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

© Springer Nature Singapore Pte Ltd. 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

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.

© Springer Nature Singapore Pte Ltd. 2018

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-

tigation. In fact, factor analysis as an important technique for analyzing multivariate data was introduced in the context of psychological investigations to identify factors or traits which explain observed correlations between different pairs of subjects in which individuals have been tested. Methods of clustering and classification also had their initial applications in social investigations to identify homogeneous groups based on the different features of the individuals. Multi-dimensional scaling as an important tool for data visualization cropped up in connection with linguistic ability studies and related aspects of human behavior. Several techniques like conjoint analysis were developed during researches on consumer behavior.

In recent times, we quite often access data on multiple attributes based on which we like to compare several entities like different locations or institutions or societies or strategies or deployment plan, etc., and assign ranks to these entities so that we can identify and concentrate on the 'best' or the 'worst' situation needing 'urgent' or 'convenient' intervention. In fact, such multi-attribute decision problems are of great interest and importance in social science research. Indeed, before we can pool data on the same phenomena from different sources—and such data could be purely qualitative in character like opinions or judgements or ranks etc—we should examine the extent to which they agree among themselves. Similarly, meta-analysis or analysis of analyses carried out on the same phenomenon by different researchers possibly following different models and methods is required to ensure consolidation of analyses to enhance the substantial content of any social research study.

We also get data on social interactions among individuals or on decisions of individuals and groups to move from one place or one profession or one job to another. Such data can reveal important latent features about the individuals as also about groups on proper analysis by techniques which have emerged over the years.

1.4 Preview of this Book

This book is not intended to be a standard textbook on the subject of statistics for social science research covering all types of phenomena studied in social sciences and the whole gamut of statistical techniques that are being used or are required to be used in that context. It is just a supplementary reading covering only some selected techniques which are widely applied and often warranted in some areas of social science research. In fact, some researches in social sciences have led to the development and use of some of the methods and techniques discussed in this book. Content Analysis is one such example. Several techniques in multivariate data analysis owe their origins to psychology, e.g., factor analysis. The same is true about scaling techniques originally used in the context of psychological tests. While Management Science may not be regarded as a component of social sciences, research in marketing has to deal with human behavior like preferences for certain brands or grades of a certain product when it comes to a purchase decision. And there should be no reservation to accept such studies as research in social science.

Thus, product scaling and multi-dimensional scaling are statistical techniques which are found useful in marketing research.

Techniques dealt with in this book range from those which relate to problems of data credibility in studies in which confidentiality is a major concern and responses are likely to be untrue to techniques involved in pooling data from different sources, from simple scoring of responses to items in a questionnaire used in an opinion survey to analysis of multivariate data. Some of these techniques are of relatively recent origin, while several others have found their way in social sciences as well as in other areas of research quite some time back.

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 throws up theoretical frameworks putting them to the test of the measurement process. Social science research is primarily empirical in character and inferences made about a whole lot of social phenomena are inductive in nature, being based on data which are quite often subjective. Such data-based inferences, taking for granted some postulates and some model behavior of the data, do naturally use relevant statistical techniques and corresponding softwares.

Sometimes a distinction is made between qualitative and quantitative research. It is difficult to illustrate purely qualitative research, except to indicate that qualitative and logical thinking to draw conclusions from the data in hand coupled with qualitative interpretation of such conclusions in the context of the phenomenon or group of related phenomena also constitute useful research in social science. There has been a growing tendency among researchers these days to quantify many constructs and features (variables) that defy direct or unequivocal quantification. While it is true that statistical methods and techniques are involved in a quantitative analysis, it must be remembered that such methods and techniques should enhance the substantive content of research and not just the technical content. The latter objective may call for application of latest available statistical techniques, while the former focuses on a pragmatic and, may be, limited use of such techniques only to derive strength from whatever constitute the premises for making inferences about the phenomena under investigation.

Right from planning a data-gathering exercise, through making the data collected and documented amenable to quantitative analysis and carrying out necessary test for 'poolability' of data gathered from different sources, getting appropriate analysis done on the data as eventually accepted, to reaching evidence-based inferences and interpreting results in the context of the research project, we need to-day Statistics in every stage—imaginatively and effectively—to enhance not merely the technical content of the study but also its substantive content. Data visualization as being somewhat similar to and still distinct from dimension reduction is quite useful in exploratory research on certain types of social phenomena like disagreements among judges in assessing relative positions of certain objects or subjects in terms of some relevant attributes.

There arise problems in getting correct or truthful responses to questions pertaining to sensitive or confidential items like consumption of drugs or income from dubious sources or unusual behavior, etc., and in such cases, randomised response techniques [RRTs] are used in some studies to extract reasonable estimates of some parameters of interest without asking direct questions on the underlying issue. RRT has been treated in Chap. 2 for both qualitative and quantitative data.

Before embarking on any quantitative analysis, qualitative analysis often helps to answer certain questions relating to disputed authorship of some piece of literature or to the trend in public opinion about some contemporary issue like limits to freedom of speech and the like. In such cases, data are scattered in some reports or recorded speeches or other artefacts like photographs or banners. We can think of a content analysis (taken up in Chap. 3) to come up with some sensible answers to some vexing questions that can evade a sophisticated approach to secure a 'correct' answer.

In many studies on opinions or skills or competencies and similar other attributes, we use tests or instruments to develop some measures in terms of scores assigned to responses to different items. And these scores in different subjects, in different environments and in different times, may not be comparable and we need to scale them properly before we can make use of the scores for any decision or action. In many socioeconomic enquiries, for example, an organizational climate survey to bring out employees' perceptions about the climate for work prevailing within the organization, we often try to seek responses from individual employees on a statement like 'I get full cooperation from my peers and colleagues in discharging my responsibilities.' Each respondent is to tick one of five possible categories to indicate his/her perception about this issue, viz. strongly disagree, disagree, undecided or indifferent, agree, and strongly agree. The number of response categories could be seven or nine or some other odd integer. Different scaling techniques have been discussed in Chap. 4.

Recent times see a wide variety of data streaming in from different sources to throw light on the same phenomenon may be dispersed over different locations or institutions or groups. In respect of each such location/institution/group, the data may not be all equally revealing about the nature and magnitude of the phenomenon under study. We are required to rank these different entities to identify the 'best' and the 'worst' situations, so that we can prioritize our interventions in them accordingly. There are a few techniques for this multi-attribute decision-making problem, and we focus on a widely used technique, viz. TOPSIS where the concept of an 'ideal' situation and distances of different situations from the 'ideal' are the components. The use of this technique for data integration has been explained with an illustration from environmental pollution data in Chap. 5.

Chapter 6 is devoted to an emerging topic of judging quantitatively agreement among different raters or experts or judges in situations like diagnosis by several medical men of a disease some patient is suffering from, or reliability of a test battery in a psychometric test as judged by a group of subject experts, or relative importance of a particular feature of a product or service in assessing the latter's quality in the eyes of several potential customers, or opinions expressed by several political analysts on the likely impact of an agreement signed between two countries on international trade, and the like. Any attempt to pool the assessments or ratings

or diagnoses has to be preceded by a statistical assessment of agreement among the opinions or judgements. This is not the same as application of Delphi or similar techniques to make the ratings or opinions or judgements converge. All this has been explained with ample illustrations in Chap. 6.

Meta-analysis is another recent paradigm in social science research. Here the idea is to make use of all the available evidence which may be in the form of several pieces of information (derived from some data) from different sources, some of which may be in the form of expert opinions. Evaluation of each piece of information enables us to determine the weight to be attached to it in pooling information. However, pooling information demands that different pieces of information are not conflicting with each other. Finally, we have to choose an appropriate method to combine the different pieces of information and express the reliability of the final conclusion. This is the content of Chap. 7.

Coming to data analysis, it must be admitted that such data are necessarily multivariate and, more often than not, the data set covers a large number of units or individuals which differ among themselves to different extents in respect of several observable or measurable features which are correlated to different extents one with the others. It will be desirable to group such units or individuals into homogeneous clusters before we analyze relations among the variable features within each cluster separately. And, we should even start looking at the variables themselves before we subject them to further analysis. Toward that reduction in dimensions, we may profitably use factor analysis as also principal component analysis, and we identify and extract artificial combinations or components that can be carefully interpreted in terms of the research objectives.

Whenever a research encompasses more than one periods of time or, say, generations of individuals, we may be interested in noting the changes or transitions of the individuals across social groups or occupations. Such mobility studies are also quite useful in market research to reveal customer loyalty to certain product or service brands using a 'mover–stayer model.' In such mobility studies, Markov Chains and related tools play an important role. In fact, Renewal-Reward Process models have been used in studies on occupational mobility. Chapter 12 is devoted to the subject of social and occupational mobility along with some related issues in manpower planning.

Coming to other aspects of data analysis, it must be admitted that such data are necessarily multivariate and, more often than not, the data set covers a large number of units or individuals which differ among themselves to different extents in respect of several observable or measurable features which are correlated to different extents one with the others. It will be desirable to group such units or individuals into homogeneous clusters before we analyze relations among the variable features within each cluster separately. And, we should even start looking at the variables themselves before we subject them to further analysis. Toward that reduction in dimensions, we may profitably use factor analysis as also principal component analysis to identify and extract artificial combinations or components that can be carefully interpreted in terms of the research objectives. In Chap. 8, we deal with clustering techniques and

Discriminant Analysis while in Chap. 9, we discuss principal component analysis, and in Chap. 10, we take up study of factor analysis.

Under multivariate analysis, two very important techniques are clustering and classification. Under the problem of clustering, we try to find out the unknown number of homogeneous inherent groups in a data set as well as the structure of the groups. But under classification, the basic problem is discrimination of objects into some known groups. One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. Classification is fundamental to most branches of science. The information on which the derived classification is based is generally a set of variable values recorded for each object or individual in the investigation, and clusters are constructed so that individuals within clusters are similar with respect to their variable values and different from individuals in other clusters. The second set of statistical techniques concerned with grouping is known as discriminant or assignment methods. Here the classification scheme is known a priori and the problem is how to devise rules for allocating unclassified individuals to one or other of the known classes.

Principal component analysis (PCA) is a dimension reduction procedure. The method is useful when we have obtained data on a number of variables (possibly a large number of variables), and believe that there is some redundancy in those variables. In this case, redundancy means that some of the variables are highly correlated with one another, possibly because they are measuring the same phenomenon. Because of this redundancy, it should be possible to reduce the observed variables into a smaller number of principal components (artificial variables) that will account for most of the variance in the observed variables.

Factor analysis presented in Chap. 10 is a statistical method used to study the dimensionality of a set of variables. In factor analysis, latent variables represent unobserved constructs and are referred to as factors. Factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables. It is often used in data reduction to identify a small number of factors that explain most of the variance that is observed in a much larger number of manifest variables. Its basic difference from principal component analysis (PCA) is that in PCA variables are replaced by a small number of linear combinations which are expected to explain a larger part of the variation, but it is usually not possible to correlate these linear combinations with some physical phenomena. But in factor analysis, the newly derived latent variables are extracted as factors representing some physical phenomena. Given a set of scores for a group of persons corresponding to aptitude tests in subjects like mathematics, physics, statistics and their performances in 100-m race, long jump, high jump, etc., one may extract two latent factors, viz. intellectual ability and physical ability.

There are situations where we like to compare entities like music, or dance or an object of art or just any product available in many variants or brands and we need to scale these entities (generally called products) to get an idea about the relative merits of the different entities or relative distances between them on a straight line or a two- or three-dimensional surface. We have one-dimensional scaling provided by Thurstone's Law of Comparative Judgment, further taken up by Mosteller and

others. To get a better visualization of the relative distances or dissimilarities among the entities, multi-dimensional scaling was introduced by Torgersen. In Chap. 11, we discuss about this aspect of data analysis.

Whenever a research encompasses more than one periods of time or, say, generations of individuals, we may be interested in noting the changes or transitions of the individuals across social groups or occupations. Such mobility studies are also quite useful in market research to reveal customer loyalty to certain product or service brands using a 'mover–stayer model.' In such mobility studies, Markov Chains and related tools play an important role. In fact, Renewal-Reward Process models have been used in studies on occupational mobility. Chapter 12 is devoted to the subject of social and occupational mobility along with some related issues in manpower planning.

Social network refers to the articulation of a social relationship, ascribed or achieved, among individuals, families, households, villages, communities, regions, etc. The study of social networks is a fast widening multidisciplinary area involving social, mathematical, statistical, and computer sciences. It has its own parameters and methodological issues and tools. Social network analysis (abbreviated SNA) means an analysis of various characteristic of the pattern of distribution of relational ties in a social group and drawing inferences about the network as a whole or about those belonging to it considered individually or in groups. Bandyopadhyay et al. (2009) have discussed in detail how graph–theoretical and statistical techniques can be used to study some important parameters of global social networks and illustrate their uses in social science studies with some examples derived from real-life surveys. In Chap. 13, we consider a few features or characteristics of a social network and explain how these features can be measured. Then we discuss the possibility of using sampling techniques in case of large networks.

References and Suggested Readings

Bandyopadhyay, A., Rao, A. R., & Sinha, B. K. (2009). *Statistical methods for social networks with applications*. California, USA: Sage Publication.

Bhattacharjee, A. (2012). Social science research: principles, methods and practices. Creative commons attribution - non-commercial-share alike.

Garrona, P., & Triacca, U. (1999). Social change: Measurement and theory. *International Statistical Review*, *67*(1), 49–62. Research Eds. H. M. Blalock, & A. B. Blalock (pp. 5–27). New York: McGraw Hill.

Lazarsfeld, P. F. (1965). Le vocabulaire des siences sociales. In R. Boudon, & P. F. Lazarsfeld (Eds.), *Methodes de la sociologie* (Vol. 1). Paris-La Hayne: Mouton.

Chapter 2
Randomized Response Techniques

2.1 Introduction

Almost half a century back, randomized response technique/methodology [RRT/RRM] was first introduced and popularized by Warner. The idea is to be able to elicit a truthful response on sensitive issues(s) from the sampled respondents, so that eventually reliable estimates of some of their feature(s) can be estimated for the population as a whole. Since then, survey theoreticians and survey practitioners have contributed significantly in this area of survey methodological research.

Warner (1965) introduced an ingenious device to gather reliable data relating to such issues that may attach unethical stigmas in a civilized society. Therefore, direct questionnaire method is likely to result in refusal/denial or occasionally masked untruthful response. In the context of a society, issues such as abortions, spouse-mishandling, finding HIV tests positive, underreporting income tax returns, false claims for social benefits may have sensitive/unethical stigmas attached. People generally tend to hide public revelations of such vices.

In such circumstances, Warner suggested a way to avoid attempting to collect direct responses (DRs) from the selected respondents—either individually or in groups. Instead, he recommended implementation of what is termed as randomized response technique (RRT) in order to collect information from each sampled respondent when a stigmatizing issue is under contemplation in a study.

There is a huge amount of the published literature in this area of applied research. We refer to an excellent expository early book on RRT by Chaudhuri and Mukerjee (1988). Hedayat and Sinha (1991), Chap. 11, also provides a fairly complete account of RRTs. Two most recent books (Chaudhuri 2011; Chaudhuri and Christofides 2013) are worth mentioning as well.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_2

2.2 Warner's Randomized Response Technique [RRT]

To fix ideas, we consider sampling of individuals from a reference survey population in order to estimate the population proportion of a specific feature such as false claims for social benefits which is likely to be stigmatizing in nature. Therefore, direct questionnaire procedure is likely to be ruled out. In this context, Warner (1965) suggested the following approach.

Note that we are addressing the issue of eliciting truthful information on a sensitive qualitative feature [SQlF], with exactly one of the binary responses [yes/no] attached to each individual in the population, and we are interested in estimation of the population proportion P of 'yes' response(s) based on our study of the sampled respondents. The problem is to provide (i) a method of ascertaining truthful responses from the respondents facing the SQlF in the surveyed population and (ii) (unbiased) estimator of P. Generally, simple random sampling with replacement of respondents from the reference population [presumably large] is contemplated.

With reference to a single SQlF, its possession [yes] will be denoted by the attribute Q and its non-possession [no] will be denoted by the negation of Q, that is, $\bar{Q}$. The simplest related question technique of Warner (1965) refers to preparation of two identical and indistinguishable decks of cards with known multiple but unequal number of copies of both. One set [Set I] will have the instruction on the back of each card: Answer Q truthfully. Naturally, the truthful response should be 'yes' in case the respondent possesses the attribute Q and 'no' otherwise. The other set [Set II] deals with the instruction: Answer $\bar{Q}$ truthfully. This time a response of 'yes' would mean the respondent does not possess Q; otherwise, the response is 'no' implying that the respondent does possess Q. We may denote by p the known proportion of cards of Q category so that $1 - p$ is the proportion of cards of $\bar{Q}$ category. A general instruction is given to all respondents: Each respondent is to select one card at random and with replacement out of the full deck and act as per the instruction given at the back of the selected card. The respondents are supposed to report only the yes/no answers—without divulging what kind of card had been selected by them. Naturally, this randomization device of selection of a card ensures that a respondent can make a choice of Q with probability p or a choice of $\bar{Q}$ with probability $1 - p$, $0 < p \neq 0.5 < 1$, being known beforehand. It is believed that this randomization mechanism will convince the respondent about retaining the confidentiality of the response [yes/no] provided by him/her, without disclosing the choice of the card bearing the label Q or $\bar{Q}$ to the interviewer! In other words, the investigator is not to be told about the specific question chosen/answered by the respondent. For obvious reason, this method is also known as mirrored question design. See Blair et al. (2015) for descriptions of this and a few more RRMs. Routine formulae are there to work out the details of estimation, etc., in this and various other complicated randomization frameworks. In this simple randomized response framework, we proceed as follows toward unbiased estimation of P:

Note that a 'yes' answer has two sources: choice of one card from Set I, followed by 'yes' response, or choice of one card from Set II, followed by 'yes' response.

Therefore, $P[yes] = pP + (1-p)(1-P) = P(2p-1) + (1-p)$. This we equate to the sample proportion of 'yes' responses among the total number of responses.

If there are n respondents and out of them, eventually, some f of them report 'yes,' then we have the defining equation:

$$f/n = P(2p-1) + (1-p)$$

whence

$$\hat{P} = \frac{f/n - (1-p)}{2p-1}.$$

It is seen from the above why we need the condition: $p \neq 0.5$. It follows that

$$(i)\, V(\hat{P}) = P(1-P)/n + p(1-p)/n(2p-1)^2$$

$$(ii)\, \hat{V} = p(1-p)/n(2p-1)^2 + [(1-p)^2 + \hat{P}(2p-1) - f(f-1)/n(n-1)]/n(2p-1)^2.$$

This last expression, when square-rooted, gives what is known as the estimated standard error (s.e.) of $\hat{P}$.

Remark 2.1 The above results are based on the fact that f follows binomial distribution with parameters (n, θ) where n is the sample size [number of respondents] and $\theta = P(2p-1) + (1-p)$, being the probability of 'yes' response by a respondent under the RRM in use. It is known that f/n serves as an unbiased estimate for θ and $f(f-1)/n(n-1)$ serves as an unbiased estimate for θ^2. The rest are simple algebraic manipulations. We will refer to this method as RRM1.

Illustrative Example 2.1 We choose $n = 120$ and $p = 0.40$. Suppose the survey yields $f = 57$. This suggests

$$\hat{P} = [57/120 - 0.60]/[-0.20] = 0.625; \; s.e.$$
$$(\hat{P}) = \sqrt{24/48 + [.36 - .125 - .2235]/48} = 0.0724.$$

Remark 2.2 Use of both versions [affirmative and negative] of the sensitive question Q may, at times, lead to confusion among the respondents. This was soon realized, and the RRT was accordingly modified by introducing what is called unrelated questionnaire method. We will designate this method as RRM2. This is described below.

Once again, we are in the framework of eliciting truthful response on the sensitive question Q but using a modified version of the RRT described above. This time, again, we form two sets of cards, and for the Set I, we keep the same instruction on the back of each card. For Set II, we rephrase the instruction by introducing a simple-minded question like: Were you born in the first quarter of a year? This question is denoted by the symbol Q^* so that it also has two forms of the true reply: 'yes' for the

affirmative reply and 'no' for its negation. When this RRT of eliciting response is executed, the chance of a 'yes' response is given by: $pP + (1-p)/4$. This is because in a random sample of respondents about 1/4th are likely to have been born in the first quarter of a year. As in the above, this is equated to the sample proportion of 'yes' responses, i.e., f/n, and thereby, we obtain $\hat{P} = [f/n - (1-p)/4]/p$. It is a routine task to work out $V(\hat{P})$, and this is given below:

$$\begin{aligned} V(\hat{P}) &= (1-p+4pP)(3+p-4pP)/16np^2 \\ &= [(1-p)(3+p) + 8p(1+p)P - 16p2P^2]/16np^2. \end{aligned}$$

To compute $\hat{V}(\hat{P})$, in the above expression, we have to replace P by $\hat{P}$ which is already shown above. Further, also we need to replace P^2 in the above by an expression to be derived from the defining equation:

$$f(f-1)/n(n-1) = [pP + (1-p)/4]^2$$

upon expansion of the RHS expression and replacement of P by $\hat{P}$ derived earlier. Once estimated variance estimate is obtained, we compute s.e. of the estimate by taking the square root of the above quantity. Note that this time the distribution of f is binomial with parameters $(n, \eta = pP + (1-p)/4)$.

Illustrative Example 2.2 We choose $n = 120$ and $p = 0.40$. Suppose survey yields $f = 57$. This suggests

$\hat{P} = [57/120 - 0.15]/[0.40] = 0.8125.$

Estimating equation for $\hat{P^2}$ is given by

$$0.2235 = p^2P^2 + p(1-p)P/2 + (1-p)^2/16 = 0.16P^2 + 0.0975 + 0.0225;$$

$$0.1035 = 0.16P^2;\ \hat{P^2} = 0.6469.$$

$$\hat{V}(\hat{P}) = [2.04 + 1.56 - 1.6560]/307.2 = 0.0063;\ s.e. = 0.0795.$$

Remark 2.3 In the above and in many such similar contexts, use of stack of cards of different colors can be conveniently replaced by use of spinner wheels marked with different colors in different parts. Thus, for example, red color may occupy 40 percent of the area in the wheel. Naturally, we are referring to the back side of the wheel for coloring purposes. This should be understood, and we will not dwell with this version of the randomization.

2.3 Generalizations of RRMs

We now introduce several generalizations of the above RRMs—these are dictated by real-life applications. Always, the idea is to provide increased and perceived protection to the respondents from the perspective of protecting their confidentiality. In RRM2, we replaced $\bar{Q}$ by a completely simple-minded question which had nothing to do with the stigmatizing question Q. It was at times felt that this might still throw some doubt in the minds of the respondents. It is advisable that we utilize a question in the Set II which is not too far removed from Q which was taken to be false claims for social benefits. What about using 'My family makes 2 or more out-of-state trips on an average every year' whose affirmative version we may denote by Q^* while the negation is denoted by $\bar{Q}^*$? This may not be totally stigmatizing in nature, and the respondents may not feel like either abstaining or giving a wrong answer if a card from Set II is actually selected in the randomization process. However, the true proportion of respondents (in the population as a whole) belonging to the category of Q^* may not be known beforehand. That simply means that this time f will still follow binomial distribution with parameters (n, η) where $\eta = pP + (1-p)P^*$ where P^* stands for the chance of Q^*, the affirmative version of the choice placed in the cards of Set II. Therefore, we may still develop the defining equation $f/n = pP + (1-p)P^*$. Whereas in the cases of RRM1 and RRM2, in this kind of equation, P was the only unknown proportion to be estimated, this time we have two unknowns, viz., P and P^*. Therefore, we need one more equation involving these two unknown parameters. This calls for the following RRM3.

We divide the whole collection of respondents into two equal/almost equal groups, say of sizes n1 and n2. For Group I, we collect information by using a version of RRM2, viz., by replacing the question related to birth by the question related to Q^* on family trips. This results in the pair $(f1, n1)$ upon implementation. For notational simplicity and for ease of making generalizations, we use $p1$ for p. Therefore, $f1$ is distributed as binomial $(n1, \eta1$ where $\eta1 = p1P + (1-p1)P^*$. Likewise, for Group II, based on the data of the form $(f2, n2)$, from the cards drawn from Set II, it turns out that $f2$ is binomial with parameters $(n2, \eta2)$ where $\eta2 = p2P + (1-p2)P^*$. Note that $\eta1$ and $\eta2$ are, respectively, the proportions of cards in the two Sets I and II bearing the affirmative versions of Q and Q^*, respectively.

We have generated two equations, viz.,

$$f1/n1 = \eta1 = p1P + (1-p1)P^* : f2/n2 = \eta2 = p2P + (1-p2)P^*.$$

From the above, we may easily solve the primary parameter P [as well as the other parameter P^*].

The solutions are linear functions of the sample proportions $f1/n1$ and $f2/n2$. Therefore, we can work out variance estimates and estimated variances in a routine manner. It must be noted that the solutions exist only when our choice is such that $p1 \neq p2$.

Illustrative Example 2.3 We choose $n1 = n2 = 120$ and $p1 = 0.40$ and $p2 = 0.60$. Suppose survey yields $f1 = 57$ and $f2 = 75$. This leads to the equations:

$$57/120 = 0.40P + 0.60P^*; 75/120 = 0.60P + 0.40P^*.$$

Therefore, the estimates for P and P^* are 0.925 and 0.175, respectively. Before proceeding further with other approaches/methods, we will digress for a moment to discuss a source of non-response and its follow-up studies.

2.4 Not-at-Homes: Source of Non-response

While extracting information through a direct response survey on some features [qualitative or quantitative] from the respondents in a survey population, it is generally believed that there would be cooperation from the respondents—at least when the features are non-sensitive in nature. Of course, for sensitive features, we need to develop RRTs. However, there are instances where we encounter non-response for various reasons even when the features are non-evasive in nature. One of such sources is attributed to 'Not-at-homes.' Survey sampling researchers attempted to study this phenomenon. Notable contributors are: Yates (1946), Hansen and Hurwitz (1946), Hartley (1946), Politz and Simmons (1949), and Deming (1953). Their studies were essentially geared toward regular features of the survey questions. The technique for extraction of 'response' is known as Hartley–Politz–Simmons technique.

Much later, Rao (2014) considered the case of handling situations, wherein it is unlikely for a respondent to reveal truthful answer(s) even when it is non-sensitive in nature. It was followed up by yet another follow-up paper by Rao et al. (2016). We will not elaborate on this issue further.

2.5 RRMs—Further Generalizations

Following Blair et al. (2015), we will now briefly discuss two more generalizations of the basic RRM.

(i) Forced Response Designs [FRD]: This RRM incorporates a forced response of yes as well as a forced response of no. The idea is to label forced yes (no) to the outcome 1(6), while for any other outcome of the throw of a regular [unbiased] six-faced die the respondent is supposed to give truthful response in terms of yes/no for possession of the sensitive stigmatizing feature. Thus eventually, we have only yes or no response from each respondent.

(ii) Disguised Response Design [DRD]: The yes response to the sensitive feature is meant to be identified as the YES Stack of black and red cards. Likewise, the no response to the sensitive feature is to be identified as the NO Stack of black and

red cards. Total number of cards is the same for both types of stacks. Further, if we have 80 percent red cards in YES Stack, then we must have 20 percent red cards in the NO Stack. This may be arbitrary but must be predetermined and be the same for all respondents. Every respondent is supposed to truthfully implement his choice of the correct stack by referring to the sensitive stigmatizing feature under study. Once this is done, he/she is supposed to draw a card at random from the correctly selected stack and only disclose the color of the card drawn—without any mention of the stack identified. Whether the respondent belongs to yes/no category [in respect of the feature under study] is his/her truthful confession to himself/herself.

Illustrative Example 2.4 Here, we discuss about FRD. We take $n = 300$, and suppose after implementation of the FRD, we obtain: yes count $= 180$ and no count $= 120$. Let P be the true proportion of persons possessing the sensitive feature in the population. Then, the chance of yes response from a respondent is given by $1/6 + 4P/6$ and we equate this to the sample proportion $= 180/300$. This yields $\hat{P} = 0.65$. Further, it can be shown that

$$\hat{V}(\hat{P}) = 9[f(n-f)/n^2(n-1)]/4 = 0.0018,$$

upon simplification. Hence, s.e. of the estimate $= 0.0424$.

Illustrative Example 2.5 We take up DRD now. We start with $n = 300$ respondents, and suppose, upon implementation of the DRD, we obtain: red count $= 180$ and black count $= 120$. Let P be the true proportion of persons possessing the sensitive feature in the population. Then, the chance of red card being drawn is given by $0.8P + 0.2(1 - P) = 0.2 + 0.6P$. We equate this to the sample proportion $= 180/300$. This yields $\hat{P} = 0.6667$. Further,

$$\hat{V}(\hat{P}) = 25[f(n-f)/n^2(n-1)]/9 = 0.0021,$$

upon simplification. Hence, s.e. of the estimate 0.0458.

2.6 RRMs for Two Independent Stigmatizing Features

In case there are two or more sensitive qualitative features of a population to be studied, one can always study them separately. However, a joint study makes more sense since less effort will be spent to capture incidence. The RRM2 discussed above can be conveniently generalized to cover this situation. In the deck of cards, we accommodate cards of three different colors: black, red, and yellow. Black [red/yellow] cards read: Answer Q1 [Q2/Q3] truthfully where Q1 refers to SQlF1: underreporting income tax returns; Q2 refers to SQlF2: false claims for social benefits; and Q3 refers to a simple-minded innocent statement like on an average my family makes 2 or

more out-of-state trips per year. All the three questions seek truthful binary response: yes/no. We presume that apart from the unknown proportions $P1$, $P2$ referring to chances of underreporting of IT returns and making false claims for social benefits, respectively, the other parameter $P3$ referring to asserting the statement about family trips is also unknown [and, may be incidentally estimated]. Thus, we are in the framework of three unknown parameters, and hence, we need three different [technically called linearly independent] estimating equations. We proceed by dividing the total number of respondents into three equal/almost equal groups. Also, we need three sets of cards with three different proportions of color compositions.

Illustrative Example 2.6 We start with $n = 301, n1 = n2 = 100, n3 = 101$. Further, color distribution of the cards in the three sets is taken as

Set I : B : R : Y : : 25, 30, and 45 percents;
Set II : B : R : Y : : 30, 45, and 25 percents;
Set III : B : R : Y : : 45, 25, and 30 percents.

Each respondent from Group I will pick up a card at random from Set I and will only communicate the truthful answer: yes/no—without divulging the color of the card drawn. Likewise, for respondents from the other two groups, same conditions apply. Suppose the proportions of yes answers are: 55/100, 43/100, and 51/101. Then, the defining equations are:

$$0.55 = 0.25P1 + 0.30P2 + 0.45P3;\ 0.43 = 0.30P1 + 0.45P2 + 0.25P3;$$

$$0.52 = 0.45P1 + 0.25P2 + 0.0.30P3.$$

From the above, we derive the estimates as

$$\hat{P}1 = 0.5155;\ \hat{P}2 = 0.1454;\ \hat{P}3 = 0.8385.$$

Remark 2.4 In the above example, it is tacitly assumed that the two sensitive features are independently distributed over the reference population. Otherwise, the two should be jointly studied in terms of 2×2 classification: [(Yes, Yes), (Yes, No), (No, Yes), (No, No)]. This and much more are discussed in the published literature. See, for example, Hedayat and Sinha (1991).

2.7 Toward Perception of Increased Protection of Confidentiality

Since the introduction of RRT, survey sampling practitioners/theoreticians have paid due attention to this area of survey methodological research. As has been mentioned, the purpose is to be able to elicit a truthful response on sensitive feature(s) from the sampled respondents, so that eventually the population proportion of incidence of

the sensitive feature can be unbiasedly estimated. Toward this, a novel technique was introduced by Raghavarao and Federer (1979) and it was termed block total response [BTR] technique. A precursor to this study was undertaken by Smith et al. (1974). We propose to discuss the basic BTR technique with an illustrative example.

As usual, we start with one SQlF, say Q [with an unknown incidence proportion P to be estimated from the survey] and along with it we also consider a collection of 8 RQlFs $[Q1, Q2, \ldots, Q8]$ which are simple-minded and yet binary response queries. We thus have a total collection of nine QlFs, including the SQlF. The steps to be followed are:

(i) We prepare several blocks of questions, i.e., a questionnaire involving, say some 4 of the RQlFs and the SQlF in each block. The only condition to be satisfied in the formation of the blocks is that each RQlF must appear the same number of times in the entire collection of blocks. Additionally, we also prepare a Master Block Bl^* : $[Q1, Q2, \ldots, Q8]$ comprising of all the RQlFs.
For example, we may choose

$$Bl\ 1 : [Q1, Q2, Q4, Q6; Q]; Bl\ 2 : [Q1, Q3, Q6, Q7; Q],$$

$$Bl\ 3 : [Q2, Q3, Q5, Q8; Q]; Bl\ 4 : [Q4, Q5, Q7, Q8; Q].$$

(ii) Since we have a total of five blocks, we need five groups of respondents. The first four groups for dealing with blocks $Bl\ 1 - Bl\ 4$ are assumed to have the same size, say 50 each. In addition, we will go for some 30 respondents, for example, for the block Bl^*. So, we are dealing with a collection of say 230 respondents—randomly divided into these five groups.

(iii) Each member of the first group of respondents is exposed to the questions contained in $Bl\ 1$, and he/she is told to respond truthfully to each of the RQlFs as also to the Q. However, he/she is supposed to report/divulge only the block total response [BTR]—the total score in terms of yes responses. This is continued for all other blocks as also for the Master Block Bl^*.

(iv) The above completes the survey aspect of the BTR technique. Suppose we end up with the following summary data in terms of average score in each block per respondent:

$$Bl\ 1 : 0.285; Bl\ 2 : 0.354; Bl\ 3 : 0.328; Bl\ 4 : 0.396; Bl^*\ : 0.395.$$

(v) An estimate of the incidence proportion P of the SQlF is given by the computational formula:

(a) Sum of average scores in the first four blocks $= 1.363$, and this is equated to $[2 \sum Pi + 4P]/5$.
(b) The average score of 0.395 in the last block is equated to $\sum Pi/8$.
(c) From the above, $\hat{P} = [5 \times 1.363 - 2 \times 8 \times 0.395]/4 = 0.324$.

Remark 2.5 The above illustration arises out of a very general approach toward BTR technique. Naturally, once a respondent is told to provide only the BTR without divulging any kind of details as to the nature of individual responses, the investigator may be assured of increased cooperation from the respondents. This basic BTR technique has been extended further with an aim to provide enhanced protection of privacy to the respondents. The details may be found in Nandy et al. (2016) and Sinha (2017).

2.8 Confidentiality Protection in the Study of Quantitative Features

Consider a situation wherein we are dealing with a finite [labeled] population of size N and there is a sensitive qualitative study variable Y for which the 'true' values are $Y_1, Y_2, \ldots, Y_N$ for the units in their respective orders. To start with, these values are unknown and we want to unbiasedly estimate the finite population mean $\bar{Y} = \sum_i Y_i/N$.

We may adopt $SRSWOR(N, n)$ or any other *suitably defined* fixed size (n) sampling design and draw a random sample of n respondents. Had the study variable been non-sensitive in nature, we could take recourse to 'direct questioning' involving the sampled respondents. In a very general setup, we may make use of the Horvitz–Thompson estimate [HTE, for short]. It simplifies $\bar{y}$ when $SRSWOR(N, n)$ is adopted. However, we are dealing with a sensitive characteristic [such as 'income accrued through illegal profession'] and we need to use a suitably defined RRT. Here, we propose an RRT for this purpose.

Assume that the true Y-values are completely covered by a pool of K known quantities like $M_1, M_2, \ldots, M_K$. The set of M-values may even comprise a larger set. Therefore, in effect, we are assuming that the N population values are discrete in nature.

We choose a small fraction δ and proceed to deploy RRT as is explained in the following example with $K = 10$ and $\delta = 0.2$.

We prepare 25 identical cards, and at the back of the cards we give instructions: For each of five cards, it reads at the back: 'Report your true income accrued through illegal profession.' For the rest of the 20 cards, we use them in pairs, and for the ith pair, it reads at the back of both the cards: 'Report M_i'; $i = 1, 2, \ldots, 10$.

Each respondent chooses a card at random out of the 25 cards, reads out the back side, and acts accordingly. We assume that the respondents act honestly and provide 'truthful' figure—no matter which card is chosen—without disclosing in any way the nature of the card.

Note that the chance of choosing a card with marking as 'Report your true income...' is $5/25 = 0.20$ which coincides with the chosen value of δ. On the other hand, chance of picking up a card corresponding to any specified value M_i is $2/25 = 0.08$ which is equal to $(1 - \delta)/K$.

