Optimal Analysis of Qualitative Data with Illustrations

Optimal Analysis of Qualitative Data with Illustrations

By

Shizuhiko Nishisato

Cambridge
Scholars
Publishing

Optimal Analysis of Qualitative Data with Illustrations

By Shizuhiko Nishisato

This book first published 2024

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-0364-1297-5
ISBN (Ebook): 978-1-0364-1298-2

Contents

Preface

Someone uttered *Garbage In Garbage Out!* some 60 years ago. It was around the time when statistics became widely used in many applied areas (*e.g.,* education, psychology, sociology, ecology, anthropology, archaeology and health sciences). The above saying was meant to draw the researchers' attention to the basic premise that proves the *outputs of statistical analysis are meaningful only when the inputs are valid.* This saying is still valid, or rather more valid today than some 60 years ago.

The current book is in response to the above warning and we will explore how inputs for data analysis should carefully be examined and analyzed. Its specific concern is not with quantitative data but qualitative (non-quantitative) data, which are typically collected in applied areas of statistics. How should we process qualitative data in a quantitative way? This question is often encountered in a wide variety of survey research, where you are often asked about gender, municipality and occupation, all of which yield qualitative data and not quantitative data.

Time has changed and nowadays we are often asked to fill in a questionnaire at medical clinics, government offices or charity organizations. This has become a data-gathering world involving every one of us. The data collectors are no longer limited to researchers.

This book is then a wake-up call that researchers are often unaware of the goldmine of precious information embedded in qualitative data, and that a much better handling of the data than in the current practice is urgently needed. We should also know then that qualitative data are more difficult to analyze than quantitative analysis. Apart from this technical aspect of analysis, we would like to emphasize that qualitative data are a goldmine of complex information, tightly packed in the data. We will provide simple means of unwrapping hidden information in qualitative data into linear, nonlinear and multidimensional information in manageable ways, namely through illustrative numerical examples. The author is of the opinion that qualitative data contain much more information than quantitative data.

Our book will start with characterizations of different types of data and then will concentrate on non-quantitative data (*e.g.,* most of data collected in applied areas of statistics). We then ask if it is at all possible to introduce multidimensional coordinates to place non-quantitative data in mathematically logical ways: the answer is definitely yes. Our objective then is to carry out analysis of non-quantitative data quantitatively in multidimensional space. Nishisato (2007a) called this type of analysis *nonlinear multidimensional analysis of qualitative data.* Try to believe that *qualitative data contain more information than quantitative data.* We will see simple illustrations that prove this belief is true.

To analyze non-quantitative data quantitatively, we require a special plan for such analysis. This transition is not always simple to follow. Therefore, we will use many small numerical examples to see every step of analysis. The book is very introductory, and we hope that simple examples will clarify every step of the transition from analysis of quantitative data to that of qualitative data. Our goal is to offer an introduction to optimal analysis of qualitative data in familiar multidimensional Euclidean framework. Please observe how this transition from quantitative multivariate analysis to multidimensional analysis of non-quantitative data can be carried out. We will not only offer quantitative analysis of qualitative data but it will also be optimal analysis. Yes, our specific goal is *optimal multidimensional analysis of qualitative data.*

Although we will talk about some complex topics such as nonlinear analysis and multidimensional decomposition of data, please be assured that our discussion will remain elementary with many numerical illustrations. The main lesson will be that even non-quantitative data can be analyzed quantitatively and furthermore it can be done in a mathematically optimal way.

It is the author's wish to offer an easy discussion of non-quantitative data with numerical examples, and show how such data can optimally be analyzed. The analytical method of our discussion will be purely data-oriented, and we will illustrate the procedure in the simplest possible way. The currently available books are mostly advanced and will require prior reading of the current book. Therefore, this is the first step towards advanced analysis of qualitative data. It is hoped that this book will be a useful introduction to everyone in the wide area of data analysis.

Shizuhiko Nishisato
Professor Emeritus, University of Toronto, Canada

Acknowledgments

First of all, I would like to thank Senior Commissioning Editor Adam Rummens, Commissioning Editor Alison Duffy, Typesetter Manager Amanda Millar and staff members of Cambridge Scholars Publishing for efficiently handling all the matters of this publication. I also would like to thank the Board Members for accepting the manuscript. My experience with the publisher has all been very positive and encouraging. At the final stage, Professor Eric J. Beh of Australia offered me an invaluable technical help.

