Statistical Survey Design and Evaluating Impact

Sample surveys are a popular and viable alternative to the official statistics and censuses to collect descriptive information on population characteristics. Sampling designs that enable the selection of a specific portion of the population have been regarded as an effective way to provide reliable and valid estimates of the population parameters.

Two types of statistical designs – sample survey and evaluation – are pivotal tools for research in different disciplines. Sample survey designs focus on representation to capture important characteristics of a population. Evaluation designs emphasize on making the two groups, treatment and control, look as similar as possible to each other in terms of the covariates. Evaluation designs strive for internal validity whereas sample survey designs try to achieve external validity.

Statistical designs are critical in the evaluation of social programmes and policies and in exploring causal relationships. This book discusses some important methodologies to design sample surveys and impact evaluation. Solved examples are included to illustrate each technique discussed in the book. Case studies from around the world highlight applications of the techniques. Sources of biases and ways to overcome them have been outlined in a clear and concise manner. Clear guidelines with remedial measures have been outlined to facilitate choosing a suitable sampling design from among many alternatives.

Tarun Kumar Roy is former Director, International Institute for Population Sciences (IIPS), Mumbai, India. He taught courses on evaluation of family welfare programs and sampling design. He was engaged in the National Family Health Survey, India and also serves as an expert on the Sampling Review Committee for the Global Adult Tobacco Survey (GATS), USA since 2008.

Rajib Acharya is an Associate with the Population Council, New Delhi. He is a statistician and demographer with more than 16 years of experience in designing large sample surveys, sample size determination, processing and managing large scale cross-sectional and longitudinal data sets and analysis of survey data on population, health and nutrition.

Arun Kumar Roy is Chief Executive of Economic Information Technology, a socio-economic research organization in Kolkata. He has served as a Monitoring and Evaluation Consultant to several organizations including the World Bank in Odisha Institutional Council on Management of Population Programs, Malaysia and German Agro Action (GAA), Bonn.

Statistical Survey Design and Evaluating Impact

Tarun Kumar Roy
Rajib Acharya
Arun Kumar Roy

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi - 110002, India

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107146457

First published 2016

Printed in India

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-14645-7 Hardback

To our parents for their invaluable support

Contents

Figures

Tables

Foreword

When Dr Tarun Kumar Roy received a major award for excellence in teaching, on the occasion of the 50th (jubilee) anniversary of the International Institute for Population Sciences in Mumbai, where he was a previous director, the applause from the large audience was thunderous and sustained. It went on for more than ten minutes. The atmosphere was exhilarating and the good will was infectious. I had never before seen such an enthusiastic display of affection for a professor and a scholar, but having worked with Dr Roy for more than twenty years I was not surprised. What causes an individual to garner such respect from a very diverse audience? Dr Roy is the consummate professor and a true guru in his field. His ability to connect with students and researchers on both a cerebral and personal level is a key factor. He is well known for the depth of his knowledge in numerous areas (including statistical survey design, demography, international development, and impact evaluation). His ability to accurately answer questions from students and colleagues in his areas of expertise in simple terms, and his honesty in not being afraid to admit when he does not know an answer, which thankfully is a rare event, has always been highly appreciated. Personally, I have worked on large-scale nationally representative household surveys in 17 low- and middle-income countries, and in all that time I have rarely come across a colleague who is as knowledgeable, dependable, and humble as Professor Roy.

Together with his distinguished co-authors (Rajib Acharya and A. K. Roy), T. K. Roy has managed to translate his ability to connect with students and faculty on multiple levels into an immensely readable book that will be equally useful to beginners and seasoned professionals. The authors' skill in getting the material across to readers comes not only from their solid background in scientific survey sampling methods and statistics but also from their on-the-ground experience with some of the largest and most complex household surveys that have ever been conducted, including four National Family Health Surveys in India (that have included interviews with a total of more than one million respondents). This experience has allowed the authors to go well beyond the theory of sampling statistics to the application of those statistics to real-life situations.