Using the notations δ and K, for the chosen sample of n respondents, we have thus collected the responses to be denoted by $R_1, R_2, \ldots, R_n$. Each response is random in nature and

$$E(R_i) = \delta Y_i + (1 - \delta)\bar{M} \tag{2.8.1}$$

where $\bar{M} = \sum_i M_i/K$ and Y_i is the true [unknown] response of the ith sampled respondent. From this, it follows that Y_i can be unbiasedly estimated as

$$\hat{Y}_i = [R_i - (1 - \delta)\bar{M}]/\delta. \tag{2.8.2}$$

Hence, an unbiased estimate for the finite population mean, based on estimates of Y_i's, is obtained by referring to HTE in general and to the sample mean of estimated Y's in case SRSWOR has been implemented during sample selection. The proof of this claim rests on the formula: $E = E_1E_2$. Therefore,

$$\hat{\bar{Y}} = \sum_i \hat{Y}_i/n. \tag{2.8.3}$$

In the above, for $K = 10$ and $\delta = 0.2$, $\hat{Y}_i = [5R_i - 4\bar{M}]$ and hence $\hat{\bar{Y}} = 5\bar{R} - 4\bar{M}$ is the estimate of population mean, under SRSWOR sampling. Here, $\bar{R}$ refers to the sample mean of the sampled R's and $\bar{M}$ refers to the mean of the given M's.

Remark 2.6 It may be noted that in the above it is tacitly assumed that each Y_i matches with one of the given values M_i's. However, no sampled respondent is supposed to divulge which M-value matched his/her true value of Y.

Below, we proceed to work out a formula for the estimated standard error [s.e.] of the estimate of the population mean based on the above procedure. In addition to $E(R_i)$ displayed in (2.8.1), we have

$$E(R_i^2) = \delta Y_i^2 + (1 - \delta)\bar{Q}_M \tag{2.8.4}$$

where $\bar{Q}_M = \sum_i M_i^2/K$ is the mean of squares of the M-values.

These suggest

$$V(R_i) = \delta(1 - \delta)Y_i^2 + (1 - \delta)[\bar{Q}_M - (1 - \delta)\bar{M}^2] - 2\delta(1 - \delta)Y_i\bar{M}. \tag{2.8.5}$$

From (2.8.4), it follows that

$$\hat{Y}_i^2 = [R_i^2 - (1 - \delta)\bar{Q}_M]/\delta \tag{2.8.6}$$

Using (2.8.6), we may deduce that

$$\hat{V}(R_i) = (1-\delta)[R_i^2 - (1-\delta)\bar{Q}_M] + (1-\delta)[\bar{Q}_M - (1-\delta)\bar{M}^2] - 2(1-\delta)\bar{M}[R_i - (1-\delta)\bar{M}] \quad (2.8.7)$$

which simplifies to

$$(1-\delta)[R_i - \bar{M}]^2 + \delta(1-\delta)[\bar{Q}_M - \bar{M}^2] \quad (2.8.8)$$

We now proceed to work out estimated standard error for the estimate of the finite population mean under $SRSWOR(N, n)$. Clearly, under $SRSWOR(N, n)$,

$$\hat{\bar{Y}} = \sum_i \hat{Y}_i/n;\ \hat{Y}_i = [R_i - \bar{M}(1-\delta)]/\delta \quad (2.8.9)$$

Moreover,

$$V[\hat{\bar{Y}}] = V_1E_2 + E_1V_2 \quad (2.8.10)$$

Note that

$$V_1E_2 = V_1\left(\sum_i Y_i/n\right) = (1/n - 1/N)S_Y^2 \quad (2.8.11)$$

where S_Y^2 refers to the population variance of Y-values with divisor $N-1$. To estimate this, we usually employ the sample counterpart of S_Y^2, viz., $s_Y^2 = sum_i(Y_i - \bar{Y})^2/(n-1)$. Here, of course, Y_i's are unknown and are being estimated in terms of the R's by an application of RRT. The expression for s_Y^2 involves square terms, i.e., Y_i^2's and cross-product terms, i.e., Y_iY_j's. From (2.8.6), we deduce expressions for $\hat{Y}_i^2$'s. Since the respondents act independently, estimates of the product terms Y_iY_j's are also derived by the product of terms of the form (2.8.4). This takes care of estimate for V_1E_2 term.

Next, note that for every sampled respondent such as the ith, V_2 refers to variance of $\hat{Y}_i$. From (2.8.4), it follows that $V(\hat{Y}_i) = V(R_i)/\delta^2$. From (2.8.7), we readily have an expression for estimate of $V(R_i)$. Therefore,

$$\hat{E}_1V_2 = \sum_i \hat{V}(\hat{Y}_i)/n^2 = \sum_i \hat{V}(R_i)/n^2\delta^2 \quad (2.8.12)$$

Illustrative Example 2.7 As before, we take $K = 10$ and $\delta = 0.2$. Let our choice of M's be [expressed in units of thousand rupees]: $1, 2, \ldots, 10$. We consider a small population and adopt $SRSWOR(N = 20, n = 5)$. Let the sampled R's [as per the respondents' reporting] be: $3, 7, 4, 8, 5$, that is our data. We show the necessary computations below.

Example 2.7 : Computational details

$R-values$	$Y-estimates$	$Y^2-estimates$
3	−7	−109
4	−2	−74
5	3	−29
7	13	91
8	18	166
Total	25	45

From (2.8.4), we obtain an estimate of the population mean of Y-values as the sample mean computed as Rs. $25/5 = 5$ thousand. To compute estimated s.e. of the estimate, we proceed as follows:

From the discussion below (2.8.11), it follows that an estimate of V_1E_2 is given by $(1/n - 1/N)$ times an unbiased estimate of s_Y^2 based on the computations in Example 2.7 above. Since $s_Y^2 = [(n-1)\sum_i Y_i^2 - \sum\sum_{inej} Y_iY_j]/n(n-1)]$, we do term by term estimation by using relevant square terms and product terms from Example 2.7. This yields:

$\hat{s_Y^2} = [4 \times 45 - 80]/20 = 5$ and hence, an unbiased estimate of V_1E_2 is given by $(1/5 - 1/20) \times 5 = 3/4 = 0.75$.

For the other term, viz., E_1V_2, an unbiased estimate is to be computed from (2.8.12) in combination with (2.8.7). For the computations, note that $\bar{Q}_M = 38.5$. In (2.8.7),

Term 1[with positive sign]:

$$\hat{V}(R_i) = (1-\delta)[R_i^2 - (1-\delta)\bar{Q}_M] = 0.8[R_i^2 - 30.8]$$

Term 2 [with positive sign]:

$$(1-\delta)[\bar{Q}_M - (1-\delta)\bar{M}^2] = 14.3$$

Term 3 [with negative sign]

$$2(1-\delta)\bar{M}[R_i - (1-\delta)\bar{M}] = 8.8[R_i - 4.4]$$

Example 2.7 : Computational details

$R-values$	$Term\ 1 = 0.8[R_i^2 - 30.8]$	$Term\ 2 = 14.3$	$Term\ 3 = 8.8[R_i - 4.4]$
3	−17.44	14.3	−12.32
4	−11.84	14.3	−3.52
5	−4.64	14.3	5.28
7	14.56	14.3	22.88
8	26.56	14.3	31.68
Total	7.20	71.5	44.00

Therefore, unbiased estimate of E_1V_2 is computed as $[7.20 + 71.50 - 44.00]/16 = 34.70$. Finally, adding the two components, an unbiased estimate of the variance $= 0.75 + 34.70 = 35.45$ so that estimated $s.e. = \sqrt{(}35.45) = 5.95$.

Remark 2.7 In a similar study, Bose (2015) took up the case of $SRSWR(N, n)$ and derived expression for the estimate of the population mean and an expression for its variance. The above study is quite general in nature and applies to any fixed size (n) sampling design.

Remark 2.8 The BTR technique discussed in the context of sensitive qualitative feature can be extended to the case of sensitive quantitative feature—without the assumption of 'closure' w.r.t. a given set of known quantities such as $[M_1, M_2, \ldots, M_K]$. This has been taken up recently in Nandy and Sinha (2018). We omit the details.

2.9 Concluding Remarks

The topic of RRT is vast and varied in terms of the published literature in the form of papers, books, and reports. We have simply introduced the basic ideas and initial methodologies that were suggested in the context of estimation of a population proportion of a sensitive feature of the members of a population. We have also presented one method w.r.t. quantitative feature. There are similar methodologies dealing with (i) more than one sensitive qualitative features, (ii) one or more sensitive quantitative features, and so on.

References and Suggested Readings

Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, *110*, 1304–1319.

Bose, M. (2015). Respondent privacy and estimation efficiency in randomized response surveys for discrete-valued sensitive variables. *Statistical Papers*, *56*(4), 1055–1069.

Chaudhuri, A. (2011). *Randomized response and indirect questioning techniques in surveys*. FL, USA: CRC Press, Chapman and Hall, Taylor & Francis Group.

Chaudhuri, A., & Mukerjee, R. (1988). *Randomized response: Theory and applications*. New York: Marcel and Dekker.

Chaudhuri, A., & Christofides, T. C. (2013). *Indirect questioning in sample surveys*. Germany: Springer.

Deming, W. E. (1953). On a probability mechanism to attain an economic balance between the resultant error of nonresponse and the bias of nonresponse. *Journal of American Statistical Association, 48*, 743–772.

Hansen, M. H., & Hurwitz, W. N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, *41*, 517–529.

Hartley, H. O. (1946). Discussion on paper by F. Yates. *Journal of the Royal Statistical Society*, *109*, 37–38.

Hedayat, A. S., & Sinha, B. K. (1991). *Design and inference in finite population sampling*. New York: Wiley.

Nandy, K., Markovitz, M., & Sinha Bikas K. (2016). In A. Chaudhuri, T. C. Christofides, & C. R. Rao (Eds.), Handbook of statistics (Vol. 54, pp. 317–330).

Nandy, K. & Sinha, B. K. (2018). Randomized response technique for quantitative sensitive features in a finite population.

Politz, A. N., & Simmons, W. R. (1949, 1950). An attempt to get not at homes into the sample without call-backs. *Journal of the American Statistical Association*, *44*, 9–16; *45*, 136–137.

Raghavarao, D., & Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society, Series B, Methodological*, *41*, 4045.

Rao, T. J. (2014). A revisit to Godambes theorem and Warners randomized response technique. In *Proceedings of the International Conference on Statistics and Information Technology for a Growing Nation* (p. 2).

Rao, T. J., Sarkar, J., & Sinha, B. K. (2016). Randomized response and new thoughts on Politz–Simmons technique. In A. Chaudhuri, T. C. Christofides, & C. R. Rao (Eds.), Handbook of statistics (Vol. 54, pp. 233–251).

Sinha, B. K. (2017). Some refinements of block total response technique in the context of RRT methodology. In H. Chandra, & S. Pal (Eds.), Randomized Response Techniques and Their Applications. Special Issue of *SSCA Journal*, *15*, 167–171.

Smith, L., Federer, W., & Raghavarao, D. (1974). A comparison of three techniques for eliciting truthful answers to sensitive questions. In *Proceedings of the Social Statistics Association* (pp. 447–452). Baltimore: American Statistical Association.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, *109*, 12–43.

Chapter 3
Content Analysis

3.1 Introduction

Some authors in the area of social science research try to distinguish between qualitative research and quantitative research, in terms of the research hypotheses, the nature of evidences generated or compiled, the method of processing evidences, and the way inferences are arrived at. It is felt such a distinction is hardly tenable in the context of present-day researches. Quantification in some way or the other has become a part of any study in any domain today. Any comparison involves quantification in most cases, especially in dealing with more or less similar entities. Apparently, qualitative aspects of certain entities, like style of writing a document or preparing a message for wide circulation that may incorporate some images, focus on customer satisfaction in drawing a quality policy or bias in reporting some incident with a significant fallout on social harmony, etc., can be and are being quantified by identifying some countable or measurable features. Subsequently, these are subjected to quantitative analysis and inferences are drawn on the basis of such analyses. Of course, results of any quantitative analysis or any inference reached on that basis should be interpreted in the context of the research problem and mostly in a qualitative manner.

In some studies of great importance in social, political, economic or even archaeological interest, written texts, images drawn, actions observed in videotaped studies, artifacts preserved in places of historic interest, newspaper reports and similar data convey some 'messages' pertaining to some context, and we need to analyze these messages to make valid and replicable inferences about the context.

The context could be an issue of some interest and may be associated with some hypothesis. Content Analysis refers to procedures for assessing the relative extent to which specified references, attitudes, or themes permeate given messages or documents. Content Analysis conforms to three basic principles of a scientific method. It is

Objective—in terms of explicit rules being followed to enable different researchers to obtain the same results from the same documents or messages.

Systematic—through inclusion of only materials which support the researcher's ideas by applying some consistent rules.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_3

Generalizable—in terms of being applicable to situations wherever data of this type only are available.

The researcher interprets the content to reveal something about the characteristics of the message, particularly its bearing on the research problem and the hypothesi(e)s.

Content Analysis enables researchers to swift through large volumes of data with relative ease in a systematic fashion. It can be a useful technique for allowing us to discover and describe the focus of individual, group, institutional, or social attention. It also allows inferences to be made which can be corroborated using other methods of data collection. In the language of Krippendorff (1980), content analysis research is motivated by the search for technique to infer from symbolic data what would be either too costly, no longer possible or too obtrusive by the use of other techniques.

3.2 Uses of Content Analysis

Content Analysis can be a powerful tool for determining authorship of a document. Styles of different authors vary in respect of paragraph length (number of sentences or of printed lines in a paragraph, sentence length (number of words in a sentence), word length (number of letters in a word), use of parenthetic clauses, repetitions of the same words within a paragraph or even within the same sentence, use of different forms of narration, etc. The frequency of nouns or function words or some particular words used in the document can be counted and compared with such frequencies in similar writings by some known authors to help build a case for the probability of each person's authorship of the document, and to apply some criterion like maximum likelihood to make a valid inference.

Content Analysis is also useful to examine trends in documents of a particular type like newspapers, covering, say, news about social evils, or crimes or about misgovernance, over a period of time. The relative emphasis on correctly (1) reporting the events as they happened, (2) analyzing the causes that led to the events, and (3) pointing out the likely fallouts on the society or the economy, can be examined to detect such trends in journalism.

Content Analysis can also provide an empirical basis for monitoring shifts in public opinion on issues of social and cultural importance. Thus, mission statements of different institutions or policy pronouncements of government departments may be compared with the programmes and projects undertaken.

3.3 Steps in Content Analysis

Content Analysis studies usually involve

* formulation of the research questions;
* selection of communication content;
* developing content categories;

* finalizing units of analysis;
* preparing a coding schedule, pilot- testing and checking inter-rater reliability;
* analyzing the coded data.

Content analysis relies heavily on coding and categorizing the symbolic data, ensuring adequate reliability and validity. Such categorical data can subsequently be analyzed by appropriate statistical tools for making inferences about the context. Categories must be mutually exclusive and exhaustive.

Two approaches for coding, viz. emergent coding where categories are established following some preliminary examination. Two persons can independently review the material in the common context and come up with a list of features, which they can compare and reconcile. Third, they use a consolidated list to be used for coding. Finally, reliability is examined in terms of the Kappa coefficient, the acceptable value being 0.8. In a priori coding, categories are established prior to the analysis based on some theory. Professional colleagues agree on the categories. Revisions are made as and when necessary and categories are tightened up to the point that maximizes mutual exclusivity and exhaustivity.

Categorization has to be relevant to the problem or hypothesis under consideration. One of the critical steps in Content Analysis involves developing a set of explicit recording instructions to ensure adequate reliability for the coded data.

3.4 Reliability of Coded Data

Reliability can imply stability or intra-rater reliability for the same rater to get the same results in repeated coding exercises, and reproducibility or inter-rater reliability to ensure that the same feature is coded in the same category by different raters.

Reliability may be calculated in terms of agreement between raters using Cohen's Kappa, lying between 1 to imply perfect reliability and 0 indicating agreement only through chance, and given by

$$\kappa = (PA - PC)/(1 - PC)$$

where PA = proportion of units on which raters agree and PC = proportion of agreement by chance.

Consider the following data on agreement between 2 raters in categorizing a number of coding units.

Example 1: Rater 1 versus Rater 2

*Rater*2*Rater*1	*Academic*	*Emotional*	*Physical*	*Total*
Academic	0.42(0.29)	0.10(0.21)	0.05(0.07)	0.57
Emotional	0.07(0.18)	0.25(0.13)	0.03(0.05)	0.35
Physical	0.01(0.04)	0.02(0.03)	0.05(0.01)	0.08
Marginal	0.50	0.37	0.13	1.00

Here $PA = 0.72$ and $PC = 0.43$ so that $\kappa = 0.51$.

To interpret this value of κ, we refer to Table due to Landis and Koch (1977) based on personal opinion and not on evidence. These guidelines may be more harmful than helpful, as the number of categories and of subjects will affect the value of κ. The value will be higher with fewer categories.

Interpretation of κ

Range of κ	*Strength of Agreement*
< 0	*poor*
$.00 - -.20$	*slight*
$.21 - -.40$	*fair*
$.41 - -.60$	*moderate*
$.61 - -.80$	*substantial*
$.81 - -1.00$	*almost perfect*

Therefore, we can possibly infer a moderate agreement between the two raters in this example. To judge inter-coder reliability, Chadwick (1984) suggested a coefficient defined as *PA* simply.

Holsti (1963) suggested another measure as

$$R = 2PA/(1 + PA).$$

Both these measures lying between 0 and 1 can be interpreted in a similar manner. In the preceding example, these two measures have values 0.72 and 0.84, respectively. The fact that these are higher than the value of κ is due to the absence of any correction for agreement by chance alone.

Assumptions in the Use of κ

(i) Units of analysis are independent.
(ii) Categories of the nominal scale are independent, mutually exclusive, and exhaustive.
(iii) Raters are coding independently.

In Scott's π to measure inter-rater reliability, expected agreement is calculated differently, using joint proportions. It makes the assumption that raters have the same distribution of responses across categories, which makes Cohen's Kappa slightly more informative. Expected agreement in the example will be found as follows. To calculate expected agreement in Scott's *pi*. We sum marginal totals across raters, divide by the total number of ratings to obtain joint proportions, then square and total these. This gives us $PC = .5352^2 + .3602^2 + .1052^2 = 0.43$ as noted earlier $PA = 0.72$. Therefore, κ comes out to be $(0.72 - 0.43)/(1 - 0.43) = 0.29/0.57 = 0.51$.

Not necessarily expected, Cohen's Kappa had exactly the same value in this example. Cohen's κ and Scott's π apply to the case of two raters or judges only.

Fleiss's κ proceeds on the same lines as Cohen's in the case of more than two raters and is defined as follows:

$$\kappa = (Pc - Pe)/(1 - Pe).$$

The denominator gives the degree of agreement that is attainable above chance, and the numerator gives the degree actually achieved above chance. While $\kappa = 1$ indicates complete agreement among raters, $\kappa = 0$ if there is no agreement. Fleiss's Kappa can be used only with binary or nominal scale ratings, No version is available for ordered-categorical ratings.

It is to be noted that Cohen's Kappa assumes that each of two raters rates all the items (subjects). Fleiss's Kappa allows different items to be rated by different individuals like Item 1 is rated by raters A, B, and C while item 2 may be rated by raters D, E, and F.

Let n raters put N items or subjects in k different categories. Let n_{ij} = number of raters who assigned the i-th item/ subject to the j-th category. Then $n = \sum_{j=1}^{k} n_{ij}$. Let $p_j = 1/[Nn] \sum_{i=1}^{N} n_{ij}$. Also let P_i indicate the extent to which raters agree for the ith item / subject (i.e., how many rater pairs are in agreement, relative to the number of all possible rater–rater pairs). Then $P = \sum_i P_i/N$ and $Pc = \sum_j p_j^2$. In fact, Pc is akin to Simpson's Index in analysis of diversity and gives the probability that any two items or subjects randomly chosen will belong to the same category. In other words, it measures the agreement between rater pairs.

$$P_j = \sum n_{ij}(n_{ij} - 1)/[n(n-1)] = \left[\sum n_{ij}^2 - n\right]/[n(n-1)].$$

Thus, the Kappa coefficient comes out as $\kappa = [P - Pc]/[1 - Pc]$.

Consider, as an example, a situation wherein 14 psychiatrists asked to look at ten patients, each gives one of possibly five diagnoses to each patient. The data will appear as a matrix

Patient Diagnoses by 14 psychiatrists

Patient	1	2	3	4	5	*Pi*
1	0	0	0	0	14	1.000
2	0	2	6	4	2	0.253
3	0	0	3	5	6	0.308
4	0	3	9	2	0	0.440
5	2	2	8	1	1	0.330
6	7	7	0	0	0	0.462
7	3	2	6	3	0	0.242
8	2	5	3	2	2	0.176
9	6	5	2	1	0	0.286
10	0	2	2	3	7	0.286
Total	20	28	39	21	32	–
pj	0.143	0.200	0.279	0.150	0.229	–

In this example, P comes out as $3.780/10 = 0.378$ and Pc has the value 0.213. Hence, the value of κ is found to be 0.210. This indicates poor agreement among the psychiatrists regarding their diagnostic classification of the patients, as can be seen from the fact except for patients numbered 1, 6, and 9 where the raters perfectly or mostly agree among themselves, patients 2, 3, 5, or 10 are assigned to different categories by different raters.

Validity of Inferences from the Data

(i) Units of analysis are independent.
(ii) Categories of the nominal scale are independent, mutually exclusive, and exhaustive.
(iii) Raters are coding independently.

* use of multiple sources of information;
* involvement of different investigators;
* use of alternative methods of analysis;
* recourse to different theories.

This approach to validation is referred to as triangulation. Search for appropriate alternatives remains a problem. Triangulation lends credibility to the findings by incorporating multiple sources of data, multiple methods, multiple investigators, or alternative theories. To cross-validate the findings of a content analysis of quality policies of different organizations, we can think of directly asking employees of the concern to respond to certain questions which truly reflect the intentions and directions about quality as are spelt out in the quality policy.

Nature of Research Question

Which election issues figured prominently during 1991 elections in the editorials and letters to the editors of selected dailies and how these dailies differed in terms of the frequency of appearance and the direction of treatment (favorable, unfavorable, and neutral) of these issues?

Let us consider the problem to compare private and public manufacturing organizations with similar product profiles in regard to the relative emphasis, they place on customer satisfaction. We can think of a hypothesis stating that a greater emphasis is placed in the public sector than in the private sector.

One could have possibly hypothesized in the opposite direction if we were dealing with service organizations. We could also look at variations across sectors of manufacturing in terms of the output profiles, rather than across ownership patterns and construct our hypotheses accordingly.

We can consider variations in focus on quality in different stages of production or on concerns for own people or for corporate social responsibility (CSR) obligations. Data for any such purpose may arise in different ways and may have different levels of cost and credibility implications.

A simple unobtrusive way to generate data here would be to just analyze the quality policy or business policy documents of the selected organizations and not

to use any questionnaires or interviews or direct observations on practices. It must be kept in mind, however, that direct observations as are involved in an audit would have provided much more reliable evidence about emphasis on customer satisfaction, since there could be gaps between what appear in the policy documents and what are followed in actual practice.

We need to have representative samples of manufacturing houses for each of the groups in which we are interested. The policy documents are the sampling units and the words or phrases or sentences which relate directly or indirectly to 'customer satisfaction' or the issue of interest constitute the coding units. We can develop certain categories into which words or phrases in the policy document bearing on, say, customer satisfaction directly or implicitly can be placed in a mutually exclusive and exhaustive manner by a reader/ rater, e.g., strong, moderate, not explicitly stated.

We have to be careful in comprehending the implication of a statement like 'we will regard every complaint as a failure,' which possibly implies a strong emphasis on customer satisfaction. The important point to note is the fact that Content Analysis in this context is not just in terms of frequency counts, e.g., 'how many times the phrase customer satisfaction appears in the policy statement' to build up a distribution of this number across units, separately for each of the groups we like to distinguish. Thereafter, such distributions could be compared, either in terms of summary measures or using the principle of dominance or applying suitable tests of homogeneity. In this approach, data would be numerical and we need not involve more than one rater or reader. However, if we look into the implications of various statements contained in the quality policy bearing on customer satisfaction and then require to put any unit into one of several categories-decided upon in some way-different raters or judges would possibly come up with different frequencies (of units) in each of the categories. In such cases, we should first examine the extent of agreement among the raters and, once a reasonable extent of agreement has been found to exist, we can proceed with the average number in each of the categories for further analysis. Thus, we could eventually land up with two such distributions, one for private and the other for public enterprises and can apply the homogeneity test using the Brandt–Snedecor formula for chi-square.

Suppose we want to judge variations in quality management systems using prefixed categories like compliance-oriented, improvement-oriented, and excellence-oriented, again from the policy documents. The entire policy documents become the coding units and the rater has to carefully examine the document as a whole in order to put a document in a given category.

Once the categories are finalized and the units, viz. the policy documents in terms of relevant words or phrases or sentences have been coded, separately for each of the groups to be compared, by two or more raters, we proceed to check inter-rater agreement and, if found adequate, we can combine the distributions of units across the classes as assigned by the different raters and proceed with the distribution on the basis of average frequencies (of course avoiding fractions). We thus get the frequency distributions for the different groups and apply appropriate statistical tools to detect variations and decide on the hypothesis or hypotheses we started with. Thus, working with a 3×2 table with ownership categories along columns and emphasis categories

along rows, we can apply the usual chi-square test for independence or no association between degree of emphasis and ownership pattern.

We can also work with several columns for sectors of manufacturing and examine possible differences across sectors. Essentially, Content Analysis will generate categorical data. In some cases, these are numerical or can be converted into numbers. Inferences based on such analysis should be validated (triangulated) by analyzing data on customer complaints or on product image or on complaint redressal, etc., collected from documents or from direct interviews of customers or potential customers or from consumers' fora and even product or manufacturer rating agencies.

3.5 Limitations of Content Analysis

It should be noted that content analysis

- does not tell us about causal connections between variables under study
- cannot explain why certain trends emerge.

It is primarily used to supplement the findings of mainstream research designs, such as survey research, where inferences are based on what Content Analysis eventually yields in terms of categorical data in situations which do not allow direct responses or evidences to be sought. However, Content Analysis can also be used in designing survey questionnaires by involving several experts or referees to go through the draft questionnaire or a part of the questionnaire to ascertain how much relevant or revealing the survey instrument id in relation to the research question to be addressed. Revision of the questionnaire may be suggested by the extent of agreement among the experts.

3.6 Concluding Remarks

When used properly, content analysis id a powerful data reduction technique. Its major benefit arises from the fact that it is a systematic and replicable technique for compressing many words of text into fewer categories based on explicit rules of coding. It has an attractive feature of being unobtrusive and being useful in

dealing with large volumes of data. The technique of content analysis goes far beyond simple counts of word frequency. However, there could be two significant flaws in content analysis that destroy its utility, viz. faulty definitions of categories and non-mutually exclusive and exhaustive categories.

References and Suggested Readings

Berelson, B. (1952). *Content analysis in communication research*. Glencoe III: Free Press.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.
Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.
Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg: Advanced Analytics LLC.
Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison Wesley.
Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. London: Sage.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. Roberts, C. W. (Ed.). (1997). *Text analysis for the social sciences: methods for drawing inferences from texts and transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.
Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *9*(3), 321–325.
Sim, J., & Wright, C. C. (2005). The Kappa statistics in reliability studies: Use, interpretation and sample size requirement. *Physical Therapy*, *85*(3), 257–268.
Weber, R. P. (1990). *Basic content analysis*. CA: Newbury Park.

Chapter 4
Scaling Techniques

4.1 Introduction

A test is generically an exercise to assess the performance of an individual (or even a group) on a task or of the aptitude or ability to perform or the attitude toward a task or a situation. Tests of intelligence or personality or attentiveness or similar other traits define a wide range of assessment or identification exercises. Usually, any such test administered on an individual will result in a measure, generally a number, which is referred to as a score. Most often, the test exercise involves an instrument like a question paper in academic examination or a questionnaire about self or about the environment. Such an instrument will again have some distinguishable items or components to take care of different aspects of the trait being identified or assessed.

The length of a test in terms of the number of items or components, the nature of administration, viz. administered individually or to a group, the manner in which responses to test items will be compared with some 'correct' or 'desired' or 'normative' responses, and such other features of a test will generally differ from one test to another.

And it is quite expected that the range of scores within which the score of any individual test taker will lie varies from test to test. The same test may be repeatedly applied to many different groups of individuals who may possess some similarity in respect of, say, age or gender or socioeconomic background or educational levels etc.

A raw score is the total number of score points a test taker obtains by answering questions correctly on a test. A percent correct score represents the percentage of questions a testee answered correctly on a test. For example, if the raw score was 35 for 35 out of 50 questions in a test correctly answered, the percent correct score will be 70. This score can be taken as an adjusted raw score to account for differences in lengths of different tests.

To achieve comparability, standardized testing programs report scaled scores. The scaled scores are obtained by a linear or nonlinear transform of the raw scores. Scaled scores have a common scale to account for differences in difficulty levels

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_4

across different forms as also across different groups of test takers placed in different contexts. Equating is the process by which the raw scores on a new form are adjusted to account for differences in form difficulty from a base or reference form. The utility of scaled scores comes from allowing for meaningful interpretations and, at the same time, minimizing misinterpretations and inappropriate inferences. Scores arise when we like to compare participants in some activity like sports and games, or music and dance, or elocution and drawing, debate or discussion. In any such case, we use some method of scoring somewhat specific to the context that takes care of the demographic features of the participants, the objectives of event, the idiosyncrasies of the evaluator(s), and extent of discrimination among the participants intended. And scaling of the raw scores as assigned to different participants lacks comparability across events, across occasions, across participant groups, and across evaluation agencies.

Apart from raw scores in educational and psychological tests, we come across problems of comparing products or similar objects with one or more features which are presented to some judges for assessing the relative differences among the different objects. Often, the judges will be required to rank the products, say n in number, from 1 for the best to n for the worst. It is possible to have tied ranks when some objects appear to be indistinguishable in respect of the feature(s) to the judges. Ranking a large or even moderately large number of objects usually becomes difficult and invites ties. As one alternative, we present to each judge each of the possible pairs of objects and ask the judge to prefer one in the pair say 'i' to the other, say 'j'. Subsequently, we proceed with the proportion of judges who prefer object k to object l within each possible object pair (k, l). Our intention is to reveal relative distances among the objects by associating with each object a scale value which helps us to visualize the objects as points on a straight line. We can involve several features of each product, several judges to scale these and present the object pairs on several occasions.

As an extension of this activity, known as product scaling, we can think of a visual representation of the relative distances among the objects by representing each object as a point in a plane or, at best, in a three-dimensional space. This activity is known as multi-dimensional scaling.

Section One deals with Scaling of Scores, while scaling of categorical responses obtained in a variety of socioeconomic surveys has been taken up in Section Two and Section Three is devoted to Product Scaling or Scaling of Concrete Entities involving directly the roles of judges who primarily assess the relative merit or strength of an entity relative to others in a group.

4.2 Scaling of Test Scores

The main defect of the prevalent system of ranking in scholastic tests consists of the adding of the raw scores of an individual on several tests to get his composite or total score and ranking all individuals on the basis of the total score. This is not a valid procedure since the same raw score x on different tests may involve different degrees

of ability and hence may not be equivalent in different tests. Hence, the raw scores have to be scaled under some assumption regarding the distribution of the trait which the test is measuring.

4.2.1 Percentile Scaling

Here we assume that the distribution of the trait under consideration is rectangular, under which we shall have percentile differences equal throughout the scale. To determine the scale value corresponding to a score x on a test, we have to find the percentile position of an individual with score x, i.e., the percentage of individuals in the group having a score equal to or less than x, which can be easily obtained from the score distribution assuming that 'score is a continuous variable. Regardless of the form of the original raw score distribution, the distribution of percentile scores will be rectangular. However, the distribution of raw scores is rarely rectangular, so that the basic assumption underlying the percentile scaling may not always be realistic. Thus, while using this scaling method, one should be aware of its limitations.

4.2.2 Z-Scaling or σ-Scaling

Here we assume that whatever differences that may exist in the forms of the raw score distributions may be attributed to chance or to the limitations of the test. In fact, the distributions of the traits under consideration are assumed to differ only in mean and s.d. Hence, the scores on different tests should be expressed in terms of the scores in a hypothetical distribution of the same form as the trait distribution with some arbitrarily chosen mean and s.d. The transformed scores are called linear derived scores. In particular, if the mean is arbitrarily taken to be zero and the s.d. to be unity, the scores are called standard scores or ?-scores or z-scores. To avoid negative standard scores, in linear derived scores the mean is generally taken to be 50 and the s.d. to be 10. If a particular test has raw score mean and s.d. equal to μ and σ, respectively, then the linear derived score (w) corresponding to a raw score x on that test is given by $(x - \mu)/\sigma = (w - 50)/10$, or, $w = 50 + 10((x - \mu)/\sigma) = 50 + 10z$, where w is the linear derived score with mean 50 and s.d. 10 and z is the standard score.

This linear transformation changes only the mean and the s.d., while retaining the form of the original distribution.

4.2.3 T-Scaling

In this case, we assume that the trait distribution is normal. The raw score distribution may deviate from normality, but the deviations from normality are attributed to

chance or to limitations of the tests. The mean and s.d. of the normal distribution of the trait may be arbitrarily taken to be 50 and 10, respectively. To get the scaled score corresponding to a raw score x, first we find, as in percentile scaling, the percentile position (P) of an individual with score x and the find the point (T) on a normal distribution with mean 50 and s.d. 10, below which the area is P/100. This is given by $\mathbf{Tau}((T-50)/10) = P/100$, where $\mathbf{Tau}(u)$ is the area under the curve of the standard normal variable from -∞ to u.

The scaled score obtained by this process is called T-score in memory of the psychologists Terman and Thorndyke. The scale is due to McCall.

The scaled score obtained by this process is called T-score in memory of the psychologists Terman and Thorndyke. The scale is due to McCall.

Normalized scores are also expressed as stanine (standard nine) scores. The stanine scale takes nine values from 1 to 9, with mean 5 and s.d. 2. When a distribution is transformed to a stanine scale, the frequencies are distributed as follows:

STANINE DISTRIBUTION

Stanine Score	1	2	3	4	5	6	7	8	9
Percentage on each score (rounded)	4	7	12	17	20	17	12	7	4

A transformation is nonlinear if it changes the form of the distribution. Normalized scores and percentile scores are merely special cases of nonlinear transformation of the raw scores. For nonlinear transformation, any form of distribution may be chosen.

4.2.4 Method of Equivalent Scores

Here we do not make any assumption about the distribution of the trait under consideration. The appropriate trait distribution is obtained by graduating the raw score distribution by an appropriate Pearsonian curve.

Let x and y be the scores on two tests, having probability density functions f and h, respectively, obtained by some process of graduation. Now, two scores x_i and y_i, on the two tests, are to be considered equivalent, in the sense that they bring into play equal amounts of the trait, if and only if

$$\int_{-\infty}^{x_i} f(u)du = \int_{-\infty}^{y_i} h(u)du.$$

For practical convenience, an equivalence curve may be obtained by computing a number of pairs of equivalent scores, (x_i, y_i) and fitting to the corresponding set of points an appropriate curve, say $y = g(x)$.

Equivalent scores can also be obtained from the score distributions for x and y without going into the process of graduation. First, two ogives are drawn on the same graph paper. Two scores x_i and y_i with the same relative cumulative frequency are then regarded as equivalent. For the purpose of comparison or combination, the raw

scores on different tests may be converted into equivalent scores on a standard test. In this method, the form of the distribution of equivalent (transformed) scores is the same as that of the standard test. If, however, the standard test score has a normal distribution, the method reduces to normalized scaling.

Example 4.1 The raw score distributions for Vernacular and English for a group of 500 students are given below. One of two students got 80 in Vernacular and 40 in English, while the other got 60 in both. Compare their performances by (i) percentile scaling, (ii) linear derived scores, (iii) T-scaling and equivalent scores ogive method.

First, we have to remember that a score of 80 is to be considered as an interval from 79.5 to 80.5 and similarly for the other scores.

To obtain the percentile positions, we obtain the cumulative frequencies (less than type) for both Vernacular and English. They are shown in Table 5.3.

Hence, the percentile positions corresponding to 80.5 and 60.5 in Vernacular are given by

$$P_{80.5}(Vern) = [(497 + 0.6)/500] \times 100 = 99.52;$$

$$P_{60.5}(Vern) = [(436 + 7.2)/500] \times 100 = 88.64.$$

Similarly, for English,

$$P_{.5}(Vern) = [(497 + 0.6)/500] \times 100 = 99.52;$$

$$P_{60.5}(Vern) = [(436 + 7.2)/500] \times 100 = 88.64.$$

Distributions of Scores in Vernacular and English of a Group of 500 Students

Score −−	*Vernacular Frequency*	*English Frequency*
0 − 4	–	3
5 − 9	–	6
10 − 14	–	12
15 − 19	6	23
20 − 24	7	35
25 − 29	18	45
30 − 34	34	74
35 − 39	56	72
40 − 44	84	78
45 − 49	74	53
50 − 54	104	46
55 − 59	53	29
60 − 64	36	18
65 − 69	16	5
70 − 74	9	1
75 − 79	0	–
80 − 84	3	–

Cumulative Distributions of Scores in Vernacular and English of a Group of 500 Students

Score	*Vernacular*	*English*
−−	*Cumulative Frequency*	*Cumulative Frequency*
0 − 4	–	3
5 − 9	–	9
10 − 14	–	21
15 − 19	6	44
20 − 24	13	79
25 − 29	31	124
30 − 34	65	198
35 − 39	121	270
40 − 44	205	348
45 − 49	279	401
50 − 54	383	447
55 − 59	436	476
60 − 64	472	494
65 − 69	488	499
70 − 74	497	500
75 − 79	497	500
80 − 84	500	500

Hence, the total scaled score for Student 1, getting 80 in Vernacular and 40 in English, is by percentile scaling $99.52 + 57.12 = 156.64$ and that of Student 2, getting 60 in both Vernacular and English, is $88.64 + 95.92 = 184.56$.

