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

उच्च संकाय प्रशिक्षण केन्द्र
CENTRE OF ADVANCED FACULTY TRAINING

सर्वेक्षण अभिकल्पना में आधुनिक प्रगति एवं सांख्यिकीय सॉफ्टवेयर द्वारा सर्वेक्षण आँकड़ों का विश्लेषण

Recent Advances in Survey Design and Analysis of Survey Data Using Statistical Software

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

संदर्भ पुस्तिका - 1
REFERENCE MANUAL-1

प्रतिदर्श सर्वेक्षण प्रभाग
DIVISION OF SAMPLE SURVEY
भाकृअनुप-भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली–110 012

भाकृअनुप
ICAR

ICAR-INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE
LIBRARY AVENUE, PUSA, NEW DELHI – 110 012
www.iasri.res.in

2014

प्राक्कथन

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

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

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

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

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

नई दिल्ली–110 012
28 अक्तूबर, 2014

FOREWORD

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

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

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

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

U C Sud
Director, IASRI

New Delhi-110012
October 28, 2014

आमुख

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

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

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

नई दिल्ली
28 अक्तूबर, 2014

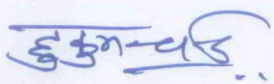
हुकुम चन्द्र
कौस्तव आदित्य

PREFACE

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

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

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



New Delhi
October 28, 2014

Kaustav Aditya
(Course Co-Coordinator)

Hukum Chandra
(Course Coordinator)

CONTENTS

AN OVERVIEW OF VARIOUS ACTIVITIES IN SAMPLE SURVEY DIVISION

U.C. Sud

Indian Agricultural Statistics Research Institute, New Delhi-110012

1.1 MANDATE

The Division of Sample Survey is mainly involved in the development of sample survey techniques for estimation of various parameters of interest relating to crops, livestock, fishery and allied fields.

1.2 MAIN ACTIVITIES

- Research
- Teaching
- Research Guidance
- Training
- Advisory & Consultancy

1.3 THRUST AREAS

- Application of Remote Sensing and Geographical Information System
- Small Area Estimation
- Area & Production Estimation
- Cost of Production Studies
- Assessment and Evaluation Studies
- Harvest and Post Harvest Losses
- Analysis of Complex Surveys
- Development of Data Bases

1.4 IMPORTANT FEATURES

1.4.1 APPLIED RESEARCH – USER ORIENTED

- **Some Important Methodologies Developed**

I. Crop production estimation through crop cutting approach which formed a sound objective method of estimating crop production in the country.

II. Estimation of livestock numbers, products and attendant practices.

III. Extent of cultivation and production of fruits and vegetables.

IV. Estimation of plantation crops like arecanut, coconut, cashewnut.

V. Estimation of fish catch both from marine and inland resources.

VI. Estimation of cost of production of crops as well as livestock products.

VII. Evaluation studies such as assessment of development programmes like IADP, HYVP, dairy improvement programmes, etAssessment of harvest and post harvest losses

1.4.2 BASIC RESEARCH

i) Successive Sampling

ii) Systematic Sampling

iii) Cluster Sampling

iv) Varying Probability Sampling

v) Controlled Selection

vi) Non – Sampling Errors

vii) Various Methods of Estimation- ratio, regression etc.

viii) Analysis of Survey Data

1.4.3 RESEARCH PROJECTS

- **COMPLETED DURING 1996-2000**

1 Pilot sample survey for estimating the energy utilisation for different levels of adoption of modern technology in agriculture.

2 A sampling study on utilization of crossbred working animal vis- a-vis non-descripts.

3 Pilot sample survey for estimation of yield of pepper and study of cultivation practices using successive sampling.

4 Statistical modeling for projection of bovine populations and prediction of milk availability.

5 Survey methodology to study economics of keeping goats.

6 Sampling methodology for estimation of fish catch from a lake.

7 Pilot sample survey for evolving a sampling methodology for estimation of area and yield of cultivated fodder crops other than berseem and jowar crop, cost of production and cultivation practices thereof Ghaziabad Distt. (U.P)

8 Sample survey to evolve suitable sampling methodology to study impact of command area irrigation project on agricultural production.

9 A methodological investigation in estimating seasonal fluctuations of post-harvest food-grains losses (wheat).

10 Pilot sample survey for developing a sampling methodology for estimation of post- production losses of milk in rural areas.

11 Estimation of crop yield for small areas.

12 Pilot sample survey to study the economics of Angora rabbits.

13 Pilot sample survey for estimating the area under waste-land.

14 Sample survey for estimation of cashewnut and cashew apple yield and study of its cultivation practices.

15 Development of estimation procedures for Agricultural by-products.

16 Chi-square tests in survey data.

17 Pilot studies for estimation of birth and death rates in ovines.

18 Sample survey for study of constraints in transfer of new agricultural technology under field conditions.

19 Studies on feed intake by bovines through stall feeding and grazing

20 A study of variability of different components of cost of production of fruits at different stages of sampling and estimation of sample sizes at given levels of precision.

21 Estimation of Regression co-efficients from Sample Survey Data.

22 An analysis of yield gap for buffaloes milk

23 Small Area Estimation of Milk Production

24 Development of data base relating to basic and current agricultural and allied statistics over time and space

- **PROJECTS completed in 2001 onwards**

1. A study for estimation of area and production of important vegetable crops on the basis of partial harvest. Use of Remote Sensing Technology in Crop Yield Estimation Surveys
2. A study of variance estimation in complex surveys
3. Estimation of flow and change in dynamic populations
4. To study the effect of various input components on the yield of important vegetable crops.
5. Use of Remote Sensing technology in crop yield estimation surveys .
6. Sample survey to evolve methodology for estimation of fish catch from rivers or streams specially of the hilly areas.
7. Pilot sample survey to develop a sampling methodology for estimation of poultry meat production.
8. Pilot sample survey for estimating the area and yield rates of ginger and potato in hilly areas.
9. Sampling procedure for selection of representative sample of fertilizer from ship (Funded through A.P. Cess Fund)
10. **State of the Indian Farmer: a millennium study. (**Min. of Agriculture, Govt. of India.)
11. A pilot study on cost of production of Coconut in Kerala **(**Funded from **Coconut Development Board, Kochi, Kerala)**

12. Study relating to formulating long term mechanisation strategy for each agro-climetic zone/ state. (funded from **DOAC, Ministry of Agriculture, Govt. of India).**
13. Study of Land Use Statistics through integrated modelling using Geographic Information System (Funded through **A.P. Cess Fund**)
14. Development of GIS based techniques for identification of potential agro-forestry area
15. Estimation of wool production - emerging data needs and a methodological reappraisal – Funded through A.P. Cess Fund (Collaborative with Central Sheep & Wool Research Institute (ICAR), Avikanager (Raj)
16. Assessment of Harvest and Post Harvest Losses **(Mission Mode project under NATP Funding)**
17. Crop Yield Estimation at Small Area Level Using Farmers's Estimates (**Ministry of Statistics & Programme Implementation.)**
18. Pilot sample survey to develop sampling methodology for estimation of area, production and productivity of important flowers on the basis of market arrivals. (**Min. of Stats. & Programme Implementation**, **CSO, Sadar Patel Bhawan, R.K.Puram, New Delhi).**
19. Developing Remote Sensing Based Methodology for Collecting Agricultural Statistics in Meghalaya.
20. A pilot study to develop an alternate methodology for estimation of area and production of horticultural crops - **Funded by CSO, MOS & PI, New Delhi.**
21. To assess the survey capabilities of private sector **- Funded by CSO, MOS & PI, New Delhi.**
22. Pilot study on small area crop estimation approach for crop yield estimates at the Gram Panchayat level - Funded by DES, DOAC, MOA, GOI, New Delhi.
23. Pilot study to develop sampling methodology for estimation of production of mushroom.
24. Developing Remote Sensing Based Methodology for Collecting Agricultural Statistics in North-East hilly region.
25. Study to investigate the causes of variation between official and trade estimates of cotton production- Funded by DES, MOA, GOI
26. Estimation of extent of farming practices, resources and activities with energy use.
27. Study on status and projection estimates of agricultural implements and machinery.
28. Evaluation of Rationalization of Minor Irrigation Statistics Scheme
29. Small area estimation for zero-inflated data
30. Consultancy Project on "Determination of optimum sample size for crop yield estimation at the gram panchayat level"

31. Sampling Methodology for Estimation of Meat Production in Meghalaya- Funded by Department of Animal Husbandry, Dairying and Fisheries, Ministry of Agriculture, Govt of India.
32. Consultancy Project on "Evaluation of Agricultural Census Scheme"
33. Consultancy project entitled "Study to develop an alternative methodology for estimation of Cotton production"- Funded by Directorate of Economics and Statistics (DES), Ministry of Agriculture, Govt of India.
34. District-level Poverty Incidence Estimation from NSSO data using Small Area Estimation Technique- Funded by CSO, MOS&PI, GOI.
35. National Initiative on Climate Resilient Agriculture (NICRA)-Agroforestry Component
36. Study of Sample Sizes for Estimation of Area and Production of Food Grain Crops
37. Study to develop methodology for crop acreage estimation under cloud cover in the satellite imageries
38. A study on calibration estimators of finite population total for two stage sampling design
39. On Small Area Inference using Survey Weights
40. Spatial Nonstationarity in Small Area Estimation under Area Level Model
41. Impact assessment of agroforestry model in Vaishali district of Bihar State
42. Farm power machinery use protocol and management for sustainable crop

1.4.4 ON-GOING RESEARCH PROJECTS

1. Small area estimation under skewed data.
2. Pilot Study for Estimation of Seed, Feed and Wastage Ratios of Major Food grains
3. Development of Innovative Approaches for Small Area Estimation of Crop Yield, Socio- economic and Food Insecurity Parameters
4. Assessment of quantitative harvest and post harvest losses of major crops/commodities in India
5. Calibration Estimators under Two Stage Sampling Design when Study Variable is Inversely Related to Auxiliary Variable
6. Study to test the developed alternative methodology for estimation of area and production of horticultural crops

1.4.5 CONSULTANCY PROJECTS FOR PREPARATION OF MANUALS - FUNDED BY CSO, MOS & PI, NEW DELHI

- Area and Crop Production Statistics
- Animal Husbandry Statistics
- Agricultural Prices and Marketing

- Cost of Cultivation Surveys
- Horticulture and Spices Statistics
- Fishery Statistics

1.4.6 COLLABORATIVE RESEARCH PROJECTS

- Survey of Agricultural Accidents for the year 2004-05 in a large sample of villages selected on the basis of statistical consideration, with AICRP on ESA (Ergonomics & Safety in Agriculture) (Old Name- AICRP on HESA (Human Engineering & Safety in Agriculture)
- Assessment of post harvest losses of crops/commodities with AICRP on PHT.

1.4.6 SOME OTHER STUDIES

1. Equines in the household economy of the poor (Study was undertaken in 1998 in collaboration with NRC Equines, Hissar and NCAP, New Delhi)
2. A pilot study of agro-forestry/social forestry in Chhachhroli block of Yamunanagar District in Haryana (Study was undertaken in 1998 in collaboration with forest department of Haryana State)
3. A study for estimation of crop yield at block level using crop-cut and farmer's estimate (being conducted in Haryana State during rabi season 1998-99)

1.4.7 LINKAGES

- Ministry of Agriculture
  - Directorate of Economics & Statistics
  - Expert Committee on Crop Forecast
  - Working Group on Improvement of Agricultural Statistics
  - Steering Committee on Agricultural Census and Input Survey
  - Causes of Variation between Official & Trade estimates of Cotton Production
  - Meeting with Major Trade & Industry Associations dealing with Oilseeds Processing and Marketing / Trading
  - Technical Committee for Reviewing the Sampling Design and Estimation Procedure followed in respect of Crop Estimation Surveys on Fruits, Vegetables and Minor Crops in different States
- Department of Animal Husbandry and Dairying
  - Technical Committee of Direction for improvement of Animal Husbandry and Dairying Statistics

- Sub-Committee for Identification of Methodology for Estimation and Requirement of Fodder, Fodder Seeds

- Ministry of Rural Development
  - Directorate of Marketing and Inspection
  - Technical Committee for Estimation of Marketable Surplus and Post-harvest Losses of Foodgrains
- Central Statistical Organization /National Sample Survey Organization
  - Expert Committee on Small Area Statistics
  - Expert Group on Examination of the Extent of Underestimation of GDP in Agriculture
- Other Research Institutes
  - National Remote Sensing Agency
  - Space Application Centre
  - Indian Institute of Remote Sensing
  - All-India Soil & Land Use Survey

Regarding Remote Sensing and GIS Applications

Collaboration in the project on Integrated Approach for Land use Statistics

1.4.8 FUTURE PROGRAMMES (EFC BASED)

a) **Statistical applications of GIS and Remote Sensing Techniques in agricultural systems.**

Application of remotely sensed data for improvement of area and production statistics and development of yield models.

b) **Development of techniques for planning and analysis of survey data related to agricultural systems.**

i) Development of model based small area techniques for estimation of parameters from agricultural and allied surveys.

ii) Studies on enquiry based farmers estimate with a view to improve the present method of crop estimation surveys.

iii) Re-appraisal of the sampling methodologies for estimation of crops livestock products, fisheries, horticulture crops and cotton.

iv) Studies of estimation and marketing of crops, livestocks and fisheries and other data gaps in agricultural sector needed for speeding up the planning process.

v) Evaluation studies for measuring the impact of development of programmes such as Krishi Vigyan Kendra agro-forestry, various rural developments programmes etc.

vi) Studies relating to theoretical aspects on analysis from survey data and applications to agricultural surveys such as regression analysis of survey data, categorical data analysis, variance estimation of non-linear statistics from complex survey, multivariate analysis.

vii) Studies on assessment and control of various non-sampling errors such as coverage errors, response and non-response errors as well as errors due to sensitive questions will be carried out.

viii) Studies on estimation of high value minor crops.

1.4.9 AGRICULTURAL RESEARCH DATA BOOKS

Agricultural research is a vital input for planned growth and sustainable development of agriculture in the country. Indian Council of Agricultural Research, being an apex scientific organization at national level, plays a crucial role in promoting and accelerating the use of science and technology programmes relating to agricultural research and education. It also provides assistance and support in demonstrating the use of new technologies in agriculture.

Information pertaining to agricultural research, education and related aspects available from different sources is scattered over various types of published and unpublished records. The Agricultural Research Data Book 2012, which is the fifteenth in the series, is an attempt to put together main components/indicators of such information. The Data Book comprising 172 Tables is organized, for the purpose of convenience of the users, into ten sections namely, Natural Resources; Agricultural Inputs; Animal Husbandry, Dairying and Fisheries; Horticulture; Production and Productivity; Agricultural Engineering & Produce Management; Export & Import; India's Position in World Agriculture; Investment in Agricultural Research & Education; and Human Resources under National Agricultural Research System (NARS). This edition contains the latest information / data as available in the country at the end of May, 2012. In ARDB 2012, some value editions like predicting the future year production of food grain crops etc., based on previous years data using statistical models, pictorial/graphical representations of data have been done. For depicting state-wise data, thematic maps have been prepared using Geographical Information System (GIS). Efforts have been made to incorporate the comments and suggestions received from various users. The first Agricultural Research Data Book was brought out in the year 1996. Subsequently, this was updated and brought out in the years 1997, 1998, 1999, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009 and 2011 (fourteenth edition)

1.4.10 TRAINING PROGRAMMES ORGANISED

1.4.10.1 INTERNATIONAL

Training Course Title	Year	Target group
International Training Programme on "Techniques of Estimation of Output of Food Crops"	1989 1991 1992 1993 1994 1995 1996	Participants from Afro-Asian Countries
Commonwealth Training Programme on "Agricultural Sample Surveys, crop Yield Modelling and Computer Programming"	1992	Participants from Commonwealth Countries
Remote Sensing and GIS Applications	1999	Officials from Nigerian Government
Training Programme/Study Tour (Funded by FAO)	2001	Participants from Eritrea, Asmara
Workshop on "Economic Accounts for Agriculture" jointly conducted by the Food and Agriculture Organization (FAO) of the United Nations and the United Nations Statistical Institute for Asia and the Pacific (SIAP)	2001	Participants from Bangladesh, Bhutan, Cambodia, Islamic Republic of Iran, Laos, Nepal, Pakistan, Sri Lanka and India
An International study tour/Training Programme on "Development of Agricultural Statistics System"- sponsored by Food & Agriculture Organisation	2006	Participants from Timor Leste
International Training Programme for Malasian Delegates	2006	Malasian Delegates
Study Visit on "Indian Agricultural Statistics System" for Afghanistan nationals sponsored by the Food and Agriculture Organization (FAO)	2008	Afghanistan nationals
An International training programme on "Agricultural Statistics System in India", funded by Central Statistical Organization (CSO), Ministry of Statistics and Programme Implementation (MOS&PI), Govt. of India (GOI).	2010	Officials of SAARC countries
An International training programme on "Applications of Remote Sensing and GIS in Agricultural Surveys"	2011	Officers of Census and Statistics Department Colombo, Srilanka

Assisted in organizing one week module of an International training programme on "Sampling Techniques and Survey Methods" organized by National Academy of Statistical Administration (NASA), MOS & PI, GOI at IASRI, New Delhi	2011	Officers of Srilanka Government
An International training programme on "Application of Remote Sensing and GIS in Agricultural Surveys", sponsored by Afro Asian Rural Development Organization (AARDO)	2012	Afro-Asian Rural Development Organizations (AARDO) member countries
An International Training programme on "Techniques of Estimation and Forecasting of Crop Production in India", funded by the Food and Agriculture Organization (FAO)	2012	Officials of the Ethiopian Central Statistical Agency
An International training programme on 'Applications of Remote Sensing and GIS in Agricultural Surveys' for African -Asian Rural Development Organization (AARDO)	2013	Afro-Asian Rural Development Organizations (AARDO) member countries
A study visit on "Agricultural statistics system and food security policy analysis in India" under the Project "Strengthening Food and Agriculture Information System" funded by UNDP.	2013	Officials of DPR Korea

1.4.10.1 NATIONAL

S. No.	Training Course Title	Year	Target group
1.	Models in Survey Sampling	1993	Associate/Asstt. Professor from SAU's/Scientists from ICAR Institutes
2.	Modern Sampling Techniques	1997	Senior/Middle level ISS Officers from CSO/NSSO
3.	Recent Advances in Survey Sampling in Relation to Agriculture	1998	Associate/Asstt. Professor from SAU's/Scientists from ICAR Institutes
4.	Recent Advances in the Analysis of Survey Data	1998	Senior/Middle level ISS Officers from CSO/NSSO
5.	Methodological Aspects in Sample Surveys	1999	Senior/Middle level ISS Officers from CSO/NSSO
6.	Small Area Estimation-Theory and Applications	1999	Senior/Middle level ISS Officers from CSO/NSSO

7.	Qualitative Aspects in Collection and Analysis of Survey Data	1999	Senior/Middle level ISS Officers from CSO/NSSO
8.	Winter School on "Recent Developments in Survey Sampling in Relation to Agricultural Research"	1999	Associate/ Asstt. Professors from SAU's/ Scientists from ICAR Institutes
9.	Computer Intensive Techniques in Agricultural Surveys	2000	Associate/ Asstt. Professors from SAU's/ Scientists from ICAR Institutes
10.	Recent Advances in the Analysis of Survey Data	2001	Scientists / Faculty Members of National Agricultural Research System
11.	Sample Surveys related to the estimation of area and production of fruits and vegetables	2001	Officials from Department of Horticulture, Govt. of Haryana
12.	Small Area Estimation Techniques in Agriculture	2002	Associate/ Asstt. Professors from SAU's/ Scientists from ICAR Institutes
13.	Summer school on "Application of Remote Sensing and GIS in Agricultural Statistics"	2003	Scientists/Assistant/Associate Professors from ICAR Institutes and State Agricultural Universities
14.	A winter school on Recent advances in survey sampling with special emphasis on computer intensive data analysis techniques	2003	Scientists/Assistant/Associate Professors from ICAR Institutes and State Agricultural Universities
15.	A training programme on "Sampling Techniques, Sample Surveys and Methodological Aspects relating to Cost of Cultivation Studies"	2004	Senior level Officers of Tariff Commission, Ministry of Commerce & Industry, Govt. of India
16.	A Winter School on "Sample Survey Techniques in Agricultural Research"	2005	Assistant / Associate Professors from SAUs and Scientists from ICAR Institutes
17.	A Refresher training programme on "Small Area Estimation Techniques" sponsored by CSO, Ministry of Statistics and Programme Implementation	2005	Indian Statistical Services (ISS) Officers
18.	A CAS sponsored Training Programme on "Recent Advances in the Analysis of Survey Data"	2005	Scientists / Asstt. / Assoc. Prof. of various SAUs under National Agricultural Research System

19.	A training programme on “Small Area Estimation Techniques”	2006	Senior level Officers of Tariff Commission, Ministry of Commerce & Industry, Govt. of India
20.	Summer school on “Sample Survey Techniques in Agricultural Research”	2006	Scientists/Assistant/Associate Professors from ICAR Institutes and State Agricultural Universities
21.	Winter School on “Sample Survey Techniques in Agricultural Research”	2008	Scientists/Assistant/Associate Professors from ICAR Institutes and State Agricultural Universities
22.	A CAS training programme on “Recent advances in sample survey and analysis of survey data”	2009	Scientists / Asstt. / Assoc. Prof. of various SAUs under National Agricultural Research System
23.	A refresher training programme on “Small Area Estimation” sponsored by Central Statistical Organization (CSO), Ministry of Statistics and Programme Implementation (MOS&PI), Govt. of India (GOI).	2010	Indian Statistical Services and other senior officers of States/UTs
24.	A refresher training course on “Research Methodology for Official Statistics” sponsored by CSO, MOS&PI, GOI.	2010	Indian Statistical Service (ISS) officers and statistical personnel
25.	A Refresher Training Programme on "Agricultural Statistical System in India" sponsored by MOSPI, GOI	2010	Statistical Personnel of States/UTs/PSUs of MOSPI, GOI
26.	A refresher training programme on “Small Area Estimation”, funded by CSO, MOS&PI, GOI.	2010	In-service ISS officers and senior officers of State Govts./UTs
27.	A refresher training programme on “Agricultural Statistics”, funded by CSO, MOS&PI, GOI.	2011	Officers of Directorate of Economics & Statistics of different States
28.	CAFT Training Programme on “Recent Advances in Sample Survey and analysis of Survey Data using Statistical Softwares”	2012	Scientists of ICAR Institutes/Asstt-Associate Professors of SAUs
29.	Referesher training programme on “Agricultural Statistics”	2012	Officers from the Department of Agriculture, Government of Andhra Pradesh State

30.	Refresher Course on "Small Area Estimation" was organized	2012	In- Service Officers of Indian Statistical Service at IASRI, New Delhi
31	Refresher training programme on "Agricultural Statistics"	2012	Officers of the Department of Agriculture, Government of Andhra Pradesh
32	Refresher training programme on "Small Area Estimation"	2012	Officers of Ministry of Agriculture, Govt. of India
33	CAFT Training Programme on Recent Advances in Sample Survey and Analysis of Survey Data using Statistical Softwares	2013	Scientists of ICAR Institutes/Asstt-Associate Professors of SAUs
34	Refresher training programme on "*Integrated Sample Survey Methodology*	Jan., 2014	Department of Animal husbandry, Dairying & Fisheries, Ministry of Agriculture, Govt. of India.
35	Refresher training programme on "*Integrated Sample Survey Methodology*	Mar., 2014	Department of Animal husbandry, Dairying & Fisheries, Ministry of Agriculture, Govt. of India.
36	Referesher training programme on "Agricultural Statistics"	2014	Officers from the Department of Animal and Husbandry, Government of Chhattisgarh

Scientists are also involved in teaching and research guidance for M.Sc/Ph.D.

BASIC STATISTICAL METHODS

A K Gupta

Indian Agricultural Statistics Research Institute, New Delhi-110012

2.1 Introduction

Statistics is a very broad subject, with applications in a vast number of different fields. Generally, one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from the data.

Statistics has been defined differently by the authors from time to time. Some authors define it as Statistical Data i.e. numerical statements of the facts while others define it as Statistical Methods i.e. principles and techniques used in collecting and analyzing the data.

But the Statistics defined as Statistical Data is inadequate because Statistics is not merely confined to the collection of data only but other aspects like presentation, analysis and interpretation etc. are also the parts of Statistics.

The best definition of Statistics was given by Croxton and Cowden according to whom the Statistics is the science which deals with the collection, analysis and interpretation of numerical data/facts.

The theoretical developments in modern statistics came during mid-seventeenth century with the introduction of 'Theory of Probability' by Pascal (1623-1662), P. Fermat (1601-1665) with the Development of Properties of Coefficient of Binomial Expansion, James Bernoulli (1654-1705), De-Movire (1667-1754), Laplace (1749-1827) with the Development of Theory of Probability, Gauss (1777-1855) with the Development of Principle of Least Square and Normal Laws of Error.

But the Modern Developments in Statistics are due to Galton (1822-1921) with the Development of regression Theory, Karl Pearson (1857-1936) with the Development of Correlation analysis and Chi-square Test and Test of Significance, W S Gosset (1908) with the discovery of t-distribution applicable in small sample test. Sir Ronald A. Fisher (1890-1962) known as Father of Statistics, placed Statistics in a very sound footing by applying it to various diversified fields such as Genetics, Biometrics, Education, Agriculture etc. and by discovery of Point Estimation (Efficiency sufficiency principle of Maximum Likelihood), Exact Sampling Distribution, ANOVA, Design of Experiment etc. His contribution made the Statistics a very responsible position among various sciences.

2.2 Classification

The process of arranging data into homogenous group or classes according to some common characteristics of the data is called Classification.

(1) Qualitative Base:

When the data are classified according to some quality or attributes such as sex, religion, literacy, intelligence etc.

(2) Quantitative Base:

When the data are classified by quantitative characteristics like heights, weights, ages, income etc.

(3) Geographical Base:

When the data are classified by geographical regions or location, like states, provinces, cities, countries etc.

(4) Chronological or Temporal Base:

When the data are classified or arranged by their time of occurrence, such as years, months, weeks, days etc.

Tabulation of Data

The process of placing classified data into tabular form is known as Tabulation. A table is a symmetric arrangement of statistical data in rows and columns.

Diagrams and Graphs of Statistical Data

One of the most effective way of representation of statistical data may is through diagrams and graphs. The commonly used diagrams and graphs are as below:

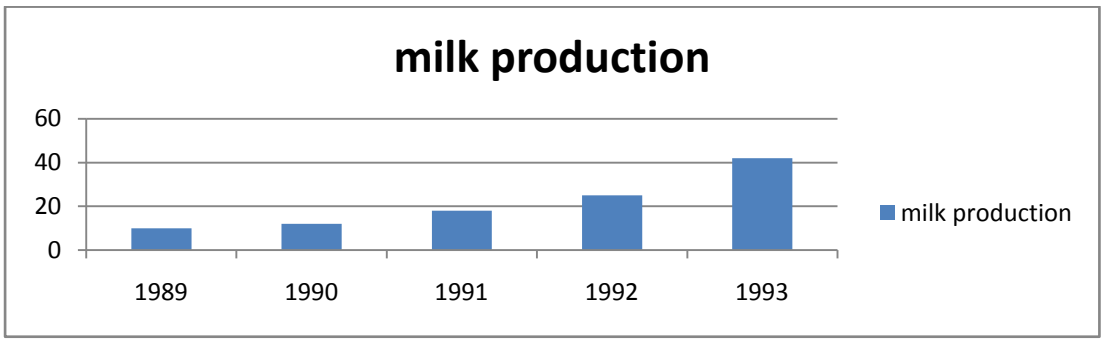
Types of Diagrams/Charts:

1. Simple Bar Chart
2. Multiple Bar Chart
3. Component Bar Chart or Sub-Divided Bar Chart
4. Simple Component Bar Chart
5. Percentage Component Bar Chart
6. Sub-Divided Rectangular Bar Chart
7. Pie Chart

Simple Bar Chart

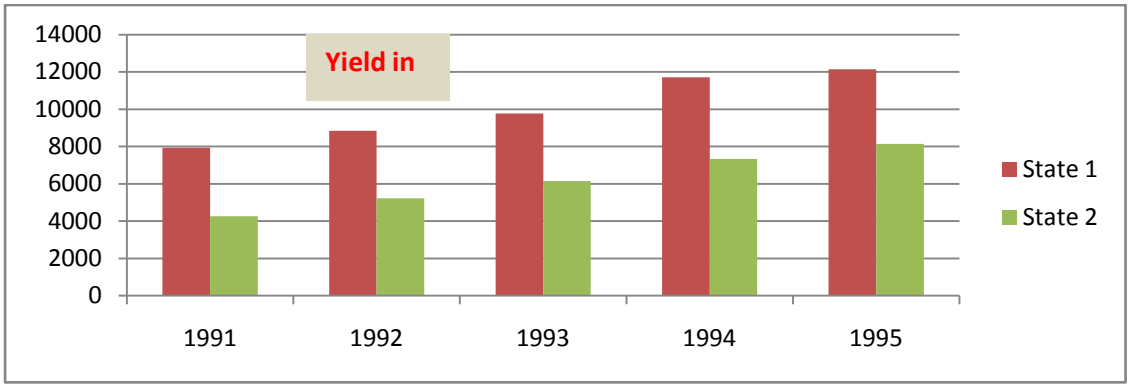
In simple bar chart, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars. Following steps are undertaken in drawing a simple bar diagram:

Simple Bar Chart showing the Milk Production



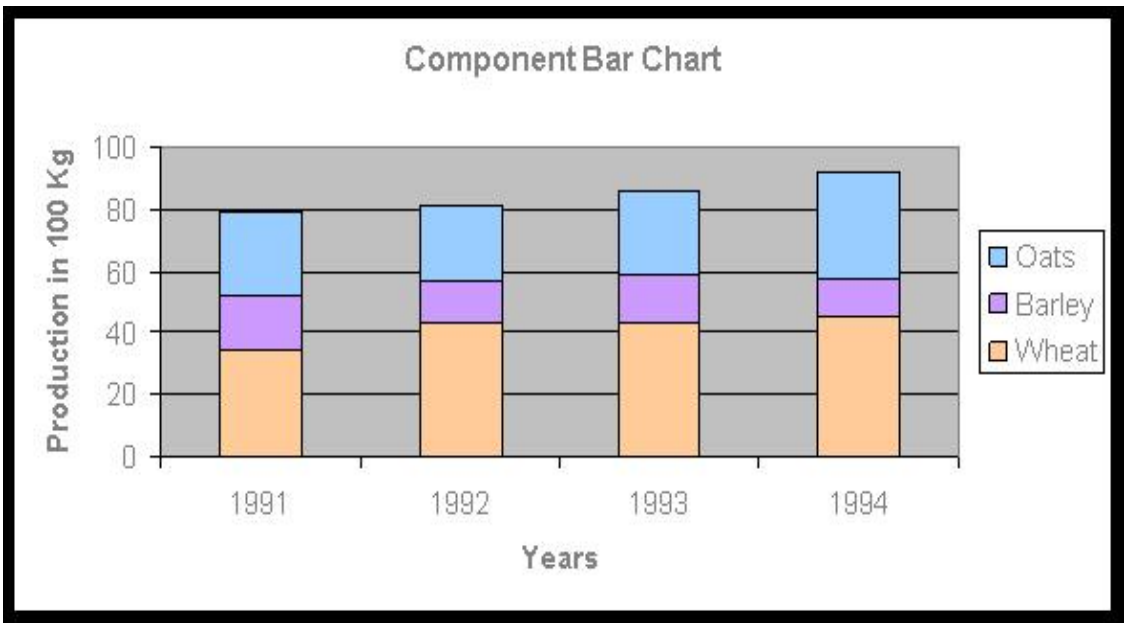
Multiple Bar Charts

By multiple bars diagram, two or more sets of inter-related data are represented.



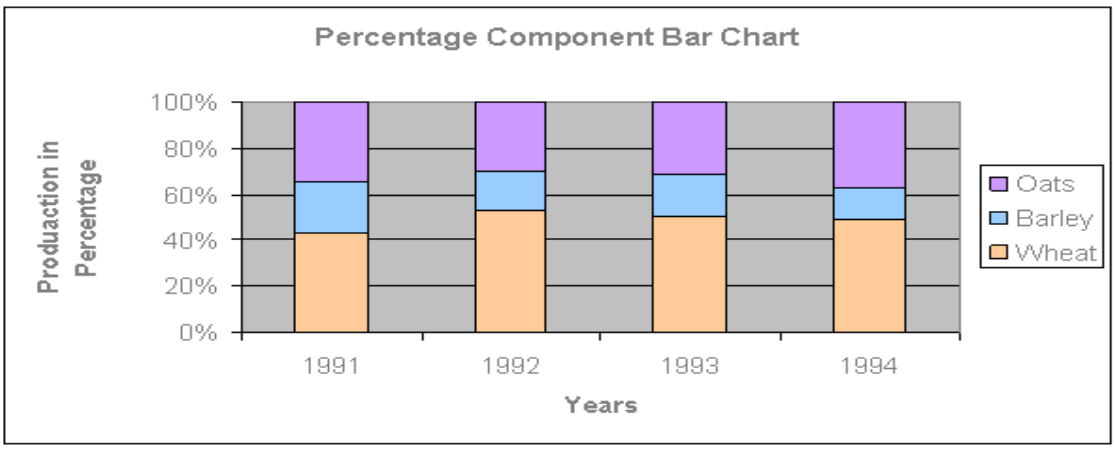
Component Bar Chart

Component bar chart is used to represent data in which the total magnitude is divided into different components.



Percentage Component Bar Chart

Component Bar Charts/Sub-divided Bar Charts may be drawn on percentage basis. To draw sub-divided bar chart on percentage basis, we express each component as the percentage of its respective total.



Pie Chart

Pie chart is used to compare the relation between the whole and its components. To construct a pie chart (sector diagram), we draw a circle with radius (square root of the total). The total angle of the circle is 360^0. The angles of each component are calculated by the formula.

$$Angle\ of\ Sector = \frac{Component\ Part}{Total} \times 360^0$$

These angles are made in the circle by mean of a protractor to show different components. The arrangement of the sectors is usually anti-clock wise.

Example:

The following table gives the details of monthly budget of a family. Represent these figures by a suitable diagram.

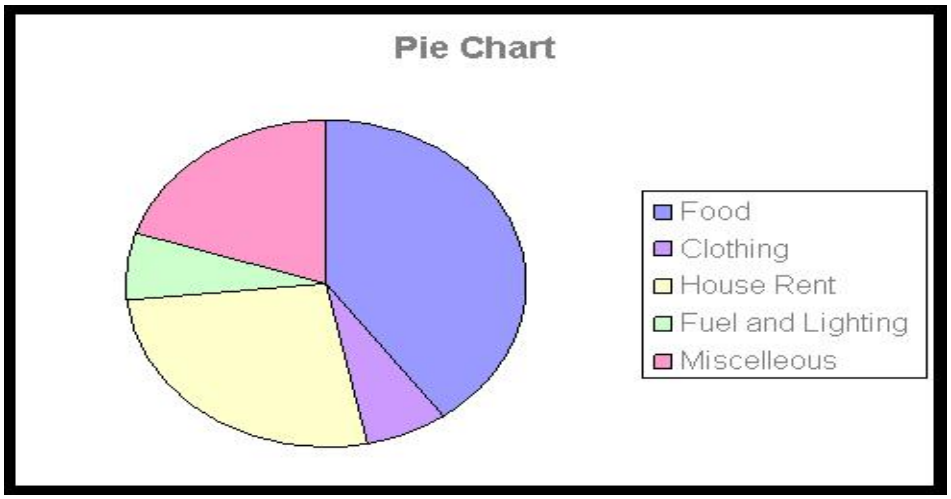
Item of Expenditure	Family Budget
Food	Rs. 600
Clothing	Rs.100
House Rent	Rs.400
Fuel and Lighting	Rs.100
Miscellaneous	Rs.300
Total	Rs.1500

Solution:

The necessary computations are given below:

$$Angle\ of\ Sector = \frac{Component\ Part}{Total} \times 360^0$$

Items	Family Budget		
	Expenditure (Rs.)	Angle of Sectors	Cumulative Angle
Food	600	144^0	144^0
Clothing	100	24^0	168^0
House Rent	400	96^0	264^0
Fuel and Lighting	100	24^0	288^0
Miscellaneous	300	72^0	360^0
Total	1500	360^0	

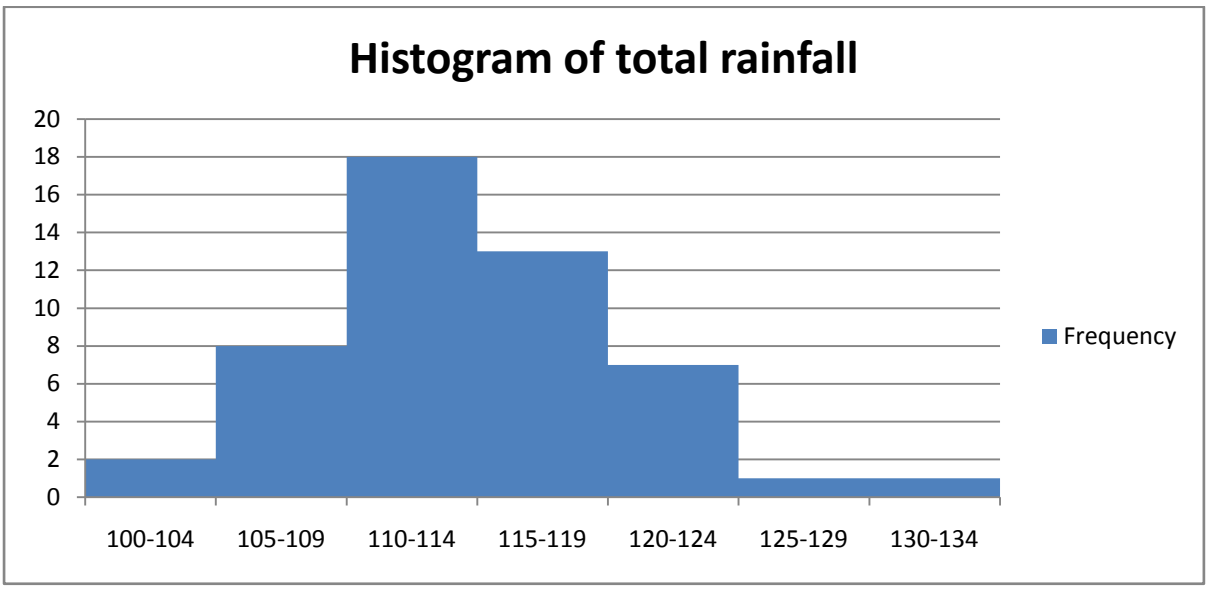


Most common graphs:

1. Histogram,
2. Frequency polygon,
3. Cumulative frequency graph or Ogive.

1. Histogram

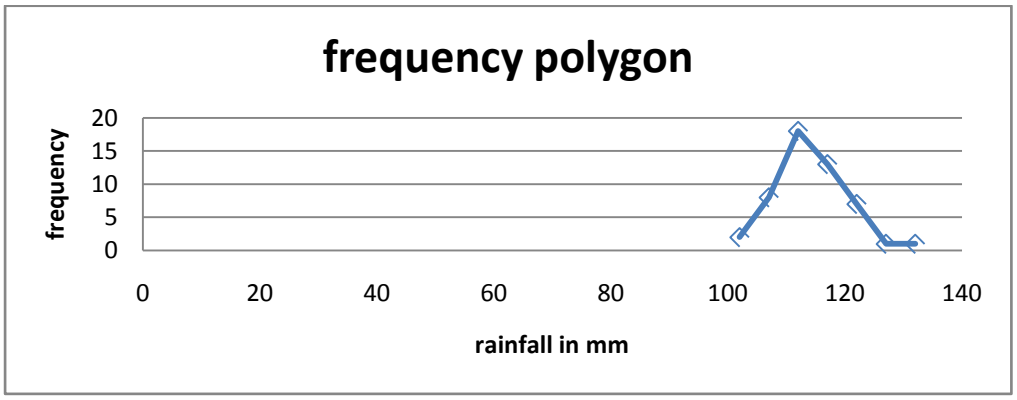
The histogram is a graph that uses contiguous vertical bars to display the frequency of the data (unless the frequency equals 0) contained in each class. The heights of the bars equal the frequency (after certain scale has been chosen) and the bases of the bars lie on the corresponding class.



2. Frequency Polygon

A frequency polygon is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. In the Cartesian

system OXY the midpoints are the first coordinates of the vertices of the polygon and the frequencies are the second coordinates.



3. OGIVE

An OGIVE is a graph that represents the cumulative frequencies for the classes in a frequency distribution. It shows how many of values of the data are below certain boundary.

2.3 MEASURES OF CENTRAL TENDENCY

Measures of central tendency are the statistical constants which enable us to comprehend the whole of the distribution/data in to a single value or it is the value of the variable under study which is representative of the entire distribution.

Ideal measures of central tendency:

An average possesses all or most of the following qualities (characteristics) is considered a good average:

(1) It should be rigidly defined.
(2) It should be easy to calculate and easy to understand.
(3) It should be based on all the observations.
(4) It should be suitable for further mathematical/algebraic treatment.
(5) It should not be affected by extreme values.
(6) It should be affected at least as possible by the fluctuations of the sample values.

Types of measures of central tendency:

1. Arithmetic Mean
2. Median

3. Mode
4. Geometric Mean
5. Harmonic Mean

Arithmetic mean

Arithmetic mean of a variable or set of given observations is quotient of sum of the given observations and the number of the observations.

The arithmetic mean can be computed for both ungroup data (raw data: a data without any statistical treatment) and grouped data (a data arranged in tabular form containing different groups). If x is a variable having n observations, arithmetic mean abbreviated as A M and denoted by $\bar{X}$ can be computed by using any of the following formula;

For ungrouped Data:

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})$ and $\bar{X} = A + \frac{1}{n}\sum_{i=1}^{n}(x_i - A)$

For grouped Data:

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} f_i\, x_i$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} f_i\,(x_i - \bar{X})$ and $\bar{X} = A + \frac{1}{n}\sum_{i=1}^{n} f_i\,(x_i - A)$

Short Cut Method:

$\bar{X} = A + \frac{\sum_{i=1}^{n} d_i}{n}$ (For ungrouped data), $\bar{X} = A + \frac{\sum_{i=1}^{n} f_i\, d_i}{\sum_{i=1}^{n} f_i} \times h$ (For grouped data)

Where

x_i – different observations of the variable under study

f_i – frequencies of different class intervals/groups

A – Assumed mean

d_i – deviations of the observations from assumed mean A

n – number of observations and

h – class interval in case of grouped data

Median

Median of a given distribution is the value of the variable which divides the distribution in to two equal parts. It is the value such that number of observations preceding as well as succeeding from the median is equal or which exceeds and exceeded by the same number of observations. Median is thus a **Positional Average** only.

First of all, the given observations of the distribution are arranged in ascending/descending order in case of ungrouped data. Median is calculated as follows;

(i) If number of observations is odd

Median = Value of $\left(\frac{n+1}{2}\right)^{th}$ item

(ii) If the number of observations is even

Median = Average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ items

Median for grouped data:

In case of grouped data (discrete frequency distribution), a separate column of cumulative frequencies is made. Find the number n/2. See the cumulative frequency in which this number n/2 falls. The corresponding x_i value will be the median of the grouped distribution.

In case of the grouped data (continuous frequency distribution), a separate column of cumulative frequencies is also made. Find the number n/2. See the cumulative frequency in which this number n/2 falls. The corresponding class interval is called the Median Class. After locating the Median Class, following formula is used for calculation of median.

$$Median = l + \frac{h}{f}(\frac{n}{2} - c)$$

Where,

$l =$ Lower class limit of the Median Class

$f =$ Frequency of the Median Class

$n = \Sigma f =$ Sum of the frequencies of various class intervals

$c =$ Cumulative frequency of the class preceding the Median Class

$h =$ Class interval size of the Median Class

Mode

Mode is the value which occurs most frequently in the given set of observations i.e. it is the value of the variable which is predominant in the given set of observations. If the data having only one mode the distribution is said to be uni-model and is said to be bi-model, if data have two modes.

For ungrouped data, mode is calculated by inspecting the given data. The value which occurs maximum number of times in the distribution is called the Mode of the given distribution.

For grouped data, locate the Modal Class/Group. The class/group which has the maximum frequency is called the Modal Class/Group. After locating the Modal Class/Group, the following formula is applied for calculation of Mode of the given frequency distribution.

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

Where,

l is the lower class limit of the model group,
f_m is the frequency of the model group
f_1 is the frequency of the class interval preceding the model group
f_2 is the frequency of the class interval preceding the model group
h is the class interval of the modal group

Geometric Mean

Geometric mean of a set of n observations is the n^{th} root of the multiplication of all the n observations. Hence the geometric mean denoted by G; of n observations x_i, i = 1, 2, .., n is given by the formula

$$G = (x_1 . x_2 . \ldots\ldots\ldots\ldots . x_n)^{1/n}$$

$$G = \text{Antilog } [\frac{1}{n} \sum_{i=1}^{n} \log x_i]$$

In case of grouped frequency distribution, geometric mean is given by the formula

$$G = \text{Antilog } [\frac{1}{n} \sum_{i=1}^{n} f_i \log x_i] \quad \text{where} \quad n = \sum_{i=1}^{n} f_i$$

The Geometric Mean of the values 10, 5, 15, 8, 12 is given by

$$G = \sqrt[5]{10 \times 5 \times 15 \times 8 \times 12}$$

$$= \sqrt[5]{72000} = (72000)^{\frac{1}{5}} = 9.36$$

By log method

x	$\log xi$
10	1.0000
5	0.6990
15	1.1761
8	0.9031
12	1.0792
Total	$\Sigma \log xi = 4.8573$

$$G = Anti\log\left(\frac{\Sigma \log xi}{n}\right)$$

$$G = Anti\log\left(\frac{4.8573}{5}\right)$$

$$G = Anti\log(0.9715) \quad = \mathbf{9.36}$$

Harmonic mean

Harmonic mean is defined as the quotient of **"number of the given values"** and **"sum of the reciprocals of the given values"**.

