S. P. Mukherjee · Bikas K. Sinha
Asis Kumar Chattopadhyay

Statistical Methods in Social Science Research

Springer

Statistical Methods in Social Science Research

S. P. Mukherjee · Bikas K. Sinha
Asis Kumar Chattopadhyay

Statistical Methods in Social Science Research

Springer

S. P. Mukherjee
Department of Statistics
University of Calcutta
Howrah, West Bengal
India

Asis Kumar Chattopadhyay
Department of Statistics
University of Calcutta
Kolkata, West Bengal
India

Bikas K. Sinha
Indian Statistical Institute
Kolkata, West Bengal
India

ISBN 978-981-13-2145-0 ISBN 978-981-13-2146-7 (eBook)
https://doi.org/10.1007/978-981-13-2146-7

Library of Congress Control Number: 2018950966

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

To a large extent, social science research involves human beings and their groups and associated abstract entities like perception, attitude, personality, analytical skills, methods of teaching, evaluation of performance, merger of cultures, convergence of opinions, extinction or near extinction of a tribe. On the other hand, research in 'hard sciences' like physics, chemistry, or biotechnology involves largely concrete entities and their observable or measurable features.

Social science is a big domain that encompasses psychology, social anthropology, education, political science, economics, and related subjects which have a bearing on societal issues and concerns. Currently, topics like corporate social responsibility (CSR), knowledge management, management of talented or gifted students, leadership, and emotional intelligence have gained a lot of importance and have attracted quite a few research workers. While we have 'special' schools for the mentally challenged children, we do not have mechanisms to properly handle gifted or talented children.

While encompassing economics and political science within its ambit, social science research today challenges many common assumptions in economic theory or political dogmas or principles. Some of the recent research is focused on the extent of altruism—as opposed to selfish motives—among various groups of individuals.

There has been a growing tendency on the part of social science researchers to quantify various concepts and constructs and to subsequently apply methods and tools for quantitative analysis of evidences gathered to throw light on the phenomena being investigated. While this tendency should not be discouraged or curbed, it needs to be pointed out that in many situations such a quantification cannot be done uniquely and differences in findings by different investigators based on the same set of basic evidences may lead to completely unwarranted confusion.

Most of the social science research is empirical in nature and, that way, based on evidences available to research workers. And even when such evidences are culled from reports or other publications on the research theme, some evidences by way of pertinent data throwing light on the underlying phenomena are generally involved. And the quality of such evidences does influence the quality of inferences derived

from them. In fact, evolution of some special statistical tools and even concepts was motivated in the context of data collection and analysis in social science research.

While the dichotomy of research as being qualitative and quantitative is somewhat outmoded, it is generally accepted that any research will involve both induction from factual evidence and deduction of general principles underlying different phenomena and is quite likely to involve both quantitative analysis and qualitative reasoning. In fact, uncertainty being the basic feature of facts and factual evidences about social phenomena, we have to use probabilistic models and statistical tools to make inductive inferences. It is this recognition that can explain two generic observations. The first is that quite a few statistical concepts, methods, and techniques owe their origin to problems which were faced by research workers investigating individual and collective behaviors of humans in different spheres of their activities and the impact of the latter on the economy, the society, and the environment. The second relates to the fact that the social science research has not always taken full advantage of the emerging concepts, methods, and tools in statistics to enhance the substantive—and not just technical—content of research and the findings thereof.

The authors felt the need to apprise research workers in the broad domain of social science about some well-known and widely used statistical techniques besides a few others which are yet to find large-scale applications. Mention may be specifically made about data integration, meta-analysis, content analysis, and multidimensional analysis—topics which have been dealt with in this book with due attention to rigor, simplicity, and user-friendliness.

It is sincerely hoped that this book will benefit research in social science and a feedback from readers will benefit the authors with inputs for improvement in both content and presentation in the future editions. During the preparation of this book, the authors used reference texts, books, journal articles, and their own authored/coauthored research papers—with due acknowledgment of the sources and seeking/securing permission/NOC from the competent persons/authorities.