Thus, we see that the relative performances of the two students are quite different although their total raw scores are equal.

For linear derived scores with mean 50 and s.d. 10, we require the means and s.d.'s of scores in the two subjects.

Hence, the w scores are given by

$$T80(Vern.) = 50 + [(80 - 47.09)/11.32] \times 10 = 79.07,$$

$$T60(Vern.) = 50 + [(60 - 47.09)/11.32] \times 10 = 61.40,$$

$$T40(Eng.) = 50 + [(40 - 37.87)/13.10] \times 10 = 51.63,$$

and

$$T60(Eng.) = 50 + [(60 - 37.87)/13.10] \times 10 = 66.89.$$

As such, the total w-score of Student 1 is $79.07 + 51.63 = 130.70$, and that of Student 2 is $61.40 + 66.89 = 128.29$. Linear derived scores, therefore, show that Student 1 is slightly superior to Student 2.

Now, for T-scaling, percentile positions have to be converted into T-scores. We have, for Vernacular,

$$T80(Vern.) = 50 + \tau_{.9952} \times 10 = 75.90, T60(Vern.) = 50 + \tau_{.8864} \times 10 = 62.08.$$

Next, for English,

$$T40(Eng.) = 50 + \tau_{.5712} \times 10 = 51.79, T60(Eng.) = 50 + \tau_{.9952} \times 10 = 67.41.$$

Hence, the total T-score of Student 1 is $75.90 + 51.79 = 127.69$, and the total T-score of Student 2 is $62.08 + 67.41 = 129.49$.

Thus T-scaling shows that Student 2 is slightly superior to Student 1.

Determination of Equivalent Scores in English and Vernacular from the Ogives

In the equivalent scores method, let us take Vernacular as the standard. From the above figure, we find that a score of 40 in English is equivalent to a score of 49.8 in Vernacular and a score of 60 in English is equivalent to a score of 66.9 in Vernacular.

Hence, the total score of Student 1 in terms of Vernacular score is $80 + 49.8 = 129.8$ and that of Student 2 is $60 + 66.9 = 126.9$.

Thus, this method again shows that Student 1 is slightly superior to Student 2.

4.3 Scaling of Categorical Responses

Attributes like perception, attitude, honesty, sincerity are unobserved latent variables. Response of an individual to a statement or a question relating to such an attribute is generally recorded as belonging to one of several prefixed categories. To a question on how did one feel about some program on Research Methodology, a respondent can state Fair or Good or Very Good or Excellent. (One may add a fifth category, viz. nothing particular or undecided.). Given a statement like My colleagues/ peers give me full cooperation in discharging my duties in an organizational climate survey, possible reactions could be completely disagree, disagree, cannot comment, agree, and completely agree.

To proceed further to analyze the ordinal data, we need to assign scale values to each category, assuming a certain probability distribution of the underlying latent variable over a support $[a, b]$. Usually, we have an odd number (5 *or* 7) of categories and we assume a normal distribution for the underlying trait or latent variable. We can have any number of categories and can assume any other trait distribution. The upper-class boundaries for the latent variable are, say $a = x_1, x_2, \ldots, x_k = b$. The task is to estimate these unknown boundaries in terms of the observed frequencies of responses in the corresponding categories and then finding midpoints of the classes.

4.3.1 Estimation of Boundaries

Given frequencies $f_1, f_2, \ldots, f_k$ in the k categories with cumulative frequencies $F_1, F_2, \ldots, F_k$, respectively, we equate F_i to $\Phi(xi)$ to get x_i from a table of Φ-values (left tail areas under the assumed normal trait distribution with mean 0 and s.d. 1). It can be shown that x's thus determined are reasonably good estimates of the unknown boundary points for the prefixed number of categories and, hence, of intervals for the trait or latent variable. Intervals for the trait are replaced by means of the trait distribution truncated between the boundaries for the intervals.

Assuming normality, the truncated mean for the class (x_{i-1}, x_i) is given by

$$s_i = (\phi(x_{i-1}) - \phi(x_i)/(\Phi(x_i) - \Phi(x_{i-1}).$$

The first class for the normal trait distribution is $(-\infty, x_1)$ while the last class is (x_{k-1}, ∞) with

$$\Phi(-\infty) = 0, \Phi(\infty) = 1, \phi(-\infty) = o = \phi(\infty).$$

4.3.2 *Finding Scale Values*

Sometimes, scale values are taken as equidistant integers like 1, 2, 3, 4, 5 or 2, 4, 6, 8. These then become data- and model-invariant. We must note that trait intervals corresponding to different response categories are not generally equal. In Likert's scaling, the scale values are data-dependent and will depend on the particular set of observed frequencies as also on the trait distribution assumed.

Example 4.2 Quality of service offered at a newly opened customer service center judged by 60 visitors is shown below.

Cumulative Distributions of Scores in Vernacular and English of a Group of 500 Students

Grade	*Frequency*	*Cumulative Frequency*
Poor	3	3
Fair	14	17
Good	26	43
VeryGood	13	56
Excellent	4	60

Assuming an $N(0, 1)$ distribution for the underlying trait, viz. perceived quality, upper-class boundaries become -1.65, -0.58, 0.55, 1.48 and ∞. Thus, scale values are -2.06, -1.04, 0.01, 1.02 and 1.90. If we take an arbitrary mean of 3 and an s.d. 1, these values become 0.94, 1.96, 3.01, 4.02, and 3.90 and are not equidistant.

In a situation, where the trait can vary over the range 0 to 5 and we can justifiably assume a uniform distribution of the trait, the scale values will be 0.25, 0.83, 2.50, 3.45, and 4.67, respectively.

4.4 Use of U-Shaped Distributions

Bose and Sen (1958) noted that in some cases respondents tend to exhibit extreme positions and there are only few responses in the middle (undecided or neutral) category. In such cases, they have advocated the use of some U-shaped distribution to work out scale values for the categories. While the Pearsonean $Type\ II$ distribution with the probability density function

$$f(x) = y_0(1 - x^2/a^2)^m, -a < x < a; -1 < m < 1$$

was a candidate, it is not possible to guess the value of m from the observed data. Bose and Sen suggested that if from a uniform distribution over the range $-\lambda < x < \lambda$ we take out the standard normal distribution, the resultant density

$$f(x) = [H - 1/\sqrt{(2\pi)}exp(-x^2/2)], -\lambda < x < \lambda; = 0, otherwise$$

will correspond to a U-shaped distribution. To make the density nonnegative, the constant H has to be greater than $1/\sqrt{(2\pi)}$, a value attained when $\lambda = 2.49$. In fact, λ can vary between 0 and 2.49. They took $\lambda = 2.054$ with $H = 0.477$ to ensure a moderate curvature of the distribution that will have the density function

$$f(x) = 0.477 - 1/\sqrt{(2\pi)}exp(-x^2/2), -2.054 < x < 2.054.$$

Starting with five categories for the response, they give expressions for the five conditional means in terms of entries in tables prepared by them. They considered Krishnan's data on study habits consisting of 39 questions in a survey covering a batch of 23 students. To each question, one of the following 5 ratings was attached by each student (though a few of the students did not attempt a few of the items):

Rating

5 if the statement is always true of you;
4 if the statement is most often true of you;
3 if the statement is often true of you;
2 if the statement is sometimes true of you;
1 if the statement is never true of you.

In this case, the statements were such that extreme types of ratings are quite likely and the underlying distribution of the trait (rating) can be reasonably assumed to be U-shaped. To avoid negative scale values, a formula like $Y = 50 + 24.39m$ was adopted, where m is the mean for a doubly truncated part of the underlying U-shaped trait distribution. These scaled scores for the different items could be added up to yield the total score of a student. By examining the total scores, we can study the attitude of the students against the set of questions.

4.5 Product Scaling

Product features like appearance or performance (judged in terms of user-friendliness or ease of maintenance or clarity of images or of sound, etc.) may be rated by different users or viewers differently, and we may be interested to work out some scale values for different brands or versions of a product which will facilitate decisions regarding a choice among them. Several different situations may arise. Handwriting, drawing,

music, etc., are well-known products whose quality cannot be directly measured. As such ranking of several such products (not to speak of placing their intrinsic merits on a cardinal scale) becomes difficult. The method of paired comparisons introduced by Thurstone (1927) and later extended by Mosteller (1951), Bradley and Terry (1952) Gibson (1953), Gibson and Burros (1994), etc., provides a way out.

The method involves a number of judges or raters—assumed equally competent—to each of whom all possible pairs of products are presented and preferences of these judges for one of the two members in each pair are noted. These preferences are then incorporated in a theoretical framework to develop sale values which can be attached to the products under study.

In scaling, a number of products which do not admit of copies or prototypes but can be presented on more than one occasions or a set of products where different copies or prototypes of the same product can be considered on different occasions, it is desirable to ensure that scaling takes due account of variations in judgment (preferences) from occasion to occasion, from one prototype to another besides variations among judges so that differences in scale values reflect true differences in quality or ability. The possible situations likely to arise are

(1) Several pairs of prototypes corresponding to the product pair are presented to each of several judges on different occasions, a separate pair being considered on each occasion. In particular, one prototype pair corresponding to one product pair can be considered by each of the judges on one occasion only.
(2) The same pair of prototypes corresponding to a product pair is presented on several occasions to each of several judges, or in particular to one judge only.
(3) Each pair of products is presented to each of several judges on each of several occasions. This is the case when prototypes of a product cannot be conceived or used.

Mukerjee (1980) provides complete derivations of scale separations among the products in each of these three cases, under appropriate assumptions. With a suitable choice of origin and scale, these separations can yield the absolute scale values for the different products.

We consider the third situation first. We assume that the n products are inherently different among themselves w.r.t. the feature or trait assessed and that the m judges are competent to bring out such differences. Thus, there will be no ties in the ranks assigned by any judge to the n products. We also assume that each product has a true value for the trait, although such a value cannot and will not be assigned to any product by any judge. Each of the $n(n-1)/2$ pairs of products will be presented to each judge on each of p occasions. On each occasion, a judge will be required to prefer one product in a pair to the other. Let $p_{ik.j}$ be the proportion of judges preferring product i to product k on the jth occasion. Thus, $p_{ik}^{-} = \sum_j p_{ik.j}$ is the average proportion. Let $e_{ik.j}^{(l)}$ be the error in judgment committed by judge l on the jth occasion when the product pair (i, k) is presented to him. We assume that these errors are jointly normally distributed with

$$Var(e_{ik.j}^{(l)}) = \delta_{ik}^2 \quad for \ all \ \ j \ \ and \ \ l \ \ and \ \ Cov(e_{ik.j}^{(l)}, e_{pq.s}^{(t)}) = 0,$$

if the set l, t, j, s contains at least three distinct integers. For any occasion, let us define the indicator variable $Y_{ik.j}^{(l)} = 1$ if product i is preferred to product k by the lst judge; $= 0$ otherwise.

Observed values of these indicator variables form the basis of the method of paired comparisons. We note that $p_{ik} = 1/p \sum Y_{ik.j}$.

Evidently, $Y_{ik.j}^{(l)} = 1$ if and only if $\mu_i + e_{ik.j}^{(l)} > \mu_k + e_{ki.j}^{(l)}$ so that

$$EY_{ik.j}^{(l)} = Prob[\mu_k + e_{ki.j}^{(l)} - \mu_i - e_{ik.j}^{(l)} < 0] = Prob[S_{ik.j<0}^{(l)}], say.$$

Under the assumptions made, $S_{ik.j}^{(l)}$ is univariate normal with mean $\mu_i - \mu_k$ and variance $2\delta_{ik}^2$. Hence,

$$E[p_{ik}^-] = \Phi[(\sqrt{2}\delta_{ik})^{-1}(\mu_i - \mu_k)]$$

Applying the method of moments for estimation, we obtain the estimated scale separation as

$$\mu_i - \mu_k = \sqrt{2}\delta_{ik}\Phi^{-1}(p_{ik}^-)$$

We can reasonably assume $\delta_{ik} = \delta$ for all i and k.

Taking $\sqrt{2}\delta$ as the unit of separation, the scale separation between product i and product k becomes $\cap\phi^{-1}(p_{ik})$. We can subsequently estimate $\mu_i - \mu^- = 1/n \sum_{k \neq i}(\mu_i - \mu_k)$. If we take the mean μ^- as the origin, we can work out scale values for the different products.

Let us consider the following example.

Example 4.3 Suppose 200 individuals were asked about their preferences for four different types of music.

Music Type	1	2	3	4
Music Type 1	.500	.770	.878	.892
2	.230	.500	.743	.845
3	.122	.257	.500	.797
4	.108	.155	.203	.500

Under the usual assumption of normality of the distribution of difference in judgments with means $S_i - S_j$ and s.d. σ_{i-j}, and with the constant σ_{i-j} taken as the unit of the scale, we get the matric of scale separations $S_i - S_j$ as follows

Music Type	1	2	3	4
Music Type 1	0	.739	1.165	1.237
2	−.739	0	.653	1.015
3	−1.165	−.653	0	.831
4	−1.237	−1.015	−.831	0
Column Mean	−.785	−.232	.247	.771

With the origin at $\bar{S}$, the mean scale value, the column means give us the corresponding scale values for the four music types. With origin at S_1, on the other hand, we get the following scale values:

$$\begin{pmatrix} Music\ Type & 1 & 2 & 3 & 4 \\ Scale\ Value & 0 & .553 & 1.032 & 1.556 \end{pmatrix}.$$

4.6 Other Unidimensional Scaling Methods

In the class of unidimensional scaling, the one that is worth mentioning besides Likert's and Thurstone's is Guttman scaling. We should remember that in any such scaling exercise, we are assigning a scale value to each individual who responds to an item in one of several categories or considers the preferences of some judges between members of each individual pair in respect of some property or trait or even take into account responses by an individual to a series of ordered items, ordered according to difficulty level or maturity level or some such trait. Guttman scaling applies to the last situation where a cumulative score or scale value is assigned to each respondent to several ordered items with binary responses. Guttman scaling developed by Louis Guttman (1944, 1950) as part of his classic work American Soldiers, this is a multi-item scaling that provides a cumulative score to each individual responding to a set of questions with binary response. The method takes into account the position of each question or item with regard to difficulty or some similar aspect also. Let us consider the example discussed in Abedi (2010) in which each of five children indicated whether he/she has mastered a topic in Mathematics with the response 1 or not with a response.

			Problems		
Child	*Counting*	*Addition*	*Subtraction*	*Multiplication*	*Division*
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	1	1	1	1	1

In such a well-structured pattern that is expected in a logical situation, we can represent the data by the following order:
Counting Child 1; Addition Child 2; Subtraction Child 3;
Multiplication Child 4; and Division Child 5.
The order can be transformed into a set of equispaced numbers like
Child 1–2; Child 2–4; Child 3–6;
Child 4–8; and Child 5–10.

The score of a child (row) is proportional to the number of nonzero entries. This is the case of a perfect Guttman scale. An imperfect scale could relate to a situation where some child puts 1 for addition but 0 for counting or puts 1 for multiplication

but 0 for subtraction. Deviations from the ideal scale could be considered as errors and the coefficient of reproducibility of the scale value assigned to an individual could be obtained as

Reproducibility = 1 − (actual number of errors)/(number of possible errors)

Considering opinions or attitudes or desires and similar traits, the items in a test have to be so framed in a logical sequence that a positive answer to anyone would imply a positive answer to all the previous items. A respondent would be asked to respond to some item in the middle and, depending on whether the response is positive or negative, responses to other items would be sought. Otherwise, with items properly graded, the respondent would be asked to start with the easiest or the most basic question and wherever he/ she stops answering positively, we can assign a cumulative score to the respondent, assuming perfect reproducibility. This saves a lot of time and effort in an opinion survey or structured interview. The score assigned to a respondent would, of course, depend on numbers that will be associated with the level of each question, levels being usually equidistant. Rasch model is a one-parameter model in the family of item response latent trait models which produces an interval scale that determines item difficulty as well as person measures of the trait, considering a set of carefully selected survey items The scale is then used to show person measure. The scale units are logits (log odds ratio units). Usually, 4 to 8 related items are considered. For convenience in understanding, logit values are transformed usually to a 10-point scale by using the transformation

Measure (new) = 10 × [old measure − minimum]/[maximum − minimum] where the old measure is the logit and the new measure has the 10-point scale and the maximum and minimum relate to item difficulty of any item. In the Rasch model, we assume that the probability of correct response to an item as a logistic function of the difference between the person (proficiency or ability) parameter and the item (difficulty or maturity level) parameter. Let the random variable X_{ni} stand for the response (1 for a correct or positive response and 0 for an incorrect or a negative response) of individual n in item i. We assume that

$$Prob(X_{ni} = 1) = exp(\beta_n - \delta_i)/[1 + exp(\beta_n - \delta_i)]$$

where β_n is the person measure (of ability or positivity of trait or proficiency) of person n and δ_i stands for difficulty of item i. For the same person, score difference on two items with difficulty levels δ_1 and δ_2 will depend only on $\delta_1 - \delta_2$. If we denote the total score of a person n ?? two items by r_n, then the item parameters can be estimated from the conditional log odds ratio for $[X_{ni} = 1 | r_n = 1]$.

Parameters in the logistic model are derived using the conditional maximum likelihood method. We eventually get a person score as also an item score which can be shown as locations on a continuous latent variable. It has been argued that the total score for an individual in all the items has a nonlinear relation with the person's ability and that is why the logistic function has been used. Among the shortcomings of the Rasch scaling, one relates to the assumption that the different items have the same

discriminatory ability. When polytomous responses are considered, Rasch scaling has some semblance with Likert scaling. And in case of dichotomous responses, this is somewhat analogous to Guttman scaling.

4.7 Concluding Remarks

Nominal data may give rise to categories, e.g., religious or linguistic or similar groups. Scaling does not apply to such categorical data. Ranks can be assigned in the case of ordinal data, and these are equidistant. Equal differences in ranks may not imply equal differences in the underlying trait. Scale values may be equidistant only when all the class frequencies are equal and the underlying trait follows a rectangular distribution. Scaling of products (like art objects or even multifaceted consumer goods) presented in pairs to a group of judges is based on a matrix where elements are proportions of judges preferring one product to the other, considering all possible pairs. Multi-dimensional scaling is a useful tool to visualize relative positions of products.

References and Suggested Readings

Bose, P. K., & Chaudhuri, S. B. (1955). Scaling procedure in scholastic and vocational test. *Sankhya*, *15*, 197–206.

Bose, P. K., & Sen, P. K. (1960). Measurement of attributes for U-shaped distributions. *Calcutta Statistical Association Bulletin*, *8*, 31–42.

Bradley, R. A., & Terry, M. E. (1952). The rank analysis of incomplete block designs 1. The method of paired comparisons. *Biometrika*, *39*, 324–345.

Dunn-Rankin, P., Knezek, G. A., Wallace, S., & Zhang, S. (2004). Lawrence Erlbaum Associates Publishers.

Gibson, W. A. (1953). A least squares solution for cas IV of the law of comparative judgement. *Psychometrika*, *18*, 15–21.

Gibson, W. A., & Burros, R. H. (1954). A least squares solution ot case III of the law of comparative judgement. *Psychometrika*, *19*, 51–54.

Krishnan, B. (1956). Psychological Studies, No. 1.

Likert, R. A. (1982). Archives of Psychology, No. 140.

Mosteller, F. (1951). Remarks on the method of paired comparisons, I: The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, *16*, 3–9.

Mukerjee, R. (1980). A generalized procedure for product scaling. *IAPQR Transactions*, *2*, 71–83.

Nunnally, J. C. (1981). *Psychometric theory*. New York: Mcgraw Hill.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286.

Chapter 5
Data Integration Techniques

5.1 Introduction

We undertake a study of ranking a given set of competing and alternative 'locations,' using different methods of data integration. Among other aggregating techniques, TOPSIS Method [TM]—a specific Multiple Criteria Decision Making [MCDM] Algorithmic Tool/Technique—will be discussed at length. This method is thoroughly discussed in the reference papers listed at the end. Another less popular ELECTRE METHOD is also discussed in the book [2]. It is relatively unexplored in the area of what is known as 'Data Integration Techniques.' There are very few published papers. We will be discussing computational details underlying this technique.

In Sect. 5.2, we start with a brief description of different data integration tools/techniques. In Sect. 5.3, the computational algorithm underlying the TM is discussed in a theoretical framework and this we do by borrowing the standard notations. This is geared toward providing maximum comfort to the readers—at least to those who are familiar with application areas. In Sect. 5.4, we work on a data set for illustrating the computations. It is only natural that we, as statisticians, provide further insights into the intricacies of application of TM in real data. We close the chapter with some remarks in Sect. 5.4.

5.2 Elementary Methods for Data Integration[1]

Assume there are 'm' locations and there are 'n' sources wherefrom 'meaningful inputs' emerge into each of these locations. Each 'input' is quantified and measured

[1]This section draws material from the author's co-published paper: 'Data Integration Techniques,' published in International Journal of Tropical Agriculture [IJTA], Serial Publications; ISSN: 0254 - 8755, Vol. 33, No. 2, April–June 2015, pp. 1339–1344. Permission for re-use was obtained from co-author, Dr. Mamunur Rashid, as well as from the Publisher.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_5

by a positive 'score' in the same scale—across all sources and locations. Assume that smaller the score, better is the 'rank' of the location. The problem is to 'integrate' the scores from all sources for each of the locations and provide an overall ranking of the locations. This is popularly known as 'data integration.'

We denote by $\mathbf{X} = ((\mathbf{x_{ij}}))$ the positive-valued score matrix of order $m \times n$—representing the locations as rows and the sources as the columns of the data matrix. In order that a location is adjudged the best with respect to a specific source, it is tacitly assumed that the score for this location has to be the least among all others in the list [of locations] with respect to this specific source. The objective of the study is to arrive at an 'overall' ranking of the locations, by taking into account their 'performance' across all the sources. It may so happen that the natural choice of one or more sources lend themselves to 'maximum-the-best' criterion. In such a case, one suggestion is to change the scores for all locations [across that column of the **X**-matrix] by taking their reciprocals. In fine, one has to ensure that all the scores for each source have the same interpretation in terms of 'min. - to - max.' going hand in hand with 'best to worst.' At times, the **X**-matrix is also termed as 'decision matrix.'

It is clear that for one single source of evaluation, the ranking of locations is trivial. Also as and when all the sources exhibit same relative positions of different sources, the solution is easy to arrive at. Non-trivial situations arise when there are 'wave-like' patterns in the data, and this is most expected scenario in practice with real data.

One natural and simple-minded approach has been to work out the average score for each location—by averaging the scores across all the sources. That means, we simply compute the row averages in the **X**-matrix of scores and use them for ranking of the locations. There are obvious limitations to this approach since it does not take into account the variations among the scores [of different locations] under each evaluation criterion i.e., source. It deals with one location at a time. Apart from this, the point to be noted is that while we are working out the average score, we are assuming that all the sources are equally important and hence they possess the same weights. This has been a point of concern to the data analysts, and they have worked out a solution to this problem. Naturally, we may call upon 'subject experts' and utilize their knowledge in ascertaining relative weights of the different sources. Failing to have access to such experts' inputs, data-driven techniques have been suggested in the literature. One such technique is based on 'Shannon Entropy Measure.' There are two other data-driven techniques for ascertaining source weights in such contexts.

We will discuss and apply two of these techniques for evaluation of weights of different sources. Once the weights are determined, the formulae for applying the weights are the same to arrive at the individual rankings of the locations.

Before we proceed further, we may also mention in passing the following:

Since the main purpose of this exercise is to find overall ranks of the locations, it has been suggested that the Data Matrix $X = ((x_{ij}))$ may as well be converted into a 'Matrix of Ranks'—by working out the ranks of the locations for each source. That would mean—observations are to be ranked column-wise. Once this is done, usual 'weighted average' technique may be employed for further analysis.

We will not pursue these discussions/computations further.

5.3 Topsis Method: Computational Algorithm in a Theoretical Framework and Related Issues

We now describe the necessary steps with reference to computation of Entropy Weight Measure.

Step 1. Transferring the Decision Matrix to a Proportion Matrix

In order to compute the entropy measure for the jth source, the related values in the decision matrix are first normalized in terms of proportion as:

$$p_{ij} = x_{ij} / \sum_{1 \le t \le m} x_{tj}; 1 \le i \le m; 1 \le j \le n.$$

Step 2. Calculating the Entropy Measure for each source

In this step, the entropy of the jth Source, E_j, is calculated as follows:

$$E_j = -\alpha \sum_{1 \le i \le m} p_{ij} ln(p_{ij}); 1 \le j \le n.$$

where, $\alpha = 1/ln(m)$; m being the total number of alternative locations.

Next, the operation of subtraction is used to measure the Degree of Diversity, D_j, relative to the corresponding anchor value (unity), using the formula: $D_j = 1 - E_j; 1 \le j \le n$.

Step 3. Defining Source-wise Entropy Weights

The entropy weight W of each source is calculated using $W_j = D_j / \sum_t D_t; 1 \le j \le n$.

We have thus ascertained the weights of each of the sources as per entropy measure. Another method is based on the notion of 'Coefficient of Variation' $[CV]$ defined as $CV = sd/mean$. Weights are taken to be directly proportional to the respective CV's or their squares [for computational simplicity]. We will also describe another method, known as 'method of reversal.'

Once the weights are chosen [by any convenient method], these weights are then incorporated into a suitable formula to calculate an overall score for each location. The Topsis Method [TM] is chosen because of its high speed, accuracy, and compatibility [5]. The algorithm of this technique is summarized as follows:

(1) Transfer the Decision Matrix to a Normalized Decision Matrix $R = ((r_{ij}))$ [in the sense of unit squared length i.e., $\sum_i r_{ij}^2 = 1$ for each $j = 1, 2, \ldots, n$]:

$$r_{ij} = x_{ij} / \sqrt{\sum_t x_{tj}^2}; 1 \le i \le m; 1 \le j \le n.$$

(2) Weigh the Normalized Decision Matrix R using the Source Weights:

$$V = ((v_{ij})); v_{ij} = W_j r_{ij}; 1 \le i \le m; 1 \le j \le n.$$

(3) Define the 'Ideal Positive' and 'Ideal Negative' solutions:

$$V_j^+ = min_{1 \le i \le m} v_{ij}; 1 \le j \le n; V_j^- = max_{1 \le i \le m} v_{ij}; 1 \le j \le n.$$

Note that in the above, 'Ideal' corresponds to minimum v-score while 'Anti-ideal' corresponds to maximum v score across all the locations for each source.

(4) Measure the distances, d_i^+ and d_i^-, from the ideal and negative ideal solutions:

$$d_i^+ = \left[\sum_j (v_{ij} - V_j^+)^2\right]^{1/2}; d_i^- = \left[\sum_j (v_{ij} - V_j^-)^2\right]^{1/2}.$$

In the above, the 'distance measure' used is referred to as 'Euclidian distance' or 'Euclidian Norm,' denoted by L_2.

(5) Determine the relative closeness of alternatives to ideal solution by computing what is known as 'Composite Index' [CI]:

$$CI_i = d_i^+ / [d_i^+ + d_i^-]; i = 1, 2, \ldots, m.$$

These composite indices are used for final ranking of the methods, the rule being: min. - to - max. for ranks 1 - to - m.

5.4 Topsis Method: Computational Details in an Illustrative Example

We take up a hypothetical example involving eight different locations and seven distinct sources. The data set is shown in Table 5.1.

In what follows, we deal with two different approaches for ascertaining the weights of the sources: (i) Entropy measure and (ii) CV^2. There is also a third approach known as 'Reversal Method' which is explained below.

1. Start with equal weights for all the sources and rank the locations, using TM and following one specific distance measure.
2. Reverse the roles of sources and locations, and rank the sources, using the ranks of the locations derived in Step 1 as their weights.
3. Now reverse their roles again and use ranks of the sources [derived in Step 2] as their weights.
4. This yields the ranks of the locations eventually.

We will not pursue this third approach here.

Before proceeding further, we display the weights as determined by the first two methods for the above data set (Tables 5.2, 5.3, 5.4, 5.5, 5.6, and 5.7).

We now proceed with the rest of the computations.

Table 5.1 Locations versus sources: data on toxic release [in percentage] across different locations and sources

Location Sl No.	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7
I	7	13	21	3	24	21	17
II	12	9	18	3	32	28	11
III	17	4	23	7	22	19	23
IV	9	11	17	15	15	23	19
V	14	10	13	8	21	19	25
VI	6	11	19	5	23	21	22
VII	15	9	13	14	18	19	18
VIII	16	11	10	5	13	20	15
$\sum_i x_{ij}$	96	78	134	60	168	170	150
$\sum_i x_{ij}^2$	1276	810	2382	602	3772	3678	2958

Table 5.2 Normalized scores $((p_{ij}^{`}s))$ and entropy-based source-specific weights

Location Sl No.	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7
I	0.073	0.166	0.156	0.0500	0.143	0.123	0.113
II	0.125	0.116	0.134	0.0500	0.190	0.165	0.073
III	0.177	0.051	0.172	0.118	0.132	0.112	0.153
IV	0.094	0.141	0.127	0.250	0.089	0.135	0.127
V	0.146	0.128	0.097	0.133	0.125	0.112	0.167
VI	0.062	0.141	0.142	0.083	0.137	0.123	0.147
VII	0.156	0.116	0.097	0.233	0.107	0.112	0.120
VIII	0.167	0.141	0.075	0.083	0.077	0.118	0.100
$\sum_i p_{ij}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
E_j	0.9723	0.9817	0.9849	0.9210	0.9834	0.9959	0.9872
$D_j = 1 - E_j$	0.0277	0.0183	0.0151	0.0190	0.0166	0.0041	0.0128
W_j	0.2438	0.1611	0.1329	0.1672	0.1462	0.0361	0.1127

Table 5.3 CV^2-based source-specific weights

Location Sl No.	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7
$\sum_i x_{ij}$	96	78	134	60	168	170	150
$\sum_i x_{ij}^2$	1276	810	2382	602	3772	3678	2958
$\sum_i (x_{ij} - \bar{x}_j)^2$	124.0	49.5	137.5	152.0	244.0	65.5	145.5
CV^2	0.1076	0.0651	0.0610	0.3377	0.0692	0.0181	0.0517
CV^2-based weights $W(CV^2)$	0.1514	0.0916	0.0859	0.4753	0.0974	0.0255	0.0728

Table 5.4 Normalized decision matrix $R = ((r_{ij}))$

Location Sl No.	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7
I	0.1960	0.4568	0.4303	0.1223	0.3908	0.3463	0.3126
II	0.3359	0.3162	0.3688	0.1223	0.5210	0.4617	0.2022
III	0.4759	0.1405	0.4713	0.2853	0.3582	0.3133	0.4229
IV	0.2519	0.3865	0.3483	0.6113	0.2442	0.3792	0.3493
V	0.3919	0.3514	0.2664	0.3261	0.3419	0.3133	0.4597
VI	0.1680	0.3865	0.3893	0.2038	0.3745	0.3463	0.4045
VII	0.4199	0.3162	0.2664	0.5756	0.2931	0.3133	0.3310
VIII	0.4479	0.3865	0.2049	0.2038	0.2117	0.3298	0.2758
$\sum_i r_{ij}^2$	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 5.5 Weighted normalized decision matrix $V = ((v_{ij})) = ((W_j r_{ij}))$ with entropy weights and ideal positive and ideal negative solutions

Location Sl No.	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7
I	0.0478	0.0736	0.0572	0.0204	0.0571	0.0125	0.0352
II	0.0819	0.0509	0.0490	0.0204	0.0762	0.0167	0.0228
III	0.1160	0.0169	0.0626	0.0477	0.0524	0.0113	0.0477
IV	0.0614	0.0623	0.0463	0.1022	0.0357	0.0137	0.0394
V	0.0955	0.0566	0.0354	0.0545	0.0500	0.0113	0.0518
VI	0.0409	0.0623	0.0517	0.0341	0.0547	0.0125	0.0456
VII	0.1024	0.0509	0.0354	0.0962	0.0428	0.0113	0.0373
VIII	0.1092	0.0623	0.0272	0.0341	0.0309	0.0119	0.0311
W_j	0.2438	0.1611	0.1329	0.1672	0.1462	0.0361	0.1127
Ideal positive solutions V^+	0.0409	0.0169	0.0272	0.0204	0.0309	0.0113	0.0228
Ideal negative solutions V^-	0.1160	0.0736	0.0626	0.1022	0.0762	0.0167	0.0518

Remark 5.1 We leave it to the interested reader to take up the exercise of determination of (i) distance measures d_i^+ and d_i^- with CV^2-based weights, (ii) composite indices, and, finally, (iii) the ranks of the locations based on CV^2-based weights.

Remark 5.2 It would be interesting to work out ranking exercise using the 'Reversal Method' as well.

Remark 5.3 The distance measure used in the above is based on the notion of 'Squared Distance' or 'Euclidian Distance,' and it is usually denoted by L_2. The expressions for $d_i^+ = [\sum_j (v_{ij} - V_j^+)^2]^{1/2}$ and $d_i^- = [\sum_j (v_{ij} - V_j^-)^2]^{1/2}$ may be explicitly written as

Table 5.6 Weighted normalized decision matrix $R = ((v_{ij})) = ((W_j r_{ij}))$ with CV^2 weights and ideal positive and ideal negative solutions

Location Sl No.	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7
I	0.1960	0.4568	0.4303	0.1223	0.3908	0.3463	0.3126
II	0.3359	0.3162	0.3688	0.1223	0.5210	0.4617	0.2022
III	0.4759	0.1405	0.4713	0.2853	0.3582	0.3133	0.4229
IV	0.2519	0.3865	0.3483	0.6113	0.2442	0.3792	0.3493
V	0.3919	0.3514	0.2664	0.3261	0.3419	0.3133	0.4597
VI	0.1680	0.3865	0.3893	0.2038	0.3745	0.3463	0.4045
VII	0.4199	0.3162	0.2664	0.5756	0.2931	0.3133	0.3310
VIII	0.4479	0.3865	0.2049	0.2038	0.2117	0.3298	0.2758
CV^2-based weights $W(CV^2)$	0.1514	0.0916	0.0859	0.4753	0.0974	0.0255	0.0728
Ideal positive solutions V^+	0.1680	0.1405	0.2049	0.1223	0.2117	0.3133	0.2022
Ideal negative solutions V^-	0.4759	0.4568	0.4713	0.6113	0.5210	0.4617	0.4597

Table 5.7 Computations of distance measures d_i^+ and d_i^- with entropy weights, composite indices and ranks

Location Sl No.	d_i^+	d_i^-	CI	Rank
I	0.07074	0.10968	0.3921	2
II	0.07344	0.09693	0.4311	3
III	0.08700	0.08265	0.5128	5
IV	0.09921	0.07191	0.5798	7
V	0.08362	0.06663	0.5565	6
VI	0.06274	0.10508	0.3738	1
VII	0.10537	0.05321	0.6645	8
VIII	0.08356	0.08388	0.4990	4

$$d_i^+ = \left[\sum_j \frac{(x_{ij} - min_j)^2 W_j}{\sum_t x_{tj}^2} \right]^{1/2};$$

$$d_i^- = \left[\sum_j \frac{(x_{ij} - max_j)^2 W_j}{\sum_t x_{tj}^2} \right]^{1/2}.$$

In the above, min_j refers to least value of x_{ij}'s for every fixed j across the jth column.

Another distance measure is also used at times. That is called 'Absolute Distance' or, L_1-norm. Using L_1-norm amounts to defining

$$d_i^+ = \sum_j \frac{(x_{ij} - min_j)W_j}{\sum_t x_{tj}};$$

$$d_i^- = \sum_j \frac{(max_j - x_{ij})W_j}{\sum_t x_{tj}}.$$

and computing the composite indices as $CI_i = d_i^+/[d_i^+ + d_i^-]; i = 1, 2, \ldots, m$.

We will not get into the computational details.

References and Suggested Readings

Cascales, G., & Lamata, T. (2012). On rank reversal and TOPSIS method. *Mathematical and Computer Modelling*, *56*, 123–132.

Filar, J. A., Ross, N. P., & Wu, M.-L. (1997). US EPA Report on Environmental Assessment Based on Multiple Indicators.

Hwang, L. C., & Yoon, K. (1981). *Multiple attribute decision making methods and applications*. New York: Springer.

Pakpour, S., Snizhana, O., Prasher, Shiv, Milani, A., & Chnier, M. (2013). DNA extraction method selection for agricultural soil using TOPSIS multiple criteria decision-making model. *American Journal of Molecular Biology*, *3*, 215–228.

Ross, N. P., & Sinha, B. K. (2001). On Some Aspects of Data Integration Techniques with Applications. Unpublished Manuscript.