Harmonic mean in mathematical terms is defined as follows:

For ungrouped data: $$HM = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

For grouped data: $$HM = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)}$$

The Harmonic Mean of the numbers: 13.5, 14.5, 14.8, 15.2 and 16.1 is given by

x	$\frac{1}{x}$
13.2	0.0758
14.2	0.0704
14.8	0.0676
15.2	0.0658
16.1	0.0621
Total	$\Sigma\left(\frac{1}{x}\right) = 0.3417$

$$HM = \frac{5}{0.3417} = 14.63$$

2.4 MEASURES OF DISPERSION

Measures of central tendency give us single figure which represent the entire distribution or set of observations or around which the observations of the set of data concentrated. But they are inadequate to give us the complete idea of the distribution because they do not tell us the extent to which the observations of the distribution vary from the central value. There may be more than one distributions having the same central value but there may be the vide variation in the different observations of the distribution. The observation may be close to the central value or they may be spread away from the central value. If the observations are close to the central value, we say that dispersion or variation is small. If the observations are spread away from the central value, we say dispersion is more.

Suppose we have three groups of students who have obtained the following marks in a test. The arithmetic means of the three groups are also given below;

Group A: 46, 48, 50, 52, 54 $\bar{X}_A = 50$

Group B: 30, 40, 50, 60, 70 $\bar{X}_B = 50$

Group C: 10, 30, 50, 70, 90 $\bar{X}_C = 50$

All the three sets of observations have the same arithmetic mean i.e. 50. But we see that the variation/dispersion of the other values to the central value is less in Group A in comparison of group B and Group C or we may also say that the variation/dispersion in the observations are more in Group C in comparison of the other two groups.

Thus in order to give a proper idea about the overall nature of the given values of a distribution or set of data, it is necessary to state how are the values of the distribution scattered/dispersed from the measures of central tendency? Therefore, the measures of dispersion may be defined as a statistics signifying the extent of the variations of items of the given set of observations around the measure of central tendency.

Measures of Dispersion

For the study of dispersion, there are some measures which show whether the dispersion is small or large. There are two types of measure of dispersion;

(a) Absolute Measure of Dispersion

(b) Relative Measure of Dispersion

Absolute Measures of Dispersion:

These measures give us an idea about the amount of dispersion in a set of observations.

1. Range
2. Quartile Deviation or Semi Inter Quartile Range
3. Mean Deviation
4. Variance and Standard deviation

Relative Measure of Dispersion:

These measures are calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measured. The relative measures of dispersion are:

1. Coefficient of Range
2. Coefficient of Quartile Deviation
3. Coefficient of Mean Deviation
4. Coefficient of Variation

Range

Range is defined as the difference between the maximum and the minimum values of the given observations. If x_m denotes the maximum value and x_0 denotes the minimum value, range is defined as:

$$Range = x_m - x_0$$

$$Coefficient\ of\ Range = \frac{x_m - x_0}{x_m + x_0}$$

Quartile Deviation/Semi Inter Quartile Range

It is based on the lower quartile Q_1 and the upper quartile Q_3.

$$Quartile\ Deviation\ = \frac{Q_3 - Q_1}{2}$$

$$Coefficient\ of\ Quartile\ Deviation = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation

The mean deviation is defined as the arithmetic mean of the absolute deviations of all the values taken from some suitable average which may be the arithmetic mean, the median or the mode.

The mean deviation of a set of sample data in which the suitable average (AM) is $\bar{X}$, is given by the relation:

$$Mean\ Deviation = \frac{\Sigma|X - \bar{X}|}{n}$$

For frequency distribution

$$Mean\ Deviation = \frac{\Sigma f|X - \bar{X}|}{\Sigma f}$$

Mean deviation is a better measure of dispersion than Range and Quartile Deviation.

Coefficient of Mean Deviation

Coefficient of Mean Deviation is given by

$$\text{Coefficeient of Mean Deviation} = \frac{\text{Mean deviation}}{\text{AM}}$$

Variance and Standard Deviation

The standard deviation is defined as the positive square root of the mean of the squares of all the deviations taken from arithmetic mean of the data. The standard deviation is denoted by σ and is given by

Population Standard Deviation is given as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Sample Standard Deviation is given as

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

The unit of standard deviation is same as the units of the original observations.

The Variance is the square of the standard deviation. The standard deviation plays a dominating role for the study of variation in the data. It is widely used for the analysis of measure of dispersion.

As far as the important statistical tools are concerned, the first important tool is the arithmetic mean $\bar{X}$ and the second important tool is the standard deviation. Both are based on all the observations and are subject to mathematical treatment.

However some alternative methods are also available to compute standard deviation. The alternative methods simplify the computation.

Assumed Mean Method

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{(\sum fd)}{\sum f}\right)^2} \qquad (d = X - A)$$

Coefficient of Standard Deviation and Coefficient of Variation

The standard deviation is the absolute measure of dispersion. Its relative measure is called standard coefficient of dispersion or coefficient of standard deviation. It is given as

$$\text{Coefficient of Standard Deviation} = \frac{\sigma}{\bar{X}}$$

The coefficient of variation (*CV*) is given by the formula

$$\text{Coefficient of Variation} = \frac{\sigma}{\bar{X}} \times 100$$

Coefficient of variation is a pure number and the unit of observations cannot be mentioned with its value. It is written in percentage. When its value is 20%, it means that when the mean of the observations is assumed equal to 100, their standard deviation will be 20.

Coefficient of variation is used to compare the degree of dispersion/variation in different sets of data particularly the data which differ in their means or differ in the units of measurement. The wages of workers may be in dollars and the consumption of meat in their families may be in kilograms. The standard deviation of wages in dollars cannot be compared with the standard deviation of the quantity of meat in kilograms. Both the standard deviations need to be converted into coefficient of

variation for comparison. Suppose the value of coefficient of variation of wages is 10% and the value of coefficient of variation of meat is 25%. This means that the wages of workers are consistent in comparison of their consumption of meat. We say that there is greater variation in their consumption of meat. The observations about the quantity of meat are more dispersed than their wages.

BASIC CONCEPTS OF SAMPLING AND SIMPLE RANDOM SAMPLING

K.K. TYAGI

Former Principal Scientist

Indian Agricultural Statistics Research Institute, New Delhi-110012

(e-mail: kkanttyagi@gmail.com)

3.1 BASIC CONCEPTS OF SAMPLING

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.**

In the collection of data there are basically two different approaches. The first is called **complete enumeration**. It consists of the collection of data on the survey items from each unit of the population. This procedure is used in censuses of population, agriculture, livestock, retail stores, industrial establishments, etc. The other approach, which is more general since the first approach can be considered as its special case, is based on the use of sampling methods and consists of collecting data on survey items from selected units of the population. **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 as a whole**. It also provides procedures for the estimation of results that would be obtained if a comparable survey were taken on all the units in the population. In other words, **a sampling method makes it possible to estimate the population totals, averages or proportions while reducing at the same time the size of survey operations**.

A distinctive feature of surveys based on the use of sampling methods called for brevity sample surveys is **sampling errors**. The feature refers to the discrepancies between the sample estimates and the population values that would be obtained from enumerating all the units in the population in the same way in which the sample is enumerated. These discrepancies are unavoidable because sample estimates are based on data for only a sample of units. The employment of sampling methods, however, enables estimates of the average magnitude of these discrepancies to be made. Sampling methods also provide the means of fixing in advance the details of survey design, such as the size of the sample, in such a way that the average magnitude of the sampling errors does not exceed

the amount allowed with a pre-assigned probability. In other words, sampling methods enable us to control the precision of sample estimates within limits fixed in advance.

Sampling methods are based on laws of chance and the application of the theory of probability. There are other methods of sampling referred to under the name of purposive selection or judgement sampling. In these methods, units are selected in the sample according to how typical they are of the population according to the judgment of specialists in the subject matter. The composition of the sample resulting from the application of such a selection procedure is influenced by the personal judgment of those responsible for selection. The procedure is not objective; neither is it based on the principles of the theory of probability. Consequently, it does not provide the possibility of estimating and controlling the magnitude of sampling errors. Here, we shall concern ourselves only with probability sampling.

A simple way of obtaining a probability sample is to draw the units one by one with a known probability of selection assigned to each unit of the population at the first and each subsequent draw. The successive draws may be made with or without replacing the units selected in the preceding draws. The former is called the procedure of sampling with replacement, the latter sampling without replacement.

The application of the probability sampling method assumes that the population can be sub-divided into a finite number of distinct and identifiable units called sampling units. It is irrelevant for the sampling procedure what the sampling units are. They may be natural units such as individuals in a human population or fields in a crop survey or natural aggregates of such units like families or villages, or they may be artificial units such as a single plant, a row of plants or a plot of specified size, in sampling a field. For sampling purposes, it is essential to be able to list all of the sampling unit in the population. Such **a list is called the frame and provides the basis for the selection and identification of the units in the sample**. Examples of a frame are a list of agricultural fields, list of households in a village, list of villages in a district, list of districts in a State, list of States in India etc. The village forms the sampling unit and provides the means for further selecting a sample of household, fields and plots. The frame also often contains information about the size and structure of the population. That information is used in sample surveys in a number of ways as will be explained subsequently.

3.2 NEED FOR STATISTICAL DATA

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conversation and controversy require numerical data for their resolution. Data resulting from the physical, chemical, and biological experiments in the form of observations are used to test different theories and hypotheses. Various social and economic investigations are carried out through the use and analysis of relevant data. The data collected and analyzed in an objective manner and presented suitably serve as basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include Government, Industry, Business, Research Institution, Public Organizations, and International 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, metrology, 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 health, mortality, and population could be used for formulating policies for checking population growth. Similarly, information on forestry and environment is needed to plan strategies for a cleaner and healthier life. Agricultural production data 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 supply positions for proper planning of production levels and sales campaigns.

The research institutions need data to verify the earlier findings or to draw new inferences. The data are used by public organizations to assess the State policies, and to point it out to the administration if these are not up to the expectations of the people. The international organizations collect data to present comparative positions of different countries in respect of economy, education, health, culture, etc. Besides, they also use it to frame their policies at the international level for the welfare of people.

3.3 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.

Definition 3.1: The data collected by the Investigator from the original source are called *primary data*.

Definition 3.2: 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.

3.4 SOME TECHNICAL TERMS

Definition 4.1: An *element* is a unit for which information is sought.

Definition 4.2: The *population or universe* is an aggregate of elements, about which the inference is to be made.

Populations are called finite or infinite, depending on the number of units constituting it.

Definition 4.3: *Sampling units* are non-overlapping collections of elements of the population.

Definition 4.4: A list of all the sampling units in the population to be sampled is termed *frame or sampling frame*.

Definition 4.5: A sub-set of population selected from a frame to draw inferences about a population characteristic is called a *sample*.

In practice, number of units selected in a sample is much less than the number in the population. Inferences about the entire population are drawn from the observations made on the study variable for the units selected in the sample.

3.5 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:

3. 5.1: 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.

3.5.2: 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.

3.5.3: 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.

3.5.4: 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.

3.5.5: 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.

3.6 SAMPLING PROCEDURES

Definition 6.1: If the units in the sample are selected using some probability mechanism, such a procedure is called *probability sampling.*

This type of survey assigns to each unit in the population a definite probability of being selected in the sample. Alternatively, it enables us to define a set of distinct samples which the procedure is capable of selecting if applied to a specific population. The sampling procedure assigns to each possible sample a known probability of being selected. One can build suitable estimators for different population characteristics for

probability samples. For any sampling procedure of this type, one is in a position to develop theory by using probability apparatus. It is also possible to obtain frequency distribution of the estimator values it generates if repeatedly applied to the same population. The measure of the sampling variation can also be obtained for such procedures, and the proportion of estimates that will fall in a specified interval around the true value can be worked out.

Definition 6.2: The procedure of selecting a sample without using any probability mechanism is termed as *non-probability sampling.*

Purposive sampling (also termed as Judgment sampling) is common when special skills are required to form a representative subset of population. For instance, auditors often use judgment samples to select items for study to determine whether a complete audit of items may be necessary. Sometimes, quotas are fixed for different categories of population based on considerations relevant to the study being conducted, and selections within the categories are based on personal judgment. This type of sampling procedure is also termed quota sampling.

Obviously, these methods are subject to human bias. In appropriate conditions, these methods can provide useful results. They are, however, not amenable to the development of relevant theory and statistical analysis. In such methods, the sampling error can not be objectively determined. Hence, they are not comparable with the available probability sampling methods.

Definition 6.3: In *with replacement* (WR) sampling, the units are drawn one by one from the population, replacing the unit selected at any particular draw before executing the next draw.

As the constitution of population remains same at each draw, some units in the with replacement sample may get selected more than once. This procedure gives rise to N^n possible samples when order of selection of units in the sample is taken into account, where N and n denote the population and sample sizes respectively.

Example 6.1

Given below are the weights (in kgs) of 4 participants of a training programme:

Participant	:	A	B	C	D
Weight (in kgs)	:	55	80	65	70

Enumerate all possible WR samples of size 2. Also, write values of the study variable (weight) for the sample units.

Solution: Here, N=4 and n=2. There will, therefore, be $4^2 = 16$ possible samples. These have been enumerated below along with the weight values for the units included in the sample.

Table 6.1: Possible samples along with their variable values

Sample	Participants in the sample	Weight for the sampled Participants	Sample	Participants in the sample	Weight for the sampled Participants
1	A, A	55, 55	9	C, A	65, 55
2	A, B	55, 80	10	C, B	65, 80
3	A, C	55, 65	11	C, C	65, 65
4	A, D	55, 70	12	C, D	65, 70
5	B, A	80, 55	13	D, A	70, 55
6	B, B	80, 80	14	D, B,	70, 80
7	B, C	80, 65	15	D, C	70, 65
8	B, D	80, 70	16	D, D	70, 70

Definition 6.4: In *without replacement* (WOR) sampling, the units are selected one by one from the population, and the unit selected at any particular draw is not replaced back to the population before selecting a unit at the next draw.

Obviously, no unit is selected more than once in a WOR sample. If the order of selection of units in the sample is ignored, then there are $\binom{N}{n}$ possible samples for this selection procedure.

Example 6.2

Using data of Example 6.1, enumerate all possible WOR samples of size 2, and also list the weight values for the respective sample units.

Solution: In this case, number of possible samples will be $\binom{4}{2}$ = 6. These have been enumerated below. Note that no samples like AA or BB appear in the list of possible sample, and also the ordered samples like AB and BA are treated as the same sample.

Table 6.2: Possible samples along with their variable values

Sample	Participants in the sample	Weight for the sampled Participants	Sample	Participants in the sample	Weight for the sampled Participants
1	A, B	55, 80	4	B, C	80, 65
2	A, C	55, 65	5	B, D	80, 70
3	A, D	55, 70	6	C, D	65, 70

3.7 NEED FOR SAMPLING

Considering that some margin of error is permissible in the data needed for practical purposes, an effective alternative to a complete enumeration survey can be a 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 a sample survey, a complete enumeration survey 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 except on a sampling basis. Besides, in case of descriptive surveys, a complete enumeration survey is just not practicable. Thus, if the interest is to obtain the average weight per student in a batch, then one will have to confine the observations, of necessity, to a part (or a sample) of the population or universe and to infer about the population as a whole on the basis of the observations on the sample. However, since an inference is made about the whole population from a part in a sample survey, the results are likely to be different from the population values and the differences would depend on the selected part or sample. Thus the information provided by a sample is subject to a kind of error which is known as sampling error. On the other hand, as only a part of the population is to be surveyed, there is greater scope for eliminating the ascertainment or observational errors by proper controls and by employing trained personnel than is possible in a complete enumeration survey. 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 for 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 a sample survey. One of the principal steps in a sample survey relate to methods of primary data collection.

3.8 METHODS OF COLLECTING PRIMARY DATA

The different methods of collecting data are:

- Physical observation or measurement
- Personal interview
- Mail enquiry
- Telephonic enquiry
- Web-based enquiry
- Method of Registration
- Transcription from records

The first six methods relate to 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 demerits 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.

3.9 VARIOUS CONCEPTS AND DEFINITIONS

a) 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.

b) 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.

c) 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 districts in a particular State in India. The frame should be up to date and free from errors of omission and duplication of sampling units.

d) 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.

e) 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 getting 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.

f) 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 $\overline{Y} = \sum_{i=1}^{N} Y_i / N$.

g) 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 $y_1, y_2, ..., y_n$. 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.

h) Sample design

A clear specification of all possible samples of a given type with their corresponding probabilities is said to constitute a sample design. For example, suppose we select a sample of n units with equal probability with replacement, the sample design consists of N^n possible samples (taking into account the orders of selection and repetitions of units in the sample) with $1/N^n$ as the probability of selection for each of them, since in each of the n draws any one of the N units may get selected. Similarly, in sampling n units with equal probability without replacement, the number of possible samples (ignoring orders of selection of units) is $\binom{N}{n}$ and the probability of selecting each of the sample is $1/\binom{N}{n}$.

i) Unbiased Estimator

Let the probability of getting the i-th sample be P_i and let t_i be the estimate, i.e., the value of an estimator t of the population parameter θ based on this sample (i=1,2,…,M_o), M_o being the total number of possible samples for the specified probability scheme. The expected value or the average of the estimator t is given by $E(t) = \sum_{i=1}^{M_o} t_i \,.\, P_i$

An estimator t is said to be an unbiased estimator of the population parameter θ if its expected value is equal to θ irrespective of the y-values. In case expected value of the estimator is not equal to population parameter, the estimator t is said to be a biased estimator of θ. The estimator t is said to be positively or negatively biased for population parameter according as the value of the bias is positive or negative.

j) Measures of error

Since a sample design usually gives rise to different samples, the estimates based on the sample observations will, in general, differ from sample to sample and also from the value of the parameter under consideration. The difference between the estimate t_i based on the i-th sample and the parameter, namely $(t_i - \theta)$, may be called the error of the estimate and this error varies from sample to sample. An average measure of the divergence of the different estimates from the true value is given by the expected value of the squared error, which is $M(t) = E(t-\theta)^2 = \sum_{i=1}^{M_0} (t_i - \theta)^2 \,.\, P_i$

and this is know as mean square error (MSE) of the estimator. The MSE may be considered to be a measure of the accuracy with which the estimator t estimates the parameter.

The expected value of the squared deviation of the estimator from its expected value is termed sampling variance. It is a measure of the divergence of the estimator from its expected value and is given by

$$V(t) = \sigma^2(t) = E\{t - E(t)\}^2 = E(t^2) - \{E(t)\}^2$$

This measure of variability may be termed as the precision of the estimator t.

The MSE of t can be expressed as the sum of the sampling variance and the square of the bias. In case of unbiased estimator, the MSE and the sampling variance are same. The square root of the sampling variance i.e. $\sigma(t)$ is termed as the standard error (SE) of the estimator t. In practice, the actual value of $\sigma(t)$ is not generally known and hence it is usually estimated from the sample itself.

k) Confidence interval

The frequency distribution of the samples according to the values of the estimator t based on the sample estimates is termed as the sampling distribution of the estimator t. It is important to mention that though the population distribution may not be normal, the sampling distribution of the estimator t is usually close to normal, provided the sample size is sufficiently large. If the estimator t is unbiased and is normally distributed, the interval $\left\{t \pm K.SE(t)\right\}$ is expected to include the parameter θ in P% of the cases where P is the proportion of the area between –K and +K of the distribution of standard normal variate. The interval considered is said to be a confidence interval for the parameter θ with a confidence coefficient of P% with the confidence limit t – K.SE(t) and t + K.SE(t). For example, if a random sample of the records of batteries in routine use in a large factory shows an average life t = 394 days, with a standard error SE(t) = 4.6 days, the chances are 99 in 100 that the average life in the population of batteries lies between

t_L = 394 - (2.58)(4.6) = 382 days

t_U = 394 + (2.58)(4.6) = 406 days

The limits, 382 days and 406 days are called lower and upper confidence limits of 99% confidence interval for t. With a single estimate from a single survey, the statement "θ lies between 382 and 406 days" is not certain to be correct. The "99% confidence" figure implies that if the same sampling plan were used many times in a population, a confidence statement being made from each sample, about 99% of these statements would be correct and 1% wrong.

l) Sampling and Non-sampling error

The error arising 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, incompleteness 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.

3.10 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:

- **Lottery Method**, and
- **Use of Random Number Tables.**

3.10.1: 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.

3.10.2: 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,** and
- **A million random digits Table.**

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-digits 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 by selecting 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 used numbers is, therefore, modified and some of these modified procedures are:

- **Remainder Approach,**
- **Quotient Approach,** and
- **Independent Choice of Digits**

3.10.3: Remainder Approach: Let N be a 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 the serial number equal to the remainder obtained on dividing k by N is selected. If the remainder is zero, the last unit is selected. As an illustration, let N = 123, the 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, the remainder is 41. Hence, the unit with serial number 41 is selected in the sample.

3.10.4: Quotient Approach: Let N be a 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 (N'-1). Dividing k by d, the quotient q is obtained and the unit bearing the serial number (q + 1) is selected in the sample. As an illustration, let N=16 and hence N'=96 and d = 96 / 16 = 6. Let the two-digit random number chosen be 65 which lies between 0 and 95. Dividing 65 by 6, the quotient is 10 and hence the unit bearing serial number (10+1) = 11 is selected in the sample.

It may be mentioned here that while using the Random Number Table, any starting point can be used, and one can move in any pre-determined direction along the rows or columns. If in any problem, more than one sample is to be selected, each should have its independent starting point.

3.11 SIMPLE RANDOM SAMPLING

Simple random sampling (SRS) is a method of selecting 'n' units out of 'N' units such that each one of the possible non-distinct samples has an equal chance of its being chosen. 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 are then drawn either by means of a Table of random numbers or by means of a computer program that produces such a Table. Sampling where each member of a population may be chosen more than once is called sampling with replacement (WR). Similarly a method of sampling in which each member cannot be chosen more than once is called sampling without replacement (WOR). Population is either finite or infinite. It can be seen that in SRSWOR, the probability of selecting the units in the sample is equal for all the units. Let Y be the characteristic of interest. The N units that comprise the population are denoted by $y_1, y_2, ..., y_N$. Let the population mean, $\bar{Y}_N$, the parameter of interest, be denoted by $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$.

We denote by $\bar{y}_n$ the sample mean, where $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ is an unbiased estimator of the population mean. In other words, the average value of $\bar{y}_n$ over all possible samples ($^{N}C_n$ in this case) is equal to $\bar{Y}_N$.

Also, the sampling variance of $\bar{y}_n$ is given by

$$V(\bar{y}_n) = (\frac{1}{n} - \frac{1}{N}) S^2 \qquad (11.1)$$

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

An unbiased estimator of this sampling variance is given by

$$\hat{V}(\bar{y}_n) = (\frac{1}{n} - \frac{1}{N}) s^2 \qquad (11.2)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y}_n)^2$ is the sample mean square.

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

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$V(\bar{y}_n) = \frac{\sigma^2}{n} \quad \text{and} \qquad (11.3)$$

$$\hat{V}(\bar{y}_n) = \frac{s^2}{n} \qquad (11.4)$$

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.

For estimating the population total $Y = \sum_{i=1}^{N} Y_i$, we have an estimator

$$\hat{Y} = N\sum_{i=1}^{n} y_i / n = N.\bar{y}_n \qquad (11.5)$$

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

This estimator can be expressed as $\hat{Y} = \sum_{i=1}^{n} w_i y_i = (N/n)\sum_{i=1}^{n} y_i$ where $w_i = N/n$.

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

The estimate of sampling variance of $\hat{Y}$ is given by $\hat{V}(\hat{Y}) = \hat{V}(N.\bar{y}_n) = N^2.\hat{V}(\bar{y}_n)$ and the standard error of $\hat{Y}$ is given by $S\hat{E}(\hat{Y}) = S\hat{E}(N.\bar{y}_n) = N.S\hat{E}(\bar{y}_n)$.

From the above, it is evident 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) $V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$

3.12 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		

(i) 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.

(ii) 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:

SAMPLE-I

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 = $\hat{\bar{Y}}_N = \bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{292}{10} = 29.2$ q

$$Estimate\ of\ Population Total = \hat{Y} = N\ x\ \bar{y}_n = 108\,x\,29.2 = 3153.6\,q$$

The estimate of standard error of $\hat{Y}$ is given by

$$S\hat{E}\left(\hat{Y}\right) = S\hat{E}\left(N.\bar{y}_n\right) = N.S\hat{E}\left(\bar{y}_n\right)$$

where $S\hat{E}\left(\bar{y}_n\right) = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}\, s$

$$where\ s^2 = \frac{1}{n-1}\sum\left(y_i - \bar{y}_n\right)^2 = \frac{1}{10-1} \times 1533.6 = 170.4\ q^2$$

$$So,\ s = \sqrt{170.4} = 13.0537\,q$$

$$Hence,\ S\hat{E}\left(\bar{y}_n\right) = \sqrt{\left(\frac{1}{10} - \frac{1}{108}\right)}\,x13.0537 = 0.3012\,x13.0537 = 3.9322$$

ii) 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 $\bar{Y}_N = \frac{3320}{108} = 30.74\,q$ 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.

SAMPLE-II

Random Number	Sampling Unit Sl. No. (Remainder of Random Number/108)	Yield (q)
798	**42**	**29**
831	**75**	**60**
074	**74**	**19**
005	**5**	**55**
423	**99**	**42**
138	**30**	**35**
971	**107**	**37**
166	**58**	**25**
455	**23**	**30**
201	**93**	**18**

Estimate of Population Average yield = $\hat{\bar{Y}}_N = \bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{340}{10} = 34.0\,\text{q}$

$Estimate\ of\ Population Total = \hat{Y} = N\,x\,\bar{y}_n = 108\,x\,34.0 = 3672\,q$

The estimate of standard error of $\hat{Y}$ is given by

$S\hat{E}\left(\hat{Y}\right) = S\hat{E}\left(N.\bar{y}_n\right) = N.S\hat{E}\left(\bar{y}_n\right)$

where $S\hat{E}\left(\bar{y}_n\right) = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}\,s$

$where\ s^2 = \frac{1}{n-1}\sum\left(y_i - \bar{y}_n\right)^2 = \frac{1}{10-1} \times (1533.6 = 170.4\ q^2$

$So,\ s = \sqrt{170.4} = 13.0537\,q$

$Hence,\ S\hat{E}\left(\bar{y}_n\right) = \sqrt{\left(\frac{1}{10} - \frac{1}{108}\right)}\,x\,13.0537 = 0.3012\,x\,13.0537 = 3.9322$

ii) The 95% confidence interval for population mean is given by

$$\bar{y}_n \pm t_{0.05/(10-1=9)df} \times S\hat{E}(\bar{y}_n) = 29.2 \pm 2.262 \times 3.9322 = 29.2 \pm 8.89$$

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 $\bar{Y}_N = \frac{3320}{108} = 30.74\,q$ 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.

SAMPLE-III

Random Number	Sampling Unit Sl. No. (Remainder of Random Number/108)	Yield (q)
034	**34**	**42**
977	**rejected**	
167	**59**	**29**
125	**17**	**29**
555	**15**	**75**
162	**54**	**08**
844	**88**	**39**
630	**90**	**44**
332	**8**	**42**
576	**36**	**11**
181	**73**	**32**

Estimate of Population Average yield = $\hat{\bar{Y}}_N = \bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{351}{10} = 35.1\,\text{q}$

$$Estimate\ of\ Population Total = \hat{Y} = N\ x\ \bar{y}_n = 108\,x\,35.1 = 3790.8\,q$$

The estimate of standard error of $\hat{Y}$ is given by

$$S\hat{E}(\hat{Y}) = S\hat{E}(N.\bar{y}_n) = N.S\hat{E}(\bar{y}_n)$$

where $S\hat{E}(\bar{y}_n) = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}\, s$

where $s^2 = \frac{1}{n-1}\sum (y_i - \bar{y}_n)^2 = \frac{1}{10-1}(3180.90 - 10x35.1^2) = \frac{1}{9} \times 3180.90 = 353.43\ q^2$

So, $s = \sqrt{353.43} = 18.80\, q$

Hence, $S\hat{E}(\bar{y}_n) = \sqrt{\left(\frac{1}{10} - \frac{1}{108}\right)} x 18.80 = 0.3012\, x 18.80 = 5.6588$

ii) The 95% confidence interval for population mean is given by

$\bar{y}_n \pm t_{0.05/(10-1=9)df} \times S\hat{E}(\bar{y}_n) = 35.1 \pm 2.262 \times 5.6588 = 35.1 \pm 12.80$

So, the 95% confidence interval for population mean is (35.1 - 12.80 to 35.1 + 12.80) i.e. (22.30, 47.90). It can be seen clearly that the population mean $\bar{Y}_N = \frac{3320}{108} = 30.74\, q$ 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.

SAMPLE_VII

Random Number	**Sampling Unit Sl. No. (Remainder of Random Number/108)**	**Yield (q)**
601	**61**	**25**
245	**29**	**31**
889	**25**	**29**
882	**18**	**38**
238	**22**	**28**
842	**86**	**22**
839	**83**	**32**
443	**11**	**25**
996	**rejected**	
802	**46**	**30**
552	**12**	**25**
Total		**285**

Estimate of Population Average yield = $\hat{\bar{Y}}_N = \bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{285}{10} = 28.5$ q

Estimate of Population Total $= \hat{Y} = N\, x\, \bar{y}_n = 108\, x\, 35.1 = 3790.8\, q$

The estimate of standard error of $\hat{Y}$ is given by

$S\hat{E}(\hat{Y}) = S\hat{E}(N.\bar{y}_n) = N.S\hat{E}(\bar{y}_n)$

Where

$$S\hat{E}(\bar{y}_n) = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}\, s$$

$$\text{where } s^2 = \frac{1}{n-1}\sum (y_i - \bar{y}_n)^2 = \frac{1}{10-1}(3180.90 - 10x35.1^2) = \frac{1}{9} \times 220.00 = 24.67\ q^2$$

$$\text{So, } s = \sqrt{24.67} = 4.97\, q$$

$$\text{Hence, } S\hat{E}(\bar{y}_n) = \sqrt{\left(\frac{1}{10} - \frac{1}{108}\right)} x\, 4.97 = 0.3012\, x\, 4.97 = 1.4971$$

ii) The 95% confidence interval for population mean is given by

$$\bar{y}_n \pm t_{0.05/(10-1=9)df} \times S\hat{E}(\bar{y}_n) = 28.5 \pm 2.262 \times 1.4971 = 28.5 \pm 3.3864$$

So, the 95% confidence interval for population mean is (28.5 - 3.3864 to 28.5 + 3.3864) i.e. (25.1136, 31.8864). It can be seen clearly that the population mean $\bar{Y}_N = \frac{3320}{108} = 30.74\, q$ 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.

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.

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.

DETERMINATION OF SAMPLE SIZE

K.K. TYAGI

Former Principal Scientist

Indian Agricultural Statistics Research Institute, New Delhi-110012

(e-mail: kkanttyagi@gmail.com)

4.1 INTRODUCTION

In the planning of a sample survey, determination of sample size is an important decision which a survey statistician has to take while deciding the sampling plan. One has to be careful while deciding the sample size, because too large a sample implies waste of resources, and too small a sample diminishes the utility of the results. An efficient sampling plan should enable an optimum utilisation of budgetary resources to provide the best estimators of the population parameters. As is well known, efficiency of an estimator is normally measured by inverse of mean square error (or variance in case of unbiased estimators). A desirable proposition would be to minimize the cost as well as variance simultaneously. But, unfortunately, it is not possible. With an increase in the sample size, normally the cost of the survey increases while the variance decreases, thereby increasing the efficiency. Thus for determination of sample size, a balance is required to be struck which is reasonable with respect to cost as well as efficiency. Sampling theory provides a framework within which the problem of determining sample size may be tackled reasonably.

We first consider the estimation of sample size in case of simple random sampling. The problem has been analysed in a very elegant way by considering hypothetical example by Cochran (1977). We quote the example:

“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 can not 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 finite population correction (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

$$\sigma_p \cong \sqrt{(P.Q / n)} \qquad (1.1)$$

where $Q = (1 - P)$.

Equating half width of the confidence interval to the permissible error, we get

$$2 . \sigma_p = 5 \qquad (1.2)$$

or $2 \sqrt{(P.Q / n)} = 5$

or $n = 4 P . Q / 25 \qquad (1.3)$

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. We therefore ask the anthropologist if he can give us some idea of the likely value of P. He replies 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.

4.2 PRINCIPAL STEPS INVOLVED IN THE CHOICE OF A SAMPLE SIZE

- A statement about the margin of error to be tolerated in the results.
- Choice of desired confidence level.
- Some equation that connects n with the desired precision of the sample should be found.
- This equation will contain, as parameters, certain unknown properties of the population. This must be estimated in order to give specific results.
- Usually in a sample survey, more than one characteristic is measured. Sometimes, the number of characteristics is large. If a desired degree of precision is prescribed

for each characteristic, the calculation leads to a conflicting values of n, one for each characteristic. Some method must be found for reconciling these values.

- Finally, the chosen value of n must be appraised to see whether it is consistent with the resources available to take the sample. This demands an estimation of the cost, labour, time and material required to obtain the proposed size of sample. It sometimes becomes apparent that n will have to be drastically reduced. One has to choose whether to proceed with a much smaller sample size, thus reducing precision, or to abandon efforts until more resources can be found. Regarding the choice of a level for tolerable margin of error and the confidence level, the user normally has only a vague idea and it is only through the discussions and clarifications that a quantitative specific measures are obtained. It may be remarked that these measures are mainly subjective and depend largely on the judgment of the user regarding the importance, applicability and vulnerability of the results.

Regarding the sample sizes in case of simple random sampling, the cases for qualitative and quantitative data are presented below:

4.3 QUALITATIVE DATA: ESTIMATION OF PROPORTIONS

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 that we are willing to incur that the actual error is larger than d; i.e., we want

$$\Pr (\mid p - P \mid \geq d) = \alpha \tag{3.1}$$

Simple random sampling is assumed, and p is taken as normally distributed. Now we know that,

$$\sigma_p = \sqrt{\left(\frac{N-n}{N-1}\right)} \sqrt{\frac{P.Q}{n}} \tag{3.2}$$

Hence the formula that connects n with the desired degree of precision is

$$d = t\sqrt{\left(\frac{N-n}{N-1}\right)} \sqrt{\frac{P.Q}{n}} \tag{3.3}$$

where t is the abscissa of the normal curve that cuts off an area of α at the tails. Solving for n, we find

$$n = \frac{\dfrac{t^2\ P\ Q}{d^2}}{1+\dfrac{1}{N}\left(\dfrac{t^2\ P\ Q}{d^2}-1\right)} \tag{3.4}$$

For practical use, an advance estimate p of P is substituted in this formula. If N is large, a first approximation is

$$n_0 = \frac{t^2\ p\ q}{d^2} \tag{3.5}$$

In practice, we first calculate n_0. If n_0/N is negligible, n_0 is a satisfactory approximation to the n of (3.4). If not, it is apparent on comparison of (3.4) and (3.5) that n is obtained as

$$n = \frac{n_0}{1+(n_0-1)/N} \cong \frac{n_0}{1+(n_0/N)} \tag{3.6}$$

4.4 EXERCISE

In the hypothetical blood groups example, d = 0.05, p = 0.5, α = 0.05, t = 2.

Thus, q = 1 – p = 0.5, and from (3.5), we have

$$n_0 = \frac{(4)(0.5)(0.5)}{(0.0025)} = 400 \tag{4.1}$$

Let us assume that there are only 3200 people on the island. The fpc is needed, and we find

$$n = \frac{n_0}{1+(n_0-1)/N} = \frac{400}{1+\frac{399}{3200}} = 356 \tag{4.2}$$

The formula for n_0 holds also if d, p and q are all expressed as percentages instead of proportions. Since the product p.q 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.

4.5 QUANTITATIVE DATA: ESTIMATION OF POPULATION MEAN

Consider a population of size N from which a simple random sample is to be selected for estimating the population mean $\bar{Y}$. Suppose, we wish to control the relative error 'r' in the estimated population total or mean. As the sample mean $\bar{y}$ estimates the population mean $\bar{Y}$, the margin of error and confidence level are specified as

$$P\left(\left| \frac{\bar{y}-\bar{Y}}{\bar{y}} \right| \geq r \right) = P\left(\left| \frac{N\bar{y}-N\bar{Y}}{N\bar{Y}} \right| \geq r \right) = P\left(\left| \bar{y}-\bar{Y} \right| \geq r\bar{Y} \right) = \alpha \tag{5.1}$$

where α is a small probability. We assume that $\bar{y}$ is normally distributed. Its standard error is given by

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N}}\ \frac{S}{\sqrt{n}} \tag{5.2}$$

Hence

$$r\bar{Y} = t.\ \sigma_{\bar{y}} = t.\ \sqrt{\frac{N-n}{N}}\ \frac{S}{\sqrt{n}} \tag{5.3}$$

Solving for n gives

$$n = \left(\frac{t \cdot S}{r \cdot \bar{Y}}\right)^2 / \left[1 + \frac{1}{N}\left(\frac{t \cdot S}{r \cdot \bar{Y}}\right)^2\right] \tag{5.4}$$

Note that the population characteristic on which n depends is its coefficient of variation $S/\bar{Y}$. This is often a more stable quantity and easier to guess in advance than S itself.

As a first approximation, we take

$$n_0 = \left(\frac{t.S}{r.\bar{Y}}\right)^2 \tag{5.5}$$

by substituting for an advance estimate of ($S/\bar{Y}$).

If n_0 / N is appreciable, we compute n as in (3.6) as

$$n = \frac{n_0}{1 + n_0/N} \tag{5.6}$$

If instead of the relative error r, we wish to control the absolute error d in $\bar{Y}$, we take $n_0 = t^2S^2/d^2$.

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

$$n = \frac{1}{\left(\frac{d^2}{C^2} + \frac{1}{N}\right)} \tag{5.7}$$

where C is the % coefficient of variation of the population.

4.6 METHODOLOGICAL ISSUES RELATING TO DETERMINATION OF SAMPLE SIZE

The determination of sample size is generally based on

- the available financial and manpower resources, and

- the required level of reliability in the estimates expected from the sample.

Generally, it would be preferable to start with the second consideration, and if the budget is a constraint to assess the precision that can be achieved under that constraint in order to decide whether the achievable precision would be acceptable, and if not, whether the budget should be increased.

In a sample survey, the sampling error associated with a given sample size varies from item to item. For major items of frequent occurrence, such as area under a crop, generally the sampling error is less that than the minor items of infrequent occurrence such as use of pesticides/ insecticides. Similarly, for items which have greater variability, the sampling error would be larger than that for items having lesser variability. To decide on the sample size for the survey, it would be necessary to calculate the sample size required for estimating with the requisite precision a few major items of interest and take the largest of the indicated size requirements as the sample size for the survey.

4.7 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 (population mean and standard deviation) required for the determination of sample size for the various items of interest. However, that may not be always possible, unless the pilot survey is taken well in advance, because estimates of financial resource required are to be made available to the government quite in advance to make the requisite budget. Thus the determination of the sample size in most cases may have to be done in advance without any pilot survey.

In such cases the survey Statistician has to make use of the information which is readily available. He may often have to depend on the results of similar survey conducted in the past, preferably in the same country or elsewhere in the neighbouring countries. If no such results are available, the Statistician has to make reasonable guesses of the different parameters which enter the formula for determination of the sample size.

If a stratified multi-stage random sampling design is used, which is usually the case, the problems are further compounded because 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. What one can do in such circumstances is to proceed in stages by working out the sample size required for a simple random sample (SRS) and to make adjustments to the sample size to take into account the effects of multi-stage sampling and possible stratification as illustrated below.

Generally the level of precision desired of an estimate is expressed as a percentage of itself or, strictly speaking of the population parameter. But the subtle difference is usually ignored.

Let Y be the characteristic under study, $\bar{Y}$ be the population mean and $\bar{y}$ be the sample mean. Clearly, $\bar{y}$ is an estimate of $\bar{Y}$.

The required sampling precision is prescribed as a percentage of $\bar{Y}$. For example, sampling precision of $\bar{y}$ should be E per cent of $\bar{Y}$ i.e., the population average should lie between $(\bar{y} + \frac{E\ x\ \bar{Y}}{100})$ *and* $(\bar{y} + \frac{E\ x\ \bar{Y}}{l00})$.

Taking 95% confidence interval, which is the usual case, this implies that

$$\frac{E\,x\,\bar{Y}}{100} = 2\,x\,SE(\bar{y}) \Rightarrow \frac{SE(\bar{y})}{\bar{Y}}\,x100 = \frac{E}{2} \qquad (7.1)$$

Hence, percentage relative SE = $\frac{E}{2}$

If the sampling precision is set at 5% relative SE, we know that in simple random sampling

$$\text{Re}\,lative\ SE(\bar{y}) = \frac{SE(\bar{y})}{\bar{Y}}\ x\ 100$$

$$= \frac{\frac{\sigma}{\sqrt{n}}}{\bar{Y}} \text{ x } 100 = \frac{\sigma}{\bar{Y}} \text{ x } 100 \text{ x } \frac{1}{\sqrt{n}} = \frac{CV\,x100}{\sqrt{n}} \qquad (7.2)$$

Hence, $\frac{E}{2} = \frac{CV}{\sqrt{n}}\ x\ 100 \Rightarrow$ n = 4 x 10000 x $\frac{(CV)^2}{E^2}$ = 40000 $\frac{(CV)^2}{E^2}$ (7.3)

If E = 5, $n = 40000\,x\frac{(CV)^2}{25} = 1600\ (CV)^2$ (7.4)

If E = 10, $n = 40000\,x\ \frac{(CV)^2}{100} = 400\ (CV)^2$ (7.5)

Thus to determine the sample sizes, we require the value of coefficient of variation, which generally being a stable quantity can be estimated on the basis of a previous survey on the subject or a closely related subject.

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.

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.

PLANNING AND EXECUTION OF SAMPLE SURVEYS

U.C. SUD

Indian Agricultural Statistics Research Institute, New Delhi-110012

5.1 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:

- Formulation of data requirements - objectives of the survey
- Ad-hoc or repetitive survey
- Method of data collection
- Questionnaire versus schedules
- Survey, reference and reporting periods
- Problems of sampling frames
- Choice of sampling design
- Planning of pilot survey
- Field work
- Processing of data, and
- Preparation of report.

The different aspects listed above are inter-dependent.

i) Formulation of Data Requirements

The users i.e. the persons or organisations 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:

- A clear statement of the desired information in statistical terms
- Specification of the domain of study
- The form in which the data should be tabulated
- The accuracy aimed at in the final results and
- 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: Adhoc or Repetitive

An adhoc 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 adhoc 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 a variety 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:

- Direct Personal Interview,
- Questionnaires Sent Through Mail,
- Interview by Enumerators and
- Telephone Interview.

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.

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.

Interviews 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.

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.

- Survey period: The time period during which the required data is collected.
- Reference period: The time period to which the collective data for all the units should refer.
- 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

One of the main requirements for efficiently designing sample survey is a well constructed sampling frame. In actual practice, quite often frames are not always perfect. Various types of imperfection such as omission, duplication etc. exist in the available frame. In multi-stage sampling, the problems of securing a good sampling frame arises for each of the stages. Usually a frame for higher stage units, such as towns, urban blocks and villages is more stable than one for lower stage units such as farms and households, which are more subject to changes. In agricultural surveys,

normally the frames of first few stages of units upto village level are used from records while the frame of households, fields etc. within the villages are prepared afresh. This approach reduces the chances of imperfection in sampling frames.

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 auxillary information which are normally available are utilised 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 part surveys. It is desirable to design and carry out a pilot survey. It will be useful for

- Testing out provisional schedules and related instructions,
- Evolving suitable procedure for field and tabulation work, and
- 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 adhoc surveys, one may plan for adhoc 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 agencies are 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:

- Tabulation and summary of data and
- 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:

- Introduction & review of literature
- Objective(s)
- Scope
- Subject coverage
- Method of data collection
- Survey references and recording
- Sampling design and estimation procedure
- Tabulation procedure
- Presentation of results
- Activity of results
- Cost structure of the survey
- Agency for conducting the survey
- References.

5.2 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. Clarifications of doubts raised by the investigators are 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.

5.3 THREE KINDS OF SCHEDULE ITEMS

The information included in the schedule may be classified under three headings:

5.3.1 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.

5.3.2 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.

5.3.3 Questions on the Subject of the Survey

These questions may be directed towards obtaining more or less objective facts or toward revealing attitudes and opinions on matters of current interest.

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.

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 adopted 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.

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.

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 barrelled questions, such as, the following question to a public transport "Do you like travelling on trains and buses?". Respondent liking a one and disliking other would be in a dilemma in answering this question. Clearly it needs to be divided into two questions.

5.4 SUGGESTIONS FOR WORDING QUESTIONS

5.4.1 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.

5.4.2 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.

5.4.3 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.

5.4.4 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.

5.4.5 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 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.

5.4.6 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.

5.4.7 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.

5.4.8 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.

5.5 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 etc. for filling in the questionnaire/ schedule under the survey and a copy of the same should be provided to each Field Investigator.

5.6 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.

5.7 CONCLUDING REMARKS

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 experienced the questionnaire designer, 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

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.

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.

NON SAMPLING ERRORS IN SURVEYS

Prachi Mishra Sahoo

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

6.1 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.

6.2 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.

6.3 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: $(N_2 / N)(\bar{y}_1 - \bar{y}_2)$. 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.

6.4 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.

6.5 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.

NON-RESPONSE IN LARGE SCALE SURVEY

Kaustav Aditya

Indian Agricultural Statistics Research Institute, New Delhi-110012

7.1 Introduction

The need for statistical information seems endless in the modern society. In particular data are regularly collected to satisfy the need for information about specified sets of elements, called as finite population. One of the most important modes of data collection for satisfying such needs is a sample survey, that is, a partial investigation of the finite population. By 'population' we mean a group of units defined according to the aims and objective of the survey. The information that we seek about the population is usually the total number of units, aggregate values of the various characteristics, averages of various characteristics etc. A sample survey costs less than a complete enumeration, is usually less time consuming, and may even more accurate than the complete enumeration. The term sample is used for the set of units or portion of the aggregate of material which has been selected with the belief that it will be representative of the whole aggregate. 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 drawn in such a manner that each unit in the population has a predetermined probability of selection. The sampling theory deals with scientific and objective procedure of choosing an appropriate sampling design, i.e. selecting a sample from the population which is representative of the population as a whole and also provides suitable estimation procedure to estimate the population parameters. Sometimes the principal goal of sampling design is to achieve a stated degree of precision for minimum cost or maximizing precision for fixed cost. A basic requirement of good survey practice is that a measure of precision be provided for each estimate derived from survey data.

