The Analysis of Cross-Classified Categorical Data

Stephen E. Fienberg

The Analysis of Cross-Classified Categorical Data

Second Edition

Springer

Stephen E. Fienberg
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213
fienberg@stat.cmu.edu

ISBN 978-0-387-72824-7

Library of Congress Control Number: 2007928366

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springer.com

To Fred

Preface to the Second Edition

The reception given to the first edition of this book, especially by nonstatisticians, has been most pleasing. Yet several readers have written to me asking for further details on, or clarifications of, methods and examples, and suggesting the preparation of sets of problems at the end of each chapter so that the book would be more useful as a text. This second edition was prepared, in large part, as a response to these requests.

Methodological research on the analysis of categorical data based on the use of loglinear models has continued at a rapid pace over the last three years. In this new edition, I have attempted to expand the discussion of several topics, by drawing selectively from this new literature, while at the same time preserving the existing structure of chapters and sections.

While not a single chapter remains completely unchanged, the bulk of the new material consists of (1) problem sets at the end of Chapters 2 through 8, (2) expanded discussion of linear logistic response models and polytomous response models in Chapter 6, (3) a further discussion of retrospective epidemiological studies in Chapter 7, and (4) a new appendix on the small-sample behavior of goodness-of-fit statistics. I have added briefer materials and references elsewhere and corrected several minor errors from the first edition. A relatively major correction has been made in connection with the theorem on collapsing tables in Section 3.8.

I gave considerable thought to the preparation of an additional appendix on computer programs for the analysis of categorical data, but in the end I resisted the temptation to do so. Many programs for maximum-likelihood estimation in connection with loglinear models are now in widespread use. These include the GLIM package prepared in England under the guidance of John Nelder and the sponsorship of the Royal Statistical Society, and various adaptations of iterative scaling programs originally prepared by Yvonne Bishop and Shelby Haberman (e.g., BMDP3F in the BMDP Programs distributed by the UCLA Health Sciences Computing Facility). Most users are likely to find one or more suitable programs available at their own computer installation that can be used to work through the examples and problems in this book. My primary reason for not providing any further guidance to computer programs is that I believe there will be major changes in both their availability and in the numerical methods they will be using within the next two to three years. Thus any explicit advice I could offer now would be out of date soon after the publication of the second edition.

Many friends, colleagues, and students provided me with suggestions, comments, and corrections for this edition. These include John Duffy, O. Dudley Duncan, David Hoaglin, J. G. Kalbfleisch, Kinley Larntz, S. Keith Lee, William Mason, Michael Meyer, Doug Ratcliff and Stanley Wasserman. The

preparation of this edition was partially supported by Office of Naval Research Contract N00014-78-C-0600 to the University of Minnesota.

For the typing and organization of the final manuscript, as well as for the updating of the indexes, I am indebted to Linda D. Anderson.

New Brighton, Minnesota
November 1979

Stephen E. Fienberg

Preface to the First Edition

The analysis of cross-classified categorical data has occupied a prominent place in introductory and intermediate-level statistical methods courses for many years, but with a few exceptions the only techniques presented in such courses have been those associated with the analysis of two-dimensional contingency tables and the calculation of chi-square statistics. During the past 15 years, advances in statistical theory and the ready availability of high-speed computers have led to major advances in the analysis of multi-dimensional cross-classified categorical data. Bishop, Fienberg, and Holland [1975], Cox [1970a], Haberman [1974a], Lindsey [1973], and Plackett [1974] have all presented detailed expositions of these new techniques, but these books are not directed primarily to the nonstatistical reader, whose background may be limited to one or two semesters of statistical methods at a noncalculus level.

The present monograph is intended as an introduction to the recent work on the analysis of cross-classified categorical data using loglinear models. I have written primarily for nonstatisticians, and Appendix I contains a summary of theoretical statistical terminology for such readers. Most of the material should be accessible to those who are familiar with the analysis of two-dimensional contingency tables, regression analysis, and analysis-of-variance models. The monograph also includes a variety of new methods based on loglinear models that have entered the statistical literature subsequent to the preparation of my book with Yvonne Bishop and Paul Holland. In particular, Chapter 4 contains a discussion of contingency tables with ordered categories for one or more of the variables, and Chapter 8 presents several new applications of the methods associated with incomplete contingency tables (i.e., tables with structural zeros).

Versions of material in this monograph were prepared in the form of notes to accompany lectures delivered in July 1972 at the Advanced Institute on Statistical Ecology held at Pennsylvania State University and during 1973 through 1975 at a series of Training Sessions on the Multivariate Analysis of Qualitative Data held at the University of Chicago. Various participants at these lectures have provided me with comments and suggestions that have found their way into the presentation here. Most of the final version of the monograph was completed while I was on sabbatical leave from the University of Minnesota and under partial support from National Science Foundation Grant SOC72-05257 to the Department of Statistics, Harvard University, and grants from the Robert Wood Johnson Foundation and the Commonwealth Fund to the Center for the Analysis of Health Practices, Harvard School of Public Health.

I am grateful to Stephen S. Brier, Michael L. Brown, Ron Christensen, David R. Cox, William Fairley, S. Keith Lee, William Mason, and Roy E. Welsch for extremely valuable comments and suggestions. Many people have provided me with examples and other materials, from both published and unpublished works, that have found their way into the final manuscript, including Albert Beaton, Richard Campbell, O. Dudley Duncan, Leo Goodman, Shelby Haberman, David Hoaglin, Kinley Larntz, Marc Nerlove, S. James Press, Ira Reiss, Thomas Schoener, and Sanford Weisberg. Most of all, I am indebted to Yvonne Bishop, Paul Holland, and Frederick Mosteller, whose collaboration over a period of many years helped to stimulate the present work.

For the typing and organization of the final manuscript, I wish to thank Sue Hangge, Pat Haswell, Susan Kaufman, and Laurie Pearlman.

New Brighton, Minnesota Stephen E. Fienberg

Contents