Kolkata, India
September 2018

S. P. Mukherjee
Bikas K. Sinha
Asis Kumar Chattopadhyay

Contents

1 Introduction 1
- 1.1 The Domain of Social Sciences 1
- 1.2 Problems in Social Science Research 2
- 1.3 Role of Statistics 4
- 1.4 Preview of this Book 6

References and Suggested Readings 11

2 Randomized Response Techniques 13
- 2.1 Introduction 13
- 2.2 Warner's Randomized Response Technique [RRT] 14
- 2.3 Generalizations of RRMs 17
- 2.4 Not-at-Homes: Source of Non-response 18
- 2.5 RRMs—Further Generalizations 18
- 2.6 RRMs for Two Independent Stigmatizing Features 19
- 2.7 Toward Perception of Increased Protection of Confidentiality 20
- 2.8 Confidentiality Protection in the Study of Quantitative Features 22
- 2.9 Concluding Remarks 26

References and Suggested Readings 26

3 Content Analysis 29
- 3.1 Introduction 29
- 3.2 Uses of Content Analysis 30
- 3.3 Steps in Content Analysis 30
- 3.4 Reliability of Coded Data 31
- 3.5 Limitations of Content Analysis 36
- 3.6 Concluding Remarks 36

References and Suggested Readings 37

4 Scaling Techniques . . . 39
4.1 Introduction . . . 39
4.2 Scaling of Test Scores . . . 40
4.2.1 Percentile Scaling . . . 41
4.2.2 Z-Scaling or σ-Scaling . . . 41
4.2.3 T-Scaling . . . 41
4.2.4 Method of Equivalent Scores . . . 42
4.3 Scaling of Categorical Responses . . . 45
4.3.1 Estimation of Boundaries . . . 45
4.3.2 Finding Scale Values . . . 46
4.4 Use of U-Shaped Distributions . . . 46
4.5 Product Scaling . . . 47
4.6 Other Unidimensional Scaling Methods . . . 50
4.7 Concluding Remarks . . . 52
References and Suggested Readings . . . 52

5 Data Integration Techniques . . . 53
5.1 Introduction . . . 53
5.2 Elementary Methods for Data Integration . . . 53
5.3 Topsis Method: Computational Algorithm in a Theoretical Framework and Related Issues . . . 55
5.4 Topsis Method: Computational Details in an Illustrative Example . . . 56
References and Suggested Readings . . . 60

6 Statistical Assessment of Agreement . . . 61
6.1 General Introduction to Agreement . . . 61
6.2 Cohen's Kappa Coefficient and Its Generalizations: An Exemplary Use . . . 62
6.3 Assessment of Agreement in Case of Quantitative Responses . . . 66
References and Suggested Readings . . . 67

7 Meta-Analysis . . . 69
7.1 Introduction . . . 69
7.2 Estimation of Common Bernoulli Parameter "p" . . . 70
7.3 Estimation of Common Mean of Several Normal Populations . . . 70
7.4 Meta-Analysis in Regression Models . . . 72
References and Suggested Readings . . . 74

8 Cluster and Discriminant Analysis . . . 75
8.1 Introduction . . . 75
8.2 Hierarchical Clustering Technique . . . 76
8.2.1 Agglomerative Methods . . . 76

8.2.2 Similarity for Any Type of Data . . . 77
8.2.3 Linkage Measures . . . 78
8.2.4 Optimum Number of Clusters . . . 79
8.2.5 Clustering of Variables . . . 79
8.3 Partitioning Clustering-k-Means Method . . . 80
8.4 Classification and Discrimination . . . 81
8.5 Data . . . 83
References and Suggested Readings . . . 94

9 Principal Component Analysis . . . 95
9.1 Introduction . . . 95
9.1.1 Method . . . 96
9.1.2 The Correlation Vector Diagram (Biplot, Gabriel 1971) . . . 99
9.2 Properties of Principal Components . . . 101
References and Suggested Readings . . . 102

10 Factor Analysis . . . 103
10.1 Factor Analysis . . . 103
10.1.1 Method of Estimation . . . 106
10.1.2 Factor Rotation . . . 108
10.1.3 Varimax Rotation . . . 108
10.2 Quartimax Rotation . . . 109
10.3 Promax Rotation . . . 109
References and Suggested Readings . . . 111

11 Multidimensional Scaling . . . 113
11.1 Introduction . . . 113
11.2 Types of MDS . . . 114
11.2.1 Non-metric MDS . . . 115
11.2.2 Replicated MDS . . . 115
11.2.3 Weighted MDS . . . 115
11.2.4 Sammon Mapping . . . 116
11.3 MDS and Factor Analysis . . . 116
11.4 Distance Matrix . . . 117
11.5 Goodness of Fit . . . 119
11.6 An Illustration . . . 120
11.7 Metric CMDS . . . 121
References and Suggested Readings . . . 122

12 Social and Occupational Mobility . . . 123
12.1 Introduction . . . 123
12.2 Model 1 . . . 126
12.2.1 Some Perfect Situations . . . 127
12.2.2 Possible Measures of Career Pattern . . . 127