For each sampling design, it is assumed that the true values of the variables of interest could be made available for the elements of the population under consideration. However, this may not be true particularly for large scale surveys. Errors can occur at almost every stage of planning and execution of a large scale survey. These errors may be attributed 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 analysed. Two kinds of errors are distinguishable, i.e., sampling error and non-sampling error. The error which arises due to only a sample being used to estimate the population parameters is termed as sampling error because sample surveys use only a part of the population where estimators based on a sample are different from the corresponding population parameter. This difference between the estimator and the population parameter is called as sampling error whereas the errors arising due to mainly misleading definitions and concepts, inadequate frames, unsatisfactory questionnaire, defective methods of data collection, tabulation, coding, decoding, incomplete coverage of sample units, etc., are called non-sampling errors.

For the choice of a suitable sampling design one has to rely upon accurate, unambiguous and unduplicated sampling frame, which is the list of all sampling units with reference to which the relevant data are collected and recorded. The sampling frame is keystone around which selection process must be designed. Appraisal of the available or obtainable frame must dominate the search for good selection procedure and choice among several alternatives. But, many a times, the sampling frames are either not available or are not in the form as desired by the sampler and suffer from various imperfections resulting incomplete, inaccurate and duplicated access to all the units in target population. Consequently, the precision of statistical results for the target population is affected by the error component which arises from the fact that the population to which results are needed does not conform to each other due to imperfection of the frame. The imperfection of the frame is attributable to various non-sampling errors like deviation in the content, incorrect frame of auxiliary information, non-coverage, measurement errors etc. Items such as interviewer bias, coding errors, false or erroneous replies and simple mistakes also fall into this class. In the simpler types of surveys in which the measuring devices are accurate and the quality of work is high, the assumption that the error of estimate arises solely from the random sampling variation that is present when a part of the population is measured instead of the whole population, holds reasonably well whereas in complex surveys, when difficult problems of

measurements are involved the estimators may also suffer from another source called as non-sampling error. In specification, three additional sources of error may be present which are as follows,

1. Failure to measure some of the units in the chosen sample. This may be attributed to oversight or, with human populations or, failure to locate some individuals or their refusal to answer the question when located.
2. Errors of measurement on a unit. The measuring device may not be perfect. When the survey involves human populations the respondents may not possess accurate information or they may give biased or misleading answers.
3. Errors introduced in editing, coding and tabulating the results.

These sources of error necessitate a modification of the standard theory of sampling. The principal aim of such modification is to provide guidance about the allocation of resources for the reduction of random sampling errors and other errors and to develop methods for computing standard errors and the confidence limits that remain valid when the other errors are present.

There are many kinds of non-sampling errors, which is present in the survey. Those are,

a. **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 do a person have can be subject to different interpretations in some cultures.

b. **Coverage or frame error:** thisoccurs mostly due to imperfect frames or non-coverage during the survey.

c. **Measurement error:** These errors arise from the fact that what is observed or measured departs from the actual values of sample units.

d. **Processing errors:** Processing errors comprise:

- Editing errors.
- Coding errors.
- Data entry errors.
- Programming errors etc.

e. **Errors of estimation:** These arise in the process of estimation of the population parameters from the sample. These include errors of sample selection and implementation, nonresponse, as well as sampling variability and estimation of bias. This group of errors centres on the process of selecting sample design, implementation and estimation.

f. **Nonresponse:** The incomplete coverage of units, mentioned above, occurs due to non-availability of information from some units included in the sample. It happens if a questionnaire is mailed to a sample of units, and some respondents fail to return the completed questionnaire. If the visits are made to a sample of households, some respondents maybe not-at-home, and others may refuse to cooperate. The inability to collect relevant information for some of the sample units due to refusal by respondents to divulge information, their being not-at-home, sample units being inaccessible, or due to any other such reason, is termed nonresponse. Nonresponse produces error in survey estimates in two ways. First, there may be decrease in sample size or in the amount of information collected in response to a particular question resulting in larger standard errors. Second, and perhaps more important, a bias is introduced to the extent that non-respondents differ from respondents within a selected sample.

7.2 Sources of Nonresponse Error

There are three primary sources of nonresponse and they can be represented as a hierarchy. First, a sampled company may not be contacted, in which case the establishment does not have an opportunity to respond. This is referred to as a non-contact. Second, a sampled unit that is contacted may fail to respond. This represents unit nonresponse. Third, the unit may respond to the questionnaire incompletely. This level is referred to as item nonresponse. Failure to contact could occur in establishment surveys due to seasonal closings. For example, during vacation the leisure industry, seashore resorts close during the winter while ski resorts and ski equipment shops close during the summer. Similarly, the food processing industry is also affected both by seasonality and disturbances in the weather. An attempted contact may also fail because of a temporary

closing due to a strike or work stoppage, a possible event in industries with strong and radical labor unions. Attempted contacts may not succeed due to a failure to locate the company. The firm may have moved or changed telephone number, or an incorrect address may have been inserted on the universe file. In the case of mail surveys, the survey form might be sent to the wrong location, the form misplaced prior to mailing, or lost during the mailing process. Non attempted contacts may result from negligence or sabotage on the part of the interviewer or in the mailing operation. Also, there may not be enough time in the collection period to reach to all sampled units. The end result is that the sampled company is never contacted in the first place.

There are roughly four types of nonresponse present in the survey which are,

i. **Non-coverage:** this is failure to visit some units in the sample. This is a problem with the areal sampling units in which the interviewer must find and list all dwellings (according to some definition) in a city block. It arises also from the use of incomplete lists. Sometimes weather or poor transportation make it impossible to reach certain units during the period of the survey.

ii. **Not-at-homes:** this may occur due to temporary movement of residents to some other places.

iii. **Unable to answer:** the respondent may not have the required information or he may deliberately choose not to answer.

iv. **The hard core:** this groupcomprisepersons with negative mindset and they may refuse to be interviewed, or persons who are incapacitated, or others who are far from home during the whole time available for field work.

7.3 Estimation In The Presence Of Nonresponse:

Nonresponse is a common problem in surveys. The presence of nonresponse may not only render the estimates biased but there may also be increase in the standard errors of the estimates. Different techniques have been suggested to tackle the problem of non-response in the context of estimation of finite population mean. Most of the proposed techniques deal with the situation of non-response for single stage sampling designs. However, multi-stage sampling designs are commonly used in surveys. It is a common situation in surveys that information in some cases is not obtained at the first attempt. Callbacks and follow-ups aim to eliminate or at least greatly reduce the problem of

nonresponse in surveys to overcome the problem of non-availability of sampling frame of the ultimate sampling units. In theory, these techniques are commendable, but in practice they are not without problems. A long series of callbacks or follow-ups may prove costly and time consuming. Nonresponse may still be unacceptably high after the ultimate call or follow-up later, this is especially a common problem in mail surveys. An estimate obtained from such incomplete data may be misleading. One approach to overcome the problem of nonresponse is to re-contact the non-respondents and obtain the information through personal interview.

Pioneering the work in this context, Hansen and Hurwitz (1946) were the first to deal with the problem of incomplete samples in mail surveys which are commonly used for data collection in advanced countries due to their low cost. The Hansen and Hurwitz (1946) technique consists in selecting sub-sample of initial non-respondents of subsequent data collection with a more expensive method. The technique is, generally, applicable to mail surveys. The problem of nonresponse is common in mail surveys. The approach consist in taking a random sub-sample of the persons who have not been reached and make a major effort to interview everyone in the sub-sample. It was shown that unbiased estimation is possible despite the non-observation of certain elements in initial sample.The summery of the proposed technique is given bellow:

1. Select a sample of the respondents and mail the questionnaire to all of them.
2. After the deadline is over, identify the non-respondents and select a sub-sample of the non-respondents.
3. Collect data from the non-respondents in the subsample through interview.
4. Combine data from the two parts of the survey to estimate the population parameters.

It is important to note that the approach used by Hansen and Hurwitz (1946) was based on a deterministic response mechanism. Let the population of size Nbe divided into two classes i.e. those who will respond at the first attempt belong to the response class and those who will not respond will be termed as representing the nonresponse class. Let N_1 and N_2be the number of units that belongs to the response and the nonresponse classes

respectively such that $N_1+N_2=N$. Let y_i be the value of the i-th response variable, i=1, 2, ..., N. The population mean $\bar{Y}$ can be written as,

$$\bar{Y} = \frac{N_1\bar{Y}_1 + N_2\bar{Y}_2}{N} = W_1\bar{Y}_1 + W_2\bar{Y}_2$$

where W_1 and W_2 are the proportions of units in the response and nonresponse classes such that $W_1 + W_2 = 1$, and $\bar{Y}_1$ and $\bar{Y}_2$ are the population means in these classes. Thus $\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_i$ and $\bar{Y}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} Y_i$

Let n be the size of the simple random sample drawn from Nunits. Let $n_1$out of n denote the number of units responding in the sample and $n_2$units denote the nonresponding units. Let h_2 be the size of subsample from the n_2 non-respondents to be interviewed so that $n_2=h_2f$. Unbiased estimators of N_1 and $N_2$are given by,

$$\hat{N}_1 = \frac{n_1 N}{n}, \hat{N}_2 = \frac{n_2 N}{n}$$

Let $\bar{y}_{h_2}$ denote the mean of h_2 observationsin the subsample and define,

$$\bar{y}_w = \frac{n_1\,\bar{y}_{n_1} + n_2\,\bar{y}_{h_2}}{n}, \text{where, } \bar{y}_{h_2} = \frac{1}{h_2}\sum_{i=1}^{h_2} y_i \text{ and } \bar{y}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$$

The estimator $\bar{y}_w$ is an unbiased estimator of $\bar{Y}$.

An interesting plan dealing with the problem of not-at-home was devised by Politz and Simmons (1949, 1950). The aim of this technique was to adjust the biases, without call-backs, which cropped up due to incomplete sample that did not distribute proportionately over the response class. The plan runs as follows: respondents to be included in the sample were visited only once by enumerators during a specific time on five week days (excluding Saturdays & Sundays). The respondent found at home was asked how many times in previous five days he was at home at the specific time of interview. If the respondent said that he was at home j (j=1, 2,..,m)number of days, the ratio $(j+1)/6$was considered as an estimate of the probability of availability of respondent in the sample. If the respondent was found not-at-home, no information was collected. Let the population consists of N units and n respondents be selected by simple random sampling with

replacement. Assuming that p_i denotes the probability that the *i*-th respondent is available at the time of call, an estimator of population mean is defined as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

where p_i is the probability of availability of the *i*-th respondent at the time of call. Of course, y_i is zero if the respondent is not available at the time of call. It may be seen that $\bar{y}$ is biased.

Hendricks (1949) discussed the method of extrapolation in mail surveys. He fitted a log linear model to estimate the population parameters. Scott (1961), Filion (1976) and Jones (1979) used linear model instead of log-linear model to estimate the population parameters. Another procedure suggested by Birnbaum and Sirken (1950) wasto fix the number of "call-backs", say, *k*, in advance for the enumeration of the selected sample. All nonresponding units were called back, up to a maximum number of *k* calls, till they responded. However sub-sampling of the non-respondents was not considered. It was assumed that all units in the population had the same probability of being available at exactly the *j*-th attempt, where $1 \le j \le k$. On the basis of available information on the different probabilities and the various cost components, the sampling error *S*, the bias *b* and the expected cost *C* can be calculated. A solution was obtained for surveys in which the interview consisted of only one question and the answer to which was either 'yes' or 'no'. The optimum values of k and the sample size n were obtained by minimizing C, for a fixed precision δ at a given probability level α, defined by,

$$P\left(\left|S+b\right| \le \delta\right) \ge 1-\alpha$$

A different procedure was suggested by Deming (1953). In this procedure the number of attempts k were fixed in advance to enumerate the selected sample. At each attempt all the available units were enumerated and a fraction of the remaining units were selected for the next attempt. The estimate was obtained as the simple average of the units which

were enumerated. The estimate was biased since the method of sampling did not ensure that all the selected units were enumerated. The bias however decreased with each subsequent attempt. To calculate the bias and the variance of the estimate, it was assumed that the population can be considered as consisting of a number of classes according to the average number of successful enumerations that can be made out of a fixed number of attempts. This was taken as the probability of a unit being available for enumeration at any attempt. For a fixed k, the bias, variance and expected cost of the survey could be calculated provided the relative size, mean and variance of each class were known.

Large scale surveys usually employ two stage or multistage sampling designs. Aditya et al. (2012, 2013) and Sud et al. (2012) extended the Hansen and Hurwitz (1946) technique under two stage sampling design.

REFERENCES

Aditya, K., Sud, U.C. and Chandra, H. (2014). Estimation of Domain Mean using Two Stage Sampling with Sub-Sampling of Non-respondents. Journal of the Indian Society of Agricultural Statistics. Accepted.

Aditya, K., Sud, U.C. and Chandra, H. (2012). Estimation of Domain Total for Unknown Domain Size in the Presence of Nonresponse. Statistics and Application, 10 (1-2), New Series, 13-25.

Belloc, B.B. (1954). Validation of morbidity survey data by comparison with hospital records. *J. Amer. Statist. Assoc., **49**, 832-846.*

Hansen, M.H. and Hurwitz, W.N. (1946). The problem of non response in sample surveys. *J. Amer. Statist. Assoc., **41**, 517-529.*

Hanson, R.H., and Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *J. Amer. Statist. Assoc., **53**, 635-655.*

Hansen, M.H., Hurwitz, W.N. and Bershad, M. (1961). Measurement errors in census and surveys. Bull. *Int. Statist. Inst. **38**, 2, 359-374.*

Hansen, M.H., Hurwitz, W.N. and Jubine, T.B. (1964). The use of imperfect lists for probability sampling at the U.S. Bureau of the Census. *Bull. Int. Statist. Inst., **40**.*

Kish, L., and Lansing, J.B. (1954). Response errors in estimating the value of homes. *J. Amer. Statist. Assoc., **49**, 520-538.*

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Statist. Soc., **109**, 325-370.*

Politz, A.N. and Simmons, W.R. (1949). An attempt to get the 'not at homes' into the sample without call backs. *J. Amer. Statist. Assoc.*, ***44***, *9-31, and 45, 136-137*.

Seal, K.C. (1962). Use of outdated frames in large scale sample surveys. *Calcutta Statist. Assoc. Bull.* ***11***.

Singh, R. (1983). On the use of incomplete frames in sample surveys. *Biometrical J.*, ***25***, *545-549.*

Singh, R. (1985). Estimation from incomplete data in longitudinal surveys. *J. Statist. Plan. Inf.*, ***7***, *163-170.*

Singh, R. (1986). Predecessor-Successor Method. *Encyclopaedia of Statistical Sciences, 7, 137-139. John Wiley & Sons Inc.*

Singh, R. (1989). Use of Imperfect frames in Census and Surveys. *Project Report, Indian Agricultural Statistics Research Institute (ICAR), New Delhi.*

Singh, R. and Rai, T. (1983). Use of Imputations for missing data in census and surveys. *Project Report, Indian Agricultural Statistics Research Institute (ICAR), New Delhi*.

Sukhatme, P.V. and Seth, G.R. (1952). Non sampling errors in surveys. *J. Ind. Soc. Agril. Statist.* ***4***, *5-41*.

Sud, U. C., Aditya, K., Chandra, H., and Parsad, R. (2012). Two Stage Sampling for Estimation of Population Mean with Sub-Sampling of Non-respondents" was published in *J. Ind. Soc. Agri. Stat.*, Vol. 66(3), pp. 447-457.

Sud, U. C., Aditya, K., Chandra, H. and Parsad, R. (2013). Two Stage Sampling with Two-phases at the Second Stage of Sampling for Estimation of Finite Population Mean under Random Response Mechanism. Journal of the Indian Society of Agricultural Statistics, 67 (3), 305-317.

RATIO AND REGRESSION METHODS OF ESTIMATION IN SAMPLE SURVEYS

Kaustav Aditya

Indian Agricultural Statistics Research Institute, New Delhi-110012

8.1 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

8.2 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.

8.2.1 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 $R_n = \dfrac{-Cov(R_n, \bar{x}_n)}{\bar{x}_N}$.

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, $C_x = \dfrac{S_x}{\bar{X}_N}$, $C_y = \dfrac{S_y}{\bar{Y}_N}$ 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.

8.2.2 Variance of the Ratio Estimator

The variance of the ratio estimator to a first approximation is given by

$$V_1(R_n) = R_N^2 \ (\frac{N-n}{Nn})(C_y^2 + C_x^2 \text{ - } 2\rho \ \mathrm{C_y} C_x),$$

and the variance of the ratio estimator of population mean to a first approximation is given by

$$V_1(\bar{y}_R) = \frac{N-n}{Nn} \left(\mathrm{S_y^2} + \mathrm{R_N^2} \ \mathrm{S_x^2} - 2R_N S_{yx} \right) .$$

8.2.3 Estimator of the variance of the ratio estimator

A consistent estimator of the relative variance of a ratio estimator is given by

$$\hat{V}_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn}\left[\frac{s_y^2}{\bar{y}_n^2} + \frac{s_x^2}{\bar{x}_n^2} \text{ - } \frac{2s_{yx}}{\bar{y}_n \bar{x}_n} \right]$$

and the estimator of variance of the ratio estimator of population mean to a first approximation is given by

$$\hat{V}_1(\bar{y}_R) = \frac{N-n}{Nn} \ \left[\mathrm{s_y^2} + R_n^2 s_x^2 - 2R_n s_{yx} \right]$$

where s_y^2, s_x^2 and s_{yx} are the corresponding sample values.

8.2.4 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.

8.3 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.

8.3.1 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

$$V_1\left(\bar{y}_{Rs}\right)=\sum_{t=1}^{K} p_t^2\left(\frac{1}{n_t}-\frac{1}{N_t}\right)\left(S_{ty}^2+R_{N_t}^2 S_{tx}^2-2R_{N_t}S_{txy}\right),$$

$$V_1(\bar{y}_{Rs})=\frac{1}{N}\sum_{t=1}^{K} p_t\ (\frac{N_t-n_t}{n_t})(S_{ty}^2+R_{N_t}^2 S_{tx}^2-2R_{N_t}S_{txy}),$$

$$V_1(\bar{y}_{Rs})=\frac{1}{N}\sum_{t=1}^{K} p_t\ (\frac{N_t-n_t}{n_t})(S_{ty}^2+R_{N_t}^2 S_{tx}^2-2R_{N_t}\rho_t S_{tx}S_{ty}).$$

The above formula is based on the assumption that n_t is large. A consistent estimator of $V_1\left(\bar{y}_{Rs}\right)$ is given by

$$\hat{V}_1\left(\bar{y}_{Rs}\right)=\frac{1}{N}\sum_{t=1}^{K} p_t(\frac{N_t-n_t}{n_t})(s_{ty}^2+R_{n_t}^2 s_{tx}^2-2R_{n_t}s_{tyx}).$$

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:

8.3.2 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

Relative Bias in $(\bar{y}_{Rc})=((E_1(\bar{y}_{Rc})-\bar{Y}_N)/\bar{Y}_N=\sum_{t=1}^{K} p_t^2(\frac{N_t-n_t}{N_t n_t})(C_{tx}^2-\rho_t C_{tx}C_{ty})$.

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 $R_n = \dfrac{\sum_{t=1}^{K} p_t \bar{y}_{n_t}}{\sum_{t=1}^{K} p_t \bar{x}_{n_t}}$

8.4 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.

8.4.1 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.

8.4.2 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 $-Cov(\hat{\beta}, \bar{x}_n)$.

8.4.3 Variance of the Simple Regression Estimate

To a first approximation,

$$V(\bar{y}_{lr}) \cong (\frac{1}{\mathrm{n}} - \frac{1}{N})\,\mathrm{S}_{\mathrm{y}}^2(1-\rho^2)$$

where ρ is the correlation coefficient between y and x in the population.

8.4.4 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 $\mathrm{r} = \frac{s_{xy}}{s_x s_y}$ is the sample correlation coefficient.

8.5 REGRESSION ESTIMATORS IN STRATIFIED SAMPLING

At first, we shall consider two difference estimates, namely

(i) Separate difference estimator

(ii) Combined difference estimate

8.5.1 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

$$\hat{V}(\bar{y}_{lrs}) = \sum_{i=1}^{K} p_i^2(\frac{1}{n_i} - \frac{1}{N_i})(s_{iy}^2 + \hat{\beta}_i^2 s_{ix}^2 - 2\hat{\beta}_i s_{ixy})$$

8.5.2 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

$$\hat{V}(\bar{y}_{lrc}) = \sum_{i=1}^{K} p_i^2 (\frac{1}{n_i} - \frac{1}{N_i})(s_{iy}^2 + \hat{\beta}^2 s_{ix}^2 - 2\hat{\beta} s_{ixy}).$$

8.6 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:

$$s_y^2 = \frac{1}{(n-1)} \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right]$$

$$s_x^2 = \frac{1}{(n-1)} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right]$$

$$s_{xy} = \frac{1}{(n-1)} \left[\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \right]$$

$$b = \frac{s_{xy}}{s_x^2} \qquad r = \frac{s_{xy}}{s_x . s_y}$$

$$\bar{y}_n = \frac{1}{n}\sum y_i \qquad \bar{x}_n = \frac{1}{n}\sum x_i$$

$$R_n = \frac{\bar{y}_n}{\bar{x}_n} \qquad \bar{X} = \frac{X_N}{N}$$

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

$$= \frac{\hat{V}\left(\bar{y}_{SRS}\right)}{\hat{V}\left(\bar{y}_R\right)} \text{ x } 100$$

(b) Estimate of Relative Efficiency of Regression estimate over Simple Mean estimate

$$= \frac{\hat{V}\left(\bar{y}_{SRS}\right)}{\hat{V}\left(\bar{y}_{lr}\right)} \text{ x}100$$

(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 selected 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:

$$\sum_{i=1}^{n} y_i = 566$$

$$\sum_{i=1}^{n} x_i = 681$$

$$\sum_{i=1}^{n} y_i^2 = 23450$$

$$\sum_{i=1}^{n} x_i^2 = 34617$$

$\sum_{i=1}^{n} x_i y_i = 26879$

$s_y^2 = 287.85$

$s_x^2 = 458.56$

$s_{xy} = 262.86$

$b = 0.57$

$r = 0.72$

$\bar{y}_n = 33.29$

$\bar{x}_n = 40.06$

$R_n = 0.83$

$\bar{X}_N = 47.98$

$\hat{\bar{y}}_R = 39.88$ $\hat{V}(\hat{\bar{y}}_R) = 9.86$ $SE(\bar{y}_R) = 3.14$ (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

$W_m = N_m / \sum N_m$, $\hat{R}_m = \bar{y}_m / \bar{x}_m$

(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 $\hat{V}(y_{Rs})$

$$\hat{V}(y_{Rs}) = \sum N_m^2 \left(\frac{1}{n_m} - \frac{1}{N_m}\right)\left(s_{y_m}^2 + \hat{R}_m^2 . s_{x_m}^2 - 2.\hat{R}_m . s_{xy_m}\right) = 2441137855.48$$

(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 $\hat{V}(y_{lc})$

$$\hat{V}(y_{lc}) = \sum_{m}^{K} \frac{W_m^2(1-f_m)}{n_m(n_m-1)} \sum_{j}^{n_m} \left[(y_{mj} - \bar{y}_m) - b_c(x_{mj} - \bar{x}_m)\right]^2 = 2020917640 \quad \text{where } f_m = \frac{n_m}{N_m}$$

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.) = $\frac{\hat{V}(y_{Rc})}{\hat{V}(y_{lc})}$. 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.) = $\frac{\hat{V}(y_{lc})}{\hat{V}(y_{ls})}$. 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.

VARIANCE ESTIMATION USING RESAMPLING TECHNIQUES

Tauqueer Ahmad

Indian Agricultural Statistics Research Institute, New Delhi-110012

9.1 INTRODUCTION

The theory and applications of sample surveys have grown considerably in the last 50 years. As the number and uses of sample surveys have increased, so has the need for methods of analyzing and interpreting the resulting data. It is often designed to analyse and interpret the resulting voluminous data by swifter methods. A basic requirement of a good survey is that a measure of precision be provided for each estimate computed from survey data collected on the basis of the survey design. The most commonly used measure of precision is the variance of the survey estimator.

There are several methods of variance estimation available in the literature. An important question now is, how to choose an appropriate variance estimator? The choice in general is a very difficult one. Factors like efficiency of the variance estimator, timeliness, cost, simplicity and other administrative conveniences must be considered. Estimate of variance should be computed in accordance with the complexity of the sample design: neglect of that complexity is a common source of serious mistakes (Kish and Frankel, 1970). On the other hand, trying to obtain more exact and complicated statistics like measure of variation (variance, standard error, mean square error) of first order statistics would become too difficult with non-linear statistics from complex surveys.

In case of complex surveys, the most commonly used methods for estimation of variance are:

i) Taylor Series Method
ii) The Method of Random Groups
iii) The Jackknife Method
iv) Balanced Repeated Replication Method and
v) Bootstrap Method.

Of these, the first method also known as Delta method or Linearization method, tries to derive linear approximation of the survey estimator and then the variance formula specific to the sampling design is obtained. The second method consists in dividing the population into several random groups using the same sample design and then computing the variance using the estimates obtained from each random group. The last three methods are based on sample reuse technique. The sample reuse approach for variance estimation involves taking replicated sub-samples from the total sample and forming survey estimates from the sub-samples. The variability of the sub-sample estimates is then used to estimate the variance of the total sample estimate. All the above stated methods are used for estimating the variance in the situation, when more complex designs and non-linear estimators are involved. Among these jackknife and bootstrap methods are the most commonly used methods for estimation of variance

The details of these two methods are given in sub-sequent sections.

9.2 JACKKNIFE METHOD

A sample reuse technique called the jackknife method has been suggested as a useful method of variance estimation. The method derives estimates of the parameter of interest from each of several sub-samples of the parent sample and then estimates the variance of the parent sample estimator from the variability between the sub-sample estimates.

Research on the jackknife method has proceeded along two distinct lines:

(1) Its use in bias reduction (Quenouille, 1949, 56) and
(2) Its use for variance estimation (Tukey, 1958)

Use of the jackknife technique in finite population estimation has been made by Durbin (1959) in ratio estimation. McCarthy (1966) discussed the application of jackknife as a method of variance estimation in complex surveys. Frankel (1971) termed the technique for the paired selection design as Jackknife Repeated Replication (JRR) method. Extensive discussion on jackknife method is given in Gray and Schucany (1972) and in a recent monograph by Efron (1982). The method briefly is as follows:

Assume, a random sample $x_1, x_2,\ldots,x_n$ from a population with unknown parameter θ. Let θ be the mean or median. Now, if θ is the mean, the sample average $\bar{x}$ and the sample standard deviation s, provide its measure of precision. However, if θ is the median, then the precision of the sample median is not so obvious from the sample. The delete one jackknife procedure is to obtain value of the statistic from various sub-samples, deleting one observation at a time.

Let $\hat{\theta}_n$ be the estimate of θ and let $\hat{\theta}^i_{n-1}$ be estimate of θ after x_i is deleted from the sample.

Let $\hat{\theta}^i_{ps} = n\hat{\theta}_n - (n-1)\hat{\theta}^i_{n-1}$ be the "pseudo sample value (PS)"

The average of P.S. values is the delete one jackknife estimate of θ,

$\hat{\theta}_j = \frac{1}{n}\sum \theta^i_{ps}$ and estimate of its precision is given by

$$\hat{V}\left(\hat{\theta}_j\right) = \frac{1}{n(n-1)}\sum\left(\hat{\theta}^i_{ps} - \hat{\theta}_j\right)^2 \qquad \ldots\ldots\ldots(2.1)$$

Extension to deleting more than one observation as well as deleting a group of observations at a time when the sample has been divided into several groups are available. Using robust methods such as the trimmed mean of the P.S. values in place of the average sometimes improves the behaviour of the jackknife estimate (Rustogi, 1990). The methodology of JRR has been demonstrated for its application in the case of several basic and non-linear estimators in Wolter's (1985).

9.3 BOOTSTRAP TECHNIQUES

Bootstrap method was proposed by Efron in 1979, who showed that bootstrap method correctly estimates the variance of a sample median, a case where jackknife is known to fail. Bootstrap methods are basically simulation methods conducted on high speed computers and aimed at generating new data sets from the observed original data set. The term 'bootstrap' - derived from the old saying about pulling yourself up by your own bootstrap-reflects the fact that one available sample gives rise to a large number of other samples. It is essentially a highly computer-intensive resampling technique to extract as much information as possible from the data on hand. It substitutes considerable amount of computation in place of theoretical analysis. The bootstrap method needs no prior assumptions about the distribution of observations as well as the estimators. It provides estimates of bias and standard error apart from estimates of other distributional properties of the estimators, however, complex it may be.
In complex survey data, often the sampling design induces a non-iid structure to the data such as without replacement sampling, stratification, multistage or unequal probability of selection. Though the techniques for variance estimation do exist, they often are cumbersome to implement or do not extend to complex designs. It would be desirable to have resampling methods that reuse the existing estimation system repeatedly, using computer power to avoid theoretical work and that can be applied to such data. In recognition of this need, various bootstrap techniques for variances estimation for complex survey data have been developed. The existing bootstrap techniques are as follows:

1. Naive bootstrap technique (BWR M1)
2. Bootstrap with replacement technique (BWR M2)
3. Rescaling bootstrap with replacement technique (RS BWR)
4. Bootstrap without replacement (Gross) technique (GS BWO)
5. Bootstrap without replacement (Sitter) technique (SR BWO)
6. Rescaling bootstrap without replacement technique (RS BWO)

Let us consider stratified random sampling in case of finite population, consisting of N units, which is partitioned into L non-overlapping strata of N_1, N_2, …, N_L units; thus $N_1+N_2+\ldots+N_L=N$. A simple random sample without replacement is taken independently from each stratum. The sample size within each stratum are denoted by $n_1,n_{2,\ldots,}n_L$ and the total sample sizes is $n=n_{1+}n_2+\ldots+n_{L.}$ The parameter of interest θ is non-linear function of the population mean vector $\overline{Y} = (\overline{Y}_1, \overline{Y}_2, \cdots, \overline{Y}_p)$ say $\theta = g(\overline{Y})$. This form includes ratios, regression and correlation coefficients. In this case, the unbiased estimator of $\overline{Y}$ is

$$\bar{y} = \sum_{h=1}^{L} W_h \bar{y}_h = \left(\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_p\right)$$

where

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} = (\bar{y}_{1h}, \bar{y}_{2h}, \cdots, \bar{y}_{ph}) \text{ and } W_h = \frac{N_h}{N}$$

For p = 1, an unbiased estimate of Var($\bar{y}$) is

$$\text{var}(\bar{y}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1 - f_h}{n_h}\right) s_h^2 \qquad \ldots\ldots\ldots(3.1)$$

Where $f_h = \frac{n_h}{N_h}$ and $s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$

9.3.1 The Naive Bootstrap Technique (BWR M1)

The Naive bootstrap technique first introduced by Efron (1979) is denoted by BWR M1. This technique is based on the i.i.d. property of the y_{hi}'s within each stratum. If the standard i.i.d. bootstrap is applied to the sample data $\{y_{hi}\}_{i=1}^{n_h}$ in each stratum, then the resulting resampling algorithm would take the following form:

(i) Draw a simple random sample $\{y^*_{hi}\}_{i=1}^{n_h}$ with replacement from the original sample $\{y_{hi}\}_{i=1}^{n_h}$ in stratum h, independently for each stratum. Calculate

$$\bar{y}^*_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y^*_{hi}, \quad \bar{y}^* = \sum W_h \bar{y}^*_h \quad \text{and} \quad \hat{\theta}^* = g(\bar{y}^*)$$

(ii) Repeat step (i), a large number of times, say B, to get $\hat{\theta}^{*1}$, $\hat{\theta}^{*2}$, . . ., $\hat{\theta}^{*B}$.

(iii) Estimate Var ($\hat{\theta}$) with

$$v_b = E_*\left(\hat{\theta}^* - E_*\hat{\theta}^*\right)^2 \qquad \ldots\ldots\ldots (3.1.1)$$

or its Monte Carlo approximation

$$v_b(a) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^*_{(.)})^2 \qquad \ldots\ldots\ldots (3.1.2)$$

where, $\hat{\theta}^*_{(.)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$ and E_* refers to the expectation with respect to bootstrap sampling.

In the linear case with p = 1, $\hat{\theta}^* = \sum W_h \bar{y}^*_h = \bar{y}^*$, v_b reduces to

$$v_b = \text{var}_*(\bar{y}^*) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{n_h - 1}{n_h}\right) s_h^2 \qquad \ldots\ldots\ldots (3.1.4)$$

Comparing (4) with the standard unbiased variance estimate var($\bar{y}$) given in equation (1), it is seen that v_b is not a consistent estimator for Var($\bar{y}$). This can be avoided by using a correction factor only if $n_h = k$ and $f_h = 1$ for all h, in which case $\frac{k}{k-1}(1-f)$ $\text{var}_*(\hat{\theta}^*)$ is consistent.

9.3.2 Bootstrap with Replacement Technique (BWR M2)

Recognizing the above mentioned scaling problem in the Naive bootstrap technique, Efron (1982) developed Bootstrap with replacement technique denoted by BWR M2. He suggested to draw bootstrap sample of size $(n_h - 1)$ with simple random sampling with replacement sampling designs instead of n_h independently from each stratum. Rest of the procedure is same as in the case of naive bootstrap technique.

9.3.3 Rescaling Bootstrap with Replacement Technique (RS BWR)

Rao and Wu (1988) proposed a rescaling of the standard bootstrap when $\hat{\theta} = g(\bar{y})$, a non-linear function of means. In this method, one applies the previously stated algorithm, with a general resample size m_h not necessarily equal to n_h, but rescales the resampled values appropriately so that the resulting variance estimator is the same as the usual unbiased variance estimator in the linear case. They have extended the method to unequal probability sampling without replacement and two stage cluster sampling without replacement with the use of some correction factor at the bootstrapping stage.

9.3.4 Bootstrap Without Replacement (Gross) Technique (GS BWO)

In addition to the above with replacement resampling schemes, a without-replacement bootstrap (BWO) was proposed by Gross (1980) in the case of a single stratum. Suppressing the h-th subscript, this method assumes N = kn for some integer k and creates a pseudo population of size N by replicating the data k times. The resample is then obtained by drawing n units without replacement from the pseudo population. Although, the BWO method is intuitively appealing, it does not yield the usual unbiased estimate of variance in the linear case.

9.3.5 Bootstrap Without Replacement (Sitter) Technique (SR BWO)

Sitter (1992) developed a method of bootstrapping denoted here by SR BWO which retains the desirable properties of BWR and BWO but extends to more complexes without replacement sampling designs. In general the method entails:

(a) Selecting a subsample without replacement from the original sample to mirror the original sampling scheme.
(b) Replacing this subsample in the original sample and
(c) Repeating this a specified number of times k_h.

This bootstrapping procedure is repeated a large number of times. The value of k_h is chosen such that the bootstrap estimate of variance matches the usual one in the linear case.

9.3.6 Rescaling Bootstrap without Replacement Technique (RS BWO)

It seems reasonable to design a resampling scheme that parallels the original sampling scheme as closely as possible. This is what so appealing about the BWO method as compared to the BWR methods and the rescaling method. Therefore, Ahmad (1997) developed a new BWO technique known as "Rescaling Bootstrap without Replacement" denoted here by RS BWO. The technique is as follows:

(i) Draw a simple random sample of size of size m_h without replacement from the given sample $\left\{y_{hi}\right\}_{i=1}^{n_h}$ of size n_h in stratum h, independently for each stratum.

Calculate

$$\tilde{y}_{hi} = \bar{y}_h + (1-f_h)^{1/2} \frac{\sqrt{m_h}}{\sqrt{n_h - m_h}} (y^*_{hi} - \bar{y}_h) \quad \ldots\ldots\ldots (3.6.1)$$

$\tilde{y}_h = \frac{1}{m_h} \sum_{i=1}^{m_h} \tilde{y}_{hi}$, where $f_h = \frac{n_h}{N_h}$, $\bar{y}^*_h$ and $\bar{y}_h$ are the bootstrap sample mean and original sample mean for the h-th stratum.

Obviously, one of the basic assumption of this technique is that n_h is sufficiently large as compared to m_h.

$$\tilde{y} = \sum_{h=1}^{L} W_h \tilde{y}_h , \qquad \tilde{\theta} = g(\tilde{y})$$

(ii) Replace the sample in the original sample and independently replicate step (i). Repeat this process a large number, say B, of times and calculate the corresponding estimates $\tilde{\theta}^1, \tilde{\theta}^2, \ldots, \tilde{\theta}^B$.

(iii) The bootstrap variance estimator of $\tilde{\theta} = g(\tilde{y})$ is given by

$$\tilde{v}_b = E_*(\tilde{\theta} - E_*\tilde{\theta})^2 \quad \ldots\ldots\ldots (3.6.2)$$

where, E_* denotes the expectation with respect to bootstrap sampling from a given sample.

The Monte-Carlo estimator $\tilde{v}_b(a)$ as an approximation to $\tilde{v}_b$ is given by

$$\tilde{v}_b(a) = \frac{1}{B-1} \sum_{b=1}^{B} (\tilde{\theta}^b - \tilde{\theta}_a)^2 \quad \ldots\ldots\ldots (3.6.3)$$

where

$$\tilde{\theta}_a = \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}^b$$

In the linear case, $\theta = \bar{Y}$ and p=1, $\tilde{v}_b$ reduces to the usual unbiased variance estimator var($\bar{y}$) for any choice of m_h, noting that

$$\text{var}_*(\tilde{y}_h) = (1-f_h)\frac{s_h^2}{n_h}$$

Hence,

$$\tilde{v}_b = \text{var}_*(\tilde{y}) = \sum_{h=1}^{L} W_h^2(1-f_h)\frac{s_h^2}{n_h} \quad = \text{var}\,(\bar{y})$$

9.4 OPTIMUM CHOICE OF BOOTSTRAP SAMPLE SIZES (m_h)

The optimum choice of bootstrap sample size (m_h) has been obtained using the rescaling factor of the pseudo value. The rescaling factor is

$$(1-f_h)^{1/2}\left(\frac{\sqrt{m_h}}{\sqrt{n_h}-\sqrt{m_h}}\right) \quad \text{and} \quad m_h = \frac{n_h}{\left(2-\frac{n_h}{N_h}\right)} = \frac{n_h}{2-f_h}$$

It can be seen that for optimum choice m_h, $\tilde{y}_{hi} = y^*_{hi}$ and the RS BWO method reduces to the naive bootstrap method, however, in the latter method's step (i) a simple random sample of size $(\frac{n_h}{2-f_h})$ is selected from $\{y_{hi}\}_{i=1}^{n_h}$ in stratum h.

Ahmad (1997) extended the RS BWO technique to two stage cluster sampling with equal probabilities and without replacement. The technique is justified by showing that in the linear case the variance estimator reduces to the usual variance estimator for the two stage sampling designs. It can be extended to other complex survey designs also, with appropriate modifications.

Ahmad (1997) has compared the RS BWO technique with the existing bootstrap techniques and jackknife method of variance estimation for stratified random sampling without replacement through a simulation study. It has been observed that the RS BWO technique performs better than the other similar bootstrap variance estimation techniques in terms of percentage relative bias and relative stability. Also its performance is almost comparable with the jackknife variance estimator.

REFERENCES

Ahmad, T. (1997). A resampling technique for complex survey data. *Journal of Indian Society of Agricultural Statistics*, **50(3)**, 364-379.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika, **46,** 477-480.*

Efron, B. (1979). Bootstrap methods: another look at jackknife. *Annals of Statistics*, **7**, 1-26.

Efron. B. (1982). Nonparametric estimates of standard error: the Jackknife, the bootstrap and other methods. *Biometrika*, **68**, 589-599.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science*, **1**, 54-77.

Gray, H.L. and Schucany, W.R. (1972). *The generalized jackknife statistics*. New York: Marcel Deckcer.

Gross, S.T. (1980). Median estimation in sample surveys. In: Proc. Sec. Sur. Res. Meth., 181-184. ASA Washington, D.C.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353-360.

Rao, J.N.K. and Wu, C.F.J. (1985). Inference from stratified samples; second order analysis of three methods for non-linear statistics. *Journal of American Statistical Association*, **80**, 620-630.

Rao, J.N.K. and Wu, C.F.J.(1988). Resampling inference with complex survey data. *Journal of American Statistical Association*, **63**, 231-241.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of American Statistical Association*, **27**, 755-765.

Wolter, K.M. (1985). *Introduction to variance estimation*. Springer-Verlag, New York.

APPLICATION OF MULTI-PHASE SAMPLING AND SUCCESSIVE SAMPLING IN SAMPLE SURVEYS

Prachi Mishra Sahoo

Indian Agricultural Statistics Research Institute, New Delhi-110012

10.1 MULTI-PHASE SAMPLING

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 first phase sample or phase I sample. Determinations for the vector **y** are made on a subsample of the original sample. The subsample is called the second phase sample or phase 2 sample. In the form originally suggested by Neyman (1938), the original sample was stratified on the basis of **x** and the stratified estimator for **y** constructed using the estimated stratum sizes estimated with the phase 1 sample. We first describe this particular, and important, case of two-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 auxiliary 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.

NOTATONS

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*;
$s^{2'}{}_x$	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 as randomized branch sampling).

10.2 SUCCESSIVE SAMPLING

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 $Cov(\bar{y},\bar{x})>0$.

On the contrary, for estimating the average of the means the latter would be better than the former in that

$$V(\bar{y}+\bar{x}) = V(\bar{y}) + V(\bar{x}) + 2Cov(\bar{y},\bar{x}) > V(\bar{y}) + V(\bar{x}),$$

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 (say S_f) of fresh units on the current occasion, and the data retaining to x on , x and y on, and y on S_c, S_f and S 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 these samples should be chosen and what procedure is 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 from by 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.

10.3 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 of n_{12} units from the sample taken at step (1) 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_2 on 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,

$$\mathrm{V}(t_c) = \frac{S_y^2}{n_1} + \frac{\sum_{j=1}^{N} P_j \left[\frac{y_j}{nP_j} - \bar{Y} \right]^2}{n_{12}} - \frac{S_y^2}{N}$$

Also in view of selection of S_f as noted in the step (2), the unbiased estimator of $\bar{Y}$ *is*,

$$\bar{y}_f = \frac{1}{n_{22}} \sum_{j=1}^{n_{22}} y_j \text{ with variance as,}$$

$$V\left(\bar{y}_f\right) = \left(\frac{1}{n_{22}} - \frac{1}{N} \right) S_y^2$$

Further, t_c and $\bar{y}_f$ are correlated so,

$$COV(t_c, \bar{y}_f) = -\frac{1}{N} S_y^2$$

So in this sampling on two successive occasion, the best minimum variance combination of t_c and $\bar{y}_f$ will be,

$$\bar{y}_{ss} = a t_c + (1-a)\bar{y}_f, \text{where}, a = \frac{V_f}{V_c + V_f}$$

with variance,

$$V(\bar{y}_{ss}) = \frac{V_c V_f}{V_c + V_f} + \mathrm{COV}(t_c, \bar{y}_f);$$

where, $V_f = V(\bar{y}_f) - \mathrm{COV}(t_c, \bar{y}_f)$ and $V_c = V(t_c) - \mathrm{COV}(t_c, \bar{y}_f)$

REFERENCES

Eckler, A. R. (1955). Rotation Sampling. *American Statistician*, **26**: 664-685.

Jessen, R. J. 91942). Statistical Investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin* No. 304.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of American Statistical Association*, **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, *Journal of American Statistical Association*, **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.

STRATIFIED AND MULTISTAGE SAMPLING IN AGRICULTURAL SURVEYS

Tauqueer Ahmad

Indian Agricultural Statistics Research Institute, New Delhi-110012

11.1. Stratified Sampling

11.1.1 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.

11.1.2 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.

11.2. Multistage Sampling

11.2.1 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.

11.2.2 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. Hansen, M.H. and Hurwitz, W.H. (1943b). On the theory of sampling from finite populations; *Ann. Math. Statist.*, **14**, 333-362.
5. 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.
6. Kish, L. (1965). Survey sampling; John Wiley & Sons, Inc.
7. Lahiri, D.B. (1954 a). Technical paper on some aspects of development of the sample design; National Sample Survey Report No. 5, Government of India, reprinted in *Sankhya*, **14**, 264-316.
8. Murthy, M.N., (1977). Sampling theory and methods; Statistical Publishing Society
9. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984). Sampling theory of surveys with applications; Indian Society of Agricultural Statistics.

CATEGORICAL DATA ANALYSIS IN COMPLEX SURVEYS

ANIL RAI

Indian Agricultural Statistics Research Institute, New Delhi-110012

12.1 INTRODUCTION

In large scale surveys the data on large number of quantitative as well as qualitative characters have been collected from the sampled units. These surveys can be broadly divided in two categories on the basis of the type of analysis of the data. First category of surveys are known as descriptive surveys in which main concern of the researchers is estimation of parametric function of the target population. Second category of the surveys are known as analytical surveys. In this category the main concern of the researchers lies in structural analysis of the population. The categorical data analysis falls in to second category, specially in case of analysis of qualitative variables of survey.

Standard statistical procedures were developed by assuming that the units in the sample are independently and identically distributed. These assumptions can be ensured when sample units are either drawn from an infinite population or selected with the help of simple random sampling with replacement (SRSWR) from a finite population. The independence of sample elements greatly facilitates in obtaining theoretical results of interest, because of mathematical simplicity, which becomes desirable in case of complex statistics such as correlation, regression etc. Independence is often assumed automatically and needlessly, even its relaxation would permit broader conclusions.

