDBETWEEN()
     - Drag down the formula to select the units in the sample.

   4.2 Systematic Sampling: In Systematic Sampling, first unit is selected with the help of random numbers, and the remaining units are selected automatically according to a predetermined pattern. The systematic sampling technique is operationally more

convenient than simple random sampling. It also ensures, at the same time that each unit has an equal probability of inclusion in the sample. Suppose there are N units in a population, which are numbered from 1 to N. Let N = nk, (n = sample size and k = an integer), $k=N/n$ and if a random number less than or equal to k is selected and every kth unit thereafter. The resulting samples are called kth systematic sampling and such a process is called Linear Systematic sampling. Steps involved to select systematic sample in MS Excel:

- Prepare the data Like in SRS, create a population list (1 to 100)
- Choose the Sampling Interval (k):
  Decide the sample size (for example, 10) and the population size (100). The sampling interval k is calculated as: N/n = 100/ 10 = 10. So, the interval is every 10th person.
- Select the First Element Randomly: Use the =RANDBETWEEN() function to select a random starting point between 1 and k, for example, if the random number between 1 and 10 is 3, then the starting point is the 3rd person.
- Select Every kth Person: From the randomly chosen starting point, select every 10th person in the list. If the random starting point is 4, select the 4th, 14th, 24th, and so on.

4.3 Stratified Sampling

In stratified random sampling, population is divided into non-overlapping groups based on specific characteristics, such as age, gender, income level, or education. Each group is called a stratum. After dividing the population into homogeneous strata, a random sample is taken from each stratum. The sample size taken from each group can be proportional to the stratum size or equal across all strata, depending on the objective. Then, these samples are used to give conclusion about the whole target population. By ensuring that each subgroup is represented, stratified random sampling helps reduce sampling error and increases the precision of the overall estimate. Steps involved to select stratified random sample in MS Excel:

- Prepare the data: Create a list of individuals and assign them to different homogeneous strata.
- Sample within Each Stratum: For each subgroup, randomly select a subset of individuals. The allocation of sample size within each stratum can be done using different methods depending on the research objectives and the relative importance of each stratum. The common methods for allocating the sample size across strata are equal allocation, proportional allocation, optimum allocation.
- Combine the analysis to get final results.

4.4 Cluster Sampling: In Cluster Sampling, the population is divided into clusters (groups), and entire clusters are randomly selected for inclusion in the sample. Steps involved to select cluster in MS Excel:

- Prepare the data: Divide the population into clusters. For example, group individuals by location or department.
- Select Clusters Randomly: Use a random number generator to select a certain number of clusters.
- Include All Members of the Selected Clusters: Once clusters are selected, include all individuals within those clusters in the final sample. If there are 10 clusters, randomly select 2 clusters and include all individuals in those clusters.

5. Conclusion

This chapter explored how to implement various sampling schemes in MS Excel, including Simple Random Sampling, Systematic Sampling, Stratified Sampling, and Cluster Sampling. Excel's built-in functions, such as =RAND(), sorting tools, and basic arithmetic operations, make it an excellent tool for performing sampling in both straightforward and more complex scenarios. The flexibility of Excel allows users to efficiently manage and analyze data, providing a practical solution for a wide range of sampling techniques.

R SOFTWARE AN OVERVIEW

Kaustav Aditya and Hukum Chandra

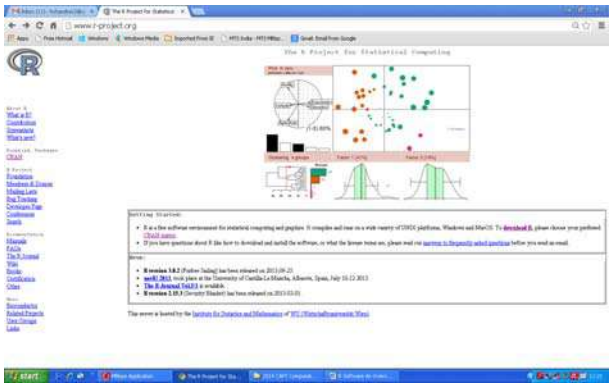
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

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 down load from **http://www.r-project.org**/. The primary purpose of this lecture is to introduce R.

2. To Download R Software

- In any web browser (e.g. Microsoft Internet Explorer), go to: **http://www.r-project.org**



- Downloads: **CRAN**

- 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-3.0.2.exe** (52 megabytes, 32/64 bit) and save it to your hard disc.

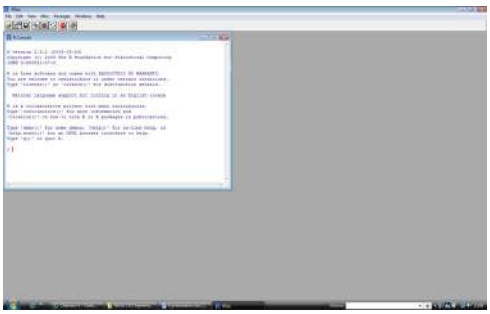
- This is the latest available version of the software. It is an '.exe' file, which you can save in your hard disc. By double clicking on the name of this file, R is automatically installed. All you need to do is follow the installation process.

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.

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.



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.

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

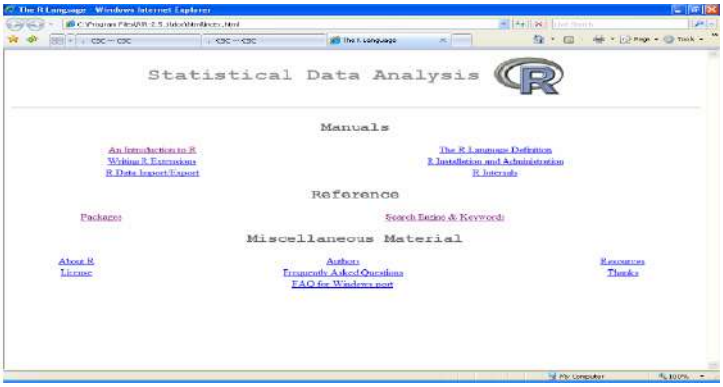
Click on the down-arrow at the top of the "Open" dialog box and change the selection to "Look in" **C:**. You should now see the filename **Intro_to_R.txt** among a list of files. Double-click on the filename to open it.

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:

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.



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  xbar= mean(weight)  then
sqrt(sum((weight- xbar)^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

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)
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.

5. 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. R is capable of creating high quality graphics. Graphs are typically created using a series of high-level and low-level plotting commands. High-level functions create new plots and low-level functions add information to an existing plot. Customize graphs (line style, symbols, color, etc) by specifying graphical parameters. Specify graphic options using the par() function. The function par() is used to set or get graphical parameters. This function contains 70 possible settings and allows you to adjust almost any feature of a graph. Graphic parameters are reset to the defaults with each new graphic

device. Most elements of par() can be set as additional arguments to a plot command, however there are some that can only be set by a call to par(), mfrow, mfcol see the documentation for others.

Scatterplot And Line Graphs

Scatter plots: are useful for studying dependencies between variables.

- The **plot()** function is used for producing scatterplots and line graphs

See ? **plot**

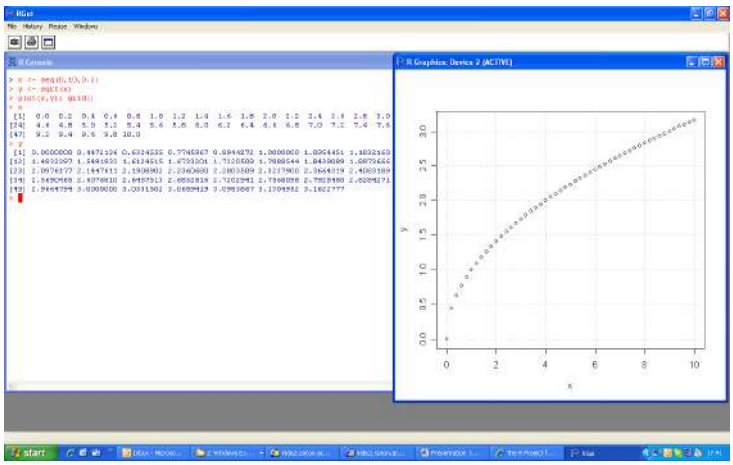
- Using the **plot command**

  x <- seq(0,10,0.2)

  y <- sqrt(x)

  plot(x,y); grid()

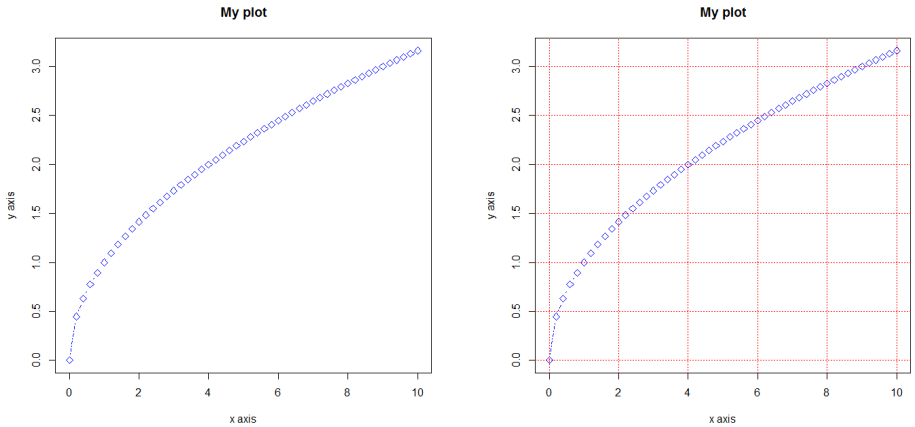
- As one might guess, the last command adds a grid to the plot.



plot(x,y); grid()

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")

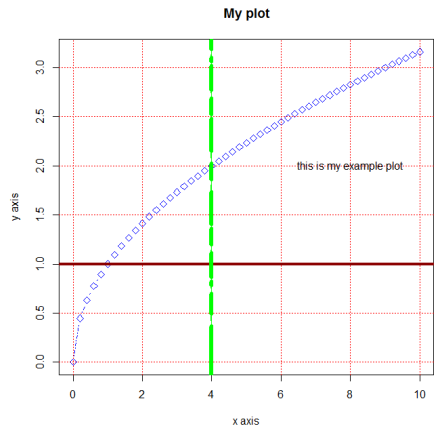


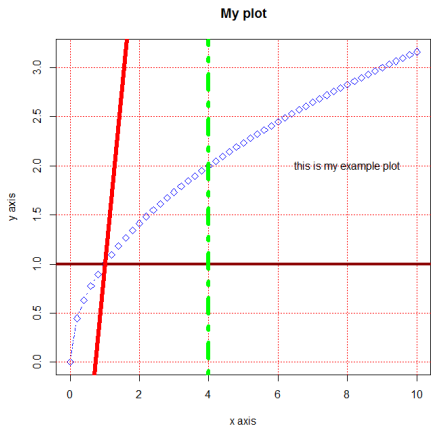
Common arguments for plot()

type	1-character string denoting the plot type

xlim	x limits, c(x1, x2)
ylim	y limits, c(y1, y2)
main	Main title for the plot
sub	Sub title for the plot
xlab	x-axis label
ylab	y-axis label
col	Color for lines and points
pch	Number referencing a plotting symbol or a character string
cex	A number giving the character expansion of the plot symbols
lty	Number referencing a line type
lwd	Line width

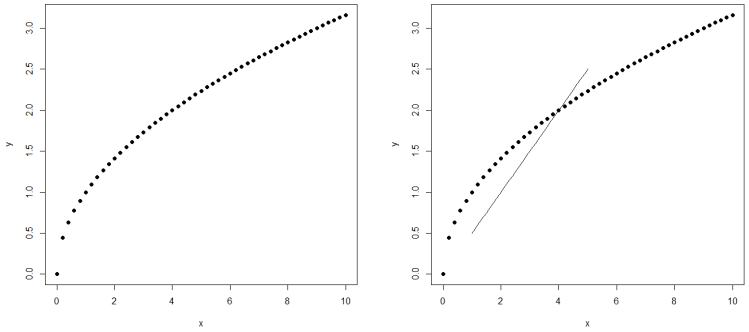
```
plot(x,y,type="b",col="blue",lwd=1,lty=4,pch=5, main="My plot", xlab="x axis", ylab="y axis")
grid(col="red")
text(8,2,"this is my example plot")
abline(h=1,v=4, col=c("darkred","green"), lty=c(1,4), lwd=c(4,6))
reg.lm=lm(x~y)
abline(reg.lm, col="red",lwd=6)                    #To add the regression line
```





There is wealth of plotting parameters you can set

```
plot(x,y)
plot(x,y, pch=16) : plot with new mark with dark circle
x1<- seq(1,5,0.1)
lines(x1,.5*x1)  #lines will add (x,y) values
```



dx<- rnorm(20,5,5) ## generate 100 random number from standard normal distribution

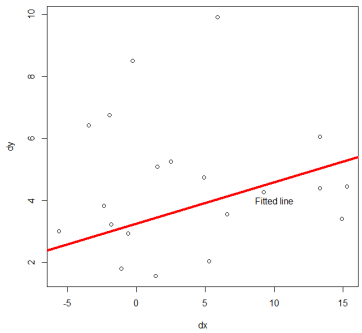
dy<- rchisq(20,5) ## generate 100 random number from chisq distribution with mean 5

plot(dx,dy,pch=1)

fit<-lm(dx~dy)

abline(fit,col="red",lwd=4)

text(10,4,"Fitted line")



See ? plot

See ? points

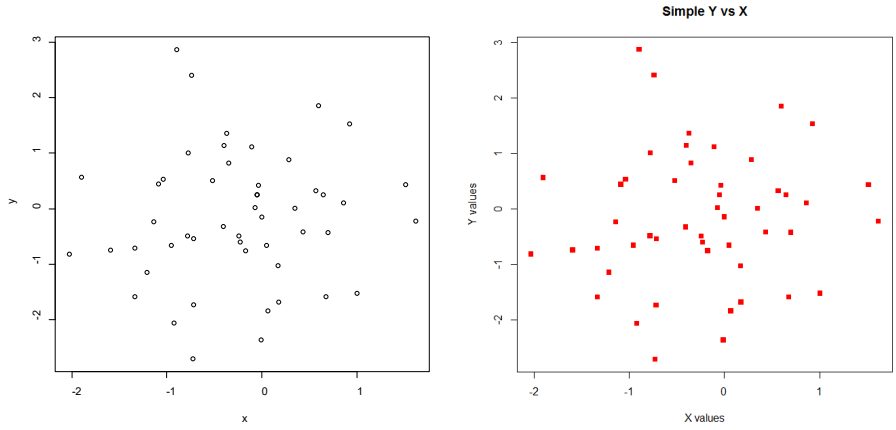
x <- rnorm(50) ;y <- rnorm(50)

group <- rbinom(50, size=1, prob=.5)

Basic Scatterplot

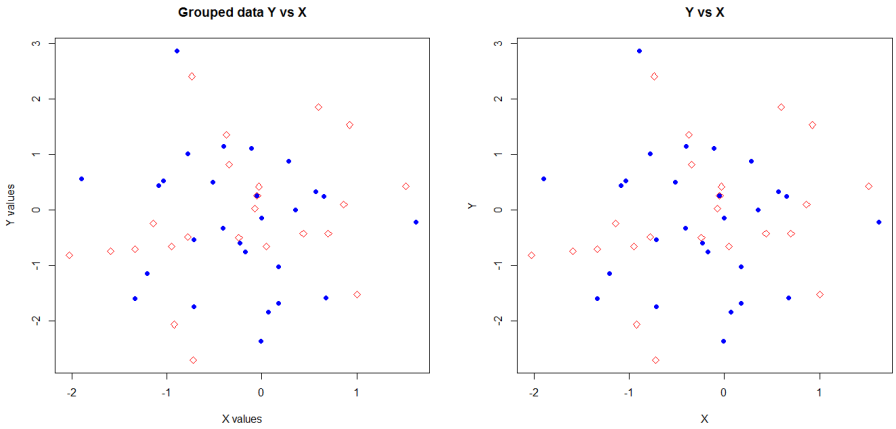
plot(x, y)

plot(x, y, xlab="X values", ylab="Y values", main="Simple Y vs X", pch=15, col="red")



Distinguish between two separate groups

plot(x, y, xlab="X values", ylab="Y values", main="Grouped data Y vs X", pch=ifelse(group==1, 5, 19), col=ifelse(group==1, "red", "blue"))



plot(x, y, xlab="X", ylab="Y", main="Y vs X", type="n")

points(x[group==1], y[group==1], pch=5, col="red")

points(x[group==0], y[group==0], pch=19, col="blue")

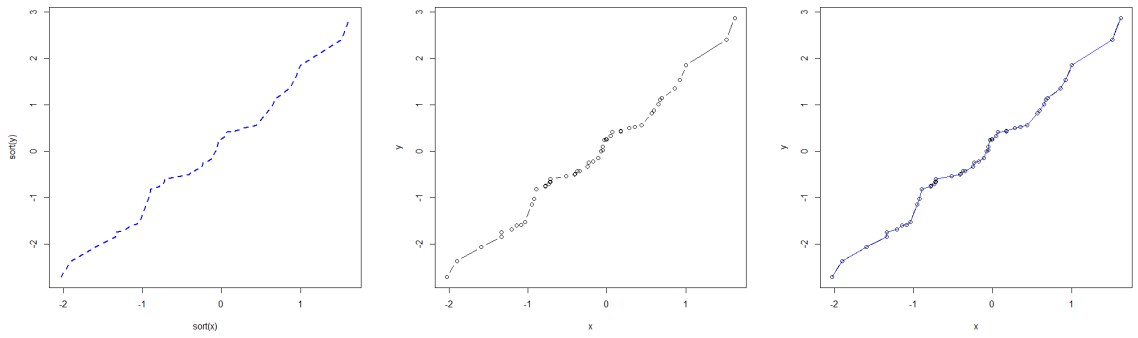
plot(x, y, xlab="X", ylab="Y", main="Y vs X", type="n")

points(cbind(x,y)[group==1,], pch=5, col="red")

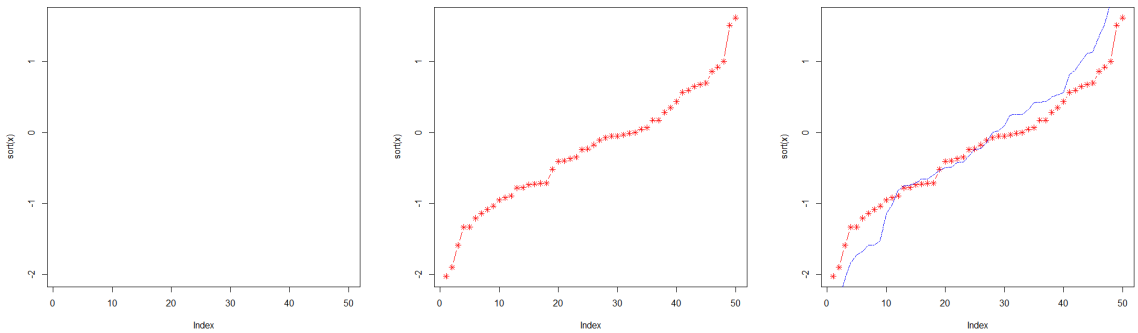
points(cbind(x,y)[group==0,], pch=19, col="blue")

Line Graphs

plot(sort(x), sort(y), type="l", lty=2, lwd=2, col="blue")




```
plot(x, y, type="n")
lines(sort(x), sort(y), type="b")
lines(cbind(sort(x),sort(y)), type="l", lty=1, col="blue")
plot(sort(x), type="n")
lines(sort(x), type="b", pch=8, col="red")
lines(sort(y), type="l", lty=6, col="blue")
```

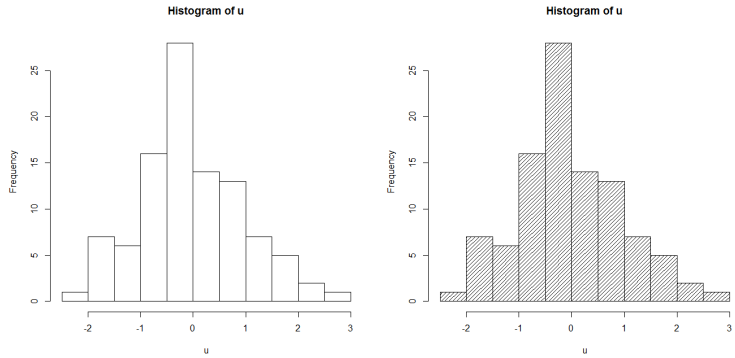


Histogram

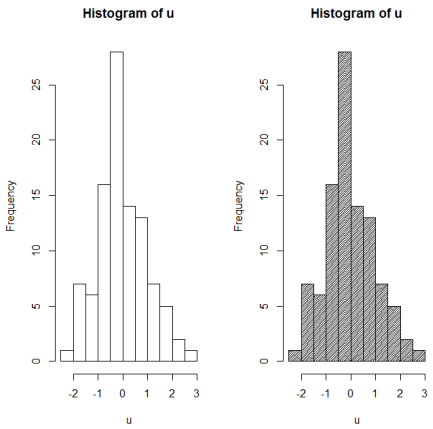
Histograms: used to study the distribution of continuous data, use command **hist**.

hist: function to plot histogram

```
u<- rnorm(100)        # generate 100 random numbers from SND
hist(u)               #default histogram
hist(u, density=20)   #with shading
```