12.2.3 Measure of Career Pattern Based on Mahalanobis Distance . . . 128
12.2.4 Measure of Career Pattern Based on Entropy . . . 129
12.2.5 An Example . . . 130
12.3 Model 2 . . . 131
References and Suggested Readings . . . 132

13 Social Network Analysis . . . 135
13.1 Introduction . . . 135
13.2 Sampling and Inference in a SN . . . 137
13.3 Data Structure in a Random Sample of Units . . . 138
13.4 Inference Procedure . . . 140
13.5 Estimation of Average Out-Degree Based on Data Type – 1/2 . . . 140
13.6 Inference Formulae for Data Type 3 Using Sample Size n . . . 142
13.7 Computations for the Hypothetical Example . . . 144
13.8 Estimation of Average Reciprocity . . . 145
References and Suggested Readings . . . 152

About the Authors

S. P. Mukherjee retired as Centenary Professor of statistics at Calcutta University, Kolkata, India, where he was involved in teaching, research, and promotional work in the areas of statistics and operational research for more than 35 years. He was Founder Secretary of the Indian Association for Productivity, Quality and Reliability and is now its Mentor. He received the Eminent Teacher Award (2006) from Calcutta University, P.C. Mahalanobis Birth Centenary Award from the Indian Science Congress Association (2000), and Sukhatme Memorial Award for Senior Statisticians from the Government of India (2013). A Fellow of the National Academy of Sciences of India, he has about 80 research papers to his credit. He was Vice-President of the International Federation of Operational Research Societies (IFORS).

Bikas K. Sinha was attached to the Indian Statistical Institute (ISI), Kolkata, India, for more than 30 years until his retirement in 2011. He has traveled extensively within USA and Europe for collaborative research and teaching assignments. He has more than 140 research articles published in peer-reviewed journals and almost 100 research collaborators worldwide. His research interests cover a wide range of theoretical and applied statistics. He has coauthored three volumes on Optimal Designs in Springer Lecture Notes Series in Statistics (vol. 54 in 1989, vol. 163 in 2002, and vol. 1028 in 2014) and another volume on Optimal Designs in Springer Textbook Series (2015).

Asis Kumar Chattopadhyay is Professor of statistics at Calcutta University, Kolkata, India, where he obtained his Ph.D. in statistics. With over 50 papers in respected international journals, proceedings, and edited volumes, he has published three books on statistics including the two published by Springer—*Statistical Methods for Astronomical Data Analysis* (Springer Series in Astrostatistics, 2014) and *Statistics and its Applications: Platinum Jubilee Conference*, Kolkata, India, December 2016 (Springer Proceedings in Mathematics & Statistics, 2018). His main interests are in stochastic modeling, demography, operations research, and astrostatistics.

Chapter 1
Introduction

1.1 The Domain of Social Sciences

Social sciences correspond to a vast and rapidly growing area that encompasses investigations into diverse phenomena happening in the society, the economy, and the environment. In fact, social sciences deal with people—individuals, groups, or firms. As Bhattacharjee (2012) puts it, social sciences taken as a single branch of knowledge define the science of people or collection of people such as cultural groups, trading firms, learned societies, or market economies and their individual or collective behavior. That way social science embraces psychology (the science of human behavior), sociology (the science of social groups), political science (dealing with political groups), and economics (the science of firms, markets, and economies).

Some of the phenomena studied in social sciences are too complex to admit concrete statements; on some we cannot have direct observations or measurements; some are culture (or region) specific while others are generic and common. Data including laboratory measurements, survey observations, responses to questions, documents, artifacts, mission and vision statements and similar entities available in social sciences for scientific investigations into the 'behavior' phenomenon are so vague, uncertain, and error-prone that methods of investigation and techniques applied in physical sciences cannot be immediately used without necessary modifications. In fact, disagreements among observers or investigators on the same features of the same individuals are quite considerable, and it becomes difficult to generalize findings or conclusions based on a single set of data.

Measurements play an important role in any scientific investigation, to the extent that the quality and adequacy of pertinent measurements do affect the credibility of findings from the investigation. Measurement in the social sciences may be conceived as a process linking abstract concepts to empirical indicators. It transforms concepts into accounting indicators or schemes. The following phases in this transformation can be clearly identified.

1. The abstract definition of the phenomenon or concept that is to be studied.

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_1

2. The breakdown of the original concept into 'constituent concepts' or 'dimensions.' The original concept corresponds, more often than not, to a complex set of phenomena rather than to a single directly observable phenomenon.
3. An indicator is assigned to each dimension.
4. Usually, an aggregate indicator is developed, unless characteristics of the phenomenon do not justify the construction of some synthetic indicator. In other cases, the aggregate indicator entails construction of an accounting scheme, as for instance a social accounting matrix or accounts of employment or of health.