In case of survey sampling, independence of sample elements is typically assumed, although it is seldom realized, when data is collected with complex survey designs. This dependence of sample elements is mainly due to effect of clustering and stratification of the sampling units in the target population. In the past, literature of survey sampling is mainly concentrated on providing estimates of simple statistics like mean, proportions, total, ratio's etc. alongwith the estimates of their standard errors of the target population. Recent advances in the field of computer technology and easy as well as cheaper availability of computer hardware and software had changed the emphasis of survey research towards computer intensive techniques.

Some of the important emerging areas are categorical data analysis, regression analysis, variance estimation techniques etc.

There are mainly two types of problems encountered in categorical data analysis. First, is the measure of association, by which the degree of relationship between any two variables can be measured, second is, testing of hypotheses, under which various hypotheses of interest like, goodness-of-fit, independence of attributes and homogeneity of proportions can be tested. The categorical data analysis in survey sampling is mostly confined to testing of hypothesis because of it's practical utility to study the structure of the population under consideration and drawing inferences accordingly. For example, an investigator may wish to compare sample proportions of farmers growing different high yielding varieties of a crop, with known estimates of population proportions from the previous surveys in a district. This can also be used to check the quality of the survey. More generally, we might make comparisons of proportions obtained from different surveys of the same population or among the surveys of different regions or countries on similar lines.

Different methods of sampling in the case of categorical data can be broadly classified into three groups. In case of method I, total sample size is fixed and units are directly selected from the population.. The classification of the units in different cells are determined by the combination of attributes possessed by selected units. In method II, the marginal frequencies of one of the variables involved in classification is fixed and sample is selected accordingly. The classification on the basis of other variables becomes known latter. Lastly, in method III, the combination of attributes on which sample units are classified in to different cells are not inherent property of the units, but it is induced by the researchers.

Log-linear models, which expresses the logarithm of the expected frequencies for categorical responses as a linear function of unknown parameters, encompass both factorial models for cross-classified categorical data and logistic models for one or more dependent categorical or continuos predictors. Bishop, Fienberg and Holland (1975) provided one of the earliest books in this rapidly expanding field. Books by Agresti (1991), Christensen (1990) etc. also provides details regarding these models. In fact log-linear model analysis has rapidly become a major tool of statistical

practice for deciphering multi-dimensional contingency tables arising through product-multinomial, multinomial or poisson sampling. It is widely accepted that general multiplicative model provides a natural framework for exploratory examination and testing of various hypothesis of statistical independence among variables or related distributional homogeneities. Three general strategies for fitting log-linear models have been widely promulgated in forms suitable for use by researchers. First is iterative proportional fitting of hirarchical analysis of variance models, analogous to factorial class. Second, is weighted least squares fitting of asymptotic regression models to log-linear functions of observed cell frequencies. Third, is function minimization techniques of a more general nature i.e. Newton-Raphson or iterative weighted least-squares method.

12.2 REVIEW OF LITERATURE

Statistical methods for the analysis of cross-classified count data are used extensively by survey researchers. In particular, the Pearson chi-squared test of independence in a two-way contingency tables is probably best known and most often used test. Analysis of multi-way contingency tables are also quite common, largely due to development of hierarchical log-linear models and related logit models as well as associated methods for systematic testing of hypotheses, similar to analysis of variance for continuos data.

Cohen (1976) investigated a very special case of general problem of testing good-ness-of-fit when data are collected with the help of complex sampling design, particularly special case of cluster sampling under the model of constant design effect. The design effect as defined by Kish and *Frankel (1974)* is the ratio of the variance under the sampling design to variance, under simple random sampling with replacement. The most sustained work on tests of independence from complex samples has been carried out by *Nathan* (*1969, 1972, 1973, 1975*). He also reviewed the work of several authors, such as *Bhapkar and Koch (1968) and Chapman (1966).*

These authors concentrated their efforts on a statistics which is closely related to null hypothesis of testing independence of attributes. Several unbiased statistics have been proposed based on sample re-use techniques specially balanced repeated replication (B R R) method. Also, the variance-covariance estimates for different cells of two-way contingency tables for statistics were estimated with the help of same sample re-

use method. The test statistics proposed by the above authors are in the form of quadratic function of above statistics. Unfortunately, test statistics proposed by this technique behaves very badly with respect to its achieved significance levels due to high correlation between numerators and denominator of the proposed test statistics. The simulation results reported by *Nathan (1973)* are flawed, as pointed out by the author in his subsequent paper in 1975.

Fellegi (1980). proposed two test statistics after careful examination of the problem, of estimating the variance of non-linear statistics from complex samples, in the light of existing literature. First test statistic is based on Taylor's approximation of the statistics considered by *Nathan (1973)* etc. Second test statistics is based on the approach of eliminating the effect of complex survey design from usual Pearson's chi-square statistics, is to divide the statistic by average of cell design effects. *Rao and Scott (1979, 1981)* have shown that in case of complex sampling designs customary test procedures are not valid even asymptotically. It has been proved that Pearson's chi-square X^2 test statistic and liklihood ratio G^2 test statistic are asymptotically distributed as a weighted sum of independent chi-square random variables, each with one degree of freedom. The weights attached to each chi-square random variable are eigen values of design effect matrix, which is based on the concept of usual design effect matrix which is based on the concept of usual design effect of individual cells. Further, first order correction to the above statistics i.e. X^2 and G^2 have been proposed so that their first moments becomes equal to its degrees of freedom as in case of usual applications for multinomial sampling. Also, a second order correction based on *Satherthwaite approximation (1946)* has been applied to modify the X^2 and G^2 when data are obtained with the help of complex survey designs. *Rao and Scoll (1984, 1987)* extended these modifications for multi-dimensional contingency tables with the help of log-linear models. *Fay (1985)* on resampling techniques such as jackkfing and BRR.

Some of relatively less important work in the field of categorical data analysis in survey sampling are by *Shuster and Downing (1976)* who proposed methods for testing independence, quasi-independence and marginal symmetry in contingency tables derived for variety of sampling schemes. *Cowan and Binder (1978)* analysed the effect of two stage sampling design on the test of independence in a 2*2 tables.

Brier (1980) used Dirichlet multinomial distribution as a model for contingency tables generated by cluster sampling schemes. *Koch et al (1975)* discussed certain aspects of multi-variate analysis of the data from possibly complex survey designs in terms of large sample methodology involving weighted least squares algorithms for the computation of Wald statistics. *Holt et al (1980)* empirically studied survey design effect on test of goodness-of-fit, test of homogeneity and test of independence for British Economic Survey (BES) and General Health Survey (GHS) data. Similar, other important empirical studies in this regard are by *Rao and Hidiroglou (1981), Kumar and Rao (1984) , Fay (1984), Thomas and Rao (1984, 1985) Singh and Kumar (1986), Fay (1989) etc.*

12.3 MODIFIED CHI-SQUARE TEST STATISTICS

In this section first and second order correction to Pearson's chi-square test statistics in case of data obtained through complex sampling designs will be discussed briefly. These techniques will be discussed following the approach of log-linear models, even for two-way contingency tables as it provides general method of estimation and testing the null hypothesis under consideration. The cells in contingency tables are numbered lexicographically as i=1,2,..,I with corresponding finite population proportions μ_i's and their estimates $\hat{\mu}_i$. A log-linear model on the μ_i 's may be written as

$$\log \mu = \tilde{\mu}(\theta)\, 1 + X\theta \qquad (1)$$

where log μ is (I x 1) vector of log probabilities, X is a known I * r matrix of full rank $r \leq (\mathbf{I}-1)$ such that $\mathbf{X}'\, 1 = \mathbf{0}$, θ is an (r*1) vector of parameters. 0 is a null vector and $\tilde{\mu}(\theta)$ is the normalizing factor which ensures that $\sum \mu_{i.} = 1$. If r = I-1 in equation (1), the saturated log-linear model is obtained.

The maximum likelihood estimator (MLE) of $\mu(\theta) = \left[\mu_1(\theta), \ldots \mu_1(\theta)\right]'$ under multinomial sampling , denoted by $\mu(\hat{\theta})$ is obtained from the likelihood equations

$$\mathbf{X}'\ \mu(\hat{\theta}) = \mathrm{X}'\,\bar{\mathrm{y}} \qquad (2)$$

where $\bar{\mathbf{y}} = \left(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \ldots\ldots, \bar{\mathbf{y}}_{\mathbf{I}}\right)'$ $\bar{\mathbf{y}}_{\mathbf{i}}$, i = 1,2, ... ,I is the mean of y_{it} indicator variable which takes value 0-1 for i-th category associated with t-th unit of the sample. The

fitted proportions from (2) are obtained by any of the three techniques for fitting log-linear models. It is customary to obtain' "pseudo MLE" of $\underset{\sim}{\mu}(\theta)$ with the help of equation (2) by substituting survey estimator of $\hat{\mu} = [\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_I]$ for $\bar{\mathbf{y}}$ as appropriate likelihood function can not be defined for general complex sampling designs.

Now with the help of above" pseudo "MLE" of $\mu(\theta)$ the test statistics can be developed for testing goodness-of-fit, homogeneity of proportions and independence of attributes under their corresponding null hypothesis. The first and second order corrections can be applied to these test statistics in case of complex sampling designs. The testing of independence of attributes is most oftenly used test statistics by survey researchers. Hence, in the following paragraphs these test statistics are obtained with the help of nested models.

The **X** matrix of the equation (1) can be partitioned as $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})'$ similarly, $\theta = (\theta_1, \theta_2.)'$. The null hypothesis of testing independence of attributes is equivalent to putting $\theta_2 = \mathbf{0}$, *(Bishop et al 1975),* given log linear model (1) and r $\underline{<}$ (I-1), where as X_1 is I * r_2 of rank r_2 ($r_1 + r_2 = r$). The general Pearson chi-square statistic in this case is given by

$$X_P^2(\hat{1}) = n_0 \sum_{i=1}^{I} \left[\mu_i(\hat{\theta}) - \mu_i(\hat{\hat{\theta}}_1)\right]^2 \Big/ \mu_i(\hat{\hat{\theta}}_1) \qquad (3)$$

where, $\mu_i(\hat{\theta}_1)$ is "psuedo MLE" under reduced model, obtained from the "psuedo likelihood equations" $\mathbf{X}_1' \mu(\hat{\hat{\theta}}_1) = \mathbf{X}_1' \hat{\mu}$. *Rao and Scott* (*1984)* showed that $X_P^2(\hat{1})$ is distributed asymptotically. as sum $\sum_{i=1}^{r_2} \delta_i(\hat{1}) W_i$, where W_i is χ_1^2 r_2 - random variables and $\delta_i(\hat{1})$ is the i-th eigen value of the design effect matrix $\underset{\sim}{\hat{\Delta}}(\hat{1})$ given below.

$$\underset{\sim}{\hat{\Delta}}(\hat{1}) = \left[\tilde{\mathbf{X}}_2' \underline{\Omega}(\hat{\hat{\theta}}_1) \tilde{\mathbf{X}}_2\right]^{-1} \left[\tilde{\mathbf{X}}_2 \hat{\Sigma} \tilde{\mathbf{X}}_2\right] \qquad (4)$$

where, $\tilde{\mathbf{X}}_2' = \left[I - \mathbf{X}_1 \left\{ X_1' \; \Omega \left(\hat{\theta}_1 \right) \mathbf{X}_1 \right\}^{-1} \mathbf{X}_1' \; \Omega \left(\theta_1 \right) \right] \mathbf{X}_2$

$$\Omega \left(\tilde{\theta} \right) = n_0^{-1} \left\{ D \left[\mu \left(\tilde{\theta} \right) - \mu \left(\tilde{\theta} \right) \mu \left(\tilde{\theta} \right)' \right] \right\}$$

$\tilde{\theta}$ equals $\hat{\theta}$ or $\hat{\theta}_1$ as appropriate. The $\hat{\Sigma}$ is estimated variance-covariance matrix of cell proportions based on survey design. The first order corrections to the test statistics $X_P^2 \left(\nu_1 \right)$ is given by

$$X_{PF}^2 \left(\nu_1 \right) = X_P^2 \left(\nu_1 \right) / \hat{\delta} \left(\nu_1 \right) \tag{5}$$

where,

$$r_2 \hat{\delta} \left(\nu_1 \right) = T_r \left[\left\{ X' \; \Omega \left(\hat{\theta}_1 \right) \mathbf{X} \right\}^{-1} \left\{ \mathbf{X}' \; \hat{\Sigma} \; \mathbf{X} \right\} \right] - T_r \left[\left\{ X_1' \; \Omega \left(\hat{\theta}_1 \right) \mathbf{X}_1 \right\} \left\{ \mathbf{X}_1' \; \hat{\Sigma} \; \mathbf{X}_1 \right\} \right]$$

Here, Tr denotes the trace of the matrix. This correction, in general depends on the full estimated covariance matrix $\hat{\Sigma} = \text{Var}(\hat{\mu})$. However, if both $\mu \left(\theta \right)$ and $\mu \left(\theta_1 \right)$

can be expressed explicitly in terms of cell probabilities μ_i and their marginals, then $\hat{\delta} \left(\nu_1 \right)$ can be written in terms of deffs of cell estimates and of their marginals, Rao and Scott (1984). If an estimate of the full co-variance matrix $\hat{\Sigma} = \text{Var} \left(\hat{\mu} \right)$ is

available, a more accurate second order correction to X^2_P can be obtained by using the well known Satherthwaite approximation to the distribution of a weighted sum of independent χ_1^2 variables. In fact this correction takes account of variability in generalised design effects $\delta_i \left(\nu_1 \right)$, unlike the first order corrections. This is given by

$$X_{PS} \left(\hat{\delta}, \hat{\mathbf{a}} \right) = \mathbf{X}_{\mathbf{PF}}^{\mathbf{2}} \left(\nu_1 \right) / \left[\mathbf{1} + \hat{\mathbf{a}} \left(\nu_1 \right) \right] \tag{6}$$

Where this statistics follows χ_V^2 , chi-square random variable with v =

$\left(\mathbf{I}-\mathbf{r}-1\right)\Big/\left(1+\hat{\mathbf{a}}^2\left(\tfrac{2}{1}\right)\right)$ degrees of freedom. Here, $\hat{a}$ is the coefficient of variation of the eigen values $\delta_i\left(\tfrac{2}{1}\right)$ of the estimated design effect matrix $\hat{\Delta}$ i.e.

$$\hat{a}^2\left(\tfrac{2}{1}\right) = \{ \sum_{\mathbf{i}=1}^{\mathbf{I}-\mathbf{r}-1} \hat{\delta}_{\mathbf{i}}^2\left(\tfrac{2}{1}\right) / \left(\mathrm{I}-\mathrm{r}-1\right) \hat{\delta}^2 \} - 1$$

where, (I-r-1) $\hat{\delta}\ (\tfrac{2}{1})$ = Tr $\hat{\Delta}$

and $\sum_{i=1}^{I-r-1} \hat{\delta}_i^2\left(\tfrac{2}{1}\right)$ = Tr $\hat{\Delta}^2$

It is expected that second order corrections should control type-I error much better than first order correction. More details regarding chi-square tests in case of data from complex survey designs is given in a book by *Skinner et al (1989)*

12.4 SIMULATION

A small simulation study is discussed to illustrate the extend of misleading inferences which can be drawn by applying ordinary Chi-square test when data is collected with help of complex sampling design. In this simulation study a multivariate normal population of 10-variables is generated with the help of data taken from 1981 census of Faridkot district of Punjab state. This population of size 5000 is divided in to five equal strata each of size 1000, after sorting the data on the basis of stratifying variable. The categorical variables in this population are created in such a way that the cells of contingency tables formed with the help of these variables consists of equal number of observations in each cell. This ensure that the different variable used in these tables are independent with each others in the population. Now, 100 samples of different sizes (say, 50, 75, 100) are selected by SRSWOR from different strata independently, the sample sizes of each strata are given in the table-1. This type of sampling leads to the selection of the sample with properties of various common sampling designs used in practice. Now, the Chi-square test of independence of attributes is applied to test the null hypothesis of independence of attributes with the help of each of the selected samples. The table-2 gives the number of samples for

which this test is found to be significant for chi-square statistics as well as other modified chi-square test statistics. It can be seen that extent of wrong inference drawn is more in case of non-self weighting designs and increases as the skewness of the design increases. The second order correction X^2_{PS} control the effect of sampling designs more effectively as compared to X^2_{PF} i.e. first order correction.

Table - 1

Allocation of sample size in to different strata for various sampling design

Design	n=50					n=75					n=100				
Stratum	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
D-1	10	10	10	10	10	15	15	15	15	15	20	20	20	20	20
D-2	6	8	10	12	14	9	12	15	18	21	12	16	20	24	28
D-3	2	8	10	12	18	3	12	15	18	27	4	16	20	24	36
D-4	2	3	5	15	25	3	4	8	22	38	4	6	10	30	50
D-5	1	1	2	8	38	2	2	3	12	56	2	2	4	16	76
D-6	15	7	5	8	15	22	11	8	12	22	30	14	10	16	30
D-7	23	2	1	2	22	34	3	2	3	33	46	4	2	4	44
D-8	24	1	1	1	23	35	2	2	2	34	48	2	2	2	46

Table - 2

Number of samples for which X_P^2 X_{PF}^2 X_{PS}^2 are significant

(X_3 * X_4 table, 5% level of significance, m=100)

Design	n=50			n=75			n=100		
	X_P^2	X_{PF}^2	X_{PS}^2	X_P^2	X_{PF}^2	X_{PS}^2	X_P^2	X_{PF}^2	X_{PS}^2
1	2	3	4	5	6	7	8	9	10
D-1	-	-	-	-	-	-	-	-	-
D-2	6	6	4	5	5	2	4	4	3
D-3	21	6	1	14	3	2	10	2	2
D-4	30	7	1	23	3	1	31	1	-
D-5	52	11	2	46	7	2	63	4	1
D-6	8	5	2	10	4	2	4	2	2
D-7	72	15	5	73	13	5	75	10	3
D-8	79	12	4	76	10	4	77	10	5

IMPORTANT REFERENCES

Agresti, A.C.(1990). Categorical data analysis. *Wiley John & Sons, New York.*

Bhapkar, V.P. and Koch. C.G.(1968). Hypothesis of no interaction in multidimensional contingency tables. *Technometrics,* 10, 107-123.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W.(1975). Discrete multivariate analysis; Theory and practice. *Cambridge: Massachusetts Institute of Technology Press.*

Chapman, D.W.(1966). An approximate test of independence based on replication of complex survey design. *Unpublished master's thesis, Cornell University.*

Fay , R. E. (1985) . A Jackknife chi-square tests for complex samples. *J. Am. Statist Assoc.* , 80, 148-157.

Fay , R. E. (1989). Additional evaluation of chi-square methods for complex samples. *Proc. Am. Statist. Assoc. Sur. Meth. Sec.* ,680-685.

Fellegi , I . P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *J. Am. Statist. Assoc*., 75, 261-268.

Kumar , S. and Rao, J. N. K. (1984) . Logistic regression analysis of Labour force survey data . *Sur. Meth.,* 10(1), 62-81.

Nathan , G. (1969) . Test of independence in contingency tables from stratified samples. In. N. L. Johnson and H. Smith eds. New Developments in Survey Sampling. *Wiley, New York,* 578-600.

Nathan , G. (1972). On the asymptotic power of tests for independence in contingency tables from stratified samples. *J. Am. Statist. Assoc.,* 6, 917-920.

Nathan , G. (1973) . Approximate tests of independence in contingency tables from stratified samples. *N. C. H. S. , Vital and Health Statistics. Series-2, No. 53, Washington D.C.*

Nathan , G. (1975) . Test of independence in contingency tables from stratified proportional samples. *Sankhya* , 37, (C) Part I , 77-87.

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

Rao , J. N.K. and Scott , A. J. (1984). on chi-square tests for multiway contingency tables with proportions estimated from survey data. *Ann. Statist.,* 12 , 46-60.

Skinner , C. J. , Holt , D. and Smith , T. M. F. (1989). Analysis of complex surveys. *John Wiley & sons , New York.*

Thomas , D. R. and Rao, J. N. K. (1984). A monte carlo study of exect levels of chi-squared goodness-of-fit statistics under cluster sampling. *Technical report , 66 ,Carlton univ., Ottawa.*

Thomas, D. R. and Rao, J. N. K. (1985). On the power of some goodness-of -fit tests under cluster sampling. *Technical report , 66, Carlton univ., Ottawa.*

Thomas, D.R. and Rao, J.N.K.(1987). Small sample comparisons of level and power for simple goodness of fit statistics under cluster sampling. *J.Am. Statist.Assoc.,* 82, *630-636.*

Thomas, D.R., Singh, A.C. and Roberts, G.(1989). Size and power of independence tests for R*C tables from Complex surveys. *Proc. Am. Statist. Assoc., 763-768.*

IMPUTATION TECHNIQUES IN SAMPLE SURVEYS

Tauqueer Ahmad

Indian Agricultural Statistics Research Institute, New Delhi-110012

13.1 INTRODUCTION

Modern complex surveys typically collect responses to a large number of items for each sampled element. The problem of missing data occurs when some or all of the responses are not collected for a sampled element or when some responses are deleted because they fail to satisfy edit constraints. It is desirable to distinguish between total (or unit) nonresponse and item nonresponse due to procedural differences in their adjustment methods.

Unit (or total) nonresponse occurs when no information is collected from a sample unit. It may be caused by a refusal, by a failure to contact the unit (not at home), by the inability of the unit to cooperate (perhaps because of illness or a language barrier), by the unit not being found (for instance, movers in a panel survey), or by completed questionnaires being lost.

Item nonresponse occurs when the unit cooperates in the survey but fails to provide answers to some of the questions. It may arise because of the following reasons.

i) The informant lacks the information necessary to answer the question. It includes his failure to make the effort required to ascertain the information by retrieving it from his memory or by consulting his records. This is usually described as “Don’t know”.

ii) The informant refuses to give an answer, perhaps on the grounds that he finds the question sensitive, embarrassing, or consider it irrelevant to his perception of the survey’s objectives.

iii) The interviewer fails to record the answer.

iv) The response is subsequently rejected at an edit check on the grounds that it is inconsistent with other responses.

With regard to compensation for nonresponse, the importance of the distinction between unit or item nonresponse resides in the amount of information available about the non-respondents. In general the only information available about total non-respondents is that on the sampling frame from which the sample was selected (e.g., the strata and PSUs in which they are located). The important aspects of this information can usually be readily incorporated into weighting adjustments that attempt to compensate for the missing data. Hence, as a rule weighting adjustments are used for total nonresponse.

In the case of item nonresponse, however, a great deal of additional information is available for the elements involved, i.e., not only the information from the sampling frame, but also their responses for other survey items. In order to retain all survey responses for elements with some item nonresponse, the usual adjustment procedure

produces analysis records that incorporate the actual responses to items for which the answers were acceptable and imputed responses for other items. Imputation, i.e., filling in for missing values is a very common technique for handling nonresponse. Imputation is done to reduce non-response bias for a variable, which occurs because the distribution of the missing values (if known) may be different from the distribution of responses. This is achieved by using relationships between the item to be imputed and other variables – but if this relationship is incorrectly modeled it could make matters worse. There are several benefits to imputation procedures.

They are as follows:

i) Imputation aims to reduce biases in survey estimates arising from missing data.

ii) Imputation makes analysis easier and the results simpler to present. Complex algorithms to estimate population parameters in the presence of missing data are not required, and hence much processing time is saved. There is neither a need to determine the different sets of cases with missing data that have to be deleted from different analyses, nor a need to provide details of the extent and treatment of missing data with each set of results.

iii) The results obtained from different analyses are bound to be consistent with one another, a feature which need not apply to results of analyses from an incomplete data set.

On the other hand, imputation of missing data does, however, have its drawbacks. It is not certain that the results obtained after imputation will be less biased than those based on the incomplete data set. It all depends on the imputation procedure and the form of estimate. Another risk with imputation is that researchers may falsely treat the completed data set as if all the data were actual responses, thereby overstating the precision of the survey estimates. Therefore, the researchers working with a data set containing imputed values should proceed with caution and be aware of the extent of imputation for the variables in their analysis as well as the details of the procedure used. The imputed values should be flagged, so that the careful researcher can assess the effect that imputation may have on the analysis. The aim of this lecture is to provide a brief overview of traditional as well as modern computer intensive techniques of imputation for handling missing survey data.

13.2 TRADITIONAL METHODS OF IMPUTATION

An imputation procedure is defined as a procedure that imputes a value for each missing value that is assumed to be quite close to the true missing value. A wide variety of imputation methods has been developed for assigning values for missing item responses (Kalton and Kasprzyk, 1986). Imputation technique may be quite useful when imputation for any missing value is done based on homogeneous imputation classes.

Deductive imputation

Sometimes the missing answer to an item can be deduced with certainty from the pattern of responses to other items. Edit checks should check for consistency between responses to related items. When the edit checks constrain a missing response to only one possible value, deductive imputation can be employed. Deductive imputation is the ideal form of imputation.

Mean imputation

Missing values are replaced by the mean of all responding values for the variable. This can be done based on the whole dataset or separately for different categories of respondents defined by combinations of selected classification variables.

Zero imputation

It is a method of imputation in which zero is substituted for the missing data when a unit fails to respond.

Regression imputation: This method uses respondent data to regress the variable for which imputations are required on a set of auxiliary variables. The regression equation is then used to predict the values for the missing responses. The imputed value may either be the predicted value or the predicted value plus some residual (stochastic regression imputation). There are several ways in which the residual may be obtained.

Cold-deck imputation

Missing values are replaced by values of older data, e.g. from a previous survey, which could furthermore be adjusted for trend.

Hot-deck imputation: In general, a hot-deck procedure is a duplication process - when a value is missing from a sample, a reported value is duplicated to represent this missing value. The adjective “hot” refers to imputing with values from the current sample. This procedure usually has some classification process associated with it. All of the sample units are classified into disjoint groups so that the units are as homogeneous as possible within each group. For each missing value, a reported value is imputed which is in the same classification group. Thus, the assumption is made that within each classification group the non-respondents follow the same distribution as the respondents. Current survey practice uses many variations of hot-deck procedures.

A sequential hot-deck procedure is one in which the sample is put in some type of order within each classification group, and for each missing value the previous reported value is duplicated. For example, the ordering might be based on a geographic variable. The result of a geographic ordering is that the reported value duplicated for a missing value is from a unit which is geographically close to the unit with the missing value. The sequential hot-deck suffers the disadvantage that it may easily make multiple uses of donors, a feature that leads to a loss of precision in survey estimates.

The above disadvantages of the sequential hot-deck are avoided in the hierarchical hot-deck method. The procedure sorts respondents and non-respondents into a large number of imputation classes from a detailed categorization of a sizeable set of auxiliary variables. Non-respondents are then matched with respondents on a hierarchical basis, in the sense that if a match cannot be made in the initial imputation class, classes are collapsed and the match is made at a lower level of detail.

Another form of hot-deck method is distance function matching which assigns a non-respondent the value of the 'nearest' respondent, where 'nearest' is defined in terms of a distance function for the auxiliary variable. Various forms of distance function have been proposed and the function can be constructed to reduce the multiple uses of donors by incorporating a penalty for each use.

Composite imputation

This method combines ideas from different methods. For example, hot deck and regression imputation can be combined by calculating predicted means from a regression but then adding a residual randomly chosen from the empirical residuals to the predicted value when forming values for imputation.

Multiple imputation

Because many imputation methods often do not preserve distributional properties, multiple imputation is advocated as a way of improving the ability to make inferences from data where imputation has been undertaken, particularly when the proportion of values missing is high. Multiple imputation retains the advantages of single imputation like completing the data set and using the expert knowledge for imputation and rectifies its major disadvantages Rubin (1986). As its name suggests, multiple imputation replaces each missing value by a vector composed of $M \geq 2$ possible values. The M values are ordered in the sense that the first components of the vectors for the missing values are used to create one complete data set, the second components of the vectors are used to create the second completed data set and so on. There are some practical difficulties with multiple imputation as there is generally a desire to produce one definitive micro data set for public use rather than a several which will give slightly different results and the typical data user may not be willing to analyze several datasets in order to obtain each answer.

13.3 IMPUTATION METHODS FOR LONGITUDINAL SURVEYS

In longitudinal surveys imputations may be carried out within each wave separately i.e. considering each time stage (wave) as a complete survey or using together multiple observations for different time stages for each unit. The imputation procedures of first type are called cross-sectional. They will be necessary if the results are to be published just after the completion of each wave. Some of the widely used imputation procedures have been studied in the previous section.

Imputation procedures of second kind i.e. involving multiple observations on each unit for imputing missing values are called cross-wave imputation methods. These methods make use of larger information and hence are expected to perform better than the cross-sectional methods. However a major disadvantage of these methods is that

one has to wait for the completion of all the waves before taking up the analysis of data. In this section some cross-wave imputation methods are discussed which are expected to be useful for imputing the missing values for any time stage.

Direct Substitution Imputation

Some variables show high stability over time between successive time stages of a survey. Under such situations the direct substitution of target variable from the previous (nearest) wave may serve well for a missing value on another wave. For example, in case of fodder yield surveys, if the data are collected weekly/fortnightly, there may not be much variation in yields between the two waves when the crop is fully mature. This method of imputation will be like deductive imputation assuming no change over time.

Average of Preceding Wave and Succeeding Wave Imputation

The basic assumption in the direct substitution imputation method discussed above is that the value of the character under study is stable over time. But in actual practice in most of the longitudinal surveys there will be some changes taking place in the value of the character under study over time. For example, milk yield of an animal, yield of vegetables etc., increase/decrease during the initial/final time stages respectively. Therefore a simple and more appropriate method of imputation in such situations will be to assign the average value of the preceding wave and succeeding wave for the missing wave value.

13.4 NEW METHODS OF IMPUTATION

Recent advances in methods and computing capabilities have made possible the application of more complex statistical modeling techniques like non-parametric regression; neural networks including multi layer perceptron, self organizing maps, support vector machines etc. for the purpose of imputation

Some important Software for Imputation

A number of computer programs have been developed for editing and imputation using traditional approach. The Canadian Census Edit and Imputation System (CANCEIS) has been developed to perform editing and imputation for the Canadian Census. The GEIS software also developed by Statistics Canada, implements methods for data editing and imputation where variables are numerical, continuous and non-negative, and edits (consistency rules) can be expressed in linear form. SOLAS 3.0 is a statistical package that allows both Random Hot Deck imputation and Group Mean imputation. Besides specialized software for imputation, general statistical packages such as SAS and SPSS also offer some methods for imputation. For example, the PROC MIXED in SAS can handle the longitudinal data with missing observations.

13.5 CONCLUSION

In real applications, a number of different imputation methods should be combined, especially when a data source has several different types of variables. Methods of imputation used for a dataset should always be well documented for the end user of data. Imputed values should always be flagged or there should be two versions of the

variable, one with imputed values and one with original values, so that it is possible to use either of the values and compare the results. Further, modern computer intensive imputation methods based on neural networks are very promising tool for helping to handle non-response in surveys. Although, much works remains to be done before these will become commonplace methods.

REFERENCES

Austin, J. and Lees, K. (2000). A Search Engine Based on Neural Correlation Matrix Memories, *Neurocomputing*, 35, 55 – 72.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of American Statistical Association*, 71, 17-35.

Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. Proceedings of the section on survey research methods. *American Statistical Association*, 22-31.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Kohonen, T. (1997). *Self-Organizing Maps*. Springer, Berlin, Heidelberg.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Nordbotten, S. (1963). *Automatic Editing of Individual Statistical Observations, Statistical Standards and Studies, No.*2. UN Statistical Commission and Economic Commission of Europe, New York.

Nordbotten, S. (1995). Editing Statistical Records by Neural Networks, *Journal of Official Statistics*, 11, 391-411.

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data, *Journal of Official Statistics*, 12, 385-401.

Rao, J.N.K. (1999). Some Current Trends in Sample Survey and Methods, *Sankhya*, 61, Series B, 1-57.

Roddick, L.H. (1993). Data Editing Using Neural Networks. Technical Report, Systems Development Division, Statistics Canada.

Rubin, D.B.(1986). *Multiple Imputation for Non response in Surveys*. New York. John Wiley & Sons.

Vapnik, V.N. (1996). *Structure of statistical learning theory. In Computational Learning and Probabilistic Reasoning (editor A. Gammerman).* New York, Wiley.

ADAPTIVE CLUSTER SAMPLING FOR RARE EVENTS

Ankur Biswas

Indian Agricultural Statistics Research Institute (IASRI)-ICAR, New Delhi-12.

14.1 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.

14.2 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

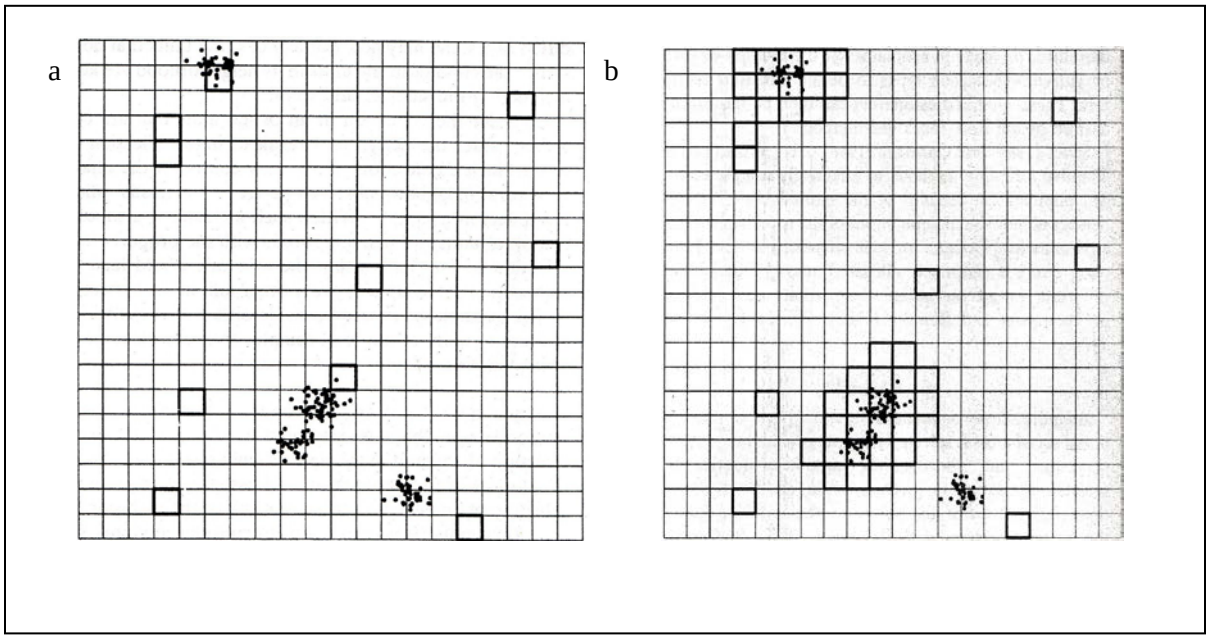


Figure 1. Adaptive cluster sampling to estimate the number of point-objects in a study region of 400 units. An initial random sample of 10 units is shown in (a). Adjacent neighbouring units are added to the sample whenever one or more of the objects of the population are observed in a selected unit. 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.

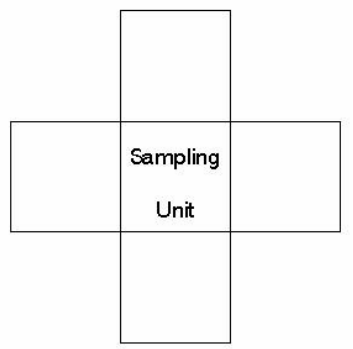


Figure 2. Neighbourhood for a sampling unit in the strawberry field study.

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

$$\alpha_i = 1 - \binom{N - m_i - a_i}{n_1} \Big/ \binom{N}{n_1} \qquad \ldots(2.1)$$

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

$$\alpha_i = 1 - (1 - p_i)^{n_1} \qquad \ldots(2.2)$$

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.

14.3 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 ***y***.

14.3.1 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.

14.3.2 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.

14.3.3 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

$$\alpha_k^* = 1 - \binom{N - m_k}{n_1} \Big/ \binom{N}{n_1} \qquad \ldots(3.3.1)$$

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

$$t_{HT^*} = \frac{1}{N} \sum_{k=1}^{\nu} y_k J_k / \alpha_k^*, \qquad \ldots(3.3.2)$$

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

$$\pi_{jh} = 1 - \left\{ \binom{N - m_j}{n_1} + \binom{N - m_h}{n_1} - \binom{N - m_j - m_h}{n_1} \right\} \Big/ \binom{N}{n_1}, \qquad \ldots(3.3.3)$$

when the initial sample is selected without replacement and

$$\pi_{jh} = 1 - \left[\left\{ 1 - m_j / N \right\}^{n_1} + \left\{ 1 - m_h / N \right\}^{n_1} - \left\{ 1 - \left(m_j + m_h \right) / N \right\}^{n_1} \right] \qquad \ldots(3.3.4)$$

when the initial sample is selected with replacement.

With the convention that $\pi_{jj} = \pi_j$, the variance of the estimator t_{HT*} is

$$\mathrm{Var}(t_{HT*}) = \frac{1}{N^2} \sum_{j=1}^{\varsigma} \sum_{h=1}^{\varsigma} y_j y_h \left(\pi_{jh} - \pi_j \pi_h \right) \Big/ \left(\pi_j \pi_h \right) \qquad \ldots(3.3.5)$$

An unbiased estimator of the variance of t_{HT*} is

$$\hat{\mathrm{V}}\mathrm{ar}(t_{HT*}) = \frac{1}{N^2} \sum_{k=1}^{\kappa} \sum_{m=1}^{\kappa} y_k y_m \left(\pi_{km} - \pi_k \pi_m \right) \Big/ \left(\pi_k \pi_m \pi_{km} \right), \qquad \ldots(3.3.6)$$

where the summation is over the κ distinct networks represented in the initial sample.

14.4 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.

14.5 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.

SURVEY FOR ESTIMATION OF AREA AND PRODUCTION OF FLORICULTURE CROPS AND MUSHROOMS

A. K. Gupta

Indian Agricultural Statistics Research Institute, New Delhi-110012

15.1 Estimation of area and production of flowers

In order to strengthen the existing database pertaining to flowers, the National Statistical Commission recommended the need to develop a cost-effective suitable sampling methodology for estimation of area and production of important flowers on the basis of market arrivals. Accordingly, a pilot study entitled "Pilot sample survey to develop sampling methodology for estimation of area, production and productivity of important flowers on the basis of market arrivals", was planned and conducted in Delhi State during September 2003 to August 2004 in which estimates of production of flowers were developed by considering two approaches; (i) market arrival and (ii) village survey.

15.1.1Methods and Material

i) Market Survey Approach

There are three flower mandis in Delhi namely, Hanuman Mandir Mandi, Khari Baoli Mandi and Mahrauli Mandi. Cut flowers of Rose, Gladiolus, Chrysanthemum, Tube-rose and Carnation etc. are mainly traded in the Hanuman Mandir Mandi and trading of loose flowers of Marigold, Rose, Margaret and Jaffrey etc. is carried out in Fatehpuri and Mahrauli Mandi by the commission agents and self-selling farmers. The commission agents and self-selling farmers were selected as per the following sampling design for collection of data on varieties of flowers sold in the three mandis.

Sampling Design: A stratified random sampling design was followed in each flower mandi. Commission agents comprise the first stratum while the self-selling farmers the second stratum. Within the first stratum, seven random groups of commission agents were formed to cover all the commission agents trading in each Hanuman Mandir Mandi and Khari Baoli Mandi. A suitable number of self-selling farmers were selected for the

purpose of making inquiry about flowers sold by them. The survey period of one year i.e. September 2003 - August 2004 was divided into three sub-periods viz. Period-1: September - December 2003, Period-2: January - April 2004 and Period-3: May - August 2004. A self-selling farmer once chosen was not repeated in the particular period of inquiry. One random group of commission agents was randomly selected and observed for a fortnight in one sub-period for data collection purpose. The remaining six random groups were observed in a similar manner. All the seven groups were observed in seven fortnights in one sub-period. This process was repeated in the other two sub-periods also. All the commission agents of Mahrauli mandi were observed for 8 days in each period.

ii) Village Survey Approach

Out of a total of 228 rural villages, there were about 92 flower growing villages in Delhi out of which a sample of 15% villages was randomly selected as per the following sampling design.

Sampling Design: The sampling design adopted for estimation of area under floriculture in flower growing villages of Delhi was one of stratified uni-stage random sampling with villages as the sampling units. For estimating production of important flowers on the basis of village survey, the sampling design was stratified two stage random sampling with villages as first stage sampling unit and farmers growing flowers as the second stage sampling unit. All the flower growing villages of Delhi were divided into three strata in each of the sub-periods as follows; Stratum I: Villages having area up to 5 ha under flower, Stratum II: Villages having area more than 5 ha and less than 10 ha and Stratum III: villages having area more than 10 ha under floriculture. Out of 92 flower growing villages in Delhi, a random sample of 15 (15%) flower growing villages was selected. Accordingly, in each of the three strata, 3, 7 and 5 villages were selected in Period-1; 5, 4 and 6 villages in Period-2; and 5, 3 and 7 villages in Period-3 respectively for the purpose of estimation of area and production under different flowers. The area estimates were obtained by complete enumeration of selected villages. Villages having 15 or less than 15 farmers were completely enumerated for compilation of production figures. The production estimates for villages having more than 15 farmers were made on the basis of a random sample of 15 farmers selected in such a way that there was an appropriate

selection of each kind of flower grown in the selected villages. The sampling units at both the stages were selected by simple random sampling without replacement.

15.1.2 Estimation Procedure

i) Market Survey Approach

The estimation procedure comprises the following steps:

Estimation of total market arrival in Delhi for a particular kind of flower in the i^{th} period

Let there be various kinds of flowers grown in Delhi. The entire survey period has been divided into 3 sub-periods viz. Period-1 of 122 days, Period-2 of 121 days and Period-3 of 123 days; two Strata viz. Stratum-1 of Commission Agents/Mashakhors and Stratum-2 of Self-selling Farmers and seven groups (Each group comprises suitable number of Commission Agents/Mashakhors from first stratum and self-selling farmers from the second stratum) were formed.

Let y_{ghad} be the market arrival of the flower for g^{th} group, h^{th} stratum, a^{th} commission agent/ self-selling farmer on d^{th} day, Mean market arrival per day for g^{th} group, h^{th} stratum, a^{th} commission agent/self-selling farmer is given by

$$\overline{y}_{gha} = \sum_{d}^{m_h} \frac{y_{ghad}}{m_h}$$

where m_h is the number of days for which selected sample of commission agents/self-selling farmers was observed.

Estimated market arrival per day per commission agent/self-selling farmer in the h^{th} stratum of g^{th} group is given by

$$\overline{y}_{gh.} = \frac{1}{n_h} \sum_{a}^{n_h} \overline{y}_{gha}$$

The variance of $\overline{y}_{gh.}$ is given by

$$V(\overline{y}_{gh.}) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \frac{1}{N_h - 1} \sum_{a}^{N_h} \left(\overline{y}_{gha} - \overline{Y}_{gh.}\right)^2$$

where $$\bar{Y}_{gh.} = \frac{1}{N_h}\sum_{a}^{N_h} \bar{y}_{gha}$$

and the estimate of V($\bar{y}_{gh.}$) is given by

$$\hat{V}(\bar{y}_{gh.}) = (\frac{1}{n_h} - \frac{1}{N_h})(\frac{1}{n_h - 1})\sum_{a}^{n_h}\left(\bar{y}_{gha} - \bar{y}_{gh.}\right)^2$$

where N_h is the total number of agents in the h^{th} stratum and n_h is the number of agents observed in the h^{th} stratum of the g^{th} group.

Estimated total market arrival per day for the mandi on the basis of g^{th} group is given by

$$\hat{Y}_{g..} = \sum_{h} N_h \bar{y}_{gh.}$$

The variance of $\hat{Y}_{g..}$ is given by

$$V(\hat{Y}_{g..}) = \sum_{h} N_h^2 V(\bar{y}_{gh.})$$

and the estimate of $V(\hat{Y}_{g..})$ is given by

$$\hat{V}(\hat{Y}_{g..}) = \sum_{h} N_h^2 \hat{V}(\bar{y}_{gh.}) .$$

Estimated total market arrival per day for the mandi averaged over all groups for the i^{th} period is

$$\hat{Y}...(i) = \frac{1}{g}\sum_{g} \hat{Y}_{g..}$$

The variance of $\hat{Y}...(i)$ is given by

$$V\left(\hat{Y}_{...(i)}\right) = \frac{1}{g^2}\sum_{g} V\left(\hat{Y}_{g..}\right)$$

and the estimate of $V\left(\hat{Y}_{...(i)}\right)$ is given by

$$\hat{V}\left(\hat{Y}_{...(i)}\right)=\frac{1}{g^2}\sum_{g}\hat{V}\left(\hat{Y}_{g..}\right)$$

Estimator of total market arrival for the entire year (366 days) for the specified kind of flower is given by

$$\hat{\mathrm{Y}}....=122\,\hat{Y}...{}_{(1)}+121\,\hat{\mathrm{Y}}_{...(2)}+\ 123\,\hat{\mathrm{Y}}_{...(3)}$$

The variance of $\hat{Y}....$ is given by

$$V(\hat{Y}....)=(122)^2V(\hat{Y}_{...(1)})+(121)^2\,V(\hat{Y}...{}_{(2)})+(123)^2V(\hat{Y}_{...(3)})\,.$$

and the estimator of variance of the estimate of total market arrival for the year is given by

$$\hat{V}(\hat{Y}....)=(122)^2\hat{V}(\hat{Y}_{...(1)})+(121)^2\,\hat{V}(\hat{Y}...{}_{(2)})+(123)^2\hat{V}(\hat{Y}_{...(3)})\,.$$

Estimates of total market arrivals for all kind of flowers grown in Delhi on the basis of market arrival were obtained on the similar lines.

ii) Village Survey Approach

The estimation procedure comprises the following steps:

Estimation of total production for a particular kind of flower for ith period in the villages of Delhi

Let y_{hij} be the production of a particular kind of flower for the jth farmer of the ith village in the hth stratum (j = 1, 2, …, M_{hi} ; i = 1, 2, …, N_h ; h = 1, 2, 3). Let M_{hi} be the number of farmers in the ith village of the hth stratum and N_h be the total number of villages in the hth stratum. The average production of a particular kind of flower in the ith village of the hth stratum is given by

$$\bar{y}_{hi}=\frac{1}{m_i}\sum_{j=1}^{m_i}y_{hij}\ ,$$

where m_i is the number of selected farmers in the ith village.