Many of my mentors whom I would like to thank are no longer with us, but it is nevertheless my wish to acknowledge their generous guidance offered to me. Among them are R. Darrell Bock (my Ph.D. thesis supervisor), Lyle V. Jones (the Director of the Psychometric Laboratory of the University of North Carolina (UNC) at Chapel Hill), Masanao Toda (my supervisor at Hokkaido University, Japan), Emir H. Shuford (my UNC academic advisor), Chikio Hayashi, Louis Guttman, Paul Horst, C. Radhakrishna Rao, John C. Gower, J. Douglas Carroll, Phipps Arabie, Yoshinori Matsuyama, Haruyo Hama, Edwin Diday, Dalbir Bindra and Yoshio Sugiyama.

Since my first academic publication in 1960, it has already been 64 years. During that period, there are so many colleagues who helped my career. In particular, I owe much to the five friends, Akinori Okada (Japan), Eric J. Beh (Australia), Rosaria Lombardo (Italy), José G. Clavel (Spain) and Pieter M. Kroonenberg (The Netherlands): Okada, as the editor of the Springer Behaviormetrics Series, helped me to publish three books in 2021, 2022 and 2023; he also conceived the idea of my Festschrift and kindly arranged its publication from Springer; Beh, Lombardo and Clavel are co-authors of our 2021 book from Springer and most importantly they devoted their precious time as expert editors of the Festschrift *Analysis of Categorical Data from Historical Perspectives. Essays in Honour of Shizuhiko Nishisato.* (Springer, 2023); and Kroonenberg, a long-time friend, wrote a scholastic preface and a chapter full of photos of many colleagues for the Festschrift, providing a prototype of what future Festschrifts should be like.

In addition to these four friends, I owe so much to many other colleagues, in alphabetical order, Kohei Adachi, Sergei Adamov, Hyung Ahn, Yasumasa Baba, Daniel Baier, Graham Bean, Hans-Hermann Bock, Ingo & Ulf Böckenholt, Giuseppe Bove, Hamparsum Bozdogan, Ruth Childs, Naohito Chino, Vartan Choulakian, Norman Cliff, Reinhold Decker, Jan de Leeuw, Wolfgang Gaul, Michael J. Greenacre,

Patrick Groenen, Gila Hanna, Willem Heiser, David Hemsworth, Takahiro, Hoshino, Lawrence Hubert, Tadashi Imaizumi, Shuichi Iwatsubo, Yu-taka Kano, Henk A. L. Kiers, Se-Kang Kim, Carlo Lauro, Daniel Lawrence, Ludovic Lebart, Kuo-Sing Leong, Brigitte Le Roux, Michael Maraun, Charles Mayenga, Jaqueline Meulman, Boris Mirkin, Oscar Milones, Masahiro Mizuta, Luis Moreno, Hans-Joachim Mucha, Takashi Murakami, Atsuho Nakayama, Maurice Odondi, S. H. Poh, James O. Ramsay, John Sachs, Gilbert Saporta, Wen-Jenn Sheu, Yoshio Takane, Yutaka Tanaka, Shigeo Tatsuki, Josh M. F. ten Berge, Michele Tenenhaus, Yukihiko Torii, Ross E. Traub, Tekefumi Ueno, Michel van de Velden, Peter van der Heijden, Jan L. A. van Rijckevorsel, Helmat Vorkauf, Hiroshi Yadohisa, Satoru Yokoyama, Ryozo Yoshino, Philip Weingarden, Ian Wiggins and Richard G. Wolfe.

This book is dedicated to my immediate families: My wife Lorraine, son Ira, his wife Samantha Dugas and my grandson Lincoln Dugas-Nishisato in Canada. My sister Michiko Soma and brother Akihiko Nishisato in Japan. I also owe so much to my late brother Tsunehiko Nishisato and his family. I cannot thank all of them enough for what they have kindly done for me for so many years.

