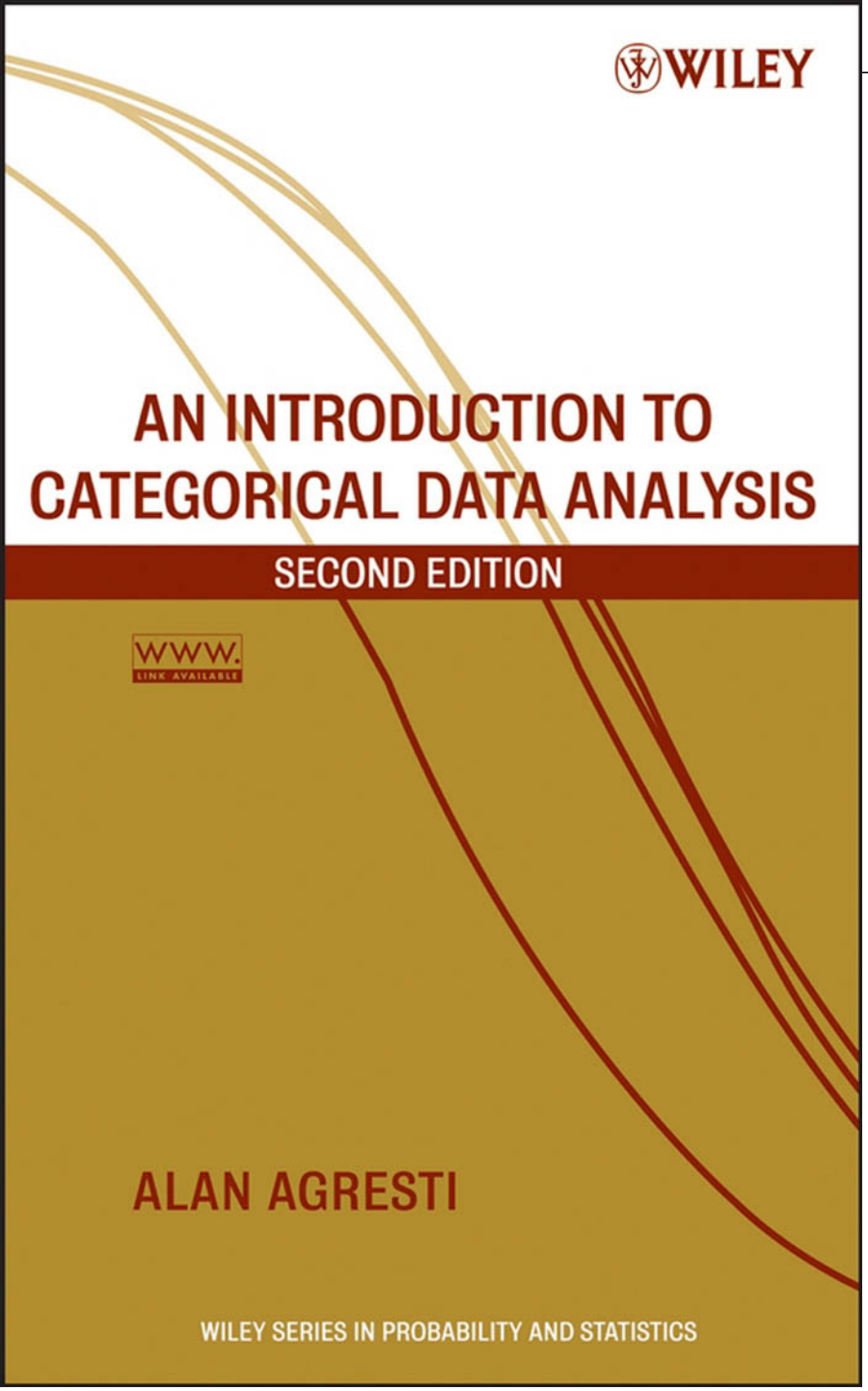

WILEY
AN INTRODUCTION TO
CATEGORICAL DATA ANALYSIS
SECOND EDITION
WWW.
LINK AVAILABLE
ALAN AGRESTI
WILEY SERIES IN PROBABILITY AND STATISTICS

An Introduction to Categorical Data Analysis

Second Edition

ALAN AGRESTI

Department of Statistics
University of Florida
Gainesville, Florida

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

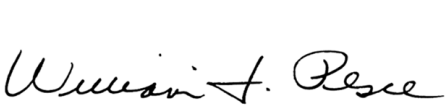
An Introduction to Categorical Data Analysis

The Wiley Bicentennial–Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!