To borrow a phrase from Amitav Banerjee, this book embodies the principle of "statistics without tears". It is both comprehensive and accessible. It covers the gamut from basic sample designs to designs for causal effects. The book also includes chapters on tackling non-sampling errors and statistical tests for measuring impact, that readers will find invaluable. Any discussion of sampling survey design would be incomplete without the attention that this book pays to problems that are likely to be encountered in designing and implementing surveys. The clear,

down-to-earth examples that accompany key topics and the more in-depth case studies will enhance the reader's understanding of the concepts and statistical equations.

This volume will serve as a valuable reference that any researcher, survey planner or student needs to help understand survey statistics, plan high-quality surveys, and make informed use of survey data. It is also an essential compendium that can be relied on to provide answers to a variety of theoretical and practical questions relating to statistical survey design and impact evaluation. The authors' decision to cover such a wide variety of topics in a single volume and to speak to a variety of different audiences will certainly make this book one of the most useful sources of information on survey statistics that has been published in recent years.

Fred Arnold
The Demographic and Health Surveys Programme
Senior Fellow
ICF International
Rockville, Maryland, USA

Preface

The twenty-first century is the *Century of Information*. Scientifically designed gathering, collation, analyses, interpretation, sharing and dissemination of information have become the primary occupation of millions, as the activity encompasses the widest possible spectrum of subjects ranging from education to development including social sciences, developmental economics and planning, medical sciences, engineering and even commercial subjects such as marketing and branding. The two major pillars in this grand edifice are 'survey designs' and 'evaluation designs'. In spite of the significant potential of statistical designs in the development of valid and reliable information, there is a wide and visible gap between development of theory and its practice.

Professor Leslie Kish, one of the world's top survey statisticians, once commented in his paper 'The Hundred Years' Wars of Survey Sampling', "my central complaint is that over 95% of statistical attention in academia, textbooks and publications is devoted to mathematical statistical analysis and only 2% to design" (Kish, 2003a). With regard to attention to designs, his comment is as apt as ever. This is a rather unfortunate situation, particularly since the two designs can compliment and nurture each other's application and growth. Professor Kish continued to remark "the consequences of that neglect are too often poor designs by non-statisticians (engineers, economists etc.)". Given this, the present book is a modest attempt to provide practitioners with tools for application of both designs, irrespective of their field of expertise.

A BRIEF HISTORY OF TWO DESIGNS

Understanding the prevalence of variables and their causal relationships has long been the prime focus of research. Initially, the interest was on obtaining a count of the population, such as adult population for tax purposes as well as for enrolment in the military, availability of land for habitation and agriculture etc. Graunt (1662) estimated the population of London around the year 1662. Utilising parish registers, he first estimated the ratio of number of burials per family. In fact, he observed that there would be 3 burials per 11 families in a year. He also obtained the total number of burials in London. With the help of the total and the ratio, he then estimated the total number of families. An assumption about the average family size then resulted in the estimated total population of the country (Bethlehem, 2009). Laplace used a type of survey to estimate the population of France in around the year 1812. Although he used an

approach similar to Graunt's, his method of estimation of the ratio (the ratio used by him was based on average number of persons per department, an aerial unit) was more systematic and based on selection of a number of areas in the country. These methods can be called *inductive reasoning*, not sampling.

Survey sampling began towards the end of the nineteenth century. During that time and till the 1930s, both probability and purposive sampling, particularly the latter, were in use and there was no basis for providing any idea about the likely error or precision of an estimate. Opinion poll surveys were used by Literary Digest, a news magazine in US between 1890 and 1938. The surveys successfully predicted the presidential candidates in a few elections, but failed to do so in the 1936 U.S. election. The sample size considered for this survey (1936 election) was quite large; more than two million persons participated in it. However, the selection was done purposively. Most respondents for the opinion poll were republicans; majority members of Literary Digest. This fact was not taken into account during the analysis, with the result that the estimate of a republican victory was grossly biased.

Work by Neyman (1934) provided the basis for modern probability sampling. He outlined a procedure for calculating the error of an estimate drawn from a probability sample and he also demonstrated the pitfalls in selection of a purposive sample (also see O'Muircheartaigh and Wong, 1981).

Studies on causal relationship, prior to the 930s, were largely concentrated on obtaining an average value of a dependent variable in the presence and absence of an independent variable. For example, in the 1920s an increase in lung cancer cases was noticed. Studies were undertaken to examine the effect of smoking on the disease. The output variable, proportion of persons suffering from the disease was compared among smokers and non-smokers. Early application of randomised experiments was largely in the field of agriculture, which began in the late 1920s after the pioneering works of Fisher (1918, 1925) and Neyman (1923).