As for my personal friends, I owe thanks to many friends and mention here only a handful of them: North-American friends (Takashi Asano, André & Gill Dugas, Clayton & Claire Ford and family, Akira Kobashigawa, Christopher Ringwalt, Setsuko Thurlow), childhood friends (Shizue Hashieda, Suketoshi Iiyama, Yasuko Miura and Toshitaka Tago), high school friends (Tetsuro Kokubu, Satoshi Kon, Hiroshi Nakamura, Tadashi Yamada), University friends (Shiro Imai, Reiko Komatsu, Jun-ichi Nakahara, Mutsumi Nakahara, Stephen J. Zyzanski), Hokkaido University ”Bon Jour” classmates (Miyuki Hasuike, Yuya Ishibashi, Reiko Kawanishi, Kouichi Murakami, Miyuki Nakamura, Hiroshi Oikawa, Hideshi Seki, Hirozumi Shibuya, Michio Takada) and friends at the Riverhouse at Old Mill in Toronto, Canada.

Best wishes as always,
Shizuhiko Nishisato (”Nishi”)

Part I

Introduction

Outline

Part 1 starts with descriptions of different kinds of data, where we will characterize data in terms of two categories, qualitative and quantitative. **Quantitative data** are what most of us are familiar with and they can directly be subjected to such statistical analysis as regression analysis, principal component analysis and the analysis of variance and covariance. Since quantitative data can be processed by the normal arithmetic operations of addition, subtraction, multiplication and division, analysis of quantitative data is most widely covered by statistics courses and there are many reference books.

We will, therefore, focus our attention on the other type of data, qualitative data, which are often collected in applied areas of statistics such as gender, districts of residence, countries of citizenship and marital statuses. As most of us know, **qualitative data** are not directly amenable to the basic arithmetic operations. Yet, qualitative data are ubiquitous in research. Furthermore, how to analyze qualitative data is not well known. If this perception is correct, there is some urgency to disseminate the pertinent knowledge among many researchers in applied areas of statistics.

We will start with Stevens' theory of measurement (Stevens, 1951), where the concept of measurement is defined and four types of measurement are explained and discussed from the analytical point of view. In terms of Stevens' terminology, we can classify nominal and ordinal measurement as qualitative data, and interval and ratio measurement as quantitative data. There is a widespread belief that quantitative data are more informative than qualitative data. But, a real intention of the current book is to convey the message that qualitative data can be more informative than quantitative data. This may be a startling and unbelievable statement to many researchers, but we believe that the statement is correct and we will prove it with numerical examples.

Because qualitative data are not quantitative, we require some pre-processing, and this task will be looked after once we glance at Stevens' theory of measurement (1951). The first question that comes to our mind is where we can find quantitative information embedded in qualitative data. The answer is that the information is embedded in joint frequencies of qualitative variables: Drinks such as tea, coffee and soda are qualitative, and age groups 1, 2 and 3 are also qualitative, but once we obtain how many people in age group 1 prefer tea to the other two, and how many people in another age group prefer coffee best, tea second best, and so on, then we will find some quantitative information in the response distribution. Once we obtain joint frequencies of age groups and the three kinds of drinks, we will be able to infer the relative positions of the qualitative variables in Euclidean space. Thus, analysis of qualitative data requires an extra step of quantification before we engage ourselves in a typical task of data analysis.

This quantification task may appear as an impossible task, but it is already solved by *quantification theory.* Our first task is to understand the nature of qualitative data and start with expressing qualitative data in terms of joint frequencies of variables and combinations of variables, which is an extra task as compared with the traditional data analysis. Once this task is done, we will introduce some fundamental principles of extracting quantitative information in a mathematically optimal way. This is a difficult topic and we need some basic principles used in data analysis such as linear combination of variables, graphical display of relations of variables, canonical reduction of a quadratic equation and principal axes of Euclidean coordinates. Some of these topics may be difficult to follow, but we will provide a simple and helpful introduction to these concepts used in our data analysis.

