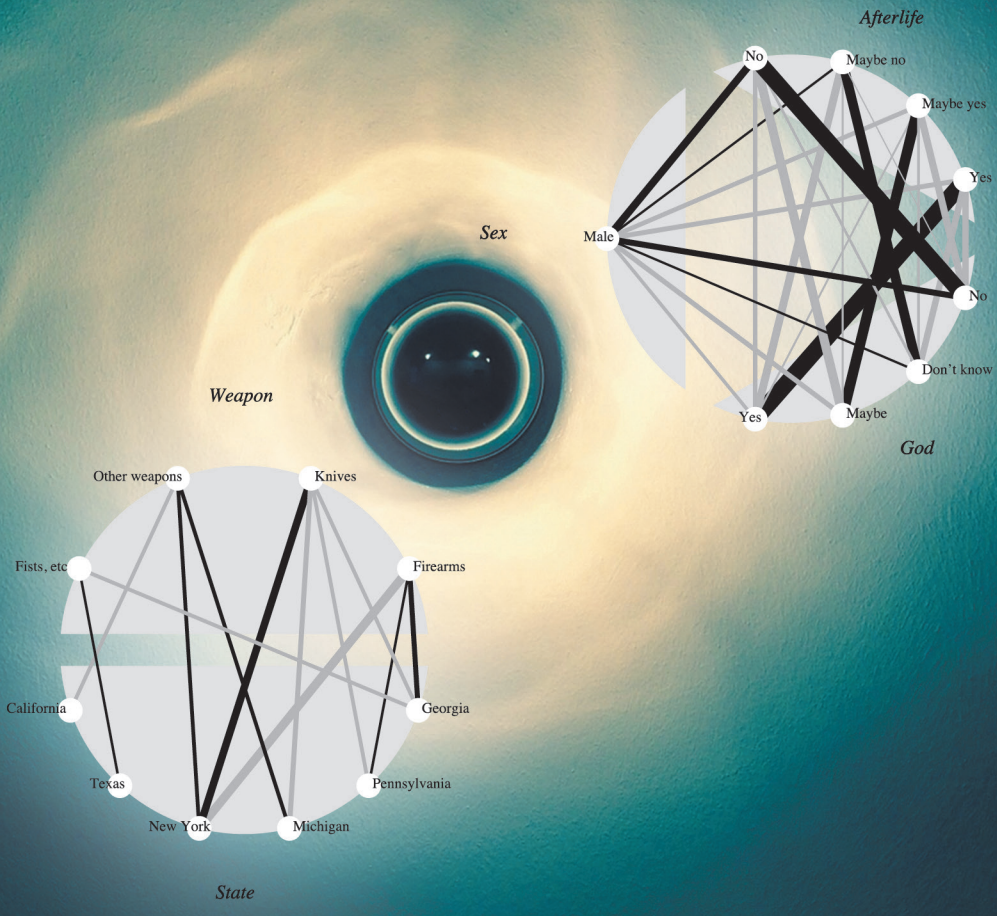
CATEGORICAL DATA ANALYSIS BY EXAMPLE

GRAHAM J. G. UPTON



WILEY

CATEGORICAL DATA ANALYSIS BY EXAMPLE

CATEGORICAL DATA ANALYSIS BY EXAMPLE

GRAHAM J. G. UPTON

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data

Names: Upton, Graham J. G., author.
Title: Categorical data analysis by example / Graham J.G. Upton.
Description: Hoboken, New Jersey : John Wiley & Sons, 2016. | Includes index.
Identifiers: LCCN 2016031847 (print) | LCCN 2016045176 (ebook) | ISBN 9781119307860 (cloth) | ISBN 9781119307914 (pdf) | ISBN 9781119307938 (epub)
Subjects: LCSH: Multivariate analysis. | Log-linear models.
Classification: LCC QA278 .U68 2016 (print) | LCC QA278 (ebook) | DDC 519.5/35–dc23
LC record available at https://lccn.loc.gov/2016031847

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

This book is aimed at all those who wish to discover how to analyze categorical data without getting immersed in complicated mathematics and without needing to wade through a large amount of prose. It is aimed at researchers with their own data ready to be analyzed and at students who would like an approachable alternative view of the subject. The few starred sections provide background details for interested readers, but can be omitted by readers who are more concerned with the "How" than the "Why."