Shah, K. R., & Sinha, Bikas K. (2002). On some aspects of data integration techniques with environmental application. *Journal of Environmetrics*, *14*, 409–416.

Yoon, K. P., & Hwang, C. (1995). Multiple attribute decision making: An introduction. Sage university paper series on quantitative applications in the social sciences (pp. 07–104). Thousands Oaks, CA: Sage.

Zeleny, M. (1982). *Multiple criteria decision making*. New York: McGraw Hill.

Zou, Z. H., Yan, Y., & Sun, J. N. (2006). Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. *Journal of Environmental Sciences-China*, *18*, 1020–1023.

Chapter 6
Statistical Assessment of Agreement

6.1 General Introduction to Agreement

Researchers have become increasingly aware of the problem of assessing agreement since more than one and a half century in the past. There are numerous examples that illustrate these situations, and here we list some of them. In clinical and medical measurement comparison of a newly developed measurement method with an established one, it is often desired to check whether they agree sufficiently and accurately enough for the new to replace the old. The new method of measurement is most often cheaper, quicker, and suboptimal; however, it needs a thorough and careful examination to see if it can effectively replace the old one. In criminal trials, a group of jurors are used and sentencing depends on the complete agreement among the jurors. Hotels receive five-star recognition only after several experts and designated visitors agree on the services and facilities rendered by the hotels. The medals and ranking in sport games are based on the ratings provided by several judges.

It has now become generally accepted that measurements of agreement are needed to assess the acceptability of new or generic process, methodology, and formulation in both science and non-science fields of laboratory performance, instrument or assay validation, method comparisons, statistical process control, goodness of fit, and individual bioequivalence. Examples include the agreement of laboratory measurements collected through various laboratory instruments, the agreement of a newly developed method with gold standard method, the agreement of manufacturing process measurements with specifications, the agreement of observed values with predicted values, and the agreement in bioavailability of a new or generic formulation with a commonly used formulation. By the way, measuring agreement has been used very often to designate the level of agreement between different data-generating sources, commonly referred to as observers or raters. A rater could be a chemist, a psychologist, a radiologist, a clinician, a nurse, a rating system, a diagnosis, a treatment, an instrument, a method, a process, a technique or a formula, to mention a few. Elementary to advanced statistical methods have been used over time to assess the level of

This chapter draws material from co-published work of one of the authors: 'Some further aspects of assessment of agreement involving bivariate normal responses,' published in Int'l Jour. of Statistical Sciences, Vol. 13, 2013, pp. 1–19. Portions have been used here with permission from the Author Dr. Ganesh Dutta as well as Publisher.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_6

agreement between different data-generating sources referred to above as observers or raters.

Cohen's Kappa statistic (1960) and weighted Kappa (1968) are the most popular indices for measuring agreement when the responses are nominal. Weighted Kappa statistic has been proposed by Landis and Koch (1977), and it is appropriate for assessing agreement when the categories of response are ordinal. Several authors have proposed guidelines for the interpretation of kappa statistic. Vide, for example, Landis and Koch (1977), Fleiss (1981), Bland and Altman (1986), and Kraemer et al. (2002). A comprehensive review paper is also worth reporting (Banerjee et al. 1999). Recently, some studies have been undertaken to critically examine certain aspects of Cohen's Kappa. These relate to its attaining the negatively extreme value and its standardization. See Pornpis et al. (2006).

Extensions have also been made to allow for more than two raters, more than two possible ratings, ordinal data and continuous data. In addition, many other applications of kappa statistic in a variety of different contexts can be found in the literature. A reference book in this area is by Eye and Mun (2005). Another book dealing with both categorical and continuous measurements for multiple raters and multiple ratings is by Shoukri (2004).

Lin (1989) introduced the concordance correlation coefficient (CCC) for measuring agreement which is more appropriate when the data are measured on a continuous scale. A weighted CCC was proposed by Chinchilli et al. (1996) for repeated measurement designs and a generalized CCC for continuous and categorical data was introduced by King and Chinchilli (2001). Lin (2000) also introduced total deviation index (TDI) for measuring individual agreement with applications in laboratory performance and bioequivalence. Further to this, Lin et al. (2002) proposed methods for checking the agreement in terms of coverage probability(CP) when the two measurements are quantitative in nature.

When the study of agreement involves three or more raters on a continuous scale, there are different approaches to follow. Two most recent references are (i) Barnhart et al. (2007) and (ii) Lin et al. The authors broadly follow (i) ANOVA and (ii) modeling approach to examine the extent of agreement. The approach proposed and studied in Lin et al. (2002) has been extended in Hedayat et al. (2009) for the case of multiple raters.

We will touch upon some of the techniques developed for study of agreement involving both types of data.

6.2 Cohen's Kappa Coefficient and Its Generalizations: An Exemplary Use

There are many instances of applications of the basic technique for assessing agreement between two raters, in case the subjects are rated according to a binary feature, to be designated as Yes and No. Cohen's Kappa (1960) was suggested in the agreement literature with this specific purpose. Generalizations and extensions to other

contexts were brought in from time to time. We will discuss at length one study carried out in a hospital in Bangkok [Pornpis et al. (2006)].

Rajavithi Hospital, Bangkok, Thailand houses Thai Screening for Diabetic Retinopathy Study Group in its Department of Ophthalmology. Three MD doctors Dr. Paisan Ruamviboonsuk, Dr. Khemawan Teerasuwanajak, and Dr. Kanokwan Yuttitham carried out a revealing diagnostic study in this specialist eye hospital having [in-house and confined to hospital beds] 600+ diabetic patients. All the patients were under treatment for diabetic retinopathy of different degrees of severity. The study was based on randomly selected 400/600+ diabetic patients and from each selected patient, one good single-field digital fundus image was taken with signed consent and with due approval by Ethical Committee on Research with Human Subjects.

The purpose was to extract information from each image on three major features:

(i) Diabetic Retinopathy Severity [6 options]:

No Retinopathy/Mild/Moderate NPDR/Severe NPDR/PDR/Ungradable;

(ii) Macular Edema [2 options]: Presence/Absence/Ungradable;

(iii) Referral to Ophthalmologists [2 options]: Referrals / Non-Referrals / Uncertain.

These features were extracted by (i) Retina Specialists [3], (ii) General Ophthalmologists [3], (iii) Photographers [3] and (iv) Nurses [3]—all engaged in their respective meaningful professions within the hospital. It thus transpires that altogether 12 raters collected data on each of the 3 features mentioned above and from each of the 400 images so collected. Therefore, the study group was loaded with massive amount of data.

The objective of the research study was to examine the extent of agreement within and between different Expert Groups and to provide adequate interpretation of the results. It is believed that all the 12 experts/raters examined the images independently of one another.

As noted from the above, items (ii) and (iii) deal mostly with binary response [Presence versus Absence or Referral versus Non-Referral] data while item (i) deals with multi-response categorical data. We will slightly modify item response for (ii) to give it a shape of binary response data. It is revealed that the first two Retina Specialists RS1 and RS2 independently counted the respective Presence–Absence responses [in respect of the Feature: Macular Edema] as: 337 versus 40 and 344 versus 33. This indeed showed remarkable agreement among them upfront [89 versus 11 percent and 91–9 percent]! It was too good to be acceptable. The study group wondered about the validity of the findings and contacted Dr Montip Tiensuwan, Statistics Faculty, Department of Mathematics, Mahidol University, Bangkok. Dr Tiensuwan had already studied the literature on Statistical Assessment of Agreement and worked with one of the authors of this article [Sinha]. Her collaboration with the Hospital Study Group was successful, and it eventually resulted in a good journal publication. We will now elaborate on the major findings of their study.

It is clear that each image was inspected by each of the three RSs, and hence, it is possible to examine the scope of agreement more closely before deciding on its extent. As is stated above, RS1 and RS2 largely agreed on classification of patients into Presence–Absence Categories w.r.t. Macular Edema. But this only reflected what

is called marginal nature of binary classification. We are also in a position to check case by case the nature of agreement or otherwise of RS1 and RS2. For example, when pairwise ratings given by $RS1$ and $RS2$ are considered for each of the 377 patients, we find that

$$[(Y, Y) : 326/377; (Y, N) : 11/377; (N, Y) : 18/400; (N, N) : 22/377]$$

- the 'marginal' totals being $[RS1(Y) : 337/377; RS2(Y) : 344/377]$, as was specified above. It transpires that there are altogether $29/377 = 89$ percent cases of disagreement between the two raters. In effect, therefore, $RS1$ and $RS2$ are in very good agreement. And this Cohen designated as observed agreement, denoted by θ_0. According to Cohen, this is only half of the story and it could as well be due to what he assigned as chancy agreement! The idea is that two so-called experts could purely agree by chance—by making assessments independently. Using elementary probability formula, he computed the contribution from chancy agreement as:

$$\theta_e = P[Y, Y] + P[N, N] = P[Y, .]P[., Y] + P[N, .]P[., N]$$

by referring to the 'marginal probabilities.' According to this formula, for the above data set, chancy agreement, denoted by θ_e is computed as 82.50 percent! Cohen then suggested 'chance-corrected' agreement index as

$$\kappa = \frac{\theta_0 - \theta_e}{1 - \theta_e}.$$

Computation yields $\kappa = 56$ percent which suggests a moderate level of agreement only. Likewise, it is a routine task to compute κ coefficient between $RS1$ and $RS3$ or, between $RS2$ and $RS3$. It may be noted that the κ coefficients do not obey any transitivity law.

This study became instantly famous because of the following special feature. For any group of 3 Experts [Retina Specialists/General Ophthalmologists/etc/etc], the purview of the study also captured Consensus Rating [CR] of the raters for each feature. Thus, for example, in respect of Macular Edema, there was a Consensus Rating given collectively by the 3 RSs as follows: [Presence: 355/400; Absence: 35/400; Ungradable: 10/400]. Subsequently, κ coefficient was computed for the RSs as against the CR[RS] one by one.

Also for that matter, we can compute κ values in respect of the feature (iii), by restricting to the 2×2 case of binary response, neglecting the uncertain category. We will skip the details.

So far as the feature in item (i) is concerned, we need to be careful in assessing the extent of agreement between any two raters [or between a rater of a category and the CR of the same category]. This is because we are now dealing with six categories of response in respect of the status of Diabetic Retinopathy [DR] as mentioned in (i). Cohen's original idea of computation of κ, based on θ_0 and θ_e, does not pose any difficulty anyway. First of all, we can visualize the response count data for a pair of experts as forming a table of order 6×6 with the percentage counts along the main

diagonal [say, f_{ii}/n for the ith category of response] serving as constituents of θ_0. By the same token, products of percentage counts [based on the notion of independence] such as $(f_{i,.}/n)(f_{.,i}/n)$ will add up to the computation for θ_e. Then the formula for κ can be routinely applied. This was done and sooner or later, it drew criticism! We will take up the data set for $RS1$ versus $RS2$ and examine the matter below.

We will follow the codes: Code I - No Retinopathy; Code II - Mild; Code III - Moderate NDPR; Code IV Severe NPDR; Code V PDR and Code VI: Ungradable.

Along the main diagonal, the percentage of observed agreement θ_0 amounts to 80.50 percent. Further, direct computation yields for $\theta_e = 48.60$ percent. Hence, $\kappa = 62$ percent a very moderate level of agreement. The criticism has been based on the following arguments: Pairwise categories

$$[(Code\ I, Code\ II), (Code\ II, Code\ I), (Code\ II, Code\ III), (Code\ III, Code\ II) etcetc]$$

represent what may be termed as 'narrowly missed' cases. Cohen's κ does not take cognizance of these narrowly missed cases/classes and attributes no credit whatsoever to the raters. It is argued that one should make a case of allowing for partial credits to be attributed to such and similar categories. In contrast to Cohen's original κ—termed henceforth as Unweighted κ—weights were assigned to all the categories and κ was modified to Weighted κ, written as $\kappa(W)$. It is computed along similar lines as

$$\kappa(W) = (\theta(W)_0 - \theta(W)_e)/(1 - \theta(W)_e)$$

where $f_{ij}W_{ij}$s are used in the computation of $\theta(W)_0$ and $f_{i.}f_{.j}W_{ij}$s are used in the computation of $\theta(W)_e$. The choice of the weight matrix $\mathbf{W} = ((\mathbf{W_{ij}}))$ has not been any smooth matter. Reasonable and acceptable choice of the weight matrix of dimension R have the elements $W_{ij} = 1 - (i - j)^2/(R - 1)^2$.

Weighted κ statistics were calculated for pairs of raters, including comparison against the CR in respect of all the three features listed in (i), (ii), and (iii). The results are shown in the Appendix.

This study also covered another important aspect of comparison of expertise across different specialist groups. In the published literature, there are formulae available to account for this kind of comparison. Applied to this case, a measure of composite performance of 3 Retina Specialists/3 Ophthalmologists/3 Technicians/3 Nurses for each of the 3 features was computed. For example, for DR, it was revealed that composite performance indices are

$$RS - 0.58;\ Oph. - 0.36;\ Tech. - 0.37,\ Nurses - 0.26.$$

Likewise, for Macular Edema, the values are: [0.58, 0.19, 0.38, 0.20] and for Referral, these are: [0.63, 0.24.0.30, 0.20].

It transpired that except for the Retina Specialists, no other categories of so-called experts showed any visible mode of agreement in any of the features. Of all 400 cases, 44 warranted Referral to Ophthalmologists due to Retinopathy Severity and 5 warranted Referral to Ophthalmologists due to uncertainty in diagnosis. Fourth Retina Specialist carried out dilated fundus examination of these 44 patients, and

substantial agreement [$\kappa = 0.68$] was noticed for DR severity examination confirmed Referral of 38/44 cases.

In conclusion, it is stated that Retina Specialists are all in active clinical practice and hence are most reliable for digital image interpretation of images. Individual Raters' background and experience play roles in digital image interpretation expertise. Unusually, high percentage of images were declared as ungradable by nonphysician raters, though only 5 out of 400 were declared as ungradable by consensus of the Retina Specialists Group. Lack of confidence of non-physicians, rather than true image ambiguity, is likely to be a realistic reason for this. For this study, other factors [blood pressure, blood sugar, cholesterol, etc.] had not been taken into account.

6.3 Assessment of Agreement in Case of Quantitative Responses

In this section, we focus on the feature of agreement involving data for two competing raters measured on a continuous scale. There are several usual approaches for evaluating agreement for such paired data such as Pearson correlation coefficient, regression analysis, paired t-tests, least-squares analysis for slope and intercept, within subject coefficient of variation, and intra-class correlation coefficient.

The concordance correlation coefficient (CCC) was first proposed by Lin (1989) for assessment of agreement in continuous data. It represents a breakthrough in assessing agreement between two raters for continuous data in that it appears to avoid all the shortcomings associated with usual approaches in some situations. In short, Lin (1989) expresses the degree of concordance between two variables X and Y by the Mean Squared Deviation (MSD), $E(X-Y)^2$ and defines the CCC as

$$\rho_c = 1 - \frac{\mathrm{E(Y-X)^2}}{\mathrm{E_{Indep}(Y-X)^2}} = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{6.3.1}$$

where $\mathrm{E_{Indep}}(.)$ represents expectation under the assumption of independence of X and Y, $\mu_x = \mathrm{E(X)}$, $\mu_y = \mathrm{E(Y)}$, $\sigma_x^2 = \mathrm{Var(X)}$, $\sigma_y^2 = \mathrm{Var(Y)}$, and $\sigma_{xy} = \mathrm{Cov(X,Y)} = \rho\sigma_x\sigma_y$.

It is readily seen that ρ_c can be expressed as

$$\rho_c = \rho \times \frac{2\sigma_1\sigma_2}{(\mu_x - \mu_y)^2 + (\sigma_x^2 + \sigma_y^2)}$$

Further to this, it follows that

$$\rho_c = 1 \; if \; and \; only \; if \; [\rho = 1, \mu_x = \mu_y; \sigma_x = \sigma_y].$$

Lin (1989) estimates this CCC [ρ_c] with data by substituting the sample moments of bivariate sample into above formula to compute the sample counterpart of CCC (ρ_c). The CCC translates the MSD into a correlation coefficient that measures the

agreement along the identity line. It has the properties of a correlation coefficient in that it ranges between -1 and $+1$, with -1 indicating perfect reversed agreement ($Y = -X$), 0 indicating no agreement, and $+1$ indicating perfect agreement ($Y = X$). Lin et al. (2002) gave a review and comparison of various measures, including the CCC, of developments in this field by comparing the powers of the tests:

(1) $\mu_x = \mu_y$, (2) $\sigma_x = \sigma_y$, and (3) $\rho = \rho_0$, where ρ_0 is a given value, assumed to be substantially high.

Their calculation is illustrated using a real data example. This work was further extended in Hedayat et al. (2009) involving multiple raters. In another direction, Yimprayoon et al. (2006) extended the work of Lin et al. (2002) by combining the problems of testing for $\mu_x = \mu_y$, $\sigma_x = \sigma_y$, and $\rho \geq \rho_0$ into one overall testing problem under bivariate normal setup and then they presented the result based on simulation study.

An intuitively clear measurement of agreement is a measure that captures a large proportion of data within a predetermined boundary from the line of agreement, i.e., $X = Y$. In other words, we want the probability of the absolute value of $D = Y - X$ less than the specified boundary, k, to be large. This probability is termed in the literature as coverage probability (CP) (cf. (Lin et al. 2002)), and it is defined as

$$\mathrm{CP}(k) = \mathrm{P}[|D| < k], \tag{6.3.2}$$

where X and Y denote random variables representing paired observations for assessing the agreement. It is generally assumed that X and Y have a bivariate normal distribution with means μ_x and μ_y, variances σ_x^2 and σ_y^2 and correlation coefficient ρ so that the covariance of X and Y is $\sigma_{xy} = \rho\sigma_x\sigma_y$.

The multiparameter hypothesis involving (6.3.1), (6.3.2), and (6.3.3) displayed above is too demanding for agreement between the two raters. Therefore, a more appropriate and plausible null hypothesis can be formulated as

$$\mathrm{H}_0 : \ |\mu_x - \mu_y| \geq \varepsilon_0, \ \frac{\sigma_x}{\sigma_y} \ or \ \frac{\sigma_y}{\sigma_x} \geq \eta_0, \ \rho \leq \rho_0 \tag{6.3.3}$$

where ε_0 is close to zero and η_0 and ρ_0 are close to unity—all are assumed to be specified. A large sample test [known as Likelihood Ratio Test] of this hypothesis has been worked out in Dutta and Sinha (2013).

References and Suggested Readings

Anderson, S., & Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, *18*, 259–273.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. New York: Wiley.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27*, 3–23.

Barnhar, H. X., Haber, M. J., & Lin, L. I. (2007). *An Overview On Assessing Agreement With Continuous Measurement*.

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *8*, 307–310.

Chinchilli, V. M., Martel, J. K., Kumanyika, S., & Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measures designs. *Biometrics, 52*, 341–353.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Dutta, G., & Sinha, B. K. (2013). Some further aspects of assessment of agreement involving bivariate normal responses. *International Journal of Statistical Sciences*, *13*.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. 2nd ed. (pp. 38–46). New York: John Wiley.

Hedayat, A. S., Lou, C., & Sinha, B. K. (2009). A statistical approach to assessment of agreement involving multiple raters. *Communications in Statistics - Theory & Methods*, *38*, 2899–2922.

Holder, D. J., & Hsuan, F. (1993). Moment-based criteria for determining bioequivalence. *Biometrika*, *80*, 835–846.

King, T. S., & Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*. pp. 2131–2147.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine* (pp. 2109–2129).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–74.

Lin, L., Hedayat, A. S., Sinha, B. K., & Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association*, *97*, 257–270.

Lin, L., Hedayat, A. S., & Wu, W. (2012). *Statistical tools for measuring agreement*. Berlin: Springer.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268.

Lin, L. I. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, *48*, 599–604.

Lin, L. I. (1997). Rejoinder to the letter to the editor by Atkinson and Nevill. *Biometrics*, *53*, 777–778.

Lin, L. I. (2000). Total deviation index for measuring individual agreement: With application in lab performance and bioequivalence. *Statistics in Medicine*, *19*, 255–270.

Lin, L. I., & Torbeck, L. D. (1998). Coefficient of accuracy and concordance correlation coefficient: New statistics for method comparison. *PDA Journal of Pharmaceutical Science and Technology*, *52*, 55–59.

Pornpis, Y., Tiensuwan, M., & Sinha, B. K. (2006). Cohen's kappa statistic: A critical appraisal and some modifications. *Calcutta Statistical Association Bulletin, 58*, 151–169.

Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics*, *51*, 615–626.

Schall, R., & Luus, H. G. (1993). On population and individual bioequivalence. *Statistics in Medicine*, *12*, 1109–1124.

Schall, R., & Williams, R. L. (1996). Towards a practical strategy for assessing individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, *24*, 133–149.

Sheiner, L. B. (1992). Bioequivalence revisited. *Statistics in Medicine*, *11*, 1777–1788.

Shoukri, M. M. (2004). *Measures of interobserver agreement*. Boca Raton: Chapman & Hall/CRC.

Von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*.

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. New York: Marcel Dekker.

Vonesh, E. F., Chinchilli, V. M., & Pu, K. (1996). Goodness-of-fit in generalised nonlinear mixed-effect models. *Biometrics*, *52*, 572–587.

Yimprayoon, P., Tiensuwan, M., & Sinha Bimal, K. (2006). Some statistical aspects of assessing agreement: Theory and applications (English summary). *Festschrift for Tarmo Pukkila on his 60th birthday* (pp. 327–346). Tampere: Department of Mathematics, Statistics and Philosophy, University of Tampere.

Chapter 7
Meta-Analysis

7.1 Introduction

The common mean estimation problem was first introduced by Cochran (1937), while he was considering combining a series of similar experiments. The general setting for this kind of problem is: Suppose we have k independent groups of normal variables with sample size n_i, for the ith group, having the sample mean $\bar{y}_i$ $N(\mu, \sigma_i^2/n_i)$ where $i = 1, 2, \ldots, k$.

The setup presupposes that there is a common unknown mean μ for the k populations, and the problem considered is that of efficient unbiased estimation of μ based on the data from the k groups.

For $k = 2$, Cochran (1937) suggested the unbiased estimator

$$\hat{\mu} = \frac{[\bar{y}_1 n_1/\sigma_1^2 + \bar{y}_2 n_2/\sigma_2^2]}{[n_1/\sigma_1^2 + n_2/\sigma_2^2]}.$$

This estimator is the best linear unbiased estimator [BLUE] of the common mean μ, assuming that the two population variances σ_1^2 and σ_2^2 are both known.

In case both the variances are unknown, a natural way out would be to replace them by their sample counterparts, i.e., their unbiased estimates $\hat{\sigma}_i^2 = s_i^2 = [\sum(y_{ij} - \bar{y}_i)^2/(n_i - 1)]$; $i = 1, 2$. Graybill and Deal (1959) were prompted by this motivation, and they introduced

$$\hat{\mu} = \frac{\sum_i \bar{y}_i n_i/\sigma_i^2}{\sum_i n_i/\sigma_i^2}.$$

and this is referred to in the literature as Graybill–Deal estimator $\hat{\mu}_{G-D}$. The samples are assumed to have been drawn from normal populations. As a consequence, sample means and sample variances are distributionally independent. That justifies, by E_1E_2 argument, that the Graybill–Deal estimator is unbiased for the mean μ. The properties of such estimators have been widely studied in the literature. In particular,

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_7

we would like to mention the works of Meier (1953), Cochran and Carroll (1953), Zacks (1966), Rao and Subrahmaniam (1971), Khatri and Shah (1974), Sinha (1985), and Krishnamoorthy and Moore (2002).

Below we will discuss some results in this fascinating area of meta-analysis and data analysis.

7.2 Estimation of Common Bernoulli Parameter "p"

This is the simplest application of meta-analysis. There are a number of independent studies toward estimation of a Bernoulli parameter p. From the ith study, we come across a total of n_i counts out of which f_i correspond to "success" counts; $i = 1, 2, \ldots, k$. We know that $\hat{p} = f_i/n_i; i = 1, 2, \ldots, k$. Using the additive property of binomial distribution, it follows that $\sum_i f_i = f\ Bin(n, p)$ where $n = \sum_i n_i$. As a result, $\hat{p} = f/n$. The same result also follows from an application of Graybill–Deal formula. From the ith source, we have

$$\hat{p}_i = f_i/n_i; V(\hat{p}_i) = p(1-p)/n_i; i = 1, 2, \ldots, k.$$

By combining these estimates according to Cochran/Graybill–Deal formula, we deduce the above result. For a single parameter p in a Bernoulli setup, the result on a combination of evidences from a number of independent sources is thus quite obvious. It is also known that for $f\ Bin(n, p)$, an unbiased estimator for $p(1-p)$ is given by $f(n-f)/n(n-1)$.

7.3 Estimation of Common Mean of Several Normal Populations

We start with k univariate normal populations with a common unknown mean μ and possibly different and unknown population variances $\sigma_i^2; i = 1, 2, \ldots, k$. We have available random samples $(y_{ij}; j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k)$ from the populations where the sample sizes $n_i; i = 1, 2, \ldots, k$ are known in advance with $n_i \geq 2$ for each i.

$$Case\ 1:\ \sigma_i^2 s\ are\ known\ beforehand$$

It follows that the sample means form a set of jointly sufficient statistics for the population mean μ and, further, that the joint distribution of the sufficient statistics is not complete. It is well known that the weighted mean

$$\bar{y} = \frac{\sum_i \bar{y}_i n_i/\sigma_i^2}{\sum_i n_i/\sigma_i^2}$$

serves as the BLUE for μ. However, this case is far from reality wherein we have no firsthand idea about the population variances—not even possibly about their relative values. This case is discussed below.

$$Case\ 2: \ \sigma_i^2 s\ are\ unknown\ beforehand$$

It follows that the sample means

$$\bar{y}_i; i = 1, 2, \ldots, k$$

and sample variances with appropriate divisors like

$$s_i^2 = \sum_j (y_{ij} - \bar{y}_i)^2/(n_i - 1); i = 1, 2, \ldots, k$$

are jointly sufficient statistics for the population mean μ and the population variances. Further, the joint distribution of the sufficient statistics is not complete. It is in this context that Graybill and Deal (1959) suggested the following estimator for the common mean μ:

$$\hat{\mu}_{G-D} = \bar{\bar{y}} = \frac{\sum_i \bar{y}_i n_i/s_i^2}{\sum_i n_i/s_i^2}.$$

The above provides a computational formula for the Graybill–Deal estimate $\hat{\mu}_{G-D}$ of the common mean μ.

We recall that in normal samples the sample means and sample variances are independently distributed. Therefore, we can work out

$$E[\hat{\mu}_{G-D}] = E_1 E_2[\bar{\bar{y}}] = E_1[\mu] = \mu.$$

In the above, we have used the fact that $E[\bar{y}_i;\ \text{given}\ s_i^2] = \mu$ for each population. Further,

$$Var[\hat{\mu}_{G-D}] = [V_1 E_2 + E_1 V_2](\hat{\mu}_{G-D}) = E_1 \left[\frac{\sum_i \sigma_i^2 n_i/s_i^4}{\left(\sum_i n_i/s_i^2\right)^2} \right].$$

since $V_1 E_2 = V_1[\mu] = 0$.

Remark 7.1 An exact analytical expression for the above quantity is hard to derive. In an infinite series form, it has been provided by Khatri and Shah (1974). Also, a first-order approximation has been provided by Meier (1953) and this is reproduced below.

$$\frac{1}{\sum n_i/\sigma_i^2}\left[1 + 2\sum c_i(1 - c_i)/(n_i - 1) + \cdots\right]; c_i = (n_i/\sigma_i^2)\left(\sum_t n_t \sigma_t^2\right)^{-1}, i = 1, 2, \ldots, k.$$

Following Meier (1953), an approximation for estimated variance of $\hat{mu}_{G-D}$ is given by

$$\frac{1}{\sum_i n_i/s_i^2}\left[1+\sum_i(4/(n_i-1))\left[\frac{n_i/s_i^2}{\sum_t n_t/\sum s_t^2}-\frac{n_i^2/s_i^4}{\left(\sum_t n_t\sum_t^2\right)^2}+\cdots\right]\right]$$

Toward unbiased estimation of the above variance, there has been an attempt by Sinha (1985) and a first-order approximation of the estimated variance is given by

$$\frac{1}{\sum_i n_i/s_i^2}\left[1+\sum_i(4/(n_i+1))\left[\frac{n_i/s_i^2}{\sum_t n_t/\sum s_t^2}-\frac{n_i^2/s_i^4}{\left(\sum_t n_t\sum_t^2\right)^2}+\cdots\right]\right]$$

There are two other variance estimators available in the published literature. These are:

$$1/\left[\sum_i n_i/s_i^2\right];$$

$$\frac{1}{k-1}\left[\sum_i\frac{(n_i/s_i^2)}{\sum_t n_t/s_t^2}(\bar{y}_i-\hat{\mu}_{G-D})^2\right].$$

Remark 7.2 Apart from the exercise on computation of estimate of the common mean and its estimated variance, research has gone in a different direction. In a theoretical framework, it is well known that the "k-population means-based common mean estimator" as weighted mean of individual estimators [which are respective sample means] is better than any "$(k-1)$-subpopulation means-based common mean estimator" as weighted mean of individual estimators for the subset of $(k-1$-populations and so on. This is a very general result whose proof is almost trivial. However, in case of Graybill–Deal estimator, variance comparison for two estimators—one based on k-populations and the other based on a subset of $(k-1)$ populations—is very much a non-trivial exercise. In fine, this comparison depends very much on the individual sample sizes. In short, we need a minimum sample size from each population so that a sense of "improvement" based on an increasing number of competing populations can be established. We will not discuss this matter any further.

7.4 Meta-Analysis in Regression Models

This time we discuss the problem of unbiased estimation of the common parameter θ involved in the linear regression models of the means of two independent normal populations with unequal and unknown variances. We work out the popular Graybill–Deal estimator for the common parameter.

We start with two simple linear regression models:

$$y_{1j} = \alpha_1 + \beta_1 x_{1j} + e_{1j};\ j = 1, 2, \ldots, n_1;\ E(e_{1j}) = 0;\ E(e_{1j}^2) = \sigma_1^2;\ E(e_{1j}e_{1t}) = 0,\ j \neq t.$$

$$y_{2j} = \alpha_2 + \beta_2 x_{2j} + e_{2j};\ j = 1, 2, \ldots, n_2;\ E(e_{2j}) = 0;\ E(e_{2j}^2) = \sigma_2^2;\ E(e_{2j}e_{2t}) = 0,\ j \neq t.$$

As usual,

$$\hat{\alpha}_1 = \bar{y}_1 - \hat{\beta}_1 \bar{x}_1;\ \hat{\alpha}_2 = \bar{y}_2 - \hat{\beta}_2 \bar{x}_2;$$

$$\hat{\beta}_1 = \frac{\sum_j (y_{1j} - \bar{y_1})(x_{1j} - \bar{x}_1)}{\sum_j (x_1 j - \bar{x}_1)^2} = \frac{SPXY_{11}}{SSX_1};$$

$$\hat{\beta}_2 = \frac{\sum_j (y_{2j} - \bar{y_2})(x_{2j} - \bar{x}_2)}{\sum_j (x_2 j - \bar{x}_2)^2} = \frac{SPXY_{22}}{SSX_2}.$$

$$V(\hat{\alpha}_i) = \sigma_i^2 [1/n_i + \bar{x}_i^2 / SSX_i];\ i = 1, 2.$$

$$\hat{\sigma}_i^2 = \sum_t [(y_{it} - \bar{y}_i) - \hat{\beta}_i (x_{it} - \bar{x}_i)]^2 / (n_i - 2);\ i = 1, 2.$$

Now suppose we are in a situation which leads to the validity of the assumption of equality of the two intercept parameters, that is, $\alpha_1 = \alpha_2 = \alpha$, say. Then, we have two estimates for α, and hence, we can combine the two to form a Graybill–Deal-type estimate. It is defined as $\hat{\alpha}$ where

$$\hat{\alpha}\left[\frac{1}{\hat{\sigma}_1^2[1/n_1 + \bar{x}_1^2/SSX_1]} + \frac{1}{\hat{\sigma}_2^2[1/n_2 + \bar{x}_2^2/SSX_2]}\right] = \frac{\hat{\alpha}_1}{\hat{\sigma}_1^2[1/n_1 + \bar{x}_1^2/SSX_1]} + \frac{\hat{\alpha}_2}{\hat{\sigma}_2^2[1/n_2 + \bar{x}_2^2/SSX_2]}.$$

Under the above linear regression framework, if it so transpires that the mean regression lines are identical, i.e., $\alpha_1 = \alpha_2$; $\beta_1 = \beta_2$, then the common intercept parameter α and the common regression parameter β are estimated by matrix version of the Graybill–Deal-type estimator which is described below.

We denote by θ the two-dimensional vector parameter α, β. Then

$$\hat{\theta} = [(\hat{\sigma}_1^2 W_1)^{-1} + (\hat{\sigma}_2^2 W_2)^{-1}]^{-1} [(\hat{\sigma}_1^2 W_1)^{-1} \hat{\theta}_1 + (\hat{\sigma}_2^2 W_2)^{-1} \hat{\theta}_2].$$

It follows that $\hat{\theta}$ is unbiased for θ. The proof is based on $E_1 E_2$ argument and the fact that estimates of the error variances are independent of the estimates of the model parameters in a regression context.

References and Suggested Readings

Chiou, W.-J., & Cohen, A. (1984). On estimating a common multivariate normal mean vector. *Annals of the Institute of Statistical Mathematics*, *37*, 499–506.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to Journal of the Royal Statistical Society*, *4*, 102–118.

Cochran, W. G., & Carroll, S. P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics*, *9*, 447–459.

Graybill, F. A., & Deal, R. B. (1959). Combining unbiased estimators. *Biometrics*, *15*, 543–550.

Khatri, C. G., & Shah, K. R. (1974). Estimation of location parameters from two linear models under normality. *Communications in Statistics*, *3*, 647–663.

Krishnamoorthy, K. & Moore, B. C. (2002). Combining information for prediction in linear regression. *Metrika*, *56*, 73–81.

Loh, L. W. (1991). Estimating the common mean of two multivariate normal distributions. *The Annals of Statistics*, *19*, 297–313.

Meier, P. (1953). Variance of the weighted mean. *Biometrics*, *9*, 59–73.

Norwood, T. E., & Hinkelmann, K, Jr. (1977). Estimating the common mean of several normal populations. *The Annals of Statistics*, *5*, 1047–1050.

Rao, J. N. K., & Subrahmaniam, K. (1971). Combining independent estimators and estimation in linear regression with unequal variances. *Biometrics*, *27*, 971–990.

Shinozaki, N. (1978). A note on estimating the common mean of k normal distributions and the stein problem. *Communications in Statistics- Theory and Method*, *7*, 1421–1432.

Sinha, Bimal K. (1985). Unbiased estimation of the variance of the Graybill–Deal estimator of the common mean of several normal populations. *The Canadian Journal of Statistics*, *13*, 243–247.

Zacks, S. (1966). Unbiased estimation of the common mean. *Journal of the American Statistical Association*, *61*, 467–476.

Chapter 8
Cluster and Discriminant Analysis

8.1 Introduction

Under multivariate analysis, two very important techniques are clustering and classification. Under the problem of clustering, we try to find out the unknown number of homogeneous inherent groups in a data set as well as the structure of the groups. But under classification, the basic problem is discrimination of objects into some known groups. One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. Classification is fundamental to most branches of science.

Cluster analysis has a variety of objectives. It is focussed on segmenting a collection of items (also called observations, individuals, cases, or data rows) into subsets such that those within each cluster are more closely related to one another than objects assigned to different clusters. The main focus in cluster analysis is on the notion of degree of similarity (or dissimilarity) among the individual objects being clustered. The two major methods of clustering are hierarchical clustering and k-means clustering. Most of the clustering methods are exploratory in nature and do not need any model assumption.

Different statistical techniques are available for clustering and classification (Fraix-Burnet et al. 2015; De et al. 2013 and references there in). But depending on the nature of the different types of data, several problems often arise and in some cases a proper solution is still not available.

Sometimes the data set under consideration has a distributional form (usually normal), and sometimes it is of non-normal nature. Based on the above point, there is a justification needed about which clustering or classification technique should be used so that it reflects the proper nature of the data set provided. This problem is more relevant for classification as most of the classification methods are model

---

Sections of this chapter draw from one of the authors' published work, 'Statistical Methods for Astronomical Data Analysis,' authored by Asis Kumar Chattopadhyay and Tanuka Chattopadhyay, and published in 2014 by Springer Science+Business Media, New York.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_8

based. For clustering, most of the methods are nonparametric in nature and as such the above problem is not very serious. But here also basic assumption is that the nature of the variables under study is continuous, whereas under practical situations, these may be categorical like binary, nominal, ordinal, and even directional (particularly for environmental and astronomical data). Under such situations, standard similarity/dissimilarity measures will not work.

The clustering techniques which require an inherent model assumption are known as model-based methods, whereas the clustering technique where no modeling assumption or distributional form is needed may be termed as non-model-based methods. Hence based on the nature of data set, one has to decide about proper application of the two types of techniques.