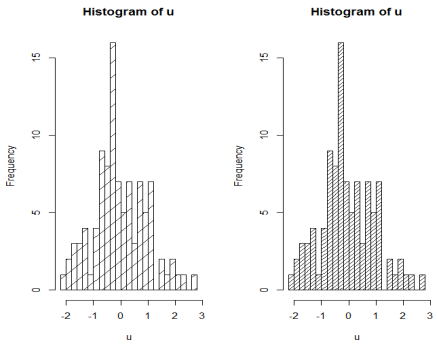


The sequence of commands below plots two histograms in one window
par(mfrow=c(a,b)) gives a rows with b plots on each row. Try
par(mfrow=c(1,2)); hist(u);hist(u, density=50)



#with specific number of bins

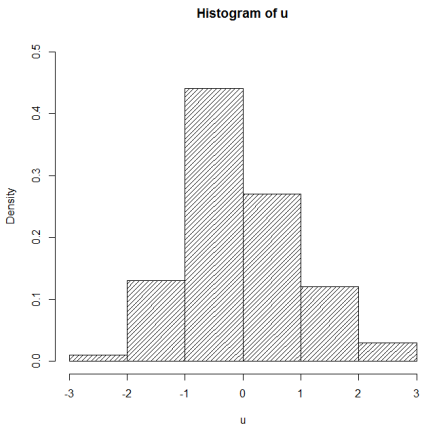
par(mfrow=c(1,2)); hist(u, density=5, breaks=20); hist(u, density=20, breaks=20)



Read in the help file about hist- **help(hist)**

Proportion, instead of frequency also specifying y-axis

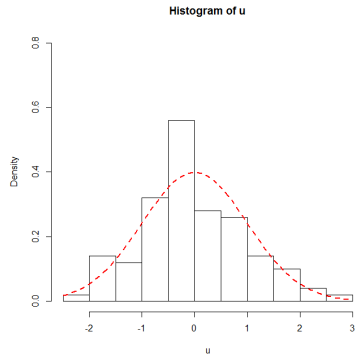
hist(u, density=20, breaks=-3:3, ylim=c(0,.5), prob=TRUE)



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

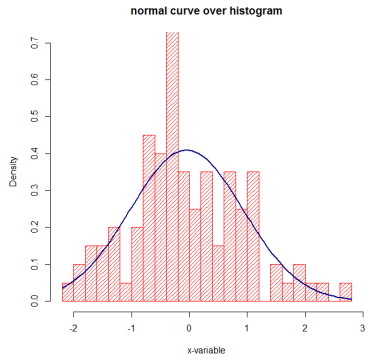
colored normal curve

Uses the observed mean and standard deviation for plotting the normal curve

m<-mean(u) ;std<-sqrt(var(u))

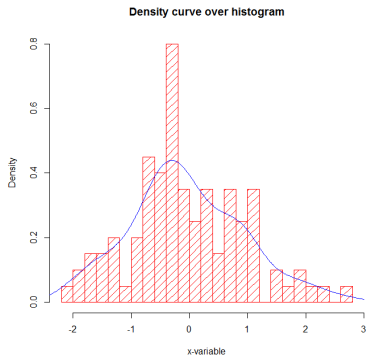
hist(u, density=20, breaks=20, prob=TRUE, xlab="x-variable", col="red", ylim=c(0, 0.7), main="normal curve over histogram")

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")

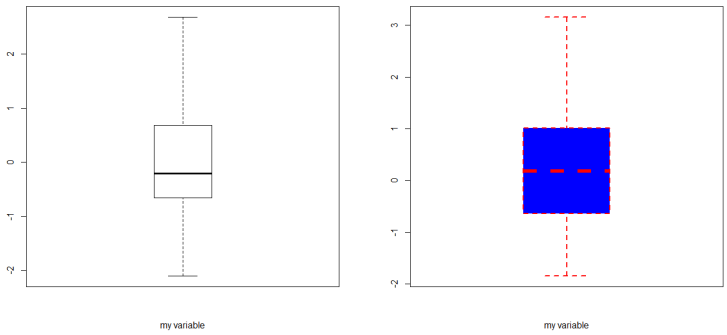


Boxplots

Boxplots: 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) # Basic boxplot

boxplot(u, xlab="my variable", boxwex=.6, col="blue", border= "red", lty=2, lwd=2)



we create data: three variables

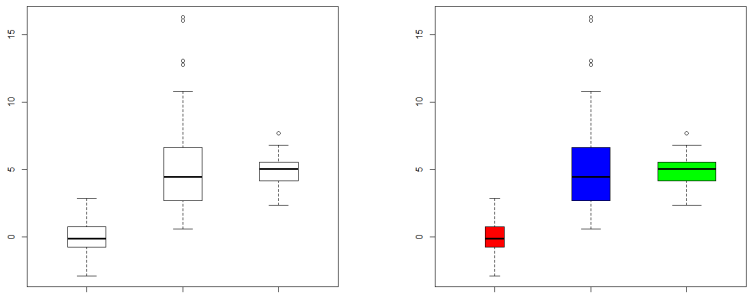
u1<- rnorm(100) ## generate 100 random number from standard normal distribution

u2<- rchisq(100,5) ## generate 100 random number from chisq distribution with mean 5

u3<- rnorm(100,5,1) ## generate 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"))

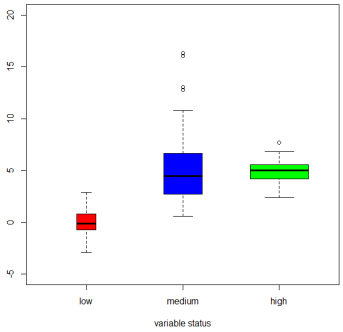


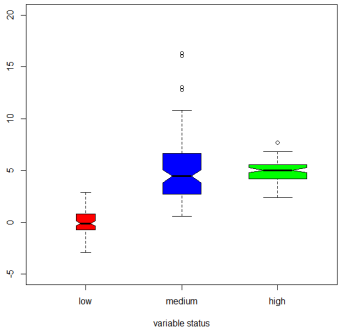
variablename<-c("low","medium", "high")

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

Barchart (Or Barplot)

The R command is barplot

MPCE <- c(400, 300,600,550,425)

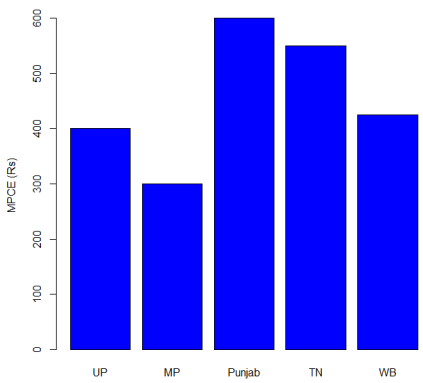
Suppose data in MPCE are average MPCE of some states whose names are to be assigned against their value. Following commands are required:

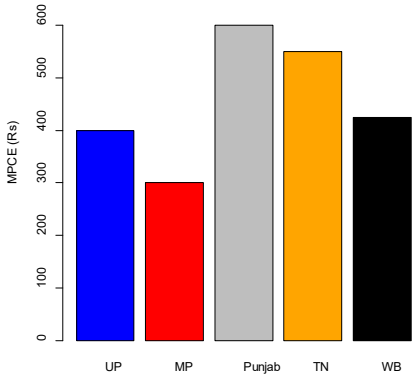
names(MPCE)<-c("UP","MP","Punjab","TN","WB")

To assign names of states. Double quotation mark " " means that names are characters not numeric.

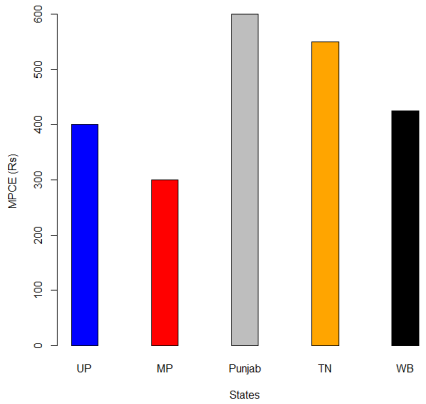
barplot(MPCE, names=names(MPCE), ylab="MPCE (Rs)",col="blue")

barplot(MPCE, names=names(MPCE),ylab="MPCE (Rs)", col = c("blue","red","gray","orange","black"))





barplot(MPCE, space=2,names=names(MPCE),xlab="States", ylab="MPCE (Rs)", col = c("blue","red","gray","orange","black"))

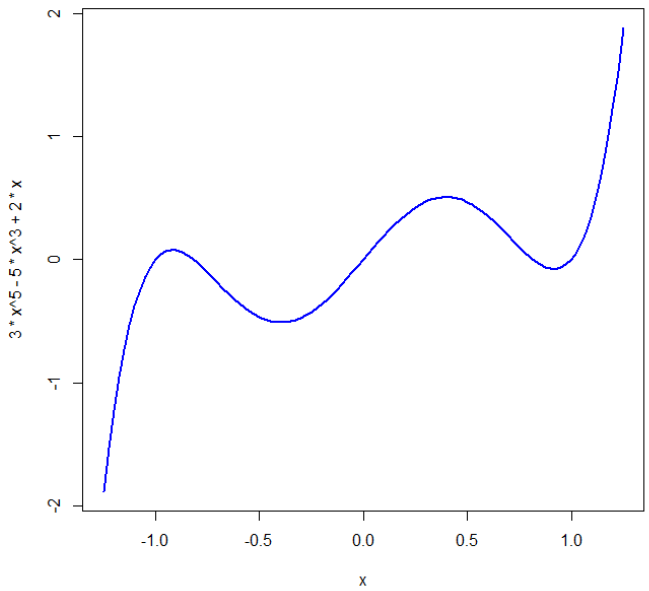


?barplot

Curve

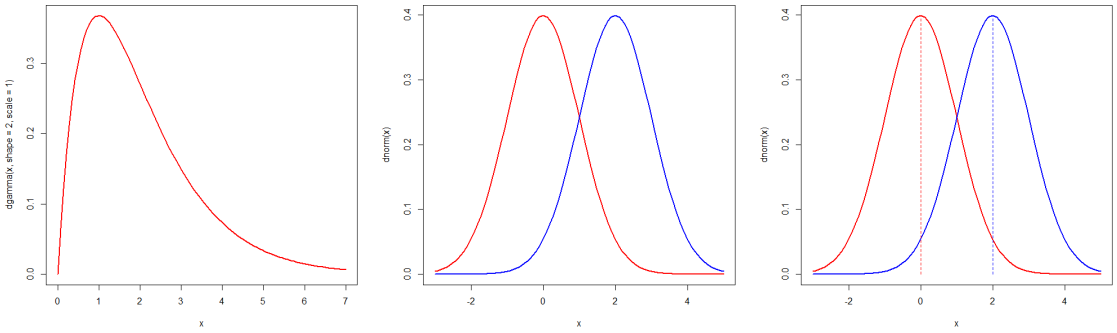
- The function **curve()** draws a curve corresponding to a given function
- If the function is written within curve() it needs to be a function of x
- If you want to use a multiple argument function, use x for the argument you wish to plot over

```
# Plot a 5th order polynomial
curve(3*x^5-5*x^3+2*x, from=-1.25, to=1.25, lwd=2, col="blue")
```




```
# Plot the gamma density
  curve(dgamma(x, shape=2, scale=1), from=0, to=7, lwd=2, col="red")
# Plot multiple curves, notice that the first curve determines the x-axis
curve(dnorm, from=-3, to=5, lwd=2, col="red")
curve(dnorm(x, mean=2), lwd=2, col="blue", add=TRUE)
```

```
# Add vertical lines at the means
lines(c(0, 0), c(0, dnorm(0)), lty=2, col="red")
lines(c(2, 2), c(0, dnorm(2, mean=2)), lty=2, col="blue")
```



Saving Graphs

- Graphs can be saved using several different formats, such as PDFs, JPEGs, and BMPs, by using pdf(), jpeg() and bmp(), respectively
- Graphs are saved to the current working directory

  Save graphics by choosing File -> Save as
- # Create a single pdf of figures, with one graph on each page

  pdf("SavingExample.pdf", width=7, height=5) # Start graphics device

  pdf("C://SavingExample.pdf", width=7, height=5)

```
x <- rnorm(100)
hist(x, main="Histogram of X")
plot(x, main="Scatterplot of X")
dev.off()                          # Stop graphics device
```

 # Create multiple pdfs of figures, with one pdf per figure

```
pdf(width=7, height=5, onefile=FALSE)
x <- rnorm(100)
hist(x, main="Histogram of X")
plot(x, main="Scatterplot of X")
dev.off()                          # Stop graphics device
```

6. Packages

- **Packages** are collections of **R** functions, data, and compiled code in a well-defined format. The directory where packages are stored is called the **library**
- The base distribution 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 etc
- Packages not included in the base distribution can be downloaded and installed directly from R prompt
- Once installed, they have to be loaded into the session to be used
- Currently, the CRAN package repository has **4348 packages**
- **library()** **# To see all installed packages**
- **help("INSTALL")** or **help("install.packages")** in R for information on how to install packages from this repository

Adding Packages

- Choose **Install Packages** from the **Packages** menu
- Select a **CRAN Mirror**
- Select a package (e.g. car)
- Then use the **library(*package*)** function to load it for use (e.g. library(car))

7. 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]
```

8. Reading Data

Reading data from files

There are a few principal functions reading data into R

- read.table, read.csv, for reading tabular data
- readLines, for reading lines of a text file
- source, for reading in R code files (inverse of dump)
- load, for reading in saved workspaces

read.csv(file, header = TRUE, sep = ",", quote="\"", dec=".", fill = TRUE, comment.char="", ...)

Specify the package

library (MASS)

Set working directory

setwd("G:/Course")

Reading ASCII Format

mydata=read.table("G:/Course/yielddata.txt")

dim(mydata)

summary(mydata)

mydata=read.table("G:/Course /yielddata.txt",header=T)

dim(mydata)

summary(mydata)

names(mydata)

mydata=read.table(file="yielddata.txt",header=T)

dim(mydata)

summary(mydata)

names(mydata)

```
 [1] "Dist"          "Yield"         "MARG_HH_F"     "HH_SIZE"
 [5] "NetArea"      "Cropedarea"    "Netirrig"      "GrossIrrigated"
 [9] "Rainfall"      "Fert"
```

mydata1=data.frame(mydata)

mydata1$Yield

mydata1$Fert

Extract district with yield less than median yield

mydata1$Yield[mydata1$Yield<median(mydata1$Yield)]

Extract data with yield less than median yield

mydata2=mydata1[mydata1$Yield<median(mydata1$Yield)]

mydata2=mydata1[mydata1$Yield<median(mydata1$Yield),]

dim(mydata2)

Read Data (Execl)

Call the require library

Load package XLConnect

The XLConnect package is part of the Comprehensive R Archive Network (CRAN). It can be easily installed by using the install.packages() command in your R session

install.packages ("XLConnect")

To load the package, use the library() or require() command in your R session

loadWorkbook() - loading/creating an Excel workbook

The loadWorkbook() function loads a Microsoft Excel workbook, so that it can then be further manipulated. Setting the create argument to TRUE will ensure the file will be created, if it does not exist yet. Both .xls and .xlsx file formats can be used.

```
loadWorkbook (filename , create = TRUE )
library(XLConnect)
library (MASS)
mydata2=loadWorkbook(file="yielddata.xls", create = TRUE)
readWorksheet(mydata3,sheet="yielddata",,header=T)
```

READING DATA IN OTHER FORMAT

library (foreign)

Read SPSS Dataset

MySpssdata=read.spss(file="yielddata.sav", use.value.labels=True, to.data.frame=True)

Read STAT Dataset

MyStatdata=read.dta(file="yielddata.dta")

Writing Data From Files

write.table(Result, ""'MyResults.txt ")

```
write(Results,"MyResults2.txt")
write(Results,"MyResults2.txt",ncolumns=2)
```

How to save R workshop

```
save.image("myworshop.RData")
```

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 _Motor Trend_ 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. _Biometrics_, *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 desribe?

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 transmisson?

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)

> par(mfrow=c(1,2)); hist(mtcars$mpg); hist(mtcars$wt)

> 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)

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)

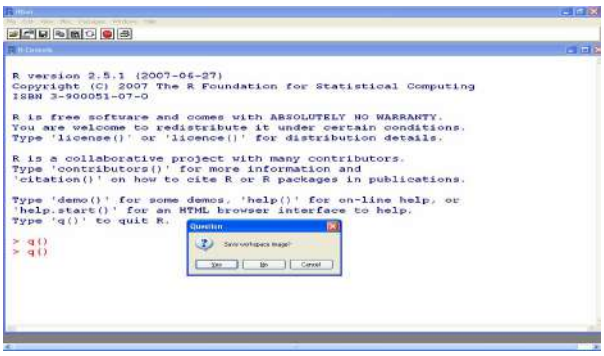
Linear regression

>lm(mtcars$wt~mtcars$mpg)

Try to see help (lm)

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.

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 (2012). R: A language and environment for statistical computing.
- R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

DATA VISUALIZATION USING R

Bharti

ICAR-Indian Agricultural Statistics Research Institute, New Delhi - 110012

1. Introduction:

Data visualization is the graphical representation of data that turns raw data into clear and meaningful visuals. These visuals like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Important principles for effective data visualization include keeping it simple, accurate, relevant, consistent, and interactive. Key benefits of data visualization include:

- Visual representations of data are often easier to comprehend than raw numbers.
- Identifying trends, outliers, and correlations is easier with visual tools.
- Well-designed visualizations help convey complex data to others in an easily digestible format.

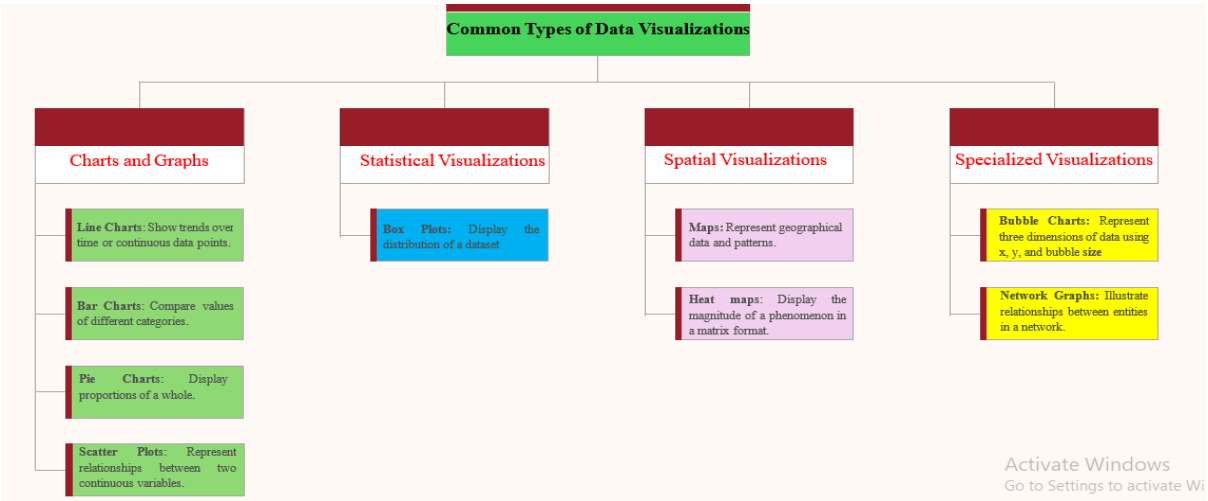
2. Getting Started with R and RStudio

Before diving into data visualization, ensure that you have **R** and **RStudio** installed on your computer. R is a language for statistical computing and graphics. RStudio is an Integrated Development Environment (IDE) for R. These can be downloaded from CRAN and RStudio from RStudio's website. Once installed, launch RStudio, where you can interact with R and visualize data effectively.

3. Basic Plotting in R

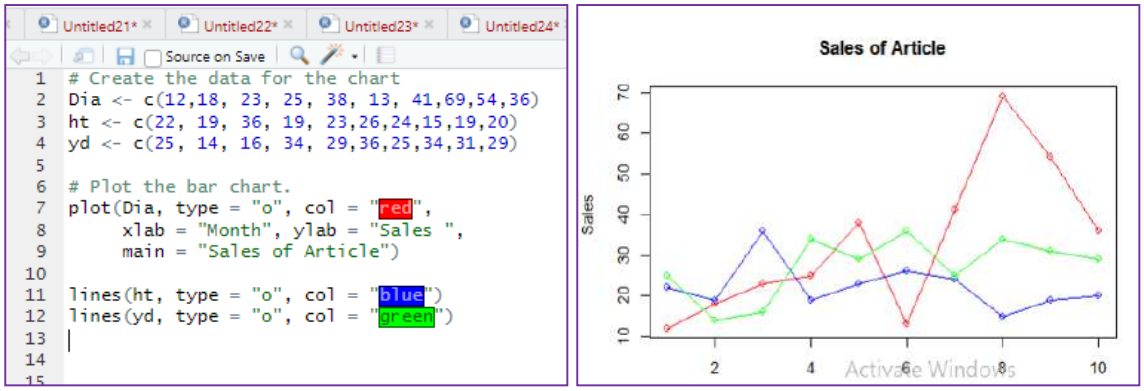
R provides a rich set of plotting functions in the **base** package, allowing users to quickly generate a variety of basic plots.

Common Types of Data Visualizations:



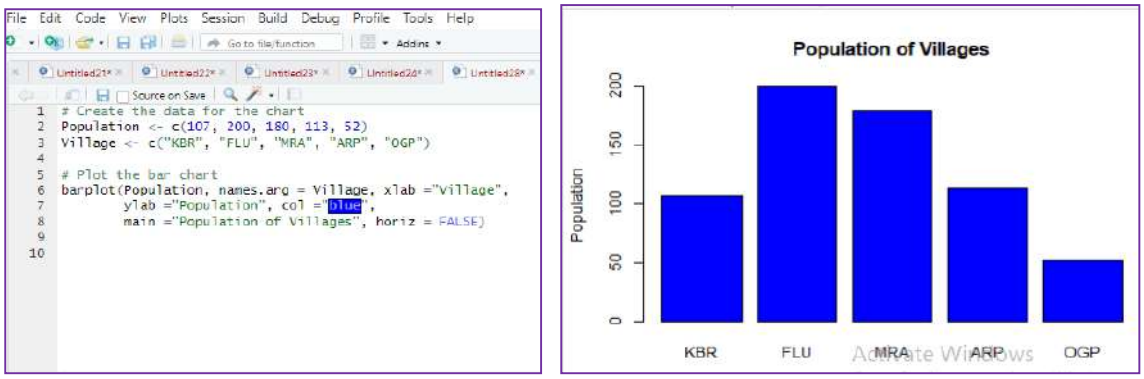
3.1 Charts and Graphs:

1. **Line Charts:** A line chart visually displays data trends over time using connected data points. It is widely used in various fields for analysing and representing data patterns.



