

Discrete Multivariate Analysis

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland
with the collaboration of
Richard J. Light and Frederick Mosteller

Discrete Multivariate Analysis Theory and Practice



Yvonne M. Bishop
Washington, DC 20015-2956
Ymbishop@verizon.net

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

Paul W. Holland
Educational Testing Service
Princeton, NJ 08541
pholland@ets.org

ISBN 978-0-387-72805-6

Library of Congress Control Number: 2007928365

© 2007 Springer Science+Business Media, LLC. This Springer edition is a reprint of the 1975 edition published by MIT Press.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springer.com

Preface

The analysis of discrete multivariate data, especially in the form of cross-classifications, has occupied a prominent place in the statistical literature since the days of Karl Pearson and Sir R. A. Fisher. Although Maurice Bartlett's pioneering paper on testing for absence of second-order interaction in $2 \times 2 \times 2$ tables was published in 1935, the widespread development and use of methods for the analysis of multidimensional cross-classified data had to await the general availability of high-speed computers. As a result, in the last ten years statistical journals, as well as those in the biological, social, and medical sciences, have devoted increasing space to papers dealing with the analysis of discrete multivariate data. Many statisticians have contributed to this progress, as a glance at the reference list will quickly reveal. We point, especially, to the sustained and outstanding contributions of Joseph Berkson, M. W. Birch, I. J. Good, Leo A. Goodman, James E. Grizzle, Marvin Kastenbaum, Gary G. Koch, Solomon Kullback, H. O. Lancaster, Nathan Mantel, and R. L. Plackett.

The one person most responsible for our interest in and continued work on the analysis of cross-classified data is Frederick Mosteller. It is not an overstatement to say that without his encouragement and support in all phases of our effort, this book would not exist. Our interest in the analysis of cross-classified data goes back to 1964 and the problems which arose during and after Mosteller's work on the National Halothane study. This work led directly to the doctoral dissertations of two of us (Bishop and Fienberg), as well as to a number of published papers. But Fred's contributions to this book are more than just encouragement; he has read and copiously commented on nearly every chapter, and while we take complete responsibility for the final manuscript, if it has any virtues they are likely to be due to him.

Richard Light enthusiastically participated in the planning of this book, and offered comments on several chapters. He prepared the earlier drafts of Chapter 11, Measures of Association and Agreement, and he made the major effort on the final version of this chapter.

We owe a great debt to many of our colleagues and students who have commented on parts of our manuscript, made valuable suggestions on aspects of our research, and generally stimulated our interest in the subject. Those to whom we are indebted include Raj Bahadur, Darrell Bock, Tar Chen, William Cochran, Joel Cohen, Arthur Dempster, O. Dudley Duncan, Hillel Einhorn, Robert Fay, John Gilbert, Anne Goldman, Shelby Haberman, David Hoaglin, Nathan Keyfitz, William Kruskal, Kinley Larntz, Siu-Kai Lee, Lincoln Moses, I. R. Savage, Thomas Schoener, Michael Sutherland, John Tukey, David Wallace, James Warram, Sanford Weisberg, Janet Wittes, and Jane Worcester.

For the production of the manuscript we are indebted to Holly Grano, Kathi Hirst, Carol Lambert, and Mary Jane Schleupner.

Preface

The National Science Foundation has provided substantial support for our research and writing under grant GS-32327X1 to Harvard University. We have also received extensive support from other research grants. These include: research grants CA-06516 from the National Cancer Institute and RR-05526 from the Division of Research Facilities and Resources, National Institutes of Health to the Children's Cancer Research Foundation; National Science Foundation research grants GP-16071, GS-1905, and a grant from the Statistics Branch, Office of Naval Research, to the Department of Statistics, University of Chicago, as well as a grant from the Alfred P. Sloan Foundation to the Department of Theoretical Biology, University of Chicago; National Science Foundation research grant GJ-1154X to the National Bureau of Economic Research, Inc., and a faculty research grant from the Social Science Research Council to Paul W. Holland.

Earlier versions of material in several chapters appeared in *The Annals of Statistics*, *Biometrics*, *Biometrika*, and *The Journal of the American Statistical Association*.

Brookline, Massachusetts
New Brighton, Minnesota
Hingham, Massachusetts

Y.M.M.B.
S.E.F.
P.W.H.