At present, big data issues related to data size are quite common. In statistical terms, these problems may be tackled in terms of both the number of observations and the variables considered. Many standard clustering techniques fail to deal with such big data sets. Thus, some dimension reduction methods may be applied at first and then clustering may be performed on the reduced data set. Some data mining techniques are very helpful under such situations.

Finally and most importantly, after all these considerations, the similarity of grouping of objects obtained from different methods should be checked in terms of some physical properties.

8.2 Hierarchical Clustering Technique

There are two major methods of clustering, viz. hierarchical clustering and k-means clustering. In hierarchical clustering, the items are not partitioned into clusters in a single step. Instead, a series of partitions takes place, which runs from a single cluster containing all objects to n clusters each containing a single object. Hierarchical clustering is subdivided into agglomerative methods, which proceed by series of combinations of the n objects into groups, and divisive methods, which separate n objects successively into smaller groups. Agglomerative techniques are more commonly used. Hierarchical clustering may be represented by a two-dimensional diagram known as dendrogram which illustrates the additions or divisions made at each successive stage of analysis.

8.2.1 Agglomerative Methods

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, G_n; G_{n-1}; ::::::::; G_1. The first G_n consists of n single-object 'clusters,' and the last G_1 consists of single group containing all n cases. The structure of the groups is not unique and depends on several factors like choice of the dissimilarity/similarity measure, choice of the linkage measure.

At each particular stage, the method adds together the two clusters which are most similar. At the first stage, we join together two objects that are closest together, since at the initial stage each cluster has only one object. Differences between methods arise because of the different ways of defining dissimilarity or similarity between clusters.

Hierarchical clustering is largely dependent on the selection of such a measure. A simple measure is Manhattan distance, equal to the sum of absolute distances for each variable. The name comes from the fact that in a two-variable case, the variables can be plotted on a grid that can be compared to city streets, and the distance between two points is the number of blocks a person would walk.

The most popular measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. In the two-variable case, the distance is analogous to finding the length of the hypotenuse in a triangle. Besides Manhattan and Euclidian distances, there are other dissimilarity measures also based on the correlation coefficients between two observations on the basis of several variables.

Alternatively, one may use a similarity measure which is complementary in nature and under the normalized set up, it may be obtained by subtracting the dissimilarity measure from one.

8.2.2 Similarity for Any Type of Data

The above-mentioned dissimilarity/similarity measures are applicable to continuous-type data only. But generally, we work with mixed-type data sets those include different types like continuous, discrete, binary, nominal, ordinal. Gower (1971) has proposed a general measure as follows:

The Gower's Coefficient of Similarity:

Two individuals i and j may be compared on a character k and assigned a score $sijk$. There are many ways of calculating $sijk$, some of which are described below.

Corresponding to n individuals and p variables, Gower's similarity index S_{ij} is defined as

$$S_{ij} = \Sigma_{k=1}^{p} s_{ijk} / \Sigma_{k=1}^{p} \delta_{ijk} (i, j = 1, 2, \ldots n)$$

$$\begin{aligned} \text{where } \delta_{ijk} &= 1 \text{ when character } k \text{ can be compared} \\ &\quad \text{for observations } i \text{ and } j \\ &= 0 \text{ otherwise} \end{aligned}$$

For continuous (quantitative) variables with values $x_{1k}, x_{2k}, \ldots, x_{nk}$ for the kth variable

$$s_{ijk} = 1 - \mid x_{ik} - x_{jk} \mid / R_k$$

where R_k is the range of the variable k and may be the total range in population or the range in the sample.

For a categorical (qualitative) character with m categories ($m = 2$ for binary variable)

$$\begin{aligned} s_{ijk} &= 0 \text{ if } i \text{ and } j \text{ are totally different} \\ &= q \text{ (positive fraction) if there is some degree of agreement} \\ &= 1 \text{ when } i \text{ and } j \text{ are same} \end{aligned}$$

8.2.3 Linkage Measures

To calculate distance between two clusters, it is required to define two representative points from the two clusters (Chattopadhyay and Chattopadhyay 2014). Different methods have been proposed for this purpose. Some of them are listed below.[1]

Single linkage: One of the simplest methods is single linkage, also known as the nearest neighbor technique. The defining feature of the method is that distance between clusters is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each cluster are considered.

In the single linkage method, d_{rs} is computed as $d_{rs} = \text{Min } d_{ij}$, where object i is in cluster r and object j is in cluster s and d_{ij} is the distance between the objects I and j. Here the distance between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s. The minimum value of these distances is said to be the distance between clusters r and s. In other words, the distance between two clusters is given by the value of the shortest link between the clusters. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is minimum, are merged.

Complete linkage: The complete linkage, also called farthest neighbor, clustering method is the opposite of single linkage. Distance between clusters is now defined as the distance between the most distant pair of objects, one from each cluster. In the complete linkage method, $d - rs$ is computed as $d_{rs} = \text{Max } d_{ij}$, where object i is in cluster r and object j is cluster s. Here the distance between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s and the maximum value of these distances is said to be the distance between clusters r and s. In other words, the distance between two clusters is given by the value of the largest distance between the clusters. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is minimum, are merged.

Average linkage: Here the distance between two clusters is defined as the average of distances between all pairs of observations, where each pair is composed of one object from each group. In the average linkage method, d_{rs} is computed as

[1] A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer' is reproduced in this part.

$d_{rs} = Trs/(Nr \times Ns)$ where Trs is the sum of all pair-wise distances between cluster r and cluster s. Nr and Ns are the sizes of the clusters r and s, respectively. At each stage of hierarchical clustering, the clusters r and s, for which d_{rs} is the minimum, are merged.

Minimax Linkage: This was introduced by Bien and Tibshirani (2011). For any point x and cluster G, define

$$d_{\max}(x, G) = \max_{y \in G} d(x, y)$$

as the distance to the farthest point in G from x. Define the minimax radius of the cluster G as

$$r(G) = \min_{x \in G} d_{\max}(x, G)$$

that is, find the point $x \in G$ from which all points in G are as close as possible. This minimizing point is called the prototype for G. It may be noted that a closed ball of radius $r(G)$ centered at the prototype covers all of G. Finally, we define the minimax linkage between two clusters G and H as

$$d(G, H) = r(GUH)$$

that is, we measure the distance between clusters G and H by the minimax radius of the resulting merged cluster.

8.2.4 Optimum Number of Clusters

Usually, the number of clusters is determined from the dendrogram and validated by the physical properties. We specify a horizontal line for a particular similarity/dissimilarity value, and the clusters below this line are selected as optimum. But some mathematical rules (thumb rules) are also available which are based on between cluster and within cluster sum of squares values. If we denote by k, the number of clusters and define by W(k) the sum of the within cluster sum of squares for k clusters then the values of W(k) will gradually decrease with increase in k and that 'k' may be taken as optimum where W(k) stabilizes. For detailed discussion, one may follow the link http://www.cc.gatech.edu/~hpark/papers/cluster_JOGO.pdf.

8.2.5 Clustering of Variables

The hierarchical clustering method can also be used for clustering of variables on the basis of the observations. Here instead of the distance matrix, one may start with the correlation matrix (higher correlation indicating similarity of variables).

The linkage measures as listed in the previous section will not be applicable for variable clustering. In order to measure similarity/dissimilarity between two clusters of variables, one may either use the correlation between first principal components corresponding to the two clusters or the canonical correlations.

8.3 Partitioning Clustering-k-Means Method

The k-means algorithm (MacQueen 1967) assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. This method can be used for clustering of objects and not variables.

This method starts with a value of k. We will discuss later the method of selection of the value of k. Then we randomly generate k clusters and determine the cluster centers, or directly generate k seed points as cluster centers. Assign each point to the nearest cluster center in terms of Euclidian distance. Re-compute the new cluster centers. Repeat until some convergence criterion is met, i.e., there is no reassignment. The main advantages of this algorithm are its simplicity and speed which allows it to run on large data sets. Its disadvantage is that it is highly dependent on the initial choice of clusters. It does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. It maximizes inter-cluster variance and minimizes intra-cluster variance.

The advantages of partitioning method are as follows (Chattopadhyay and Chattopadhyay 2014):

(a) A partitioning method tries to select best clustering with k groups which is not the goal of hierarchical method.
(b) A hierarchical method can never repair what was done in previous steps.
(c) Partitioning methods are designed to group items rather than variables into a collection of k clusters.
(d) Since a matrix of distances (similarities) does not have to be determined and the basic data do not have to be stored during the computer run, partitioning methods can be applied to much larger data sets.

For k-means algorithms, the optimum value of k can be obtained in different ways.

On the basis of the method proposed by Sugar and James (2003), by using k-means algorithm first determine the structures of clusters for varying number of clusters taking $k = 2, 3, 4$, etc. For each such cluster formation, compute the values of a distance measure

$$dK = (1/p) \min_x E[(x_k - c_k)'(x_k - c_k)]$$

which is defined as the distance of the x_k vector (values of the parameters) from the center c_k (which is estimated as mean value), p is the order of the x_k vector.

Then the algorithm for determining the optimum number of clusters is as follows. Let us denote by d'_k the estimate of d_k at the kth point which is actually the sum of within cluster sum of squares over all k clusters. Then d'_k is the minimum achievable distortion associated with fitting k centers to the data. A natural way of choosing the number of clusters is plot d'_k versus k and look for the resulting distortion curve. This curve is always monotonic decreasing. Initially, one would expect much smaller drops, i.e., a leveling off for k greater than the true number of clusters because past this point adding more centers simply partitions within groups rather than between groups.

According to Sugar and James (2003) for a large number of items the distortion curve when transformed to an appropriate negative power, will exhibit a sharp "jump" (if we plot k versus transformed d'_k). Then calculate the jumps in the transformed distortion as

$$J_k = (d_k'^{-(p/2)} - d_{k-1}'^{-(p/2)})$$

Another way of choosing the number of clusters is plot J_k versus k and look for the resulting jump curve. The optimum number of clusters is the value of k at which the distortion curve levels off as well as its value associated with the largest jump.

The k-means clustering technique depends on the choice of initial cluster centers (Chattopadhyay et al. 2012). But this effect can be minimized if one chooses the cluster centers through group average method (Milligan 1980). As a result, the formation of the final groups will not depend heavily on the initial choice and hence will remain almost the same according to physical properties irrespective of initial centers. In MINITAB package, the k-means method is almost free from the effect of initial choice of centers as they have used the group average method.

8.4 Classification and Discrimination

Discriminant[2] analysis and classification are multivariate techniques concerned with separating distinct sets of objects and with allocating new objects to previously defined groups. Once the optimum clustering is obtained by applying the method discussed under previous section, one can verify the acceptability of the classification by computing classification/misclassification probabilities for the different objects. Although the k-means clustering method is purely a data analytic method, for classification it may be necessary to assume that the underlying distribution is multivariate normal. The method can be illustrated as follows for two populations (clusters). The method can be easily generalized for more than two underlying populations.

[2]A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical Methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer' is reproduced in this part.

Let $f_1(x)$ and $f_2(x)$ be the probability density functions associated with the $p \times 1$ random vector X for the populations π_1 and π_2 respectively. Let Ω be the sample space, i.e., collection of all objects. Let us denote by x the observed value of X. Let R1 be that set of x values for which we classify objects as π_1 and $R_2 = \Omega R_1$ be the remaining x values for which we classify objects as π_2. Since every object must be assigned to one and only one of the two groups, the sets R_1 and R_2 are disjoint and exhaustive. The conditional probability of classifying an object as π_2 when in fact it is from π_1 (error probability) is,

$$P(2 \mid 1) = P[X \in R_2 \mid \pi_1] = f_{R2} f_1(x) dx$$

Similarly, the other error probability can be defined. Let p_1 and p_2 be the prior probabilities of π_1 and π_1, respectively, $(p_1 + p_2 = 1)$. Then the overall probabilities of correctly and incorrectly classifying objects can be derived as

P (correctly classified as π_1) = P (Observation actually comes from π_1 and is correctly classified as π_1) $= P[X \in R_1 \mid \pi_2]p_2$.
P (misclassified as π_1) $= P[X \in R_1 \mid \pi_2]p_2$.

The associated cost of misclassification can be defined by a cost matrix

	Classified as	
True population	π_1	π_2
π_1	0	$C(2 \mid 1)$
π_2	$C(1 \mid 2)$	0

For any rule, the average or Expected Cost of Misclassification (ECM) is given by

$$ECM = C(2 \mid 1)P(2 \mid 1)p_1 + C(1 \mid 2P(1 \mid 2)p_2$$

A reasonable classification rule should have ECM as small as possible.

Rule: The regions R_1 and R_2 that minimize the ECM are defined by the value of x for which the following inequalities hold.

$$R_1 : \frac{f_1(x)}{f_2(x)} > \frac{C(1 \mid 2)p_2}{C(2 \mid 1)p_1}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{C(1 \mid 2)p_2}{C(2 \mid 1)p_1}$$

If we assume $f_1(x)$ and $f_2(x)$ are multivariate normal with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2, respectively, then a particular object with observation vector x_0 may be classified according to the following rule (under the assumption $\Sigma_1 = \Sigma_2$)

Allocate x_0 to π_1 if

$$(\mu_1 - \mu_2)'\Sigma^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \frac{C(1 \mid 2)p_2}{C(2 \mid 1)p_1}$$

allocate x_0 to π_2 otherwise.

If we choose $C(1 \mid 2) = C(2 \mid 1)$ and $p_1 = p_2$, then the estimated minimum ECM rule for two Normal populations will be as follows:

Allocate x_0 to π_1 if

$$(m_1 - m_2)'S_{\text{pooled}} - 1x_0 - \frac{1}{2}(m_1 - m_2)'\Sigma^{-1}(m_1 + m_2) \geq 0$$

where m_1 and m_2 are sample mean vectors of the two populations and S_{pooled} is pooled (combined) sample covariance matrix. Allocate x_0 to π_2 otherwise. The LHS is known as the linear discriminant function. One can easily generalize the method for more than two groups.

8.5 Data

Example 8.5.1 The Fisher's *Iris* data set is a multivariate data set introduced by Fisher (1936). It is also known as Anderson's *Iris* data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris setosa (type-3), Iris versicolor (type-2), and Iris virginica (type-1)). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters (Table 8.1).

We have performed k-means clustering of the data on the basis of the first four variables, viz. sepal length, sepal width, petal length, and petal width. Choosing $k = 3$, we have divided the 150 observations into three groups in order to verify whether we can identify three groups corresponding to three species. From columns 6 and 7, it is clear that k-means method has correctly identified Iris setosa (type-3) species for all the 50 cases, whereas there are some errors corresponding to types 1 and 2. For type 2, three cases and for type 1 fourteen cases had wrongly identified. The summary result for k-means clustering is given below:

Table 8.1 Results of k-means clustering for Iris data

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
5.1	3.5	1.4	0.2	I. setosa	3	3
4.9	3	1.4	0.2	I. setosa	3	3
4.7	3.2	1.3	0.2	I. setosa	3	3
4.6	3.1	1.5	0.2	I. setosa	3	3
5	3.6	1.4	0.2	I. setosa	3	3
5.4	3.9	1.7	0.4	I. setosa	3	3
4.6	3.4	1.4	0.3	I. setosa	3	3
5	3.4	1.5	0.2	I. setosa	3	3
4.4	2.9	1.4	0.2	I. setosa	3	3
4.9	3.1	1.5	0.1	I. setosa	3	3
5.4	3.7	1.5	0.2	I. setosa	3	3
4.8	3.4	1.6	0.2	I. setosa	3	3
4.8	3	1.4	0.1	I. setosa	3	3
4.3	3	1.1	0.1	I. setosa	3	3
5.8	4	1.2	0.2	I. setosa	3	3
5.7	4.4	1.5	0.4	I. setosa	3	3
5.4	3.9	1.3	0.4	I. setosa	3	3
5.1	3.5	1.4	0.3	I. setosa	3	3
5.7	3.8	1.7	0.3	I. setosa	3	3
5.1	3.8	1.5	0.3	I. setosa	3	3
5.4	3.4	1.7	0.2	I. setosa	3	3
5.1	3.7	1.5	0.4	I. setosa	3	3
4.6	3.6	1	0.2	I. setosa	3	3
5.1	3.3	1.7	0.5	I. setosa	3	3
4.8	3.4	1.9	0.2	I. setosa	3	3
5	3	1.6	0.2	I. setosa	3	3
5	3.4	1.6	0.4	I. setosa	3	3
5.2	3.5	1.5	0.2	I. setosa	3	3
5.2	3.4	1.4	0.2	I. setosa	3	3
4.7	3.2	1.6	0.2	I. setosa	3	3
4.8	3.1	1.6	0.2	I. setosa	3	3
5.4	3.4	1.5	0.4	I. setosa	3	3
5.2	4.1	1.5	0.1	I. setosa	3	3
5.5	4.2	1.4	0.2	I. setosa	3	3
4.9	3.1	1.5	0.2	I. setosa	3	3
5	3.2	1.2	0.2	I. setosa	3	3
5.5	3.5	1.3	0.2	I. setosa	3	3
4.9	3.6	1.4	0.1	I. setosa	3	3
4.4	3	1.3	0.2	I. setosa	3	3

(continued)

Table 8.1 (continued)

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
5.1	3.4	1.5	0.2	I. setosa	3	3
5	3.5	1.3	0.3	I. setosa	3	3
4.5	2.3	1.3	0.3	I. setosa	3	3
4.4	3.2	1.3	0.2	I. setosa	3	3
5	3.5	1.6	0.6	I. setosa	3	3
5.1	3.8	1.9	0.4	I. setosa	3	3
4.8	3	1.4	0.3	I. setosa	3	3
5.1	3.8	1.6	0.2	I. setosa	3	3
4.6	3.2	1.4	0.2	I. setosa	3	3
5.3	3.7	1.5	0.2	I. setosa	3	3
5	3.3	1.4	0.2	I. setosa	3	3
7	3.2	4.7	1.4	I. versicolor	2	1
6.4	3.2	4.5	1.5	I. versicolor	2	2
6.9	3.1	4.9	1.5	I. versicolor	2	1
5.5	2.3	4	1.3	I. versicolor	2	2
6.5	2.8	4.6	1.5	I. versicolor	2	2
5.7	2.8	4.5	1.3	I. versicolor	2	2
6.3	3.3	4.7	1.6	I. versicolor	2	2
4.9	2.4	3.3	1	I. versicolor	2	2
6.6	2.9	4.6	1.3	I. versicolor	2	2
5.2	2.7	3.9	1.4	I. versicolor	2	2
5	2	3.5	1	I. versicolor	2	2
5.9	3	4.2	1.5	I. versicolor	2	2
6	2.2	4	1	I. versicolor	2	2
6.1	2.9	4.7	1.4	I. versicolor	2	2
5.6	2.9	3.6	1.3	I. versicolor	2	2
6.7	3.1	4.4	1.4	I. versicolor	2	2
5.6	3	4.5	1.5	I. versicolor	2	2
5.8	2.7	4.1	1	I. versicolor	2	2
6.2	2.2	4.5	1.5	I. versicolor	2	2
5.6	2.5	3.9	1.1	I. versicolor	2	2
5.9	3.2	4.8	1.8	I. versicolor	2	2
6.1	2.8	4	1.3	I. versicolor	2	2
6.3	2.5	4.9	1.5	I. versicolor	2	2
6.1	2.8	4.7	1.2	I. versicolor	2	2
6.4	2.9	4.3	1.3	I. versicolor	2	2
6.6	3	4.4	1.4	I. versicolor	2	2
6.8	2.8	4.8	1.4	I. versicolor	2	2

(continued)

Table 8.1 (continued)

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
6.7	3	5	1.7	I. versicolor	2	1
6	2.9	4.5	1.5	I. versicolor	2	2
5.7	2.6	3.5	1	I. versicolor	2	2
5.5	2.4	3.8	1.1	I. versicolor	2	2
5.5	2.4	3.7	1	I. versicolor	2	2
5.8	2.7	3.9	1.2	I. versicolor	2	2
6	2.7	5.1	1.6	I. versicolor	2	2
5.4	3	4.5	1.5	I. versicolor	2	2
6	3.4	4.5	1.6	I. versicolor	2	2
6.7	3.1	4.7	1.5	I. versicolor	2	2
6.3	2.3	4.4	1.3	I. versicolor	2	2
5.6	3	4.1	1.3	I. versicolor	2	2
5.5	2.5	4	1.3	I. versicolor	2	2
5.5	2.6	4.4	1.2	I. versicolor	2	2
6.1	3	4.6	1.4	I. versicolor	2	2
5.8	2.6	4	1.2	I. versicolor	2	2
5	2.3	3.3	1	I. versicolor	2	2
5.6	2.7	4.2	1.3	I. versicolor	2	2
5.7	3	4.2	1.2	I. versicolor	2	2
5.7	2.9	4.2	1.3	I. versicolor	2	2
6.2	2.9	4.3	1.3	I. versicolor	2	2
5.1	2.5	3	1.1	I. versicolor	2	2
5.7	2.8	4.1	1.3	I. versicolor	2	2
6.3	3.3	6	2.5	I. virginica	1	1
5.8	2.7	5.1	1.9	I. virginica	1	2
7.1	3	5.9	2.1	I. virginica	1	1
6.3	2.9	5.6	1.8	I. virginica	1	1
6.5	3	5.8	2.2	I. virginica	1	1
7.6	3	6.6	2.1	I. virginica	1	1
4.9	2.5	4.5	1.7	I. virginica	1	2
7.3	2.9	6.3	1.8	I. virginica	1	1
6.7	2.5	5.8	1.8	I. virginica	1	1
7.2	3.6	6.1	2.5	I. virginica	1	1
6.5	3.2	5.1	2	I. virginica	1	1
6.4	2.7	5.3	1.9	I. virginica	1	1
6.8	3	5.5	2.1	I. virginica	1	1
5.7	2.5	5	2	I. virginica	1	2

(continued)

Table 8.1 (continued)

Sepal length	Sepal width	Petal length	Petal width	Species	Type	k-means Clus no.
5.8	2.8	5.1	2.4	I. virginica	1	2
6.4	3.2	5.3	2.3	I. virginica	1	1
6.5	3	5.5	1.8	I. virginica	1	1
7.7	3.8	6.7	2.2	I. virginica	1	1
7.7	2.6	6.9	2.3	I. virginica	1	1
6	2.2	5	1.5	I. virginica	1	2
6.9	3.2	5.7	2.3	I. virginica	1	1
5.6	2.8	4.9	2	I. virginica	1	2
7.7	2.8	6.7	2	I. virginica	1	1
6.3	2.7	4.9	1.8	I. virginica	1	2
6.7	3.3	5.7	2.1	I. virginica	1	1
7.2	3.2	6	1.8	I. virginica	1	1
6.2	2.8	4.8	1.8	I. virginica	1	2
6.1	3	4.9	1.8	I. virginica	1	2
6.4	2.8	5.6	2.1	I. virginica	1	1
7.2	3	5.8	1.6	I. virginica	1	1
7.4	2.8	6.1	1.9	I. virginica	1	1
7.9	3.8	6.4	2	I. virginica	1	1
6.4	2.8	5.6	2.2	I. virginica	1	1
6.3	2.8	5.1	1.5	I. virginica	1	2
6.1	2.6	5.6	1.4	I. virginica	1	1
7.7	3	6.1	2.3	I. virginica	1	1
6.3	3.4	5.6	2.4	I. virginica	1	1
6.4	3.1	5.5	1.8	I. virginica	1	1
6	3	4.8	1.8	I. virginica	1	2
6.9	3.1	5.4	2.1	I. virginica	1	1
6.7	3.1	5.6	2.4	I. virginica	1	1
6.9	3.1	5.1	2.3	I. virginica	1	1
5.8	2.7	5.1	1.9	I. virginica	1	2
6.8	3.2	5.9	2.3	I. virginica	1	1
6.7	3.3	5.7	2.5	I. virginica	1	1
6.7	3	5.2	2.3	I. virginica	1	1
6.3	2.5	5	1.9	I. virginica	1	2
6.5	3	5.2	2	I. virginica	1	1
6.2	3.4	5.4	2.3	I. virginica	1	1
5.9	3	5.1	1.8	I. virginica	1	2

Number of clusters: 3

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	39	25.414	0.732	1.552
Cluster2	61	38.291	0.731	1.647
Cluster3	50	15.151	0.482	1.248

We have also performed **linear discriminant analysis** by considering types as the true groups.

Linear Method for Response: Type
Predictors: Sepal le Sepal wi Petal le Petal wi
Summary of Classification

Put into	....True Group....		
Group	1	2	3
1	49	2	0
2	1	48	0
3	0	0	50
Total N	50	50	50

Summary of Classification with Cross-validation

Put into	....True Group....		
Group	1	2	3
1	49	2	0
2	1	48	0
3	0	0	50
Total N	50	50	50
N Correct	49	48	50
Proportion	0.980	0.960	1.000

N = 150 N Correct = 147 Proportion Correct = 0.980
Squared Distance Between Groups

	1	2	3
1	0.000	17.201	179.385
2	17.201	0.000	89.864
3	179.385	89.864	0.000

Linear Discriminant Function for Group

	1	2	3
Constant	−103.27	−71.75	−85.21
Sepal le	12.45	15.70	23.54
Sepal wi	3.69	7.07	23.59
Petal le	12.77	5.21	−16.43
Petal wi	21.08	6.43	−17.40

Variable Pooled Means for Group

	Mean	1	2	3
Sepal le	5.8433	6.5880	5.9360	5.0060
Sepal wi	3.0573	2.9740	2.7700	3.4280
Petal le	3.7580	5.5520	4.2600	.4620
Petal wi	1.1993	2.0260	1.3260	0.2460

Variable Pooled StDev for Group

	StDev	1	2	3
Sepal le	0.5148	0.6359	0.5162	0.3525
Sepal wi	0.3397	0.3225	0.3138	0.3791
Petal le	0.4303	0.5519	0.4699	0.1737
Petal wi	0.2047	0.2747	0.1978	0.1054

Pooled Covariance Matrix
Sepal le Sepal wi Petal le Petal wi
Sepal le 0.26501
Sepal wi 0.09272 0.11539
Petal le 0.16751 0.05524 0.18519
Petal wi 0.03840 0.03271 0.04267 0.04188
Here we see that only three observations are wrongly classified. The corresponding probabilities are given by

Observation	True Group	Pred Group	Group	Probability Predicted
71 **	2	1	1	0.75
			2	0.25
			3	0.00
84 **	2	1	1	0.86
			2	0.14
			3	0.00
134 **	1	2	1	0.27
			2	0.73
			3	0.00

Example 8.5.2 The following data are related to a survey on environmental pollution level. The following variables were observed in suitable units at 111 selected places. The four variables under study were Ozone content, Radiation, Temperature, and Wind speed in some proper units. We have performed hierarchical clustering with Euclidian distance and single linkage. The data set as well as the cluster membership is shown in the following table.

The summary of results and the dendrogram are given below the table. By considering similarity level at 93, six clusters were found of which three (4, 5, and 6) may omitted as outliers containing 2, 1, and 1 observations. Hence clusters 1, 2, and 3 are the main clusters. Figures corresponding to radiation, temperature, wind speed, ozone content and H-cluster number of 111 places.

Table 8.2 Results of hierarchical clustering for pollution data

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
190	67	7.4	41	1
118	72	8	36	2
149	74	12.6	12	2
313	62	11.5	18	1
299	65	8.6	23	1
99	59	13.8	19	2
19	61	20.1	8	3
256	69	9.7	16	1
290	66	9.2	11	1
274	68	10.9	14	1
65	58	13.2	18	3
334	64	11.5	14	1
307	66	12	34	1
78	57	18.4	6	3
322	68	11.5	30	1
44	62	9.7	11	3
8	59	9.7	1	3
320	73	16.6	11	1
25	61	9.7	4	3
92	61	12	32	2
13	67	12	23	3
252	81	14.9	45	1
223	79	5.7	115	1
279	76	7.4	37	1
127	82	9.7	29	2
291	90	13.8	71	1

(continued)

Table 8.2 (continued)

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
323	87	11.5	39	1
148	82	8	23	2
191	77	14.9	21	1
284	72	20.7	37	1
37	65	9.2	20	3
120	73	11.5	12	2
137	76	10.3	13	2
269	84	4	135	4
248	85	9.2	49	1
236	81	9.2	32	1
175	83	4.6	64	1
314	83	10.9	40	1
276	88	5.1	77	1
267	92	6.3	97	1
272	92	5.7	97	1
175	89	7.4	85	1
264	73	14.3	10	1
175	81	14.9	27	1
48	80	14.3	7	3
260	81	6.9	48	1
274	82	10.3	35	1
285	84	6.3	61	1
187	87	5.1	79	1
220	85	11.5	63	1
7	74	6.9	16	3
294	86	8.6	80	1
223	85	8	108	1
81	82	8.6	20	3
82	86	12	52	3
213	88	7.4	82	1
275	86	7.4	50	1
253	83	7.4	64	1
254	81	9.2	59	1
83	81	6.9	39	3
24	81	13.8	9	3
77	82	7.4	16	3
255	89	4	122	4
229	90	10.3	89	1
207	90	8	110	1

(continued)

Table 8.2 (continued)

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
192	86	11.5	44	1
273	82	11.5	28	1
157	80	9.7	65	1
71	77	10.3	22	3
51	79	6.3	59	5
115	76	7.4	23	2
244	78	10.9	31	1
190	78	10.3	44	1
259	77	15.5	21	1
36	72	14.3	9	3
212	79	9.7	45	1
238	81	3.4	168	6
215	86	8	73	1
203	97	9.7	76	1
225	94	2.3	118	1
237	96	6.3	84	1
188	94	6.3	85	1
167	91	6.9	96	1
197	92	5.1	78	1
183	93	2.8	73	1
189	93	4.6	91	1
95	87	7.4	47	3
92	84	15.5	32	3
252	80	10.9	20	1
220	78	10.3	23	1
230	75	10.9	21	1
259	73	9.7	24	1
236	81	14.9	44	1
259	76	15.5	21	1
238	77	6.3	28	1
24	71	10.9	9	3
112	71	11.5	13	2
237	78	6.9	46	1
224	67	13.8	18	1
27	76	10.3	13	3
238	68	10.3	24	1
201	82	8	16	1
238	64	12.6	13	1
14	71	9.2	23	3

(continued)

Table 8.2 (continued)

Radiation	Temperature	Wind speed	Ozone content	H-cluster number
139	81	10.3	36	2
49	69	10.3	7	3
20	63	16.6	14	3
193	70	6.9	30	1
191	75	14.3	14	1
131	76	8	18	2
223	68	11.5	20	1

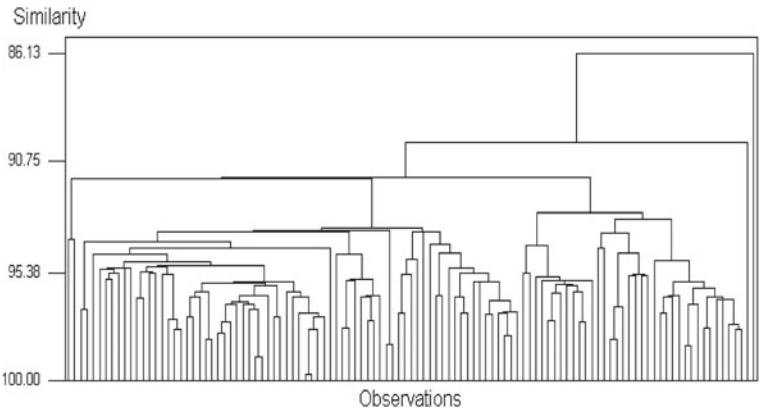

Fig. 8.1 Dendrogram of pollution data

Number of main clusters: 3

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	71	202337.219	48.851	101.003
Cluster2	12	5151.429	18.929	35.732
Cluster3	24	26269.208	30.505	58.654

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centroid
Radiatio	240.7606	123.9167	46.6250	184.8018
Temperat	80.1831	73.5833	71.9167	77.7928
Wind spe	9.6577	10.2583	11.5292	9.9387
Ozone Co	49.2535	22.1667	17.7500	42.0991

The dendrogram of the pollution data is shown below. The centroids of the first three clusters are widely separated corresponding to all the variables; the 24 places falling in cluster 3 may be considered to be least polluted, whereas the 71 places falling in cluster 1 are most polluted (Fig. 8.1).

References and Suggested Readings

Bien, J., & Tibshirani, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, *106*(495), 1075–1084.

Chattopadhyay, A. K., & Chattopadhyay, T. (2014). *Statistical methods for astronomical data analysis*., Springer series in astrostatistics New York: Springer.

Chattopadhyay, T., et al. (2012). Uncovering the formation of ultracompact dwarf galaxies by multivariate statistical analysis. *Astrophysical Journal*, *750*, 91.

De, T., Chattopadhyay, T., & Chattopadhyay, A. K. (2013). Comparison among clustering and classification techniques on the basis of galaxy data. *Calcutta Statistical Association Bulletin*, *65*, 257–260.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Fraix-Burnet, D., Thuillard, M., & Chattopadhyay, A. K. (2015). Multivariate approaches to classification in extragalactic astronomy. *Frontiers in Astronomy and Space Science*, *2*, 1–17.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, p. 281).

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325–342.

Sugar, A. S., & James, G. M. (2003). Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, *98*, 750.

Chapter 9
Principal Component Analysis

9.1 Introduction

Under the high-dimensional setup with p variables, the problem that often arises is the critical nature of the correlation or covariance matrix. When p is moderately or very large it is generally difficult to identify the true nature of relationship among the variables as well as observations from the covariance or correlation matrix. Under such situations, a very common way to simplify the matter is to reduce the dimension by considering only those variables (components) those are truly responsible for the overall variation.

Principal component analysis (PCA) is a dimension reduction procedure. PCA was developed in 1901 by Karl Pearson, as an analogue of the principal axis theorem in mechanics. It was later independently developed by Harold Hotelling (1933, 1936). Several authors considered PCA in different forms (Joliffe 1982, 2002). There are several case studies and applications (Jeffers 1967; Chattopadhyay and Chattopadhyay 2006). The method is useful when we have a large number of variables, and some variables are of less or no importance. In this case, redundancy means that some of the variables are highly correlated with one another, possibly because they are measuring the same phenomenon. Because of this redundancy, it should be possible to reduce the observed variables into a smaller number of principal components (derived variables) that will account for most of the variance in the observed variables.

Being a dimension reduction technique, principal component analysis has similarities with exploratory factor analysis. The steps followed when conducting a principal component analysis are almost the same as those of exploratory factor analysis. However, there are significant conceptual differences between the reduction procedure that gives a relatively small number of components those account for most of the variance in a set of observed variables. In summary, both factor analysis and principal component analysis have important roles to play in social science research, but their conceptual foundations are quite different.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_9

More recently, Independent component analysis (ICA) has been identified as a strong competitor for principal component analysis and factor analysis. ICA finds a set of source data that are mutually independent (not only with respect to the second moment), but PCA finds a set of data that are mutually uncorrelated and the principal components become independent only under Gaussian setup. ICA was primarily developed for non-Gaussian data in order to find independent components responsible for a larger part of the variation. ICA separates statistically independent original source data from an observed set of data mixtures.

9.1.1 Method

In PCA, primarily[1] it is not necessary to make any assumption regarding the underlying multivariate distribution but if we are interested in some inference problems related to PCA then the assumption of multivariate normality is necessary (Chattopadhyay and Chattopadhyay 2014). The eigenvalues and eigenvectors of the covariance or correlation matrix are the main contributors of a PCA. Of course, the eigenvalues of covariance and correlation matrices are different and they coincide when we work with standardized values of the variables. So the decision whether one should start work covariance or correlation matrix is important. Normally, when all the variables are of equal importance, one may start with the correlation matrix. The eigenvectors determine the directions of maximum variability, whereas the eigenvalues specify the variances. In practice, decisions regarding the quality of the principal component approximation should be made on the basis of eigenvalue–eigenvector pairs. In order to study the sampling distribution of their estimates, the multivariate normality assumptions became necessary as otherwise it is too difficult. Principal components are a sequence of projections of the data. The components are constructed in such a way that they are uncorrelated and ordered in variance. The components of a p-dimensional data set provide a sequence of best linear approximations. As only a few of such linear combinations may explain a larger percentage of variation in the data, one can take only those components instead of p variables for further analysis.

A PCA is concerned with explaining the variance–covariance structure through a few linear combinations of the original variables. Its general objectives are data reduction and interpretation. Reduce the number of variables from p to $k < (kp)$. Let the random vector $X' = (X_1 \ldots X_p)$ have the covariance matrix Σ (or correlation matrix R) with ordered eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p \geq 0$ and corresponding eigenvectors $e'_1, e'_2, \ldots, e'_p$, respectively.

[1]This section draws from one of the authors' published work, 'Statistical Methods for Astronomical Data Analysis,' authored by Asis Kumar Chattopadhyay and Tanuka Chattopadhyay, and published in 2014 by Springer Science+Business Media New York.

Consider the linear combinations

$$\begin{aligned} Y_1 &= l_{11}X_1 + l_{21}X_2 + \cdots + l_{p1}X_p = e_1'X \\ Y_2 &= l_{12}X_1 + l_{22}X_2 + \cdots + l_{p2}X_p = e_2'X \\ &\vdots \\ Y_p &= l_{1p}X_1 + l_{2p}X_2 + \cdots + l_{pp}X_p = e_p'X \end{aligned}$$

Then we have the following result:

Result: Let $X' = (X_1 \ldots X_p)$ have covariance matrix Σ with eigenvalue–eigenvector pairs $(\lambda_1, e_1) \ldots (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq .\lambda_p \geq 0$.