William J. Pesce
President and Chief Executive Officer

Peter Booth Wiley
Chairman of the Board

An Introduction to Categorical Data Analysis

Second Edition

ALAN AGRESTI

Department of Statistics
University of Florida
Gainesville, Florida

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

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

Library of Congress Cataloging-in-Publication Data

Agresti, Alan
An introduction to categorical data analysis / Alan Agresti.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-471-22618-5
1. Multivariate analysis. I. Title.

QA278.A355 1996
519.5'35 - - dc22 2006042138

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Contents

Preface to the Second Edition

In recent years, the use of specialized statistical methods for categorical data has increased dramatically, particularly for applications in the biomedical and social sciences. Partly this reflects the development during the past few decades of sophisticated methods for analyzing categorical data. It also reflects the increasing methodological sophistication of scientists and applied statisticians, most of whom now realize that it is unnecessary and often inappropriate to use methods for continuous data with categorical responses.

This book presents the most important methods for analyzing categorical data. It summarizes methods that have long played a prominent role, such as chi-squared tests. It gives special emphasis, however, to modeling techniques, in particular to logistic regression.

The presentation in this book has a low technical level and does not require familiarity with advanced mathematics such as calculus or matrix algebra. Readers should possess a background that includes material from a two-semester statistical methods sequence for undergraduate or graduate nonstatistics majors. This background should include estimation and significance testing and exposure to regression modeling.

This book is designed for students taking an introductory course in categorical data analysis, but I also have written it for applied statisticians and practicing scientists involved in data analyses. I hope that the book will be helpful to analysts dealing with categorical response data in the social, behavioral, and biomedical sciences, as well as in public health, marketing, education, biological and agricultural sciences, and industrial quality control.

The basics of categorical data analysis are covered in Chapters 1–8. Chapter 2 surveys standard descriptive and inferential methods for contingency tables, such as odds ratios, tests of independence, and conditional vs marginal associations. I feel that an understanding of methods is enhanced, however, by viewing them in the context of statistical models. Thus, the rest of the text focuses on the modeling of categorical responses. Chapter 3 introduces generalized linear models for binary data and count data. Chapters 4 and 5 discuss the most important such model for binomial (binary) data, logistic regression. Chapter 6 introduces logistic regression models

for multinomial responses, both nominal and ordinal. Chapter 7 discusses loglinear models for Poisson (count) data. Chapter 8 presents methods for matched-pairs data.

I believe that logistic regression is more important than loglinear models, since most applications with categorical responses have a single binomial or multinomial response variable. Thus, I have given main attention to this model in these chapters and in later chapters that discuss extensions of this model. Compared with the first edition, this edition places greater emphasis on logistic regression and less emphasis on loglinear models.

I prefer to teach categorical data methods by unifying their models with ordinary regression and ANOVA models. Chapter 3 does this under the umbrella of generalized linear models. Some instructors might prefer to cover this chapter rather lightly, using it primarily to introduce logistic regression models for binomial data (Sections 3.1 and 3.2).

The main change from the first edition is the addition of two chapters dealing with the analysis of clustered correlated categorical data, such as occur in longitudinal studies with repeated measurement of subjects. Chapters 9 and 10 extend the matched-pairs methods of Chapter 8 to apply to clustered data. Chapter 9 does this with marginal models, emphasizing the generalized estimating equations (GEE) approach, whereas Chapter 10 uses random effects to model more fully the dependence. The text concludes with a chapter providing a historical perspective of the development of the methods (Chapter 11) and an appendix showing the use of SAS for conducting nearly all methods presented in this book.

The material in Chapters 1–8 forms the heart of an introductory course in categorical data analysis. Sections that can be skipped if desired, to provide more time for other topics, include Sections 2.5, 2.6, 3.3 and 3.5, 5.3–5.5, 6.3, 6.4, 7.4, 7.5, and 8.3–8.6. Instructors can choose sections from Chapters 9–11 to supplement the basic topics in Chapters 1–8. Within sections, subsections labelled with an asterisk are less important and can be skipped for those wanting a quick exposure to the main points.