```
# Create the data for the chart
Dia <- c(12,18, 23, 25, 38, 13, 41,69,54,36)
ht <- c(22, 19, 36, 19, 23,26,24,15,19,20)
yd <- c(25, 14, 16, 34, 29,36,25,34,31,29)

# Plot the bar chart.
plot(Dia, type = "o", col = "red",
     xlab = "Month", ylab = "Sales ",
     main = "Sales of Article")

lines(ht, type = "o", col = "blue")
lines(yd, type = "o", col = "green")
```


2. **Bar Charts:** A bar chart is a visual representation of data where individual bars represent different categories, and the length of each bar corresponds to the value it represents. It is commonly used to compare and show the relationships between different data sets.



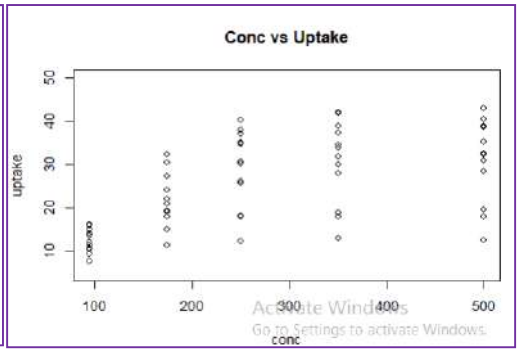
```
# Create the data for the chart
Population <- c(107, 200, 180, 113, 52)
Village <- c("KBR", "FLU", "MRA", "ARP", "OGP")

# Plot the bar chart
barplot(Population, names.arg = Village, xlab ="Village",
        ylab ="Population", col ="blue",
        main ="Population of Villages", horiz = FALSE)
```


3. **Pie Charts:** A pie chart is a circle divided into sectors to show the proportion of different categories in a dataset. Each sector's size corresponds to the percentage it represents, making it effective for visualizing relative proportions.

```
# Create data for the graph.
data <- c(63, 26, 60, 53, 29)
labels <- c("Irrigation", "Fertilizer", "Plant Protection", "Land Preparation", "Harvesting")

piepercent<- round(100 * data / sum(data), 1)

# Plot the chart.
pie(data, labels = piepercent,
    main = "Energy Use Patttern in Crop", col = rainbow(length(data)))
legend("topright", labels,
       cex = 0.5, fill = rainbow(length(data)))
```

4. **Scatter Plots:** A scatter plot is a graph showing points in a coordinate system, with each point representing a pair of values for two variables. Scatter plots are crucial in statistical analysis for identifying associations and understanding data distribution.

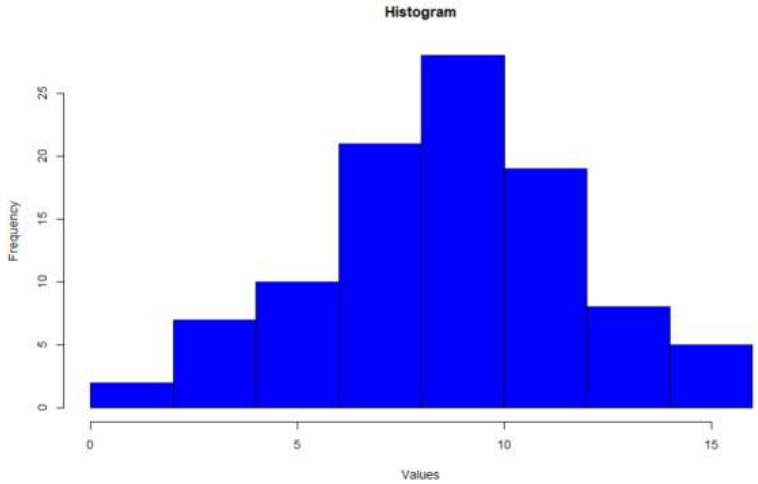
```
# Get the input values.
datam <- CO2[, c('conc', 'uptake')]

# Plot the chart for cars with
plot(x = datam$conc, y = datam$uptake,
     xlab = "conc",
     ylab = "uptake",
     xlim = c(95, 500),
     ylim = c(5, 50),
     main = "Conc vs Uptake"
)
```



3.2 Statistical Visualizations:

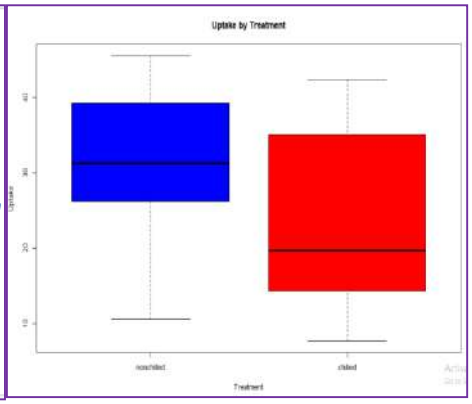
1. **Histogram:** It visually displays the distribution of data by illustrating how values are distributed across different ranges or bins. It provides insights into the central tendency and spread of the data, making it a valuable tool for understanding patterns and trends within a dataset.

```
hist(a, col = "blue", xlab='Values', ylab='Frequency', main='Histogram')
```



2. **Box Plots:** A box plot graphically represents the five-number summary, offering a concise overview of the data's central tendency and spread. The five-number summary includes the minimum, maximum, median (50th percentile), lower quartile (Q1, 25th percentile), and upper quartile (Q3, 75th percentile). In a box plot, a central box spans the interquartile range (from Q1 to Q3), with a line inside marking the median. Lines extend from the box to the smallest and largest observations. Box plots can be oriented either horizontally or vertically. Additionally, a box plot may identify potential outliers. They serve as effective tools for conveying information about the location and variation within datasets, particularly for highlighting changes between different data groups.

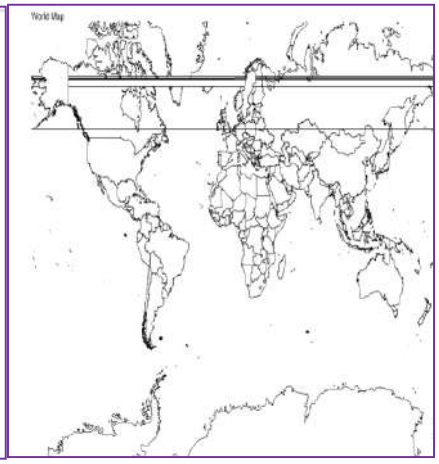
```r
# Load the dataset
data(CO2)

# Set up plot colors
my_colors <- c("blue","red")

# Create the box plot with customized aesthetics
boxplot(uptake ~ Treatment, data = CO2,
        main = "Uptake by Treatment", xlab = "Treatment", ylab = "Uptake",
        col = my_colors, border = "black")

# Add a legend
legend("topright", legend = unique(CO2$Treatment),
       fill = my_colors, border = "black", title = "Treatment")
```



3.3 Spatial Visualizations:

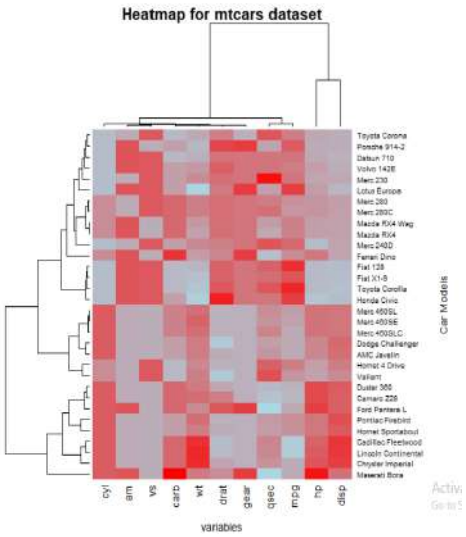
1. **Maps:** Represent geographical data and patterns.

```r
library(ggplot2)
library(maps)
world_map <- map_data("world")
ggplot(world_map, aes(long, lat, group = group)) +
  geom_polygon(fill = "white", color = "black") +
  coord_map("mercator") +
  ggtitle("World Map") +
  theme_void()
```



2. **Heatmaps:** Display the magnitude of a phenomenon in a matrix format.

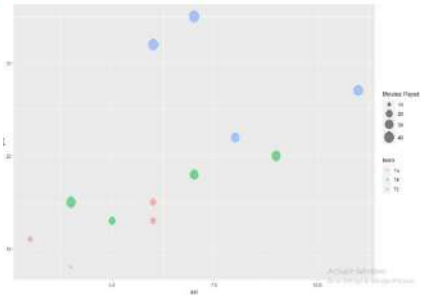
```r
# Heatplot from Base R
# using default mtcars dataset from the R
x  <- as.matrix(mtcars)

# custom colors
new_colors <- colorRampPalette(c("lightblue", "red"))

# plotting the heatmap
plt <- heatmap(x,
               # assigning new colors
               col = new_colors(100),

               # adding title
               main = "Heatmap for mtcars dataset",
               margins = c(5,10),
               # adding x-axis and y-axis labels
               xlab = "variables",
               ylab = "Car Models",
               scale = "column"
)
```



3.4 Specialized Visualizations:

Bubble Charts: Represent three dimensions of data using x, y, and bubble size.

```
#create data frame
fd <- data.frame(team=c('TA', 'TA', 'TA', 'TA','TB', 'TB', 'TB', 'TB','TC', 'TC', 'TC', 'TC'),
                 pts=c(8, 11, 13, 15, 13, 15, 18, 20, 22, 27, 32,35),
                 ast=c(4, 3, 6, 6, 5, 4, 7, 9, 8, 11, 6, 7),
                 min=c(9, 12, 15, 18, 20, 36, 30, 35, 31, 40, 43, 49))

#view data frame
fd
library(ggplot2)

#create bubble chart and color circles based on value of team variable
ggplot(fd, aes(x=ast, y=pts, size=min, color=team)) +
  geom_point(alpha=0.5)  +
  scale_size(range=c(2, 10), name='Minutes Played')
```



4. Conclusion

Data visualization is an essential skill for effective data analysis, as it allows complex data to be communicated in an intuitive and accessible way. R, a powerful programming language for statistical computing, offers an extensive range of visualization tools for various needs. It includes basic charting options like bar graphs, histograms, and line plots, which are ideal for simple data exploration. These tools allow users to easily examine trends, distributions, and relationships in data. With R's straightforward approach, users can quickly generate clear and meaningful visuals, making it a great choice for beginners and those seeking simplicity. This ease of use, combined with its versatility, ensures that data insights are presented in a clear and effective manner, helping analysts and decision-makers interpret information with confidence. R's basic visualization tools lay the foundation for any data story, supporting a wide range of analytical purposes.

ANALYSIS OF SURVEY DATA USING R SOFTWARE

Raju Kumar and Deepak Singh

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

A sample survey is a method for collecting data from or about the members of a population so that inferences about the entire population can be obtained from a subset, or sample, of the population members. In other words, it is a way of collecting information from a random sample of observations drawn from a population of interest using a probability-based sample design. Certain strategies are frequently employed in sample surveys to improve precision and control survey data collection expenses. These methods introduce a complexity to the analysis, which must be taken into account in order to create unbiased estimates and their associated precision levels. This paper gives a quick overview of how these design complications affect sampling variance and then outlines how to use the survey function in software R to analyze sample survey data.

2. Complex Sample Designs

Statistical methods are involved in carrying out a study include planning, designing, collecting data, analyzing, drawing meaning interpretation and reporting of the research findings. 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 post stratification 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 (e.g. 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.

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.

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 estimates 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.

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.	cds	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.	pcttest	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
24.	acs.k3	average class size years K-3
25.	acs.46	average class size years 4-6
26.	acs.core	Number of core academic courses
27.	pct.resp	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
37.	api.stu	number of students tested.

Type **summary(apipop)** 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. **apipop** 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"     "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"       "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"    "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" "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"    "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, 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.

DEVELOPMENT OF R PACKAGE

Pankaj Das

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

R is a powerful statistical programming language widely used in data science, machine learning, and research. One of its key strengths lies in its extensibility, allowing users to develop their own **R packages** to share code, functions, and datasets efficiently. Developing an R package is essential for creating reusable and maintainable code, contributing to the open-source community, and enhancing reproducibility in research. This article provides a step-by-step guide to developing an R package, covering package structure, documentation, testing, and publishing on **CRAN (Comprehensive R Archive Network)** or **GitHub**.

Why Develop an R Package?

Developing an R package offers several benefits:

1. **Code Reusability** – Functions can be easily shared and reused in different projects.
2. **Collaboration** – Team members can work with standardized functions.
3. **Documentation** – Well-structured packages enhance usability and understanding.
4. **Contribution to the Community** – Packages can be published for global use.

Steps to Develop an R Package

1. Setting Up the Package

To create an R package, start by installing the devtools and usethis packages:

```
#r console
install.packages("devtools")
install.packages("usethis")
library(devtools)
library(usethis)
```

Now, create the package structure using:

```
#r console
create_package("path/to/package_name")
```

This command generates a folder structure with necessary files.

2. Understanding Package Structure

An R package consists of:

- **DESCRIPTION** – Metadata about the package (title, author, dependencies).
- **NAMESPACE** – Specifies which functions are exported.

• **R/** – Contains all R scripts with functions.
• **man/** – Stores documentation for functions.
• **tests/** – Includes unit tests for checking function correctness.
• **vignettes/** – Provides long-form documentation and use cases.

3. Writing Functions

Develop functions inside the R/ directory. Example:

```
#r console
# Save this in R/myfunction.R
my_function <- function(x, y) {
  return(x + y)
}
```

4. Documenting Functions

Use **roxygen2** for documentation. Add comments like:

```
#r console
#' Add Two Numbers
#'
#' This function takes two numbers and returns their sum.
#'
#' @param x First number.
#' @param y Second number.
#' @return Sum of x and y.
#' @examples
#' my_function(3, 5)
#' @export
my_function <- function(x, y) {
  return(x + y)
}
```

Run document() to generate documentation:

```
#r console
devtools::document()
```

5. Testing the Package

Testing ensures reliability. Use testthat to create test cases:

```
#r console
usethis::use_testthat()
```

Write tests inside tests/testthat/:

```
#r console
test_that("my_function works correctly", {
  expect_equal(my_function(2, 3), 5)
  expect_equal(my_function(-1, 1), 0)
})
```

Run tests using:

```
#r console
devtools::test()
```

6. Checking and Building the Package

Before releasing, check the package:

```
#r console
devtools::check()
```

To build the package:

r

devtools::build()

7. Publishing the Package

Publishing on GitHub

If you want to share your package on GitHub, use:

```
#r console
usethis::use_git()
usethis::use_github()
```

Users can install your package via:

```
#r console
devtools::install_github("username/package_name")
```

Publishing on CRAN

To publish on CRAN:

1. Check your package with devtools::check()
2. Submit using usethis::use_cran_submission()
3. Follow CRAN guidelines and respond to reviewer feedback.

Conclusion

Developing an R package is a structured process that enhances reproducibility, usability, and collaboration in research and data science. By following best practices in documentation, testing, and distribution, you can create robust and valuable R packages for personal or community use.

PYTHON – AN OVERVIEW

Md Ashraful Haque

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012

1. Introduction:

Python is the one of the most popular programming languages now-a-days. It is a high-level, interpreted, interactive, object-oriented programming language. Python language was created by Guido van Rossum in 1991 at the National Research Institute for Mathematics and Computer Science in the Netherlands. Python programming language is mainly used for-

- Data handling and visualization
- Analysis of variety of data such as numerical, textual, image, videos, audio etc.
- Performing complex mathematical computations
- Server-side scripting for developing web applications.
- Standalone software development etc.

Why Python?

Python is very easy learn language. It can work in any system irrespective of the operating system. The syntax of python language is very simple and allows programmers to write programs in very few lines. Python runs on an interpreter system, which means that the code is being executed as soon as it is written. And last but not least, that python has a very large and mature community for the developers. There are lots of blogs, tutorials, documents, guide videos available online for python developers.

Python Installation:

Most of the latest computer systems have python already installed. To check if you have python installed on a Windows PC, search in the start bar for Python or run the following on the Command Line (cmd.exe):

C:\your\python\installation\folder>python --version

If not, then one can download the latest version of python (latest version is 3.9.2) from https://www.python.org/downloads/ for the particular operating system and follow the guidelines while installation.

Getting Started with Python:

Any python script or file is saved with .py file extension. Let's write the first python program that prints 'Hello, Everyone!!!'. So, first open a text editor and write the following code in it:

e.g.

```
print("Hello, Everyone!!!")
```

Now save it as 'first.py'. Now open command prompt, go to the python installation folder and type the following command:

C:\your\python\installation\path>python /your/program/path/first.py

The output should read:

Hello, Everyone!!!

Python from Command Line:

In case of python, it is possible to run the code as a command line itself using the command prompt.

Type the following on the Windows, Mac or Linux command line:

C:\your\python\installation\path>python

From there one can write any python code, including our first example from earlier in the:

C:\your\python\installation\path>python

Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>

Which will write "Hello, Everyone!!!" in the command line:

C:\your\python\installation\path>python

Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> print("Hello, Everyone!!!")

Hello, Everyone!!!

Whenever you are done in the python command line, you can simply type the following to quit the python command line interface:

exit()

Python Syntax:

The major syntactical rules of python programs has been provided below-

Execution of code

a. python can be executed directly from command line.

>>> print("Hello, Everyone!!!")

Hello, Everyone!!!

b. Python can also be executed using a file with '.py' extension

C:\your\python\installation\path>python /your/program/path/first.py

Indentation

The indentation refers to the spaces at the beginning of a program line. Indentation is very important and stricter in python. Python uses indentation as a block of code.

e.g.

```
if 5 > 2:
  print("Five is greater than two!")
```

Comments

In python, comments can be included in the code by using '#' symbol. Comments can be used in the beginning, middle, or in the end pf the code. Comments can be multiline. For multiline comments one can use triple quotes (""").

Variables in Python:

In python, the variables are simple storage structures for storing data values. There is no requirement of *type* declaration for the variables in python. The *type* of any variable can be acquired by *type()* function.

e.g.

```
x = 5
y = "python"
print(type(x))
print(type(y))
```

In python variables names -

- are case sensitive
- Must starts with a letter or underscore
- Can be alphanumeric

Python variables can store different types of data.

Text Type:	str
Numeric Types:	int, float, complex
Sequence Types:	list, tuple, range
Mapping Type:	dict
Set Types:	set, frozenset
Boolean Type:	bool
Binary Types:	bytes, bytearray, memoryview

Operators in Python:

Python divides the operators in the following groups:

Arithmetic operators	+, -, /, *, %, **, //
Assignment operators	=, +=, -=, *=, /=
Comparison operators	==, !=, >, <, >=, <=
Logical operators	And, or, not
Identity operators	is, is not,
Membership operators	in, in not
Bitwise operators	&, |, ^, ~, >>, <<

Data structures in python:

Data Structures are the way of organizing, storing, manipulating, and accessing data in better way. The data structures enable us to can be access and update data in a more efficient manner depending upon the situation. Data Structures are fundamentals of any programming language around which a program is built. There are mainly four types of built-in data structures in python. Python helps to learn the fundamental of these data structures in a simpler way as compared to other programming languages. These data structures are-

- List
- Tuple
- Set
- Dictionary

1. List Data Structures:

List are used to store more than one data in single variable. In python lists are flexible i.e. it can store multiple data type in a single list.

The characteristics of Lists Data Structures in python are:

- Items are indexed (starting from 0)
- Items are ordered
- Items are changeable
- Lists allow duplicate values of items

Creation of List: Lists are created by placing the comma separated items inside the square brackets.

```
## creation of lists
list1 = ["apple", "banana", "cherry"]
list2 = [1, 5, 7, 9, 3]
list3 = [True, False, False]
list4 = [1, 2, 3, "GFG", 2.3]
list5 = [1,2,3,4,4,]
```

Accessing Items from List: Items of the lists can be access by mentioning the index or indices inside the square brackets.