Average production of a particular kind of flower in the hth stratum is given by

$$\bar{y}_h\ =\ \frac{1}{n_h}\sum_{i=1}^{n_h}M_{hi}\ \bar{y}_{hi}\,,$$

where n_h denotes the number of villages selected in the h^{th} stratum.

Accordingly, an estimator of the total production of a particular kind of flower is given by $\hat{Y} = \frac{N_1}{n_1}\sum_{i=1}^{n_1} y'_{1i} + \sum_{h=2}^{3}\frac{N_h}{n_h}\sum_{i=1}^{n_h} M_{hi}\,\bar{y}_{hi}$

The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = N_1^2\left(\frac{1}{n_1} - \frac{1}{N_1}\right) S_{b1}^2 + \sum_{h=2}^{3} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)S_{bh}^2 + \sum_{h=2}^{3}\frac{N_h}{n_h}\sum_{i=1}^{N_h} M_{hi}^2\left(\frac{1}{m_i} - \frac{1}{M_{hi}}\right)S_{hi}^2$$

where $S_{b1}^2 = \frac{1}{(N_1-1)}\sum_{i=1}^{N_1}(y'_{1i} - \bar{Y}_1)^2$, $S_{bh}^2 = \frac{1}{(N_h - 1)}\sum_{i=1}^{N_h}\left(M_{hi}\bar{Y}_{hi} - \bar{Y}_h\right)^2$ and

$S_{hi}^2 = \frac{1}{(M_i - 1)}\sum_{j=1}^{M_i}(y_{hij} - \bar{Y}_{hi})^2$.

The estimator of the $V(\hat{Y})$ is given by

$$\hat{V}(\hat{Y}) = N_1^2\left(\frac{1}{n_1} - \frac{1}{N_1}\right) s_{b1}^2 + \sum_{h=2}^{3} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)s_{bh}^2 + \sum_{h=2}^{3}\frac{N_h}{n_h}\sum_{i=1}^{n_h} M_{hi}^2\left(\frac{1}{m_i} - \frac{1}{M_{hi}}\right)s_{hi}^2$$

where $s_{b1}^2 = \frac{1}{(n_1-1)}\sum_{i=1}^{n_1}(y'_{1i} - \bar{y}'_1)^2$, y'_{1i} be the produce reported from i^{th} farmer in the 1^{st} stratum and $\bar{y}'_1$, the average produce from the 1^{st} stratum is given by

$$\bar{y}'_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y'_{1i}$$

and $s_{bh}^2 = \frac{1}{(n_h - 1)}\sum_{i=1}^{n_h}\left(M_{hi}\bar{y}_{hi} - \bar{y}_h\right)^2$, $s_{hi}^2 = \frac{1}{(m_i-1)}\sum_{j=1}^{m}(y_{hij} - \bar{y}_{hi})^2$.

15.1.3 Results and Discussion

Estimates of arrival of flowers on the basis of Market Survey Approach

Estimates of the total market arrival of loose flowers in Metric Tonnes as well as of cut flowers in lakh numbers along with their percentage standard errors in the three flower mandi of Delhi are presented in Table 1. A close perusal of Table 1 reveals that the estimate of the total market arrival of loose flowers from the villages of Delhi in the

flower mandis was 14570.910 MT with 2.51% standard error (S E). The corresponding figures for cut flowers was 670.68820 lakhs with 1.53% S E.

Table 1: Estimate of different kind of flowers in Delhi on the basis of market arrivals

Flowers	Loose (MT)	Cut (Lakh Numbers)
	Estimate	Estimate
Rose	1896.550 (1.02)	571.46046 (1.74)
Marigold	1727.171 (3.62)	-
Guldawari	33.651 (*)	26.93842 (7.75)
Rajnigandha	13.520 (*)	10.81733 (6.11)
Jaffrey	8897.820 (3.85)	-
Margaret (White/Yellow)	1899.035 (5.69)	-
Gladiolus	-	14.05386 (3.77)
Gerbera	-	-
Orchid	-	-
Carnation	-	-
Tube Rose (Double)	-	2.76533 (*)
Others	103.161 (2.61)	44.65279 (2.63)
Total production	**14570.910 (2.51)**	**670.68820 (1.53)**

Note: *Figures in parentheses indicate corresponding percent standard errors.*
** Estimates are based on very few observations.*

15.1.4 Estimates of area and production of flowers on the basis of Village Survey Approach

Table 2 provides period-wise as well as stratum-wise area in ha of loose and cut flowers separately in the villages of Delhi. The area under loose flowers was estimated to the tune of 2583.28 ha and that of cut flowers 442.59 ha. Thus, during the survey period, total 3025.87 ha area was estimated under flower crops in Delhi. Out of this, 85.37% area was under loose flowers while 14.63% was under cut flowers.

Period-wise and stratum-wise estimates of production of loose and cut flowers have been presented in Table 3. Estimated production of loose flowers was observed maximum, to the tune of 1359.1 MT with 7.03% SE in period-2 of stratum I; 668.551 MT with 7.30% SE in period-1 of stratum II and 8277.3 MT with 4.31% SE in period-2 of stratum III. Pooled over all periods, these figures were 2992.1 MT with 6.80% SE, 1159.8 MT with 4.95% SE and the highest 13540.7 MT with 8.18% SE for the three strata respectively. The period-wise pooled estimates of production of loose flowers was significantly higher in period-2 of the order of 10163.7 MT with 3.73% SE followed by 6272.4 MT with 7.47% SE in period -1 and 1292.5 MT with 10.62% SE in period-3. The overall estimated production of loose flowers in the villages of Delhi was to the tune of 17728.7 MT with 4.09% SE. The period-wise estimates of the production of cut flowers were 359.11495 lakhs with 1.98% SE, 257.38033 lakhs with 5.84% SE and 116.91642 lakhs with 17.97% SE respectively. The overall production of cut flowers was to the tune of 733.41170 lakhs with 3.54% SE.

Table 2: Area (ha) under important flowers in the villages of Delhi

Period	Stratum I			Stratum II			Stratum III			Total		
	Loose	Cut	Total	Loose	Cut	Total	Loose	Cut	Total	Loose	Cut	Total
1	130.08	0	130.08	61.72	23.60	85.32	627.90	98.95	726.85	819.70	122.55	942.25
2	123.05	0	123.05	34.16	75.43	109.59	796.74	85.26	882.00	953.95	160.69	1114.64
3	136.01	0	136.01	0	84.47	84.47	673.62	74.88	748.50	809.63	159.35	968.98
Total	**389.14**	**0**	**389.14**	**95.88**	**183.50**	**279.38**	**2098.26**	**259.09**	**2357.35**	**2583.28**	**442.59**	**3025.87**

Table 3: Estimates of production of important flowers in the villages of Delhi

Period	Stratum I		Stratum II		Stratum III		Total	
	Loose (MT)	Cut (Lakh)	Loose (MT)	Cut (Lakh)	Loose (MT)	Cut (Lakh)	Loose (MT)	Cut (Lakh)
1	1325.385 (13.74)	0	668.551 (7.30)	36.82095 (6.10)	4278.511 (17.28)	322.29400 (*)	6272.447 (7.47)	359.11495 (1.98)
2	1359.087 (7.03)	0	527.328 (*)	91.41663 (15.22)	8277.332 (4.31)	165.96370 (*)	10163.747 (3.73)	257.38033 (5.84)
3	307.654 (17.32)	0	0	72.15765 (30.54)	984.832 (3.79)	44.75877 (*)	1292.487 (10.62)	116.91642 (17.97)
Total	**2992.126 (6.80)**	**0**	**1159.879 (4.95)**	**200.39523 (9.19)**	**13540.675 (8.18)**	**533.01647 (*)**	**17728.681 (4.09)**	**733.41170 (3.54)**

Note: Figures with in parentheses indicate the percent standard errors.

** Estimates are based on very few observations.*

Comparison of estimates of production of flowers on the basis of Market Arrival and Village Survey Approaches

A comparative study of the estimates of production of loose and cut flowers from the market arrivals survey approach and village survey approach is presented in Table 4. The results reveal that a maximum 91.4% of loose flowers produced in the flower growing villages of Delhi arrived for trading in the flower mandis of Delhi in period-2 (peak period of flowers production) while 98.7% cut flowers produced in the villages of Delhi arrived for trading in the flower mandis of Delhi in Period-1. When pooled over the three periods, it was to the tune of 82.2% and 91.5% respectively.

Table 4: Estimates of production of important flowers based on Market Arrival and Village Survey Approaches

Period	Loose (MT)		Cut (Lakh Numbers)	
	Market Arrivals Survey Approach	**Village Survey Approach**	**Market Arrivals Survey Approach**	**Village Survey Approach**
1	4393.147 (70.0%)	6272.447	354.35129 (98.7%)	359.11495
2	9290.062 (91.4%)	10163.747	231.60071 (90.0%)	257.38033
3	887.699 (68.7%)	1292.487	84.73619 (72.5%)	116.91642
Total	**14570.908 (82.2%)**	**17728.681**	**670.68820 (91.5%)**	**733.41170**

Note: Figures in parentheses indicate the percentage of estimated flower production of flowers based on market arrivals approach to the estimated production of flowers based on village survey approach.

15.2 ESTIMATION OF AREA AND PRODUCTION OF MUSHROOMS

A lot of emphasis has been given in our country for the development of agro based industry as it has not only tremendous potential of rural employment generation but it can also gainfully utilize natural and farm resources. It is estimated that about 170 million tonnes of crops residues are left unused for burning in our country. If even a fraction of it is utilized for the production of mushroom, it can make India one of the major mushrooms producing country in the world. Though mushroom production in India started in 1961, it was only after 1990 that some noticeable progress was made.

Mushroom is a perfect health food recommended for use to enrich diet with proteins, vitamins, minerals and fibres. The cultivation of mushroom has varied advantages as each operation is a full-fledged enterprise in itself like compost making, spawn preparation, cultivation, processing and marketing which will provide the employment opportunities to the burgeoning population. For successful production of mushroom, it is necessary for mushroom cultivators to produce as economically and efficiently as possible better quality of mushroom.

The data on mushroom production was collected by the State Government Departments on complete enumeration basis. Production of mushroom involves multiple pickings. Thus, the enumerator is required to make repeated visits for the data collection of each and every picking. The whole process is time consuming, costly and cumbersome. There is thus a possibility of non-sampling errors creeping in the statistics on production of mushroom crop. A sample survey based approach therefore appears desirable for developing estimates of production of mushroom. The National Statistical Commission (Report 2001) has recommended that a sampling methodology be developed for estimation of production of mushroom. Accordingly, a study entitled “Pilot study to develop sampling methodology for estimation of production of mushroom crop” was taken up by Indian Agricultural Statistics Research Institute in Sonepat district of Haryana State to examine the feasibility of sample survey based approach for estimation of production of mushroom.

15.2.1 METHODS AND MATERIAL

The primary data was collected in Sonepat district of Haryana state pertaining to Button Mushroom crop from November 2007 to April 2008. Haryana is the third leading state in the production of mushroom crop after Tamil Nadu and Karnataka and mushroom is extensively cultivated in the Sonepat district of the state. The other mushroom producing states are Kerala, Jammu & Kashmir, Himachal Pradesh, Punjab and Uttar Predesh. The data pertaining to shed area for raising mushroom, number of beds in each shed, weight of wet compost used, spawn consumed, wheat/paddy straw used in preparation of compost, processing of mushroom after picking, disposal of produce etc. have been collected from the selected mushroom growers in each of the selected village by enquiry method.

15.2.2 Sampling Design

The sampling design was stratified two-stage random sampling with blocks/group of blocks as strata, mushroom-growing villages as primary stage sampling units and mushroom growing cultivators as the ultimate stage unit of selection. The Sonepat district comprising of 6 blocks namely; Ganaur, Gohana, Sonepat, Kharkhoda, Rai and Mudlana, was subdivided into three strata by combing the adjacent blocks. These strata were Ganaur (Ganaur and Gohana), Sonepat (Sonepat and Kharkhoda) and Rai (Rai and Mudlana). There were 23 mushroom growing villages in Ganaur, 22 in Sonepat and 8 in Rai. A total of 8 villages, 3 from each of Ganaur and Sonepat and 2 from Rai were selected by simple random sampling without replacement. All mushroom growing cultivators in each of the selected villages were categorized into three categories as Small (Wet compost used up to 500 qtls), Medium (Wet compost used from 500 to 800 qtls) and Large (Wet compost used more than 800 qtls). Six cultivators were selected from these categories with proportional allocation by simple random sampling without replacement for intensive data collection on production of mushroom in the district. Table 1 gives the stratum-wise total number of mushroom growing villages, number and name of randomly selected mushroom growing villages and number of mushroom growing cultivators.

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

Stratum	Total number of villages (N_h)	Selected villages (n_h)	Name of the selected villages	Total number of mushroom growers (M_{hi})	Selected mushroom growers (m_{hi})
Ganaur (Stratum -1) (Ganaur & Gohana)	**23**	**3**	Ahirmajra	24 (5, 9, 10)	6 (1, 2, 3)
			Ganaur	08 (7, 0, 1)	6 (5, 0, 1)
			Rajlugarhi	23 (23, 0, 0)	6 (6, 0, 0)
Sonepat (Stratum –2) (Sonepat & Kharkhoda)	**22**	**3**	Rohat	26 (23, 0, 3)	6 (4, 0, 2)
			Kakroi	18 (18, 0, 0)	6 (6, 0, 0)
			Baiyapur	21 (6, 6, 9)	6 (2, 2, 2)
Rai (Stratum-3)	**08**	**2**	Sersa	06 (0, 0, 6)	6 (0, 0, 6)
			Aterna	27 (27, 0, 0)	6 (6, 0, 0)
TOTAL	**53**	**08**		**153 (109, 15, 29)**	**48 (30, 4, 14)**

15.2.3 Estimation Procedure

Let y_{hij} be the production of mushroom for the j^{th} mushroom growing cultivator of the i^{th} village in the h^{th} stratum (j = 1, 2, …, M_{hi} ; I = 1, 2, …, N_h ; h = 1, 2, 3) where M_{hi} is the number of mushroom growing cultivators in the i^{th} village of the h^{th} stratum and N_h is the total number of mushroom growing villages in the h^{th} stratum. The estimate of productivity (qt/ha) of mushroom in the i^{th} village of the h^{th} stratum is given by

$$\bar{y}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \quad , \qquad (1)$$

where m_{hi} is the number of selected mushroom growing cultivators in the h^{th} stratum of the i^{th} village.

The estimate of average yield (qt/ha) of mushroom in the h^{th} stratum is given by

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\bar{y}_{hi}$$

The variance of $\bar{y}_h$ is given by

$$V(\bar{y}_h)=\left(\frac{1}{n_h}-\frac{1}{N_h}\right)S_{bh}^2+\frac{1}{n_h}\sum_{i=1}^{N_h}\left(\frac{1}{m_{hi}}-\frac{1}{M_{hi}}\right)S_{hi}^2$$

where $S_{bh}^2=\frac{1}{(N_h-1)}\sum_{i=1}^{N_h}\left(\bar{Y}_{hi}-\bar{Y}_h\right)^2$, $S_{hi}^2=\frac{1}{(M_{hi}-1)}\sum_{j=1}^{M_{hi}}(Y_{hij}-\bar{Y}_{hi})^2$

The estimate of variance of $\bar{y}_h$ is given by

$$\hat{V}(\bar{y}_h)=\left(\frac{1}{n_h}-\frac{1}{N_h}\right)s_{bh}^2+\frac{1}{n_h N_h}\sum_{i=1}^{n_h}\left(\frac{1}{m_{hi}}-\frac{1}{M_{hi}}\right)s_{hi}^2 \qquad (2)$$

where $s_{bh}^2=\frac{1}{(n_h-1)}\sum_{i=1}^{n_h}\left(\bar{y}_{hi}-\bar{y}_h\right)^2$, $s_{hi}^2=\frac{1}{(m_{hi}-1)}\sum_{j=1}^{m_{hi}}(y_{hij}-\bar{y}_{hi})^2$

Accordingly, an estimator of the average yield (qt/ha) of mushroom in the district is given by

$$\hat{\bar{Y}}=\frac{1}{N}\sum_{h=1}^{3}N_h\,\bar{y}_h \qquad (3)$$

where $N=\sum_{h=1}^{3}N_h$

The variance of $\hat{\bar{Y}}$ is given by

$$V\left(\hat{\bar{Y}}\right)=\frac{1}{N^2}\left[\sum_{h=1}^{3}N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)S_{bh}^2+\sum_{h=1}^{3}\frac{N_h^2}{n_h}\sum_{i=1}^{N_h}\left(\frac{1}{m_{hi}}-\frac{1}{M_{hi}}\right)S_{hi}^2\right]$$

The estimate of variance of $\hat{\bar{Y}}$ is given by

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\left[\sum_{h=1}^{3} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_{bh}^2 + \sum_{h=1}^{3} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \left(\frac{1}{m_{hi}} - \frac{1}{M_{hi}}\right) s_{hi}^2\right] \qquad (4)$$

15.2.4 Results

The results on estimates of productivity of mushroom, spawn used, weight of wet compost used in production of mushroom etc. are summarized below:

Estimate of productivity (qtl/ha) of mushroom

The stratum-wise and pooled over all the strata results of the estimates of productivity are presented in Table 2. The results revealed that the productivity of mushroom was maximum 383.62 qtl/ha with 5.17% standard error (SE) in stratum III followed by 325.44 qtl/ha with 9.62% SE in stratum I and 312.34 qtl/ha with 4.76% SE in stratum II. Pooled over all the strata, the estimate of productivity of mushroom in Sonipat district was to the tune of 328.78 qtl/ha with 4.63% SE. The estimate of productivity (kg/tray) was observed highest in stratum III (4.63 kg/tray with 4.31% SE) and lowest in stratum II (4.35 kg/ha with 1.98% SE). The productivity in stratum I was observed 4.50 kg/tray with 1.38% SE. The productivity of mushroom in Sonepat district was estimated as 4.46 kg/tray with 1.21% SE.

Table 2: Estimates of productivity of mushroom along with percentage standard error

Stratum	Estimates of productivity of mushroom	
	According to area under mushroom (qtl/ha)	**Per tray (kg)**
I	325.44 (9.62%)	4.50 (1.38%)
II	312.34 (4.76%)	4.35 (1.98%)
III	383.62 (5.17%)	4.63 (4.31%)
Over all	**328.78 (4.63%)**	**4.46 (1.21%)**

Note: Figures within parentheses represents percentage standard errors of corresponding estimates.

Estimates of productivity of mushroom on the basis of sample survey approach as well as complete enumeration approach are presented in Table 3. It may be seen from the Table that production of mushroom in the Sonepat district was estimated as 2063 Metric Tonnes with 416810 trays in 2007-08 by District Horticulture Office, Sonepat on the basis of complete enumeration approach. The productivity of mushroom was observed to be 4.94 kg/tray on the basis of complete enumeration approach while the same obtained through sample survey approach was observed to be 4.46 kg/tray. In view of the closeness of the estimates from two sources it can be inferred that sample survey approach appears to be suitable for generating estimates on productivity of mushroom crop.

Table 3: Estimate of productivity of mushroom on the basis of complete enumeration approach and sample survey based approach

Estimate of productivity on the basis of complete enumeration approach			**Sample survey approach**
No. of trays	**Total Production (MT)**	**Productivity/Tray (kg)**	**Productivity/Tray (kg)**
416810	2063	4.94	4.46 (1.21%)

Note: Figures within parentheses represent percentage standard errors of corresponding estimates.

Estimate of wet compost used in cultivation of mushroom

The results obtained for the estimates of wet compost used in mushroom production in the selected villages of Sonepat district are presented in Table 4. It may be seen that maximum wet compost was observed in stratum III. It was 327.2 Kg/ha with 14.18% SE. In stratum I, it was observed minimum i.e. 258.9 Kg/ha with 9.88% SE. Pooled over all the three strata, the wet compost used in mushroom cultivation was observed to be to the tune of 269.8 Kg/ha with 6.50% SE.

Table 4: Estimates of wet compost used in cultivation of mushroom

Stratum	Estimates of wet compost weight used in mushroom cultivation (Kg/Ha)
I	258.9 (9.88%)
II	260.4 (10.75%)
III	327.2 (14.18%)
Over all	**269.8 (6.50%)**

Note: Figures within parentheses represents percentage standard errors of corresponding estimates.

Estimate of spawn used in cultivation of mushroom

The estimates of spawn used in cultivation of mushroom are presented in Table 5. The results revealed that estimate of spawn used in cultivation of mushroom crop in Sonepat district was observed maximum in stratum III (23.86 qtl/ha with 9.39% SE) followed by 20.90 qtl/ha with 3.53% SE in stratum II. It was 19.20 qtl/ha with 9.74% SE in stratum I. It was of the order of 20.61 qtl/ha with 4.52% SE for the entire Sonepat district. This figure was obtained by pooling over all the strata.

Table 5: Estimates of spawn used in cultivation of mushroom along with % S E

Stratum	Estimates of spawn used in mushroom cultivation (Qtl/ha)
I	19.20 (9.74%)
II	20.90 (3.53%)
III	23.86 (9.39%)
Over all	**20.61 (4.52%)**

Note: Figures within parentheses represents percentage standard errors of corresponding estimates.

Estimates of different supplements used in preparation of compost

The role of compost preparation in cultivation of mushroom crop is very important. The supplements wheat/paddy straw, urea, calcium ammonium nitrate/ di-amonium phosphate (CAN/DAP), super phosphate, murate of potash, wheat bran, chicken manure, gypsum and furadan are used in preparation of compost. The estimates of each of the above mentioned supplements are summarized as follows:

Estimators of wheat/paddy straw, urea, CAN/DAP, super phosphate and murate of potash are presented in Table 6. It was observed that the maximum wheat/paddy straw was used

by the mushroom growers in stratum III. This was 30.21 qtl/ha with 8.21% SE in stratum III while it was observed minimum in stratum I (24.20 qtl/ha with 6.15% SE). The pooled estimate for wheat/paddy straw was of the order of 25.79 qtl/ha with 4.78% SE. Estimate of urea used in preparation of compost was observed 19.77 qtl/ha with 8.50% SE, the maximum was in stratum III (17.47 qtl/ha with 6.00% SE) and the minimum in stratum II. The estimate of urea for the entire district was of the tune of 18.51 qtl/ha with 4.66% SE.

The estimate of CAN/DAP was found maximum in stratum III (18.64 qtl/ha with 7.31% SE) followed by stratum I and stratum II. It was of the order of 15.66 qtl/ha with 3.37% SE and 13.64 qtl/ha with 2.48% SE in stratum I and stratum II respectively. The over all estimate of CAN/DAP used in preparation of compost was of the tune of 15.27 qtl/ha with 2.22% SE. The estimates of super phosphate in preparation of compost vary from 17.16 qtl/ha to 24.19 qtl/ha in different strata with an average of 17.79 qtl/ha for the district. It was observed maximum in stratum III while the minimum was in stratum II. The percentage standard error varies from 5.03% to 9.52% from stratum to stratum and pooled over all the strata, the estimated value of super phosphate used in preparation of compost was to the tune of 18.89 qtl/ha with 4.87% SE. Estimate of murate of potash was found maximum in stratum III (20.11 qtl/ha with 3.30% SE) followed by stratum I (17.91 qtl/ha with 9.27% SE) and stratum II (16.83 qtl/ha with 7.82% SE) respectively. The pooled estimate was 17.79 qtl/ha with 5.11% SE.

Table 6: Estimates of Wheat/Paddy straw, Urea, CAN/DAP, Super phosphate and Murate of Potash used in preparation of compost

Stratum	Estimates of Wheat/Paddy straw, Urea, CAN/DAP, Super phosphate and Murate of Potash used in preparation of compost				
	Wheat/Paddy Straw (qt/ha)	**Urea (qt/ha)**	**CAN/DAP (qt/ha)**	**Super phosphate (qt/ha)**	**Murate of Potash(qt/ha)**
I	24.20 (6.15%)	19.07 (8.46%)	15.66 (3.37%)	18.70 (9.52%)	17.91 (9.27%)
II	25.77 (8.93%)	17.47 (6.00%)	13.64 (2.48%)	17.16 (6.50%)	16.83 (7.82%)
III	30.21 (8.21%)	19.77 (8.50%)	18.64 (7.31%)	24.19 (5.03%)	20.11 (3.30%)
Over all	**25.79 (4.78%)**	**18.51 (4.66%)**	**15.27 (2.22%)**	**18.89 (4.87%)**	**17.79 (5.11%)**

Note: Figures within parentheses represents percentage standard errors of corresponding estimates.

Estimates of wheat bran, chicken manure, gypsum and furaden used in preparation of compost by the mushroom growers are presented in Table 7. The estimate of wheat bran used in preparation of compost was of the order of 59.63 qtl/ha with 9.78% SE in stratum III (maximum), 45.87 qtl/ha with 5.58% SE in stratum II (minimum) and 58.46 qtl/ha with 8.90% SE in stratum III. Over all estimate of wheat bran was observed to be 53.48 qtl/ha with 5.19% SE. From the data collected on use of chicken manure in compost preparation it was observed that maximum chicken manure was mixed by the mushroom growers of stratum II. It was estimated 62.85 qtl/ha with 7.01% SE in stratum II followed by 59.57 qtl/ha with 9.80% SE in stratum III and 56.71 qtl/ha with 3.65% SE in stratum I. It was to the tune of 59.81 qtl/ha with 3.80% SE for whole of the district.

It can be seen from Table 7 that the estimate of gypsum was observed maximum in stratum I (75.12 qtl/ha with 6.26% SE) followed by 70.36 qtl/ha with 7.47% SE in stratum III and 68.61 qtl/ha with 4.80% SE in stratum II respectively. Pooled over all the three strata, the estimate of gypsum used in preparation of compost was of the order of 71.57qtl/ha with 3.56% SE. The estimates of use of furaden in preparation of compost vary from 0.66 qtl/ha to 0.75 qtl/ha in the three strata with the variation in percentage standard error from 6.84 to 9.64. The estimate of furaden was observed maximum in stratum I and the minimum in stratum II. It was estimated as 0.71 qtl/ha with 5.06% SE for the entire district.

Table 7: Estimates of Wheat Bran, Chicken Manure, Gypsum and Furaden used in preparation of compost

Stratum	Estimates of Wheat Bran, Chicken Manure, Gypsum and Furaden used in preparation of compost			
	Wheat Bran (qt/ha)	**Chicken Manure (qt/ha)**	**Gypsum (qt/ha)**	**Furaden (qt/ha)**
I	59.63 (9.78%)	56.71 (3.65%)	75.12 (6.26%)	0.75 (6.84%)
II	45.87 (5.58%)	62.85 (7.01%)	68.61 (4.80%)	0.66 (9.64%)
III	58.46 (8.90%)	59.57 (9.80%)	70.36 (7.47%)	0.73 (7.22%)
Over all	**53.48 (5.19%)**	**59.81 (3.80%)**	**71.57 (3.56%)**	**0.71 (5.06%)**

Note: Figures within parentheses represents percentage standard errors of corresponding estimates.

Recommendations

Sample survey based approach appears to be appropriate for developing estimate of production of mushroom but some more such like studies needed to be carried out before it can be recommended for adoption. Percentage SEs estimated for different parameters for

estimating mushroom productivity is estimated under the reasonable limits, it is recommended that the sample sizes at both the stages may be taken same as in this study for the future surveys. Mushrooms are grown in sheds/beds. Therefore, reporting the estimates of different parameters on the basis of area (quintal per ha) is a better way than is currently the case in the district i.e. reporting made on the basis of kg per tray. The area under mushroom may be calculated as the total area in ha of the beds used in sheds for cultivating mushroom crop.

REFERENCES

Gupta, A. K., Jain. V. K., Narang, M. S., Tyagi, K. K. and Sud, U. C. (2004): *"Pilot sample survey to develop sampling methodology for estimation of area, production and productivity of important flowers on the basis of market arrivals"*, Project Report published by IASRI, New Delhi. The project was funded by CSO, Ministry of Statistics & Programme Implementation, Government of India.

Murthy, M. N. (1977): *"Sampling theory and methods"*, Statistical Publishing Society, Calcutta-700 035, India.

Report of the National Statistical Commission (2001): National Statistical Commission, Government of India, Volume 1, p 66-67.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Asok, C. (1984): *"Sampling theory of Surveys with applications"*, Iowa State University Press, AMES, Iowa, U.S.A. and Indian Society of Agricultural Statistics, New Delhi, India.

REGRESSION ANALYSIS FROM SAMPLE SURVEY DATA

U.C. Sud

Indian Agricultural Statistics Research Institute, New Delhi-110012

16.1 INTRODUCTION

Sample Surveys are generally multivariate and at times the surveyor's interest is to establish the pattern of relationships between the variables on which data are collected. In this context the correlation and regression analysis is most appropriate. The usual theory of least squares for estimation of regression coefficients assumes that the observations are identically, independently distributed. This assumption is satisfied for with replacement designs . But surveys are generally conducted using multistage random sampling designs where in there are stratum, clusters and units are selected without replacement. Thus, the assumption of independence of units do not hold good in the context of survey data.The usual least squares approach needs to be modified to account for the sampling design. Standard statistical packages like SPSS, BMDP, SAS etc. which are based on usual least squares approach will provide misleading inferences about the estimates of regression coefficients when the data have been collected from a complex sampling design.

When the population is very heterogeneus a single regression equation for the entire population may not fit. One may have to fit separate regression line for each group. If the groups are strata there are no problems. However, if the groups happen to be clusters then there may be serious problem as only a sub-set of clusters are represented in the sample.. Problems of this type have been tackled by Pfeffarmann and Nathan (1981) using a random coefficient regression model.

16.2 THEORETICAL DEVELOPMENTS

Assuming that a single regression equation suffice for the entire population, there can be two types of inferences; descriptive inference and analytic inference. Descriptive inferences relate to parameters which are functions of the values attached to the 'N' units in the population. When the parameter of interest relates to the super population, inferences are analytic.

Descriptive inferences can be design based or model based. The former is based on the distribution generated by random sampling and details can be found in standard text such as Cochran (1977). In the model based approach models are postulated to represent the population structure- and inferences are based on the probability distribution specified in the model, the so called ξ-distribution. Important references in this context are Smith (1976) and Sarndal (1978). A combined design and model based approach is also proposed like by Godambe and Thompson (1973). A detailed review of various approaches is given by Cassel, Sarndal and Wretman (1977).

When the object of inference is the model parameter it is assumed that the finite population constitutes a random sample from the super population. Since survey populations are usually large a finite population estimator based on all values will be "close to" the unknown super population parameter. This estimator constituting the

finite population parameter can be estimated from the sample. However, the assumption of an underlying super population is critical to interpretation.

The approaches described above can be put in a tabular form depending upon the parameter of interest and type of inferences. Hartley and Sielken (1975) classify regression models according to these two criteria.

SAMPLING THEORIES CLASSIFIED BY SAMPLING PROCEDURES AND TARGET PARAMETERS

Target Parameters	**Sampling Procedure**	
	Repeated sampling from a fixed finite population	Repeated two-step sampling from an infinite population
	Case 1	Case 2
Parameters of finite Population **B**	Classical finite population sampling theory. p- inference	Super population theory for finite population ξ-inference
	Case 3	Case 4
Parameters of infinite super population β	Infeasible	Inference on infinite population parameters two-step sampling procedure ξ-inference

Kish and Frankel (1974), Shah Holt and Folsom (1977) and Jonrup and Rennermalm (1976) consider that the only relaxant inference concerns the finite population parameter. The approach of estimation is purely design based. Kish and Frankel (1974) suggest these methods for estimating the sampling variance of the estimator. These are Taylor series method, the Jacknife method and balanced repeated replication. They have made computations of design effect (deft) of sample regression co-efficients in the case of clustered population with high intra cluster correlation and shown that the deff's are greater than one. Therefore, standard errors based on usual approach will be serious underestimate. The case where the parameter of interest is the finite population regression co-efficient and the finite population is a realization of a super population has been studied by Hartley and Seilken (1975). Hence both the sampled regression co-efficient and population regression coefficients are random variables in this approach.

Scott and Holt (1982) consider model parameters as object of inference. Since clustering effects are qualitative rather than qualitative such effects are best represented by incorporating intra-cluster correlations into the covariance mapix of the regression residuals.

Under this model where observations are assumed to have come from a two-stage sample the effect of using ordinary least-squares estimators are examined. The OLS estimator is unbiased, but there is loss in efficiency and estimate of variance of the estimated co-efficients are serious under estimate.

If the variable used for survey design is quantitative (design variable) in nature then the effect of survey design can be incorporated by including the design variable as another explanatory variable. Nathan and Holt (1980) have followed this approach. OLS estimator is shown to be biased. The maximum likelihood estimator derived under normality assumption by Demets and Halperin (1977) is shown to be asymptotically unbiased. In order to examine the effect of a quantitative design variable a simulation study was carried out which compared the probability weighted estimator and its variance, the ordinary least squares estimator and its least square variance and the maximum likelihood estimator and its normal theory variance in terms of the coverage probability of confidence intervals. The design variable was used to construct strata and to vary the selection probabilities within strata. The results of the study clearly demonstrated that the maximum likelihood estimator performs well over the other two estimators. However, in some other situation the p-weighted estimators also performed fairly well.

Dwelling on the controversy relating to the model based and design-based approaches of inference Dumouchel and Duncan (1983) enumerate the efficiency advantage of a model-based estimator where the model holds and the consistency property of design-based estimator whether or not the model holds. They also cite earlier references pertaining to the both sides of controversy. In addition a test is proposed, which can be performed with computer packages for linear regression to examine whether the model-based (un weighted) and design -based (weighted) regressions are different. If the null hypothesis, that there is no significance difference between weighted and unweighted estimators, is rejected the authors suggest using a weighted estimator while in the other case a unweighted estimator based purely on the assumptions of the model is suggested. The paper of Dumouchel and Duncan clearly illustrates the most-essential consideration in choosing between model based and design - based inference, namely, efficiency under a correctly specified model versus consistency under failure of the assumptions of the model.

Nathan (1981) points out the deficiencies of both the approaches of making inference. His criticism of the model based approach stems from the fact that it relies heavily on the assumption that the presumed model is correct - an assumption not realized very often practically indicating thereby that the model based inferences are not robust to departures from model. On the other hand design-based inferences are made on finite population parameters having little descriptive values in themselves. As a remedial action he suggests a compromise by taking only those finite population parameters which are close approximations of model parameters as target parameter of interest. In line with this argument he proposes estimating separate regression coefficients for sub-populations rather than estimating a single regression coefficient for the whole of the population. If the sub-populations are ,large enough this will ensure that the finite population regression closely approximate the super-population parameter, so that any inference relating to the finite population parameters can be considered as relating to the super-population parameter.

Commenting on the controversly, Kalton (1983) mentions that any approach to be adopted should depend on the purpose of analysis. If prediction is the ultimate aim the design-based approach should be adopted. However, if one is aiming higher i.e. developing a causal model that applies to other populations as well then model based approach is appropriate.

Regression models in Surveys

The model very widely used is

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{B} + \underset{\sim}{e}$$

where $\underset{\sim}{Y}$ denotes the vector of response variable, $\underset{\sim}{X}$ the matrix of explanatory variables, $\underset{\sim}{e}$ the vector of residuals, and the vector $\underset{\sim}{B}$ denotes the regression coefficients obtained by minimizing the residual sums of squares over 'N' units in the population. The value of $\underset{\sim}{B}$ is given by

$$\underset{\sim}{B} = \left(\underset{\sim}{X}' \underset{\sim}{X} \right)^{-1} \underset{\sim}{X}' \underset{\sim}{Y}$$

Kish and Frankel (1974) adopt this model. However, Duomouchel and Duncan (1983) note that this is not truly a model and it simply gives the definition of finite population parameter.

Fuller (1975) and Hatley and Seilken (1975) consider current finite population as a random sample from a super-population with structure $\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{B} + \underset{\sim}{e}$, where $\underset{\sim}{\beta}$ is a vector of fixed constants and $\underset{\sim}{e}$ are random error terms with $E_\zeta\left(\underset{\sim}{e}\right) = \underset{\sim}{0}$ and $V_\zeta\left(\underset{\sim}{e}\right) = \sigma^2 \underset{\sim}{V}$ and expectations are taken with respect to the super-population or ζ distribution. The target parameter of interest is

$$\underset{\sim}{B} = \left(\underset{\sim N}{X_N'}\, \underset{\sim N}{V_N^{-1}}\, \underset{\sim}{X_N} \right)^{-1} \underset{\sim N}{X'}\, \underset{\sim N}{V^{-1}}\, \underset{\sim N}{Y}$$

when $\underset{\sim N}{V} = \underset{\sim N}{I}$, $\underset{\sim}{B}$ reduces to $\left(\underset{\sim}{X_N}' \underset{\sim}{X_N} \right)^{-1} \underset{\sim}{X_N}' \underset{\sim}{Y_N}$ where $\underset{\sim}{X_N'}, \underset{\sim}{Y_N}$ are of dimension Nxp and pX 1 respectively.

There are others who use the same model as above that is

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{e}$$

with usual assumption but the target parameter of interest is $\underset{-}{\beta}$ not $\underset{-}{B}$. The estimator is obtained by direct application of least squares.

Some additional models are discussed in the survey sampling literature. These models can be considered in combination with each of the three cases discussed above. One of these models is of the form $\underset{-}{Y} = \underset{-}{X}\,\underset{-j}{\beta} + \underset{-}{e}$, or $\underset{-}{Y} = \underset{-}{X}\,\underset{-j}{B} + \underset{-}{e}$ where $\underset{-j}{\beta}$ or $\underset{-j}{B}$ are the regression coefficients in a subset of the population, such as a stratum or cluster. As mentioned earlier if the subsets are strata the $\underset{-j}{\beta}$ or $\underset{-j}{B}$ may be estimated separately for each stratum, but in the case of clusters this is not possible. Several authors have suggested an average of the $\underset{-j}{\beta}$ or $\underset{-j}{B}$ across subsets, such as $\sum \pi_j B_j$, where π_j is the proportion of the population in the jth subset , as a target parameter of interest (Konijn, (1962); Pfeffaramann and Nathan, (1977). Porter (1973), in considering a situation with several observations per unit, uses a model with B_i specific to unit i, and he estimates the mean of B_i , $\sum B_i / N$ .The Porter model is similar to the random coefficients regression model (see, for instance, Swamy, 1970) but the two models are not equivalent : in Porter's model, the B_i's are fixed for a given unit, but for the random coefficients model they are random variables.

Regression Analysis of Survey Data

If $\underset{-}{B}$ is the parameter of interest, it is estimated from a survey sample of size 'n' by

$$\underset{-w}{\hat{B}} = \left(\underset{-}{X}' \underset{-}{W}^{-1} \underset{-}{X} \right)^{-1} \underset{-}{X}' \underset{-}{W}^{-1} \underset{-}{Y}$$

where $\underset{-}{X}$ is the matrix of sample values of the explanatory variables. $\underset{-}{Y}$ is (nx1) sampled vector of response variable, and $\underset{-}{W}$ is (nxn) diagonal matrix of inclusion probabilities on the diagonal. For a super-population model with heteroscedastic variance covariance matrix of residuals, a weighted least squares may be used for sample estimator

$$\underset{-}{\beta}^{\wedge} = \left(\underset{-}{X}' \underset{-}{V}^{-1} \underset{-}{X} \right)^{-1} \underset{-}{X}' \underset{-}{V}^{-1} \underset{-}{Y}$$

The difference between the two estimators arise due to the weights used i.e. in the first case $\underset{-}{W}^{-1}$ and in other case $\underset{-}{V}^{-1}$

In case of homoscedasticity and for self-weighting designs $\underset{-W}{B}^{\wedge}$ and $\underset{-}{\hat{\beta}}$ coincide. The estimator $\underset{-W}{\hat{B}}$ can be obtained directly from standard computer programs which provide for weighted regression (e.g. BMDP) by using the weights $\underset{-}{W}^{-1}$ or from packages like, SPSS by carrying out unweighted regression on the transformed

variables $\sqrt{W_i^{-1}}\,X_i$, $\sqrt{W_i^{-1}}\,Y_i$ However, under either alternative the reported variances and covariances of the estimators are incorrect.

The model variance of $\hat{\underset{-}{\beta}}$ is

$$\sigma^2\left(\underset{-}{X}'\underset{-}{X}\right)^{-1}$$

which is the result given by standard unweighted regression program. However, the model variance of $\hat{\underset{-W}{B}}$ is

$$\sigma^2\left(\underset{-}{X}'\underline{\underset{-}{W^{-1}}}\underset{-}{X}\right)^{-1}\underset{-}{X}'\underset{-}{W}^{-1}\underset{-}{W}'^{-1}\underset{-}{X}\left(\underset{-}{X}'\underset{-}{W}^{-1}\underset{-}{X}\right)^{-1}$$

The weighted regression programs with weights $\underset{-}{W}^{-1}$ will give a different value i.e. $\sigma^2\left(\underset{-}{X}'\underset{-}{W}^{-1}\underset{-}{X}\right)^{-1}$ for the model variance of $\underset{-W}{B}^{\wedge}$ which equals $\mathrm{V}\left(\underset{-W}{B}^{\wedge}\right)$ only if $\underset{-}{W}^{-1}=\underset{-}{I}$ .The design variance of $\underset{-W}{B}^{\wedge}$ which should be the relevant measure of accuracy of $\underset{-W}{B}^{\wedge}$ as an estimator of $\underset{-}{B}$ can be estimated by techniques such as Taylor series linearization or sample reuse approach. The latter includes techniques such as balanced repeated replication, Jackknife repeated replications and boot strap method. The programs SURREGR from Research triangle Institute and SUPERCARP, (a PC version of it called PCCARP is now also available) from IOWA State University obtains variance estimates for regression statistics using the Taylor Series approach. The REPERR program in the Survey Research Center's OSIRIS IV software system computes sampling errors of regression statistics by either the BRR or JRR methods (Lepkowski, 1982).

Empirical comparisons of the variance estimators are given by Kish and Frankel (1974) and by Richards and Freeman (1980) and theoretical comparisons by Krowski and Rao (1981). Four methods of estimating variances of non-linear estimators have been described in detail by Kalton (1977) namely Taylor Expansion method, Simple Replicated Sampling, BRR and JRR. His empirical study reveals that the Taylor's method ,BRR , and JRR method perform satisfactorily with Taylor's method having slight edge. However, the replicated sampling method was found to be inferior.

From the above discussion it is clear that the question central to regression analysis of survey data is choosing the appropriate approach and then the corresponding parameters of interest. Once this is sorted out the estimators, and the variance estimators are readily available. It can be seen that different solutions are available in the literature. Thus Dumouchel and Duncan suggest carrying out a test to settle the controversy, Nathan offers a compromise package, Brewer and Mellor emphasize on standardizing the model while Kalton stresses on keeping ones aim in view while making a decision regarding the approach to be followed. But no clear cut solutions are provided and ultimate decision rests on the person who is working on the problem.

16.3 MEASUREMENT ERRORS IN CONTEXT WITH REGRESSION ANALYSIS

It is well known that data collected in sample surveys particularly that collected from human respondents are subject to measurement errors. The US Bureau of the census (1972) has reported estimates of response variance, as a percentage of total variance ranging from 0.5 to 40 percent. Battese et al (1972) report response variances of a similar magnitude for item associated with farm operations. In addition to response errors, coding and processing errors also occur. Also, imputation can be viewed as giving rise to measurement error - the measurement error in this case being the difference between imputed value and the true value. A sizable literature exists on the effects of random measurement errors on regression and correlation analysis especially in econometrics and psychometrics. If the response variable is subject to measurement errors, the regression coefficients are unaffected (although the estimates are less precise). If, however, the explanatory variables are subject to response error the least squares estimators are biased. If the response errors in the independent and dependent variables are correlated the bias is increased or decreased depending on the signs of error correlation. Cochran (1968) and chai (1971), discussed the effect of response variance on regression statistics. Fuller (1971), investigated the properties of errors in variables estimators of regressions parameters under super population model with normal errors. While estimating regression parameters in the presence of measurement errors Fuller (1975) utilising Frankel's data (1971) found out that adjustment of the covariance matrix for response errors resulted in about ten percent increase in the estimated coefficients. The computer package SUPERCARP (Hidiroglou, Fisher and Hickman, 1980) provides the option of specifying uncorrelated measurement errors in model development.

In practice measurement errors are unlikely to be random but they may well be correlated with other variables. The possible effect of correlated measurement errors on regression analysis is illustrated by Duncan and Hill (1984). In a survey data on labor earnings it was found that regression co-efficient of tenure variable was substantially underestimated when the survey response were used. They attribute this under estimation to the fact that there was a sizable negative correlation between the measurement error in reported earnings and the level of tenure. This emphasizes the importance of paying serious attention to the possible effect of measurement errors on the interpretation of regression analysis of survey data.