Through the steps of quantifying qualitative variables, we will find out such a big discovery that qualitative data are more informative than quantitative data, a fact that not many researchers may agree with. But, the readers will be convinced that the above statement is true, once we look at some examples of analysis. This may be your biggest discovery in analysis of qualitative data. Indeed, the fact that we must assign some values to non-quantitative variables give us ample opportunity to enhance the chance of capturing more information from the given non-quantitative data than from quantitative data. You will be convinced that qualitative data are more quantitatively informative than quantitative data. This will be the most important message of the book.

Measurement is defined as assignment of numerals to the data according to certain rules. How to use the knowledge of measurement effectively is a challenge for the researchers. We will be looking into the analysis of qualitative data to offer the fact that qualitative data are a goldmine of information. Thus, we should be encouraged to collect qualitative data as much as possible. By processing qualitative data in a rational way, let us avoid the *garbage-in garbage-out* frame.

Chapter 1

Measurement

There are many kinds of data such as genders, baseball players' back numbers, family sizes, weather data (temperature, atmospheric conditions (sunny, rainy, windy, foggy)), numbers of COVID patients in different provinces, vital signs (body temperature, blood pressure, heart-rate), numbers of yearly sunny days in different provinces, numbers of babies born this year in European countries, longevity statistics of the world, and so on. We can easily tell that there are many data types, some of which can directly be subjected to the arithmetic operations (*e.g.,* the number of COVID patients) and others are totally inappropriate for arithmetic operations (*e.g.,* baseball players' back numbers). Considering that data analysis is a procedure to analyze all kinds of data, it is imperative to pay some attention to the nature of data, in particular whether or not the data we collected can be processed by the normal arithmetic operations. No matter what types of data we may acquire, the task of data analysis is to process them, extract information and interpret the results. Let us start by asking what kinds of data we handle in data analysis.

1.1 Four Kinds of Measurement

It is convenient to adopt Stevens' classification of measurement (Stevens, 1951). This classification of measurement became well-known and popular in the social sciences in the 1950s. It is strange, however, that the topic of measurement seems to have gradually disappeared from the curriculum of social science education, although it is still very important. There is a famous definition of measurement by Stevens: *Measurement is assignment of numbers to objects or observations according to some rules.* He classified measurement into four types:

1. Nominal measurement
2. Ordinal measurement
3. Interval measurement
4. Ratio measurement

Let us describe each type of measurement with some examples and consider what mathematical operations are appropriate to process them.

1.1.1 Nominal Measurement

Nominal measurement is obtained or created by assigning numbers to groups, players, countries and so on for *identification purposes.* For example, numbers 1, 2, 3, 4 and 5 are assigned to five groups. These numbers are only for the identification purpose, thus they are not amenable to the operations of addition, subtraction, multiplication or division.

Example: Consider

1=Alberta, 2=British Columbia
3=Manitoba, 4=New Brunswick
5=Newfoundland and Labrador
6=Northwest Territories
7=Nova Scotia, 8=Nunavut, 9=Ontario
10=Prince Edward Island, 11=Quebec
12=Saskatchewan, 13=Yukon

These are Canadian provinces and territories. The numbers do not convey any quantitative information, but are meaningful only as identification numbers.

Nominal measurement is one of the two main data types which we will discuss how to analyze in a logical way. If we wish to process nominal measurement, our task then is to consider how to assign numbers to the original identification numbers, with the desideratum being to extract as much quantitative information as possible from the original nominal data. But how?

It is important to note that nominal measurement itself is not amenable to computation, but that the counting is one operation we can use for nominal measurement. Using counts or frequencies of nominal variables, we will transform the data so as to make them amenable to normal arithmetic operations. This may sound like an impossible task, but we will show you that it is possible and how it is done.

Remember that survey data are often full of nominal measurements (*e.g.,* gender, occupation). We cannot afford discarding them, for survey research is typically very expensive. In this book, we will amply show that nominal data are full of quantitative information to analyze. At this moment, we only mention that we can analyze nominal measurement in a mathematically optimal way. Once we realize that nominal measurement can quantitatively be analyzed, this opens up an extensive horizon for analysis of survey data.

1.1.2 Ordinal Measurement

Ordinal measurement is best represented by ranks, obtained by arranging objects in order of a certain criterion.

Example: Ask a group of people to rank the following movies according to the order of their preference from 1 (best) to the last rank (worst).