```
## accessing items
list1 = ["apple", "banana", "cherry"]
list2 = [1, 5, 7, 9, 3, 6, 9, 2, 1, 10]
x = list1[0]
print(x)
y = list2[1:4]
print(y)
```

```
## new list
new_list = [1, 2, 3, 'example', 3.132, 10, 30]
#access all elements
print(new_list)
#access index 3 element
print(new_list[3])
#access elements from 0 to 1 and exclude 2
print(new_list[0:2])
#access elements in reverse
print(new_list[::-1])
```

Updating the list: Items in the list at particular position can be updated by mentioning the values in the left-hand side of the assignment operator.

```
list2 = [1, 5, 7, 9, 3, 6, 9, 2, 1, 10]
list2[2] = 34
print(list2)
```

Remove items: Items in the list at particular position can be deleted by del statement.

```
list2 = [1, 5, -12, 9, 3, 6, 9, 2, 1, 10]
del list2[2] print(list2)
```

Some common functions operate on list data structures:

```
## append(): adds an items or a list of items in at the end of a list
list1.append(list2)
## insert(): adds an items at a particular location of a list
list1.insert(1,'mango')
## remove(): deletes an item by its value from a list
list1.remove('banana')
## clear(): deletes all the elements from the
list list1.clear()
## index(): finds the index of the given element in the list
list1.index('mango')
## finds the count of the given element present in the list
list1.count('mango')
#sorted(): temporarily sorts the elements of the list
sorted(list1)
#sort(): permanantly sorts the elements of the list
list1.sort(reverse=True)
```

Some basic operations on list data structures:

```
## Get number of items in a list
n = len(list1)
## Concatenate two lists together
list_new = list1 + list2
## check membership of an item in a list
100 in list2 # (gives true or false)
```

2. Tuple Data Structure:

Tuples are sequence of immutable objects in python. Tuples can store more than one datatype in a single instance of tuple.

The characteristics of Tuple Data Structures in python are:

- Items are indexed (starting from 0)
- Items are ordered
- Items are non-changeable
- Tuples allow duplicate values for the items

Creation of tuples: Tuples are created by placing the comma separated items inside the round brackets or parenthesis.

```
tuple1 = ("apple", "banana", "cherry")
tuple2 = (1, 5, 7, 9, 3)
tuple3 = (True, False, False)
```

Accessing items from Tuple: Items of the tuple can be access by mentioning the index or indices inside the square brackets.

```
## accessing items
tuple1 = ("apple", "banana", "cherry")
tuple2 = (1, 5, 7, 9, 3, 6, 9, 2, 1, 10)
x = tuple1[0]
print(x)
y = tuple2[1:4]
print(y)
```

Updating the tuple: Items in the tuples can't be changes once the tuple is created.

```
tuple2 = (1, 5, 7, 9, 3, 6, 9, 2, 1, 10)
tuple2[2] = 34 ## will raise an error print(tuple2)
```

Remove items: Items in the tuple can't be deleted as tuples are immutable. However, del statement can be used to delete whole tuple instead.

```
tup = ('physics', 'chemistry', 1997, 2000)
```

```
print(tup)
del tup
print(tup) ## will raise an error
```

Some basic operations on tuple data structures:

```
## Get number of items in a tuple
n = len(tuple1)
tuple_new = tuple1 + tuple2
## check membership of an item in a tuple
100 in tuple2 # (gives true or false)
```

3. Set Data Structure:

Mathematically, a set is a collection of items in any order. The sets in python are typically used for mathematical operations like union, intersection, difference and complement etc.

The characteristics of Set Data Structures in python are:

- Items are unindexed
- Items are unordered
- Items are non-changeable.
- Sets doesn't allow duplicate values

Creation of Sets: Sets are created by placing the comma separated items inside curly brackets.

```
set1 = {"apple", "banana", "cherry"}
set2 = {1, 5, 7, 9, 3}
set3 = {True, False, False}
```

Accessing items in Sets: Items in the sets can't be access by mentioning the index number. For accessing the items in the Sets one can use any loop structure.

```
Days=set(["Mon","Tue","Wed","Thu","Fri","Sat","Sun"]
for d in Days:
        print(d)
```

Adding and deleting items: In Sets, a new item can be added using *add()* function and an existing item can be deleted by *discard()* function.

```
Days=set(["Mon","Tue","Wed","Thu","Fri","Sat"])
Days.add("Sun")
print(Days)
Days.discard("Mon")
print(Days)
```

Different set operations:

Union of Sets: The union operation on two sets produces a new set containing all the distinct elements from both the sets. In the below example the element "Wed" is present in both the sets. Here, pipe (|) operator is used.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Wed","Thu","Fri","Sat","Sun"])
AllDays = DaysA|DaysB
print(AllDays)
```

Intersection of Sets: The intersection operation on two sets produces a new set containing only the common elements from both the sets. Here, ampersand (&) operator is used.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Wed","Thu","Fri","Sat","Sun"])
AllDays = DaysA & DaysB
print(AllDays)
```

Difference of Sets: The difference operation on two sets produces a new set containing only the elements from the rst set and none from the second set. Here, minus (-) operator is used.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Wed","Thu","Fri","Sat","Sun"])
AllDays = DaysA - DaysB
print(AllDays)
```

Compare Sets: We can check if a given set is a subset or superset of another set.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Mon","Tue","Wed","Thu","Fri","Sat","Sun"])
SubsetRes = DaysA <= DaysB
SupersetRes = DaysB >= DaysA
print(SubsetRes)
print(SupersetRes)
```

4. Dictionary Data Structure:

Dictionaries are the type of data structure that are used to store data in **key:value** pair. In Dictionary, each key is separated from its value by a colon (:), the items are separated by commas, and the whole thing is enclosed in curly braces.

The characteristics of Dictionaries Data Structures in python are:

- Items are ordered
- Items are changeable
- Dictionary doesn't allow duplicate values

Creation of dictionaries: Dictionaries are created by placing the comma separated **key:values** pairs inside curly brackets.

```
dictionary1 = {"brand": "Ford",
               "model": "Mustang",
               "year": 1964}
x = dictionary1 ["model"]
print(x)
```

Accessing items: Items ca be access by mentioning the key name inside the square bracket.

```
dict = {'Name': 'Zara', 'Age': 7, 'Class': 'First'}
print ("dict['Name']: ", dict['Name'])
print ("dict['Age']: ", dict['Age'])
```

Updating Dictionary: One can update a dictionary by adding a new entry or a key-value pair, modifying an existing entry, or deleting an existing entry.

```
dict = {'Name': 'Zara', 'Age': 7, 'Class': 'First'}
dict['Age'] = 8;
# update existing entry
dict['School'] = "DPS School"
# Add new entry
print ("dict['Age']: ", dict['Age'])
print ("dict['School']: ", dict['School'])
```

Delete Dictionary Elements: One can either remove individual dictionary elements or clear the entire contents of a dictionary.

```
dict = {'Name': 'Zara', 'Age': 7, 'Class': 'First'}
del dict['Name'] # remove entry with key 'Name'
dict.clear() # remove all entries in dict
del dict  # delete entire dictionary
print ("dict['Age']: ", dict['Age'])
print ("dict['School']: ", dict['School'])
```

Control Structures in Python:

There is mainly one control structure that is *if...else* . The *if...else* structures are used to implement the logical conditions of the program and allow the program to branch based on the evaluation of an expression.

General syntax of *if...else* :

```
if expression :
   statement 1
   statement 2
   ...
```

```
   statement n else:
   statement 1
   statement 2
   ...
statement always executed
```

N.B. Indentation in the control and loop structures are very crucial in case of python programming language.

```
## examples of if..else
## if statement value = 5
threshold= 4
print("value is", value, "threshold is ",threshold)
if value > threshold :
      print(value, "is bigger than ", threshold)
## if..else statement a = 330 b = 200 if b > a:
      print("b is greater than a") else:   print("error")
```

Nested control structures: The *if..else* structures can be used in nested manner by using *elif* statement.

```
## nested if.. statements
### if ... elif ... else ...
a = 5
b = 4
print("a = ", a, "and b = ", b)
if a > b :
   print(a, " is greater than ", b)
elif a == b :
   print(a, " equals ", b)
else :
   print(a, " is less than ", b)
```

Loop Structures in Python:

In python generally two types of loop structures are used: while loop and for loop.

1. *while* loop:

With the 'while' loop, a set of statements can be executed repeatedly long as a condition is true . For the loop to terminate, there has to be some termination criteria mentioned in the code which will potentially change the condition and stop the iteration.

```
   ## Simple example
```

```
i=1
while i < 6:
    print(i)
    i = i + 1
## sum of n numbers using a while loop
n = 10
cur_sum = 0
i = 1
while  i <= n :
   cur_sum = cur_sum + i
   i = i + 1
print("The sum of the numbers from 1 to", n, "is ", cur_sum)
```

Points to note:

- Here, the conditional clause (i <= n) in the *while* statement can be anything which would return a boolean value of either *True* or *False* upon execution.
- Initially *i* has been set to 1 (before the start of the loop) and therefore the condition is *True*.
- The clause can be made more complex by using parentheses, *and* and *or* operators amongst others
- The statements after the *while* clause are only executed if the condition evaluates as *True*.
- Within the statements after the *while* clause there should be something which potentially will make the condition evaluate as *False* next time around. If not the loop will never end.
- In this case the last statement in the loop changes the value of *i* which is part of the condition clause, so hopefully the loop will end.

2. *for* loop:

A *for* loop is used for iterating over a sequence (that is either a list, a tuple, a dictionary, a set, or a string) for executing a set of statements. The difference between *while* and *for* loop is that in *for* loop we know that at the outset how often the statements in the loop will be executed, we don't have to rely on a variable being changed within the looping statements as in *while* loop.

General syntax of *for* loop:

```
for variable_name in some_sequence :
   statement1
   statement2
   ...
   statementn
## simple example
```

```
for i in [1,2,3] :
   print(i)
print("\nExample 1\n")
fruits = ["apple", "banana", "cherry"]
for x in fruits:
 print(x)

print("\nExample 2\n")
for x in "banana":
 print(x)

print("\nExample 3\n")
for name in ["Tom", 42, 3.142] :
   print(name)

print("\nExample 4\n")
for i in range(10) :
   print(i)

print("\nExample 5\n")
longString = "The quick brown fox jumped over the lazy sleeping dog"
for word in longString.split() :
   print(word)
```

Python Functions:

A function is a block of code that contains a set of statements and runs only when it is called explicitly. One can pass data, known as parameters, into a function. A function can return data as a result.

e.g.

```
def my_function(str):
  print(str + "! Welcome to the class.")
my_function("Bob")
```

Packages in Python:

Package or module is a python object with arbitrarily named attributes that one can bind and reference. Packages allows us to logically locate the python code. Simply a package or module is file containing a set of python codes. Packages are also referred as library

Packages or modules or libraries can be imported by using the *'import'* keyword.

e.g.

```
import os
import sys
```

PIP is a package manager available in python. PIP is used to install, upgrade, or uninstall a packages in python environment.

C:\your\python\installation\path>pip install numpy

Some important packages or modules in Python:

NumPy:

NumPy is python library or packages used for working with arrays. NumPy was created by Travis Oliphant in 2005 and it is open source.

In python, the concepts of arrays is served by the List data structure but it is too slow in processing. NumPy provides a 50x faster access speed for the array objects in python than the List. NumPy has a lots of applications in the domain of -

- Arrays
- Matrices
- Linear Algebra
- Fourier Transformation

Creating Arrays

The object of NumPy that deals with the arrays is known as '*ndarray'*. One can create a *'ndarray'* object by using *array()* function. One can pass any type of array-like object in the *array()* function.

e.g.

```
import numpy as np
array_var = np.array([1, 2, 3, 4, 5])
```

Array can be of 0, 1, 2 or 3 dimensions.

e.g.

```
import numpy as np
array0 = np.array(42) #0 dimension
array1 = np.array([1, 2, 3, 4, 5, 6, 7, 8]) # 1 dimension
array2 = np.array([[1, 2, 3], [4, 5, 6]]) # 2 dimension
array3 = np.array([[[1, 2, 3], [4, 5, 6]], [[1, 2, 3], [4, 5, 6]]]) #3 dimension
```

Accessing Array elements

Array elements can be accessed by its index number

```
print(array1 [2]) #accessing the 3rd item from the array 'array1'
```

Slicing an Array

Slicing in python means taking elements from one given index to another given index.

```
print(array1 [1:3]) #slicing from 2nd item to the 4th element
print(array1 [2:]) #slicing from 3rd item to the last element
print(array1 [:6]) #slicing from beginning to the 5th element
```

Properties and functions:

dtype- returns the type of values stored in the array object

shape- gives the number of elements in each dimension of the array object

reshape– allows to change the shape of the array either by adding adding/removing dimensions or changing the number of elements in each dimension

concatenate()- joins two or more arrays axis wise.

array_split()– splitting an array into two or more parts

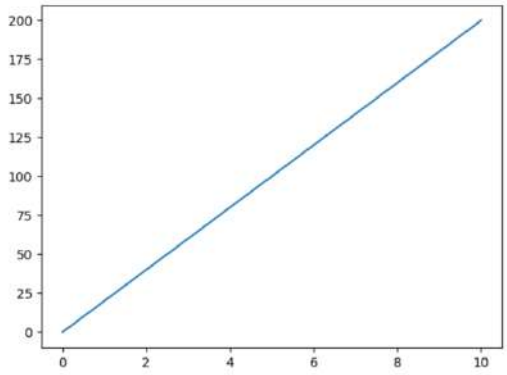
Matplotlib:

Matplotlib is a low level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter. Matplotlib is open source and we can use it freely.

Most of the Matplotlib utilities lies under the pyplot submodule, and are usually imported under the plt alias.

e.g. Draw a line in a diagram from position (0,0) to position (10, 200):

```
import matplotlib.pyplot as plt
import numpy as np
xpoints = np.array([0, 10])
ypoints = np.array([0, 200])
plt.plot(xpoints, ypoints)
plt.show()
```



Properties and functions:

marker- keyword argument to emphasize each point in the plot

linestyle/ls- keyword argument to change the style of the plotted line

xlabel()- functions for setting a label for x-axis

ylabel()- function for setting a label for y-axis

title() - function for giving the title for the plot

grid() -function to add grid lines to the plot

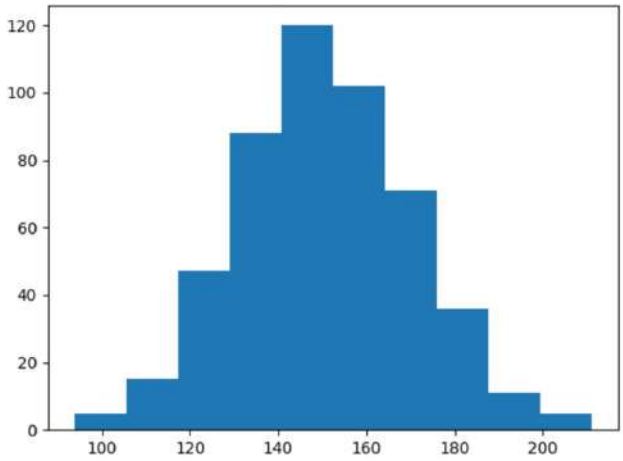
scatter()-function to draw a scatter plot

bar()- function to draw bar graphs

hist()- function to create histograms

e.g.

```
import matplotlib.pyplot as plt
import numpy as np
x = np.random.normal(150, 20 , 250)
plt.hist(x)
plt.show()
```



Pandas:

Pandas is a one of the most popular python package providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word 'Panel Data' – an Econometrics from Multidimensional data. Pandas is well suited for many different kinds of data:

- Fast and efficient DataFrame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and subsetting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations

There are mainly two data structures of pandas which handle the majority of typical use cases in finance, statistics, social science and Engineering are Series (1-dimensional) and DataFrame (2-dimensional).

DataFrame

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns.

Features of DataFrame:

- Potentially columns are of different types
- Size – Mutable
- Labelled axes (rows and columns)
- Can Perform Arithmetic operations on rows and columns

e.g.1

```
import pandas as pd
data = [1,2,3,4,5]
df = pd.DataFrame(data)
print df
```

```
   0
0  1
1  2
2  3
3  4
4  5
```

e.g. 2

```
import pandas as pd
data = [['Alex',10],['Bob',12],['Clarke',13]]
df = pd.DataFrame(data,columns=['Name','Age'],dtype=float)
print df
```

```
    Name   Age
0   Alex   10.0
1   Bob    12.0
2   Clarke 13.0
```

Importing data files using pandas

Pandas provides the means for datafiles to be imported to the python environment. External files in any format (.csv, .xls, .txt, .pdf, etc.) can be imported using pandas.

e.g. 1: .csv file can be imported by read_csv() function

```
data = pd.read_csv('/content/sample_data/california_housing_test.csv')
```

e.g. 2: .xls file can be imported by read_excel() function

```
data = pd.read_excel('/content/sample_data/shishamharvesteddata.xls')
```

Measure of central tendency

Mean, Median and Mode of the dataset can be calculated using mean(), median() and mode() functions available in Pandas

e.g.:

```
## mean
```

```
data[].mean()
## median
data[].median()
## mode
data[].mode()
```

Description statistics

Description statistics can be calculated by describe()function available in Pandas

e.g.:

```
data[['dbhcm','Branchkg','Stemkg']].describe()
```

output:

	dbhcm	Branchkg	Stemkg
count	42.000000	42.000000	42.000000
mean	18.927701	27.347262	91.985714
std	4.520851	14.871299	36.560946
min	10.828025	7.630000	20.640000
25%	16.037675	16.015000	66.947500
50%	19.135000	24.550000	96.705000
75%	21.702500	35.378750	114.047500
max	29.681529	70.235000	171.460000

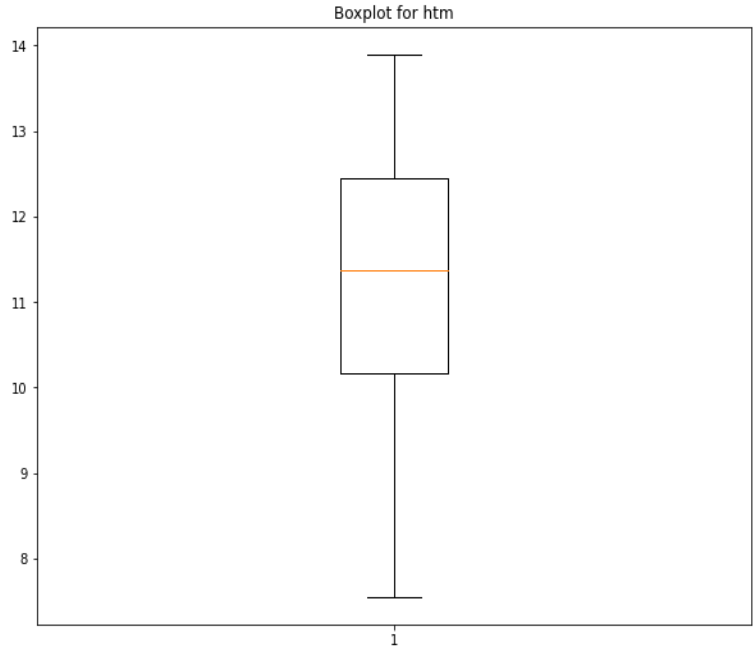
Boxplot

The boxplots can be drawn with the help of pyplot.boxplot function available with matplotlib.

e.g.:

```
## Boxplot
from matplotlib import pyplot as plt
fig = plt.figure(figsize=(10,8))
plt.boxplot(data['htm'])
plt.title('Boxplot for htm')
plt.show()
```

Output:



References:

1. https://colab.research.google.com/github/tensorflow/examples/blob/master/courses/udacity_intro_to_tensorflow_for_deep_learning/l01c01_introduction_to_colab_and_python.ipynb#scrollTo=F8YVA_634OFk
2. https://docs.python.org/3/tutorial/
3. https://numpy.org/
4. https://pandas.pydata.org/
5. https://www.guru99.com/python-tutorials.html
6. https://www.programiz.com/python-programming
7. https://www.tutorialspoint.com/python/index.htm
8. https://www.w3schools.com/python/default.asp

SPSS – AN OVERVIEW

Ankur Biswas

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

SPSS is a widely used software package for statistical analysis in social science. The original SPSS manual (Nie *et al.*, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis. Originally it is an acronym of *Statistical Package for the Social Science* but now it stands for *Statistical Product and Service Solutions*. The current versions (2015) are officially named IBM SPSS Statistics. Long produced by SPSS Inc., it was acquired by IBM in 2009. During 2009 and 2010 it was called *PASW* (*Predictive Analytics Software*) *Statistics*. It is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with rather simple instructions. This package of programs is available for both personal as well as mainframe 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.

Some versions of SPSS released in recent years are

- SPSS Statistics 17.0.1 - December 2008
- PASW Statistics 17.0.3 - September 2009
- PASW Statistics 18.0, 18.0.1, 18.0.2, 18.0.3
- IBM SPSS Statistics 19.0 - August 2010
- IBM SPSS Statistics 19.0.1, 20.0, 20.0.1, 21.0

Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services). Purpose of this chapter is to introduce the basic features of the SPSS and also to provide some basic statistical analysis using this software.

2. Key features of SPSS

Some of the key features of SPSS are

- It is easy to learn and use with its pull-down menu features
- It includes a full range of data management system and editing tools
- It offers comprehensive range of plotting, reporting and presentation features.
- It provides in-depth statistical analysis capabilities

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary stored in the data file) are features of the base software. There are varieties of statistics included in the base software. Some of the important statistics are:

Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics etc.

Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests etc.

Prediction for numerical outcomes: Linear regression, Multiple Regression

Prediction for identifying groups: Factor analysis, Cluster analysis (two-step, K-means, hierarchical), Discriminant analysis etc.

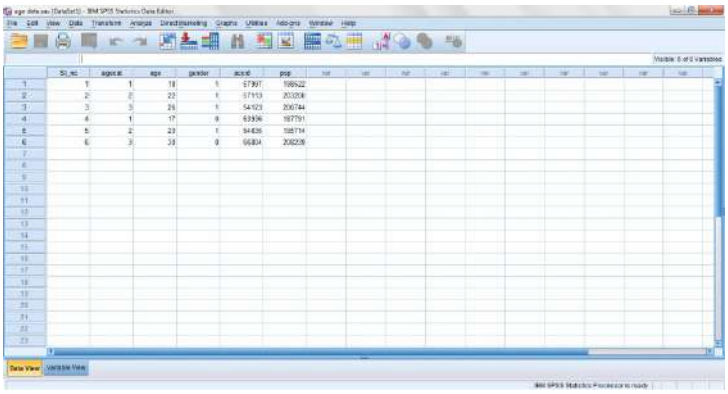
3. Basic features of SPSS

SPSS makes statistical analysis manageable for the naive user and convenient for the experts. There are a number of different types of windows in SPSS. The data editor offers a simple and efficient spreadsheet-like facility for entering data and browsing the working data file.

Data Editor: This graphical user interface displays the contents of the data file. One can create new data files or modify existing ones. The Data Editor window opens automatically when an SPSS session is started. This editor has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS window.

- ✓ **Data view**: Displays the actual data values or defined value labels. The 'Data View' shows a spreadsheet view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells. 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;
- ✓ **Variable view**: Displays variable definition information contained or metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics. One can modify variable properties in the Variable view for example, add and delete variables, change the order of variables etc.

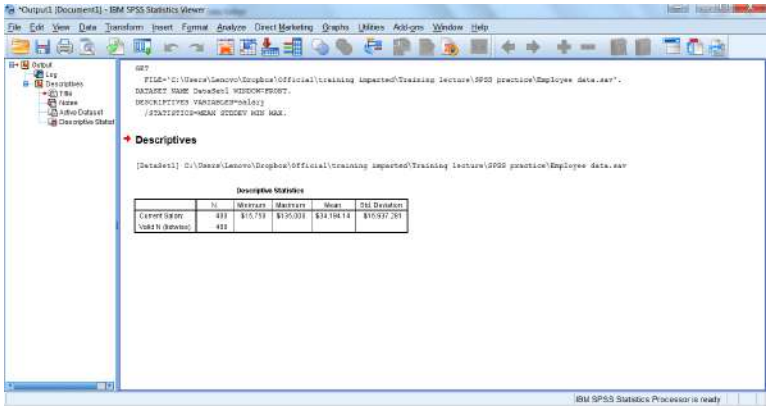
Extension of the saved data file will be ".sav".



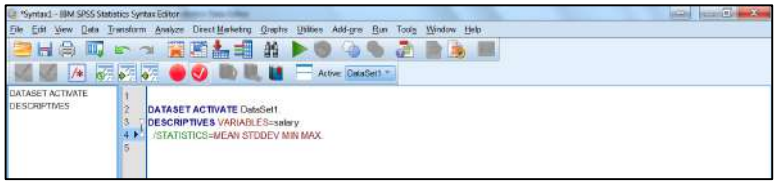
Viewer: All results, tables, and charts performed by different statistical analysis are displayed in the Viewer. Extension of the saved output file will be ".spv". One can use the Viewer to browse results, show or hide selected tables and charts, change the display order of results by moving selected items or move items between the Viewer and other applications. The output presented in Viewer can be edited and saved for later use. A

Viewer window opens automatically the first time a procedure is run that generates output. The Viewer is divided into two panes:

- ✓ The left pane contains an outline view of the contents. One can click an item in the outline to go directly to the corresponding table or chart.
- ✓ The right pane contains statistical tables, charts, and text output.



Syntax Editor: The pull-down menu interface generates command syntax: this can be displayed in the output. These command syntax can also be pasted into a syntax file in a syntax window using the "paste" button present in each menu. 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. Extension of the saved syntax file will be ".sps". Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses.



Pivot Table Editor: The results from most statistical procedures are displayed in pivot tables. These pivot tables outputs can be modified in many ways with pivot table editor. One can edit text, swap data in rows and columns, create multidimensional tables, and selectively hide and show results. Changing the layout of the table does not affect the results. Instead, it's a way to display information in a different or more desirable manner.

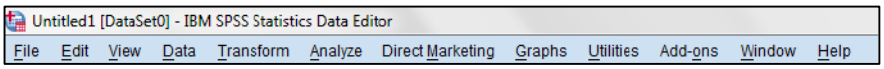
Text Output Editor: Text output not displayed in pivot tables can be modified with the Text Output Editor. One can edit the output and change font characteristics (type, style, colour, size).

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, switch the horizontal and vertical axes, rotate 3-D scatterplots, and even change the chart type.

Script Window: It provides the opportunity to write full-blown programs, in a BASIC-like language. It is a text editor for syntax composition. Extension of the saved script file will be ".sbs"

Many features of SPSS Statistics are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. 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 menu options available in SPSS are



Most menu selections open dialog boxes. These dialog boxes can be used to select variables and various options for analysis. The main dialog box usually contains the minimum information required to run a procedure. Additional specifications are made in sub-dialog boxes. All these above mentioned options have further sub-options. 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 dropdown menus to select. There are five standard command push buttons in most dialog boxes.

OK: It runs the procedure. After the variables and additional specifications are selected, click OK to run the procedure.

Paste: It generates command syntax from the dialog box selections and pastes the syntax into a syntax window.

Reset: It deselects any variables in the selected variable list and resets all specifications in the dialog box.

Cancel: It cancels any changes in the dialog box settings since the last time it was opened and closes the dialog box.

Help: It contains information about the current dialog box.

Basic Steps in Data Analysis using SPSS

- **Get data into SPSS**. There are several ways to get data in the SPSS. One can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter data directly in the Data Editor.
- **Select a procedure**. Select an appropriate procedure from the menus in order to perform appropriate analysis on the data file and calculate statistics or create charts.
- **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.

4. 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.

To be able to retrieve a file, the file must be saved with a proper name. The default extension name for saving files is **sav**. To save this file on hard disk, we carry out the following sequence:

> **File →Save As...** [opens **Save Data As** dialog box]→box under **File Name:** delete the asterisk and type file name →**OK**

The output file can also be printed and saved. The extension name for output file is .**spo**.

To retrieve this file, use the following procedure:

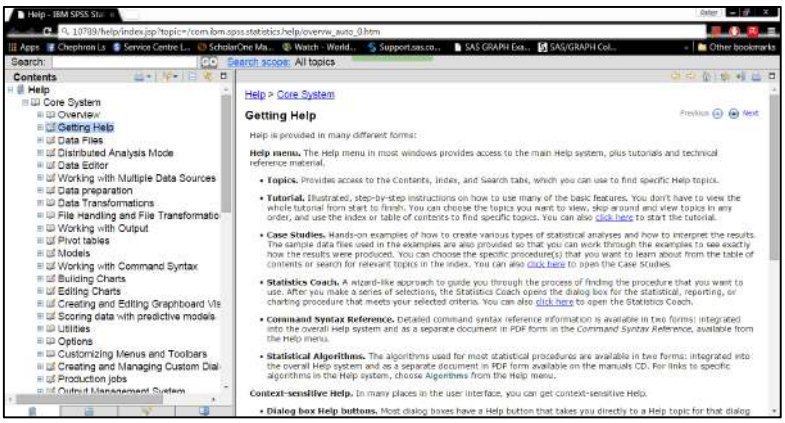
> **File→Open→Data...**[opens the **Open Data File** dialog box] →choose drive from options listed →type name under **File Name: → OK**

5. Statistical Procedures

After entering the data set in **Data Editor** or reading an ASCII data file, we are now ready to analyse it. The **Analyse** option has the following sub options:

Reports, Descriptive Statistics, Tables, Compare means, General Linear model, Mixed Models, Correlate, Regression, Loglinear, Neural Networks, Classify, Dimension Reduction, Scale, Non parametric tests, Forecasting, Time Series, Survival, Multiple response, Missing value analysis, Multiple imputation, Complex samples, Quality control, ROC curve.

Help topics available at IBM SPSS Statistics is so enriched that it helps naive users to manage their desired statistical analysis efficiently.



Some of the important statistical analysis options are described in detail as follows

5.1. Descriptive Statistics:

This submenu provides techniques for summarising data with statistics, charts, and reports. This is most useful for providing useful descriptive statistics for different types of dataset. There various sub-sub menus under this submenu.

Frequencies option helps in generating information about the relative frequency of the occurrence of each category of a variable. To compute summary statistics for each of several groups of cases, Means procedure or the Explore procedure can be used.

Descriptives option carry out statistical analysis 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. Several other additional features like 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 using this submenu.

Crosstabs is used to carry out cross-tabulation in order to count the number of cases that have different combinations of values of two or more variables, and to calculate summary statistics and tests.

P-P plots provides the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.

Q-Q plots provide the quantiles of a variable's distribution against the quantiles of the normal distribution.

5.2. 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.

One-sample t test procedure tests whether the mean of a single variable differs from a specified constant.

Independent sample t test is used to test 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 nonparametric 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 data obtained using completely randomised design can be analysed through this option.

5.3. General Linear Model:

This submenu provides techniques for testing univariate and multivariate Analysis of Variance models, including repeated measures.

Univariate sub-option could be used to analyse the experimental designs like Completely randomised design, Randomised 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.

Multivariate analyses analysis of variance and analysis of covariance designs when there are 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, one 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 which are related to each other. For a single dependent variable or unrelated dependent variables, the Univariate ANOVA procedures can be adopted.

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.

5.4. 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 nonparametric 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 nonparametric measures which are particularly useful when the data contain outliers or when the distribution of the variables is markedly non-normal.

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 value of a dependent variable from a set of independent variables is 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.

Distances calculates statistics measuring either similarities or dissimilarities (distances), either between pairs of variables or between pairs of cases. These similarity or distance measures can then be used with other procedures, such as factor analysis, cluster analysis, or multidimensional scaling, to help analyze complex datasets. Dissimilarity (distance) measures for interval data are Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, or customized; for count data, chi-square or phi-square; for binary data, Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, shape, or Lance and Williams. Similarity measures for interval data are Pearson correlation or cosine; for binary data, Russel and Rao, simple matching, Jaccard, etc.

5.5. 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.

Curve Estimation produces curve estimation regression statistics and related plots for 11 different curve estimation regression models. A separate model is produced for each dependent variable. One can also save predicted values, residuals, and prediction intervals as new variables.

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. In recent versions there are two options **Binary Logistic** as well as **Multinomial Logistic.**

Probit performs probit analysis which 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.

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.

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 which 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.

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.

5.6. Classify:

This submenu provides cluster and discriminant analysis.

Two Step Cluster performs Two Step Cluster Analysis procedure which is an exploratory data analysis tool designed to reveal natural clustering within a dataset that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques. The Log-

likelihood and Euclidean Distance Measures are used as the similarity measure between two clusters.

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.

Nearest Neighbor performs Nearest Neighbor Analysis for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

5.7. Dimension 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.

Correspondence Analysis analyses correspondence tables (such as cross-tabulations) to best measure the distances between categories or between variables. This command is in the Categories option.

Distances computes 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.

15.8. Nonparametric Tests:

This submenu provides nonparametric 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 runs test is appropriate only when the order of cases in the data file is meaningful.

1-Sample K-S 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 runs tests sorts the data values from smallest to largest and then performs a runs test on the group's 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 tests 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 nonparametric 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.

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 nonparametric 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.

5.9. Forecasting:

This submenu provides create models, seasonal decomposition, spectral analysis, autocorrelations, cross-correlations etc.

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.

Cross-correlations calculates and plots the cross-correlation function of two or more series for positive, negative, and zero lags.

Spectral analysis 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.

15.10. Complex Samples:

This submenu provides procedures for Sampling from Complex Designs. The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

18. Graphs

The Chart Builder available in Graph menu allows to build charts from predefined gallery charts or from the individual parts (for example, axes and bars). Build a chart by dragging and dropping the gallery charts or basic elements onto the canvas, which is the large area to the right of the Variables list in the Chart Builder dialog box.

Legacy Dialogs submenu provides following graph options

Bar generates a simple, clustered, or stacked bar chart of the data.

3-D Bar Charts generates bar graph in 3-dimensional axis.

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.

Boxplot generates boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.

Error Bar Charts plot the confidence intervals, standard errors, or standard deviations of individual variables.

Scatter/dot 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.

19. Exercises

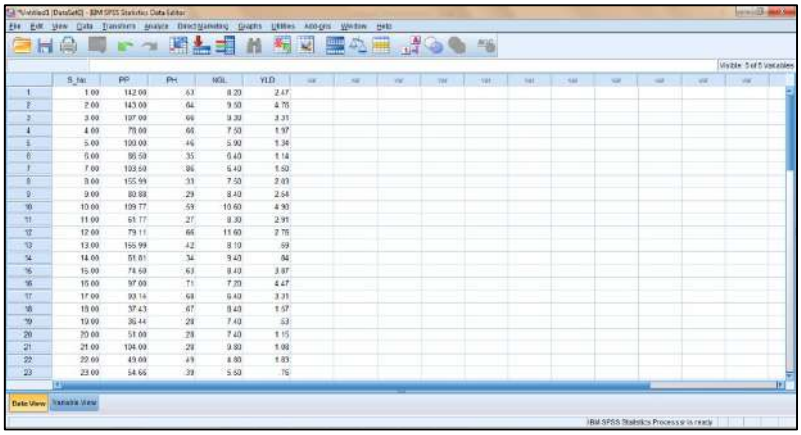
Exercise 1. The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

S.No.	PP	PH	NGL	Yield	S.No.	PP	PH	NGL	Yield
1	142.00	0.525	8.2	2.470	24	55.55	0.265	5.0	0.430
2	143.00	0.640	9.5	4.760	25	88.44	0.980	5.0	4.080
3	107.00	0.660	9.3	3.310	26	99.55	0.645	9.6	2.830
4	78.00	0.660	7.5	1.970	27	63.99	0.635	5.6	2.570
5	100.00	0.460	5.9	1.340	28	101.77	0.290	8.2	7.420
6	86.50	0.345	6.4	1.140	29	138.66	0.720	9.9	2.620
7	103.50	0.860	6.4	1.500	30	90.22	0.630	8.4	2.000
8	155.99	0.330	7.5	2.030	31	76.92	1.250	7.3	1.990
9	80.88	0.285	8.4	2.540	32	126.22	0.580	6.9	1.360
10	109.77	0.590	10.6	4.900	33	80.36	0.605	6.8	0.680
11	61.77	0.265	8.3	2.910	34	150.23	1.190	8.8	5.360
12	79.11	0.660	11.6	2.760	35	56.50	0.355	9.7	2.120
13	155.99	0.420	8.1	0.590	36	136.00	0.590	10.2	4.160
14	61.81	0.340	9.4	0.840	37	144.50	0.610	9.8	3.120
15	74.50	0.630	8.4	3.870	38	157.33	0.605	8.8	2.070
16	97.00	0.705	7.2	4.470	39	91.99	0.380	7.7	1.170
17	93.14	0.680	6.4	3.310	40	121.50	0.550	7.7	3.620
18	37.43	0.665	8.4	1.570	41	64.50	0.320	5.7	0.670
19	36.44	0.275	7.4	0.530	42	116.00	0.455	6.8	3.050
20	51.00	0.280	7.4	1.150	43	77.50	0.720	11.8	1.700
21	104.00	0.280	9.8	1.080	44	70.43	0.625	10.0	1.550
22	49.00	0.490	4.8	1.830	45	133.77	0.535	9.3	3.280
23	54.66	0.385	5.5	0.760	46	89.99	0.490	9.8	2.690

Source: Design Resources Server. ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on <05-05-2015>).

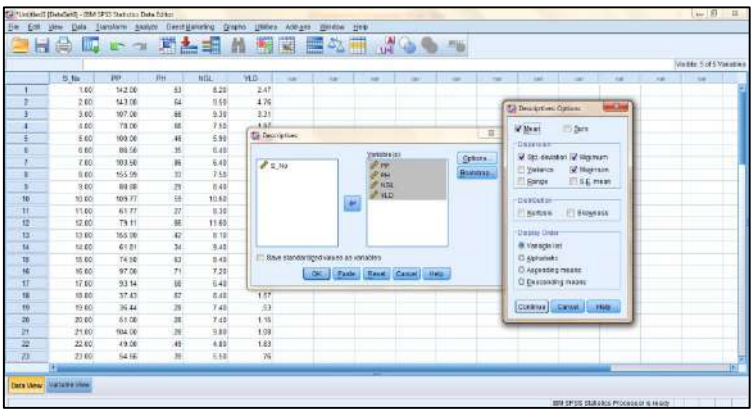
1. Find mean, standard deviation, minimum and maximum values of all the characters.
2. Find correlation coefficient between each pair of the variables.
3. Give a scatter plot of the variable PP with dependent variable yield.
4. Fit a multiple linear regression equation where yield is dependent variable whereas all other characters as independent variables.

At first enter the entire data in the data editor as given below,

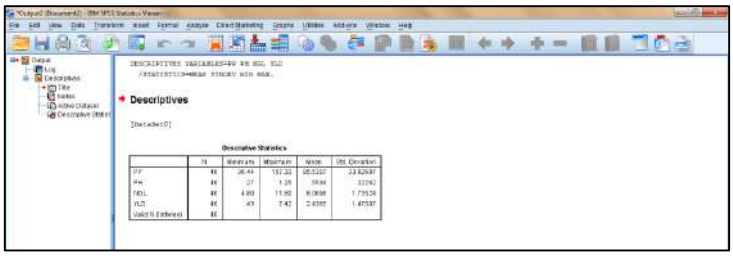