REFERENCES

Battese, G.E., Fuller, W.A. and Hickkman, R.P. (1972). 'Interviewer effects and response errors in a replicated survey of farm operators in selected Iowa Counties." *Report to Statistical Reporting Service,U.S. Deptt. of Agri. Iowa State Univ., Ames, Iowa.*

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

Cochran, W.G. (1977). 'Sampling techniques.' 3rd ed. *New York : Wiley.*

Demets, D. and Halperin, M. (1977). 'Estimation of a single regression co-efficient in samples arising from a sub-sampling procedure'. *Biometrics. 33, 47 - 56.*

Dumouchel, W.H. and Duncan, G.J. (1983). Using sample surveys weights in multiple regression analysis of a stratified samples'. *J. Amer. Statist. Ass. 78, 535 - 543.*

Duncan, G.J. and Hill, D.H. (1984). 'An investigation of the extent and consequences of measurement error in labor economic survey data.' *Unpublished working paper. Survey Research Center, The University of Michigan, Ann Arbor.*

Frankel, M.R. (1971). 'Inference from survey samples.' *Institute for Social Research, University of Michigan, Ann Arbor.*

Fuller, W.A. (1975). 'Regression analysis for sample survey.' *Sankhya C: The Indian Journal of Statistics.' 37:3, 117 -132.*

Hartley, H.O. and Seilkon, Jr., R.L. (1975). 'A super population viewpoint for finite population sampling .' *Biometrics, 31, 411 -422.*

Hidiroglou, M.A. (1982). 'Computerization of complex survey estimates.' *Survey methodology, 8, 102 - 121.*

Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980). 'Supercarp'. *Proc. of Sect. Sur. Res. Meth. Amer. Statist. Ass., 449-450.*

Jonrup, H. and Rennermalm, B. (1976) . 'Regression analysis in samples from finite populations.' *Scand. J.Statist. 3, 33 - 36.*

Kalton, G. (1983).'Models in the practice of survey sampling.' *Int. Stat. Rev. 51,175-188.*

Kish, L. and Frankel, M.R.(1974). 'Inference from complex samples.' *J. R. Stat. Soc. B, 36, 1 - 37.*

Konijn, H.(1962).'Regression analysis in sample surveys.'*J. Amer. Stat. Ass. 57,590-605.*

Lepkowski, J.M. (1982). 'The use of OSIRIS IV to analyse complex sample survey data.' *Proc. of Sect. Sur. Res. Meth. Amer. Statist. Ass. 38 - 43.*

Nathan, G. (1981). 'Notes on inference based on data from complex sample designs.' *Survey methodology, 7, 109 - 129.*

Nathan, G and Holt D. (1980). 'The effect of survey design on regression analysis.' *J.R. Statist. Soc. B, 42, 377 - 386.*

Pfefferman, D. and Nathan, G. (1977). 'Regression analysis of data from complex samples.' *Bull. Int. Statist. Inst. 47.*

Porter, R.D. (1973). 'On the use of survey sample weights in the linear model.' *Ann. of Eco. and Soc. Meas. 2:2, 141 - 158.*

Sarndal, C.E. (1978). 'Design based and model-based inference in survey sampling.' *Scand. J. Statist., 5, 27 - 52.*

Sarndal,C.E., Swensson,B., and Wretman, J. (1991). Model assisted survey sampling, *Springer-Verlag, New York.*

Scott, A.J. and Holt, D.(1982) ' The effect of two -stage sampling on ordinary least squares method., *J. Amer. Statist. Ass. 77, 844-854.*

Shah B.V , Holt ,M.M.and folsom R.E.(1977).'Inference about regression models from sample survey data.' *Bull. Int. Statist.Inst.,47(3), 43-57.*

Survey Research Center Computer Support Group (1982). 'The OSIRIS IV User's Manual, *7th ed. Institute for Social Research, Ann Arbor, Michigen.*

Swamy,P.A.V.B.(1970). 'Efficient Inference in a random coefficient regression model.' *Econometrica. 38, 311-323*

Scheuer,E.M. and Stoller,D.S.(1962). On the generation of normal random vectors, *Technometrics, 1962,4 : 278-281*.

Smith, T.M.F. (1976). The foundations of survey sampling: a review. *J.Roy. Statist Soc. A, 139, 183-195.*

REGRESSION ANALYSIS

Lalmohan Bhar

Indian Agricultural Statistics Research Institute, New Delhi-110012
e-mail: lmbhar@gmail.com

17.1 Introduction

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other, or others. This methodology is widely used in business, the social and behavioral sciences, the biological sciences including agriculture and fishery research. For example, fish weight at harvest can be predicted by utilizing the relationship between fish weights and other growth affecting factors like water temperature, dissolved oxygen, free carbon dioxide etc. There are other situations in fishery where relationship among variables can be exploited through regression analysis.

Regression analysis serves three major purposes: (1) description (2) control and (3) prediction. We frequent use equations to summarize or describe a set of data. Regression analysis is helpful in developing such equations. For example we may collect a considerable amount of fish growth data and data on a number of biotic and abiotic factors, and a regression model would probably be a much more convenient and useful summary of those data than a table or even a graph. Besides prediction, regression models may be used for control purposes. A cause and effect relationship may not be necessary if the equation is to be used only for prediction. In this case it is only necessary that the relationships that existed in the original data used to build the regression equation are still valid.

A functional relation between two variables is expressed by a mathematical formula. If X denotes the independent variable and Y the dependent variable, a functional relation is of the form

$$Y = f(X)$$

Given a particular value of X, the function f indicates the corresponding value of Y. A statistical relation, unlike a function is not a perfect one. In general, the observations for a statistical relation do not fall directly on the curve of relationship.

Depending on the nature of the relationships between X and Y, regression approach may be classified into two broad categories viz., linear regression models and nonlinear regression models. The response variable is generally related to other causal variables through some parameters. The models that are linear in these parameters are known as linear models, whereas in nonlinear models parameters are appear nonlinearly. Linear models are generally satisfactory approximations for most regression applications. There are occasions, however, when an empirically indicated or a theoretically justified nonlinear model is more appropriate. In the present lecture we shall consider fitting of linear models only.

17.2 Linear Regression Models

We consider a basic linear model where there is only one predictor variable and the regression function is linear. Model with more than one predictor variable is straight forward. The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

Where Y_i is the value of the response variable in the i^{th} trial β_0 and β_1 are parameters, X_i is a known constant, namely, the value of the predictor variable in the i^{th} trial, ε_i is a random error term with mean zero and variance σ^2 and ε_i and ε_j are uncorrelated so that their covariance is zero.

Regression model (1) is said to be simple, linear in the parameters, and linear in the predictor variable. It is "simple" in that there is only one predictor variable, "linear in the parameters" because no parameters appears as an exponent or its multiplied or divided by another parameter, and "linear in predictor variable" because this variable appears only in the first power. A model that is linear in the parameters and in the predictor variable is also called first order model.

17.2.1 Meaning of Regression Parameters

The parameters β_0 and β_1 in regression model (1) are called regression coefficients, β_1 is the slope of the regression line. It indicates the change in the mean of the probability distribution of Y per unit increase in X. The parameter β_0 in Y intercept of the regression line. When the scope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = 0$. When the scope of the model does not cover $X = 0$, β_0 does not have any particular meaning as a separate term in the regression model.

17.2.2 Method of Least Squares

To find "good" estimates of the regression parameters β_0 and β_1, we employ the method of least squares. For each observations (X_i, Y_i) for each case, the method of least squares considers the deviation of Y from its expected value, $Y_i - \beta_0 - \beta_1 X_i$. In particular, the method of least squares requires that we consider the sum of the n squared deviations. This criterion is denoted by Q:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2 \qquad (2)$$

According to the method of least squares, the estimators of β_0 and β_1 are those values b_0 and b_1, respectively, that minimize the criterion Q for the given observations.

Using the analytical approach, it can be shown for regression model (1) that the values of b_0 and b_1 that minimizes Q for any particular set of sample data are given by the following simultaneous equations:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 .$$

These two equations are called normal equations and can be solved for b_0 and b_1:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$b_0 = \frac{1}{n}(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i) = \overline{Y} - b_1 \overline{X} ,$$

where $\overline{X}$ and $\overline{Y}$ are the means of the X_i and the Y_i observations, respectively.

17.2.3 Properties of Fitted Regression Line

Once the parameters estimates are obtained, the fitted line would be

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (3)$$

The ith residual is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$, i.e., $e_i = Y_i - \hat{Y}_i$.

The estimated regression line (3) fitted by the method of least squares has a number of properties worth noting.

1. The sum of the residuals is zero, $\sum_{i=1}^{n} e_i = 0$.
2. Sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$ is a minimum.
3. Sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$, $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$.
4. Sum of the weighted residuals is zero, weighted by the level of the predictor variable in the i^{th} trial: $\sum_{i=1}^{n} X_i e_i = 0$.
5. Sum of the weighted residuals is zero, weighted by the fitted value of the response variable in the i^{th} trial: $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$.
6. The regression line always goes through the points $(\overline{X}, \overline{Y})$.

17.2.4 Estimation of Error Term Variance σ^2

The variance σ^2 of the error terms ε_i in regression model (1) needs to be estimated to obtain an indication of the variability of the probability distribution of *Y*. In addition, a variety of inferences concerning the regression function and the prediction of *Y* require an estimate of σ^2.

Denote by SSE = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$, is the error sum of squares or residual sum of squares. Then an estimate of σ^2 is given by,

$$\hat{\sigma}^2 = \frac{SSE}{n-p},$$

(4)

where *p* is the total number of parameters involved in the model. We also denote this quantity by MSE.

17.2.5 Inferences in Linear Models

Frequently, we are interested in drawing inferences about β_1, the slope of the regression line. At times, tests concerning β_1 are of interest, particularly one of the form:

$$H_0 = \beta_1 = 0$$
$$H_1 = \beta_1 \neq 0$$

The reason for interest in testing whether or not $\beta_1 = 0$ is that, when $\beta_1 = 0$, there is no linear association between *Y* and *X*. For normal error regression model, the condition $\beta_1 = 0$ implies even more than no linear association between *Y* and *X*. $\beta_1 = 0$ for the normal error regression model implies not only that there is no linear association between *Y* and *X* but also that there is no relation of any kind between *Y* and *X*, since the probability distribution of *Y* are then identical at all levels of *X*.

An explicit test of the alternatives is based on the test statistic:

$$t = \frac{b_1}{s(b_1)},$$

where $s(b_1)$ is the standard error of b_1 and calculated as $s(b_1) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$.

The decision rule with this test statistic when controlling level of significance at α is

if $|t| \leq t(1-\alpha/2; n-p)$, conclude H_0,

if $|t| > t(1-\alpha/2; n-p)$, conclude H_1.

Similarly testing for other parameters can be carried out.

17.2.6 Prediction of New Observations

The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of X for the new trial as X_h and the new observation on Y as Y_h. Of course, we assume that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation.

The distinction between estimation of the mean response, and prediction of a new response, is basic. In the former case, we estimate the mean of the distribution of Y. In the present case, we predict an individual outcome drawn from the distribution of Y. Of course, the great majority of individual outcomes deviate from the mean response, and this must be taken into account by the procedure for predicting $Y_{h(new)}$. We denote by $\hat{Y}_h$, the predicted new observation and by $\sigma^2(\hat{Y}_h)$ the variance of $\hat{Y}_h$. An unbiased estimator of $\sigma^2(\hat{Y}_h)$ is given by $\hat{\sigma}^2(\hat{Y}_h) = \hat{\sigma}^2 + s^2(\hat{Y}_h)$, where $s^2(\hat{Y}_h)$ is the estimate of variance of prediction at X_h and given by

$$s^2(\hat{Y}_h) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right). \tag{5}$$

Confidence interval of $\hat{Y}_h$ can be constructed by using t-statistic namely,

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)\ \sigma^2(\hat{Y}_h).$$

17.2.7 Measure of Fitting, R^2

There are times when the degree of linear association is of interest in its right. Here we describe one descriptive measure that is frequently used in practice to describe the degree of linear association between Y and X.

Denote by SSTO = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, total sum of squares which measures the variation in the observation Y_i, or the uncertainty in predicting Y, when no account of the predictor variable X is taken. Thus SSTO is a measure of uncertainty in predicting Y when X is not considered. Similarly, SSE measures the variation in the Yi when a regression model utilizing the predictor variable X is employed. A natural measure of the effect of X in reducing the variation in Y, i.e., in reducing the uncertaintity in predicting Y, is to express the reduction in variation (SSTO-SSE=SSR) as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{6}$$

The measure R^2 is called coefficient of determination, $0 \le R^2 \le 1$. In practice R^2 is not likely to be 0 or 1 but somewhere between these limits. The closer it is to 1, the greater is said to be the degree of linear association between X and Y.

17.2.8 Diagnostics and Remedial Measures

When a regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application, any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this section we discuss some simple graphic methods for studying the appropriateness of a model, as well as some remedial measures that can be helpful when the data are not in accordance with the conditions of the regression model.

17.2.8.1 Departures From Model to be Studied

We shall consider following six important types of departures from linear regression model with normal errors:

(i) The linearity of regression function.
(ii) The constancy of error variance.
(iii) The independency of error terms.
(iv) Presence of one or a few outlier observations.
(v) The normal distribution of error terms.
(vi) One or several important predictor variables have been omitted from the model.
(vii) Presence of multicollinearity.

17.2.8.2 Graphical Tests for Model Departures Nonlinearity of Regression Model

Whether a linear regression function is appropriate for the data being analyzed can be studied from a residual plot against the predictor variable or equivalently from a residual plot against the fitted values.

Figure 1(a) shows a prototype situation of the residual plot against X when a linear regression model is appropriate. The residuals then fall within a horizontal band centred around 0, displaying no systematic tendencies to be positive and negative.

Figure 1(b) shows a prototype situation of a departure from the linear regression model that indicates the need for a curvilinear regression function. Here the residuals tend to vary in a systematic fashion between being positive and negative.

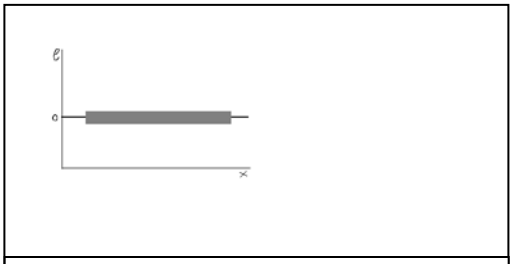


Fig. 1(a)

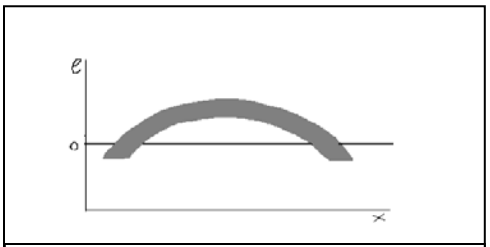


Fig. 1(b)

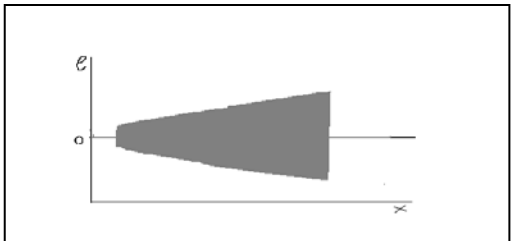


Fig. 1(c)

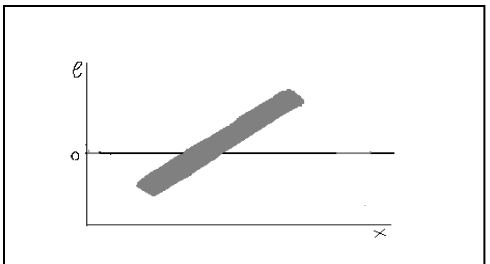


Fig. 1(d)

Nonconstancy of Error Variance

Plots of residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant

The prototype plot in Figure 1(a) exemplifies residual plots when error term variance is constant. Figure 1(c) shows a prototype picture of residual plot when the error variance increases with X. In many biological science applications, departures from constancy of the error variance tend to be of the "meghaphone" type.

Presence of Outliers

Outliers are extreme observations. Residual outliers can be identified from residual plots against X or $\hat{Y}$.

Nonindependence of Error Terms

Whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographical areas, it is good idea to prepare a sequence plot of the residuals. The purpose of plotting the residuals against time or some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence.

A prototype residual plot showing a time related trend effect is presented in Figure 1(d), which portrays a linear time related trend effect. When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the base line 0.

Nonnormality of Error Terms

Small departures from normality do not create any serious problems. Major departures, on the other hand, should be of concern. The normality of the error terms can be studied informally by examining the residuals in a variety of graphic ways.

Comparison of frequencies: when the number of cases is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 90% of the residuals fall between $\pm 1.645\ \sqrt{MSE}$.

Normal probability plot: Still another possibility is to prepare a normal probability plot of the residuals. Here each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

Omission of Important Predictor Variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response. The purpose of this additional analysis is to determine whether there are any key variables that could provide important additional descriptive and predictive power to the model. The residuals are plotted against the

additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable.

17.2.8.3 Statistical Tests for Model departures

Graphical analysis of residuals is inherently subjective. Nevertheless, subjective analysis of a variety of interrelated residuals plots will frequently reveal difficulties with the model more clearly than particular formal tests.

Tests for Randomness

A run test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specially designed for lack of randomness in least squares residuals, is the

Durbin-Watson test:

The Durbin-Watson test assumes the first order autoregressive error models. The test consists of determining whether or not the autocorrelation coefficient (ρ, say) is zero. The usual test alternatives considered are:

$H_0 : \rho = 0$

$H_0 : \rho > 0$

The Durbin-Watson test statistic D is obtained by using ordinary least squares to fit the regression function, calculating the ordinary residuals: $e_t = Y_t - \hat{Y}_t$, and then calculating the statistic:

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{n \sum_{t=1} e_t^2}$$

(7)

Exact critical values are difficult to obtain, but Durbin-Watson have obtained lower and upper bound d_L and d_U such that a value of D outside these bounds leads to a definite decision. The decision rule for testing between the alternatives is:

if $D > d_U$, conclude H_0

if $D < d_L$, conclude H_1

if $d_L \le D \le d_U$, test is inconclusive.

Small value of D lead to the conclusion that $\rho > 0$.

Tests for Normality

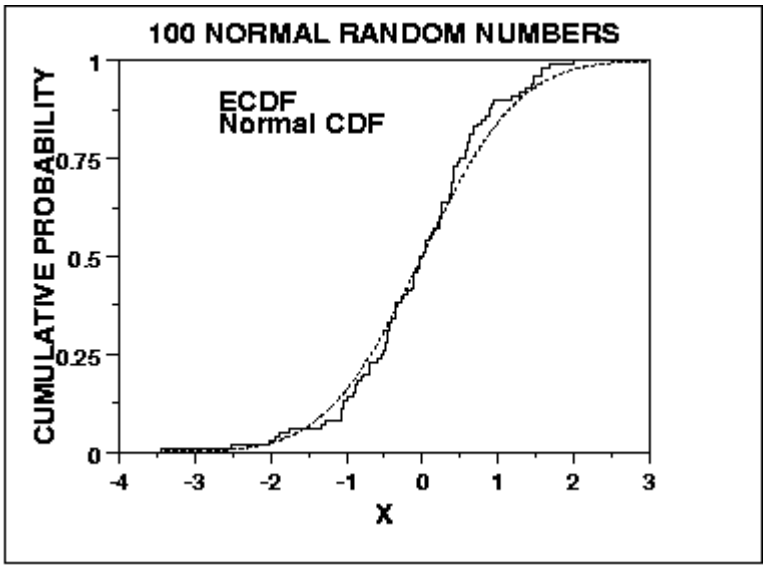
Correlation Test for Normality: In addition to visually assessing the appropriate linearity of the points plotted in a normal probability plot, a formal test for normality of the error terms can be conducted by calculating the coefficient of correlation between

residuals e_i and their expected values under normality. A high value of the correlation coefficient is indicative of normality.

Kolmogorov-Smirnov test : The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution. The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given *N ordered* data points $Y_1, Y_2, ..., Y_N$, the ECDF is defined as

$$E_N = n(i) / N ,$$

where $n(i)$ is the number of points less than Y_i and the Y_i are ordered from smallest to largest value. This is a step function that increases by 1/*N* at the value of each ordered data point. The graph below is a plot of the empirical distribution function with a normal cumulative distribution function for 100 normal random numbers. The K-S test is based on the maximum distance between these two curves.



An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test has several important drawbacks:

1. It only applies to continuous distributions.
2. It tends to be more sensitive near the center of the distribution than at the tails.
3. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

Due to limitations 2 and 3 above, many analysts prefer to use the Anderson-Darling goodness-of-fit test.

The Kolmogorov-Smirnov test is defined by:

H_0: The data follow a specified distribution

H_1: The data do not follow the specified distribution

The Kolmogorov-Smirnov test statistic is defined as

$$D = \max_{1 \le i \le N} (F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)) \quad (9)$$

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data).

The hypothesis regarding the distributional form is rejected if the test statistic, ***D***, is greater than the critical value obtained from a table. There are several variations of these tables in the literature that use somewhat different scalings for the K-S test statistic and critical regions. These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated.

Anderson-Darling Test: The Anderson-Darling test is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Currently, tables of critical values are available for the normal, lognormal, exponential, Weibull, extreme value type I, and logistic distributions.

The Anderson-Darling test is defined as:

H_0: The data follow a specified distribution.

H_1: The data do not follow the specified distribution

The Anderson-Darling test statistic is defined as $A^2 = -N - S$,

where, $S = \sum_{i=1}^{N} \frac{(2i-1)}{N} [\ln F(Y_i) + \ln(1 - F(Y_{N+1-i})]$

(10)

F is the cumulative distribution function of the specified distribution. Note that the Y_i are the *ordered* data. The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas are available in literature for a few specific distributions (normal, lognormal, exponential, Weibull, logistic, extreme value type 1). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A, is greater than the critical value.

Tests for Constancy of Error Variance

Modified Levene Test : The test is based on the variability of the residuals. Let e_{i1} denotes the i^{th} residual for group 1 and e_{i2} denotes the i^{th} residual for group 2. Also we denote n_1 and n_2 to denote the sample sizes of the two groups, where: $n_1 + n_2 = n$.

Further, we shall use $\tilde{e}_1$ and $\tilde{e}_2$ to denote the medians of the residuals in the two groups. The modified Levene test uses the absolute deviations of the residuals around their median, to be denoted by d_{i1} and d_{i2}:

$$\bar{d}_{i1} = |e_{i1} - \tilde{e}_1|, \quad \bar{d}_{i2} = |e_{i2} - \tilde{e}_2|$$

With this notation, the two-sample t test statistic becomes:

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (11)$$

Where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2}, respectively, and the pooled variance s^2 is:

$$s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n-2}.$$

If the error terms have constant variance and n_1 and n_2 are not too small, t_L^* follows approximately the t distribution with n-2 degrees of freedom. Large absolute values of t_L^* indicate that the error terms do not have constant variance.

White Test In statistics, the White test is a statistical test that establishes whether the residual variance of a variable in a regression model is constant: that is for homoscedasticity. This test, and an estimator for heteroscedasticity-consistent standard errors, were proposed by Halbert White in 1980. These methods have become extremely widely used, making this paper one of the most cited articles in economics. To test for constant variance one undertakes an auxiliary regression analysis: this regresses the squared residuals from the original regression model onto a new set of regressors, which the contains the original regressors, the cross-products of the regressors and the squared regressors. One then inspects the R^2. The LM test statistic is the product of the R^2 value and sample size:

$$LM = n.R^2 \qquad (13)$$

This follows a chi-square distribution, with degrees of freedom equal to the number of estimated parameters (in the auxiliary regression) minus one.

Tests for Outlying Observations

(i) *Elements of* ***Hat Matrix*** **:** The Hat matrix is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{X}$ is the matrix for explanatory variables. The larger values reflect data points are outliers.

(ii) ***WSSD$_i$***: $WSSD_i$ is an important statistic to locate points that are remote in *x*-space. $WSSD_i$ measures the weighted sum of squared distance of the i^{th} point from the center of the data. Generally if the $WSSD_i$ values progress smoothly from small to large, there are probably no extremely remote points. However, if there is a sudden jump in the magnitude of $WSSD_i$, this often indicates that one or more extreme points are present.

(iii) ***Cook's D_i:*** Cook's D_i is designed to measure the shift in $\hat{y}$ when i[th] obsevation is not used in the estimation of parameters. D_i follows approximately $F_{(p,n-p-1)}(1\text{-}\alpha)$. Lower 10% point of this distribution is taken as a reasonable cut off (more conservative users suggest the 50% point). The cut off for D_i can be taken as $\frac{4}{n}$.

(iv) ***$DFFITS_i$:*** *DFFIT* is used to measure difference in i[th] component of $\left(\hat{y} - \hat{y}_{(i)}\right)$. It is suggested that $DFFITS_i \geq 2\left(\frac{p+1}{n}\right)^{1/2}$ may be used to flag off influential observations.

(v) ***$DFBETAS_{j(i)}$***: Cook's D_i reveals the impact of i[th] observation on the entire vector of the estimated regression coefficients. The influential observations for individual regression coefficient are identified by $DFBETAS_{j(i)}, j = 1,2,...,p+1,$ where each $DFBETAS_{j(i)}$ is the standardized change in b_j when the i[th] observation is deleted.

(vi) ***$COVRATIO_i$:***The impact of the i[th] observation on variance-covariance matrix of the estimated regression coefficients is measured by the ratio of the determinants of the two variance-covariance matrices. Thus, COVRATIO reflects the impact of the i[th] observation on the precision of the estimates of the regression coefficients. Values near 1 indicate that the i[th] observation has little effect on the precision of the estimates. A value of COVRATIO greater than 1 indicates that the deletion of the i[th] observation decreases the precision of the estimates; a ratio less than 1 indicates that the deletion of the observation increases the precision of the estimates. Influential points are indicated by $\left|COVRATIO_i - 1\right| > \frac{3(p+1)}{n}$.

(vii) ***$FVARATIO_i$:*** The statistic detects change in variance of $\hat{y}_i$ when an observation is deleted. A value near 1 indicates that the i[th] observation has negligible effect on variance of y_i. A value greater than 1 indicates that deletion of the i[th] observation decreases the precision of the estimates, a value less than one increases the precision of the estimates.

Tests for Multicollinearity

The use and interpretation of a multiple regression model depends implicitly on the assumption that the explanatory variables are not strongly interrelated. In most regression applications the explanatory variables are not orthogonal. Usually the lack of orthogonality is not serious enough to affect the analysis. However, in some situations the explanatory variables are so strongly interrelated that the regression results are ambiguous. Typically, it is impossible to estimate the unique effects of individual variables in the regression equation. The estimated values of the coefficients are very sensitive to slight changes in the data and to the addition or deletion of variables in the equation. The regression coefficients have large sampling errors which affect both inference and forecasting that is based on the regression model. The condition of severe non-orthogonality is also referred to as the problem of multicollinearity.

The presence of multicollinearity has a number of potentially serious effects on the least squares estimates of regression coefficients as mentioned above. Some of the effects may be easily demonstrated. Multicollinearity also tends to produce least squares estimates b_j that are too large in absolute value.

Detection of Multicollinearity

Let $R = (r_{ij})$ and $R^{-1} = (r^{ij})$ denote simple correlation matrix and its inverse. Let $\lambda_i, i = 1,2,...,p \ (\lambda_p \le \lambda_{p-1} \le ....\lambda_1)$ denote the eigen values of *R*. The following are common indicators of relationships among independent variables.

1. Simple pair-wise correlations $|r_{ij}| = 1$
2. The squared multiple correlation coefficients

$R_i^2 = 1 - \frac{1}{r^{ii}} > 0.9$, where R_i^2 denote the squared multiple correlation coefficients for the regression of x_I on the remaining *x* variables.

3. The variance inflation factors, $VIF_i = r^{ii} > 10$ and
4. eigen values, $\lambda_i = 0$.

The first of these indicators, the simple correlation coefficients between pairs of independent variables r_{ij}, may detect a simple relationship between x_i and x_j. Thus $|r_{ij}| = 1$ implies that the ith and jth variables are nearly proportional.

The second set of indicators, R_i^2, the squared multiple correlation coefficient for the regression of x_i on the remaining *x* variables indicates the degree to which x_i is explained by a linear combination of all of the other input variables.

The third set of indicators, the diagonal elements of the inverse matrix, which have been labeled as the Variance Inflation Factors, VIF_i. The term arises by noting that with standardized data (mean zero and unit sum of squares), the variance of the least squares estimate of the ith coefficient is proportional to r^{ii}, $VIF_i > 10$ is probably based on the simple relation between R_i and VIF_i. That is $VIF_i > 10$ corresponds to $R_i^2 > 0.9$.

17.2.8.4 Overview of Remedial Measures

If the simple regression model (1) is not appropriate for a data set, there are two basic choices:

1. Abandon regression model and develop and use a more appropriate model.
2. Employ some transformation on the data so that regression model (1) is appropriate for the transformed data.

Each approach has advantages and disadvantages. The first approach may entail a more complex model that could yield better insights, but may also lead to more complex procedure for estimating the parameters. Successful use of transformations, on the other hand, lead to relatively simple methods of estimation and may involve fewer parameters than a complex model, an advantage when the sample size is small. Yet transformation may obscure the fundamental interconnections between the variables, though at times they may illuminate them.

Nonlinearity of Regression Function

When the regression function is not linear, a direct approach is to modify regression model (1) by altering the nature of the regression function. For instance, a quadratic regression function might be used.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

or an exponential regression function:

$$Y_i = \gamma_0 \gamma_1^{X_i} + \varepsilon_i .$$

When the nature of the regression function is not known, exploratory analysis that does not require specifying a particular type of function is often useful.

Nonconstancy of Error Variance

When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the method to allow for this and use the method of weighted least squares to obtain the estimates of the parameters.

Transformations is another way in stabilizing the variance. We first consider transformation for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformation on X should be attempted. The reason why transformation on Y may not be desirable here is that a transformation on Y, such as $Y' = \sqrt{Y}$, may materially change the shape of the distribution and may lead to substantially differing error term variance.

Following transformations are generally applied for stabilizing variance.

(1) when the error variance is rapidly increasing $Y' = \log_{10} Y$ or $Y' = \sqrt{Y}$

(2) when the error variance is slowly increasing, $Y' = Y^2$ or $Y' = Exp(Y)$

(3) when the error variance is decreasing, $Y' = 1/Y$ or $Y' = Exp(-Y)$.

Box - Cox Transformations: It is difficult to determine, which transformation of Y is most appropriate for correcting skewness of the distributions of error terms, unequal error variance, and nonlinearity of the regression function. The Box-Cox transformation automatically identifies a transformation from the family of power transformations on Y. The family of power transformations is of the form: $Y' = Y^{\lambda}$, where is a parameter to be determined from the data. Using standard computer programme it can be determined easily.

Nonindependence of Error Terms

When the error terms are correlated, a direct approach is to work with a model that calls for error terms. A simple remedial transformation that is often helpful is to work with first differences.

Nonnormality of Error terms

Lack of normality and non-constant error variance frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms. It is therefore, desirable that the transformation for stabilizing the error variance be utilized first, and then the residuals studied to see if serious departures from normality are still present.

Omission of Important Variables

When residual analysis indicates that an important predictor variable has been omitted from the model, the solution is to modify the model.

Outlying Observations

Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect. On the other hand, outliers may convey significant information, as when an outlier occurs because of an interaction with another predictor omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents in error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstances. When outlying observations are present, use of the least squares and maximum likelihood estimates for regression model (1) may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations. Robust Regression falls under such methods.

Multicollinearity

i) **Collection of additional data:** Collecting additional data has been suggested as one of the methods of combating multicollinearity. The additional data should be collected in a manner designed to break up the multicollinearity in the existing data.

ii) **Model respecification:** Multicollinearity is often caused by the choice of model, such as when two highly correlated regressors are used in the regression equation. In these situations some respecification of the regression equation may lessen the impact of multicollinearity. One approach to respecification is to redefine the regressors. For example, if x_1, x_2 and x_3 are nearly linearly dependent it may be possible to find some function such as $x = (x_1+x_2)/x_3$ or $x = x_1x_2x_3$ that preserves the information content in the original regressors but reduces the multicollinearity.

iii) Ridge Regression: When method of least squares is used, parameter estimates are unbiased. A number of procedures have been developed for obtaining biased estimators

of regression coefficients to tackle the problem of multicollinearity. One of these procedures is ridge regression. The ridge estimators are found by solving a slightly modified version of the normal equations. Each of the diagonal elements of $\mathbf{X'X}$ matrix are added a small quantity.

Example

Table 1

Case	X_{11}	X_{21}	X_{31}	Y_i
1	12.980	0.317	9.998	57.702
2	14.295	2.028	6.776	59.296
3	15.531	5.305	2.947	56.166
4	15.133	4.738	4.201	55.767
5	15.342	7.038	2.053	51.722
6	17.149	5.982	-0.055	60.446
7	15.462	2.737	4.657	60.715
8	12.801	10.663	3.048	37.447
9	17.039	5.132	0.257	60.974
10	13.172	2.039	8.738	55.270
11	16.125	2.271	2.101	59.289
12	14.340	4.077	5.545	54.027
13	12.923	2.643	9.331	53.199
14	14.231	10.401	1.041	41.896
15	15.222	1.220	6.149	63.264
16	15.740	10.612	-1.691	45.798
17	14.958	4.815	4.111	58.699
18	14.125	3.153	8.453	50.086
19	16.391	9.698	-1.714	48.890
20	16.452	3.912	2.145	62.213
21	13.535	7.625	3.851	45.625
22	14.199	4.474	5.112	53.923
23	15.837	5.753	2.087	55.799
24	16.565	8.546	8.974	56.741
25	13.322	8.589	4.011	43.145
26	15.949	8.290	-0.248	50.706

Table 2: Indicators of Influential Observations

Case	r_i	t_i	$t_I^*=s.t/s_i$	h_{ii}	D_i	$WSSD_i$
1	0.460	0.289	0.281	0.215	0.005	39*
2	1.253	0.732	0.724	0.093	0.013	12
3	0.377	0.215	0.210	0.048	0.001	1
4	0.044	0.025	0.026	0.042	0.000	1
5	-0.256	-0.146	-0.141	0.053	0.000	3
6	1.010	0.611	0.602	0.155	0.017	20
7	0.389	0.226	0.221	0.081	0.001	7
8	0.132	0.088	0.086	0.301	0.001	41
9	0.432	0.262	0.256	0.155	0.003	18
10	0.589	0.355	0.347	0.147	0.005	23
11	-3.302	-2.021	-2.193	0.173	0.214	14
12	-0.406	-0.232	-0.226	0.053	0.001	3
13	0.194	0.118	0.117	0.163	0.001	24
14	-0.268	-0.164	-0.161	0.175	0.001	23
15	0.802	0.476	0.469	0.122	0.007	15
16	-0.482	-0.295	-0.289	0.177	0.005	26
17	3.756	2.134	2.343	0.041	0.048	0
18	-6.072	-3.589	-5.436	0.114	0.412	8
19	-1.198	-0.727	-0.719	0.160	0.025	24
20	1.126	0.666	0.658	0.114	0.014	11
21	0.449	0.266	0.259	0.119	0.003	12
22	0.791	0.453	0.444	0.055	0.003	3
23	-0.060	-0.035	-0.032	0.059	0.000	3
24	0.574	1.181	1.188	0.927	4.409	19
25	0.268	0.163	0.158	0.159	0.001	19
26	-0.606	-0.356	-0.350	0.101	0.004	11

Table 3: Indicators of Influential Observations

Case	Cov Ratio	Dffits	Intercept	X1	X2	X3
				DFBETAS		
1	1.512	0.148	0.056	-0.053	-0.006	0.006
2	1.203	0.232	0.062	-0.042	-0.042	-0.050
3	1.254	0.047	-0.005	0.010	-0.008	-0.007
4	1.257	0.005	0.000	0.000	-0.001	0.000
5	1.267	-0.033	-0.001	-0.001	-0.006	0.006
6	1.331	0.258	-0.095	0.132	-0.042	-0.050
7	1.299	0.068	-0.005	0.015	-0.036	-0.005
8	1.721	0.057	0.027	-0.034	0.026	-0.006
9	1.408	0.109	-0.030	0.048	-0.035	-0.031
10	1.380	0.144	0.058	-0.058	-0.041	0.016
11	0.639	-1.004	-0.154	-0.045	0.776	0.525
12	1.260	-0.054	-0.017	0.014	0.014	0.000
13	1.435	0.051	0.017	-0.19	-0.004	0.013
14	1.452	-0.074	-0.026	0.031	-0.35	0.015
15	1.315	0.175	-0.008	0.033	-0.105	0.002
16	1.441	-0.134	-0.014	0.014	-0.044	0.047
17	0.496	0.482	0.061	-0.17	-0.107	-0.046
18	0.410	-1.945	0.362	-0.308	-0.220	-1.177
19	1.301	-0.341	0.031	-0.045	-0.080	0.094
20	1.252	0.236	-0.055	0.097	-0.105	-0.051
21	1.350	0.095	0.054	-0.061	0.024	-0.018
22	1.228	0.108	0.052	-0.048	-0.028	-0.020
23	1.279	-0.008	0.001	-0.002	0.001	0.002
24	12.715	4.230	-3.642	3.276	3.180	3.934
25	1.426	0.069	0.031	-0.039	0.029	-0.003
26	1.309	-0.117	0.000	-0.007	-0.016	0.043

Table 4: Regression Coefficients and Summary Statistics

Description	b_0	b_1	b_2	b_3	s	R^2	Max VIF	Min e.v.	Max R_i^2
All Data (n=26)	8.11	3.56	-1.63	0.34	1.80	0.94	2.82	0.210	0.65
Delete (11, 17, 18)	7.17	3.66	-1.79	0.40	0.51	0.99	2.85	0.210	0.65
Delete (24)	30.91	2.39	-2.14	-0.36	1.78	0.94	30.64	0.017	0.97
Delete (11, 17, 18, 24)	24.27	2.79	-2.11	-0.16	0.50	0.99	171.90	0.003	0.99
Ridge k=0.05 (n=22)	14.28	3.22	-1.73	0.25	0.66	0.99	10.20	0.053	0.90
Delete X3 (n=22)	19.50	3.03	-2.00		0.49	0.99	1.02	0.863	0.02

Some Selected References

Belsley, D.A., Kuh, E. and Welsch, R.E. (2004). Regression diagnostics – Identifying influential data and sources of collinearity, New York.: Wiley

Barnett, V. and Lewis, T. (1984). Outliers in Statistical Data, New York: Wiley Ltd.

Chatterjee, S. and Price, B (1977). Regression analysis by example, New York: John Wiley & sons

Draper, N.R. and Smith, H. (1998). Applied Regression analysis, New York: Wiley Eastern Ltd.

Kleinbaum, D.G. & Kupper, L.L. (1978). Applied Regression analysis and other multivariate methods, Massachusetts: Duxbury Press

Montgomery, D.C., Peck, E. and Vining, G. (2003). Introduction to linear regression analysis, 3rd Edition, New York: John Wiley and Sons Inc.

LOGISTIC REGRATION

LALMOHAN BHAR

Indian Agricultural Statistics Research Institute, New Delhi-110012
lmbhar@gmail.com

18.1 Introduction

Regression analysis is a method for investigating functional relationships among variables. The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variables. Most of the variables in this model are quantitative in nature. Estimation of parameters in this regression model is based on four basic assumptions. First, response or dependent variable is linearly related with explanatory variables. Second, model errors are independently and identically distributed as normal variable with mean zero and common variance. Third, independent or explanatory variables are measured without errors. The last assumption is about equal reliability of observations.

In case, our response variable in model is qualitative in nature, then probabilities of falling this response variable in various categories can be modeled in place of response variable itself, using same model but there are number of constraints in terms of assumptions of multiple regression model. First, since the range of probability is between 0 and 1, whereas, right hand side function in case of multiple regression models is unbounded. Second, error term of the model can take only limited values and error variance are not constants but depends on probability of falling response variable in a particular category.

Generally, conventional theory of multiple linear regression (MLR) analysis has been applied for a quantitative response variable, while for the qualitative response variable or more specifically for binary response variable it is better to consider alternative models. As for example, considering following scenarios:

- A pathologist may be interested whether the probability of a particular disease can be predicted using tillage practice, soil texture, date of sowing, weather variables etc. as predictor or independent variables.

- An economist may be interested in determining the probability that an agro-based industry will fail given a number of financial ratios and the size of the firm (i.e. large or small).

Usually discriminant analysis could be used for addressing each of the above problems. However, because the independent variables are mixture of categorical and continuous variables, the multivariate normality assumption may not hold. Structural relationship among various qualitative variables in the population can be quantified using number of alternative techniques. In these techniques, primary interest lies on dependent factor

which is dependent on other independent factors. In these cases the most preferable technique is either probit or logistic regression analysis as it does not make any assumptions about the distribution of the independent variables. The dependent factor is known as response factor. In this model building process, various log odds related to response factors are modelled. As a special case, if response factor has only two categories with probabilities p_1 and p_2 respectively then the odds of getting category one is (p_1 / p_2). If log (p_1 / p_2) is modelled using ANalysis Of VAriance (ANOVA) type of model, it is called logit model. Again, if the same model is being treated as regression type model then it is called logistic regression model. In a real sense, logit and logistic are names of transformations. In case of logit transformation, a number p between values 0 and 1 is transformed with log {p/(1-p)}, whereas in case of logistic transformation a number x between - ∞ to + ∞ is transformed with $\{e^x /(1 + e^x)\}$ function. It can be seen that these two transformation are reverse of each other i.e. if logit transformation is applied on logistic transformation function, it provides value x and similarly, if logistic transformation is applied to logit transformation function it provides value p. Apart from logit or logistic regression models, other techniques such as CART i.e. Classification And Regression Trees can also be used to address such classification problems. A good account of literature on logistic regression are available, to cite a few, Fox(1984), Klienbaum (1994) etc.

18.2 Violation of Assumptions of Linear Regression Model when Response is Qualitative

Linear regression is considered in order to explain the constraints in using such model when the response variable is qualitative. Consider the following simple linear regression model with single predictor variable and a binary response variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , i = 1, 2, \ldots, n$$

where the outcome Y_i is binary (taking values 0,1), $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and are independent and n is the number of observations.

Let π_i denote the probability that Y_i =1 when X_i = x, i.e.

$$\pi_i = P(Y_i = 1 | X_i = x) = P(Y_i = 1)$$

thus $P(Y_i = 0) = 1-\pi_i$.

Under the assumption $E(\varepsilon_i) = 0$, the expected value of the response variable is

$$E(Y_i) = 1.(\pi_i) + 0.(1-\pi_i) = \pi_i$$

If the response is binary, then the error terms can take on two values, namely,

$$\varepsilon_i = 1 - \pi_i \qquad \text{when } Y_i = 1$$

$$\varepsilon_i = -\pi_i \qquad \text{when } Y_i = 0$$

Because the error is dichotomous (discrete), normality assumption is violated. Moreover, the error variance is given by:

$$V(\varepsilon_i) = \pi_i(1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2$$
$$= \pi_i(1-\pi_i)$$

It can be seen that variance is a function of π_i's and it is not constant. Therefore the assumption of homoscadasticity (equal variance) does not hold.

18.3 Binary Logistic regression

Logistic regression is normally recommended when the independent variables do not satisfy the multivariate normality assumption and at the same time the response variable is qualitative. Situations where the response variable is qualitative and independent variables are mixture of categorical and continuous variables, are quite common and occur extensively in statistical applications in agriculture, medical science etc. The statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model, developed primarily by a researcher named Cox during the late 1950s. Processes producing sigmoidal or elongated S-shaped curves are quite common in agricultural data. Logistic regression models are more appropriate when response variable is qualitative and a non-linear relationship can be established between the response variable and the qualitative and quantitative factors affecting it. It addresses the same questions that discriminant function analysis and multiple regression do but with no distributional assumptions on the predictors. In logistic regression model, the predictors need not have to be normally distributed, the relationship between response and predictors need not be linear or the observations need not have equal variance in each group etc. A good account on logistic regression can be found in Fox (1984) and Kleinbaum (1994).

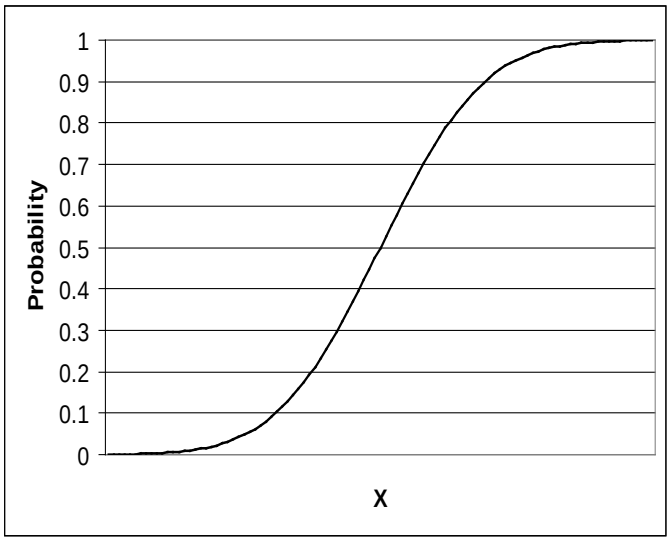
The problem of non-normality and heteroscadasticity (see section 2) leads to the non applicability of least square estimation for the linear probability model. Weighted least square estimation, when used as an alternative, can cause the fitted values not constrained to the interval (0, 1) and therefore cannot be interpreted as probabilities. Moreover, some of the error variance may come out to be negative. One solution to this problem is simply to constrain π to the unit interval while retaining the linear relation between π and regressor X within the interval. Thus

$$\pi = \begin{cases} 0 & ,\beta_0 + \beta_1 X < 0 \\ \beta_0 + \beta_1 X & ,0 \le \beta_0 + \beta_1 X \le 1 \\ 1 & ,\beta_0 + \beta_1 X > 1 \end{cases}$$

However, this constrained linear probability model has certain unattractive features such as abrupt changes in slope at the extremes 0 and 1 making it hard for fitting the same on data. A smoother relation between π and X is generally more sensible. To correct this problem, a positive monotone (i.e. non-decreasing) function is required to transform ($\beta_0 + \beta_1 x_i$) to unit interval. Any cumulative probability distribution function (CDF) P, meets this requirement. That is, respecify the model as $\pi_i = P(\beta_0 + \beta_1 x_i)$. Moreover, it is advantageous if P is strictly increasing, for then, the transformation is one-to-one, so that model can be rewritten as $P^{-1}(\pi i) = (\beta 0 + \beta 1 xi)$, where P^{-1} is the inverse of the CDF P. Thus the non-linear model for itself will become both smooth and symmetric, approaching $\pi = 0$ and $\pi = 1$ as asymptotes. Thereafter maximum likelihood method of estimation can be employed for model fitting.

18.3.1 Properties of Logistic Regression Model

The Logistic response function resembles an S-shape curve, a sketch of which is given in the following figure. Here the probability π initially increases slowly with increase in X, and then the increase accelerates, finally stabilizes, but does not increase beyond 1.



The shape of the S-curve can be reproduced if the probabilities can be modeled with only one predictor variable as follows:

$\pi = P(Y=1|X= x) = 1/(1+e^{-z})$

where $z = \beta_0 + \beta_1 x$, and e is the base of the natural logarithm. Thus for more than one (say r) explanatory variables, the probability π is modeled as

$\pi = P(Y=1|X_1= x_1 ... X_r = x_r)$

$=1/(1+e^{-z})$

where $z = \beta_0+\beta_1 x_1+...+\beta_r x_r$.