(1) Sound of Music
(2) Love in the Afternoon
(3) Manchurian Candidate
(4) King Solomon's Mine
(5) To Kill a Mocking Bird
(6) Moby-Dick
(7) Roman Holiday
(8) The Graduate

Ordinal measurement is quasi-quantitative since ordinal numbers indicate numerical orders. Because of this property, we may be tempted to assign the score of 8 to rank 1 movies out of 8 in the above list, score of 7 to rank 2 movies, and so on, to the score of 1 to rank 8 in this example. This is a typical way that ordinal measurement is handled. However, this widely used procedure is extremely problematic from the information-retrieval point of view, and if one really wants to criticize this widely used practice, one can say that this common-sense procedure will create a nasty problem in data analysis. Why?

Such a procedure drastically limits the amount of information one can extract from ordinal measurement and misleads the interpretation of the information in the original measurement. This is a substantial enough topic to write a paper on, thus we will delve into it right now.

We can now look at a few mistakes about ordinal measurement that anyone would endorse.

First, someone's rank 1 cannot be equated to another person's rank 1, for their criteria may be totally different (*e.g.,* thriller, love story, societal problem). Then, this means that we cannot compare the ranked data of one person with those of another person. What can we do then? This is a serious problem for data analysis because we typically wish to arrive at the analytical results based on the data from all the subjects.

Secondly, the distance between two adjacent ranks is not constant even within a person, not to mention those across all the respondents. For example, one person thinks that Sound of Music as number 1 is very close to Roman Holiday as number 2, which is much better than Love in the Afternoon in the third place, while another person may think that Roman Holiday is number one, much farther ahead of Love in the Afternoon in the second place, which is slightly better than Sound of Music in number 3. Unless the ranks are equally spaced the arithmetic operation of addition and subtraction would not result in numbers that we can trust and interpret.

In brief, ordinal measurement contains some quantitative information, but it is nevertheless not amenable to the basic arithmetic operations. Then, how can we subject ordinal measurement to data analysis? If we are logically minded, this will create a serious problem for analyzing ordinal measurement. But, as we will discuss later, we can develop a method to deal with ranked data or we can treat ordinal data as if they are nominal.

Nominal measurement and ordinal measurement are abundant in survey data, and they are the most problematic types for data analysis. Considering their abundance and difficult problems from the information-retrieval point of view, we have included them in the main domain of data for the current book.

1.1.3 Interval Measurement

Interval measurement is everywhere in data analysis. The main characteristic of interval measurement is that it has an *equal unit, but no rational origin.*

Examples:

IQ (intelligence quotient)
Temperature (Celsius, Fahrenheit)
Achievement test scores

The equality of the unit of measurement allows us to process interval measurement by addition and subtraction to yield some meaningful statistics. However, we should note that the *ratios of interval measurements are meaningless* because the rational origin of zero is not defined.

This point may be hard to understand, but consider, for instance, the following example: 20 Celsius does not mean that it is twice as hot as 10 Celsius. Why not? Well, consider changing these numbers to Fahrenheit. Then we find that 20 Celsius is 68 Fahrenheit and 10 Celsius is 50 Fahrenheit. See that 68 Fahrenheit is no longer twice 50 Fahrenheit.

Two more examples are that of achievement scores and IQ scores: we cannot say that an IQ of zero means no intelligence, or that the person with an IQ of 100 is twice as intelligent as the person of an IQ 50. The same applies to achievement scores. All of these erroneous statements are due to the fact the interval measurement does not have a rational origin.

It is ironical to see in education that teachers often treat achievement scores as if it is reasonable to multiply or divide them for evaluations of students. If we want to avoid the logical problem in handling educational data, consider using statistics which are specifically geared for interval measurement. A good example of an appropriate use of mathematics for interval measurement is to calculate Pearsonian correlation (ı.e., product-moment correlation) which is defined in terms of deviation scores (original scores minus the mean, where the effect of the arbitrary mean is completely removed). Similarly standard deviation, variance and correlation ratio are appropriate statistics for interval measurement because they are defined in terms of deviation scores.