WHY THIS BOOK AND HOW IT IS ORGANIZED

There are two major purposes behind gathering information in the field of social sciences. The first is to estimate prevalence or level of a particular population characteristic and the other is to understand its dispersion in the population. Estimating prevalence can encompass understanding levels of various socio-economic characteristics of a population like education, school attendance among children, household income, occupation etc. or about health related indicators such as prevalence of a particular disease, disability, death or availability of health facilities.

Some of this data is routinely collected by official registration systems, while some other is availed through censuses conducted periodically. Of late, sample surveys have become a viable and important tool for obtaining a variety of information that is required for planning purposes or to serve one time-specific research purposes. In essence, a sample survey entails selection of a portion (sample) from a population so that the collected information from the sample would permit one to infer or generalize about the population. For example, a small sample consisting of a few thousands households would suffice to provide valid information on many of the above for a populace consisting of several millions households.

However estimates obtained from a sample survey are affected by two broad types of errors — the sampling and the non-sampling error. The sampling error occurs because it provides an estimate of a population parameter based on a sample of population units and is

not based on all the units. The non-sampling error, on the other hand, occurs due to reasons other than sampling only a few units from the whole population such as extent of training of the surveyors, strictness with which standard survey procedures were followed, interviewer bias or even bias by respondents. Non-sampling errors are even present in the complete enumeration of the population. With significant cost advantage, an appropriately designed sample survey is often the most preferred option to collect data. In fact, it may be that the total error consisting of both sampling and non-sampling errors in a sample estimate is less than that in the complete enumeration. The demand for sample surveys is continuously increasing as countries get modernized and progress economically. The discussion in this book focuses on how to design (particularly, the procedures for selection of units) and estimation of parameters in a sample survey design. Part I of the book discusses various sample survey designs and related issues.

Understanding dispersion or variations in a characteristic is basically to know why such variations occur. For example, why some people in an area suffer from a disease while others do not, why some people prefer to use a commodity or service while others do not, why people have varying number of children or marry at different ages and so on. This essentially means examining the role of other characteristics that could be related to the study characteristic and thus explaining the variations in it. The study characteristic is often called the 'dependent' or 'output' variable and the other characteristics that are believed to be related to the dependent variable are called 'independent' or 'input' variables.

Examination of associations or correlations between variables can facilitate comprehending the relationship between variables and hence how one influences the other. Higher the association between an independent variable and a dependent variable, greater is the likelihood that it affects the dependent variable and therefore could help in understanding its variations. However, simple associations are not always enough to reveal the "true" or "causal" relationships between variables, particularly in social science since social phenomena are rarely so simple as to be explained by a single factor (Freedman, 1999). The other factors that obfuscate the measurement of causal relationship between the dependent variable and the study variable (independent variable) are known as 'covariates'. Their presence can hamper the measurement of the causal relationship between a study variable on the desired output. It is thus important to understand particular designs that can handle such problems. Such designs are varyingly termed as experimental or evaluation studies and in part II, this book discusses various designs to generate necessary data to facilitate obtaining a causal effect. In this book, we term this design as evaluation design.

The two designs – sample survey design and evaluation design – can help each other in attaining their objectives. Use of evaluation designs, particularly the recent emphasis on evaluating public programs, can be strengthened using appropriate survey designs. It is often necessary to control large number of covariates to isolate a program effect. One viable alternative is to have samples of both treatment and control group selected randomly from a population. Knowledge of survey design will be an added advantage in such cases. Measurement of program effect requires an estimate of its sampling variance. Familiarity with survey design can facilitate this. As surveys become popular, it will be pertinent to ask whether the survey data can be utilized more profitably to understand a causal relationship. The survey data can provide additional assurance about the external validity. Therefore, if a survey can provide the required data to obtain a treatment effect, should it not be preferred? For example, a large scale survey on health might provide the required information to study the effect of smoking on health.