As the title suggests, each new topic is illustrated with an example. Since the examples were as new to the writer as they will be to the reader, in many cases I have suggested preliminary visualizations of the data or informal analyses prior to the formal analysis. Any model provides, at best, a convenient simplification of a mass of data into a few summary figures. For a proper analysis of any set of data, it is essential to understand the background to the data and to have available information on all the relevant variables. Examples in textbooks cannot be expected to provide detailed insights into the data analyzed: those insights should be provided by the users of the book in the context of their own sets of data.

In many cases (particularly in the later chapters), R code is given and excerpts from the resulting output are presented. R was chosen simply because it is free! The thrust of the book is about the methods of analysis, rather than any particular programming language. Users of other languages (SAS, STATA, ...) would obtain equivalent output from their analyses; it would simply be presented in a slightly different format. The author does

not claim to be an expert R programmer, so the example code can doubtless be improved. However, it should work adequately as it stands.

In the context of log-linear models for cross-tabulations, two "specialties of the house" have been included: the use of cobweb diagrams to get visual information concerning significant interactions, and a procedure for detecting outlier category combinations. The R code used for these is available and may be freely adapted.

GRAHAM J. G. UPTON

Wivenhoe, Essex
March, 2016

ACKNOWLEDGMENTS

A first thanks go to generations of students who have sat through lectures related to this material without complaining too loudly!

I have gleaned data from a variety of sources and particular thanks are due to Mieke van Hemelrijck and Sabine Rohrmann for making the NHANES III data available. The data on the hands of blues guitarists have been taken from the *Journal of Statistical Education*, which has an excellent online data resource. Most European and British data were abstracted from the UK Data Archive, which is situated at the University of Essex; I am grateful for their assistance and their permission to use the data. Those interested in election data should find the website of the British Election Study helpful. The US crime data were obtained from the website provided by the FBI. On behalf of researchers everywhere, I would like to thank these entities for making their data so easy to re-analyze.

GRAHAM J. G. UPTON

CHAPTER 1

INTRODUCTION

This chapter introduces basic statistical ideas and terminology in what the author hopes is a suitably concise fashion. Many readers will be able to turn to Chapter 2 without further ado!

1.1 WHAT ARE CATEGORICAL DATA?

Categorical data are the observed values of variables such as the color of a book, a person's religion, gender, political preference, social class, etc. In short, any variable other than a *continuous variable* (such as length, weight, time, distance, etc.).

If the categories have no obvious order (e.g., Red, Yellow, White, Blue) then the variable is described as a *nominal variable*. If the categories have an obvious order (e.g., Small, Medium, Large) then the variable is described as an *ordinal variable*. In the latter case the categories may relate to an underlying continuous variable where the precise value is unrecorded, or where it simplifies matters to replace the measurement by the relevant category. For example, while an individual's age may be known, it may suffice to record it as belonging to one of the categories "Under 18," "Between 18 and 65," "Over 65."

If a variable has just two categories, then it is a *binary variable* and whether or not the categories are ordered has no effect on the ensuing analysis.

Categorical Data Analysis by Example, First Edition. Graham J. G. Upton.

1.2 A TYPICAL DATA SET

The basic data with which we are concerned are *counts*, also called *frequencies*. Such data occur naturally when we summarize the answers to questions in a survey such as that in Table 1.1.

TABLE 1.1 Hypothetical sports preference survey

Sports preference questionnaire
(A) Are you:- Male ☐ Female ☐?
(B) Are you:- Aged 45 or under ☐ Aged over 45 ☐?
(C) Do you:- Prefer golf to tennis ☐ Prefer tennis to golf ☐?