This equation is called the logistic regression equation. It is nonlinear in the parameters β_0, $\beta_1 \ldots \beta_r$. Modeling the response probabilities by the logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression. The method of estimation generally used is the maximum likelihood estimation method.

To explain the popularity of logistic regression, let us consider the mathematical form on which the logistic model is based. This function, called f (z), is given by

$$f(z) = 1/ (1+e^{-z}) , -\infty < z < \infty$$

Now when $z = -\infty$, f (z) =0 and when $z = \infty$, f (z) =1. Thus the range of f (z) is 0 to1. So the logistic model is popular because the logistic function, on which the model is based, provides

- Estimates that lie in the range between zero and one.
- An appealing S-shaped description of the combined effect of several explanatory variables on the probability of an event.

18.3.2 Maximum Likelihood Method of Estimation of Logistic Regression

For simplicity, a simple binary logistic regression model with only one explanatory variable is considered. The model is given by

$$\pi_i = P(Y_i=1|X_i=x_i) = 1/(1+e^{-z})$$

where $z = \beta_0 + \beta_1 x_i$, and e is the base of the natural logarithm. The binary response variable Y_i takes only two values (say 0 and 1). Since each Y_i observation is an ordinary Bernoulli random variable, where:

$$P(Y_i = 1) = \pi_i$$

and $$P(Y_i = 1) = 1 - \pi_i,$$

the probability distribution function is represented as follows:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \quad , Y_i = 0,1; i=1,2,...,n$$

Since ^{Y_i}'s are independent, then the joint probability density function is:

$$g(Y_1...Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$\log_e g(Y_1...Y_n) = \log_e \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$= \sum_{i=1}^{n} Y_i \log_e \pi_i/(1-\pi_i)$$

Since $E(Y_i) = \pi_i$, for a binary variable it follows that

$$1-\pi_i = \left[1+e^{-(\beta_0+\beta_1 X_i)}\right]^{-1}$$

Then,

$$\log_e\left[\pi_i/(1-\pi_i)\right] = \beta_0+\beta_1 X_i$$

Hence the log likelihood function can be expressed as follows:

$$\log_e L(\beta_0,\beta_1) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 X_i) - \sum_{i=1}^{n} \log_e\left[1+e^{-(\beta_0+\beta_1 X_i)}\right]$$

where $L(\beta_0, \beta_1)$ replaces $g(Y_1 \ldots Y_n)$ to show explicitly that the function can now be viewed as the likelihood function of the parameters to be estimated, given the sample observations.

The maximum likelihood estimates β_0 and β_1 in the simple logistic regression model are those values of β_0 and β_1 that maximize the log-likelihood function. No closed-form solution exists for the values of β_0 and β_1 that maximize the log-likelihood function. Computer intensive numerical search procedures are therefore required to find the

maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Standard statistical software programs such as SAS (PROC LOGISTIC), SPSS (Analyze- Regression-Binary Logistic) provide maximum likelihood estimates for logistic regression. Once these estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are found, by substituting these values into the response function the fitted response function, say, $\hat{\pi}_i$, can be obtained. The fitted response function is as follows:

$$\hat{\pi}_i = \left(\frac{1}{1+e^{-\left(\hat{\beta}_0+\hat{\beta}_1 X_i\right)}} \right)$$

When log of the odds of occurrence of any event is considered using a logistic regression model, it becomes a case of logit analysis. Here the thus formed logit model will have its right hand side as a linear regression equation.

18.4 Model Validation

The model validation can be done by employing various tests on any fitted logistic regression model. The tests related to the significance of the estimated parameters, goodness of fit and predictive ability of the models are discussed subsequently.

Wald, Likelihood ratio and Score tests are three commonly used tests for testing the overall significance of the logistic regression model.

18.4.1 Wald test

Let $\hat{\boldsymbol{\beta}}$ be the vector of parameter estimates obtained. Let a set of restrictions be imposed in the form of a hypothesis H_0: $\beta = 0$. If the restrictions are valid, then at least approximately $\hat{\boldsymbol{\beta}}$ should satisfy them. The Wald statistic is then defined as

$$W = \hat{\boldsymbol{\beta}}' \left[\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) \right]^{-1} \hat{\boldsymbol{\beta}}$$

Under H_0, in large samples, W has a Chi-square distribution with degrees of freedom equal to the number of restrictions imposed.

18.4.2 Likelihood Ratio (LR) Test

The LR statistic is defined as two times the logarithm of the ratio of the likelihood functions of two different models evaluated at their MLEs. The LR statistic is used for testing the overall significance of the model. Assuming that there are r_1 variables in the model under consideration which can be considered as the full model, based on the MLEs of the full model, L (full) is calculated. Beside this, the likelihood function L (reduced) is calculated for the constant only model. The LR statistic is then defined as: $LR = -2\left[\ln\{L(\text{reduced})\} - \ln\{L(\text{full})\} \right]$. LR is asymptotically distributed as Chi-square with degrees of freedom equal to the difference between the number of parameters estimated in the two models.

18.4.3 Goodness of Fit in Logistic Regression

Among various testing problems, goodness of fit is one of the most important aspects in the context of the logistic regression analysis for testing whether the model fitted well or not. Hosmer-Lemeshow goodness-of-fit test is one of the most common tools conveniently used in logistic regression analysis. This test is performed for a binary logistic regression model by first sorting the observations in increasing order of their estimated event probabilities. The observations are then divided into approximately ten groups on the basis of the estimated probabilities. Comparison between the numbers actually in each group (observed) to the numbers predicted by the logistic regression model (predicted) is carried out subsequently. The number of groups may be smaller than 10 if there are fewer than 10 patterns of explanatory variables. There must be at least three groups in order that the Hosmer-Lemeshow statistic can be computed.

The Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating the Pearson chi-square statistic from the (2×g) table of observed and expected frequencies where g is the number of groups.

The statistic is written as:

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{\left(O_i - N_i \bar{\hat{\pi}}_i\right)^2}{N_i \bar{\hat{\pi}}_i \left(1 - \bar{\hat{\pi}}_i\right)} \sim \chi^2_{g-2}$$

Where

O_i = the observed number of events in the i^{th} group

N_i = the number of subjects in i^{th} group

and $\bar{\hat{\pi}}_i$ = the average estimated probability of an event in the i^{th} group.

18.4.4 Predictive ability of the model

Once models are fitted and relevant goodness of fit measures are employed, judging the predictive ability of the model can be done. In logistic regression modeling setup, predictive ability of models can be judged by employing various measures such as Somers'D, Gamma, Kendall's Tau (Tau-a) and c. Here two measures viz. Gamma and Somers'D have been discussed. Gamma statistic is the simplest one. The measures Gamma, and Somers'D are based on concordance and discordance. By observing the ordering of two subjects on each of two variables, one can classify the pair of subjects as concordant or discordant. The pair is concordant if the subject ranking is higher on both the variables. The pair is discordant if the subject ranking is higher on one variable and lower on the other. The pair is tied if the subjects have the same prediction on both of the variables.

The Gamma is defined as $\dfrac{N_s - N_d}{N_s + N_d}$

where N_s is the number of same pairs and N_d the number of different pairs. Gamma ignores all tied pairs of cases. It therefore may exaggerate the "actual" strength of association. Gamma lies between -1 to1.

The Somers'D is a simple modification of gamma. Unlike gamma, the Somers' D includes tied pairs in one way or another. Somers'D is defined as

$$\frac{N_s - N_d}{N_s + N_d + T_y}$$

where T_y is the number of pairs tied on the dependent variable, Y. Somers' d ranges from -1.0 (for negative relationships) to 1.0.

18.4.5 Classificatory ability of the models

Comparison between various logistic regression models fitted and with other classification methods such as discriminant function and decision tree methods can be made with respect to their classifying ability with the help of (2 x 2) classification tables in case of a binary response group variable. The columns are the two observed values of the dependent variable, while the rows are the two predicted values of the dependent. In a perfect model, all cases will be on the diagonal and the overall percent correct will be 100%.

Critical terms associated with classification table are as follows:

Hit rate: Number of correct predictions divided by sample size. The hit rate for the model should be compared to the hit rate for the classification table for the constant-only model.

Sensitivity: Percent of correct predictions in the reference category (usually 1) of the dependent. It also refers to the ability of the model to classify an event correctly.

Specificity: Percent of correct predictions in the given category (usually 0) of the dependent. It also refers to ability of the model to classify a non event correctly.
False positive rate: It is the proportion of predicted event responses that were observed as nonevents

False negative rate: It is the proportion of predicted nonevent responses that were observed as events.

Higher the sensitivity and specificity lower the false positive rate and false negative rate, better the classificatory ability.

18.5 Association between attributes/ variables

An association exists between two variables if the distribution of one variable changes when the level (or value) of the other variable changes. If there is no association, the distribution of the first variable is the same regardless of the level of other variable. Odds ratio is usually used for measuring such associations. For example, consider the following table having two attributes 'Weather' and 'Mood of boss' each at two levels.

	Mood of boss	
	Good	Bad
Weather		
Rain	82	18
Shine	60	40

Odds of an event is the ratio of the probability of an event occurring to the probability of it not occurring. That is,

Odds=P(event)/{1-P(event)} = P(event=1)/P(event=0)

In the above table, there is 82% probability that the mood of the boss will be 'Good' in case of 'Rain'. The odds of 'Good mood' in 'Rain' category =0.82/0.18 =4.5. The odds of 'Good mood' in 'Shine' category =0.60/0.40 =1.5. The odds ratio of 'Rain' to 'Shine' equals (4.5/1.5) =3 indicating that the odds of getting 'Boss in good mood' during 'Rain' is three times those during 'Shine'. Also there is 18% probability that mood of boss will be 'Bad' in case of 'Rain'; the odds of 'Bad mood' in 'Rain' =0.18/0.82 =0.22. Thus, in case the probability is very small (0.18 in this case), there is no appreciable difference in mentioning the same as probability or odds.

The importance of odds ratio is case of logistic regression modeling can be further explained by taking a simple case of influence of an attribute "Gender" X with two levels (Male or Female) on another attribute "opinion towards legalized abortion" Y with two levels (Yes=1, No=0). Logistic regression when written in its linearised form takes the following 'logit' form:

$$logit\{Y=1|X=x\} = log(\pi/(1-\pi)) = log(odds) = \beta 0+\beta 1*x$$

Now,
Odds (Females)= exp($\beta 0+\beta 1$) and Odds (Males)= exp($\beta 0$). Hence

Odds ratio = exp($\beta 0+\beta 1$)/exp($\beta 0$) = $exp(\beta 1)$.
Thus here regression coefficient of Y on X i.e. $\beta 1$ is not directly interpreted but after taking exponentiation of it.

18.6 Multinomial logistic regression modeling

Let **X** is a vector of explanatory variables and π denotes the probability of binary response variable then logistic model is given by

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \mathbf{X}\beta = g(\pi)$$

where, 'alpha' is the intercept parameter and 'beta' is a vector of slope parameters. In case response variable has ordinal categories say 1,2,3,--------, I, I+1 then generally logistic model is fitted with common slope based on cumulative probabilities of response

categories instead of individual probabilities. This provides parallel lines of regression model with following form

$$g\,[\text{Prob}(\ \mathbf{y} \le \mathbf{i}(\mathbf{x})\ )] = \alpha_i + \chi\beta\ ,\ 1 \le i \le I$$

where, $\alpha_1, \alpha_2, \text{------} \alpha_k$, are k intercept parameters and β is the vector of slope parameters.

Multinomial logistic regression (taking qualitative response variable with three categories, for simplicity) is given by

$$\text{logit}[\Pr(Y \le j - 1 / \mathbf{X})] = \alpha_j + \boldsymbol{\beta}^T \mathbf{X}\ , \quad j = 1,2$$

where α_j are two intercept parameters ($\alpha_1 < \alpha_2$), $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots\ldots, \beta_k)$ is the slope parameter vector not including the intercept terms, $\mathbf{X}^T = (X_1, X_2, \ldots, X_k)$ is vector of explanatory variables. This model fits a common slope cumulative model i.e. 'parallel lines' regression model based on the cumulative probabilities of the response categories.

$$\text{logit}(\pi_1) = \log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k\,,$$

$$\text{logit}(\pi_1 + \pi_2) = \log\left(\frac{\pi_1 + \pi_2}{1-\pi_1-\pi_2}\right) = \alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k$$

where

\

$$\pi_1(X) = \frac{e^{\alpha_1 + \beta^T X}}{1 + e^{\alpha_1 + \beta^T X}}$$

$$\pi_1(X) + \pi_2(X) = \frac{e^{\alpha_2 + \beta^T X}}{1 + e^{\alpha_2 + \beta^T X}}$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

π_j (X) denotes classification probabilities Pr(Y=j-1 / X) of response variable Y, j = 1,2,3, at X^T.
These models can be fitted through maximum likelihood procedure.

18.7 Application of binary logit models in agriculture and other sciences

Sometimes quantitative information on pests and diseases is not available but is available in qualitative form such as occurrence / non-occurrence, low / high incidence etc. The statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model. It can be used to describe the relationship of several

independent variables to the binary (say, named 0 & 1) dependent variable. The logistic regression is used for obtaining probabilities of occurrence, say E, of the different categories when the model is of the form: $P(E=1) = \frac{1}{1+\exp(-z)}$ where z is a function of associated variables, if $P(E=1) \geq 0.5$ then there is more chance of occurrence of an event and if $P(E=1) < 0.5$ then probability of occurrence of the event is minimum. If the experimenter wants to be more stringent, then the cutoff value of 0.5 could be increased to, say, 0.7.

Consider the dataset given in the Table given below. Weather data during 1987-97 in Kakori and Malihabad mango (*Mangifera indica* L.) belt (Lucknow) of Uttar Pradesh is used here to develop logistic regression models for forewarning powdery mildew caused by *Oidium mangiferae* Berthet and validated the same using data of recent years. The forewarning system thus obtained satisfactorily forewarns with the results obtained comparing well with the observed year-wise responses. The status of the powdery mildew (its epidemic and spread) during 1987-97 are given in the following table, with the occurrence of the epidemic denoted by 1 and 0 otherwise. The variables used were maximum temperature (X_1) and relative humidity (X_2). The model is given by

$$P(Y=1) = 1/[1+\exp\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)\}]$$

Table: Epidemic status (Y) of powdery mildew fungal disease in Mango in U.P.

Year	Third week(Y) of March	Average weather data in second week of March	
		X_1	X_2
1987	1	30.14	82.86
1988	0	30.66	79.57
1989	0	26.31	89.14
1990	1	28.43	91.00
1991	0	29.57	80.57
1992	1	31.25	67.82
1993	0	30.35	61.76
1994	1	30.71	81.14
1995	0	30.71	61.57
1996	1	33.07	59.76
1997	1	31.50	68.29

Logistic regression models were developed using the maximum likelihood estimation procedure in SAS. Consider 1987-96 model based on second week of March average weather data using which forewarning probability is obtained for the year 1997. The parameter estimates corresponding to intercept, X1 and X2 are obtained as $\hat{\beta}_0 = -72.47$; $\hat{\beta}_1 = 1.845$; $\hat{\beta}_2 = 0.22$.

Then the model becomes

$$P(Y=1) = 1/\{1+\exp(-(-72.47+(1.845*x_1)+(0.22*x_2))\}$$

Plugging in the values X_1 = 31.50 and X_2 = 68.29, of year 1997 it can be seen that P(Y=1) = 0.66. This is the forewarning probability of occurrence of powdery mildew in mango using logistic regression modeling for 1997. The logistic regression model yielded good results. If P (Y=1) <0.5, then probability that epidemic will occur is minimal, otherwise there is more chance of occurrence of epidemic and this can be taken as objective procedure of forewarning the disease. As we were having the information that there was epidemic during the year 1997, it can be seen that the logistic regression model forewarns the actual status correctly.

Consider one example, data from the field of medical sciences relating to Occurrence or Non-occurrence of Coronary Heart Disease (CHD) in human beings as given in the following table.

Group	Age	No. of observations	Presence of CHD
1	25	10	1
2	30	15	2
3	35	12	3
4	40	15	5
5	45	13	6
6	50	8	5
7	55	17	13
8	60	10	8

The result is given below:

Response Variable (Events) CHD
Response Variable (Trials) n
Model binary logit
Optimization Technique Fisher's scoring

Number of Observations Used 8
Sum of Frequencies Used 100

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	43
2	Nonevent	57

Model Fit Statistics

	Intercept	
	Intercept	and

Criterion	Only	Covariates
AIC	138.663	112.178
SC	141.268	117.388
-2 Log L	136.663	108.178

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.4851	1	<.0001
Score	26.0782	1	<.0001
Wald	21.4281	1	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.1092	1.0852	22.1641	<.0001
age	1	0.1116	0.0241	21.4281	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.118	1.066	1.172

Association of Predicted Probabilities and Observed Responses

Percent Concordant	74.6	Somers' D	0.588
Percent Discordant	15.7	Gamma	0.651
Percent Tied	9.7	Tau-a	0.291
Pairs	2451	c	0.794

Partition for the Hosmer and Lemeshow Test

		Event		Nonevent	
Group	Total	Observed	Expected	Observed	Expected
1	10	1	0.90	9	9.10
2	15	2	2.20	13	12.80
3	12	3	2.77	9	9.23
4	15	5	5.16	10	9.84
5	13	6	6.22	7	6.78
6	8	5	4.93	3	3.07
7	17	13	12.53	4	4.47
8	10	8	8.30	2	1.70

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
0.2178	6	0.9998

The Interpretation of the above output is given subsequently.

The fitted model is given by

P (CHD=1) = 1/ (1+ exp (-z)) where z = $\beta 0+\beta 1$ * (age group)

Testing of overall Null Hypothesis that BETA = 0 using Likelihood and other tests indicate that they are highly significant and hence there is considerable effect on age on CHD disease.

The Hosmer-Lemeshow Goodness of Fit Test with 6 degrees of freedom suggests that the fitted model is adequate. Here one has to see for a large p-value (>0.05). in order to infer that the model is very well fitted.

Table : Classification Table for Predicted Event frequencies

Correct		Incorrect		Percentages				
Event	Non-event	Event	Non-event	Hit rate	Sensitivity	Specificity	False POS	False NEG
48	26	17	9	74.0	84.2	60.5	26.2	25.7

Here "Correct" columns list the numbers of subjects that are correctly predicted as events and nonevents. Also "Incorrect" columns list both the number of nonevents incorrectly predicted as events and the number of events incorrectly predicted as nonevents.

FALSE positive and FALSE negative rates are low, sensitivity (the ability of the model to predict an event correctly) (84.2%) and specificity (the ability of the model to predict a nonevent correctly) (60.5%) of the model are high enough and hence the fitted model is very effective for prediction/ classification.

References:

Fox, J. (1984). *Linear statistical models and related methods with application to social research,* Wiley, New York.

Kleinbaum, D.G. (1994). *Logistic regression*: A self learning text, New York: Springer.

RANKED SET SAMPLING

Tauqueer Ahmad

Indian Agricultural Statistics Research Institute, New Delhi-110012

19.1 Introduction

The need for statistical information seems endless in modern society. In particular, data are regularly collected to satisfy the need for information about specified sets of elements, called finite population. The collection of all units having certain properties as per the objectives of the study at particular point or period of time is known as ‘population’. One of the most important modes of data collection for satisfying such needs is a sample survey. Sample survey is a technique of selection of a part of an aggregate to represent the whole, that is, a partial investigation of the finite population and is frequently used in everyday life in all kinds of investigations. A sampling method is a scientific and objective procedure of selecting units from the target population. The method provides a sample that is expected to be representative of the population as a whole. There is another approach called ‘complete enumeration’ or ‘census survey’ in which data is collected from each unit of the population. This approach is usually used in census of population, agriculture etc. A sample survey costs less than a census survey, is usually less time consuming and even more accurate than a census survey.

The technique of selecting a sample is of fundamental importance in sampling theory and usually depends on the nature of the investigation. The sampling methods which are commonly used may be broadly bifurcated into two approaches, viz., purposive sampling (or non-probability sampling) and random sampling (or probability sampling). The method of purposive sampling is purely based on judgement of the sampler. Thus, it induces judgemental or personal bias in selection of sampling units. Due to this, random sampling approach is generally adopted in most of the surveys. The method of random sampling involves assigning predetermined probabilities to every unit of the population and selecting the units according to the predetermined probabilities. The sum of probabilities of all the units is one.

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”.

19.2 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 design with set size $m = 3$ and no. of sampling cycles $r = 4$. Although 36 sample units have been selected from the population, only the 12 circled units are actually included in the final sample for quantitative analysis.

19.3 RSS Estimator and its Variance

Let us consider only one cycle first. Let, $X_{11}, X_{12}, \ldots, X_{1m}, X_{21}, X_{22}, \ldots, X_{2m}, \ldots, X_{m1}, X_{m2}, \ldots, X_{mm}$ be independent random variables all having the same cumulative distribution function F(x). Also let $X_{i(1)}, X_{i(2)}, \ldots, X_{i(m)}$ be the corresponding order statistics of $X_{i1}, X_{i2}, \ldots, X_{im}$ (for all i=1,2,…,m). Then $X_{1(1)}, X_{2(2),} \ldots, X_{i(i),\ldots}, X_{m(m)}$ 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 ith order statistics from ith sample in jth 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

$$V[\hat{\mu}_{RSS}]=\frac{1}{mr}[\sigma^2-\frac{1}{m}\sum_{i=1}^{m}\{\mu_{(i:m)}-\mu\}^2] \qquad \text{…….(3)}$$

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.

19.4 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,

$$V(\hat{\mu}_{SRS}) = \sigma^2 / mr \qquad .....(4)$$

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,

$$RS = \frac{1}{m}\sum_{i=1}^{m}(\frac{\tau_{(i)}}{\sigma})^2 \qquad .....(6)$$

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.

TECHNIQUES OF CROP CUTTING EXPERIMENT

MAN SINGH

Indian Agricultural Statistics Research Institute, New Delhi-110012

20.1 Introduction

Agricultural statistics has great importance for the planners for planning of the economic policies for betterment of the people of the country. The agricultural statistics includes land utilization, agricultural production including livestock and fisheries, cost and prices, number & size of holdings, composition of agricultural population, ownership & tenancy of holdings, agricultural machinery & power etc. Crop area and crop production is an important constituent of agricultural statistics system. The total production of a crop is based on acreage under the crop and average yield per hectare. In India, crop area figures are compiled on the basis of complete enumeration while the crop yield is estimated on the basis of sample survey approach. The whole country is divided into three broad categories i.e. temporarily settled states, permanently settled states and non-reporting areas on the basis of the procedure adopted for recording the **crop area statistics.**

20.2 Crop Area Statistics

20.2.1 Temporary Settled States: The system of temporary settlements was introduced in our country in 1892, with a view to fix land revenue for a period. Generally, it is revised at the time of next settlement. The States and Union Territories which have adopted this system as of now covering 86% reporting area. These states are cadastrally surveyed and having a primary reporting agency for collecting the statistics of crop area. The crop area statistics are being collected by complete enumeration method. The primary worker called Patwari is responsible for collection of crop area statistics under his jurisdiction. The crop area statistics is collected through field to field inspection during each of the agricultural season. This exercise is known as *Girdawari*. The register in which crop area is recorded is known as "Khasra Register". The States and Union Territories covered under temporarily settled states are Andhra Pradesh, Telangana, Assam, Andaman & Nicobar Islands, Bihar, Chandigarh, Chhattisgarh, Delhi, Dadar & Nagar Haveli, Gujarat, Himachal Pradesh, Haryana, Jharkhand, Jammu & Kashmir, Karnataka, Madhya Pradesh, Maharashtra, Punjab, Puducherry, Rajasthan, Tamil Nadu, Uttar Pradesh and Uttrakhand.

20.2.2 Permanently Settled States: The states/areas known as permanently settled states where land revenue was permanently fixed and question of revision ordinarily did not arise. These states are cadastrally surveyed but they do not have primary reporting agency. There are three permanently settled states West-Bengal, Orissa and Kerala covering 9% reporting area. Crop area statistics are compiled by sample survey approach through a scheme entitled "Establishment of an Agency for Reporting of Agriculture

Statistics" (EARAS) by the regular reporting agency. Every year a sample of 20% villages is selected and the selected villages are completely enumerated for the purpose of reporting crop area statistics. Next year a fresh sample of 20% villages is selected and data collected. Thus, all the villages in the respective state are covered in five years.

20.2.3 Non-reporting Areas: The regions for which there is no system of reporting crop area in the country are covered under this category. Mostly NEH States are covered under this category. In this region the reporting crop area is 5%. The crop area estimates in NEH region except Assam are not based on any systematic approach. Here, the statistics of land records are collected on a sample basis. The revenue/agriculture officer collects the information on the basis of his personal belief and knowledge.

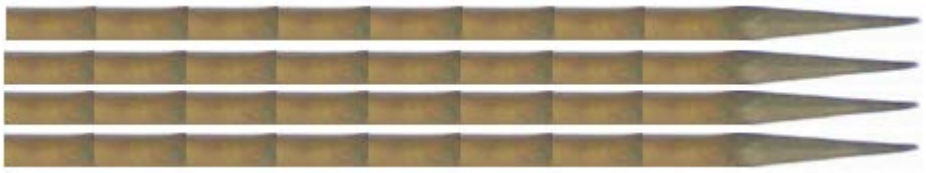
20.3 Crop Yield Estimation

The crop yield estimation in the country is carried out on the basis of sample survey approach using Crop Cutting Experiment (CCE). A series of studies were carried out by the statisticians and a scientific technique was developed to conduct the crop cutting experiment for obtaining the reliable yield rates Hubback (1925), Mahalanobis (1939), Sukhatme and Panse (1951). The estimates of yield rates are obtained on the basis of scientifically designed CCE under a scheme of the Directorate of Economics and Statistics (DES), Ministry of Agriculture entitled "General Crop Estimation Surveys" (GCES).

20.4 Equipment/Material

The CCE is conducted with the help of equipments/tools. These equipments/tools and related material are as under:

20.4.1 Pegs



20.4.2 Measuring tape

20.4.3 Rope

20.4.4 Weighing balance

4.4 (a): Spring Balances

20.4.4 (b): Beam balance

20.4.5 Set of weights

20.4.6 Hessian Cloth

The Hessian cloth is a coarsely woven fabric usually made from vegetable fibers and jute. Known for its plain weaving and durable quality, these are eco-friendly and are used for the packaging of various varieties of goods like grains, sugar, pulses and others.

20.4.7 Cloth Bag

The other items required for smooth running of CCE are two strong water proof bags (one for CCE equipments and other for CCE records), blank schedules, instruction manual, random number table, stationary. The other essential tools/machines for harvesting, threshing and winnowing/cleaning of experimental crop are easily available with owner of field.

20.5 Size and shape of the CCE plot

During the earlier attempt in crop surveys greater attention was paid in deciding the size of the CCE plot in a selected field. Various plot sizes were tried varying from 1/160 of an acre for paddy in Orissa to 1/10 of an acre for cotton in Madhya Pradesh. The plot size adopted in the earlier attempts by Hubback and Mahalanobis was very small, being of the order of 1/2000 of an acre. Attempts were made since 1944 to study the relative efficiency of various plot sizes for yield rates. The first investigation for testing the relative efficiency of different plot sizes was conducted on wheat crop in Moradabad district of Uttar Pradesh during 1944-45 (Sukhatme, 1946a, 1946b, 1947a). In this study, altogether five different plots sizes were compared; three equilateral triangles each of side 33', 16½' and $8^{1/4}$ '; two circular plots of radius 2' and 3'. The result showed that plot area less than 30 sq. ft. gave serious overestimates of yield. The bias towards overestimation diminished with the increase in the size of the experimental plot but plots of size of 118 sq. ft. were not free from perceptible bias. Therefore more sizes and shapes of plots were tried on other crop in different parts of country (Sukhatme, 1946a, 1947a). In Madras, different plot sizes were studied on paddy crop; a rectangular plot of 50 links x 20 links (area 435.6 sq. ft.), two circular plots of radius 3' each (area 28.3 sq. ft.), two circular plots of radius 2' each (area 12.6 sq. ft.) and in addition, the whole of the remaining field was harvested. Results showed that while the yield estimate from the plot size of 50 links x 20 links was in close agreement with that from harvesting the whole field, those from small plots were considerable overestimates.

Another study was taken on paddy crop by Sukhatme to the comparison of the plot size adopted by Mahalanobis in Bihar (Mahalanobis, 1945) with plot size 50 links x 25 links (Sukhatme, 1947 a). Altogether five plots were in each selected field. The plots were:

(a) one rectangular plot of size 50 links x 25 links (area 544.5 sq. ft.),

(b) two isosceles right-angled triangles with equal sides each equal to 5 ft. (area 12.5 sq. ft.) and

(c) two equilateral triangles of side 15 links each (area 42.5 sq. ft.).

The result of investigation confirmed the previous results that are smaller plot size overestimated the yield. Similar results were reported on cotton (Panse, 1946 b, 1947) which shows that even with the best supervision and training, it is not unlikely that plots up to 200 sq. ft. may give biased estimates. In an experiment conducted in Orissa on jute, the plot of size 10 links x 10 links were found to give a significant overestimation as compared to plot of 25 links x 25 links size.

The size and shape of the CCE plot for various crops, in respect of different States are specified. The shapes of the cuts for various crops vary to some extent in different States. In most of the States and for many crops, the plots are either square of size **5 meters x 5 meters, 10 meters x 10 meters** or rectangle of size **10 meters x 5 meters**. In the State of U.P., the experimental plot is equilateral triangle of side **10 meters** for most of the crops and in West Bengal, it is circle of radius 1.7145 meters approximately. For some crops, specially fruits, it consists of either specific number of trees. The plot size adopted for different food and non-food crops is as under.

Name of the crop	Shape	Length (Meter)	Breadth (Meter)	Diagonal (Meter)
Paddy, Wheat, Jowar, Bajra, Ragi, Maize, Groundnut, Tobacco, Sugarcane, Korra, Greengram, Chillies, Mesta, Horsegram, Blackgram, Bengalgram, Sunflower	Square	5	5	7.07
Redgram, Sesamum, Caster, Cotton	Square	10	10	14.14

20.6 Crop cutting experiment technique

There are following essential steps to be followed in the conduct of CCE.

- Selection of villages
- Selection of field
- Identification of South-West corner of the selected field

- Identification South-West corner of the experimental plot in the selected field
- Marking of the experimental plot of a given size and shape
- Harvesting of experimental crop of CCE plot
- Threshing of experimental crop
- Winnowing and cleaning of experimental crop
- Weighing of the produce (wet weight) just after cleaning,
- Drying of produce, in case of excess moisture, and
- Weighing of produce after drying (dry weight).

20.6.1 Selection of villages

The number of experiment has to be conducted in a state for a particular crop is to be decided at State level. The districts of the State are divided into two groups **"Major"** and **"Minor"** on the basis of area under particular crop. Those districts having area of crop under study more than 80000 hectare or between 40000 to 80000 hectare or average area more than the other districts is called "**Major"** district. Generally 80 to 120 CCE are to be conducted for particular crop in major district. The rest of the districts are called "Minor" district for a particular crop. In these districts 44 to 46 CCE are to be organized in such a way to obtained high precision of yield rates and also to manage the work load of the primary worker.

Number of CCE allotted to district is to be allocated to each stratum i.e. Community Development Block/Tehsil on the basis of area under particular crop. Maximum 16 CCE is to be conducted in stratum. Two CCE is recommended to conduct in village. Therefore, number of villages is half of the number of CCE. Stratum wise list of each district is prepared at State by the state level officer and send to the concerned district level officer to distribute among the primary workers engaged in the selected villages.

20.6.2 Selection of field

Field is a distinct piece of land growing the crop under study which is clearly demarcated on all its sides either by bunds or by patch of other crops or left un-cultivated.

As per the existing methodology of estimation of yield rates of crops, two fields of selected crop has to be selected in each selected village and one experimental plot of the selected crop has to be conducted in each selected field (Sukhatme and Panse, 1951). For selecting two fields in each selected village, two random numbers are assigned to the primary worker. The complete land of the selected village is divided into fields. Each field has its own identification number called **survey number or Khasra number.** The highest survey number in the selected village may be higher, equal or less than the random number assigned for selection of the field. If the assigned random number is

equal or less than the highest survey number, the survey number corresponding to the random number is selected and if it is higher than the highest survey number, the assigned random number is divided by the highest survey number and the survey number corresponding to the remainder is selected. In case, the remainder is 0, the highest survey number is selected. Crop under study is not grown in the selected survey number, the next survey number has to be selected.

If the selected survey number is further divided into sub-divisions, only one sub-division has to be selected randomly. In case the selected survey/sub-division number contains more than one field growing the crop under study, the field nearest to the south west corner of survey/sub-division number is to be selected. The selected field must satisfy the following conditions.

- The area of the selected field should be more than the area of CCE plot, so that the CCE plot of recommended size must be accommodated in the selected field.
- If the selected field is sown with mixed crops, the experimental crop must constitute at least 10% of its crops area.
- The experimental crop in the field is not meant for prize competition or seed production or demonstration.
- The experimental crop is not grown for fodder purpose.

The field must be considered for selection for conducting crop cutting experiment and yield obtained from the CCE plot must be recorded, if

- the experimental crop has not germinated or has failed but its area is recorded by the Village Accountant, or
- the field growing the experimental crop is grazed by cattle or damaged partially or completely by wild animals, or
- the experimental crop is affected by pests/diseases/heavy rainfall/inadequate rainfall.
- the yield must be recorded as zero in case the experimental crop is damaged completely.

The field need not be considered for selection for conducting crop cutting experiment, if

- the experimental crop has not germinated or has failed and its area is not recorded by the Village Accountant, or
- the experimental crop has withered or dried up and another crop has been raised in its place in the same season, the area of latter has been recorded by the Village Accountant.

Substitution of fields is not allowed on the plea of poor growth or prior harvest by cultivators without intimation or due to late visit by primary worker. Further, if a part or

whole of the selected field has been already harvested, the experiment should not be conducted in that field, and it has to be treated as lost.

20.6.3 Identification of South-West corner of the selected field

Fixing of the south-west corner of the selected field has made mandatory for all to make the similarity. After selection of the field, the South-West corner of the field is to be identified. If you stand at South-West corner facing north of the selected field, the selected field will be in front of you and your right hand side. In case, the selected field is not exactly in north-south and east-west direction, the corner which is approximately south-west may be taken as south-west corner of the selected field.

20.6.0 South-West corner of the selected field

20.6.4 Measurement of the length and breadth of the selected field

The field selected for CCE may or may not be in regular shape. For the purpose of identifying the south-west corner of the field, it is essential that the field should be ether in square or rectangular shape. The procedure for measuring the length and breadth in both the situation i.e. regular and irregular shape of the field is as followed.

20.6.4.1 Regular shape of the selected field

When the selected field is in a regular shape i.e. square or rectangular, then we measured the longest side as a length and shorter side as a breadth in the normal steps from the south-west corner of the field (Figure-20.6.4.1).

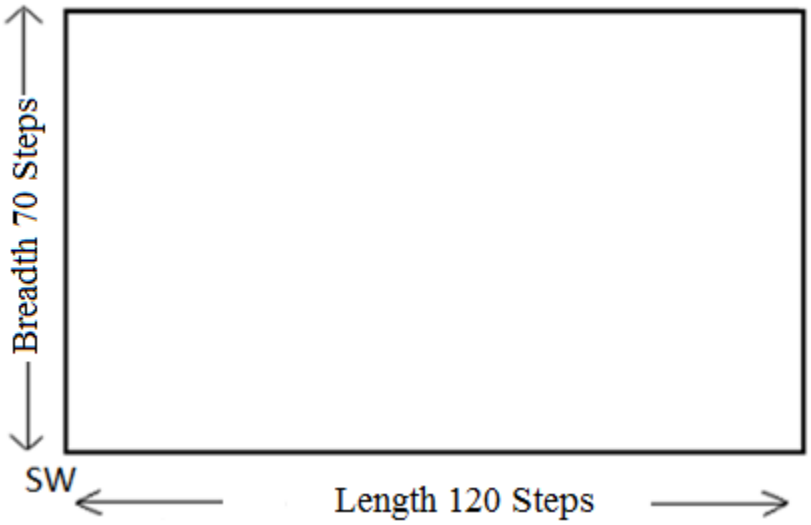
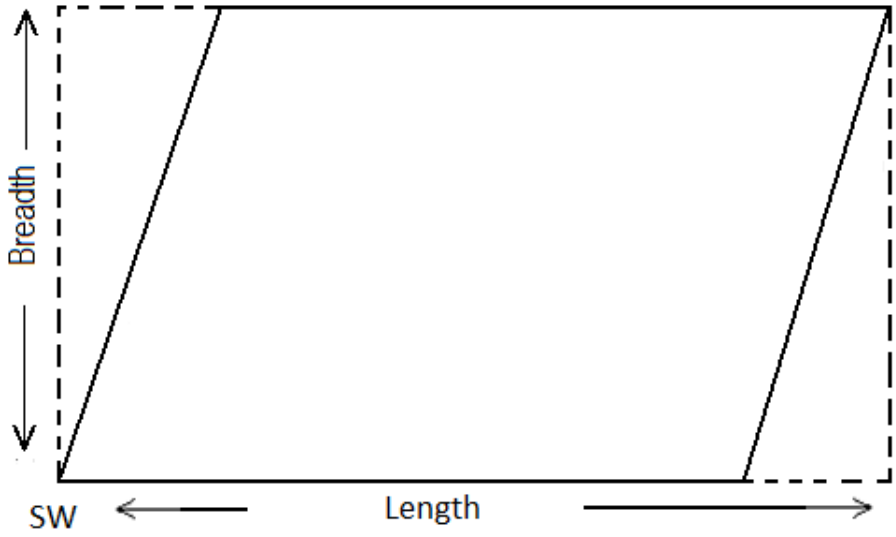


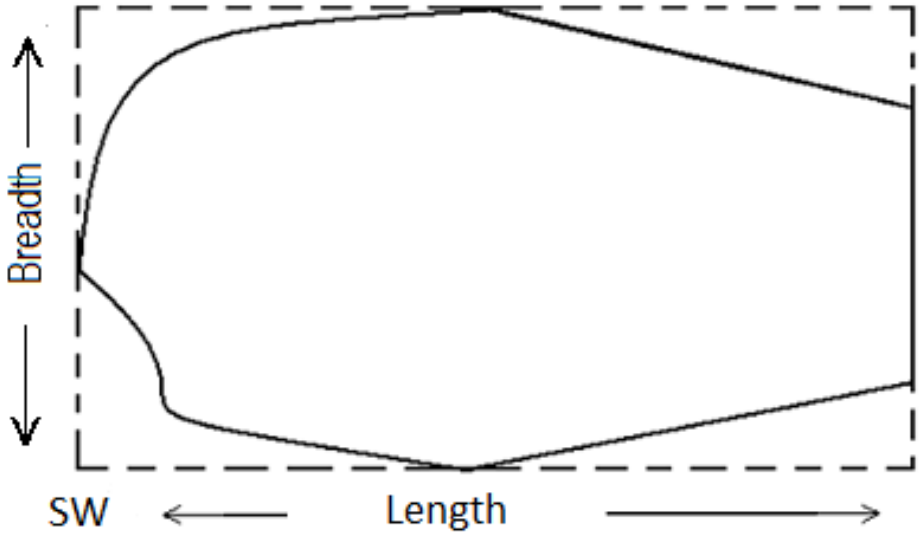
Figure-20.6.4.1: Regular shape of the field

20.6.4.2 Irregular shape of the selected field

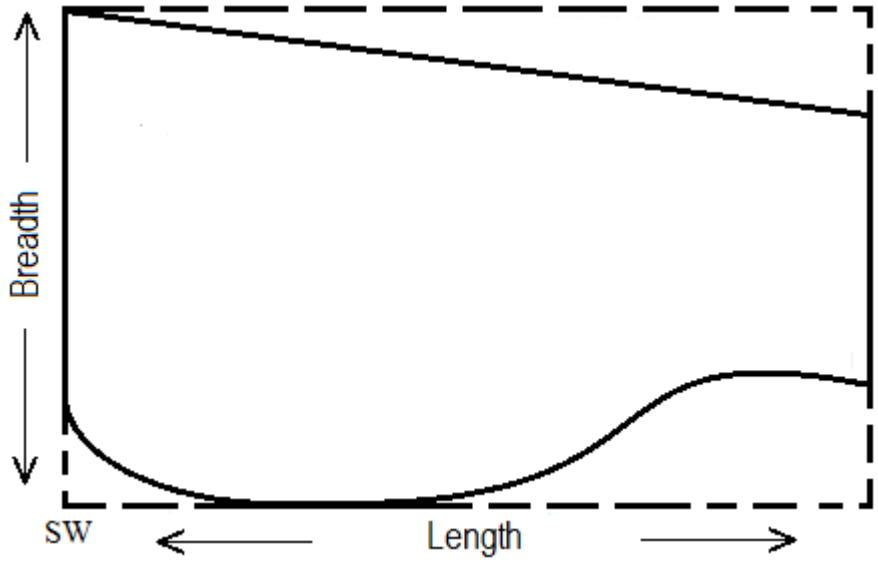
In case, the selected field is not regular shape, then enclose the selected field in an either in square or rectangular shape by touching outer least possible dimensions of the selected field. It is essential for locating the south-west corner of the selected field and also south-west corner of the experimental plot within the selected field. The south-west corner of the experimental plot should be fixed with reference to the south-west corner of the outer regular shape of the irregular selected field. Measure the longest side as a length and shorter side as a breadth of the outer regular shape of the irregular selected field in the normal steps (Figure-20.6.4.2 (a, b & c).



(a)



(b)



(c)

Figure-20.6.4.2 (a, b, c): Irregular shape of the field

20.6.5 Making of the experimental plot

The crops are sown in rows in one and two direction and without line. Keeping in view the proper representation of each plant either sown in rows or otherwise, the three different methods are recommended for making the experimental plot for CCE. Marking of the CCE plot may be done on the date of harvesting.

20.6.5.1 Making of experimental plot for CCE when crop is sown without line

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 making the experimental plot for conducting the CCE is as under.

20.6.5.1.1 Determination of the random number pair

Two random numbers, one for length and the other for breadth have to be selected with the help of random number table. These random numbers are to be selected using column number assigned to the primary worker. To ensure that the whole experimental plot gets accommodate in the selected field, steps in the length and breadth of the experimental plot have to be deducted from length and breadth of the selected field, respectively. Suppose the shape of experimental plot is square of side 5 meter. (7 steps are in 5 meter approximately).

Example:

Length of the selected field	120 Steps
Steps in length of CCE	007 Steps
Length of the selected field **minus** number of steps in length of CCE	113 Steps
Breadth of the selected field	70 Steps
Steps in breadth of CCE	07 Steps
Breadth of the selected field **minus** steps in breadth of CCE	63 Steps

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. The random number 058 is selected 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. This random number is less than 63. Accordingly, random number 51 is selected for breadth. (58, 51) is the pair of random number selected for locating the south-west corner of the experimental 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 experimental plot goes beyond the boundary of the selected field owing to irregular shape of the selected field, the pair of random number should be rejected and a new pair of random number should be selected till whole experimental plot accommodate within the field.

20.6.5.1.2 Marking of the experimental plot

20.6.5.1.2.1 Marking of south-West corner of the experimental plot

The selected random number for length is 58. Therefore, measure 58 steps along the length of the selected field from its south-west corner and the point where reached, measure 51 steps perpendicular to the length and parallel to breadth of the selected field. Thus, the point "A" where reached, is the south-west corner of the experimental plot (Figure-20.6.3.1.2). The point "A" is also called as the key point or first corner of the experimental plot. Fix a peg at the key point of the experimental plot.

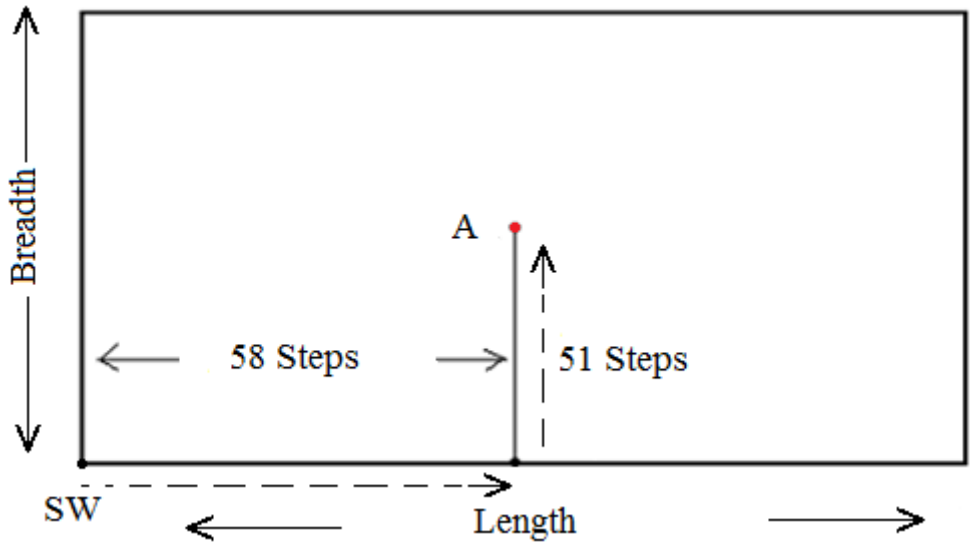


Figure-20.6.5.1.2.1: South-west corner of experimental plot (Step-1)

20.6.5.1.2.2 Marking of second corner of the experimental plot

We measured five meter along the length of the selected field from corner "A". The corner which is 5 meter away from corner "A" is the second corner "B" of the experimental plot. Fix a peg at corner "B" (Figure-20.6.3.1.2.2). The line joining the point "A" and "B" is the base of the experimental plot.