This book is of a lower technical level than my book *Categorical Data Analysis* (2nd edition, Wiley, 2002). I hope that it will appeal to readers who prefer a more applied focus than that book provides. For instance, this book does not attempt to derive likelihood equations, prove asymptotic distributions, discuss current research work, or present a complete bibliography.

Most methods presented in this text require extensive computations. For the most part, I have avoided details about complex calculations, feeling that computing software should relieve this drudgery. Software for categorical data analyses is widely available in most large commercial packages. I recommend that readers of this text use software wherever possible in answering homework problems and checking text examples. The Appendix discusses the use of SAS (particularly PROC GENMOD) for nearly all methods discussed in the text. The tables in the Appendix and many of the data sets analyzed in the book are available at the web site http://www.stat.ufl.edu/∼aa/intro-cda/appendix.html. The web site http://www.stat.ufl.edu/∼aa/cda/software.html contains information about the use of other software, such as S-Plus and R, Stata, and SPSS, including a link to an excellent free manual prepared by Laura Thompson showing how to use R and S-Plus to

conduct nearly all the examples in this book and its higher-level companion. Also listed at the text website are known typos and errors in early printings of the text.

I owe very special thanks to Brian Marx for his many suggestions about the text over the past 10 years. He has been incredibly generous with his time in providing feedback based on using the book many times in courses. He and Bernhard Klingenberg also very kindly reviewed the draft for this edition and made many helpful suggestions. I also thank those individuals who commented on parts of the manuscript or who made suggestions about examples or material to cover. These include Anna Gottard for suggestions about Section 7.4, Judy Breiner, Brian Caffo, Allen Hammer, and Carla Rampichini. I also owe thanks to those who helped with the first edition, especially Patricia Altham, James Booth, Jane Brockmann, Brent Coull, Al DeMaris, Joan Hilton, Peter Imrey, Harry Khamis, Svend Kreiner, Stephen Stigler, and Larry Winner. Thanks finally to those who helped with material for my more advanced text (*Categorical Data Analysis*) that I extracted here, especially Bernhard Klingenberg, Yongyi Min, and Brian Caffo. Many thanks to Stephen Quigley at Wiley for his continuing interest, and to the Wiley staff for their usual high-quality support.

As always, most special thanks to my wife, Jacki Levine, for her advice and encouragement. Finally, a truly nice byproduct of writing books is the opportunity to teach short courses based on them and spend research visits at a variety of institutions. In doing so, I have had the opportunity to visit about 30 countries and meet many wonderful people. Some of them have become valued friends. It is to them that I dedicate this book.

ALAN AGRESTI

London, United Kingdom
January 2007

CHAPTER 1

Introduction

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions on various controversial issues, scientists today are finding myriad uses for methods of analyzing categorical data. It's primarily for these scientists and their collaborating statisticians – as well as those training to perform these roles – that this book was written. The book provides an introduction to methods for analyzing categorical data. It emphasizes the ideas behind the methods and their interpretations, rather than the theory behind them.

This first chapter reviews the probability distributions most often used for categorical data, such as the *binomial distribution*. It also introduces *maximum likelihood*, the most popular method for estimating parameters. We use this estimate and a related *likelihood function* to conduct statistical inference about proportions. We begin by discussing the major types of categorical data and summarizing the book's outline.

1.1 CATEGORICAL RESPONSE DATA

Let us first define categorical data. A *categorical* variable has a measurement scale consisting of a set of categories. For example, political philosophy may be measured as "liberal," "moderate," or "conservative"; choice of accommodation might use categories "house," "condominium," "apartment"; a diagnostic test to detect e-mail spam might classify an incoming e-mail message as "spam" or "legitimate e-mail."

Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical scales also occur frequently in the health sciences, for measuring responses such as whether a patient survives an operation (yes, no), severity of an injury (none, mild, moderate, severe), and stage of a disease (initial, advanced).

Although categorical variables are common in the social and health sciences, they are by no means restricted to those areas. They frequently occur in the behavioral

An Introduction to Categorical Data Analysis, Second Edition. By Alan Agresti

sciences (e.g., categories "schizophrenia," "depression," "neurosis" for diagnosis of type of mental illness), public health (e.g., categories "yes" and "no" for whether awareness of AIDS has led to increased use of condoms), zoology (e.g., categories "fish," "invertebrate," "reptile" for alligators' primary food choice), education (e.g., categories "correct" and "incorrect" for students' responses to an exam question), and marketing (e.g., categories "Brand A," "Brand B," and "Brand C" for consumers' preference among three leading brands of a product). They even occur in highly quantitative fields such as engineering sciences and industrial quality control, when items are classified according to whether or not they conform to certain standards.