1.1.4 Ratio Measurement

This measurement has the properties of *equal interval and the rational origin.* As such, ratio measurement can be processed by the arithmetic operations of addition, subtraction, multiplication and division. No transformation is necessary before processing them by mathematical operations. Thus, we can say that ratio measurement is a full-fledged quantitative measurement. Most inferential statistical procedures are geared for data of ratio measurement. Unfortunately, however, many books on statistics do not contain any discussion on measurement, a key starting point for reasonable use of statistics in data analysis.

Examples:

Inter-city distance data
Weight and height data
Rain fall and snow fall data
Economic data of income and expenditure
Agricultural crops

Since our concern is about the analysis of nominal and ordinal data, we will not expand our discussion on these last two kinds of *quantitative data*, that is, interval and ratio measurements. We will just remember that ratio measurement can immediately be subjected to quantitative analysis and that we must be careful with handling interval measurement, which does not have the rational origin. As mentioned above, however, we can still subject interval measurement for computing such statistics as the standard deviation, the variance, the correlation and the correlation ratio, all of which are based on deviation scores, quantities free from the effect of an arbitrary origin. Thus, ratio measurement and interval measurement are typically referred to as quantitative data.

Let us summarize Stevens' theory of measurement (Table 1.1) from the data analytic point of view.

In Chapter 2, we will look at the current author's belief that *qualitative data contain much more quantitative information than quan-*

Table 1.1: Information for Analysis

Data	Classification	Mathematics	Analyzable Information
Nominal	Qualitative	Identification (*e.g.*, gender)	**Frequency distribution** of nominal variables
Ordinal	Qualitative	Order (*e.g.*, ranking)	**Frequency distribution** of ordinal variables
Interval	Quantitative	Equal unit, no origin (only addition, subtraction) (*e.g.*, test scores)	measurement itself
Ratio	Quantitative	Equal unit, rational origin(all arithmetic operations (*e.g.*, distance)	measurement itself

titative data, a view that most researchers may reject as a debatable assertion. It is hoped that the readers will agree with the author's assertion and move on to find the goldmine of information in qualitative data.

1.2 Analysis of Qualitative Data

We looked at four kinds of measurement, and the readers may have discovered that qualitative data (nominal and ordinal measurement) are more difficult to analyze than others. Indeed, if we consider applying normal arithmetic operations to the analysis, qualitative data are by far more difficult than quantitative data.

Thanks to the past 100 years of research endeavor, however, we now have am ample resource of information on how to analyze qualitative data in a quantitative way. The main technique has been referred to by many aliases such as correspondence analysis, dual scaling, optimal scaling, homogeneity analysis, canonical analysis, taxi-cab correspondence analysis and so on.

Until around 1990, researchers used to carry out thorough historical surveys of available tools, often going back to as early as Gleason (1926), Lenoble (1937) and Ramensky (1930). But for some reason or other, recent studies often lack such a historical survey of the available tools. To rectify this situation, we will expose the readers to many relevant historical articles so that they may choose favorite references for quantification theory. Thus, it was decided to present as much historical information as possible to the readers. Some of the references in Chapter 17 are not specifically cited in the book but are referred to as other relevant books. Not to overwhelm the readers with a long list of publications, we will present the readers with only some representative sources of information to start with. It is hoped that the readers will find these selected ones and others in the bibliography sufficient for further study of quantification theory.

1.2.1 Useful Books

During the past 80 years or so, many researchers in diverse disciplines have published books, elementary or advanced, in different languages. Readers can choose any one of these books to know what quantification theory is for and about. We hope that this list will be greatly useful. Choose what you like and then move onto further research.