Reducing bias, estimation of errors, and addressing external validity are common concerns in both the designs. Understanding of both, sample surveys and evaluation designs, can go a long way in successfully conducting fruitful researches. The present book makes a humble attempt to emphasize the role of the two designs to promote scientific research. It provides basic knowledge of the various designs to facilitate their applications. In this process, it is necessary to utilize a few propositions that are the basis of measuring an underlying concept. Mention may be made of few concepts such as 'sampling variance', 'bias', and 'error' of an estimate. The emphasis here is to comprehend a concept and know how to measure and interpret it. The details of derivation of a concept have been avoided. Although such theories are essential for the development of a discipline, their avoidance here is mainly to support applications of the designs.

The book is divided into two parts and ten chapters. In part I (Chapters 1–5) we discuss survey design covering several concepts. In Chapter 1, we discuss probability and non-probability sampling designs, concepts of bias and error, and a guideline to choosing desirable design. Chapter 2 discusses simple random sampling, stratification, stratified random sampling, method of and gain in stratification, probability proportional to size (PPS) sampling, cluster sampling and estimation of mean and proportion and their variances in different designs. It also includes an important discussion on role of weights and its calculation. Chapter 3 deals with multistage design, cluster sampling, effect of clustering and provides several useful guidelines in achieving desired design. In Chapter 4, we consider the application of probability sampling in the presence of defective frames or when no frame is available. It also talks about household listing, its importance and alternatives to household listing. We bring in detailed discussions on non-sampling error in Chapter 5.

In part II (Chapters 6–9) we discuss evaluation design. In Chapter 6 we cover types of evaluation designs including a discussion on bias and error. We discuss designs for measuring cause and effect relationship and related estimation issues in Chapter 7. Chapter 8 deals with issues of allocation in order to achieve balance between treatment arms in an evaluation. Chapter 9 discusses, in brief, various statistical tests for measuring impact.

Chapter 10 looks into the practical application of both the designs.

Acknowledgments

This body of work could never be completed without inspiration and support from several individuals who not only motivated the authors to pen down their practical experiences with sampling design, but often also challenged them intellectually thereby enriching the contents of the book. The authors would like to express their gratitude to some of these individuals:

The book would have been impossible without intellectual stimulants from Professor Faujdar Ram, Director and Senior Professor, International Institute for Population Sciences (IIPS), Mumbai, Mr S. K. Das, Ex-Director General, Central Statistical Organisation (CSO), India, Dr Fred Arnold, Senior Fellow, ICF International, Rockville, Maryland, USA, Dr Arvind Pandey, Director, National Institute for Medical Statistics (NIMS), New Delhi, and Dr Shireen Jeejebhoy, Senior Associate, Population Council, New Delhi. The authors are also grateful for the technical support provided by Dr Abhishek Singh, Dr Aparajita Chattopadhyay, Dr Laxmikant Dwivedi, and Kaushalendra Kumar - all from the faculty of IIPS, Mumbai. D. D. Mestri, Library and Information Officer, IIPS, Mumbai helped in multiple ways during the course of drafting the book. Our Special thanks also go to Dr Anupendu Gupta, Ex Deputy Director General, Geological Survey of India for motivating us to write the book in the first place.

The authors are also grateful to a number of organizations for their support before and during the writing of the book. Special mention needs to be made of organizations such as the International Institute of Population Sciences (IIPS), Mumbai, the CDC Foundation, Atlanta, the National Council of Applied Economic Research, New Delhi, the Population Council, New Delhi and Economic Information Technology, Kolkata where the authors developed the necessary skills and expertise for writing this book.

We are also indebted to the unknown reviewers who had devoted their precious time to go through the manuscript and give their valuable comments and suggestions. We also thank all the authors whose works we have referred to in the book.

We will be failing in our duty if we do not appreciate the efforts of Manish Choudhary at Cambridge University Press, New Delhi for seeing us through this book. Thank you Manish Choudhary, for your constant inspiration and support and for being patient with us when we missed several deadlines during the submission and evaluation process of the book.

Special thanks go to Jayati Roy, Madhumita Das and Rita Roy, wives of the authors, without whose supportive roles this book would not have seen the light of the day.

We hope that the book provides sufficient knowledge and stimulants to the readers to develop their own statistical designs for applications in real life situations. It would be fulfilling to see this book being equally helpful to users with or without knowledge of statistics, and users from the development sector as well as academics.