1.1.1 Response/Explanatory Variable Distinction

Most statistical analyses distinguish between *response* variables and *explanatory* variables. For instance, regression models describe how the distribution of a continuous response variable, such as annual income, changes according to levels of explanatory variables, such as number of years of education and number of years of job experience. The response variable is sometimes called the *dependent variable* or *Y variable*, and the explanatory variable is sometimes called the *independent variable* or *X variable*.

The subject of this text is the analysis of categorical response variables. The categorical variables listed in the previous subsection are response variables. In some studies, they might also serve as explanatory variables. Statistical models for categorical response variables analyze how such responses are influenced by explanatory variables. For example, a model for political philosophy could use predictors such as annual income, attained education, religious affiliation, age, gender, and race. The explanatory variables can be categorical or continuous.

1.1.2 Nominal/Ordinal Scale Distinction

Categorical variables have two main types of measurement scales. Many categorical scales have a natural ordering. Examples are attitude toward legalization of abortion (disapprove in all cases, approve only in certain cases, approve in all cases), appraisal of a company's inventory level (too low, about right, too high), response to a medical treatment (excellent, good, fair, poor), and frequency of feeling symptoms of anxiety (never, occasionally, often, always). Categorical variables having ordered scales are called *ordinal* variables.

Categorical variables having unordered scales are called *nominal* variables. Examples are religious affiliation (categories Catholic, Jewish, Protestant, Muslim, other), primary mode of transportation to work (automobile, bicycle, bus, subway, walk), favorite type of music (classical, country, folk, jazz, rock), and favorite place to shop (local mall, local downtown, Internet, other).

For nominal variables, the order of listing the categories is irrelevant. The statistical analysis should not depend on that ordering. Methods designed for nominal variables give the same results no matter how the categories are listed. Methods designed for

ordinal variables utilize the category ordering. Whether we list the categories from low to high or from high to low is irrelevant in terms of substantive conclusions, but results of ordinal analyses would change if the categories were reordered in any other way.

Methods designed for ordinal variables *cannot* be used with nominal variables, since nominal variables do not have ordered categories. Methods designed for nominal variables *can* be used with nominal or ordinal variables, since they only require a categorical scale. When used with ordinal variables, however, they do not use the information about that ordering. This can result in serious loss of power. It is usually best to apply methods appropriate for the actual scale.

Categorical variables are often referred to as *qualitative*, to distinguish them from numerical-valued or *quantitative* variables such as weight, age, income, and number of children in a family. However, we will see it is often advantageous to treat ordinal data in a quantitative manner, for instance by assigning ordered scores to the categories.

1.1.3 Organization of this Book

Chapters 1 and 2 describe some standard methods of categorical data analysis developed prior to about 1960. These include basic analyses of association between two categorical variables.

Chapters 3–7 introduce models for categorical responses. These models resemble regression models for continuous response variables. In fact, Chapter 3 shows they are special cases of a generalized class of linear models that also contains the usual normal-distribution-based regression models. The main emphasis in this book is on *logistic regression* models. Applying to response variables that have two outcome categories, they are the focus of Chapters 4 and 5. Chapter 6 presents extensions to multicategory responses, both nominal and ordinal. Chapter 7 introduces *loglinear* models, which analyze associations among multiple categorical response variables.

The methods in Chapters 1–7 assume that observations are independent. Chapters 8–10 discuss logistic regression models that apply when some observations are correlated, such as with repeated measurement of subjects in longitudinal studies. An important special case is matched pairs that result from observing a categorical response for the same subjects at two separate times. The book concludes (Chapter 11) with a historical overview of categorical data methods.

Most methods for categorical data analysis require extensive computations. The Appendix discusses the use of SAS statistical software. A companion website for the book, http://www.stat.ufl.edu/∼aa/intro-cda/software.html, discusses other software.

1.2 PROBABILITY DISTRIBUTIONS FOR CATEGORICAL DATA

Inferential statistical analyses require assumptions about the probability distribution of the response variable. For regression and analysis of variance (ANOVA)