Adachi & Murakami (2011, in Japanese), Beh & Lombardo (2012), Beh, Lombardo & Clavel (2023), Benzécri (1982, 1992), Benzécri et al. (1973), Blasius & Greenacre (1998), Bouroche (1977), Carlier & Pagés (1976), Clausen (1992), Coppi & Bolasco (1980), Cuadras & Rao (1993), de Leeuw (1984), Escofier-Cordier (1969), Escofier (2003), Escofier & Pagés (1988), Gauch (1982), Gifi (1990), Gower & Hand (1966), Greenacre (1984, 1993), Greenacre & Blasius (1994, 2006), Heiser (1981), Iwatsubo (1987, in Japanese), Kiers (1989), Komazawa (1978, 1982, in Japanese), Koster (1982), Kroonenberg (1983, 2008), Lebart, Morineau & Tabard (1977), Lebart & Fénelon (1971), Lebart, Morineau & Fénelon (1971), Lebart, Morineau & Tabard (1977), Lebart, Morineau & Warwick (1984), Le Roux & Rouanet (2004), Lingoes (1973), Meulman (1982, 1986), Murtagh (2005), Nishisato (1975 in Japanese, 1980a, 1994, 2007a, 2007b in Japanese, 2010 in Japanese, 2020, 2023, 2024), Nishisato, Beh, Lombardo & Clavel (2021), Nishisato & Nishisato (1984, 1994), Phsumi, Kebart, Morineau & Baba, 1994 in Japanese), Rouanet & Le Roux (1993), Saporta (1979, 1990), Takane (1980a in Japanese), Takeuchi & Yanai (1972 in Japanese), Takeuchi, Yanai & Mukherjii (1982), Tenenhaus (1994), van Buuren (1990), van de Geer (1993), van der Burg, 1988), van der Heijden (1987), van Rijckevorsel (1987), van Rijckevorsel & de Leuuw (1988), Verboon (1994), Weller & Romney (1990) and Whittaker (1978b).

1.2.2 Useful Papers

As for individual papers, there are too many to cite here as good sources. Please refer to the bibliography, where not only those cited in the book, but also those referred to by ”others” are included for the benefit of the reader.

Throughout the book, we will find out that qualitative data contain a large amount of quantitative information, and we are here to explore this great source of information from qualitative data. Our analysis is appropriately called *optimal analysis* (Bock, 1960) of qualitative data.

Chapter 2

Exploring Gold Mine

This chapter is an introduction to analysis of qualitative data, and its aim is to demonstrate that qualitative data contain a large amount of information, or rather they are a goldmine of information.

When Stevens (1951) proposed his classification of measurement, many universities adopted the topic of measurement in their courses in psychology, sociology and education. But, it is the current author's observation as a student then that the teaching involved only an introduction to Stevens' measurement theory and it was never extended to what to do with nominal and ordinal measurement in data analysis.

The present author was lucky enough to meet Chikio Hayashi (one of the pioneers of quantification theory) in 1960 at a conference in Tokyo and then the mentor and my thesis supervisor R. Darrell Bock in 1961, who taught his famous optimal scaling course at the University of North Carolina (UNC) at Chapel Hill. In aprticular, Bock taught students how to quantify nominal and ordinal measurement, a topic of great importance that most universities ignored teaching after Stevens' theory of measurement (see Bock & Jones, 1967). During the student days in Chapel Hill, the current author also met other important contributors to quantification theory, Louis Guttman (another pioneer of quantification theory) at the Psychometric Laboratory, UNC, and Paul Horst (another pioneer) at an annual meeting of the Psychometric Society. After completing his Ph.D. thesis, Minimum Entropy Clustering of Test Items, in 1965, he chose the topic of quantifying qualitative data as his life-long project (*e.g.*, Nishisato, 1975, 1980, 1982, 1994, 2007a, 2007b, 2010, 2022, 2023; Nishisato &

Nishisato, 1994; Nishisto, Beh, Lombardo & Clavel, 2021), with the belief that qualitative data are much more informative than quantitative data, a view that not many researchers would agree with. The author's belief is a key motivation for writing yet another book.

Let us look at a numerical example where we can amply see the author's assertion about *informative qualitative data.*

In Chapter 1, we observed that even the basic arithmetic operations of addition, subtraction, multiplication and division can not justifiably be applied to qualitative data. Then, there remains an unavoidable question of where we can find information from qualitative data. Let us discuss this problem in the next section.