Let $Y_1 = e_1'X$, $Y_2 = e_2'X \ldots Y_p = e_p'X$.

Then

$$\text{var}(Y_i) = \lambda_i (i = 1, 2, \ldots p) \text{ and}$$

$$\sigma_{11} + \sigma_{22} \cdots + \sigma_{pp} = \sum\nolimits_1^p \text{var}(Xi)$$

$$= \lambda_1 + \cdots + \lambda_p$$

$$= \sum\nolimits_1^p \text{var}(Yi)$$

Here $Y_1, Y_2, \ldots, Y_p$ are called **principal components**. In particular, Y_1 is the **first principal component** (having the largest variance), Y_2 is the second principal component (having the second largest variance), and so on.

(For proof of the above result, one may consult any standard textbook.)

Here instead of original p variables $X_1 \ldots X_p$, only a few principal components $Y_1, Y_2, \ldots, Y_k (k < p)$ are used which explains maximum part of the total variation. There are several methods to find the optimum value of k.

The specific aim of the analysis is to reduce a large number of variables to a smaller number of components by retaining the total variance (sum of the diagonal components of the covariance matrix) almost same among the observations. The analysis therefore helps us to determine the optimum set of artificial variables (viz. linear combinations) explaining the overall variations in the nature of objects.

Many criteria have been suggested by different authors for deciding how many principal components (k) to retain. Some of these criteria are as follows:

1. Include just enough components to explain some arbitrary amount (say 80%) of the total variance which is the sum of the variances (diagonal elements of the covariance matrix) of all the variables.
2. Exclude those principal components with eigenvalues below the average. For principal components calculated from the correlation matrix, this criterion excludes components with eigenvalues less than 1.

3. Use of the screen plot (plotting eigenvalues against components) technique.[2]

Example 9.1.1 (http://openmv.net/) The following data set gives the relative consumption of certain food items in European and Scandinavian countries. The numbers represent the percentage of the population consuming that food type corresponding to 15 countries and 9 food types. As there are 9 food types corresponding to only 15 countries, it is necessary to reduce the dimension in order to search for major food types.

		Instant		Powder
Country	coffee	Tea	Biscuits	soup
Germany	49	88	57	51
Italy	10	60	55	41
France	42	63	76	53
Holland	62	98	62	67
Belgium	38	48	74	37
Luxembourg	61	86	79	73
England	86	99	91	55
Portugal	26	77	22	34
Austria	31	61	29	33
Switzerland	72	85	31	69
Denmark	17	92	66	32
Norway	17	83	62	51
Finland	12	84	64	27
Spain	40	40	62	43
Ireland	52	99	80	75

			Frozen		
Country	Potatoes	fish	Apples	Oranges	Butter
Germany 21	27	81	75	91	
Italy	2	4	67	71	66
France	23	11	87	84	94
Holland	7	14	83	89	31
Belgium	9	13	76	76	84
Luxembourg	7	26	85	94	94
England	17	20	76	68	95
Portugal	5	20	22	51	65
Austria	5	15	49	42	51
Switzerland	17	19	79	70	82
Denmark	11	51	81	72	92
Norway	17	30	61	72	63
Finland	8	18	50	57	96
Spain	14	23	59	77	44
Ireland	2	5	57	52	97

[2] A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer' is reproduced in this part.

Table 9.1 Eigen analysis of the correlation matrix

Components	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalue	3.4129	1.5591	1.3412	1.0164	0.7587	0.3294	0.2633	0.2027	0.1162
Proportion	0.379	0.173	0.149	0.113	0.084	0.037	0.029	0.023	0.013
Cumulative	0.379	0.552	0.701	0.814	0.899	0.935	0.965	0.987	1.000

Table 9.2 Coefficients of 15 variables in 9 principal components

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Instant	0.383	−0.367	−0.035	−0.247	−0.335	−0.360	−0.573	0.140	0.258
Tea	0.241	−0.229	0.596	−0.326	−0.274	0.000	0.441	0.304	0.256
Biscuits	0.368	0.062	0.069	0.571	−0.249	−0.650	0.135	−0.058	−0.155
Powder s	0.389	−0.467	−0.053	−0.170	−0.047	0.225	0.119	−0.561	−0.467
Potatoes	0.284	0.412	−0.078	−0.219	0.667	−0.144	0.476	−0.058	−0.032
Frozen f	0.079	0.575	0.305	−0.452	−0.298	−0.155	−0.377	−0.248	−0.221
Apples	0.465	0.174	−0.205	0.057	−0.108	0.360	−0.123	0.632	−0.389
Oranges	0.394	0.208	−0.424	−0.013	−0.340	0.201	0.058	−0.258	0.629
Butter	0.224	0.134	0.561	0.471	0.297	0.427	−0.240	−0.197	0.169

From the screen plot and Table 9.1, it is clear that **4 components** have variances (i.e., eigenvalues of the correlation matrix) **greater than one** and these four components explain **more than 80% of the total variation, i.e., the sum of the variances of all the variables**. Hence, one can work with four principal components instead of the original nine variables.

From Table 9.2, it is clear that most of the variables have similar importance in all the first four components so that it is difficult to associate a particular component to a subset of variables. So here it is not possible to identify the physical nature of the components. This feature is generally true for principal component analysis. In order to find inherent factors, one can take help of factor analysis if the nature of the covariance matrix admits (Figs. 9.1 and 9.2).

9.1.2 The Correlation Vector Diagram (Biplot, Gabriel 1971)

A matrix of rank 2 can be displayed as a biplot consisting of a vector for each row and a vector for each column, chosen so that each element of the matrix is exactly the inner product of the vectors corresponding to its row and its column (Gabriel 1971). If a matrix is of higher rank, one may display it approximately by a biplot of a matrix of rank 2 that approximates the original matrix. In PCA, a biplot can show inter-unit distances and indicate the clustering of units, as well as displaying the variances and correlations of the variables.

Fig. 9.1 Screen plot used to decide about the number of significant principal components. The components with eigenvalues greater than 1 are usually taken as significant

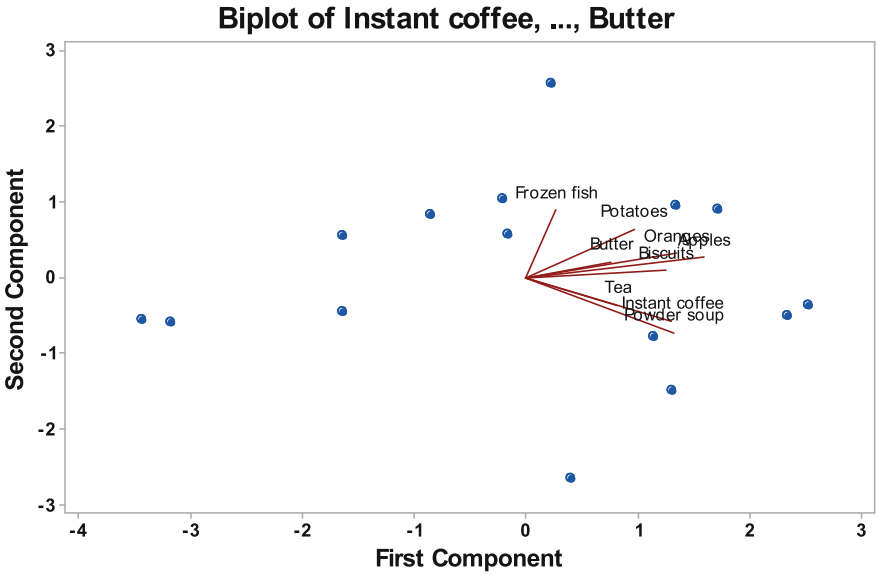

Fig. 9.2 Biplot for the data used in Example 9.1.1. The vector lengths represent variances of corresponding variables, and the angles show correlations of the variables (smaller angles indicate higher correlations). Dot points indicate the positions of the 15 countries with respect to their first and second component values. The origin represents the average value for each variable; that is, it represents the object that has an average value in each variable

Any matrix of observations y of order mXn can be written by singular value decomposition as

$$y = \Sigma_1^r \lambda_i piqi'(\lambda_1 \geq \lambda_2 \cdots \geq \lambda_r)$$

where r is the rank of the matrix y and λ_i, p_i, and q_i' are the singular value, singular column, and singular row, respectively. Then by the method of least-squares fitting a matrix of rank 2, an approximation of y is given by

$$y = \Sigma_1^2 \lambda_i piqi'$$

and the corresponding measure of goodness of fit is given by

$$\rho(2) = \frac{\lambda_1^2 + \lambda_2^2}{\sigma_1^r \lambda i^2}$$

If $\rho(2)$ is near to 1, then such a biplot will give a good approximation to y. If we denote by

$$\begin{aligned} \boldsymbol{S}^{mXm} &= (1/n)\boldsymbol{y'y} = (s_{ij}) = \text{variance–covariance matrix and} \\ \boldsymbol{R}^{mXm} &= (r_{ij}) = \text{correlation matrix} \end{aligned}$$

then it can be shown that

$$\boldsymbol{y}^{nXm} \sim \boldsymbol{G}^{nX2}\boldsymbol{H}'^{2Xm}$$

where

$$\boldsymbol{G}^{nX2} = (p1'p2')\sqrt{n} = (g_1^{nX1} g_2^{nX1})$$

and

$$\boldsymbol{H}^{mX2} = \left(\frac{1}{\sqrt{n}}\right)(\lambda_{1q1}\lambda_{2q2}) = (h_1^{mX1} h_2^{mX1}).$$

Further,

$$s_{ij} \sim h_i' h_j$$

$$s_j^2 \sim ||h_j||^2$$

$$r_{ij} \sim \cos(h_i h_j).$$

9.2 Properties of Principal Components

In PCA, the first component extracted explains the maximum amount of total variance in the observed variables. Under some conditions, this means that the first component will be correlated with at least some of the observed variables. It may be correlated with many. The second component will have two important characteristics. First,

this component explains a maximum amount of variance in the data set that was not accounted for by the first component. Again under some conditions, this means that the second component will be correlated with some of the observed variables that did not display strong correlations with the first component.

The second characteristic of the second component is that it will be uncorrelated (orthogonal) with the first component. The remaining components that are extracted in the analysis display the same two characteristics: Each component accounts for a maximum amount of variance in the observed variables which was not accounted for by the preceding components, and is uncorrelated with all of the preceding components. A principal component analysis proceeds in this fashion, with each new component accounting for progressively smaller and smaller amounts of variance (this is why only the first few components are usually retained and interpreted). When the analysis is complete, the resulting components will display varying degrees of correlation with the observed variables (https://support.sas.com/publishing/pubcat/chaps/55129.pdf), but are completely uncorrelated with one another.

Since no correlation does not generally imply that the components are independent, principal components are not generally independent except for normal distribution under which zero correlation implies independence. This is the reason why PCA works more successfully for Gaussian data. For non-Gaussian data, the independent component analysis is a better option.

References and Suggested Readings

Chattopadhyay, A. K. (2013). Independent component analysis for the objective classification of globular clusters of the galaxy NGC 5128. *Computational Statistics and Data Analysis*, *57*, 17–32.

Chattopadhyay, A. K., & Chattopadhyay, T. (2014). *Statistical methods for astronomical data analysis*., Springer series in astrostatistics New York: Springer.

Chattopadhyay, T., & Chattopadhyay, A. K. (2006). Objective classification of spiral galaxies having extended rotation curves beyond the optical radius. *Astronomical Journal*, *131*, 2452.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441; *24*(7), 498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–77.

Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *16*(3), 225–236.

Joliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *31*(3), 300–303.

Jolliffe, I. T. (2002). Principal component analysis. Springer series in statistics (2nd ed., XXIX, 487 p. illus.). New York: Springer. ISBN 978-0-387-95442-4.

Chapter 10
Factor Analysis

10.1 Factor Analysis

Factor analysis (Chattopadhyay and Chattopadhyay 2014) is a statistical method used to study the dimensionality of a set of variables. In factor analysis, latent variables represent unobserved constructs and are referred to as factors or dimensions. Factor analysis attempts to identify underlying variables or factors that explain the pattern of correlations within a set of observed variables. Factor analysis is often used in data reduction to identify a small number of factors that explain most of the variance that is observed in a much larger number of manifest variables.

Suppose the observable random vector X with p components has mean vector μ and covariance matrix Σ. In the factor model, we assume that X is linearly dependent upon a few unobservable random variables $F_1, F_2 \ldots F_p$ called common factors and p additional sources of variation $\in_1, \in_2, \ldots, \in_p$ called the errors (or specific factors). Then the factor model is

$$\overset{p\times 1}{X} = \overset{p\times 1}{\mu} + \overset{p\times m}{L}\overset{m\times 1}{F} + \overset{p\times 1}{\in} \tag{10.1.1}$$

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \in_1 \\ X_2 - \mu_1 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \in_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \in_p \end{aligned}$$

The coefficients l_{ij}s are called the loading of the ith variable on the jth factor so the matrix L is the matrix of factor loadings. Here $\in_i$ is associated only with the ith response X_i. Here the p deviations $X_1 - \mu_1 \ldots X_p - \mu_p$ are expressed in terms of

A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical Methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer Science+Business Media New York' is reproduced in this chapter.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_10

$p + m$ random variables $F_1, F_2, \ldots, F_m, \in_1, \ldots \in_p$ which are unobservable (but in multivariate regression independent variables can be observed).

With same additional assumption on the random vectors F and $\in$, the model w implies certain covariance relationships which can be checked.

We assume that

$$E(P) = 0^{m\times 1} \quad cov(F) = E(FP') = I^{m\times m}$$

$$E(\in) = 0^{p\times 1} \quad cov(\in) = E(\in\in') = \psi = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ 0 & 0 & \ldots & \psi_p \end{pmatrix}$$

$$\text{and } cov(\in, F) = E(\in, F) = 0^{p\times m} \tag{10.1.2}$$

The model $X - \mu = LF + \in$ is linear in the common factors. If the p response of X are related to the underlying in factors in a nonlinear form $[X_1 - \mu_1 = F_1F_3 + \in_1]$ Then the covariance structure $LL' + \psi$ may not be adequate. The assumption of linearity is inherent here.

These assumption and the relation (10.1.1) constitute the orthogonal factor model.

The orthogonal factor model implies a covariance structure for X.

$$\begin{aligned} \text{Here } (X - \mu)(X - \mu') &= (LF + \in)(LF + \in)' \\ &= (LF + \in)((LF)' + \in') \\ &= LF(LF)' + \in (LF)' + LF \in' + \in\in' \\ &= LFF'L' + \in F'L' + LF \in' + \in\in' \end{aligned}$$

$$\begin{aligned} \Sigma &= \text{ covariance matrix of } X \\ &= E(X - \mu)(X - \mu)' \\ &= LE(FF')L' + E(\in F)'L' + LE(F \in') + E(\in\in') \\ &= LIL' + \psi = LL' + \psi \end{aligned}$$

$$\text{Again } (X - \mu)F' = (LF + \in)F' = LFF' + \in F'$$

$$\text{or, } \operatorname{cov}(X, F) = E(X - \mu)F' = E(LF + \in)F' = LE(FF') + E(\in F') = L$$

Now $\Sigma = LL' + \psi$ implies

$$\left.\begin{aligned} \operatorname{var}(X_i) &= {l_{i1}}^2 + \cdots + {l_{im}}^2 + \psi_i \\ \operatorname{cov}(X_i X_k) &= l_{i1}l_{k1} + \cdots + l_{im}l_{km} \end{aligned}\right\} \tag{10.1.3}$$

$$\operatorname{cov}(XF) = L \Rightarrow cov(X_i F_j) = l_{ij}$$

$$\Rightarrow V(X_i) = \delta_{ii} = {l_{i1}}^2 + \cdots + {l_{im}}^2 + \psi_i$$

Let ith communality $= {h_i}^2 = {l_{i1}}^2 + \cdots + {l_{im}}^2$

Then $\delta_{ii} = {h_i}^2 + \psi_i \quad (i = 1 \ldots p)$

${h_i}^2 =$ sum of squares of loadings of ith variable on the m common factors.

Given a random sample of observations $\overset{b\times 1}{x_1}, x_2 \ldots \overset{p\times 1}{x_p}$. The basic problem is to decide whether Σ can be expressed in the form (10.1.3) for reasonably small value of m, and to estimate the elements of L and ψ.

Here the estimation procedure is not so easy. Primarily, we have from the sample data estimates of the $\frac{p(p+1)}{2}$ distinct elements of the upper triangle of Σ but on the RHS of (10.1.3) we have $pm + p$ parameters, pm for L and p for ψ. The solution will be indeterminate unless $\frac{p(p+1)}{2} - p(m+1) \geq 0$ or $p > 2m$. Even if this condition is satisfied L is not unique.

Proof Let $\overset{m\times m}{T}$ be any $\perp$ matrix so that $TT' = T'T = I$

Then (10.1.1) can be written as

$$X - \mu = LF + \in = LTT'F + \in = L^*F^* + \in \tag{10.1.4}$$

$$\text{where } L^* = LT \text{ and } F^* = T'F$$

$$\text{Since } E(F^*) = T'E(F) = 0$$

$$\text{and } \operatorname{cov}(F^*) = T'\operatorname{Cov}(F)T = T'T = I$$

It is impossible to distinguish between loadings L and L^* on the basis of the observations on X. So the vectors F and $F^* = T'F$ have the same statistical properties and even if the loadings L and L^* are different, they both generate the same covariance matrix Σ, i.e.,

$$\Sigma = LL' + \psi = LTT'L' + \psi = L^*L^{*\prime} + \psi \tag{10.1.5}$$

The above problem of uniqueness is generally resolved by choosing an orthogonal rotation T such that the final loading L satisfies the condition that $L'\psi^{-1}L$ is diagonal with positive diagonal elements. This restriction requires L to be of full rank m. With a valid ψ viz. one with all positive diagonal elements it can be shown that the above restriction yields a unique L. □

10.1.1 Method of Estimation

Given n observations vectors $x_1 \ldots x_n$ on p generally correlated variables, factor analysis seeks to verify whether the factor model (10.1.1) with a small number of factors adequately represent the data.

The sample covariance matrix is an estimator of the unknown covariance matrix Σ. If Σ appears to deviate significantly from a diagonal matrix, then a factor model can be used and the initial problem is one of estimating the factor loadings l_{ij} and the specific variances. ψ_i.

Principal Component Method[1]

Let Σ has eigenvalue–eigenvector pairs (λ_i, e_i) with $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$. Then by specified decomposition

$$\begin{aligned}\Sigma &= \lambda_1 e_1 e_1{}' + \lambda_2 e_2 e_2{}' + \cdots + \lambda_p e_p e_p{}' \\ &= (\sqrt{\lambda_1} e_1 \cdots \sqrt{\lambda_p} e_p) \begin{pmatrix} \sqrt{\lambda_1} e_1{}' \\ \vdots \\ \sqrt{\lambda_p} e_p{}' \end{pmatrix} e_1 \sqrt{\lambda_1} \cdots e_p \sqrt{\lambda_p} \qquad (10.1.6) \\ &= {}^{p\times p}L \;\; {}^{p\times p}L' + 0^{p\times p}\end{aligned}$$

[in (10.1.6) $m = p$ and jth column of $L = \sqrt{\lambda_j} e_j$].
Apart from the scale factor $\sqrt{\lambda_j}$, the factor loadings on the jth factor are the pp^n jth principal component.

The approximate representation assumes that the specific factors $\in$ are of minor importance and can be ignored in factoring Σ. If specific factors are included in the model, their variances may be taken to be the diagonal elements of $\Sigma - LL'$.

Allowing for specific factors, the approximation becomes

$$\begin{aligned}\Sigma &= LL' + \psi \\ &= (\sqrt{\lambda_1} e_1 \sqrt{\lambda_2} e_2 \cdots \sqrt{\lambda_m} e_m) \begin{pmatrix} \sqrt{\lambda_1} e_1{}' \\ \sqrt{\lambda_2} e_2{}' \\ \vdots \\ \sqrt{\lambda_m} e_m{}' \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{pmatrix} \qquad (10.1.7)\end{aligned}$$

where $m \leq p$.
(we assume that last $p - m$ eigenvalues are small)
and $\psi_{ii} = \delta_{ii} - \sum_{j=1}^{m} l_{ij}{}^2$ for $i = 1 \ldots p$.

[1] A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer' is reproduced in this part.

For the principal component solution, the estimated factor loadings for a given factor do not change as the number of factors is increased. If $m = 1$

$$L = \left(\sqrt{\lambda_1}\widehat{e_1}\right)$$

if $m = 2$

$$L = \left(\sqrt{\widehat{\lambda_1}}\widehat{e_1}\sqrt{\widehat{\lambda_2}}\widehat{e_2}\right)$$

where $(\widehat{\lambda_1}, \widehat{e_1})$ and $(\widehat{\lambda_2}, \widehat{e_2})$ are the first two eigenvalue–eigenvector pairs for S (or R).

By the definition of $\widehat{\psi_i}$, the diagonal elements of S are equal to the diagonal elements of $\widehat{L}\widehat{L}' + \psi$. How to determine m?

The choice of m can be based on the estimated eigenvalues.
Consider the residual matrix $S - (LL' + \psi)$

Here the diagonal elements are new and if the off-diagonal elements are also small we may take that particular value of m to be appropriate.

Analytically, we chose that m for which

$$\text{Sum of squared entries of } (S - (LL' + \psi)) \leq \widehat{\lambda}_{m+1}^2 + \cdots + \widehat{\lambda}_p^2 \quad (10.1.8)$$

Ideally, the contribution of the first few factors to the sample variance of the variables should be large. The contribution to the sample variance s_{ii} from the first common factor is ${l_{ii}}^2$. The contribution to the total sample variance $s_{11} + \cdots s_{pp} = h(S)$ from the first common factor is

$$\widehat{l}_{11}^2 + \widehat{l}_{21}^2 + \cdots + \widehat{l}_{p1}^2 = (\sqrt{\lambda_1}\widehat{e_1})'(\sqrt{\lambda_1}\widehat{e_1}) = \widehat{\lambda}_1$$

Since the eigenvectors $\widehat{e_1}$ has unit length.

In general,

$$\begin{pmatrix} \text{Propertion of total} \\ \text{sample variance due} \\ \text{to the factor} \end{pmatrix} = \begin{cases} \dfrac{\widehat{\lambda}_j}{s_{11}+\cdots+s_{pp}} \text{ for a factor analysis of } S \\ \dfrac{\widehat{\lambda}_j}{p} \text{ for a factor analysis of } R \end{cases} \quad (10.1.9)$$

Criterion (10.1.9) is frequently used as a heuristic device for determining the appropriate number of common factors. The value of m is gradually increased until a suitable proportion of the total sample variance has been explained.

Other Rules Used in Package

No. of eigenvalue of R greater than one (when R is used)
No. of eigenvalue of S that are positive (when S is used).

10.1.2 Factor Rotation

If $\widehat{L}$ be the $p \times m$ matrix of estimated factor loadings obtained by any method, then

$$L^* = \widehat{L}T \text{ where } TT' = T'T = I$$

is a $p \times m$ matrix of **rotated loadings**.

Moreover, the estimated covariance (or correlation) matrix remains unchanged since

$$\widehat{L}\widehat{L}' + \widehat{\psi} = \widehat{L}TT'\widehat{L}' + \widehat{\psi} = \widehat{L}^*\widehat{L}^{*'} + \widehat{\psi}$$

The above equation indicates that the residual matrix $S_n - \widehat{L}\widehat{L}' - \widehat{\psi} = S_n - \widehat{L}^*\widehat{L}^{*'} - \widehat{\psi}$ remains unchanged. Moreover, the specific variances $\widehat{\psi}_i$ and hence the communication $\widehat{h}_i^2$ are unaltered. Hence, mathematically it is immaterial whether $\widehat{L}$ or L^* is obtained.

Since the original loadings may not be readily interpretable, it is usual practice to rotate them until a 'simple structure' is achieved.

Ideally, we should like to see a pattern of loadings of each variable loads highly on a single factor and has small to moderate loading on the remaining factors.

The problem is to find an **orthogonal rotation** which compounds to a 'simple structure.'

There can be achieved if often rotation the orthogonality of the factor still exists. This is maintained of we perform orthogonal rotation. Among these (1) Variance rotation, (2) Quartimax rotation, (3) Equamax rotation are important.

Oblique rotation does not ensure the orthogonality of factors often rotation. There are several algorithms like oblimax, Quartimax.

10.1.3 Varimax Rotation

Orthogonal Transformation on L

$$L^* = LT \quad TT' = I$$

L^* is the matrix of orthogonally rotated loadings and let $d_j = \sum_{i=1}^{p} l_{ij}^{*2} \quad j = 1 \ldots m$

Then the following expression is maximized

$$\sum_{j=1}^{m} \left\{ \sum_{i=1}^{p} \left(l_{ij}^{*4} - {d_j}^2/p \right) \right\}$$

Such a procedure tries to give either large (in absolute value) or zero values in the columns of Ł*. Hence, the procedure tries to produce factors with either a stray association with the responses or no association at all.

The communality

$h_i{}^2 = \sum_{j=1}^{m} l^{*2}_{ij} = \sum_{j=1}^{m} l_{ij}{}^2$ remains constant under rotation.

10.2 Quartimax Rotation

The factor pattern is simplified by forcing the variables to correlate highly with one main factor (the so-called G-factor of 1Q studies) and very little with remaining factors. Here all variables are primarily associated with a single factor.
Interpretation of results obtained from factor analysis. It is usually difficult to interpret. Many users should significant coefficient magnitudes on many of the retained factors (coefficient greater than —.60— are often considered large and coefficients of —0.35— are often considered moderate). And especially on the first factor.

For good interpretation, factor rotation is necessary. The objective of the rotation is to achieve the most 'simple structure' though the manipulation of factor pattern matrix.

The most simple structure can be explained in terms of five principles of factor rotation.

1. Each variable should have at least one zero (small) loadings.
2. Each factor should have a set of linearly independent variables where factor loadings are zero (small).
3. For every pair of factors, there should be several variables where loadings are zero (small) for one factor but not the other.
4. For every pair of factors, a large proportion of variables should have zero (small) loading on both factors whenever more than about four factors are extracted.
5. For every pair of factors, there should only be a small number of variables with nonzero loadings on both.

In orthogonal rotation,

(1) Factors are perfectly uncorrelated with one another.
(2) Less parameters are to be estimated.

10.3 Promax Rotation

Factors are allowed to be correlated with one another.

Step I. Rotate the factors orthogonally.

Step II. Get a target matrix by raising the factor coefficients to an exponent (3 or 4). The coefficients secure smaller but absolute distance increases.
Step III. Rotate the original matrix to a best-fit position with the target matrix.

Here many moderate coefficients quickly approaches zero [$.3 \times .3 = .09$] then the large coefficients ($\geq .6$).

Example 10.1 Consider the data set related to the relative consumption of certain food items in European and Scandinavian countries considered in the chapter of principal component analysis.

If we do factor analysis with varimax rotation, then the output is as follows:
Rotated Factor Loadings and Communalities Varimax Rotation

Variable	Factor1	Factor2	Factor3	Factor4	Communality
coffee	0.336	0.807	0.018	−0.095	0.774
Tea	−0.233	0.752	0.330	0.370	0.866
Biscuits	0.502	0.124	0.712	−0.177	0.806
Powder	0.317	0.856	0.047	−0.230	0.889
Potatoes	0.595	0.047	0.060	0.485	0.595
Frozen fish	0.118	−0.100	0.050	0.918	0.869
Apples	0.832	0.284	0.251	0.097	0.846
Oranges	0.903	0.148	0.004	0.036	0.839
Butter	−0.004	0.089	0.900	0.172	0.847
Variance	2.3961	2.0886	1.4969	1.3480	7.3296
% Var	0.266	0.232	0.166	0.150	0.814

Factor Score Coefficients

Variable	Factor1	Factor2	Factor3	Factor4
coffee	0.038	0.408	−0.144	−0.040
Tea	−0.311	0.456	0.119	0.319
Biscuits	0.165	−0.141	0.506	−0.252
Powder	0.026	0.426	−0.109	−0.144
Potatoes	0.253	−0.047	−0.089	0.331
Frozen fish	0.006	−0.019	−0.072	0.692
Apples	0.339	−0.008	0.045	0.006
Oranges	0.431	−0.064	−0.129	−0.026
Butter	−0.132	−0.080	0.674	0.029

We see that according to percentage of variation about 80% variation is explained by first four components (as in case of PCA). But here the advantage id unlike PCA we can physically explain the factors. According to rotated factor loading, we can say that the first factor is composed of 'apples, oranges, and potatoes,' similarly the other three factors are composed of 'coffee, tea, and powder soup,' 'butter and biscuits,' and 'potatoes and frozen fish,' respectively.

Except Potato there is no overlapping, and the groups are well defined and may correspond to types of customers preferring 'fruits,' 'hot drinks,' 'snacks' and 'proteins, vitamins, and minerals.'

The most significant difference between PCA and factor analysis is regarding the assumption of an underlying causal structure. Factor analysis assumes that the covariation among the observed variables is due to the presence of one or more latent variables known as factors that impose causal influence on these observed variables. Factor analysis is used when there exit some latent factors which impose causal influence on the observed variables under consideration. Exploratory factor analysis helps the researcher identify the number and nature of these latent factors. But principal component analysis makes no assumption about an underlying causal relation. It is simply a dimension reduction technique.

References and Suggested Readings

Albazzas, H., & Wang, X. Z. (2004). *Industrial & Engineering Chemistry Research*, *43*(21), 6731.
Babu, J., et al. (2009). *The Astrophysical Journal*, *700*, 1768.
Chattopadhyay, A. K., & Chattopadhyay, T. (2014). *Statistical methods for astronomical data analysis*., Springer series in astrostatistics New York: Springer.
Chattopadhyay, A. K., Chattopadhyay, T., Davoust, E., Mondal, S., & Sharina, M. (2009). *The Astrophysical Journal*, *705*, 1533.
Chattopadhyay, A. K., Mondal, S., & Chattopadhyay, T. (2013). *Computational Statistics & Data Analysis*, *57*, 17.
Comon, P. (1994). *Signal Processing*, *36*, 287.
Dickens, R. J. (1972). *Monthly Notices of Royal Astronomical Society*, *157*, 281.
Fusi Pecci, F., et al. (1993). *Astronomical Journal*, *105*, 1145.
Gabriel, K. R. (1971). *Biometrika*, *5*, 453.
Hastie, T., & Tibshirani, R. (2003). In S. Becker, & K. Obermayer (Eds.). *Independent component analysis through product density estimation in advances in neural information processing system* (Vol. 15, pp. 649–656). Cambridge, MA: MIT Press.
Hyvarinen, A., & Oja, E. (2000). *Neural Networks*, *13*(4–5), 411.

Chapter 11
Multidimensional Scaling

11.1 Introduction

Multidimensional scaling (MDS) is a method to display (visualize) relative positions of several objects (subjects) in a two- or (three-)dimensional Euclidean space, to explore similarities (dissimilarities) revealed in data pertaining to the objects/subjects. Such similarities (dissimilarities) refer to pairs of entities, as judged objectively in terms of physical parameters or assessed subjectively in terms of opinions. An MDS algorithm starts with a matrix of item–item similarities and ends in assigning a location to each item in a multidimensional space. With two or three dimensions, the resulting locations may be displayed in a graph or 3D visualization. Also known as Torgerson scaling, MDS is a set of statistical techniques used in information visualization.

Multidimensional scaling (MDS) is not much of an analytical tool, but is quite useful for visualizing distances/differences among, say N units with p characteristics noted for each, starting with an $n \times n$ matrix input based on ordinal or cardinal measures of similarities or dissimilarities (considering the p characteristics) and ending up in a two- or three-dimensional representation of a unit as a point.

In MDS, the entities could be different cities located at different points on the non-Euclidean surface of the earth, and flying or driving distances from one city to another could indicate dissimilarities among pairs of cities. They could be manufactured products, or services, or handwriting, or softwares, or even modes of presentation, or participants in an essay competition, or any other entities which can be presented to a judge or a panel of judges who can rate or rank these entities or assign scores in respect of some feature(s) or trait(s) which reveal similarities or distances among the entities.

In some sense, multidimensional scaling, introduced by Torgerson, may be regarded as an extension of product scaling introduced by another psychologist Thurstone. In product scaling, we consider a number k of concrete entities which are presented pair-wise to a panel of n judges, each of whom is required to prefer one entity within a pair to the other in terms of some prespecified features. The final

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_11

data appear as a $k \times k$ matrix in which the (i, j)th element is p_{ij} = proportion of judges preferring entity j to entity i. Using the Law of Comparative Judgment and the method of discriminal dispersion, a scale value is finally found out for each entity so that their relative positions can be shown as points on a real line. In multidimensional, the entities are displayed as points in a two- (or at most three-) dimensional Euclidean plane.

Torgerson proposed the first MDS to help understand people's judgment of the similarity of members in assessment of objects. Currently, MDS finds applications in diverse fields such as marketing, sociology, physics, political science, and biology. Potential customers are asked to compare pairs of products and make judgments about their similarity. Whereas other techniques (such as factor analysis, discriminant analysis, or conjoint analysis) obtain underlying dimensions from responses to product attributes identified by the researcher, MDS obtains the underlying dimensions from respondents' judgments about the similarity of products. This is an important advantage. t does not depend on researchers' judgments. The underlying dimensions come from respondents' judgments about pairs of products. Because of these advantages, MDS is the most common technique used in perceptual mapping.

MDS pictures the structure of a set of objects from data that represent or approximate the distances between pairs of the objects. The data are also called similarities or dissimilarities. In addition to human judgment, the data can be objective similarity measures like driving or flying distances between pairs of cities or an index calculated from multivariate data, e.g., proportion of agreement in the votes cast by pairs of senators. Each object is represented by a point in a multidimensional space—Euclidean or not.

11.2 Types of MDS

MDS is a generic term and includes many types, classified according to the nature of dissimilarity data being qualitative (in non-metric MDS) or quantitative (in metric MDS). Further, we can have classical MDS with one dissimilarity or distance matrix and no weights assigned to object pairs, replicated MDS using several distance matrices but without any weights attached to the different matrices or replicated MDS using different weights for the different distance matrices.

In classical metric MDS, data are dissimilarities, complete (no missing entries), and symmetric. Measurements are at the ratio level. The Euclidean distances in D are so determined that they are as much like the dissimilarities in S. The common approach is

$l(S) = D + E$ where $l(S)$ is a linear transformation E is a matrix of residuals. Elements in D are functions of coordinates; the aim is to determine the coordinates of X so that the sum of squares of elements in E is minimized, subject to suitable normalization of X. Torgerson's method does not actually minimize this sum of squares.

11.2.1 Non-metric MDS

Sometimes, it may be realistic to assume a less stringent relationship between the observed distances or dissimilarities d_{ij} and the true distances δ_{ij} such that $d_{ij} = f(\delta_{ij} + e_{ij})$ where e_{ij} represents errors of measurements, distortions, etc. Also, we assume that $f(x)$ is an unknown monotonically increasing function. And the purpose could be to retain the relative order among the individuals. The data could be at the ordinal level of measurement. In addition, the matrix S could be either complete or incomplete, symmetric or otherwise, and pertain to similarities or dissimilarities. Non-metric MDS extends the distance model to the Minkowski case and enables defining.

$m(S) = D + E$ where $m(S)$ is a monotonic transformation of the similarities. With dissimilarities $m(S)$ preserves order, and with similarities it reverses order. Thus, in non-metric MDS we need to find a monotonic (order-preserving) transformation and the coordinates of X which together minimize the sum of squares of errors in E (after normalization of X). This is a much more complex optimization problem.

11.2.2 Replicated MDS

RMDS uses the same distance model as CMDS but considers several distance matrices to visualize locations of the individuals. In metric RMDS, we consider the representation $l_k\{S_k\} = D + E_k$ where the left-hand side gives a linear transform of the kth similarity (distance) matrix S_k which best fits the distances D. The data could be in ratio or interval scale, and the analysis minimizes the sum of squared elements in all the error matrices E_k simultaneously. In case of a non-metric RMDS, the representation becomes $m_k\{S_k\} = D + E_k$ where $m_k\{S_k\}$ is the monotonic transformation of the similarity matrix S_k that provides a least-square fit to the distances in the matrix D. It may be pointed out that RMDS tests all the matrices as being related to each other (through D) by a systematic linear or monotonic transformation (except for a random error component). Jacobowitz used RMDS to study how language development takes place as children grow to adulthood. The study involved a good number of children in each of several age groups as judges and parts of the human body as the objects to be ranked for their closeness.

11.2.3 Weighted MDS

In classical MDS (CMDS), we start with a symmetric, complete (with no missing entry) matrix of distances $D = ((d_{ij}))$ and try to approximate these distances by Euclidean distances δ_{ij} between points on a two- or three-dimensional plane, so that

the sum of squares of standardized deviations between these two sets of distances (often called stress) is minimized. In non-metric MDS, ranks of the distances are reproduced increasing function.