There are several ways to answer the Q no. 1in SPSS. Commands following first way is as follows,

Analyze → Descriptive Statistics → Descriptives…→ Put PP, PH, NGL, YLD in the variables list→ Choose appropriate options from Options tab→Press Continue → Press Ok

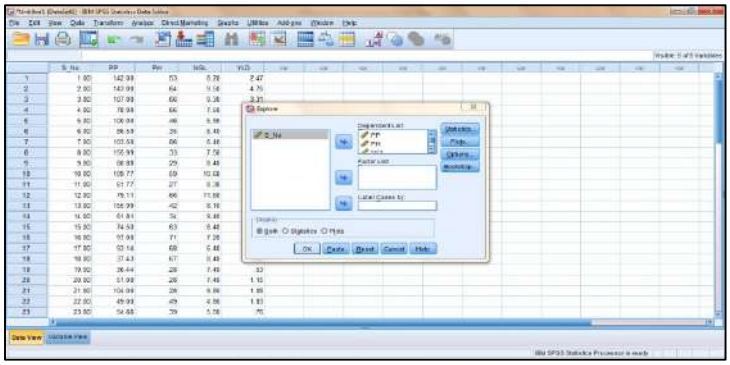


Output:

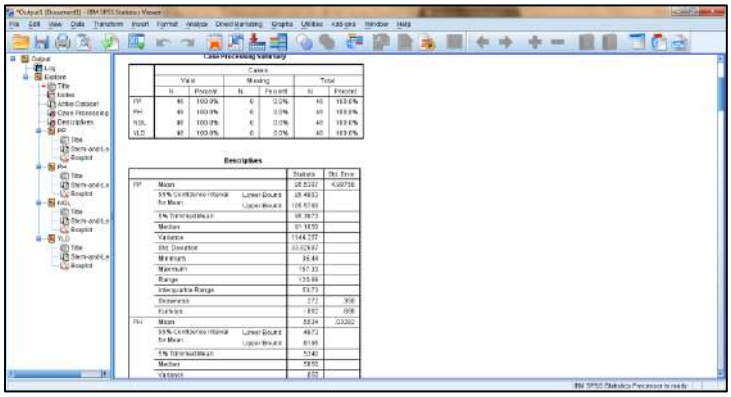


Another way:

Analyze → Descriptive Statistics → Explore…→ Put PP, PH, NGL, YLD in the Dependent list→ Choose both Statistics and plot→ Press Ok

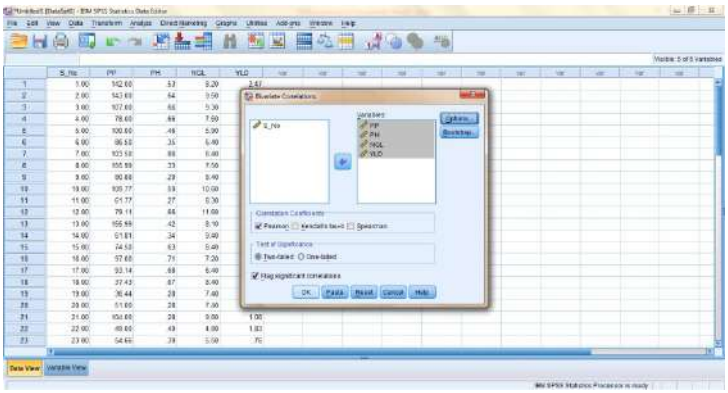


Output:

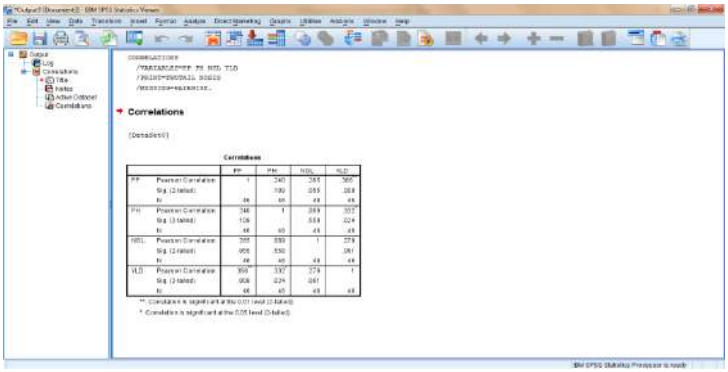


To answer Q no. 2 follow the following steps

Analyze → Correlate → Bivariate→ Put PP, PH, NGL, YLD in the Valiables list→ Choose Pearson's correlation coefficient→ Press Ok

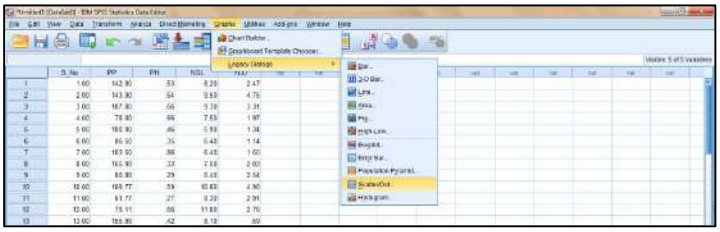


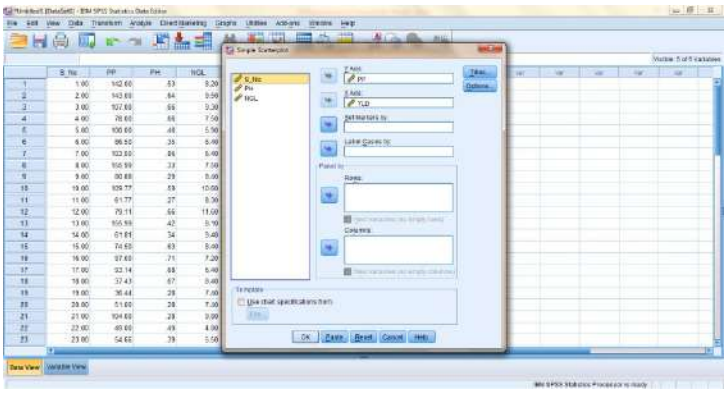
Output:



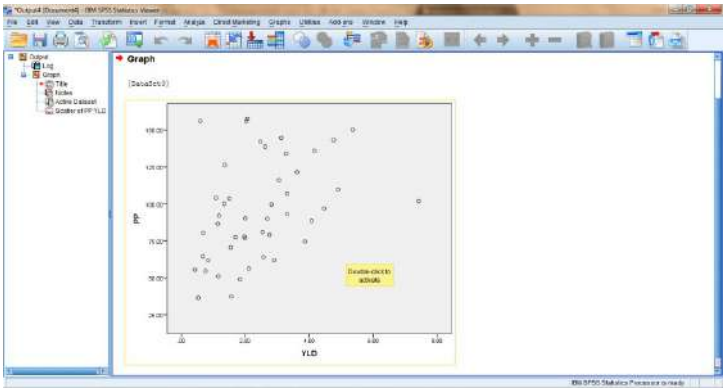
To give the scatter plot of the variable PP with dependent variable yield use following steps:

Graphs → Legacy dialogs→ Scatterplot→ Put PP at Y axis and YLD at X axis→ Ok





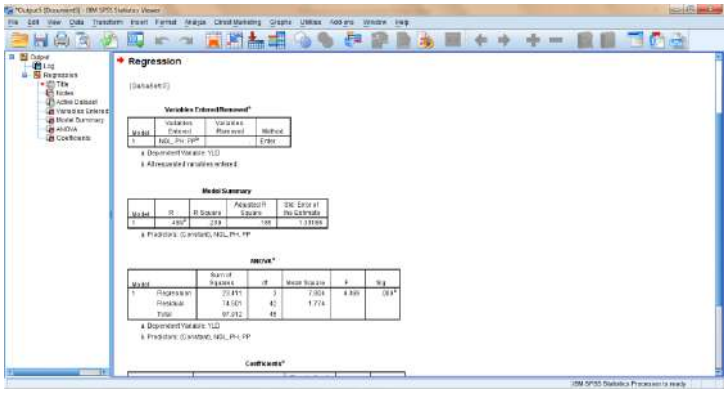
Output:



To fit a multiple linear regression equation taking yield as dependent variable and all other characters as independent variables perform following steps

Analyze → Regression → Linear → Put Yld in Dependent variable and PP, PH, NGL in independent variable list → Press Ok

Output:



Exercise 2. Practical exercise using SPSS for Survey Data

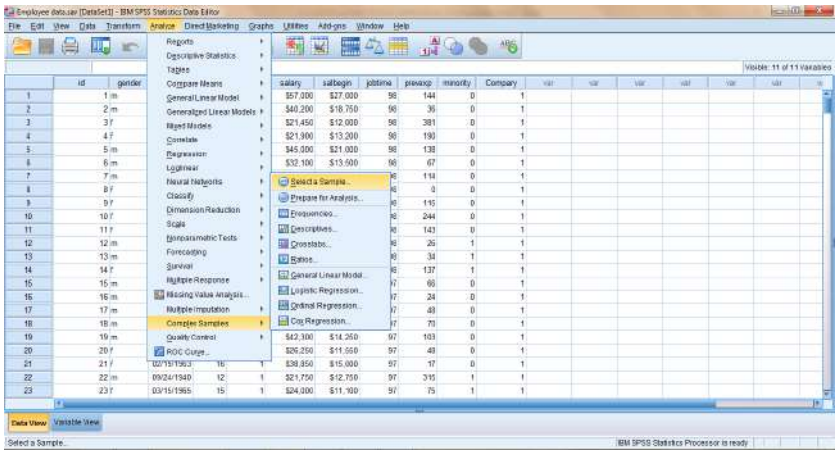
In this section, a practical exercise is provided which has analyzed using a popular statistical software, SPSS. For illustration purpose, we are going to use "Employee data" from the Sample folder of SPSS available at C:\Program Files (x86)\IBM\SPSS\Statistics\20\Samples\English. In addition a new variable "Company" has been added to the dataset which is having values from 1,2,…,10. Thus, there are 400 data-points clustered into 10 clusters each of size 40 usu. This dataset has been considered as population used for further illustration (available at https://www.dropbox.com/s/rxxccpuk3iieepa/Employee%20data.sav?dl=0).

The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind. The Complex Samples option allows to select a sample according to a complex design and incorporate the design specifications into the data analysis.

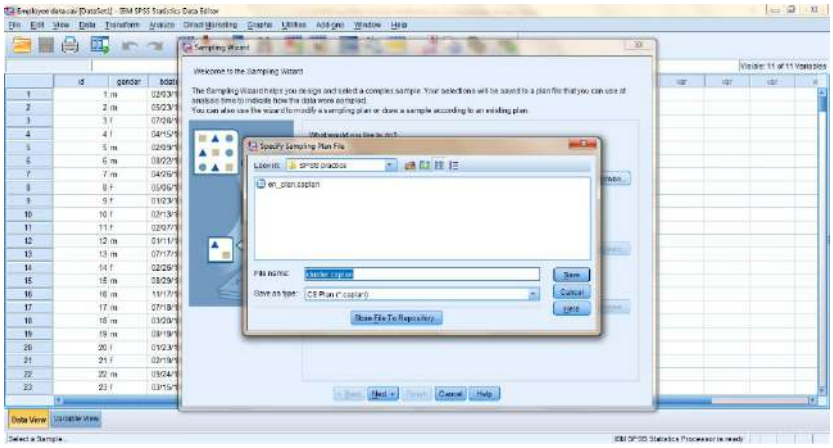
Creating a New Sample Plan

1. From the menus choose

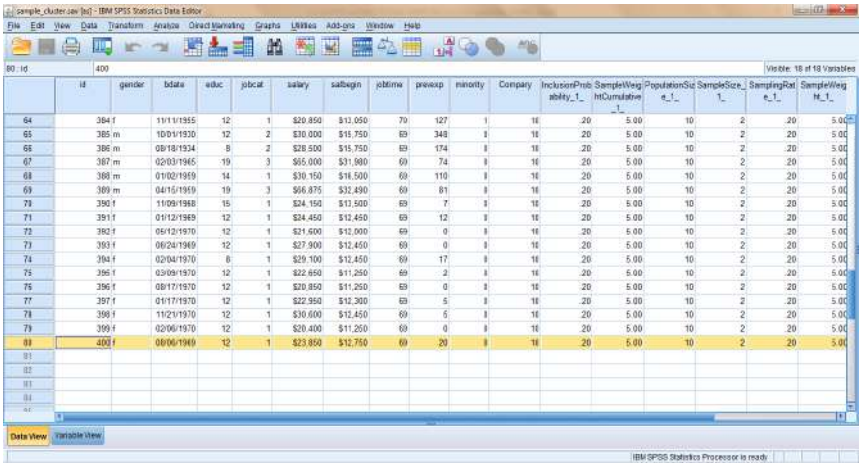
   Analyze → Complex Samples → Select a Sample....



2. Select *Design a sample* and choose a plan filename to save the sample plan.



3. Click *Next* to continue through the Wizard.
4. Optionally, in the Design Variables step, one can define strata, clusters, and input sample weights. Select the variable "Company" as cluster. Then, click Next.
5. Optionally, in the Sampling Method step, one can choose a method for selecting items.
    - If one select PPS Brewer or PPS Murthy, one can click Finish to draw the sample. Otherwise, click Next.
6. In the Sample Size step, specify the number or proportion of units to sample.
7. Optionally, in further steps one can:
    - Choose output variables to save.
    - Add a second or third stage to the design.
    - Set various selection options, including which stages to draw samples from, the random number seed, and whether to treat user-missing values as valid values of design variables.
    - Choose where to save output data.
8. Now click Finish to draw the sample.



Developed Sample Plan can be used for furthermore random sample selection as follows

Analyze → Complex Samples → Draw the Sample….

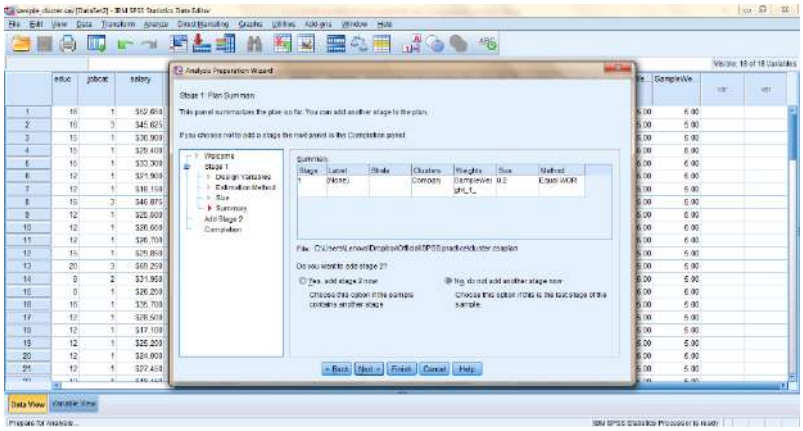
After selection of Sample next step is to prepare the sample for analysis. The Analysis Preparation Wizard guides through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, one should have a sample drawn according to a complex design.

Creating a New Analysis Plan

1. From the menus choose:

   Analyze → Complex Samples → Prepare for Analysis...

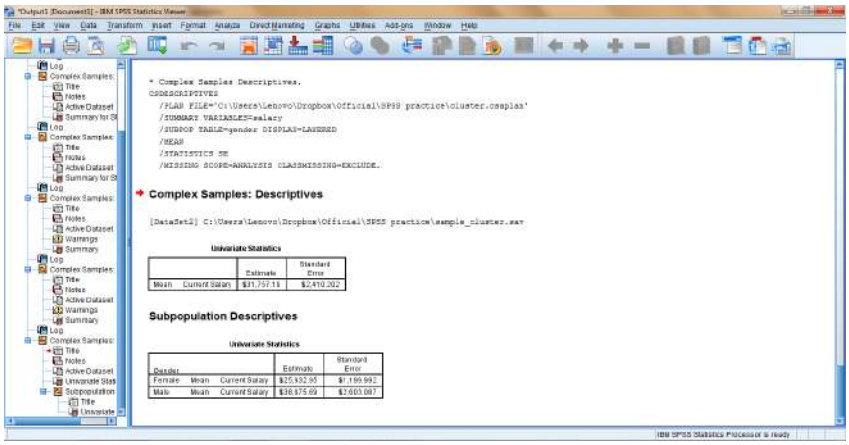
2. Select Create a plan file, and choose a plan filename to save the analysis plan.
3. Click Next to continue through the Wizard.
4. Specify the variable containing sample weights in the Design Variables step, optionally defining strata and clusters.
5. Optionally, in further steps one can:
   a. Select the method for estimating standard errors in the Estimation Method step.
   b. Specify the number of units sampled or the inclusion probability per unit in the Size step.
   c. Add a second or third stage to the design.
6. Now click Finish to save the plan.



Now using this Analysis Plan file one generates several types of outputs available in the *Complex Samples* option like

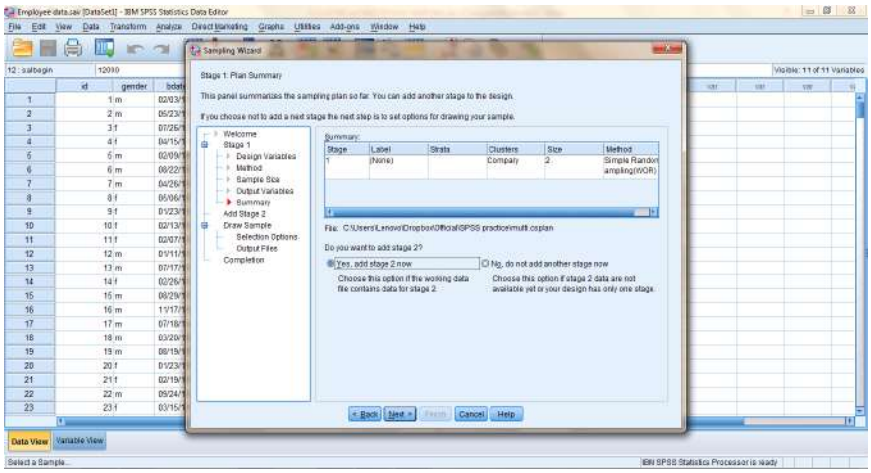
- Frequencies
- Descriptive
- Crosstabs
- Ratios
- General Linear Model
- Logistic Regression
- Ordinal Regression
- Cox Regression

Results from the Descriptive options using the "Current Salary" is given by

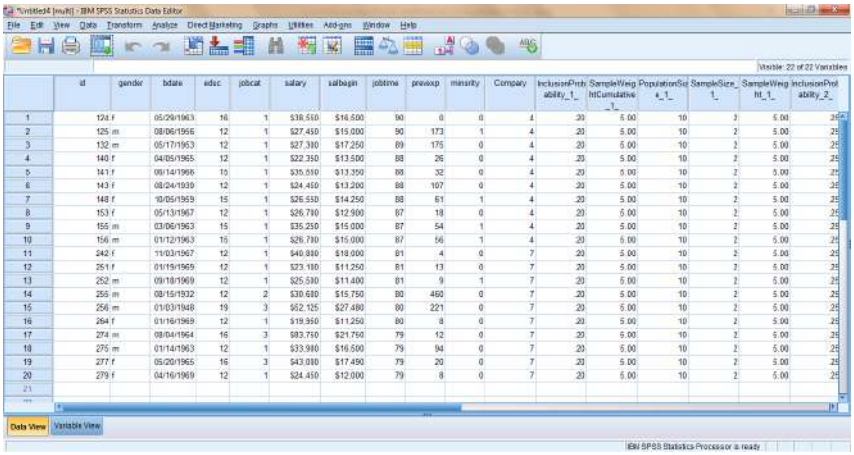


For selection of samples by Multistage sampling design one can edit the existing Sample Plan for cluster sampling or prepare new sampling plan according to Multistage sampling.

At the seventh step of the earliar shown "***Creating a New Sample Plan***", one should select "*Yes, add stage 2 now*" when the question "*Do you want to add Stage 2*" pops up in the sampling wizard as shown below:



Then define "*sample size*" for the stage 2 and path where to save the output file. An output file is given below when, first, 2 clusters are selected by SRSWOR and, the, within each selected cluster 10 units are selected by SRSWOR.



For analysis as per two stage sampling design, New Analysis Plan shall be created and further analysis of the sample shall be carried out.

REFERENCES:

1. Design Resources Server. Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on <05-05-2015>).
2. Morgan, G.A., Leech, N.L., Gloeckner G.W. and Barrett, K. C. (2012).*IBM SPSS for Introductory Statistics: Use and Interpretation*. Fifth Edition, Routledge.
3. Nie, N. H., Bent, D. H. and Hull, C. H.(1970). *SPSS: Statistical Package for the Social Sciences*. New York: McGraw-Hill.

ANALYSIS OF SURVEY DATA USING SPSS

Deepak Singh and Raju kumar

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. INTRODUCTION

SPSS is a widely used software package for statistical analysis in social science. SPSS is capable of handling large amounts of data and can perform all of the analyses covered in the text and much more. The current versions (2015) are officially named IBM SPSS Statistics. Long produced by SPSS Inc., it was acquired by IBM in 2009. During 2009 and 2010 it was called *PASW* (*Predictive Analytics Software*) *Statistics*. It is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with rather simple instructions. 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. Purpose of this chapter is to introduce the basic features of the SPSS for its application in survey data analysis.

When surveying a population, choosing a simple random sample may not be the best approach. A probability sample that uses strategies like stratification, clustering, and multistage sampling has many advantages over simple random sample under certain conditions like to increase precision, decrease cost, ensuring sub-populations are included etc. Under these situations, it is recommended to use techniques dedicated to producing correct estimates for complex sample data.

IBM SPSS Complex Samples can compute statistics and standard errors from complex sample designs by incorporating the designs into survey analysis. It offers planning tools such as stratified, clustered or multistage sampling. From the planning stage and sampling through the analysis stage, SPSS Complex Samples allows one to select a sample according to a complex design and incorporate the design specifications into the data analysis making it easy to obtain accurate and reliable results. SPSS Complex Samples considers up to three states when analyzing data from a multistage design therefore multistage analysis up to three stages is possible through it.

2. STATISTICAL PROCEDURE FOR SURVEY DATA ANALYSIS IN SPSS

Complex Samples submenu under the Analyze menu provides procedures for Sampling from Complex Designs and incorporate the design specifications into the data analysis, thus ensuring the results are valid. The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

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:

3.1 **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.

3.2 **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.

3.3 **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.

3.4 **NON RANDOM SAMPLING**

When selection at random is difficult to obtain, units can be sampled systematically (at a fixed interval) or sequentially.

3.5 **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.

3.6 **UNRESTRICTED SAMPLING**

Unrestricted sampling selects units with replacement (WR). Thus, an individual unit can be selected for the sample more than once.

3.7 **SAMPLING WEIGHTS**

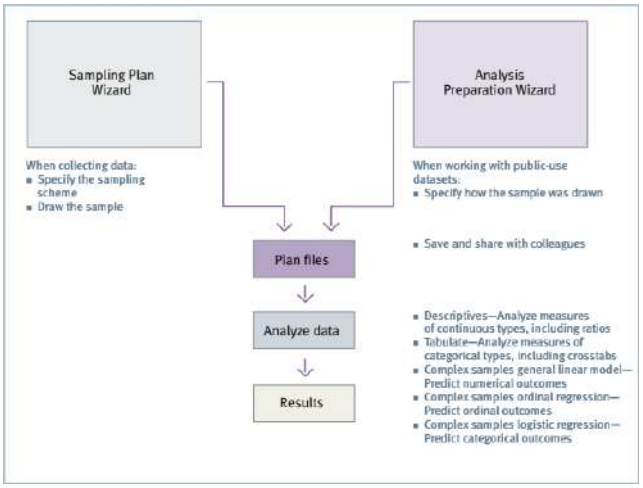
Sampling weights are automatically computed while drawing a complex sample and ideally correspond to the "frequency" that each sampling unit represents in the target population. Therefore, the sum of the weights over the sample should estimate the population size. Complex Samples analysis procedures 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.

4. 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 first step for SPSS Complex Samples is to use the wizards. If you are creating your own samples, use the Sampling Wizard to define the sampling scheme but if using datasets that have been sampled, such as those provided by the CDC, DHS surveys etc. use the Analysis Preparation Wizard to specify how the samples were defined and how to estimate standard errors. Once you create a sample or specify standard errors, you can create plans, analyze your data, and produce results.



SPSS complex samples helps to obtain correct estimates such as Population totals, means, ratios, standard errors, produce correct confidence intervals and hypothesis tests and predict outcomes.

4.1 **Complex Samples Plan (CSPLAN)**

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 need to supply design information to Complex Samples procedures. This information is stored in a **plan file** for easy reuse. CSPLAN does not actually extract the sample or analyze data.

PLAN FILES

A plan file contains complex sample specifications. There are two types of plan files:

Sampling Plan To sample cases, sample design created by CSPLAN is used as input to the CSSELECT (discussed next) procedure.

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. To analyze sample data, use an analysis design created by CSPLAN as input to the CSDESCRIPTIVES, CSTABULATE, CSGLM, CSLOGISTIC, or CSORDINAL procedures.

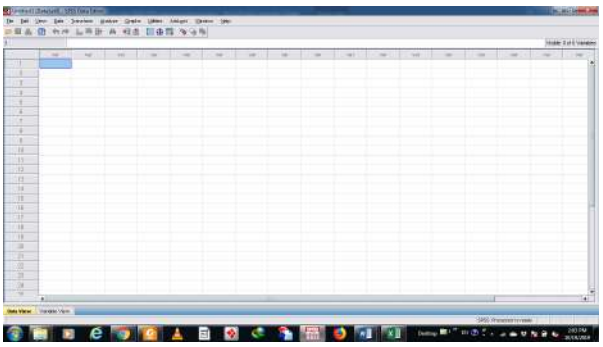
There are several advantages to saving your specifications in a plan file, including:

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.

Steps for Drawing the Sample and Analysis of Sampled Data

- START – IBM SPSS for windows



- Prepare a file from which data to be sampled in **SPSS Data Editor**
  or browse your data file by using following procedure:
  File - Open - Data

4.1.1 **SAMPLING WIZARD FOR COMPLEX DESIGN**

The sampling wizard is used to create, modify or executing the sampling plan file. We should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind for carrying this feature in SPSS.

Creating a New Sample Plan

1. From the menus choose:
   Analyze > **Complex Samples** > **Select a Sample...**
2. Select **Design a sample** and choose a plan filename to save the sample plan
   (demo_cs1)

3. Click Next to continue through the Wizard.
4. Optionally, in the Design Variables step, you can define strata, clusters, and input sample weights. After you define these, click Next.

5. Optionally, in the Sampling Method step, you can choose a method by which items can be selected like Simple random sampling with or without replacement, probability proportional to size etc. If you select **PPS Brewer** or **PPS Murthy**, you can click **Finish** to draw the sample. Otherwise, click **Next** and then:
6. In the Sample Size step, specify the number or proportion of units to sample. You can now click Finish to draw the sample.

4.1.2 **Sampling Wizard: Design Variables**

This step allows you to select stratification and clustering variables. In addition, if the current sample design is part of a larger sample design, you may have sample weights from a previous stage of the larger design. You can specify a numeric variable containing these weights in the first stage of the current design. Sample weights are computed automatically for subsequent stages of the current design.

This step has four options viz;

Stratify By; for stratification,

Clusters; for clustering variables,

Input sample weight; when sample weights of each unit are available

Stage label; for specifying an optional string label for each stage. This is used in the output to help identify stage wise information.

4.1.3 **Sampling Wizard: Sampling 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.

✓ **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.

4.1.4 **Sampling Wizard: Sample Size**

This step allows you to specify the number or proportion of units to sample within the current stage. The sample size can be fixed or it can vary across strata. For specifying sample size, clusters chosen in previous stages can be used to define strata.

- **Units.** You can specify an exact sample size or a proportion of units to sample.
- **Value.** A single value is applied to all strata. If **Counts** is selected as the unit metric,

you should enter a positive integer. If **Proportions** is selected, you should enter a non-negative value. Unless sampling with replacement, proportion values should also be no greater than 1.

- **Unequal values for strata.** Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows you to select a numeric variable that contains size values for strata.

4.1.5 **Sampling Wizard: Output Variables**

This step allows you to choose variables to save when the sample is drawn.

Population size. The estimated number of units in the population for a given stage. The root name for the saved variable is *Population Size*.

Sample proportion. The sampling rate at a given stage. The root name for the saved variable is *Sampling Rate_*.

Sample size. The number of units drawn at a given stage. The root name for the saved variable is *Sample Size_*.

Sample weight. The inverse of the inclusion probabilities. The root name for the saved variable is *Sample Weight_*.

Some stage wise variables are generated automatically. These include:
Inclusion probabilities. The proportion of units drawn at a given stage. The root name for the saved variable is *Inclusion Probability_*.

Cumulative weight. The cumulative sample weight over stages before and including the current one. The root name for the saved variable is *Sample Weight Cumulative_*.

Index. Identifies units selected multiple times within a given stage. The root name for the saved variable is *Index_*.

4.1.6 **Sampling Wizard: Plan Summary**

This is the last step within each stage, providing a summary of the sample design specifications through the current stage. From here, one can either proceed to the next stage (creating it, if necessary) or set options for drawing the sample.

4.2 **Complex Samples Selection (CSSELECT)**

This step selects complex, probability-based samples from a population. One can also control other sampling options, such as the random seed and missing-value handling. It chooses units according to a sample design created through the CSPLAN procedure.Write sampled units to an external file using an option to keep/drop specified variables. It has two sub-parts i.e. DRAW SAMPLE SELECTION OPTIONS and DRAW SAMPLE OUTPUT FILES, which are discussed below

4.2.1 **Sampling Wizard: Draw Sample Selection Options**

Draw sample. In addition to choosing whether to draw a sample, one can also choose to execute part of the sampling design. Stages must be drawn in order that is, stage 2 cannot be drawn unless stage 1 is also drawn. When editing or executing a plan, one cannot resample locked stages.

Seed. This allows one to choose a seed value for random number generation.

Include user-missing values. This determines whether user-missing values are valid. If so, user-missing values are treated as a separate category.

Data already sorted. If your sample frame is pre-sorted by the values of the stratification variables, this option allows one to speed the selection process.

4.2.2 **Sampling Wizard: Draw Sample Output Files**

This step allows one to choose where to direct sampled cases, weight variables, joint probabilities, and case selection rules.

Sample data. These options let one determine where sample output is written. It can be added to the active dataset, written to a new dataset, or saved to an external IBM® SPSS® Statistics data file. Datasets are available during the current session but are not available in subsequent sessions unless one explicitly save them as data files.

Joint probabilities. These options let one determine where joint probabilities are written. They are saved to an external SPSS Statistics data file. Joint probabilities are produced if the PPS WOR, PPS Brewer, PPS Sampford, or PPS Murthy method is selected and WR estimation is not specified.

Case selection rules. If one are constructing oner sample one stage at a time, one may want to save the case selection rules to a text file. They are useful for constructing the subframe for subsequent stages.

4.2.3 **Sampling Wizard: Finish**

This is the final step. One can save the plan file and draw the sample now or paste your selections into a syntax window.

4.3 **Preparing a Complex Sample for Analysis: The Analysis Preparation Wizard**

After selection of Sample next step is to prepare the sample for analysis. The Analysis Preparation Wizard guides one 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 do not have access to the sampling plan file used to draw the sample. If one do 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 your changes to a new file.

Complex Samples analysis procedures require analysis specifications from an analysis or sample plan file in order to provide valid results.

Plan. Specify the path of an analysis or sample plan file.

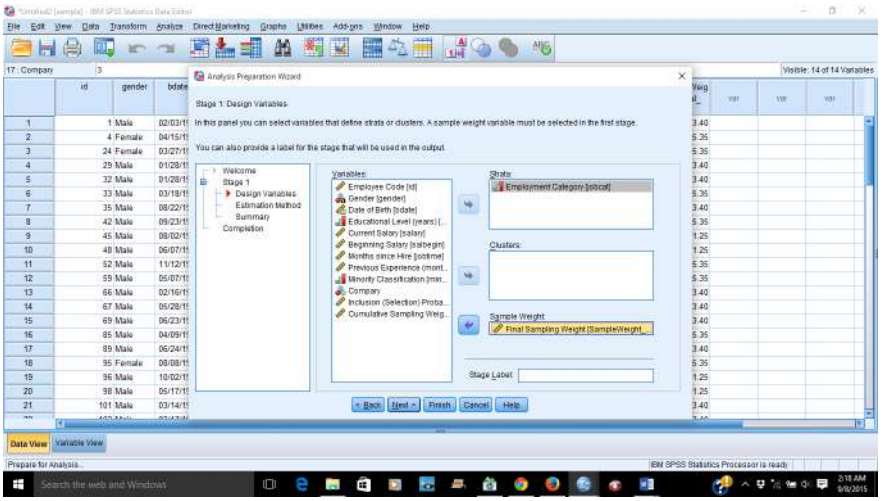