2.1 Frequencies: A Source of Information

As we have seen, nominal measurement carries only identification information. As such, nominal measurement itself cannot be subjected to any mathematical operations. Once we collect data, however, we will typically have repeated observations of nominal variables, that is, joint frequencies of nominal variables. For example, if we find out that there are

1. Many coffee-drinkers, many tea-drinkers and very few soda-drinkers among the senior group.

2. Many coffee-drinkers, a few tea-drinkers and a few soda-drinkers among the middle-age group.

3. A few coffee-drinkers, a few tea-drinkers and many soda-drinkers among the teenagers.

Or, more concretely, suppose that we obtain such data as in Table 2.1 where actual numbers of choice frequencies are given (Table 2.1). Then, would you not expect that this joint frequency distribution of the drinks and the age groups can tell something about their drinking preference? The answer to this query is definitely *yes.* Indeed, we can use the joint frequencies of qualitative variables to evaluate the relations between, for example, age groups and the drinks. This transition from a non-quantitative state to a quantitative state may be too hard to comprehend right now, but by the end of this book, we will find out that not only qualitative data can be quantitatively

Table 2.1: Realistic Data on Nominal Measurement

	Tea	Coffee	Soda	Total
Age 10-15	0	5	49	54
Age 40-45	8	40	6	54
Age 70-75	31	22	1	54
Total	39	67	56	162

analyzed, but also they offer us even a better way to explore the hidden information in data than quantitative data. This extra finding of quantitative information from qualitative data is what we call the gold-mine of information. We have unknown quantities given to the age groups and the three drinks which we can manipulate so as to optimize a certain statistic, for example, the correlation between the age groups and the drinks. The key is how to assign numerical values to the age groups and the drinks. We will use small examples to convince the readers that joint frequencies of qualitative variables are of vital importance for the quantitative analysis of nominal and ordinal measurement. Let us see some evidence of our assertion without discussing its mathematics, or logical foundations.

2.2 Qualitative Data and Our Objectives

Let us use the data in Table 2.1 as a sample data set. Notice that the data are obtained for nominal variables of age groups and drinks. An important point we want to make here is that we have *frequencies* of choices for the combinations of categories of two nominal variables (age groups and drinks). Since those age groups and drinks do not have any numerals, our analysis starts with assigning some unknown values to them. For example, we assign unknown values, x_1, x_2, x_3, to the three age groups, respectively, and three unknowns, y_1, y_2, y_3, to the drinks. Our task is how to determine the best values to these unknowns so that the relation between the age and the drinks may be as clearly observable as possible.

Let us glance at our table of data. From the joint frequencies, we see Age 70-75 is close to Tea and Coffee, Age 40-45 is close to Age 70-75 over Coffee and Age 10-15 is close to Soda. This assertion is made in terms of the joint frequencies of the elements of the data. Thanks to the frequency distribution, we can now see how each of one categorical variable occurs jointly with another set of the categorical variables. Our task is to *assign some values* to these two sets of the unknown variables in such a way that the relationship between the two sets of variables may maximally be revealed. Can we assign such values to these unknowns that make the correlation between the age and the drink be a maximum? The answer is of course yes.

Rather than involving us in the detailed analysis, let us just look at what qualitative data can be expressed in terms of pairs of unknown numerals to the rows and the columns of the data in Table 2.1. In terms of the unknowns, our data can be represented as in Tables 2.2 and 2.3. In these tables, there are six unknowns (three age groups and three drinks). Our task is how to determine the values of these unknowns so that the information in the data may be most efficiently and validly determined, and once this is done, we will move on to interpret the information embedded in qualitative data. So, when we look at Table 2.3, we can immediately see the difference between quantitative data and qualitative data.

Most importantly, (1) **quantitative data** have those elements of the table being all numerals, and they can directly be subjected to numerical analysis, while (2) **qualitative data** have those elements of the table being unknown quantities and our task is to determine the values of the unknowns in such a way that we can most efficiently extract information and interpret the relations between the categories of the two variables.

This manipulation of the unknowns suggests greater possibility of finding the closest relation between the two sets of variables. Therefore, we should note that the role of the unknowns is to allow us to obtain the most desirable outcome.

This is the challenging aspect of analyzing qualitative data. The unknowns present us a challenge of determining them in the best possible way. This is the task of analyzing qualitative data quantitatively.

2.3 Data Are Mostly Multidimensional

Imagine that we can manipulate those unknowns given to qualitative variables, following some mathematical rationale. One consequence of this handling of unknown quantities is that analysis is bound to lead