Whereas RMDS only accounts for individual differences in response bias, WMDS incorporates a model to account for individual differences in the fundamental perceptual or cognitive process that generate the responses. For this reason, WMDS is often called *individual differences scaling* (INDSCAL) and is often regarded as the second major breakthrough in multidimensional scaling.

WMDS invokes the following definition of weighted Euclidean distance:

$$d_{ij} = \left[\sum_{\alpha} w_{k\alpha} \left(x_{i\alpha} - x_{j\alpha} \right)^2 \right]^{\frac{1}{2}}$$

RMDS generates a single distance matrix D, while WMDS generates m unique distance matrices D_k, one of each data matrix S_k. The distances D_k are calculated so that they are all as much like their corresponding data matrices S_k as possible.

Thus, for WMDS, we need to solve for the matrix of coordinates X, the m diagonal matrices of weights Wk, and the m transformations Mk or 1. We wish to do this so that the sum of squared elements in all error matrices, E, is minimal subject to normalization constraints on X and Wk.

11.2.4 Sammon Mapping

Sammon mapping is a generalization of the usual metric MDS. Sammon's stress (to be minimized) is

$$\left[1 / \sum_{j<k} d_{jk} \right] \left[\sum_{i<j} \frac{(\delta_{ij} - d_{ij})^2}{d_{ij}} \right].$$

This weighting system normalizes the squared errors in pair-wise distances by using the distance in the original space. As a result, Sammon mapping preserves the small d_{ij}, giving them a greater degree of importance in the fitting procedure than for larger values of d_{ij}.

Optimal solution is found by numerical computation (initial value by CMDS). Sammon mapping better preserves interdistances for smaller dissimilarities, while proportionally squeezes the interdistances for larger dissimilarities.

11.3 MDS and Factor Analysis

Multidimensional Scaling (MDS) has been sometimes regarded as an alternative to factor analysis (FA), and important contributions to both were made initially by

psychometricians. The common goal of analysis either is to identify meaningful underlying dimensions that can explain observed similarities or dissimilarities between objects (or variables). In factor analysis, such similarities among objects or descriptive variables are indicated by the correlation matrix. In MDS, we can analyze any kind of similarity or distance matrix, and not only correlation matrix.

Despite the common objective, MDS and FA use basically different methods. While FA is an inferential tool, MDS is primarily an exploratory tool. In FA, we assume the data to follow a multivariate normal distribution and the interrelations among the variables to be linear. MDS does not require any of these assumptions. As long as similarities or dissimilarities can be rank ordered, MDS can be validly applied. Judged by their outputs, FA rings out more factors (dimensions) than MDS. MDS with two dimensions usually takes results in easily interpretable solutions. MDS can be applied to look at any kind of distances or similarities, while FA requires the computation of a correlation matrix. In fact, distances or similarities among objects in MDS can be direct assessments as perceived by the subjects, while in FA we have to first rate the perceived distances or proximities in terms of several attributes (for which FA is carried out). MDS proceeds directly from proximities or distances, while FA starts from a set of descriptive variables and a correlation matrix for these variables. Thus, MDS has a much wider applicability compared to FA.

The MDS algorithm involves an optimization exercise explicitly to minimize the overall standardized difference between the original and reproduced distances (sometimes called the stress). Factor analysis implicitly use some optimization exercise to estimate factor loadings.

In MDS, the focus is on individuals who are to be displayed as points in a plane, while factor analysis is focused on variable features (scores) and on factors to explain intercorrelations among them. Thus in the classical setup of psychological tests, factor analysis starts with a matrix of interitem or intertest correlations based on scores on m items or tests obtained by, say, n individuals or testees and proceeds to explain these correlations or similarities in terms of loadings of the different items or test on some k underlying latent factors. Possibly, though not conventionally, each item or test can be displayed as points on this k-dimensional plane. In MDS, however, we like to reveal pair-wise dissimilarities among the individuals in terms of score differences or rank differences on each item or test separately. We can use replicated MDS to eventually obtain n points on a two-dimensional or three-dimensional plane to display relative positions of the n individuals.

11.4 Distance Matrix

As in all multivariate analysis, we start with the matrix $X(n \times p)$ where the element x_{ij} denoted the value of the jth feature/characteristic for the ith unit/individual. We can develop a distance matrix $D(n \times n)$ considering only the kth feature where the element d_{ij} can be taken as $|x_{ik} - x_{jk}|$. We can generate such a matrix by considering the total or average of these absolute differences or by taking absolute differences

between the total and mean scores. In the non-metric approach, we can first convert any column of the data matrix to a vector of ranks of the n units according to the particular feature and then construct an $n \times n$ matrix of absolute rank differences. Thus, we can generate p distance matrices corresponding to the p features We can, alternatively, develop one single absolute average rank difference.

In any case, we start with a single distance matrix or a set of distance matrices.

Elements in the starting distance matrix could be differences in ranks or scores assigned to pairs of objects w.r.t. some features or properties Depending on the nature of the elements, distances could be normed to make them comparable with the Euclidean distances between points representing the objects which could be defined over unit lengths of the coordinate axes; e.g., if n objects are ranked by a judge, the maximum difference in ranks for a pair could be n-1, while for two points in a unit square the maximum distance between two points could be 21/2. In a unit cube, this would be 31/3. Using these two as divisors, we can make original differences/distances for objects in a pair with which we start and the (Euclidean) distances between points as reproduced by MDS comparable.

If we start with similarities or affinities δ_{ij} between two objects or entities i and j, we can deduce dissimilarities d_{ij} by choosing a maximum similarity $c \geq max\delta_{ij}$ and taking $d_{ij} = c - \delta ij$ for i and j different; and d_{ij} is zero otherwise. One apparent problem will arise with the choice of c, since the ultimate picture will vary from one choice to another, which is an undesirable situation. However, non-metric MDS takes care of this problem and even $cMDS$ and Sammon mapping fare pretty well in this context.

Distance, dissimilarity, or similarity (or proximity) are defined for any pair of entities (objects) in space. In mathematics, a distance function (that provides a distance between two given objects) is also called a metric $d(X, Y)$, satisfying

$$(i)\, d(X, Y) \geq 0; \quad (ii)\, d(X, Y) = 0\, if\ and\ only\ if\ X = Y;$$

$$(iii)\, d(X, Y) = d(Y, X); \quad and \quad (iv)\, d(X, Z) \leq d(X, Y) + d(Y, Z).$$

A set of dissimilarities need not be distances or, even if so, may not admit to interpretation as Euclidean distances.

The starting distances could be physical distances (driving or flying) between cities, taken in pairs, within a region like a continent or a big country, as are shown by points on the plane of a Cartesian map or on the surface of a globe. These distances are those on a non-Euclidean surface, and we may like to locate the cities as points on an Euclidean plane where the Euclidean distances between any two cities will be as close as possible to the actual distances between them.

Moving to a completely different area, we can consider the example of perception of color in human vision (studied by Ekman 1954). A total of 31 persons were asked to rate on a five-point scale from 0 (no similarity at all) to 4 (identical) each of 14_{c_2} pairs of 14 colors differing only in their hues (i.e., wavelengths from 434 to 674 m). The average rating over 31 persons for each pair (representing similarity) was then

scaled and subtracted from 1 to represent dissimilarities, resulting in the following dissimilarity matrix ($d_{ij} \times 100$) for all $i > j$.

Higher / Lower	434	445	465	472	490	504	537	555	584	600	610	628	651
445	14												
465	58	50											
472	58	56	19										
490	82	78	53	46									
504	94	91	83	75	39								
537	93	93	90	90	69	38							
555	96	93	92	91	74	55	27						
584	98	98	98	98	93	86	78	67					
600	93	96	99	99	98	92	86	81	42				
610	91	93	98	100	98	98	96	96	63	26			
628	88	89	99	99	99	98	98	97	73	50	24		
651	87	87	95	98	98	98	98	98	80	59	38	15	
674	84	86	97	96	100	99	100	98	77	72	45	45	32	24

Ekman could show two color clusters lying vertically on a two-dimensional plane one with eight colors to the left and the other with six to the right. In regard to the second coordinate, distances between points in the second cluster were smaller than those in the first cluster.

11.5 Goodness of Fit

It is expected that the fit between the original distances and the Euclidean distances reproduced by multidimensional scaling on a lower- dimensional space will depend on the number of dimensions (usually two or three) retained in the procedure and the method used for dimension reduction. It is thus important to use a measure of goodness of fit that will help us in choosing the appropriate number of dimensions of the space on which the units will be shown as points. Some of the possible measures could be based on.

Squared Differences: This is the sum of the squared differences between the given dissimilarities and the reproduced distances. The magnitude will depend on how large are the given dissimilarities and hence cannot be used to compare across situations. In fact it is not an index, similar to the adjusted R^2 in regression analysis which indicates the percentage of the sum of squared differences (corrected for the mean) accounted for by a certain number of dimensions. A value above 80 percent is expected for a good fit.

Pseudo R-squared: This is an index similar to the adjusted R^2 in regression analysis.

Stress or Normalized Stress: This is the most widely accepted measure of goodness of fit and is defined as

$$S = \left[\sum_{i<j}(d_{ij} - \delta_{ij})^2\right] \Big/ \left[\sum_{i<j} d_{ij}^2\right]$$

where d and δ correspond to the given (actual) and the reproduced distances, respectively. Stress could also be defined by replacing the denominator by the sum of squared reproduced distances. Smaller the value of stress, better the fit. It is obvious that MDS is able to reproduce the original relative positions in the map as stress approaches zero. Kruskal (1964) suggested the following advice about the stress values.

Stress	Goodness of Fit
0.200	Poor
0.100	Fair
0.50	Good
0.25	Excellent
0	Perfect

More recent articles caution against using such a table like this, since acceptable values of stress depend on the quality of the distance matrix and the number of objects as also the number of dimensions used.

11.6 An Illustration

Consider a set of eight brands of TV sets and get their performance ranks given by a judge, based on several aspects of performance like picture clarity, sound control. We can involve several judges and take the average ranks.

It is also possible that for each set we get the proportion of judges who consider the given set as the best.

For the first case, suppose we have

Set	1	2	3	4	5	6	7	8
Rank	7	4	1	5	2	8	6	3

We can now construct the absolute rank difference or distance matrix as follows.

Set	1	2	3	4	5	6	7	8
1	0	3	6	2	5	1	1	4
2	3	0	3	1	2	4	2	1
3	6	3	0	4	1	7	5	2
4	4	1	4	0	3	3	1	2
5	5	2	1	3	0	6	4	1
6	1	4	7	3	6	0	2	5
7	1	2	5	1	4	2	0	3
8	4	1	2	2	1	5	3	0

• We have to find out the coordinates of eight points on a two-dimensional plane so that the matrix of Euclidean distances between pairs of points and the matrix just now obtained is a minimum. One simple way is to consider the sum of squared differences between these two distance measures over all possible pairs divided by the sum of the original distances as the criterion of fit and the best-fitting representation is one in which this sum (stress) is minimized.

11.7 Metric CMDS

Starting with a matrix of distances or dissimilarities $D = ((d_{ij}))$ among pairs of n objects or subjects (to be compared) in respect of some feature or features, MDS comes up with a matrix of estimated distances $D' = ((d'_{ij}))$, and estimation being done to minimize the difference between D and D" is as small as possible. MDS produces a map of the n objects usually in two dimensions so as to preserve the relative positions of the objects. In classical or metric MDS, the original metric or distance is reproduced, whereas in non-metric MDS, ranks of the distances are reproduced. Going by a two-dimensional map to be provided by MDS, each object will be eventually reproduced by a pair of coordinates or a point on a scatter plot from which we can visually assess the distance between the two points in each pair which can be conveniently taken as the Euclidean distance

$d'_{ij} = \sqrt{\ } \Sigma(x_{ik} - x_{jk})^2$ (summed over k = 1, 2) where x_{ik} and x_{jk} are the coordinates of the points corresponding to objects i and j.

To work out values of d_{ij} amounts to determining values of x_{ik} and $x_{jk}, k = 1, 2, \ldots n$ so that we get points on a two-dimensional plane in such a way that d'_{ij} "is as close as possible to d_{ij}." The task is not just to find out distances d_{ij}s but the coordinates. This is given by the following algorithm

1. From **D** calculate $\mathbf{A} = \{-1/2d_{ij}^2\}$.
2. From **A** calculate $\mathbf{B} = \{a + ij - a_{i0} - a_{0j} + a_{00}\}$, where a_{i0} is the average of all a_{ij} across j.
3. Find the p largest eigenvalues $\lambda_1 > \lambda_2 > \lambda_3 \cdots > \lambda_p$ of B and corresponding eigenvectors $\mathbf{L} = \{L_{(1)}, L_{(2)}, \ldots, L_{(p)}\}$ which are normalized so that $L_{(i)}/L_{(i)} = \lambda_I$. We are assuming that p is selected so that the eigenvalues are all relatively large and positive.
4. The coordinate vectors of the objects are the rows of **L**.

References and Suggested Readings

Carroll, J. D., & Chang, J. J. (1970). *Psychometrika*, *35*, 238–319. (A key paper: Provides the first workable WMDS algorithm and one that is still in very wide use. Generalizes singular value (Eckart-Young) decomposition to N-way tables).

Jacobowitz, D. (1973). *Development of semantic structures*. Unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill.

Kruskal, J. B. (1964). *Psychometrika*, *29*, 1–27; 115–129. (Completes the second major MDS breakthrough started by Shepard by placing Shepard's work on a firm numerical analysis foundation. Perhaps the most important paper in the MDS literature).

Kruskal, J. B., & Wish, M. (1977). *Multidimensional Scalling*. Beverly Hills, CA: Sage Publications. (Very readable and accurate brief introduction to MDS that should be read by everyone wanting to know more).

Richardson, M. W. (1938). *Psychological Bulletin*, *35*, 659–660.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to Multidimensional Scaling*. New York: Academic Press.

Shepard, R. N. (1962). *Psychometrika*, *27*, 125–140; 219–246. (Started the second major MDS breakthrough by proposing the first nonmetric algorithm. Intuitive arguments placed on firmer ground by Kruskal).

Takane, Y., Young, F. W., & de Leeuw, J. (1977). *Psychometrika*, *42*, 7–67. (The third major MDS breakthrough. Combined all previous major MDS developments into a single unified algorithm).

Torgerson, W. S. (1952). *Psychometrika*, *17*, 401–419. (The first major MDS breakthrough).

Young, F. W. (1981). *Psychometrika*, *46*, 357–388. (A readable overview of nonmetric issues in the context of the general linear model and components and factor analysis).

Young, F. W. (1984). In H. G. Law, C. W. Snyder, J. Hattie, & R. P. MacDonald (Eds.). *Research methods for multimode data analysis in the behavioral sciences*. (An advanced treatment of the most general models in MDS. Geometrically oriented. Interesting political science example of a wide range of MDS models applied to one set of data).

Young, F. W., & Hamer, R. M. (1994). *Theory and applications of multidimensional scaling*. Hillsdale, NJ: Eribaum Associates. (The most complete theoretical treatment of MDS and the most wide ranging collection of).

Chapter 12
Social and Occupational Mobility

12.1 Introduction

The best way of quantifying human populations is by classifying their members on the basis of some personal attribute. One may classify families according to where they reside or workers by their occupations. Thus to study the dynamics of social processes, it is natural to start by looking at the movement of people between categories. Since such moves are largely unpredictable at the individual level, it is necessary for a model to describe mechanism of movement in probabilistic terms. The earliest paper in which social mobility was viewed as stochastic processes appears to be that of Prais (1955a, b). Since then, their has been grown up a large literature. A distinction has to be made between intergenerational mobility and intra-generational mobility. The former refers to changes of social class from one generation to another. Here the generation provides a natural discrete time unit. This phenomenon is usually called social mobility. Intra-generational mobility refers to changes of classes which take place during an individual's life span. This type of movement is called occupational or labor mobility since it is usually more directly concerned with occupations. Many deterministic and stochastic models have been developed to study social and occupational mobility situations in the different parts of the world. Several empirical studies of mobility have been published.

To study the movement of individuals over occupational categories, it is natural to start by looking at the movement of people between different categories and also at the process of recruitment of new entrants. Since such moves are largely unpredictable at the individual level, it is necessary to find a model to describe the mechanism of movement in probabilistic terms.

Study on occupational mobility is an important part of manpower planning. Such studies always help different organization, institutes, companies to properly build up their future plans regarding the number of new recruits and also help their staff members to properly plan their career. Different organization having the same setup may use the same model to study the promotion pattern. All such studies together are really helpful for proper manpower planning of the country.

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_12

Studies related to the dynamics of social systems whose members move among a set of classes are of great importance for manpower planning. In manpower planning, the classes represent grades whose sizes are fixed by the budget or amount of work to be done at each level. Recruitment and promotion can only occur when vacancies arise through leaving or expansion. The stochastic element in such processes occurs due to loss mechanisms. Individuals leaving or moving create vacancies which generate sequence of internal moves. There may also be randomness in the method by which vacancies are filled. Development of such models has been done using replacement theory. Originally, such problems have arisen in connection with the renewal of human population through deaths and births. In recent years, the main application has been in the context of industrial replacement and reliability theory. The key random variables in all cases are the length of time an entity that remains active in a particular grade.

Let us start with mobility models and some related measures. There are several models and measures based on Markov chain. Prais (1955a, b) was probably the first author to apply Markov chain theory to social mobility. The society is characterized by the transition probability matrix P, and most of the measures proposed are based on this matrix. Some examples are listed in Matras (1960). In a completely immobile society, 'son' will have the same class as their father and P will be an identity matrix. Prais (1955a, b) defined a perfectly mobile society as one in which the 'son's' class is independent of his/her'father's.' For such a society the rows of P will be identical. A third situation can be identified as extreme movement in which every 'son' has a different class from his/her 'father.'

Bartholomew (1982) proposed an idea of social mobility based on the matrix P and the elements of π (vector giving the limiting distribution of the population among the classes).

A measure of generation dependence can be developed by considering the extent to which a son's class depends on that of his/her father's. A method of doing this is suggested by considering spectral representation of P in the form

$$P = \sum_{r=1}^{k} \theta_r A_r$$

The matrices $\{A_r\}$ constitute the spectral set. The coefficients $\{\theta_r\}$ are the eigenvalues of P and since P is stochastic $1 = \theta_1 \geq |\theta_2| \geq \cdots \geq |\theta_k|$

A measure proposed by both Shorrocks (1978) and Sommers and Conlisk (1979) is based on the second largest, in absolute values, of the θ_s. If it is denoted by θ_{max}, then the measure in $\mu_1(P) = \theta_{max}$.

Bartholomew (1982) proposed two other measures given by

$$\mu_2(P) = \frac{1}{k-1} \sum_{r=2}^{k} \theta_r$$

$$\mu_3(P) = |\theta_2\theta_3 \ldots \theta_k|^{\frac{1}{(k-1)}}$$

By regarding the distribution of the population at times t and $(t+1)$ as two multinomial populations, by using Bhattacharyya distance (1945–46) a measure of divergence has been suggested by Mukherjee and Basu (1979) as below

$$cos\Delta = \sum_i \sqrt{\pi_i^{(t)} \pi_i^{(t+1)}}$$

where $\Delta = 0$ if $P = I$

By considering measures of association as inverse measure of mobility, Mukherjee and Chattopadhay (1986) proposed a number of measures based on different coefficients of association. They also proposed another measure based on minimum discrimination information statistic $(m\ d\ i\ s)$ given by

$$J(1,2) = \sum_{j=1}^{k} (\pi_j^{(t)} - \pi_j^{(t+1)}) log_e \left(\frac{\pi_j^{(t)}}{\pi_j^{(t+1)}} \right)$$

Occupational mobility refers to the movement of employees between jobs or job categories. For job changes over different organizations, the time interval between successive changes is likely to be random. As a result, for such situations simple Markov model does not provide a satisfactory representation. Attempts have been made to describe occupational mobility patterns in terms of semi-Markov processes (Ginsberg 1971, 1972). Bartholomew (1982) suggested one measure based on the transition probability matrix and the stochastic process $\{m(T)\}$ where $m(T)$ is the random number of decision points in the interval $(0, T)$. Mukherjee and Chattopadhay (1986) developed one measure based on renewal process. Starting with semi-Markov process, Mukherjee and Chattopadhyay proposed another measure in terms of the number of occupation changes during an interval of time. The same authors have also considered the situations where the job categories may be ordered in some sense.

Chattopadhyay and Gupta (2003) considered a discrete time Markov process where states of the system denote grades of the employees in an organization. The total size of the system is fixed. The recruitment needs are determined by the losses, together with any change in the size of the system. A specific version of the model with a fixed total size is due to Young and Almond (1961) who applied the model to the staff structure of a British University. The proposed model has been developed to study the career prospect on the basis of the staff categories and promotion pattern for non-teaching staff members of University of Calcutta.

Chattopadhyay and Khan (2004) has extensively studied the nature of job changes of staff members of University of Southern Queensland, Australia, on the basis of stochastic model. Khan and Chattopadhay (2003) have also derived corresponding prediction distribution on the basis of job offers received by the employees. Such type of works are very useful to investigate the manpower planning condition in different organization.

12.2 Model 1

The present model has been developed on the basis of the staff categories and promotion pattern for non-teaching staff members of University of Calcutta. Suppose that the members of the organization are divided into k strata (grades) in which there are r strata where direct appointments from outside are allowed (together with the promotion of existing staff members) and in the remaining $(k - r)$ strata posts no new appointment from outside is allowed. Only internal staff members are promoted to those positions. Let $n_i(t)$ denote the number of people in the first type of category i at time t $(i = 0, 1, 2, \ldots, r)$ and $z_j(t)$ denote the number of people in the second type of category j at time $t (j = r + 1, r + 2, r + 3, \ldots, k)$. The initial grade sizes are assumed to be given. Also let

$$N(t) = \sum_{i=1}^{r} n_i(t) + \sum_{j=r+1}^{k} z_j(t) \tag{12.2.1}$$

$$N(0) = \sum_{i=1}^{r} n_i(0) \tag{12.2.2}$$

where $N(0)$ is the total number of first type vacancies available. For $t > 0$, the grade sizes are random variables. Let us denote by $e(t)$ the number of new entrants in the system at time t and by p_{ij}, the probability of transition from grade i to grade j for an individual. These transition probabilities are assumed to be time homogeneous. The system from 1 to r be open system and grade (r + 1) to k be closed system. The allocation of new entrants in the system is specified by p_{oj} which gives the expected proportion of new entrants to the jth grade $(j = 0, 1, 2, \ldots, r)$. Here, we also assume that a person only moves to the next grade.
when $j \leq r - 1$,

$$E(n_j(t+1)) = \sum_{i=1}^{r-1} p_{ij} E(n_i(t)) + e(t+1)\ p_{oj}, \quad t = 1, 2, 3 \ldots$$
$$j = 1, \ldots, r-1 \tag{12.2.3}$$

when $j = r$,

the expected value of $n_j(t + 1)$ has been divided into two parts, one part due to changes by promotion and new appointment (n_{1_r}) and other part only due to promotion (n_{2_r}),

$$E(n_{1_r}(t+1)) = p_{r-1r} E(n_{r-1}(t)) + e(t+1) p_{0r} \quad t = 1, 2, 3 \ldots \tag{12.2.4}$$

$$E(n_{2_r}(t+1)) = p_{rr} E(n_r(t)) \tag{12.2.5}$$

$$E(n_r(t+1)) = E(n_{1_r}(t+1)) + E(n_{2_r}(t+1)) \tag{12.2.6}$$

when $j \geq (r+1)$,

$$E(z_j(t+1)) = \sum_{i=r+1}^{k} p_{ij} E(z_i(t)) \quad t = 1, 2, 3 \ldots$$
$$j = r+1, \ldots, k \tag{12.2.7}$$

12.2.1 Some Perfect Situations

Let us define the following two perfect situations regarding promotion.

I. Perfect promotion situation

Under this situation, a particular individual has the equal chances of moving into two successive categories.

$$P^{k\times k} = \begin{pmatrix} 1/2 & 1/2 & 0\ldots & 0 \\ 0 & 1/2 & 1/2\ldots & 0 \\ \vdots & & & \\ 0 & \ldots & 1/2 & 1/2 \\ 0 & \ldots & \ldots & \ldots 1 \end{pmatrix}$$

II. No promotion situation

Under this situation, a particular individual has no chance of promotion.

$$P^{k\times k} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & & \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

12.2.2 Possible Measures of Career Pattern

The extent to which an individual changes his/her job from one category to another higher category can be measured by using different indices.

A measure can be defined as a continuous function $M(.)$ defined over the set of transition matrices $\mathcal{P}$ such that

$$0 \leq M(P) \leq 1 \quad \text{for all} \quad P \in \mathcal{P}.$$

For this, the function $M(.)$ requires no significant constraint on the set of potential measures since a change of origin and scale can always be found such that the transformed variables take values within the chosen interval. The function $M(.)$ is monotonic in nature because the probability of movements between grades are given by the off-diagonal elements of the transition matrix. The increasing off-diagonal elements indicates the higher level of mobility among the career pattern of individuals and hence

$$M(P) > M(P') \text{ when } P > P' \tag{12.2.8}$$

12.2.3 Measure of Career Pattern Based on Mahalanobis Distance

We introduce a new measure of occupational mobility in terms of distance of two populations. Here,we consider occupational situation at time t as one population and at time (t+1) as another. A common distance measure is *Mahalanobis distance*.
Let $\upsilon = (X_1(t), X_2(t), X_3(t), \ldots, X_k(t))'$

$$\upsilon \sim Multinomial(N(t), \pi_1(t), \pi_2(t), \ldots, \pi_k(t)), \sum_{i=1}^{k} \pi_i(t) = 1;$$

$$uY = (Y_1(t+1), Y_2(t+1), Y_3(t+1), \ldots, Y_k(t+1))'$$

$$uY \sim Multinomial(N(t+1), \pi_1(t+1), \pi_2(t+1), \pi_3(t+1), \ldots, \pi_k(t+1)), \sum_{i=1}^{k} \pi_i(t+1) = 1;$$

where,

$X_i(t)$ = Number of persons belonging in category i at time t;
$\pi_i(t)$ = Prob. of a person belonging in category i at time t;
N(t) = Total Number of persons in entire system at time t;
$Y_i(t+1)$ = Number of persons belonging in category i at time (t + 1);
$\pi_i(t+1)$ = Prob. of a person belonging in category i at time (t + 1);
$N(t+1)$ = Total Number of persons in entire system at time (t + 1);

$$E(uX) = N(t)u\pi(t)$$

$$E(uY) = N(t+1)u\pi(t+1)$$

$$uA_1 = \begin{pmatrix} \pi_1(t)(1-\pi_1(t)) & -\pi_1(t)\pi_2(t) & -\pi_1(t)\pi_3(t)\ldots\ldots & -\pi_1(t)\pi_{k-1}(t) \\ -\pi_1(t)\pi_2(t) & \pi_2(t)(1-\pi_2(t) & -\pi_2(t)\pi_3(t)\ldots & -\pi_2(t)\pi_{k-1}(t) \\ \vdots & \vdots & \vdots & \vdots \\ -\pi_1(t)\pi_{k-1}(t) & \ldots & \ldots & -\pi_{k-1}(t)(1-\pi_{k-1}(t)) \end{pmatrix}^{(k-1)\times(k-1)}$$

Let, $t' = t + 1$;

$$uA_2 = \begin{pmatrix} \pi_1(t')(1-\pi_1(t')) & -\pi_1(t')\pi_2(t') & -\pi_1(t')\pi_3(t')\ldots & -\pi_1(t')\pi_{k-1}(t') \\ -\pi_1(t')\pi_2(t') & \pi_2(t')(1-\pi_2(t')) & -\pi_2(t')\pi_3(t')\ldots & -\pi_2(t')\pi_{k-1}(t') \\ \vdots & \vdots & \vdots & \vdots \\ -\pi_1(t')\pi_{k-1}(t') & \ldots & \ldots & \pi_{k-1}(t')(1-\pi_{k-1}(t')) \end{pmatrix}^{(k-1)\times(k-1)}$$

$$M - D = (u\pi(t) - u\pi(t+1))' uS^{-1}(u\pi(t) - u\pi(t+1)). \tag{12.2.9}$$

where

$$uS^{(k-1)*\times(k-1)} = uA_1 + uA_2 \tag{12.2.10}$$

Here actually we measure the shifting of the mean of population to study the mobility pattern.

12.2.4 Measure of Career Pattern Based on Entropy

Entropy as defined in a thermodynamical context arises naturally as additive quantity. Under this setup, probabilities are multiplicative. It can be shown that if the entropy S is a function of the probability P of a state, then S must be proportional to lnP. When we come to consider information as a function of probability, the same kind of relationship will apply.

Information is a statistical property of the set of possible messages, not of an individual message. If the probability of occurrences of symbol i in a system is p_i, Kendall (1973) observed the following requirements for a measure H of 'information' produced, which is continuous in the p_i. He then showed that only measure confirming to these requirements is

$$H_1 = -const \sum_{i=1}^{n} p_i ln(p_i) \tag{12.2.11}$$

where p_i is the probability of a person belonging to the ith category.

Under the present setup, an appropriate measure on the absolute difference between the entropies of the classifications (distributions) corresponding to t and $t + 1$ is defined as,

$$E = |\sum_{i=1}^{k} \pi_i(t) ln(\pi_i(t)) - \sum_{i=1}^{k} \pi_i(t+1) ln(\pi_i(t+1))| \qquad (12.2.12)$$

$M - D$ and E can be estimated by replacing the $\pi(t)$ and $\pi(t+1)$ values by their corresponding MLE. $M - D$ and E are exactly equal to zero under no promotion situation. Under perfect promotion situation, the values may be obtained by using the relation $u\pi(t+1) = u\Pi' u\pi(t)$. The exact value will depend on the estimated values of $u\pi(t)$ and the number of categories. $M - D$ measure completely depends upon the data set and on the distribution of population. But E measure depends upon only data set. The value of $M - D$ for perfect promotion situation also depends upon the data set, but the value of E under the perfect promotion situation does not depend upon the data set and it is always equal to unity.

12.2.5 An Example

Consider the following real-life example on the non-teaching stuff of University of Calcutta officer class of the year 1990 and 2000. This was a six-grade hierarchical system. Here, r = 3, i.e., direct appointment be allowable upto third category. The estimated transition probability matrix from the flow data for the years 1990–91 and 2000–01 is as follows:

$$P = \begin{pmatrix} .7 & .3 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & .5 & .5 & 0 & 0 \\ 0 & 0 & 0 & .834951 & .165049 & 0 \\ 0 & 0 & 0 & 0 & .457143 & .542857 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}^{(6\times 6)}$$

$$\hat{\pi}(t) = (0.134, 0.109, 0.369, 0.276, 0.093, 0.276, 0.093, 0.016)'$$

$$\hat{\pi}(t+1) = (.093, .1501, .184987, .41555, .088472, .067024)'$$

$$P_1 = \begin{pmatrix} .7 & .3 \\ 0 & 1 \end{pmatrix}^{2\times 2}$$

$$P_2 = \begin{pmatrix} .834951 & .165049 & 0 \\ .457143 & .542857 & \\ 0 & 0 & 1 \end{pmatrix}^{3\times 3}$$

$$p_{33} = .5$$

$$p_{34} = .5$$

$$p_{23} = 0$$

For the above example, the values of $M - D$ and E measures as well as the values of those measures under perfect promotion and no promotion are given in table:

	Observed	Perfect Promotion	No promotion
$\widehat{M-D}$	0.05	0.085	0
$\widehat{E}$	0.024201	0.0468460	0

Since the observed value of $M - D$ and E measures is near to the values under perfect promotion situation, it may be inferred that the chances of promotion are very high.

12.3 Model 2

The following measure was developed by Chattopadhyay and Khan (2004). Suppose that the service life of a person be comprised of k intervals of equal fixed width, t. The person gets at least one job offer within each such interval. The worth of an offer being determined by the associated salary (reward). The individual (assumed to be in service already) decides to leave the present job or not, at the end of each interval. One moves to a new job for the first time at the end of an interval in which the maximum of the remunerations associated with different job offers (within that interval) exceeds a fixed amount. This is the minimum wage at which the individual is willing to enter the job market for the first time. Subsequently, one changes the current job at the end of a particular interval only when the maximum of the wages associated with the offers received during that interval exceeds the wage of the current job. A change of job in this paper means that an individual may move from one occupation to another or within the same occupation. Let the individual gets N_i new job offers in the ith interval and let X_{ij} be the salary corresponding to the jth job offer in the ith interval, for $j = 1, 2, \ldots, n_i$, and $i = 1, 2, \ldots, k$. Note that to reflect the real-life situation, it is necessary to assume that n_i is strictly greater than zero since none can enter into the job market without a job offer. Both X_{ij} and N_i are assumed to be independently and identically distributed with pdf $g(x)$, $0 < x < \infty$, and pmf $h(y)$, $y = 1, 2, \ldots, \infty$, respectively. Define

$$Z_i = max(X_{i1}, X_{i2}, \ldots, X_{in_i}). \tag{12.3.13}$$

Here Z_i is the maximum wage of all job offers during the ith interval. Since Z_i is the largest order statistic, for a given n_i, the pdf of the conditional distribution of Z_i is

$$f(z_i|n_i) = n_i[G(x_{ij})]^{n_i-1} g(z_i)$$

where $G(\cdot)$ is the cdf of the distribution of X_{ij}. Hence, the distribution of Z_i is given by

$$f(z_i) = \sum_{n_i=1}^{\infty} n_i [G(z_i)]^{n_i - 1} g(z_i) h(n_i) \tag{12.3.14}$$

where $g(\cdot)$ and $h(\cdot)$ have the same specifications as before.

Let $F_{Z_i}(z)$ denote the corresponding cdf. Let z_0 be the minimum wage for which the individual accepts the first job offer at the i^{ih} interval. Then we can define

$$F_{Z_i}(z_0) = P[Z_i < z_0] \tag{12.3.15}$$

and its complement

$$\bar{F}_{Z_i}(z_0) = 1 - F_{Z_i}(z_0) = P[Z_i > z_0]. \tag{12.3.16}$$

Chattopadhyay and Khan defined a measure of occupational mobility as below. Define $N(k) =$ total number of job changes within the service life of the individual and $p_r^{(k)} =$ the probability of r job changes in the entire service life of the individual. Then

$$p_r^{(k)} = P[N(k) = r]. \tag{12.3.17}$$

A measure of occupational mobility using $p_r^{(k)}$ can be defined as

$$\mathrm{E}[N(k)] = \sum_{r=0}^{k} r p_r^{(k)} = [(k+1)\bar{F} - \bar{F}F^k - F(1 - F^k)]/2\bar{F}. \tag{12.3.18}$$

In the computation of $\mathrm{E}[N(k)]$, different binomial and geometric series are involved. After normalization with respect to k, the measure becomes $\mathrm{E}[N(k)/k]$.

References and Suggested Readings

Bartholomew, D. J. (1982). *Stochastic models for social processes* (3rd ed.). New York: Wiley.

Chattopadhay, A. K., & Baidya, K. (2001). Study of career pattern and promotion of Individuals in an open system. *Calcutta Statistical Association Bulletin*, *51*(1), 201–202.

Chattopadhyay, A. K., & Gupta, A. (2003). A model based study on the career prospects of individuals in an Indian university. *Journal of Statistical Research*, *37*(2), 231–239.

Chattopadhyay, A. K., & Khan, S. (2004). A statistical model of occupational mobility- A salary based measure (with Shahjahan Khan). *Hacettepe Journal of Mathematics and Statistics, Turkey*, *33*, 77–90.

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions*. New York: Wiley.

Ginsberg, R. B. (1971). Semi-Markov processes and mobility. *The Journal of Mathematical Sociology*, *1*, 233–262.
Ginsberg, R. B. (1972). Critique of probabilistic models: Application of the semi-Markov model to migration. *The Journal of Mathematical Sociology*, *2*, 63–82.
Kendall, M. G. (1973). Entropy, probability and information. *International Statistical Review*, *1*.
Khan, S., & Chattopadhay, A. K. (2003). Predictive analysis of occupational mobility based on number of job offers. *Journal of Applied Statistical Science*, *12*(1), 11–22.
Matras, J. (1960). Comparison of intergenerational occupational mobility patterns. *An Application to the Formal Theory of Social Mobility, Population Studies*, *14*, 163–169.
Mukherjee, S. P., & Basu R. (1979). Measures of social and occupational mobility. *Demography India*, *VIII*(1–2), 236–246.
Mukherjee, S. P., & Chattopadhay, A. K. (1986). Measures of mobility and some associated inference problems. *Demography India*, *15*(2), 269–280.
Prais, S. J. (1955a). Measuring social mobility. *Journal of the Royal Statistical Society Series A*, *118*, 56–66.
Prais, S. J. (1955b). The formal theory of social mobility. *Population Studies*, *9*, 72–81.
Shorrocks, A. F. (1978). The measurement of mobility. *Econometrics*, *46*, 1013–1024.
Sommers, P. M., & Conlisk, J. (1979). Eigenvalue immobility measures for Markov Chains. *The Journal of Mathematical Sociology*, *6*, 169–234.
Young, A., & Almond, G. (1961). Prediction distribution of staff. *Computer Journal*, *3*, 246–250.