February 1974

CONTENTS

PREFACE	v
1 INTRODUCTION	1
1.1 The Need.....	1
1.2 Why a Book?	1
1.3 Different Users.....	2
1.4 Sketch of the Chapters	2
1.5 Computer Programs	7
1.6 How to Proceed from Here.....	7
2 STRUCTURAL MODELS FOR COUNTED DATA	9
2.1 Introduction	9
2.2 Two Dimensions—The Fourfold Table.....	11
2.3 Two Dimensions—The Rectangular Table.....	24
2.4 Models for Three-Dimensional Arrays.....	31
2.5 Models for Four or More Dimensions.....	42
2.6 Exercises.....	48
2.7 Appendix: The Geometry of a 2 x 2 Table.....	49
3 MAXIMUM LIKELIHOOD ESTIMATES FOR COMPLETE TABLES	57
3.1 Introduction	57
3.2 Sampling Distributions	62
3.3 Sufficient Statistics	64
3.4 Methods of Obtaining Maximum Likelihood Estimates	73
3.5 Iterative Proportional Fitting of Log-Linear Models	83
3.6 Classical Uses of Iterative Proportional Fitting.....	97
3.7 Rearranging Data for Model Fitting	102
3.8 Degrees of Freedom	114
4 FORMAL GOODNESS OF FIT: SUMMARY STATISTICS AND MODEL SELECTION	123
4.1 Introduction	123
4.2 Summary Measures of Goodness of Fit.....	124
4.3 Standardized Rates.....	131
4.4 Internal Goodness of Fit.....	136
4.5 Choosing a Model	155
4.6 Appendix: Goodman's Partitioning Calculus	169
5 MAXIMUM LIKELIHOOD ESTIMATION FOR INCOMPLETE TABLES.....	177
5.1 Introduction	177
5.2 Incomplete Two-Way Tables.....	178
5.3 Incomplete Two-Way Tables for Subsets of Complete Arrays	206
5.4 Incomplete Multiway Tables.....	210
5.5 Representation of Two-Way Tables as Incomplete Multiway Arrays.....	225

6	ESTIMATING THE SIZE OF A CLOSED POPULATION	229
6.1	Introduction	229
6.2	The Two-Sample Capture-Recapture Problem.....	231
6.3	Conditional Maximum Likelihood Estimation of N	236
6.4	The Three-Sample Census	237
6.5	The General Multiple Recapture Problem	246
6.6	Discussion	254
7	MODELS FOR MEASURING CHANGE	257
7.1	Introduction	257
7.2	First-Order Markov Models.....	261
7.3	Higher-Order Markov Models.....	267
7.4	Markov Models with a Single Sequence of Transitions	270
7.5	Other Models	273
8	ANALYSIS OF SQUARE TABLES: SYMMETRY AND MARGINAL HOMOGENEITY	281
8.1	Introduction	281
8.2	Two-Dimensional Tables	282
8.3	Three-Dimensional Tables.....	299
8.4	Summary.....	309
9	MODEL SELECTION AND ASSESSING CLOSENESS OF FIT: PRACTICAL ASPECTS.....	311
9.1	Introduction	311
9.2	Simplicity in Model Building.....	312
9.3	Searching for Sampling Models.....	315
9.4	Fitting and Testing Using the Same Data.....	317
9.5	Too Good a Fit	324
9.6	Large Sample Sizes and Chi Square When the Null Model is False	329
9.7	Data Anomalies and Suppressing Parameters	332
9.8	Frequency of Frequencies Distribution	337
10	OTHER METHODS FOR ESTIMATION AND TESTING IN CROSS-CLASSIFICATIONS.....	343
10.1	Introduction	343
10.2	The Information-Theoretic Approach	344
10.3	Minimizing Chi Square, Modified Chi Square, and Logit Chi Square	348
10.4	The Logistic Model and How to Use It	357
10.5	Testing via Partitioning of Chi Square	361
10.6	Exact Theory for Tests Based on Conditional Distributions	364
10.7	Analyses Based on Transformed Proportions	366
10.8	Necessary Developments.....	371
11	MEASURES OF ASSOCIATION AND AGREEMENT.....	373
11.1	Introduction	373
11.2	Measures of Association for 2×2 Tables.....	376
11.3	Measures of Association for $I \times J$ Tables.....	385
11.4	Agreement as a Special Case of Association.....	393

12	PSEUDO-BAYES ESTIMATES OF CELL PROBABILITIES	401
12.1	Introduction	401
12.2	Bayes and Pseudo-Bayes Estimators.....	404
12.3	Asymptotic Results for Pseudo-Bayes Estimators.....	410
12.4	Small-Sample Results	416
12.5	Data-Dependent λ 's.....	419
12.6	Another Example: Two Social Mobility Tables	426
12.7	Recent Results and Some Advice.....	429
13	SAMPLING MODELS FOR DISCRETE DATA	435
13.1	Introduction	435
13.2	The Binomial Distribution	435
13.3	The Poisson Distribution.....	438
13.4	The Multinomial Distribution.....	441
13.5	The Hypergeometric Distribution	448
13.6	The Multivariate Hypergeometric Distribution	450
13.7	The Negative Binomial Distribution.....	452
13.8	The Negative Multinomial Distribution	454
14	ASYMPTOTIC METHODS	457
14.1	Introduction	457
14.2	The O , o Notation	458
14.3	Convergence of Stochastic Sequences.....	463
14.4	The O_p , o_p Notation for Stochastic Sequences.....	475
14.5	Convergence of Moments.....	484
14.6	The δ Method for Calculating Asymptotic Distributions	486
14.7	General Framework for Multinomial Estimation and Testing.....	502
14.8	Asymptotic Behavior of Multinomial Maximum Likelihood Estimators.....	509
14.9	Asymptotic Distribution of Multinomial Goodness-of-Fit Tests	513
	REFERENCES.....	531
	INDEX TO DATA SETS	543
	AUTHOR INDEX.....	547
	SUBJECT INDEX.....	551