Joint Probabilities. In order to use Unequal WOR estimation for clusters drawn using a PPS WOR method, one need to specify a separate file or an open dataset containing the joint probabilities. This file or dataset is created by the Sampling Wizard during sampling.

Creating a New Analysis Plan

1. From the menus choose:

   Analyze → Complex Samples → Prepare for Analysis...

2. Select Create a plan file, and choose a plan filename to save the analysis plan.
3. Click Next to continue through the Wizard.
4. Specify the variable containing sample weights in the Design Variables step. Select the variable "Employee category" as strata.

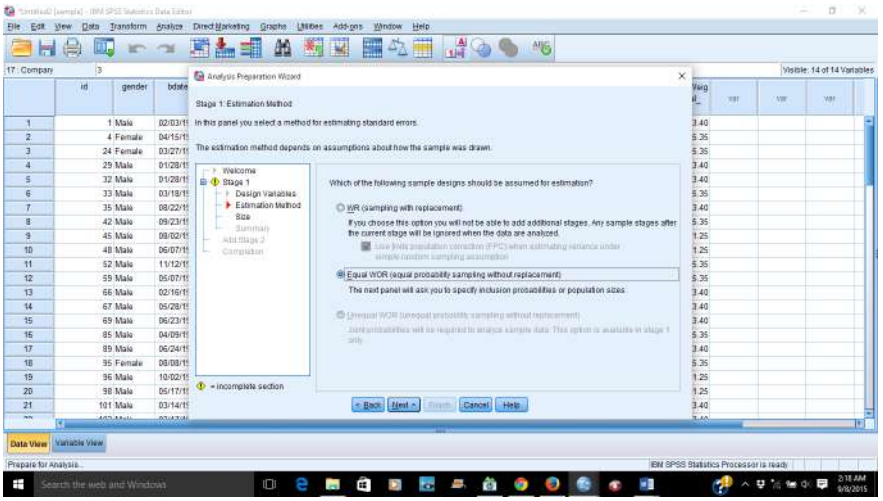


5. Optionally, in further steps one can:

a. Select the method for estimating standard errors in the Estimation Method step.

b. Specify the number of units sampled or the inclusion probability per unit in the Size step.

c. Add a second or third stage to the design.



6. Now click Finish to save the plan.

   Analysis Preparation Wizard: Design Variables

   This step allows one to identify the stratification and clustering variables and define sample weights. One can also provide a label for the stage.

Strata. The cross-classification of stratification variables defines distinct subpopulations, or strata. Oner total sample represents the combination of independent samples from each stratum.

Clusters. Cluster variables define groups of observational units, or clusters. Samples drawn in multiple stages select clusters in the earlier stages and then subsample units from the selected clusters. When analyzing a data file obtained by sampling clusters with replacement, one should include the duplication index as a cluster variable.

Sample Weight. One must provide sample weights in the first stage. Sample weights are computed automatically for subsequent stages of the current design.

Stage Label. One can specify an optional string label for each stage. This is used in the output to help identify stagewise information.

4.4 **Analysis Preparation Wizard: Estimation Method**

This step allows one to specify an estimation method for the stage.

WR (sampling with replacement). WR estimation does not include a correction for sampling from a finite population (FPC) when estimating the variance under the complex sampling design. One can choose to include or exclude the FPC when estimating the variance under simple random sampling (SRS).

Equal WOR (equal probability sampling without replacement). Equal WOR estimation includes the finite population correction and assumes that units are sampled with equal probability. Equal WOR can be specified in any stage of a design.

Unequal WOR (unequal probability sampling without replacement). In addition to using the finite population correction, Unequal WOR accounts for sampling units (usually clusters) selected with unequal probability. This estimation method is available only in the first stage.

4.5 **Analysis Preparation Wizard: Size**

This step is used to specify inclusion probabilities or population sizes for the current stage. Sizes can be fixed or can vary across strata. For the purpose of specifying sizes, clusters specified in previous stages can be used to define strata. Note that this step is necessary only when Equal WOR is chosen as the Estimation Method.

- **Units.** One can specify exact population sizes or the probabilities with which units were sampled.
- **Value.** A single value is applied to all strata. If ***Population Sizes*** is selected as the unit metric, one should enter a non-negative integer. If ***Inclusion Probabilities*** is selected, one should enter a value between 0 and 1, inclusive.
- **Unequal values for strata.** Allows one to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows one to select a numeric variable that contains size values for strata.

4.6 **Analysis Preparation Wizard: Plan Summary**

This is the last step within each stage, providing a summary of the analysis design specifications through the current stage. From here, one can either proceed to the next stage (creating it if necessary) or save the analysis specifications.

Analysis Preparation Wizard: Finish

This is the final step. One can save the plan file now or paste your selections to a syntax window.

When making changes to stages in the existing plan file, one can save the edited plan to a new file or overwrite the existing file. When adding stages without making changes to existing stages, the Wizard automatically overwrites the existing plan file. If one want to save the plan to a new file, choose to ***Paste the syntax generated by the Wizard into a syntax window*** and change the filename in the syntax commands.

4.7 **Analysis Preparation Wizard: Plan Summary**

This step allows one to review the analysis plan and remove stages from the plan.

Remove Stages. One can remove stages 2 and 3 from a multistage design. Since a plan must have at least one stage, one can edit but not remove stage 1 from the design.

5. Analysis Outputs

Now using this Analysis Plan file one generates several types of outputs available in the *Complex Samples* option like

Frequencies	General Linear Model
Descriptives	Logistic Regression
Crosstabs	Ordinal Regression
Ratios	Cox Regression

5.1 **Complex Samples Frequencies (CSFREQUENCIES)**

The Complex Samples Frequencies procedure produces frequency tables for selected variables and displays univariate statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables. Variables for which frequency tables are produced should be categorical. Subpopulation variables can be string or numeric but should be categorical.
Complex Samples Frequencies-

1. From the menus choose:
   Analyze > Complex Samples > Frequencies
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Click Continue.
4. In the Complex Samples Frequencies dialog box, select at least one frequency variable.

5.2 **Complex Samples Descriptives (CSDESCRIPTIVES)**

CSDESCRIPTIVES estimates means, sums, and ratios, and computes their standard

errors, design effects, confidence intervals, and hypothesis tests for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSDESCRIPTIVES performs analyses for subpopulations.

Descriptives-

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for Descriptive analysis
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples Descriptives Wizard – Measures - Sub Population – ok
5. Output- **SPSS Viewer**

5.3 Complex Samples Tabulate (CSTABULATE)

CSTABULATE displays one-way frequency tables or two-way cross tabulations and associated standard errors, design effects, confidence intervals, and hypothesis tests for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSTABULATE creates tables for subpopulations.

Crosstabs-

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for Crosstabs analysis Wizard
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples Crosstabs Wizard – Rows –Columns- Sub Population – ok
5. Output- **SPSS Viewer**

5.4 Complex Samples Ratios

The Complex Samples Ratios procedure displays univariate summary statistics for ratios of variables. Optionally, one can request statistics by subgroups, defined by one or more categorical variables.

Statistics. The procedure produces ratio estimates, *t* tests, standard errors, confidence intervals, coefficients of variation, unweighted counts, population sizes, design effects, and square roots of design effects.

Data. Numerators and denominators should be positive-valued scale variables. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Complex Samples Ratios-

1. From the menus choose:
2. Analyze > Complex Samples > Ratios...

3. Select a plan file. Optionally, select a custom joint probabilities file.
4. Click ***Continue***.
5. Select at least one numerator variable and denominator variable, here take “income” and “Age” respectively.
6. Optionally, one can specify variables to define subgroups for which statistics are produced, here take “ed”(ed is for sducation).

5.5 Complex Samples General Linear Model (CSGLM)

This procedure enables you to build linear regression, analysis of variance (ANOVA), and analysis of covariance (ANCOVA) models for samples drawn using complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSGLM performs analyses for subpopulations.

General Linear Model-

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for General Linear Model
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples General Linear Model Wizard – Dependent variable –Factors- Covariates – Subpopuation variable (Category, if category wise analysis is required)-ok
5. Output- **SPSS Viewer**

5.6 Complex Samples Ordinal (CSORDINAL)

CSORDINAL performs regression analysis on a binary or ordinal polychromous dependent variable using the selected cumulative link function for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, as well as WR and WOR sampling procedures. Optionally, CSORDINAL performs analyses for a subpopulation.

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for Complex Samples for Ordinal regression
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples for Ordinal regression– Dependent variable –Factors- Covariates – link function- Subpopulation variable (Category, if category wise analysis is required)-ok
5. Output- **SPSS Viewer**

5.7 Complex Samples Logistic Regression (CSLOGISTIC)

This procedure performs binary logistic regression analysis, as well as multinomial logistic regression (MLR) analysis, for samples drawn by complex sampling methods. CSLOGISTIC estimates variances by taking into account the sample design used to

select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSLOGISTIC performs analyses for subpopulations.

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Logistic Regression
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples for Ordinal regression– Dependent variable –Factors- Covariates – link function- Subpopulation variable (Category, if category wise analysis is required)-ok
5. Output- **SPSS Viewer**

5.8 **Complex Samples Cox Regression**

The Complex Samples Cox Regression procedure performs survival analysis for samples drawn by complex sampling methods. Optionally, one can request analyses for a subpopulation.

Examples. A government law enforcement agency is concerned about recidivism rates in their area of jurisdiction. One of the measures of recidivism is the time until second arrest for offenders. The agency would like to model time to re-arrest using Cox Regression but are worried the proportional hazards assumption is invalid across age categories.

To Obtain Complex Samples Cox Regression

This feature requires the Complex Samples option.

From the menus choose:

Analyze > ***Complex Samples*** > ***Cox Regression***...

- Select a plan file. Optionally, select a custom joint probabilities file.
- Click ***Continue***.
- Specify the survival time by selecting the entry and exit times from the study.
- Select an event status variable.
- Click ***Define Event*** and define at least one event value.

SAS – AN OVERVIEW

Ankur Biswas

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

SAS is a collection of modules that are used to process and analyze data. It began in the late ’60s and early ’70s as a statistical package (the name *SAS* originally stood for Statistical Analysis System). However, unlike many competing statistical packages, SAS is also an extremely powerful, general-purpose programming language. SAS is a predominant software in many industries. In recent years, it has been enhanced to provide state-of-the-art statistical tools for analysis. The only way to really learn a programming language is to write lots of programs, make some errors, correct the errors, and then make some more. If one already has access to SAS at work or school, he/she is ready to go. SAS Learning Edition 4.1 is useful for those who are learning SAS by themselves and do not have a copy of SAS to play with. This is a relatively inexpensive, fully functional version of SAS.

2. Getting Data into SAS

SAS can read data from almost any source. Common sources of data are raw text files, Microsoft Office Excel spreadsheets, Access databases, and most of the common database systems such as DB2 and Oracle. Most of this book uses either text files or Excel spreadsheets as data sources.

3. Components of SAS Programs

SAS programs often contain DATA steps and PROC steps. DATA steps are parts of the program where you can read or write the data, manipulate the data, and perform calculations. PROC (short for procedure) steps are parts of your program where you ask SAS to run one or more of its procedures to produce reports, summarize the data, generate graphs, and much more. DATA steps begin with the word DATA and PROC steps begin with the word PROC. Most DATA and PROC steps end with a RUN statement. SAS processes each DATA or PROC step completely and then goes on to the next step.

SAS also contains *global* statements that affect the entire SAS environment and remain in effect from one DATA or PROC step to another. In the program above, the OPTIONS and TITLE statements are examples of global statements. It is important to keep in mind that the actions of global statements remain in effect until they are changed by another global statement or until you end your SAS session.

All SAS programs, whether part of DATA or PROC steps, are made up of statements. Here is the rule: all SAS statements end with semicolons. This is an important rule because if you leave out a semicolon where one is needed, the program may not run correctly, resulting in hard-to-interpret error messages. Let’s discuss some of the basic rules of SAS statements. First, they can begin in any column and can span several lines, if necessary. Because a semicolon determines the end of a SAS statement, you can place more than one statement on a single line (although this is not recommended as a matter of style).

SAS is not case sensitive. Well, this is almost true. Of course references to external files must match the rules of your particular operating system. So, if you are running SAS under UNIX or Linux, file names will be case-sensitive. As you will see later, you get to name the variables in a SAS data set. The variable names in Program 1 are Name, Code, Days,

Number, Price, and CostPerSeed. Although SAS doesn't care whether you write these names in uppercase, lowercase, or mixed case, it does "remember" the case of each variable the *first* time it encounters that variable and uses that form of the variable name when producing printed reports.

4. SAS Names

SAS names follow a simple naming rule. All SAS variable names and data set names can be no longer than 32 characters and must begin with a letter or the underscore (_) character. The remaining characters in the name may be letters, digits, or the underscore character. Characters such as dashes and spaces are not allowed. Here are some valid and invalid SAS names.

Valid SAS Names

Parts, LastName, First_Name, Ques5, Cost_per_Pound, DATE, time, X12Y34Z56

Invalid SAS Names

8_is_enough - Begins with a number,

Price per Pound - Contains blanks,

Month-total - Contains an invalid character (-),

Num% - Contains an invalid character (%),

5. SAS Data Sets and SAS Data Types

When SAS reads data from anywhere (for example, raw data, spreadsheets), it stores the data in its own special form called a SAS data set. Only SAS can read and write SAS data sets. If you opened a SAS data set with another program (Microsoft Word, for example), it would not be a pretty sight. Even if SAS is reading data from Oracle tables or DB2, it is actually converting the data into SAS data set format in the background. The good news is that you don't ever have to worry about how SAS is storing its data or the structure of a SAS data set. However, it is important to understand that SAS data sets contain two parts: a descriptor portion and a data portion. Not only does SAS store the actual data values for you, it stores information about these values (things like storage lengths, labels, and formats). SAS has two types of variables: *character* and *numeric*. This makes it much simpler to use and understand than some other programs that have many more data types (for example, integer, long integer, and logical).

6. The SAS Display Manager and SAS Enterprise Guide

Because SAS runs on many different platforms (mainframes, microcomputers running various Microsoft operating systems, UNIX, and Linux), the way you write and run programs will vary. You might use a general-purpose text editor on a mainframe to write a SAS program, submit it, and send the output back to a terminal or to a file. On PCs, you might use the SAS Display Manager, where you write your program in the *Enhanced Editor (Editor* window*)*, see any error messages and comments about your program and the data in the *Log* window, and view your output in the *Output* window. In addition to the Enhanced Editor, an older program, simply called the *Program Editor*, is available for Windows and UNIX users. As an alternative to the Display Manager, you may enter the SAS environment using *SAS Enterprise Guide*, which is a front-end to SAS that allows you to use a menu-driven system to write SAS programs and produce reports. There are many excellent books published by SAS that offer detailed instructions on how to run SAS programs on each specific platform and the appropriate access method into SAS.

7. A Sample SAS Program

Let's start out with a simple SAS program that reads data from a text file and produces some basic reports to give you an overview of the structure of SAS programs. For this example, we have a text file with data on vegetable seeds. Each line of the file contains the following pieces of information (separated by spaces):

- Vegetable name
- Product code
- Days to germination
- Number of seeds
- Price

In SAS terminology, each piece of information is called a *variable*. (Other database systems, and sometimes SAS, use the term *column*.) A few sample lines from the file are shown here:

File **c:\my folder\crop.txt**

Crop_1 50104 55 30 195

Crop_1 51789 56 30 225

Crop_2 50179 68 150 395

Crop_2 50872 65 150 225

Crop_3 57224 75 200 295

Crop_3 62471 80 200 395

Crop_3 57828 66 200 295

Crop_4 52233 70 30 225

In this example, each line of data produces what SAS calls an *observation* (also referred to as a *row* in other systems). A complete SAS program to read this data file and produce a list of the data, a frequency count showing the number of entries for each crop, the average price per seed, and the average number of days until germination is shown here.

Program - A sample SAS program