Chapter 13
Social Network Analysis

13.1 Introduction

In a finite [small or large] reference population of individuals or households [HHs], in socioeconomic census or surveys, we primarily deal with individuals or HHs as 'responding units' and the designated questionnaire is geared toward the respondent and to the HH he/she belongs to. In social network analysis (SNA), the questionnaire involves, among other matters dealing with the specific responding individual/HH, specific queries dealing with *pairs of HHs* in the reference population. The word 'query' is contextual and has a very broad interpretation and use. It is the *dyadic* nature of the query involving pairs of individuals or HHs, i.e., societal units whatsoever, in a reference societal population that builds a social network [SN]. This dyadic relationship, marked by the 'direction of flow,' is what SNA in concerned with. Hence, the proper setting for studying social networks is what is called a directed graph, abbreviated as digraph, illustrated below with a small hypothetical example.

In a miniature form of a population consisting of only ten mostly marginal farmer HHs, a question about their financial needs and inter-dependence at times of financial crisis was raised. Specifically, every farmer was asked to respond to the question: During the last farming season, did you approach any of your fellow-farmers for any kind of financial help/support? The answers are summarized below.

In the above table, a 'Yes' response corresponding to $Row\ i$ and $Column\ j$ is to be interpreted as $HH\ i\ has\ sought\ help\ from\ HH\ j$ in some manner during the reference farming season. Note that whereas HH 1 has approached HH 5, HH 6, HH 7, HH 10, these four HHs did not necessarily 'reciprocate' and only HH 6 and HH 10 got to HH 1 for meeting their needs.

The above exhibits interesting features of what is termed as a social network. There are one-way, i.e., 'directed' help/support relations, two-way or 'symmetrical' or 'reciprocal' relations, and no [in either direction] relation at all. We are referring to pairs of HHs and their interrelations so far as this particular query is concerned.

In graph–theoretical terminology, the above response profiles may be viewed as generating a directed graph with one-way and two-way 'ties' or 'arrows' or 'arcs'

© Springer Nature Singapore Pte Ltd. 2018

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_13

connecting the HHs which are represented as vertices of the graph. In the above, we have ten vertices and of them $Vertex\ \#\ 2$ has a special role in the graph. It is what is termed as 'isolated vertex,' and it separates itself from the rest of the vertices. This particular HH does not approach any other HH nor any other HH does approach the head of $HH\ \#\ 2$. Such type of vertices are designated as 'isolates.' There are sociological explanations for existence of such HHs distancing themselves from the rest in a community. Otherwise, there are one-way relations as also two-way or symmetrical relations.

The 'out-degree' of a vertex [HH] is the number of other vertices [HHs] approached by the specific reference vertex. Likewise 'in-degree' of a vertex is the number of other vertices [HHs] approaching the reference vertex [HH].

In the above, out-degree of $HH\ \#\ 1$, for example, is 4 while its in-degree is 5. Sometimes, out-degree is characterized by 'Expansiveness' of the vertex [HH] while in-degree is characterized by its 'Popularity.' We may denote by E_i and P_i the out-degree and in-degree of HH_i. Again a reciprocal relation is characterized by the presence of a two-way or symmetrical tie.

We may also introduce the notion of an 'incidence matrix' of the underlying digraph. Let $I(i, j) = 1$ if there is an arc originating at i and terminating at j. In other words, $I(i, j) = 1$ if and only if HH_i reaches out to HH_j. Similarly, $I(r, s) = 1$ if and only if HH_r reaches out to HH_s. A reciprocal tie between HH_i and HH_j exists if and only if $I(i, j) = I(j, i) = 1$. Such incidence patterns give rise to the incidence matrix $I = ((I(i, j)))$ of order n where n is the number of vertices in the digraph, i.e., number of HHs in our study. It readily follows that

$$E_i = \sum_j I(i, j);\ P_j = \sum_i I(i, j), s(i, j) = I(i, j) \times I(j, i).$$

In the above, $s(i, j)$ is the score attached to the pair of HHs i and j in terms of the existence of a symmetrical relation between the two. This means: $s(i, j) = 1$ if and only if $I(i, j) = I(j, i) = 1$; otherwise, $s(i, j) = 0$.

Average out-degree is defined as $\bar{E} = \sum E_i/n$ and similarly, average in-degree is defined as $\bar{P} = \sum P_i/n$. On the other hand, average reciprocity is defined as

$$\bar{R} = \sum_i \sum_{j>i} 2s(i, j)/n(n-1).$$

In a digraph, these three quantities [average out-degree, average in-degree, and average reciprocity] are known as 'key parameters.' It is interesting to note that average out-degree = average in-degree, irrespective of the nature of the digraph.

In the literature, however, graph–theoretical [deterministic] and statistical [stochastic] models have been introduced and discussed at length. The validity of the statistical models has also been examined with reference to observed networks in terms of tests of goodness of fit and other valid statistical tests based on relevant data arising out of the networks.

Also discussed at length in the literature are graph–theoretical and statistical measures and methods. These cover (i) local measures of ego-centric characteristics; (ii) local-cum-global measures of ego-centric and global characteristics, and (iii) global characteristics. Collectively, reciprocity, cohesion (density), expansiveness (out-degrees), popularity (in-degrees) and power, connectedness and fragmentation (strong and weak components), reachability, cliques, centrality and hierarchy **and** some such measures.

We will now pass on to the sampling aspect of a digraph and examine methods of estimation of the above three key parameters of a given digraph—assumed to have arisen out of a large number of vertices [HHs]. Sampling from a finite [labeled] population of HHs and thereby developing tools and techniques for estimation of the key parameters is regarded as a very important aspect of study of a digraph.

It has to be understood that for large populations, it is not at all an easy task to enumerate all the HHs and compile data on above types of networks [without any non-response or any reporting errors whatsoever!]. Random sampling of some of the HHs for collecting necessary data seems to be a viable alternative as it can be conducted competently and more cautiously to avoid any misreporting or non-response. By doing so, we are not make any attempt whatsoever to mimic the population network and create a prototype. This is simply not possible. However, for some of the population parameters, we may attempt to provide 'reasonably accurate' estimates based on a sample network.

In the next section, we propose to discuss some aspects of sampling and inference in the context of digraphs.

13.2 Sampling and Inference in a SN

In a social network, we are dealing with queries involving 'dyads,' i.e., based on pairs of HHs. In a way, therefore, single-HH-related information is not of much direct relevance. For large villages, i.e., villages involving a large number of HHs, collection of relevant and reliable dyadic data from all pairs of the HHs is quite prohibitive and may invite missing/incomplete dyadic elements as well.

Instead, if we are interested in some specific descriptive features of the network [such as the average out-degree or average in-degree, the average reciprocity, for example], sample survey techniques, adequately applied, may provide relevant information based on a sample of HHs, suitably chosen and surveyed.

With reference to a social network, sampling of population units and estimation of dyadic parameters are extremely fascinating topics. This area of survey sampling research, though very much different from standard survey topics, has created enough interest among survey statisticians. It is impressive indeed to note that several researchers have formulated general estimation problems involving dyads [and triads, too] and applied random sampling techniques for proposing solutions to the problems. We will briefly discuss some basic results and present them in the right framework.

Following Cochran (1977), Frank (1977a, b, 1978), and Thompson (2006), we now intend to discuss some aspects of sampling and inference in a social network. In the language of sampling, we may refer to the HHs [or vertices in a digraph] as 'sampling units,' or, simply as units. To conceptualize, let us think of a network having a total of N population units with an incidence matrix I. We refer to this network as a *population network* and imagine a situation where it is *not* possible to enumerate the whole of it in terms of the out-degrees and in-degrees of all the units in the population. In such a situation, we may take recourse to a *sample*, comprising of, say n sample units, where $n << N$. We assume that all the N units in the population network are *identifiable* and may be labeled as $1, 2, \ldots, N$. Therefore, selection of a sample of n units should be a routine task. We may use *simple random sampling without replacement* procedure [abbreviated henceforth as $SRSWOR(N, n)$ procedure] and come up with a random sample of n units out of the total of N population units. According to this procedure, the units can be selected one by one at random and without replacement, each time from the rest of the population units; or else, the units can be selected all at a time. Vide Cochran (1977), for example. Our presentation is at a basic level, and it involves specific parameters of the network. Complicated sampling methodologies are to be found in Goswami et al. (1990) and Sinha (1977), among others.

In case of 'complete enumeration' of the population, i.e., hundred percent selection of the units [like HHs, in the context of a village network], we collect complete and accurate information regarding the flow of out-degrees and in-degrees associated with each unit. This, in its turn, leads to the conceptualization, visualization, and construction of the incidence matrix I, mentioned above, in its totality. From this, in principle, the entire network can be drawn and all graph–theoretical, sociological, and statistical measures can be ascertained. However, as has been pointed above, more often than not, we may end up with rather discouraging scenarios involving missing or unreliable dyadic components, unless we are dealing with small-size networks.

While illustrating the notions and basic results of sampling and inference for network data, we have conceptualized the network of a hypothetical population of $N = 53$ HHs.

13.3 Data Structure in a Random Sample of Units

For a finite population of $N = 53$ HHs, let us consider a random sample of $n = 10$ HHs—selected according to $SRSWOR(N, n)$ procedure. Let us serially number the sampled HHs as $1, 2, \ldots, n = 10$.

What kind of data do we extract from these 10 selected HH units?

The following possibilities may be enumerated:

Data Type − 1. Data are in the form of a submatrix of size $n \times N$, being a submatrix of the incidence matrix I, comprising of the rows labeled $i_1, i_2, \ldots, i_n$, corresponding to the units $i_1, i_2, \ldots, i_n$ included in the selected sample.

Referring to the conceptualized network of 53 HHS, this suggests that each HH selected in the sample is requested to release information in respect of his/her out-degree status with reference to all other $(N-1) = 54$ HHs in the population at large, no matter which HHs are included in the sample and which HHs are not included in the sample. There are two issues worth mentioning in this context. On a positive note, this provides the out-degree value E_i for each of the $n = 10$ sampled units—besides the detailed description of I_{ij}-values for each $j \neq i$. On the negative side, once sampling and data collection are over, we are not in a position to cross-verify the results of the statements made by the sampled HHs involving the unsampled HHs.

Data Type − 2. This time we confine to the collection of the out-degrees of the selected units without any details. Note that the out-degrees correspond to the row totals in the submatrix of the incidence matrix referred to above.

Referring to a village network, this suggests that each selected HH declares the value of his/her out-degree, without any details. Again there is no way to cross-verify the statements—even in respect of the HHs included in the sample.

Data Type − 3. Data are in the form of a matrix of size $n \times n$, being the submatrix of the incidence matrix I, comprising of the rows *and* columns based on the selected units, i.e., HHs. Note that this data set exclusively refers to the selected units only. Referring to a village network, we thus collect incidence scores for all pairs of HHs covered by the selection of a random sample of n HHs.

Naturally, in no case, we are in a position to extract the 'behavior' of the unsampled HHs in respect of their out-degree movements in the entire network. It must be noted that based on a sample of selected units, it is not our aim to 'reproduce' the entire network ! That is virtually impossible. Instead, our objective is to focus on some specific parameters of the population network and seek information on such parameters, based on the sample network.

Further to this, it may be argued that **Data Type − 2**, in a way, provides a summary of **Data Type − 1**. Also it transpires that **Data Type − 3** is least informative. It would be interesting to know how **Data Type − 3** provides information about the key parameters we are interested in, viz. population average out-degree (same as population average in-degree) and population average reciprocity.

At this point, let us clarify that there is a fundamental difference between survey sampling problems addressed in the literature and the one we are discussing here. In sample surveys, we extract information [on study variables as also on auxiliary variables] primarily from and on the individual sampling units themselves [without any regard to other sampling units] and the information relates to the selected sampling units. In case of study of a network, since our primary focus in on the dyadic relationships, selecting a sample of units and confining to these selected units only will shatter the pattern of dyadic relations among those sampled and those not sampled. This calls for different techniques while dealing with such sample networks. Frank (1977a, b, 1978) and others in a series of articles have systematically studied these sampling and inference problems involved in network sampling.

Our discussion here is based on the studies available in the literature and to start with, we keep it at a basic level for estimation of basic dyadic-relational parameters and we only present specific formulae based on the concept of simple random sampling.

13.4 Inference Procedure

The population parameters to be considered are (i) average out-degree [which is the same as the average in-degree] and (ii) average reciprocity.

We will exclusively confine to the population of $N = 53$ HHs and the sample of $n = 10$ HHs—selected according to $SRSWOR(N = 53, n = 10)$ procedure. It may be readily seen that Table 13.1 exhibits the sample network based on $Data\ Type - 3$. Since $Data\ Type - 2$ provides a summary of the explicit format of the network under $Data\ Type - 1$, we refrain ourselves from exhibiting $Data\ Type - 1$ and instead only tabulate below the summary statistics under $Data\ Type - 2$.

13.5 Estimation of Average Out-Degree Based on Data Type – 1/2

Procedure I: Upon selection of the n units according to $SRSWOR(N, n)$ procedure, for estimation of the population average out-degree, we propose the sample average out-degree based on the available data (Table 13.2).

Procedure I uses the realized E-values from the n selected sample units and the estimate of the overall $\bar{E}$ is its sample analogue. The sample mean is usually

Table 13.1 Response profiles on 'Seeking Financial Help from Neighboring HHs'

HH sl no.	HH 1	HH 2	HH 3	HH 4	HH 5	HH 6	HH 7	HH 8	HH 9	HH 10	Total (Yes)
1	–	No	No	No	Yes	Yes	Yes	No	No	Yes	4
2	No	–	No	No	No	No	No	No	No	No	0
3	Yes	No	–	No	Yes	No	No	No	Yes	No	3
4	Yes	No	No	–	No	Yes	No	Yes	No	Yes	4
5	No	No	No	No	–	No	No	Yes	No	No	1
6	Yes	No	No	No	Yes	–	No	Yes	No	No	3
7	No	No	No	Yes	No	Yes	–	No	No	Yes	3
8	Yes	No	No	Yes	No	No	Yes	–	No	Yes	4
9	No	No	No	No	Yes	No	No	Yes	–	No	2
10	Yes	No	No	No	No	Yes	No	No	Yes	–	3
Total (Yes)	5	0	0	2	4	4	2	4	2	4	27

Table 13.2 Summary of response profiles on 'Seeking Financial Help from Neighboring HHs'

Selected HH sl no.	Out-degree (E-value)
1	21
2	13
3	9
4	17
5	5
6	23
7	31
8	28
9	29
10	18
Total (Out-degree)	194

denoted by $\bar{e}_I$, without explicitly showing the units in the sample composition. As a numerically computed quantity, this is known as an 'estimate,' while the form suggested above refers to an 'estimator.' We will not distinguish between the two terms hereinafter. This estimator enjoys the property of unbiasedness. In the above, the suffix I is used for the estimate to denote its dependence on the Data Type I.

In the above example, we compute an estimate for population average out-degree as $\bar{e}_I = 194/10 = 19.4$.

Remark 13.1 We should point out at this stage that the sample of selected units does not provide any virtual representation of the original network at any rate. For some features of the population network, such as the average out-degree, for example, suitable estimates [such as its sample analogues]—enjoying some properties such as unbiasedness—are constructed based on the sample data. Of course, such an estimate would be useful if it possesses a smaller standard error [s.e.].

Below we provide a formula for computation of standard error of the sample average.

$$V(\bar{e}_I) = (1/n - 1/N)S_e^2,\ S_e^2 = \sum_i (E_i - \bar{E})^2/(N-1).$$

$$\hat{V} = (1/n - 1/N)s_e^2,\ s_e^2 = \sum_i (E_i - \bar{e}_I)^2/(n-1).$$

For the above example, we compute estimate of variance of the sample average as $(1/10 - 1/53)[4004 - 3763.6]/9 = 2.166$ and hence estimated s.e. $= 1.472$.

Remark 13.2 We find from the above that the sample size is a governing factor to control and/or reduce the s.e. This brings out a very important contrasting scenario viz., cost versus precision. Embarking on more units [i.e., increasing the sample

size n] will lead to increased cost but there is gain in precision in terms of reduced standard error of the resulting estimate. These are some of the issues extensively dealt with in the literature of survey sampling theory and methods. Vide Cochran (1977) and Hedayat and Sinha (1991), for example.

13.6 Inference Formulae for Data Type 3 Using Sample Size n

Procedure II: Upon selection of the sample of size n by $SRSWOR(N, n)$ procedure from the population of N units, for estimation of the population average out-degree, we propose the estimator

$$\bar{e}_{II} = [(N-1)/n(n-1)] \left[\sum_{i \neq j} E_{ij} \right],$$

based on the available Data of Type-3, i.e., confining to the n selected units only.

The derivation of this result rests on a general theory for unbiased estimation based on data drawn according to any method of sampling. Then we specialize to simple random sampling and derive the above result. We refer to Frank (1978) for technical details.

For the network considered above, an estimate of the population average out-degree is then computed as $(52/90)[27] = 15.6$.

It turns out that the above estimator is unbiased for the population average out-degree. Computation of its variance and the variance estimate from the sample data [of Type-3] is a routine task for a survey statistician. We refer to Frank (1977a, b, 1978) for necessary technical details. One can follow the general and specific results presented in these papers.

We present below the necessary formulae—taken from Bandyopadhyay et al. (2011)—for computation of estimated s.e.

Note that the estimator above can be expressed as

$$\bar{e}_{II} = [(N-1)/n(n-1)] \sum_{i} E_i^c$$

where E^c refers to the 'curtailed' E-value based on $Data\ Type-3$. In other words, the estimate is the sample average out-degree, 'inflated' by the factor $(N-1)/(n-1)$—out-degree computations being confined to and based on the selected HHs only.

These curtailed E^c values are already shown in the list of 10 sample HHs in Table 13.1.

Expression for the variance and its estimator follow.

We use the notation $b_{(i,j)} = b_{(j,i)} = E_{ij} + E_{ji}$,
for all ordered pairs of units $[i < j]$ in the population [as also in a random sample].

Note that the population average out-degree can be expressed as

$$\bar{E} = \left[\sum \sum_{i<j} b_{(i,j)} \right] / N,$$

and, further, its estimate is given by

$$\hat{\bar{E}} = (N-1) \left[\sum \sum_{i<j} b_{(i,j)} \right] / n(n-1),$$

which coincides with $\bar{e}_{II}$ given above. The variance of the estimate is, by definition, given by

$V(\bar{e}_{II}) = E[\bar{e}_{II}^2] - E^2[\bar{e}_{II}]$.

To find an unbiased estimate of this variance of $\bar{e}_{II}$ based on random sample data, one way is to find an unbiased estimate of the square of $\bar{E}$, say $\hat{\bar{E}}^2$. Then an unbiased estimate of the above variance is given by

$\bar{e}_{II}^2 - \hat{\bar{E}}^2$

Next note that $(\bar{E}^2)$ can be expressed as

$(1/N^2)[T_1 + T_2 + T_3]$

where

$$T_1 = \left[\sum \sum_{i<j} b_{(i,j)}^2 \right];$$

$$T_2 = \left[\sum \sum \sum_{i<j<k} (b_{(i,j)} b_{(i,k)} + b_{(i,j)} b_{(j,k)} + b_{(i,k)} b_{(j,k)}) \right];$$

$$T_3 = \sum \sum \sum \sum_{i<j<k<t} (b_{(i,j)} b_{(k,t)} + b_{(i,k)} b_{(j,t)} + b_{(i,t)} b_{(j,k)})].$$

It follows that an unbiased estimate of $(1/N^2)[T_1 + T_2 + T_3]$ is given by

$$(1/N^2)[\hat{T}_1 + \hat{T}_2 + \hat{T}_3]$$

where the estimates are based exclusively on sample data [of Type 3] viz.,

$$\hat{T}_1 = \left[\sum \sum_{i<j} b_{(i,j)}^2 \right] N^{(2)} / n^{(2)};$$

$$\hat{T}_2 = \left[\sum\sum \sum_{i<j<k} (b_{(i,j)}b_{(i,k)} + b_{(i,j)}b_{(j,k)} + b_{(i,k)}b_{(j,k)}) \right] N^{(3)}/n^{(3)};$$

$$\hat{T}_3 = \left[\sum\sum\sum \sum_{i<j<k<t} (b_{(i,j)}b_{(k,t)} + b_{(i,k)}b_{(j,t)} + b_{(i,t)}b_{(j,k)}) \right] N^{(4)}/n^{(4)}.$$

In the above, $N^{(2)} = N(N-1)$; $n^{(2)} = n(n-1)$ and so on for other higher powers.

13.7 Computations for the Hypothetical Example

We now turn back to the sample network for the example considered above. The sampled HHs as also their out-degree sets are already displayed above in Table 13.1. These arise from Data Type-3.

We list below in Table 13.3 the $b_{(i,j)}$'s defined above.

Computations yield: $[\sum\sum_{i<j} b_{(i,j)}] = 27$.

Note that an estimate of the population average out-degree is already computed as $\hat{\bar{E}}_{II} = (N-1)[\sum\sum_{i<j} b_{(i,j)}]/n(n-1) = (52)(27)/90 = 15.6$.

Next, we display in Table 13.4 computations of $\sum_{j(>i)} b_{ij}^2$ values for all $i < n$ in the sample.

Therefore, $\hat{T}_1 = [N^{(2)}/n^{(2)}][\sum\sum_{i<j} b_{(i,j)}^2] = [53 \times 52/90] \times 32 = 0.9569$.

Next, we need to compute
$\hat{T}_2 = [\sum\sum\sum_{i<j<k}(b_{(i,j)}b_{(i,k)} + b_{(i,j)}b_{(j,k)} + b_{(i,k)}b_{(j,k)})]N^{(3)}/n^{(3)}$.

Toward this, we display in Table 13.5 computations of the above type for all possible triplets.

Table 13.3 Computation of b_{ij} values for all pairs (i, j) with i < j

HH sl no.	HH 1	HH 2	HH 3	HH 4	HH 5	HH 6	HH 7	HH 8	HH 9	HH 10	Row Totals
1	–	0	1	1	1	2	1	1	0	2	9
2	–	–	0	0	0	0	0	0	0	0	0
3	–	–	–	0	1	0	0	0	1	0	2
4	–	–	–	–	0	1	1	2	0	1	5
5	–	–	–	–	–	1	0	1	1	0	3
6	–	–	–	–	–	–	1	1	0	1	3
7	–	–	–	–	–	–	–	1	0	1	2
8	–	–	–	–	–	–	–	–	1	1	2
9	–	–	–	–	–	–	–	–	–	1	1

Table 13.4 Computation of $\sum_{j(>i)} b_{ij}^2$ -values for all $i < n$ in the sample

HH sl no.	Sum of squares
1	13
2	0
3	2
4	6
5	3
6	3
7	2
8	2
9	1
Total	32

The above yields a total of 144.

Therefore,

$\hat{T}_2 = [53 \times 52 \times 51]/[10 \times 9 \times 8] \times 144 = 28, 111.20.$

Finally, for

$$\hat{T}_3 = \left[\sum\sum\sum \sum_{i<j<k<t} (b_{(i,j)}b_{(k,t)} + b_{(i,k)}b_{(j,t)} + b_{(i,t)}b_{(j,k)}) \right] N^{(4)}/n^{(4)}$$

necessary computations are shown in Table 13.6.

It turns out that the above yields a total of 204.

Hence,

$\hat{T}_3 = [53 \times 52 \times 51 \times 50]/[10 \times 9 \times 8 \times 7] \times 204 = 2, 84, 458.57.$

Therefore, it follows that $(\bar{E}^2)$ can be computed as

$$(1/N^2)[\hat{T}_1 + \hat{T}_2 + \hat{T}_3] = [0.9569 + 28, 111.20 + 2, 84, 458.57]/[53 \times 53] = 111.2747.$$

Finally, an unbiased estimate of the variance estimate is given by $15.6^2 - 111.2747 = 132.0853$.

Thus, estimated s.e. of the estimated average out-degree $= 11.4928$.

13.8 Estimation of Average Reciprocity

We now turn to the problem of unbiased estimation of the measure of reciprocity t.e., average reciprocity denoted by $\bar{R}$. For completeness, we reproduce the definition below.

We refer to the adjacency matrix $\mathbf{I} = ((I_{ij}))$ where $I_{ij} = 1$ whenever there is a tie originating at i and ending at j. Likewise, $I_{ji} = 1$ whenever there is a tie originating

Table 13.5 Computation of $\sum\sum\sum_{i<j<k}(b_{(i,j)}b_{(i,k)} + b_{(i,j)}b_{(j,k)} + b_{(i,k)}b_{(j,k)})]$

HH Sl No. in Pairs (i < j)	sum over k (> j) of products of pairs
(1, 2)	0
(1, 3)	11
(1, 4)	19
(1, 5)	12
(1, 6)	18
(1, 7)	8
(1, 8)	6
(1, 9)	2
(2, 3)	0
(2, 4)	0
(2, 5)	0
(2, 6)	0
(2, 7)	0
(2, 8)	0
(2, 9)	0
(3, 4)	0
(3, 5)	5
(3, 6)	0
(3, 7)	0
(3, 8)	1
(3, 9)	1
(4, 5)	3
(4, 6)	11
(4, 7)	8
(4, 8)	7
(4, 9)	1
(5, 6)	6
(5, 7)	1
(5, 8)	4
(5, 9)	1
(6, 7)	6
(6, 8)	4
(6, 9)	1
(7, 8)	4
(7, 9)	1
(8, 9)	3

Table 13.6 Computation of $\sum\sum\sum\sum_{i<j<k<t}(b_{(i,j)}b_{(k,t)} + b_{(i,k)}b_{(j,t)} + b_{(i,t)}b_{(j,k)})]$ for all possible quadruplets

HH Sl No. in triplets	Sum over highest suffix (t) in products of pairs
(1, 2, 3)	0
(1, 2, 4)	0
(1, 2, 5)	0
(1, 2, 6)	0
(1, 2, 7)	0
(1, 2, 8)	0
(1, 2, 9)	0
(1, 3, 4)	7
(1, 3, 5)	11
(1, 3, 6)	5
(1, 3, 7)	3
(1, 3, 8)	3
(1, 3, 9)	3
(1, 4, 5)	8
(1, 4, 6)	15
(1, 4, 7)	8
(1, 4, 8)	7
(1, 4, 9)	1
(1, 5, 6)	11
(1, 5, 7)	4
(1, 5, 8)	5
(1, 5, 9)	3
(1, 6, 7)	9
(1, 6, 8)	7
(1, 6, 9)	2
(1, 7, 8)	5
(1, 7, 9)	1
(1, 8, 9)	3
(2, -, -)	0
(3, 4, 5)	5
(3, 4, 6)	1
(3, 4, 7)	1
(3, 4, 8)	2
(3, 4, 9)	1
(3, 5, 6)	4
(3, 5, 7)	2
(3, 5, 8)	3
(3, 5, 9)	1
(3, 6, 7)	1

(continued)

Table 13.6 (continued)

HH Sl No. in triplets	Sum over highest suffix (t) in products of pairs
(3, 6, 8)	1
(3, 6, 9)	1
(3, 7, 8)	1
(3, 8, 9)	1
(4, 5, 6)	6
(4, 5, 7)	2
(4, 5, 8)	3
(4, 5, 9)	1
(4, 6, 7)	7
(4, 6, 8)	5
(4, 6, 9)	1
(4, 7, 8)	5
(4, 7, 9)	1
(4, 8, 9)	3
(5, 6, 7)	4
(5, 6, 8)	4
(5, 6, 9)	2
(5, 7, 8)	2
(5, 7, 9)	1
(5, 8, 9)	2
(6, 7, 8)	4
(6, 7, 9)	1
(6, 8, 9)	2
(7, 8, 9)	2

at j and terminating at i. Reciprocity between the pair of units i and j takes place whenever $I_{ij} = I_{ji} = 1$, i.e., whenever the units are involved in a reciprocal relation. We may denote the 'reciprocity' score for a pair of units i and j as $s_{(i,j)} = I_{ij} I_{ji}$ so that $s_{(i,j)} = 1$ whenever the units are involved in a reciprocal relation. Otherwise, $s_{(i,j)} = 0$. Therefore, average reciprocity in a population network, denoted in the above by $\bar{R}$, is the average of s-scores over all such $N(N-1)/2$ pairs of units in the population of N units

i.e., $\bar{R} = \sum \sum_{i<j} 2s_{(i,j)}/N(N-1)$.

We now embark upon the problem of unbiased estimation of this population parameter $\bar{R}$ based on a sample network of size n.

We propose an estimate of population $\bar{R}$ by its sample analogue $\bar{R}(s)$ based on the average of s-scores of $n(n-1)/2$ pairs of units in the sample denoted by s. It follows that the estimate $\bar{R}(s)$ is unbiased. Its variance computation is a routine but non-trivial exercise. Further, deriving an expression for an estimate [based on

sample reciprocity scores] of the variance of the estimator so derived is also a highly non-trivial exercise. We omit the derivations altogether and simply state the results.

Below $T(R)$ denotes the total reciprocity score in the population as a whole, i.e., $T(R) = \sum \sum_{i<j} s_{(i,j)}$ so that $\bar{R} = 2T(R)/N(N-1)$. Moreover, while providing sample-based estimators, we use the notation s for the selected sample. Thus, for example, while $i < j$ refers to all pairs of population units, $i < j \in s$ refers to all pairs of units in the selected sample s.

$$(1) T(R) = \sum \sum_{i<j} s_{(i,j)};$$

$$(2) \hat{T}(R) = N(N-1) \left[\sum \sum_{i<j \in s} s_{(i,j)} \right] / n(n-1);$$

$$\begin{aligned}(3) T^2(R) = &\left[\sum \sum_{i<j} s^2_{(i,j)} + 2 \sum \sum \sum_{i<j<k} (s_{(i,j)}s_{(i,k)} + s_{(i,j)}s_{(j,k)} + s_{(i,k)}s_{(j,k)}) \right. \\ &\left. + 2 \sum \sum \sum \sum_{i<j<k<t} (s_{(i,j)}s_{(k,t)} + s_{(i,k)}s_{(j,t)} + s_{(i,t)}s_{(j,k)}) \right] \\ &= [T_1 + T_2 + T_3];\end{aligned}$$

$$(4) \hat{T}_1 = N(N-1)T_1(s)/n(n-1);$$

$$\hat{T}_2 = N(N-1)(N-2)T_2(s)/n(n-1)(n-2);$$

$$\hat{T}_3 = N(N-1)(N-2)(N-3)T_3(s)/n(n-1)(n-2)(n-3);$$

$$T_1(s) = \sum \sum_{i<j \in s} s^2_{(i,j)};$$

$$T_2(s) = 2 \sum \sum \sum_{i<j<k \in s} [s_{(i,j)}s_{(i,k)} + s_{(i,j)}s_{(j,k)} + s_{(i,k)}s_{(j,k)}];$$

$$T_3(s) = 2 \sum \sum \sum \sum_{i<j<k<t \in s} [s_{(i,j)}s_{(k,t)} + s_{(i,k)}s_{(j,t)} + s_{(i,t)}s_{(j,k)}]$$

$$(5) V(\hat{T}(R)) = E[(\hat{T}(R))^2)] - T^2(R)$$

$$(6) \hat{V}(\hat{T}(R)) = [(\hat{T}(R))]^2 - (\hat{T}^2(R)) = Square\ of\ (2) - Expression\ from\ (3)\ and\ (4)$$

Table 13.7 Computation of $s_{i,j}$-values for all pairs (i, j) with i < j

HH Sl No.	HH 1	HH 2	HH 3	HH 4	HH 5	HH 6	HH 7	HH 8	HH 9	HH 10	Row totals
1	–	0	0	0	0	1	0	0	0	1	2
2	–	–	0	0	0	0	0	0	0	0	0
3	–	–	–	0	0	0	0	0	0	0	0
4	–	–	–	–	0	0	0	1	0	0	1
5	–	–	–	–	–	0	0	0	0	0	0
6	–	–	–	–	–	–	0	0	0	0	0
7	–	–	–	–	–	–	–	0	0	0	0
8	–	–	–	–	–	–	–	–	0	0	0
9	–	–	–	–	–	–	–	–	–	0	0

$$(7)\bar{R} = 2T(R)/N(N-1) = 2\sum\sum_{i<j} s_{(i,j)}/N(N-1);$$

$$(8)\hat{\bar{R}} = \bar{R}(s) = 2\sum\sum_{i<j\ \epsilon\ s} s_{(i,j)}/n(n-1);$$

$$(9)V(\hat{\bar{R}}) = 4V(\hat{T}(R))/N^2(N-1)^2;$$

$$(10)\hat{V}(\hat{\bar{R}}) = 4\hat{V}(\hat{T}(R))/N^2(N-1)^2.$$

Computations for the illustrative example are shown below (Table 13.7).

From the above table, it follows that

$$s_{1,6} = s_{1,10} = s_{4,8} = 1,$$

the rest being all 0's.

Therefore, an estimate of the population average reciprocity $[\bar{R}]$, as given by the sample average reciprocity, is computed as

$2 \times 3/10 \times 9 = 6.67$

Further to this, an estimate of total reciprocity $T(R)$ is computed as $53 \times 52 \times 0.0667 = 7$, when approximated to a whole number.

For the computation of s.e. of the estimate of the population average reciprocity, we follow the computational formulae given above. Note that $s^2_{(i,j)} = s_{(i,j)}$ since $s_{(i,j)}$ is equal to either 0 or 1. Further to this, our computations are largely simplified since only three of the s_{ij}'s are 1.

We now proceed toward computation of estimated variance using (4).

$$T_1(s) = \sum \sum_{i<j \,\epsilon\, s} s_{(i,j)}^2 = \sum \sum_{i<j \,\epsilon} s_{(i,j)} = 3;$$

$$T_2(s) = 2 \sum \sum \sum_{i<j<k \,\epsilon\, s} (s_{(i,j)}s_{(i,k)} + s_{(i,j)}s_{(j,k)} + s_{(i,k)}s_{(j,k)})$$
$$= 2 \times s_{1,6}s_{1,10} = 2;$$

$$T_3(s) = 2 \sum \sum \sum \sum_{i<j<k<t \,\epsilon\, s} (r_{(i,j)}r_{(k,t)} + r_{(i,k)}r_{(j,t)} + r_{(i,t)}r_{(j,k)})]$$
$$= 2 \times [s_{1,6}s_{4,8} + s_{1,10}s_{4,8}] = 4.$$

Therefore,

$$\hat{T}_1 = (53)(52)(3)/(10)(9) = 91.87;$$

$$\hat{T}_2 = (53)(52)(51)(2)/(10)(9)(8) = 390.4333;$$

$$\hat{T}_3 = (53)(52)(51)(50)(4)/(10)(9)(8)(7) = 5577.6190.$$

Hence,

$$\hat{T}^2(R) = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 = 6059.9223.$$

From this, we compute estimated $(\bar{R})^2$ as

$$\hat{\bar{R}}^2 = 4\hat{T}^2(R)/N^2(N-1)^2 = 0.0032.$$

Further, $\hat{\bar{R}}^2 = 0.0667^2 = 0.0044$.
Hence, finally,

$$\text{estimated variance } = 0.0044 - 0.0032 = 0.0012.$$

Thus, estimated s.e. = 0.0346.

It turns out that 95 percent confidence limits to the population average reciprocity is given by sample average reciprocity plus/minus 2 times the estimated s.e. This results into $[0.0667 - 0.0692, 0.0667 + 0.0692] = [0, 0.136]$.

Remark 13.3 It was already observed that there are only three reciprocal pairs of HHs in the random sample of 10 HHs from the hypothetical population of 53 HHs. Accordingly, the population average reciprocity has been estimated as 0.0667. Therefore, reciprocity is a 'rare' event for this population of HHs. This suggests that one

may hardly expect any reciprocal pair of HHs among the selected 10 HHs and that is indeed the case with the sample data exhibited before! In case of such 'negligible' incidence of 'rare events,' simple random sampling of a few HHs [like 10 in our example] may not produce any substantial number of reciprocal pairs.

Other sampling methods are more resourceful in such cases. One such method is the so-called inverse sampling method. We continue drawing HHs—at random and one by one—and at each stage, we compute out-degree, in-degree, and reciprocity values involving the most recent HH when included in the study. We continue sampling of HHs and doing the counting till a specified reciprocity value, say 8, is achieved or exceeded. We refer to Sinha (1977) and Goswami et al. (1990)—apart from Frank et al. papers for discussions on these useful sampling strategies in the context of dyadic networks.

References and Suggested Readings

Bandyopadhyay, S., Rao, A.R., & Sinha, B. K. (2011). *Models for Social Networks with Statistical Applications*. Sage Publication .

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, *1*, 235–264.

Frank, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, *4*, 81–89.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, *1*, 91–101.

Goswami, A., Sengupta, S., & Sinha, B. K. (1990). Optimal strategies in sampling from a social network. *Sequential Analysis*, *9*, 1–18.

Hedayat, A. S., & Sinha, B. K. (1991). *Design and inference in finite population sampling*. New York: Wiley.

Sinha, B. K. (1997). Some inference aspects of a social network. *Applied Statistical Science, II* (pp. 77–86). Commack, NY: Nova Science Publishers.

Thompson, S. K. (2006). Targeted random walk designs. *Survey Methodology*, *32*, 11–24.