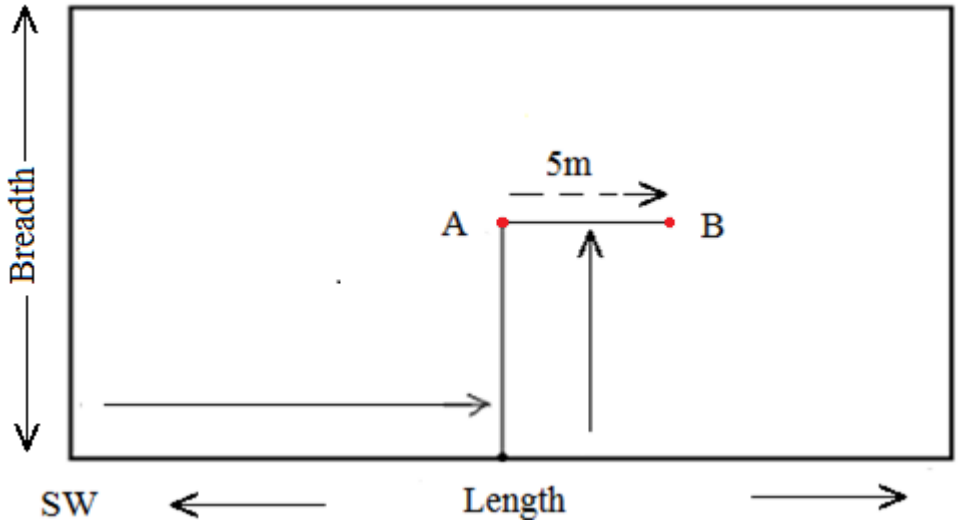


Figure-20.6.5.1.2.2: Second corner of experimental plot (Step-2)

20.6.5.1.2.3 Marking of third corner of the experimental plot

Third and fourth corner of the experimental are to marked with the help of right angle triangle method. To mark the third corner, let first person stand at corner "A" by holding the measuring tape at 0 meter mark and second person must has to stand at corner "B" holding at 12.07 (7.07+5.0) 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 reached shall be the third corner "C" of the experimental 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-20.6.5.1.2.3).

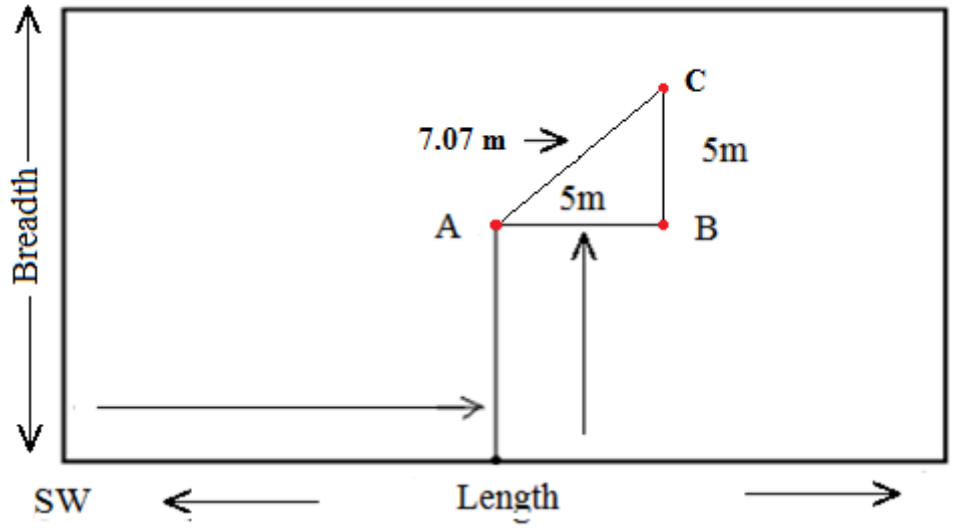


Figure-20.6.5.1.2.3: Third corner of experimental plot (Step-3)

20.6.5.1.2.4 Marking of fourth corner of the experimental plot

For locating the fourth corner of the experimental plot the third person standing at corner "C" now hold the measuring tape on 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. The point where he reached is the fourth corner "D" of the experimental plot. Fix a peg at corner "D" (Figure-20.6.5.1.2.4).

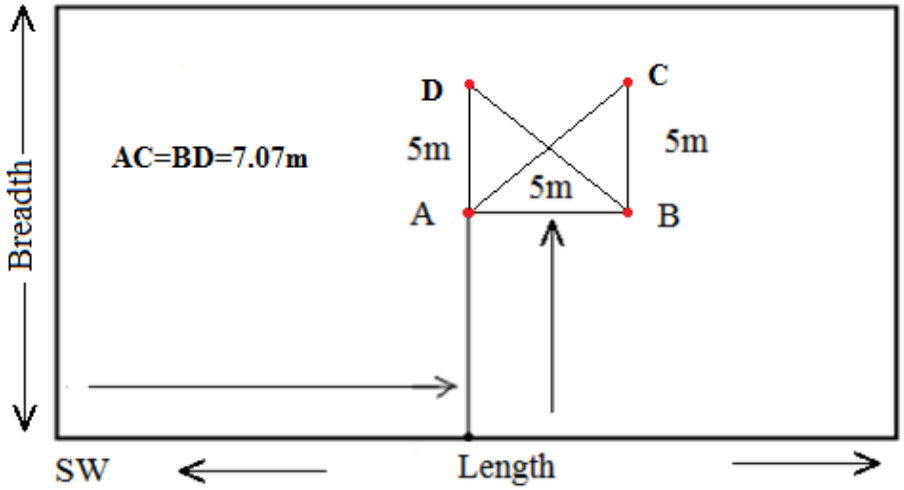


Figure-20.6.5.1.2.4: Fourth corner of experimental plot (Step-4)

20.6.5.1.2.5 Experimental plot

A, B, C and D are the four corners of the experimental plot. Check the distance between the corners. The distance between A & B, B & C, C & D and A & D should be 5 meter. The distance between both the diagonals AC and BD should also be checked and it should be 7.07 meter for each diagonal (Figure- Figure-20.6.5.1.2.5).

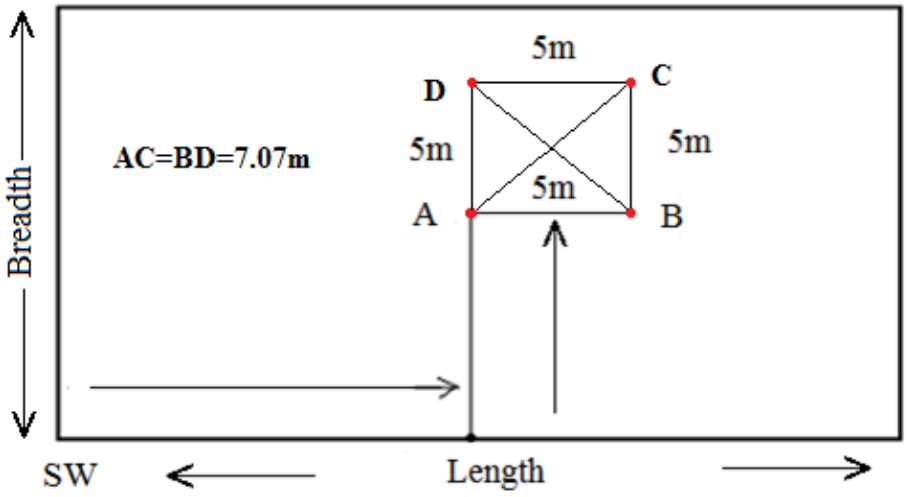


Figure-20.6.5.1.2.5: Experimental plot (Step-5)

20.6.5.2 Making of experimental plot for CCE 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. Method of making the experimental plot for conducting the CCE is as under.

20.6.5.2.1 Counting of rows

Rows are to be counted starting from the south-west corner of the field. Conventionally, this side may be identified as breadth of the field (Figure-20.6.5.2.1).

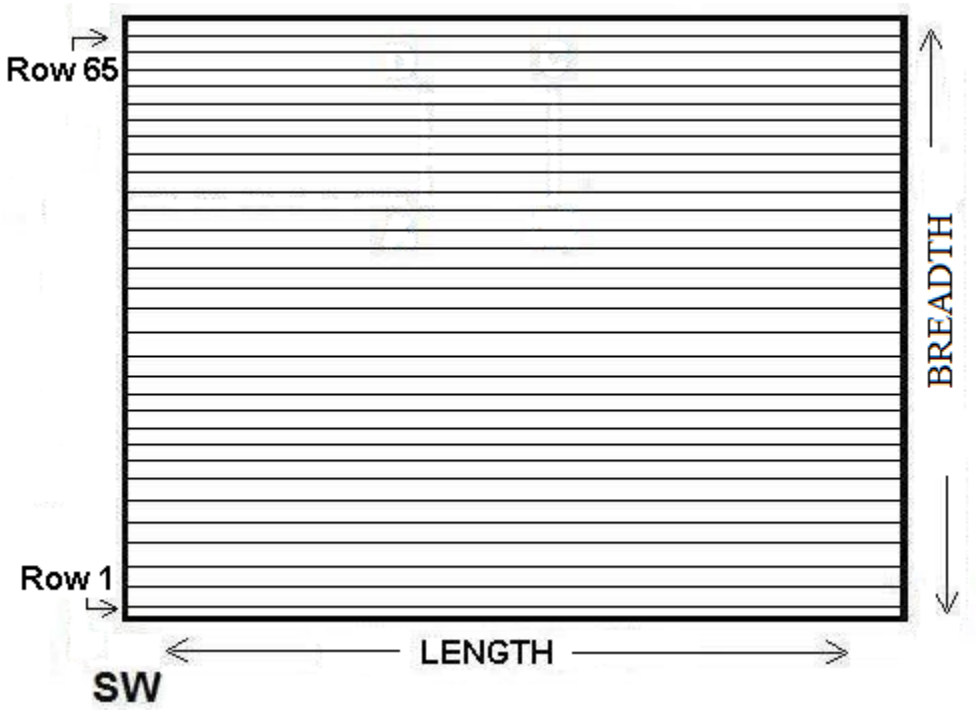


Figure-20.6.5.2.1: Enumeration of Rows

20.6.5.2.2 Measurement of length

Enclose the selected field in a regular shape (rectangle or square) in case it is not in regular shape and measure the length of longest row in normal steps in the selected field (Figure-20.6.5.2.2).

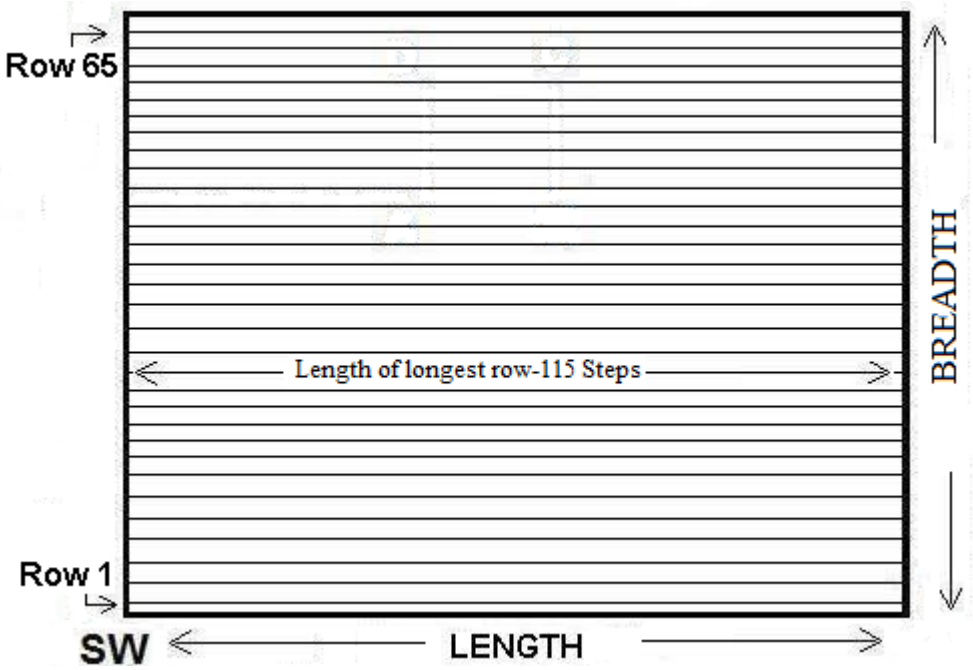


Figure-20.6.5.2.2: Length of longest row

20.6.5.2.3 Average number in the breadth of CEE

Take three physical observations randomly in the specified breadth of the experimental plot (5 meters or 10 meters in plains and 2 meters in hills) at three places in the selected field and workout the average number in the specified breadth of CCE plot.

20.6.5.2.4 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 also. Deduction of average rows in the CCE plot is necessary for ensuring that the CCE plot may fall in the selected field. Addition of one is necessary for inclusion of last row in of the selected field in the sample of CCE. A random number less than or equal to the number obtained by 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 is 6 in the specified breadth of CEE plot. For selection of random number for random row, we may calculate the number as follows.

Total number of rows in the selected field	65
Average number of rows in 5 meter breadth	6
(Number of rows in the selected field **minus** Average number of rows) + One	60

Suppose column number **one** is assigned column, Since 60 is the two digit number, therefore, using assigned column number **one** of two digit random number table, 22 is appeared first which is less than 60. The 22 is selected as random number for identifying the random row. The random row will be the first row of the experimental plot.

20.6.5.2.5 Determination of random number for random step

Steps in specified length of CCE plot may be deducted from the length (step) of longest row. Deduction of steps in length of CCE is essential for ensuring that the CCE plot may fall 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 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. For selection of random number for random step, we may calculate the number as follows.

Length of longest row in steps	115
Steps in CCE length	7
Length of longest row in steps **minus** Steps in CCE length	108

As per 108, using allotted column number **one** of three-digit random number table, random number 10 is appeared first which is less than 108, therefore, random number 10 may be considered as random step.

The selected random number pair is (22, 10)

20.6.5.2.6 Marking of the experimental plot

20.6.5.2.6.1 Marking of south-west corner of the experimental plot

Starting from south west corner of the selected field, count the rows up to random row i.e. 22. 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 experimental plot (Figure-20.6. 5.2.6.1). This may also called as first corner or key point of CCE plot. Fix a peg “A” at this point “A” in between the inter-space of the selected row and its preceding row.

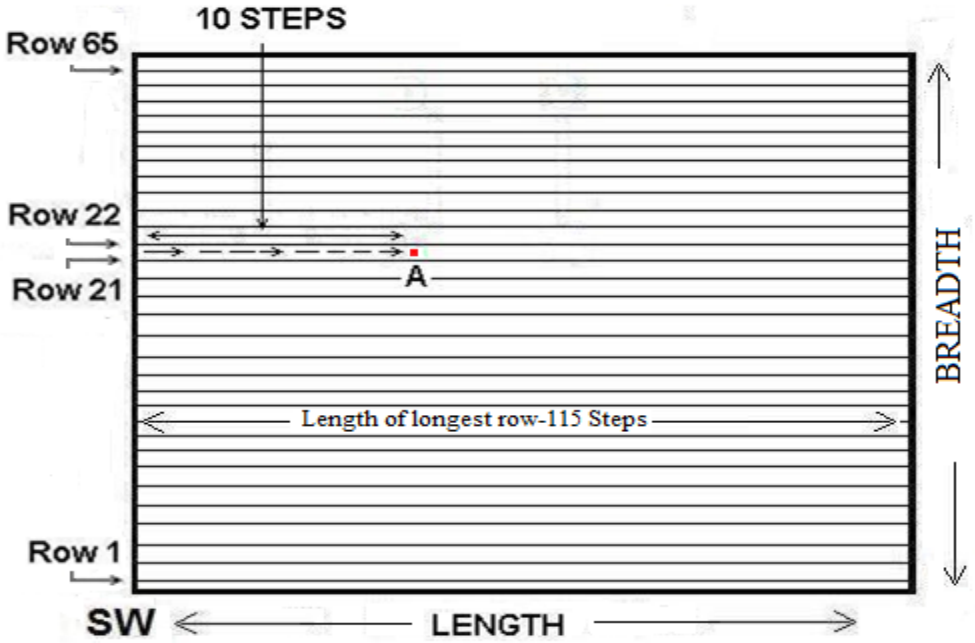


Figure-20.6.5.2.6.1: SW of experimental plot (Step-1)

20.6.5.2.6.2 Marking of second corner of experimental plot

As per length of CCE plot, measure it in meters from the key point (First corner of CCE) moving in between random row and its preceding row toward the length of row and fix second peg “B” at other corner. It is the second corner of the experimental plot (Figure-20.6.5.2.6.2).

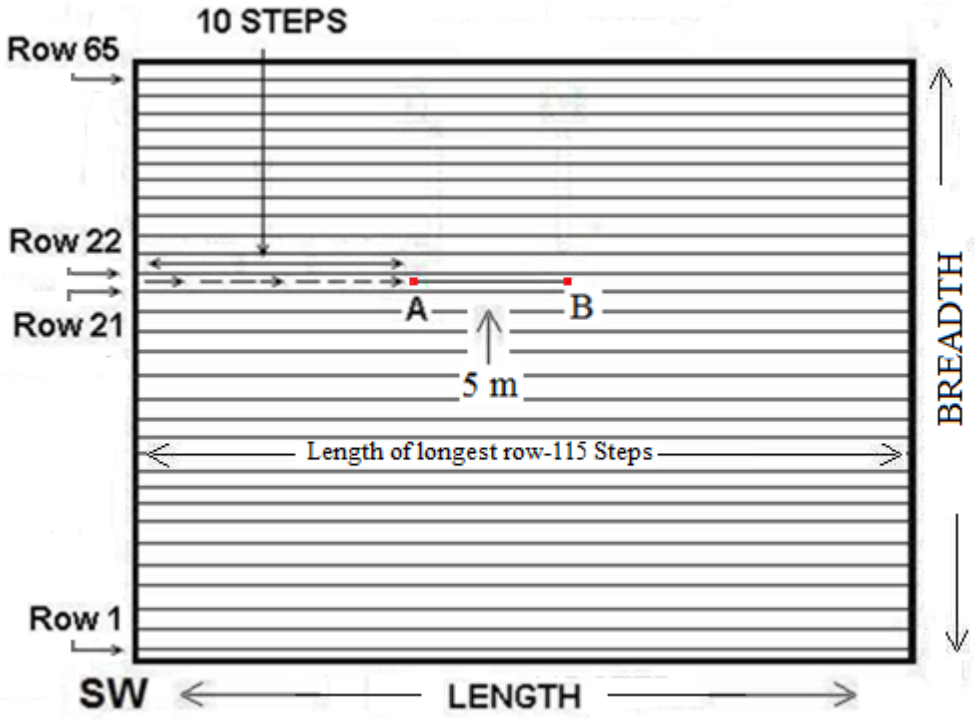


Figure-20.6.5.2.6.2: Second corner of experimental plot (Step-2)

20.6.5.2.6.3 Marking of third corner of experimental plot

Now count average number of rows i.e. six from random row 22 (second corner "B") which is the first row of CCE moving perpendicular direction to the row towards inner side of the selected field and fix third peg "C" in between the inter-space of last row (i.e. 6^{th}) to be included in the CCE plot or 27^{th} row of the selected field and its succeeding row i.e. 28^{th} (Figure-20.6.5.2.6.3).

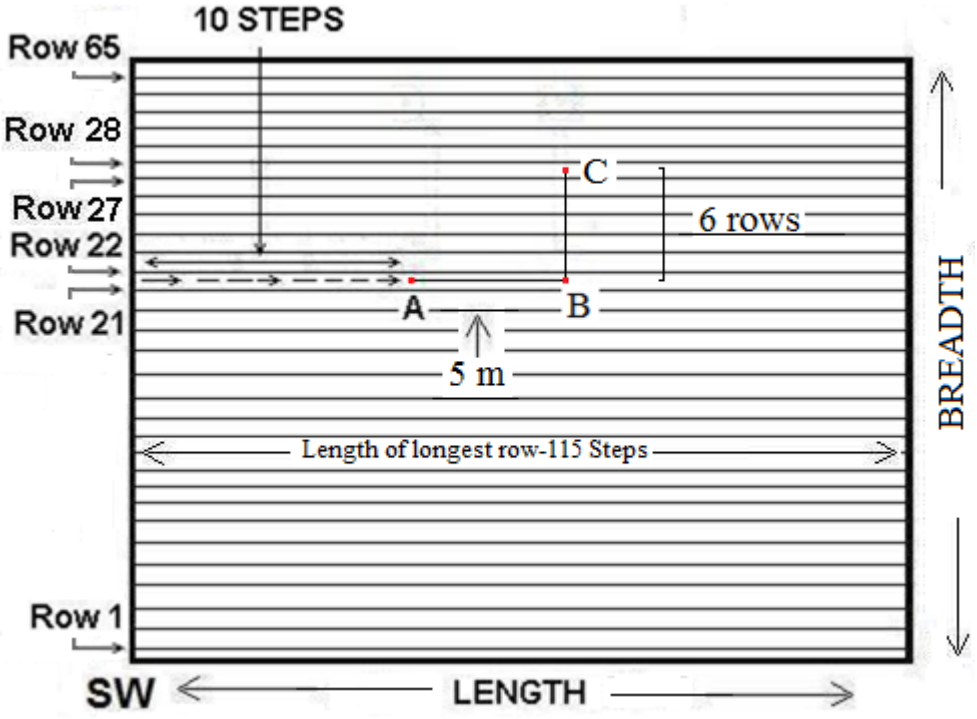


Figure-20.6.5.2.6.3: Third corner of the CCE plot (Step-3)

20.6.5.2.6.4 Marking of fourth corner of experimental plot

Measure length in meters as in the length of CCE plot, moving in between last row i.e. sixth of CCE or 27^{th} of selected field and its forward row 28^{th} from third corner "C" parallel to "B"-"A" in the direction of Key point "A", the other end where we reached, it is the fourth corner of the CCE plot, fix forth peg "D"at this point (Figure-20.6.5.2.6.4).

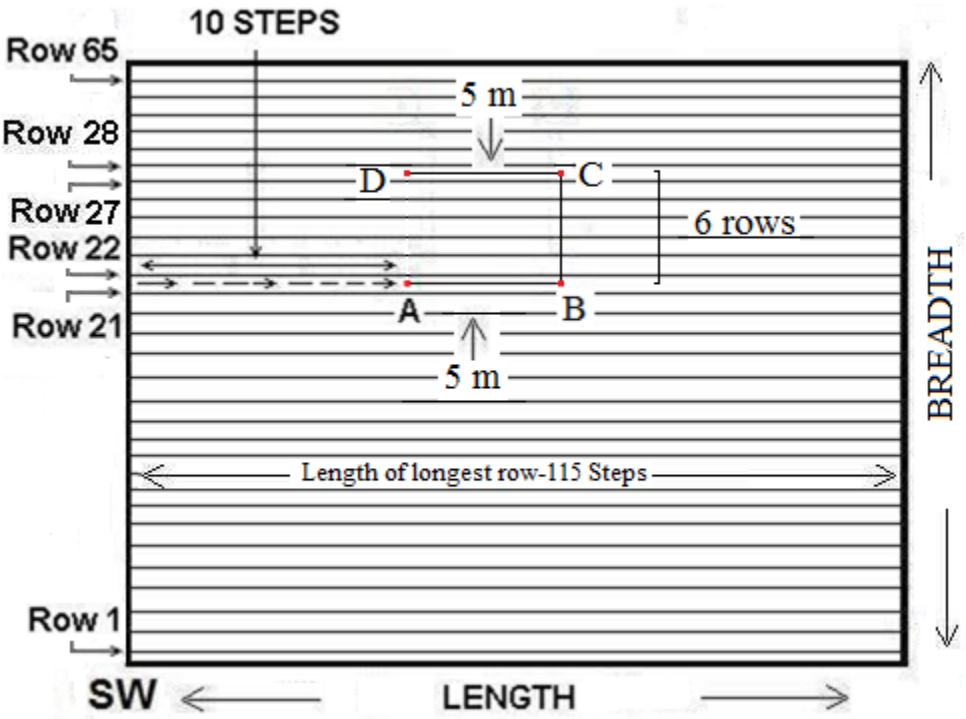


Figure-20.6.5.2.6.4: Fourth corner of the CCE plot (Step-4)

20.6.5.2.6.5 Experimental plot

Count the rows between corner "B"&"C" and "A" & "D". These are equal to average number of rows i.e. six. The distance between "A" & "B" and & "C" "D" may also be verified. It should be equal to the length of CCE plot. The distance of all the sides and diagonal may be measured and recorded (Figure-20.6.5.2.6.5). The side AD and BC may not be equal to 5 meter. Actual length of diagonals (AC and BD) may be measured and noted.

If the experimental plot does not fall wholly within the field due to irregular shape of the field, reject the first pair of random number and select a new random number pair.

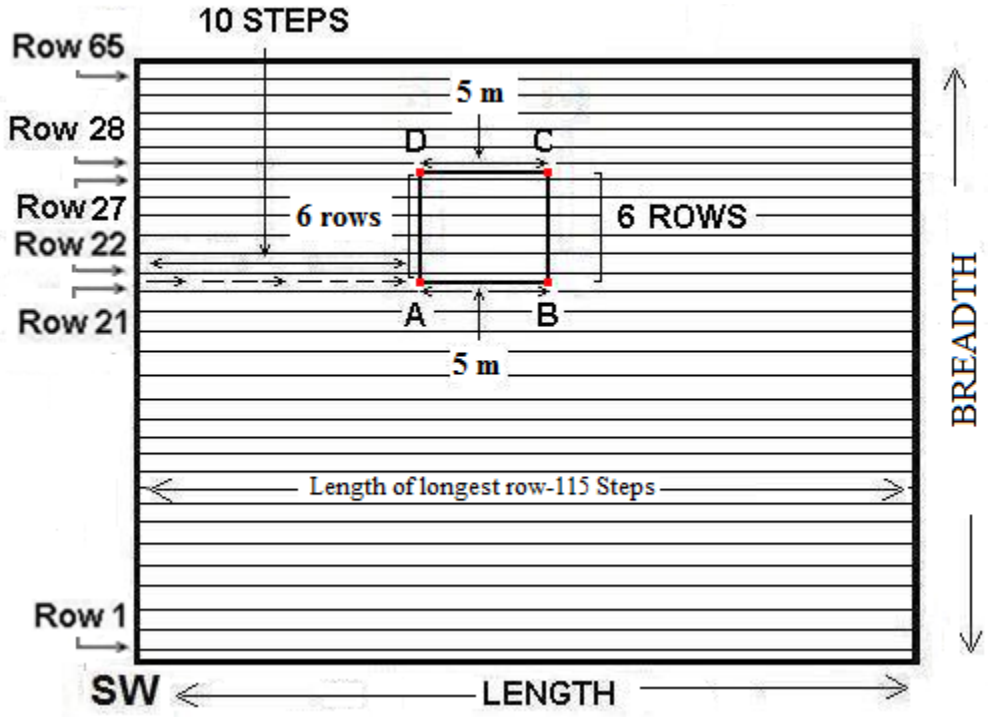


Figure-20.6.5.2.6.5: CCE plot (Step-5)

20.6.5.3 Making of experimental crop for CCE when crop is sown in line in two directions

The crop like tobacco is sown in both the directions in line. The procedure for making CCE plot is lightly differ from the procedure of making the CCE plot when crop is sown in one direction in line. The procedure of making the CCE plot is as under.

20.6.5.3.1 Enumeration of rows

Rows are to be counted in both the direction i.e. length and breadth of the selected field starting from its south-west corner (Figure-20.6.5.3.1).

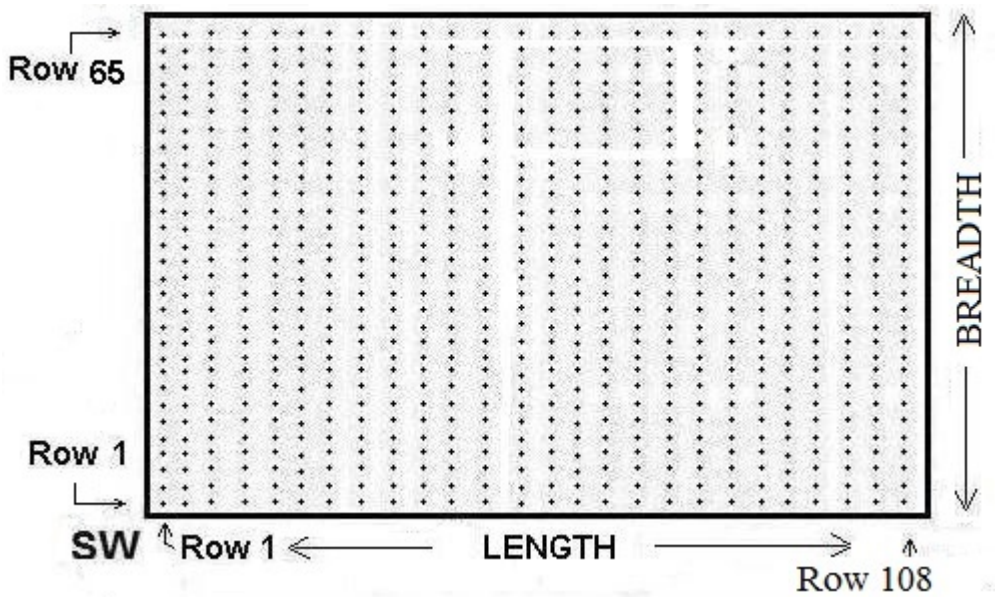


Figure-20.6.5.3.1: Enumeration of Rows

20.6.5.3.2 Average number of rows

Take three physical observations at random at three places in the selected field and count the rows in both the side i.e. length and breadth of experimental plot. The size of experimental plot for physical observation may be the same as for CCE plot (5 x 5 meters or 10 x10 meters or 10 x 5 meter in plains and 10 x 2 meters in hills).Workout the average number in the specified length and breadth of CCE plot.

20.6.5.3.3 Determination of random number for random row in the length direction

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 sample. Deduction of average rows is essential for ensuring that the CCE plot may fall in the selected field. Random number for length side is to be selected using assigned column number. The random number table has to be used as the digit in rows in length side.

Example:

Let, the average number of rows in 5 meter length is 6 and total rows are 108 in the direction of length. For selection of random number for random row, we may calculate the number as follows.

Total number of rows in the direction of length	108
Average rows in 5 meter length of CCE	6
(Total number of rows **minus** average rows in the length of CCE) + one	103

103 is three digit number, therefore, the assigned column number one of three digit random number table may be referred and a number less than or equal to 103 is to be selected. By using the column number one of three digit random number table, 48 is appeared first which is less than 103 and it will be considered as selected.

20.6.5.3.4 Determination of random number for random row in the breadth direction

The similar exercise as done for selection of random row for length may also be followed for selection of random row for breadth.

Example:

Let, the average number of rows in 5 meter breadth is 8 and total rows are 65 in the direction of breadth. For selection of random number for random row, we may calculate the number as follows.

Total number of rows in the direction of breadth	65
Average rows in 5 meter breadth of CCE	8
(Total number of rows **minus** average rows in the breadth of CCE) + one	58

58 is two digit number, therefore, the assigned column number **one** of two digit random number table may be referred and a number less than or equal to 58 is to be selected. By using the column number one of two digit random number table, 22 is appeared first which is less than 58 and it will be considered as selected.

The random number pair is (48, 22)

20.6.5.3.5 Marking of the experimental plot

20.6.5.3.5.1 Marking of south-west corner of the experimental plot

Starting from south-west corner of the selected field, move towards the direction of length of the field by counting and stop at last random row (i.e. 48). From this point move towards the direction of breadth (perpendicular to the length) of the field in between the inter-space of selected row (i.e. 48) and its preceding row (i.e. 47) by counting the rows and stop at random row selected for breadth (i.e. 22). Fix first peg “A” between the interspace of row selected for breadth (i.e. 22) and the preceding row (i.e. 21). The point “A” is the south-west corner (key point) or first corner of the experimental plot (Figure-20.6.5.3.5.1).

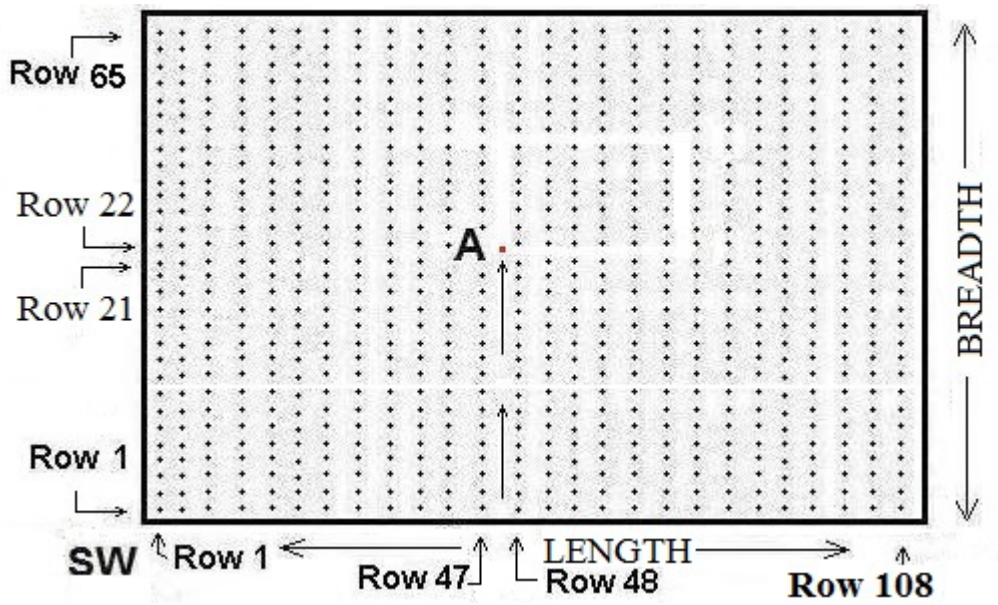


Figure-20.6.5.3.5.1: South-West Corner of Experimental Plot

20.6.5.3.5.2 Marking of second corner of experimental plot

From the key point "A" move in between the interspace of preceding row (i.e. 21) and selected row (i.e. 22) by counting the average number of rows (i.e. 6) in the direction of length of the selected field and stop at random row i.e. 6 which is to be included in CCE plot. Row number 48th of the selected field will be the first row of the CCE plot and last row of CCE plot will be 53th row of the selected field. Fix second peg "B" between the interspace of last row (i.e. 6th and 53th row of the selected field) and its succeeding row number 54th (Figure-20.6.5.3.5.2).

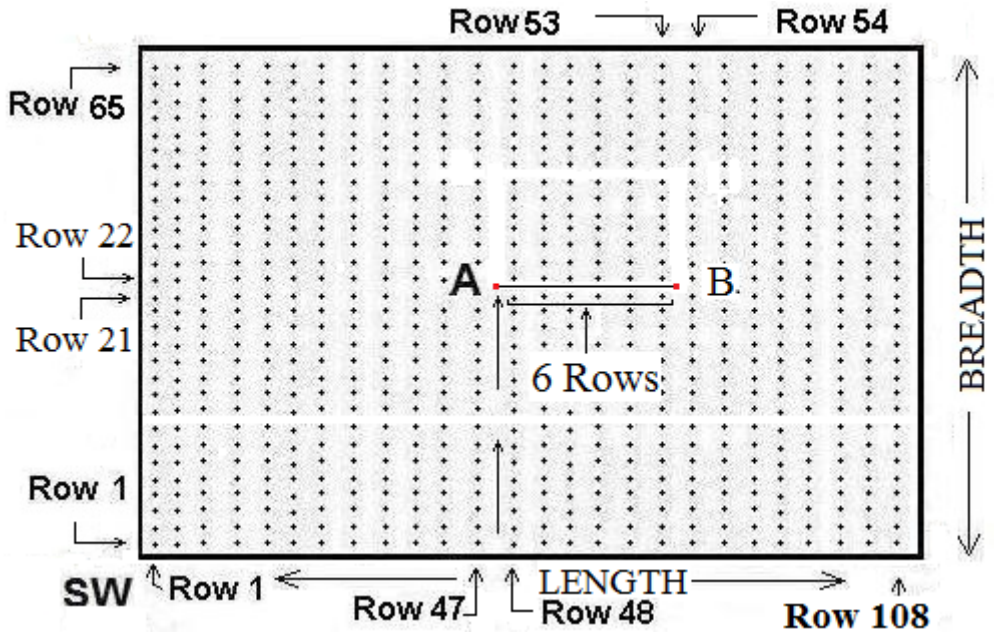


Figure-20.6.5.3.5.2: Second Corner of Experimental Plot

20.6.5.3.5.3 Marking of third corner of experimental plot

From second corner "B" proceed along the breadth of the field in between the interspace of last row of CCE (i.e. 6th row of CCE or row number 53th of selected field) and its succeeding row (i.e. 54) by counting the average number of rows workout in 5 meter in breadth (i.e. 8) and stop at last 8th row to be included in CCE plot (or row number 29 of selected field). We reached in between the interspace of last 8th row 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. Fix third peg here at "C" point (Figure-20.6.5.3.5.3).

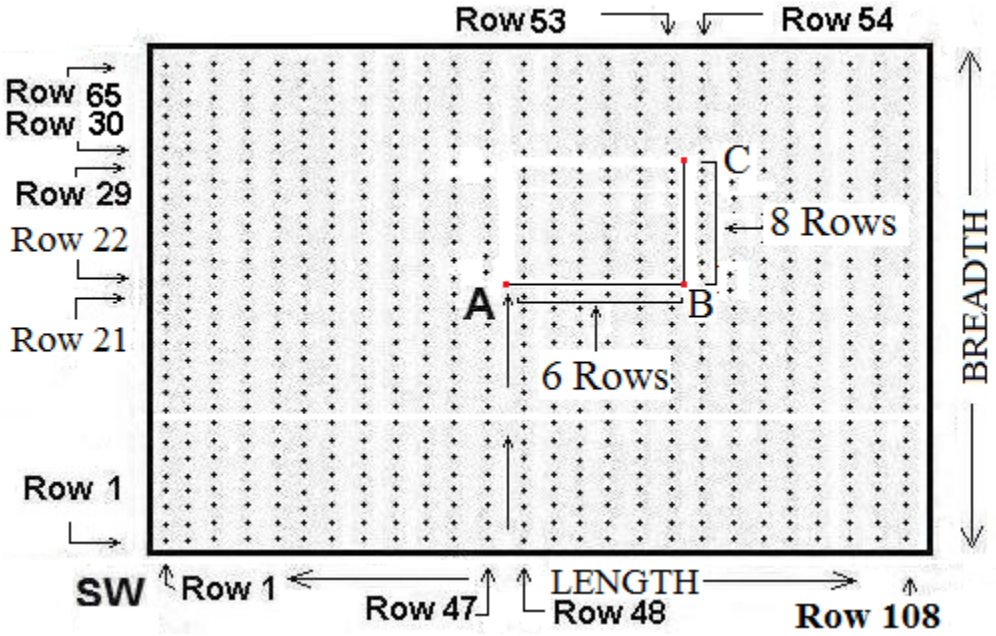


Figure-20.6.5.3.5.3: Third corner of experimental plot

20.6.5.3.5.4 Fourth corner of experimental plot

Proceed from third corner "C" along the interspace of last row (i.e. row number 29) of CCE plot and its succeeding row (i.e. row number 30 of selected field) parallel to "A" & "B" and towards south-west corner of the experimental plot by counting average number of rows in the length of CCE plot (i.e. 6 in 5 meter). We reached back in between the interspace of selected row for length (i.e. 48) and preceding row (i.e. 47). This is the forth corner of CCE plot. Fix the fourth peg "D" here (Figure-20.6.5.3.5.4). Count the rows between "A"-"D". These may be equal to rows between "B"-"C".

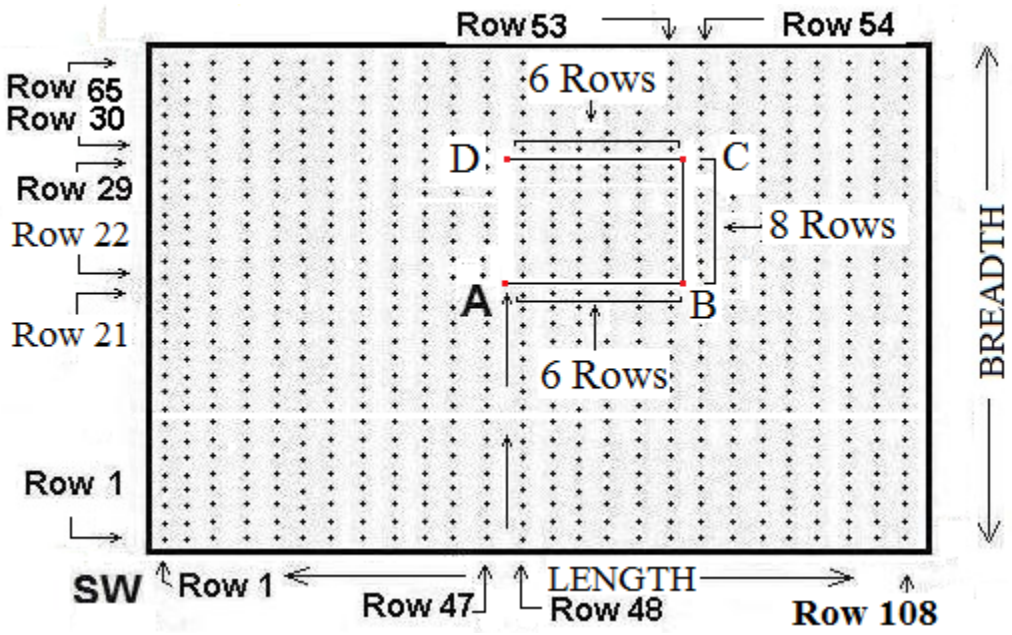


Figure-20.6.5.3.5.4: Fourth corner of experimental plot

20.6.5.3.5.5 Experimental plot

Verify the rows in both the directions i.e. length and breadth. These may be equal to the average number of rows in the sides i.e. 6 rows in length and 8 rows in breadth side. The distance between "A" & "B", "B" & "C", "C" & "D", A & D, "A" & "C" and "B" & "D" may also be measured and noted also. (Figure-6.5.3.6.5). Sides of CCE plot may not be equal to specified length and breadth of CCE plot. Actual length of diagonals (AC and BD) may also be measured and noted.

If the experimental plot does not fall wholly within the field due to irregular shape of the field, reject the first pair of random number and select a new random number pair.

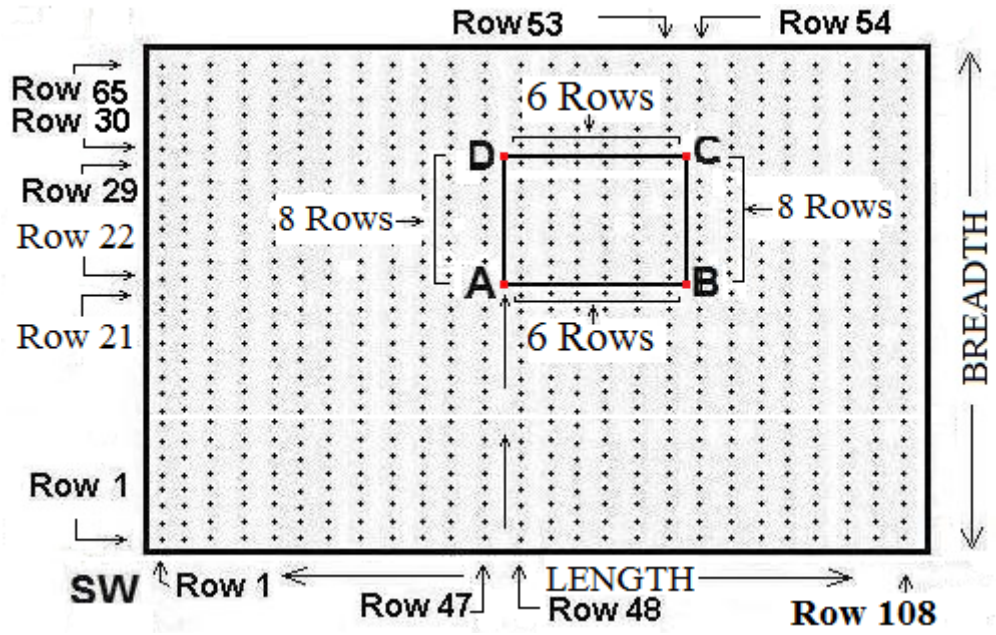


Figure-20.6.5.3.5.5: Experimental plot

20.7.0 Harvesting of experimental crop

It is important that the pegs should be tall, straight and firmly fixed on the ground. The distance outer side of one peg to another peg may be checked carefully. A rope/string should be used for demarcating the boundary of the CCE plot and its length should not be increase on stretching. A well stretch string should be tied around the pegs and it should be lowered gradually to the ground level. The position of the string on the ground demarcates the boundary of the experimental plot. The decision for harvesting the plants is based on the position of their roots, if they are on boundary line. The plants on the boundary line of the experimental plot will be harvested only if the roots are more than half inside the experimental plot. Care should be taken to collect all the harvested plants, bundled it, marked with proper and transported to the threshing floor. No plant and ear head should be fallen during harvesting, bundling and transporting.

20.7.0 Harvesting of experimental crop

20.8.0 Threshing

The harvested experimental crop should be spread on a piece of hessian cloth for drying and threshing the experimental crop. After proper drying, it should be threshed carefully as per the usual method. All grains should be with threshed crop.

20.8.0 Threshing

20.9.0 Winnowing and cleaning

Grains from straw should be separated by winnowing with the help of wind, winnowing fan and other cleaning tool. Materials like seed of other crops, weed seed, dust particles, stone, husk etc. should not be in the produce.

20.9.0 Winnowing and cleaning

20.10.0 Weight of wet produce

Weight of clean produce should be taken just after its winnowing/ cleaning. At this time most of the crops have excess moisture, therefore, this wet 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.

20.10.0 Weight of wet produce

20.11.0 Driage

Driage experiments are necessary, if the produce has more moisture. Sample of recommended quantity of the produce has to be taken in cloth bag and kept for. Driage experiments are necessary to obtain final estimates of yield in terms of dry produce. Driage experiments for different crops are to be conducted by the district statistical officer. Driage experiments for different crops are to be 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 of 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 experimental plot is less than one kilogram, the entire produce is to be taken. In the case of sugarcane, the final produce is expressed in terms of cane only. 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.

20.10.0 Weight of dry produce

Weight of dry produce should be taken after its proper drying. This wet is called as dry 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.

REFERENCES

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.