```
*Comment 1: SAS Program to read veggie data file and to produce several reports;
    * Comment 2: Entering data using program editor;
    data crop;
        input Name $ Code Days Number Price;
            *$ for character variable;
        cards;
        Crop_1 50104 55 30 195
        Crop_1 51789 56 30 225
        Crop_2 50179 68 150 395
```

```
            Crop_2 50872 65 150 225
            Crop_3 57224 75 200 295
            Crop_3 62471 80 200 395
            Crop_3 57828 66 200 295
            Crop_4 52233 70 30 225
            ;
Run;

* Comment 3: To print the inserted data;
title "List of the Raw Data";
footnote "Overview of SAS";
proc print data= crop;
run;

* Comment 4: Alternative way for running data;
DATA crop;
            input Name $ Code Days Number Price @@;
            cards;
            Crop_1 50104 55 30 195 Crop_1 51789 56 30 225
            Crop_2 50179 68 150 395 Crop_2 50872 65 150 225
            Crop_3 57224 75 200 295 Crop_3 62471 80 200 395
            Crop_3 57828 66 200 295 Crop_4 52233 70 30 225
            ;
Run;

* Comment 5: To import from external sources - txt;
data crop;
            infile "c:\mywork\crop.txt";
            input Name $ Code Days Number Price;
run;

* Comment 6: To import from external sources - csv;
data crop;
            infile 'c:\mywork\crop.csv'  dlm=',' ;
            input Name $ Code Days Number Price;
```

```
run;

* Comment 7: Alternative way using IMPORT procedure*/
proc import datafile = 'c:\mywork\crop.csv'
        out = crop dbms=csv replace;
        getnames=no;
run;

* Comment 8: To import from external sources - xls */
proc import datafile = 'c:\mywork\crop.xls'
        out = crop dbms=excel replace;
        getnames=yes;
run;
* Comment 9: To modify the data;
data crop;
        set crop;
        CostPerSeed = Price / Number;*add new variable;
        *drop days;                    *delete variable;
        rename Number=Number_seeds;  *change variable name;
run;

* Comment 10: To sort the data;
proc sort data=crop;
        by Code;
run;

* Comment 11: To find the frequency counts;
title "Frequency Distribution of crop Names";
proc freq data= crop;
        tables Name;
run;

* Comment 12: To find means of the variables;
title "Average Cost of Seeds";
proc means data= crop;
```

```
        var Price CostPerSeed;
run;

* Comment 13: To find Scatter Plot;
proc plot data = crop;
        plot Days*Price = '*';
run;

* Comment 14: To fit linear regression;
proc reg data = crop;
        model Price = Days;
run;
```

References

Cody, R. (2018). *Learning SAS® by Example: A Programmer's Guide*. 2nd ed., Cary NC: SAS Institute Inc.

Delwiche, L. D., & Slaughter, S. J. (2002). *The little SAS book: A primer*. Cary, NC: SAS Institute.

ANALYSIS OF SURVEY DATA USING SAS

Ankur Biswas

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

A sampling method is a scientific and objective procedure of selecting units from the population and provides a sample that is expected to be representative of the population. A sampling method makes it possible to estimate the population parameters while reducing at the same time the size of survey operations. Some of the advantages of sample surveys as compared to complete enumeration are reduction in cost, greater speed, wider scope and higher accuracy. A function of the unit values of the sample is called an estimator. Various measures, like bias, mean square errors, variance etc. are used to assess the performance of the estimator. See Lohr (2010), Kalton (1983), Sukhatme *et al.* (1984), Cochran (1977), Murthy (1977), Raj (1968) and Kish (1965) for more information about statistical sampling and analysis of complex survey data.

The prime objective of a sample survey is to obtain inferences about the characteristic of a population. Population is defined as a group of units defined according to the objectives of the survey. The population may consist of all the households in a village / locality, all the fields under a particular crop. We may also consider a population of persons, families, fields, animals in a region, or a population of trees, birds in a forest depending upon the nature of data required. The information that we seek about the population is normally, the total number of units, aggregate values of various characteristics, averages of these characteristics per unit, proportions of units possessing specified attributes etc. The data can be collected in two different ways. The first one is complete enumeration which means collection of data on the survey characteristics from each unit of the population.

The main problem in sample surveys is the choice of a proper sampling strategy, which essentially comprise of a sampling method and the estimation procedure. In the choice of a sampling method there are some methods of selection while some others are control measures which help in grouping the population before the selection process. In the methods of selection, schemes such as simple random sampling, systematic sampling and varying probability sampling are generally used. Among the control measures are procedures such as stratified sampling, cluster sampling and multi-stage sampling etc. A combination of control measures along with the method of selection is called the sampling scheme.

2. Use of SAS Software for Survey Data Analysis

Researchers often use sample survey methodology to obtain information about a large population by selecting and measuring a sample from that population. Due to variability among items, researchers apply scientific probability-based designs to select the sample. This reduces the risk of a distorted view of the population and enables statistically valid inferences to be made from the sample. To select probability-based random samples from a study population, you can use the SURVEYSELECT procedure, which provides a variety of methods for probability sampling. To analyze sample survey data, you can use the SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures, which incorporate the sample design into the analyses.

Many SAS/STAT procedures, such as the MEANS, FREQ, GLM, LOGISTIC, and PHREG procedures, can compute sample means, produce crosstabulation tables, and estimate regression relationships. However, in most of these procedures, statistical

inference is based on the assumption that the sample is drawn from an infinite population by simple random sampling. If the sample is in fact selected from a finite population by using a complex survey design, these procedures generally do not calculate the estimates and their variances according to the design actually used. Using analyses that are not appropriate for your sample design can lead to incorrect statistical inferences.

The SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures properly analyze complex survey data by taking into account the sample design. These procedures can be used for multistage or single-stage designs, with or without stratification, and with or without unequal weighting. The survey analysis procedures provide a choice of variance estimation methods, which include Taylor series linearization, balanced repeated replication (BRR), and the jackknife.

2.1 Proc SURVEYSELECT Procedure

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, which is the list of units from which the sample is to be selected. The sampling units can be individual observations or groups of observations (clusters). You also specify the selection method, the desired sample size or sampling rate, and other selection parameters. PROC SURVEYSELECT selects the sample and produces an output data set that contains the selected units, their selection probabilities, and their sampling weights. When you select 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.

PROC SURVEYSELECT 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.

PROC SURVEYSELECT provides the following equal probability sampling methods:

- simple random sampling (without replacement)
- unrestricted random sampling (with replacement)
- systematic random sampling
- sequential random sampling

This procedure also provides the following probability proportional to size (PPS) sampling 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. PROC SURVEYSELECT can perform stratified sampling by selecting samples independently within strata, which are 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 use a systematic or sequential selection method, PROC SURVEYSELECT can also sort by control variables within strata for the additional control of implicit stratification.

For stratified sampling, PROC SURVEYSELECT provides survey design methods to allocate the total sample size among the strata. Available allocation methods include proportional, Neyman, and optimal allocation. Optimal allocation maximizes the estimation precision within the available resources, taking into account stratum sizes, costs, and variances.

PROC SURVEYSELECT provides replicated sampling, where the total sample is composed of a set of replicates, and each replicate is 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 estimate standard errors for combined sample estimates and to perform a variety of other resampling and simulation tasks.

Simple Random Sampling

The following PROC SURVEYSELECT statements select a probability sample of customers from the '*Customers'* data set by using simple random sampling:

```
title1 'Customer Satisfaction Survey';
title2 'Simple Random Sampling';
proc surveyselect data=Customers method=srs n=100
        out=SampleSRS;
run;
```

The PROC SURVEYSELECT statement invokes the procedure. The DATA= option names the SAS data set Customers as the input data set from which to select the sample. The METHOD=SRS option specifies simple random sampling as the sample selection method. In simple random sampling, each unit has an equal probability of selection, and sampling is without replacement. Without-replacement sampling means that a unit cannot be selected more than once. The N=100 option specifies a sample size of 100 customers. The OUT= option stores the sample in the SAS data set named SampleSRS.

Figure 1 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A sample of 100 customers is selected from the data set Customers by simple random sampling. With simple random sampling and no stratification in the sample design, the selection probability is the same for all units in the sample. In this sample, the selection probability for each customer equals 0.007423, which is the sample size (100) divided by the population size (13,471). The sampling weight equals 134.71 for each customer in the sample, where the weight is the inverse of the selection probability. If you specify the STATS option, PROC SURVEYSELECT includes the selection probabilities and sampling weights in the output data set. (This information is always included in the output data set for more complex designs.)

The random number seed is 39647. PROC SURVEYSELECT uses this number as the initial seed for random number generation. Because the SEED= option is not specified in the PROC SURVEYSELECT statement, the seed value is obtained by using the time of day from the computer's clock. You can specify SEED=39647 to reproduce this sample.

Figure 1. Sample Selection Summary

Customer Satisfaction Survey

Simple Random Sampling

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Input Data Set	CUSTOMERS
Random Number Seed	39647
Sample Size	100
Selection Probability	0.007423
Sampling Weight	134.71
Output Data Set	SAMPLESRS

The sample of 100 customers is stored in the SAS data set SampleSRS. PROC SURVEYSELECT does not display this output data set. The following PROC PRINT statements display the first 20 observations of SampleSRS:

```
title1 'Customer Satisfaction Survey';
title2 'Sample of 100 Customers, Selected by SRS';
title3 '(First 20 Observations)';
proc print data=SampleSRS(obs=20);
run;
```

Figure 2 displays the first 20 observations of the output data set SampleSRS, which contains the sample of customers. This data set includes all the variables from the DATA= input data set Customers. If you do not want to include all variables, you can use the ID statement to specify which variables to copy from the input data set to the output (sample) data set.

Figure 2. Customer Sample (First 20 Observations)

Customer Satisfaction Survey
Sample of 100 Customers, Selected by SRS
(First 20 Observations)

Obs	CustomerID	State	Type	Usage
1	036-89-0212	FL	New	74
2	045-53-3676	AL	New	411
3	050-99-2380	GA	Old	167
4	066-93-5368	AL	Old	1232
5	082-99-9234	FL	New	90
6	097-17-4766	FL	Old	131
7	110-73-1051	FL	Old	102
8	111-91-6424	GA	New	247
9	127-39-4594	GA	New	61
10	162-50-3866	FL	New	100
11	162-56-1370	FL	New	224
12	167-21-6808	SC	New	60
13	168-02-5189	AL	Old	7553
14	174-07-8711	FL	New	284
15	187-03-7510	SC	New	21
16	190-78-5019	GA	New	185
17	200-75-0054	GA	New	224
18	201-14-1003	GA	Old	3437
19	207-15-7701	GA	Old	24
20	211-14-1373	AL	Old	88

Stratified Sampling

In this section, stratification is added to the sample design for the customer satisfaction survey. The sampling frame, which is the list of all customers, is stratified by *State* and *Type*. This divides the sampling frame into non-overlapping subgroups formed from the

values of the State and Type variables. Samples are then selected independently within the strata.

PROC SURVEYSELECT requires that the input data set be sorted by the STRATA variables. The following PROC SORT statements sort the *Customers* data set by the stratification variables *State* and *Type*:

```
proc sort data=Customers;
  by State Type;
run;
```

The following PROC SURVEYSELECT statements select a probability sample of customers from the *Customers* data set according to the stratified sample design:

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling';
proc surveyselect data=Customers method=srs n=15
                  seed=1953 out=SampleStrata;
  strata State Type;
run;
```

The STRATA statement names the stratification variables State and Type. In the PROC SURVEYSELECT statement, the METHOD=SRS option specifies simple random sampling. The N=15 option specifies a sample size of 15 customers for each stratum. If you want to specify different sample sizes for different strata, you can use the N=SAS-data-set option to name a secondary data set that contains the stratum sample sizes. The SEED=1953 option specifies '1953' as the initial seed for random number generation. Figure 3 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 120 customers are selected.

Figure 3. Sample Selection Summary

Customer Satisfaction Survey
Stratified Sampling

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Strata Variables	State
	Type
Input Data Set	CUSTOMERS
Random Number Seed	1953
Stratum Sample Size	15

Number of Strata	8
Total Sample Size	120
Output Data Set	SAMPLESTRATA

2.2 Proc SURVEYMEANS Procedure

The SURVEYMEANS procedure produces estimates of population means and totals from sample survey data. The procedure also computes estimates of proportions for categorical variables, estimates of quantiles for continuous variables, and ratio estimates of means and proportions. For all of these statistics, PROC SURVEYMEANS provides standard errors, confidence limits, and *t* tests.

PROC SURVEYMEANS provides domain analysis, which computes estimates for domains (subpopulations), in addition to analysis for the entire study population. Formation of subpopulations can be unrelated to the sample design, and so the domain sample sizes can actually be random variables. Domain analysis takes this variability into account by using the entire sample to estimate the variance of domain estimates. Domain analysis is also known as subgroup analysis, subpopulation analysis, and subdomain analysis.

Simple Random Sampling

This example illustrates how you can use PROC SURVEYMEANS to estimate population means and proportions from sample survey data. The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. Researchers want to know how much these students spend weekly for ice cream, on average, and what percentage of students spend at least $10 weekly for ice cream.

To answer these questions, 40 students were selected from the entire student population by using simple random sampling (SRS). Selection by simple random sampling means that all students have an equal chance of being selected and no student can be selected more than once. Each student selected for the sample was asked how much he or she spends for ice cream per week, on average. The SAS data set *IceCream* saves the responses of the 40 students:

```
data IceCream;
  input Grade Spending @@;
  if (Spending < 10) then Group='less';
  else Group='more';
  datalines;
7 7  7  7  8 12  9 10  7  1  7 10  7  3  8 20  8 19  7 2
7 2  9 15  8 16  7  6  7  6  7  6  9 15  8 17  8 14  9 8
9 8  9  7  7  3  7 12  7  4  9 14  8 18  9  9  7  2  7 1
7 4  7 11  9  8  8 10  8 13  7  2  9  6  9 11  7  2  7 9
;
```

The variable Grade contains a student's grade. The variable Spending contains a student's response regarding how much he spends per week for ice cream, in dollars. The variable Group is created to indicate whether a student spends at least $10 weekly for ice cream: Group='more' if a student spends at least $10, or Group='less' if a student spends less than $10.

You can use PROC SURVEYMEANS to produce estimates for the entire student population, based on this random sample of 40 students:

```
title1 'Analysis of Ice Cream Spending';
title2 'Simple Random Sample Design';
proc surveymeans data=IceCream total=4000;
  var Spending Group;
run;
```

The PROC SURVEYMEANS statement invokes the procedure. The TOTAL=4000 option specifies the total number of students in the study population, or school. The procedure uses this total to adjust variance estimates for the effects of sampling from a finite population. The VAR statement names the variables to analyze, Spending and Group. Figure 4 displays the results from this analysis. There are a total of 40 observations used in the analysis. The "Class Level Information" table lists the two levels of the variable Group. This variable is a character variable, and so PROC SURVEYMEANS provides a categorical analysis for it, estimating the relative frequency or proportion for each level. If you want a categorical analysis for a numeric variable, you can name that variable in the CLASS statement.

Figure 4. Analysis of Ice Cream Spending

Analysis of Ice Cream Spending
Simple Random Sample Design

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	40

Class Level Information		
Class Variable	Levels	Values
Group	2	less more

Statistics

Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
Spending		40	8.750000	0.845139	7.04054539	10.4594546
Group	**less**	23	0.575000	0.078761	0.41568994	0.7343101
	more	17	0.425000	0.078761	0.26568994	0.5843101

2.3 The SURVEYFREQ Procedure

This procedure produces one-way to *n*-way frequency and cross tabulation 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 provides a variety of options to customize the table display.

2.4 PROC SURVEYREG PROCEDURE

The SURVEYREG procedure performs regression analysis for sample survey data. The procedure fits linear models and computes regression coefficients and their variance-covariance matrix. The procedure enables you to specify classification effects by using the same syntax as in the GLM procedure.

2.5 PROC SURVEYLOGISTIC PROCEDURE

The SURVEYLOGISTIC procedure provides logistic regression analysis for sample survey data. Logistic regression analysis investigates the relationship between discrete responses and a set of explanatory variables. PROC SURVEYLOGISTIC fits linear logistic regression models for discrete response survey data by the method of maximum likelihood and incorporates the sample design into the analysis. The SURVEYLOGISTIC procedure enables you to specify categorical classification variables (also known as CLASS variables) as explanatory variables in the model by using the same syntax for main effects and interactions as in the GLM and LOGISTIC procedures.

Table 1: Survey Sampling and Analysis Procedures in SAS/STAT Software

PROC SURVEYSELECT	
Selection Methods	Simple random sampling (without replacement)
	Unrestricted random sampling (with replacement)
	Systematic
	Sequential
	Probability proportional to size (PPS) sampling, with and without replacement
	PPS systematic
	PPS for two units per stratum
	PPS sequential with minimum replacement
Allocation Methods	Proportional
	Optimal

	Neyman
Sampling Tools	Cluster sampling
	Replicated sampling
	Serpentine sorting
PROC SURVEYMEANS	
Statistics	Estimates of population means and totals
	Estimates of population proportions
	Estimates of population quantiles
	Ratio estimates
	Standard errors
	Confidence limits
	Hypothesis tests
	Domain analysis
PROC SURVEYFREQ	
Tables	One-way frequency tables
	Two-way and multiway crosstabulation tables
	Estimates of population totals and proportions
	Standard errors
	Confidence limits
Analyses	Tests of goodness of fit
	Tests of independence
	Risks and risk differences
	Odds ratios and relative risks
Graphics	Weighted frequency and percent plots
	Odds ratio, relative risk, and risk difference plots
PROC SURVEYREG	
Analyses	Linear regression model fitting
	Regression coefficients
	Covariance matrices
	Confidence limits
	Hypothesis tests
	Estimable functions
	Contrasts
	Least squares means (LS-means) of effects
	Custom hypothesis tests among LS-means
	Regression with constructed effects
	Predicted values and residuals
	Domain analysis
PROC SURVEYLOGISTIC	
Analyses	Cumulative logit regression model fitting
	Logit, probit, and complementary log-log link functions

Generalized logit regression model fitting
Regression coefficients
Covariance matrices
Confidence limits
Hypothesis tests
Odds ratios
Estimable functions
Contrasts
Least squares means (LS-means) of effects
Custom hypothesis tests among LS-means
Regression with constructed effects
Model diagnostics
Domain analysis

References

Cochran, W. G. (1977). *Sampling Techniques*. Third Edition. John Wiley and Sons.

Raj, D. (1968). *Sampling Theory*. TATA McGRAW-HILL Publishing Co. Ltd.

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills, CA and London: Sage Publications.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, Second Edition, Pacific Grove, CA: Duxbury Press.

Murthy, M.N. (1977). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Singh, Daroga and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. Wiley Eastern Limited.

Sukhatme, P. V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Third Revised Edition, Iowa State University Press, USA.

Digital India
Power To Empower

हर कदम, हर डगर
किसानों का हमसफर
भारतीय कृषि अनुसंधान परिषद
Agrisearch with a human touch

स्वच्छ भारत
एक कदम स्वच्छता की ओर

भा.कृ.अनु.प. - भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली-110012

ICAR-INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE
LIBRARY AVENUE, PUSA, NEW DELHI-110012
https://iasri.icar.gov.in