The people answering this (fictitious) survey will be classified by each of the three characteristics: gender, age, and sport preference. Suppose that the 400 replies were as given in Table 1.2 which shows that males prefer golf to tennis (142 out of 194 is 73%) whereas females prefer tennis to golf (161 out of 206 is 78%). However, there is a lot of other information available. For example:

- There are more replies from females than males.
- There are more tennis lovers than golf lovers.
- Amongst males, the proportion preferring golf to tennis is greater amongst those aged over 45 (78/102 is 76%) than those aged 45 or under (64/92 is 70%).

This book is concerned with models that can reveal all of these subtleties simultaneously.

TABLE 1.2 Results of sports preference survey

Category of response	Frequency
Male, aged 45 or under, prefers golf to tennis	64
Male, aged 45 or under, prefers tennis to golf	28
Male, aged over 45, prefers golf to tennis	78
Male, aged over 45, prefers tennis to golf	24
Female, aged 45 or under, prefers golf to tennis	22
Female, aged 45 or under, prefers tennis to golf	86
Female, aged over 45, prefers golf to tennis	23
Female, aged over 45, prefers tennis to golf	75

1.3 VISUALIZATION AND CROSS-TABULATION

While Table 1.2 certainly summarizes the results, it does so in a clumsily long-winded fashion. We need a more succinct alternative, which is provided in Table 1.3.

TABLE 1.3 Presentation of survey results by gender

	Male				Female		
Sport	45 and under	Over 45	Total	Sport	45 and under	Over 45	Total
Tennis	28	24	52	Tennis	86	75	161
Golf	64	78	142	Golf	22	23	45
Total	92	102	194	Total	108	98	206

A table of this type is referred to as a *contingency table*—in this case it is (in effect) a three-dimensional contingency table. The locations in the body of the table are referred to as the *cells* of the table. Note that the table can be presented in several different ways. One alternative is Table 1.4.

In this example, the problem is that the page of a book is two-dimensional, whereas, with its three classifying variables, the data set is essentially three-dimensional, as Figure 1.1 indicates. Each face of the diagram contains information about the 2×2 category combinations for two variables for some particular category of the third variable.

With a small table and just three variables, a diagram is feasible, as Figure 1.1 illustrates. In general, however, there will be too many variables and too many categories for this to be a useful approach.

TABLE 1.4 Presentation of survey results by sport preference

	Prefers tennis				Prefers golf		
Gender	45 and under	Over 45	Total	Gender	45 and under	Over 45	Total
Female	86	75	161	Female	22	23	45
Male	28	24	52	Male	64	78	142
Total	114	99	213	Total	86	101	187

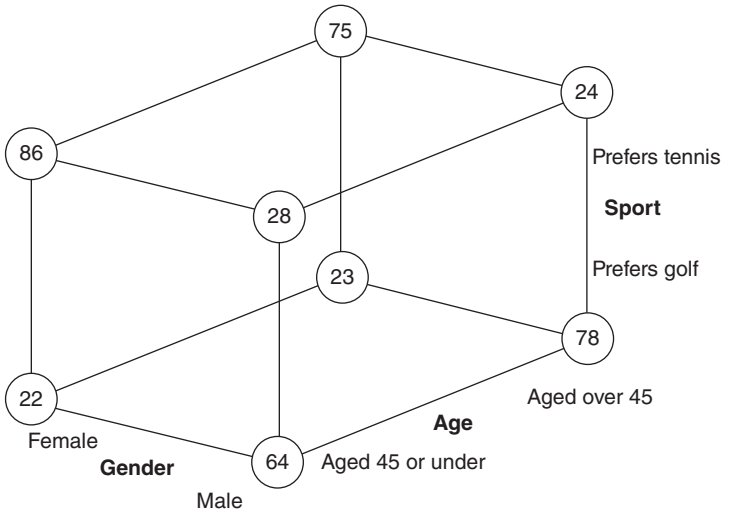


FIGURE 1.1 Illustration of results of sports preference survey.