All this implies that measurement in the context of social phenomena involves aspects of both a theoretical and an empirical character. Data are needed to construct and validate theories, at the same time theories are needed to generate and validate data.

The breakdown of a phenomenon into measurable dimensions is rarely unique, in terms of the number of dimensions—preferably non-overlapping or un-correlated—and their identification in terms of data-based indicators. The problem becomes more complicated when the phenomenon is dynamic, and we can develop a reasonable breakdown at any point of time which may not be a reasonable representation of the phenomenon at a subsequent time point. In some cases, the dimensions are not really amenable to a direct enumeration or even identification. For example, when we have to deal with feelings, aptitudes, and perceptions, we construct scales by assuming certain continua and by noting the responses to some questions believe to reveal the chosen dimension.

1.2 Problems in Social Science Research

While scientific studies are invariably concerned with 'variations' in some features or characteristics across individuals and groups, over time and over space, in the context of social sciences many of these features which vary randomly are only 'latent' variables, unlike 'manifest' variables studied in physical or biological phenomena.

Let us consider a typical theme for research, viz. greater frustration among highly educated young persons about the prevailing employment situation than among people with lower levels of education and/or with lesser ambitions in life. To examine the applicability or validity of this proposition in a particular society or region or some suitably defined group, we need evidences bearing on entities like 'ambition,' 'levels of education,' 'frustration,' and 'perceived employment situation' in respect of some individuals in a 'sample' that adequately represents the group or population in relation to which the validity of the proposition was to be examined. And the first and the third features defy unique and objective definitions and, subsequently, measures. Evidently, any form of analysis based on some evidences collected on such latent variables will attract a lot of uncertainty. However, we cannot take our hands off and have to try out some reasonable surrogates or substitutes which are manifest and can be quantified. Of course, the choice of surrogates for 'ambitions' and 'frustration' is not unique, and the responses that are likely to arise to some questions carefully

constructed to reflect on these latent variables cannot be scaled uniquely and cannot be subsequently summarized uniquely. We have to keep in mind this non-uniqueness associated with evidences that are most often necessitated in social science research.

In psychology, we talk of psychophysical experiments essentially dealing with responses to various stimuli. In education, we sometimes conduct an experiment to find out which of several alternative ways of teaching a new language is the most effective. In political science, we can think of an experiment to conduct an election in several alternative ways to identify the most preferred alternative. And rarely will experts or referees or judges will agree on the most effective or most preferred or most likely alternative. Such differences in assessment is just natural, and the confusion or inconsistency arising from such disagreement is unavoidable.

Dealing quite often with latent variables which are quantified in various equivocal terms and based on relatively small sample sizes, conclusions reached in many social science research studies are hardly 'reproducible' and hence are hardly 'scientific.' At the same time, we cannot drop all such latent variables or variables which defy unique quantification from our investigations and we deal with multiple variables in any study that make it difficult to determine the sample size that will be adequate to provide credible inferences regarding the many parameters that have to be estimated or all the hypotheses to be tested, except in terms of a number (of units) that will be too resource-intensive to really canvas.

Several so-called international agencies which have recently mushroomed and which attempt to rank different countries in terms of 'abstract' entities like 'charity-giving' only serve to dish out unscientific findings that cannot carry any conviction, but can be used wrongly by some interest groups to portray some countries poorly or in a lofty manner.

The choice of indicators based most often on some proxies or surrogates of the feature or characteristic understudy is not unique, and there is hardly any criterion to accept on in preference to another. Sometimes, a wrongly chosen indicator has led to lack of credibility of the final result based on an index that combines the various indicators. Earlier, the United Nations Development Programme took 'mean years of schooling' as an indicator of educational attainment of a country, to be taken along with the percentage of literates among adults. One should note that mean years of schooling for an individual as also for a group may increase as a consequence of stagnation and, that way, may be a negative indicator of educational attainment.

Evidences bearing on different social or cultural phenomena are mostly gathered through sample surveys, and an important decision to be taken in this regard is the choice of an appropriate sampling design to come up with an adequate sample size that can ensure credible estimates of the different parameters of interest and tests of different hypotheses with reasonable power. It is not uncommon to find a small sample used to come up with a general statement that can hardly beget any credibility.

The choice and use of an appropriate sampling design to suit the purpose of a sample survey throwing up adequate evidences of reasonable quality to make valid inferences is a bad necessity. And the inferences are to be valid in respect of a certain 'population' in which the investigator is interested and from which the sample has to be drawn. Thus, delineating the population of interest is a primary task, and in

social science investigations there could arise situations where this task is quite complicated. For example, if a national sample survey is to be conducted for getting a good estimate of the number of persons suffering from a certain disease which attracts some taboo, the problem of delineating the population of interest—which should not be the general population—poses serious problems.