1.4 SAMPLES, POPULATIONS, AND RANDOM VARIATION

Suppose we repeat the survey of sport preferences, interviewing a second group of 100 individuals and obtaining the results summarized in Table 1.5.

As one would expect, the results are very similar to those from the first survey, but they are not identical. All the principal characteristics (for example, the preference of females for tennis and males for golf) are again present, but there are slight variations because these are the replies from a different set of people. Each person has individual reasons for their reply and we cannot possibly expect to perfectly predict any individual reply since there can be thousands of contributing factors influencing a person's preference. Instead we attribute the differences to *random variation*.

TABLE 1.5 The results of a second survey

Prefers tennis				Prefers golf			
Gender	45 and under	Over 45	Total	Gender	45 and under	Over 45	Total
Female	81	76	157	Female	16	24	40
Male	26	34	60	Male	62	81	143
Total	107	110	217	Total	78	105	183

Of course, if one survey was of spectators leaving a grand slam tennis tournament, whilst the second survey was of spectators at an open golf tournament, then the results would be very different! These would be *samples* from very different *populations*. Both samples may give entirely fair results for their own specialized populations, with the differences in the sample results reflecting the differences in the populations.

Our purpose in this book is to find succinct models that adequately describe the populations from which samples like these have been drawn. An effective model will use relatively few parameters to describe a much larger group of counts.

1.5 PROPORTION, PROBABILITY, AND CONDITIONAL PROBABILITY

Between them, Tables 1.4 and 1.5 summarized the sporting preferences of 800 individuals. The information was collected one individual at a time, so it would have been possible to keep track of the counts in the eight categories as they accumulated. The results might have been as shown in Table 1.6.

As the sample size increases, so the observed proportions, which are initially very variable, becomes less variable. Each proportion slowly converges on its limiting value, the *population probability*. The difference between columns three and five is that the former is converging on the probability of randomly selecting a particular type of individual from the whole population while the latter is converging on the *conditional probability* of selecting the individual from the relevant subpopulation (males aged over 40).

TABLE 1.6 The accumulating results from the two surveys

Sample size	Number of males over 40 who prefer golf	Proportion of sample that are males aged over 40 and prefer golf	Number of males	Proportion of males aged over 40 who prefer golf
10	3	0.300	6	0.500
20	5	0.250	11	0.455
50	8	0.160	25	0.320
100	22	0.220	51	0.431
200	41	0.205	98	0.418
400	78	0.195	194	0.402
800	159	0.199	397	0.401

1.6 PROBABILITY DISTRIBUTIONS

In this section, we very briefly introduce the distributions that are directly relevant to the remainder of the book. A variable is described as being a *discrete variable* if it can only take one of a finite set of values. The probability of any particular value is given by the *probability function*, P.

By contrast, a *continuous variable* can take any value in one or more possible ranges. For a continuous random variable the probability of a value in the interval (a, b) is given by integration of a function f (the so-called *probability density function*) over that interval.

1.6.1 The Binomial Distribution

The binomial distribution is a discrete distribution that is relevant when a variable has just two categories (e.g., Male and Female). If the probability of a randomly chosen individual has probability p of being male, then the probability that a random sample of n individuals contains r males is given by

$$\mathrm{P}(r) = \begin{cases} \binom{n}{r} p^r (1-p)^{n-r} & r = 0, 1, \ldots, n, \\ 0 & \text{otherwise}, \end{cases} \tag{1.1}$$

where

$$\binom{n}{r} = \frac{n!}{r!(n-r)!},$$

and

$$r! = r \times (r-1) \times (r-2) \times \cdots \times 2 \times 1.$$

A random variable having such a distribution has *mean* (the average value) np and *variance* (the usual measure of variability) $np(1-p)$. When p is very small and n is large—which is often the case in the context of contingency tables—then the distribution will be closely approximated by a Poisson distribution (Section 1.6.3) with the same mean. When n is large, a normal distribution (Section 1.6.4) also provides a good approximation.

This distribution underlies the logistic regression models discussed in Chapters 7–9.