Another big issue concerns the size and selection of the sample used in surveys to collect data on both measurable features of individual respondents as also on traits possessed by them that evade direct measurements. The sample must be large enough to make the findings reproducible, and the data must be collected with due care to secure proper evidences that can throw light on the underlying phenomenon or phenomena. Findings of many investigations fail to become reproducible because of shortcomings in such surveys.

1.3 Role of Statistics

Statistics, being a scientific method—as distinct from a 'science' related to one type of phenomena—is called for to make inductive inferences regarding various phenomena like social tension, frustration among educated youths, exploitation and consequent feeling of alienation among neglected tribals, erosion of patriotic feelings among the young these days, religious fanaticism leading to tensions in the society, loyalty of middle-income customers to some brands of a consumer good, loss of credibility of democratic institutions over time, etc., based on evidences gathered.

In the context of a growing public demand for more credible and insightful view of distributive justice, and better and more comprehensive analysis of long-term and wide-area effects and outcomes of social expenditure by different agents, contemporary research has to come up with reasonable and defensible answers to such questions as: How does education affect employment? Does business development have an impact on crimes? To what extent are family formations and decisions are affected by economic prospects and employment security? What are the implications of a forward-looking prevention policy in health, long-term care, and the elderly?

It is true that social scientists are aware of the fact that answers to such questions are bound to be somewhat specific about time, space, culture, and other considerations. However, howsoever the group of interest may be defined, it will be surely larger—and, in some cases, much larger—than the 'sample' that can be conveniently canvassed in any research investigation. Thus, the need for inductive inferences based on evidences and some models is strongly felt.

Inductive inferences are made or have to be made in several distinct situations, viz.

(1) we have limited evidences available on a phenomenon, and we like to go from this sample of evidences to make a conclusion about the phenomenon itself (that really corresponds to an infinite population of evidences that can arise, at least in theory).

(2) we observe the currently available in a damaged or an altered set of evidences pertaining to a phenomenon that occurred in the past and we like to infer about some aspect(s) of that past phenomenon.

(3) we have observations relating to a phenomenon revealed in the recent past or currently and we like to infer about how it unfolds in the future.

We must bear in mind the fact that in induction - unlike in deduction - premises provide some support to the conclusion or inference made on the basis of evidences available along with some 'model' for processing the evidences. In Deduction, the conclusion is warranted by the premises. This implies that with any inductive inference is assumed associated some amount of uncertainty, due both to uncertainties in the evidences made use of as also the uncertainty inherent in the use of statistical tools for processing the evidences.

This inferential uncertainty has to be quantified if alternative ways for processing of evidences or even if different sets of evidences bearing on the same phenomenon are to considered. And the concept of probability is brought in to quantify uncertainty involved in a given exercise in inductive inference. Evidential uncertainty is also handled in terms of fuzziness and related measures.

While statistical methods and techniques deal essentially with 'variations' in some features or characteristics across individuals and groups, over time and over space, to bring out a pattern behind such variations which can be taken further to offer an explanation of the observed variation, in the context of social sciences many of these features which vary randomly are only 'latent' variables, unlike 'manifest' variables studied in physical or biological phenomena and even those which are 'manifest' may be mostly 'categorical' or even 'nominal' to which standard statistical techniques cannot directly apply without some necessary modification. More often than not, social phenomena reveal interrelations among constructs or variables bearing on them which cannot be studied in terms of usual dependence analysis. Variables involved can be classified as endogenous and exogenous, after delineating the boundaries of the system in which the study is embedded, while the classification as dependent and independent is not pertinent.

Statistics—meaning both statistical data as also statistical reasoning—are becoming active partners in the world of social science research, promoting and supporting, using and questioning ongoing theoretical studies. Statistics not only provides valuable empirical evidence against which theoretical constructs can be tested, but also theoretical frameworks putting them to the test of the measurement process. Theories, in fact, are the main ingredients for developing the conceptual frameworks underlying the quantification of social phenomena. Their viability and effectiveness to cope with the dynamism and comprehensiveness of social change represents a crucial test of their validity. Theories are validated by empirical data and, therefore, the quality of data made use of in this context is a vital issue. Only close collaboration between social scientists and statisticians can bring about improvements in social statistics and, that way, in social science researches.

As is the case with researches in other domains, social science research generally—if not necessarily—involves collection, aggregation, and analysis of multiple characteristics or features exhibited by the individuals or units in the group under inves-