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 multidimensional 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

1
Introduction

1.1 The Analysis of Categorical Data

A variety of biological and social science data come in the form of cross-classified tables of counts, commonly referred to as contingency tables. The units of a sampled population in such circumstances are cross-classified according to each of several categorical variables or sets of categories such as sex (male, female), age (young, middle-aged, old), or species. Intermediate-level statistics textbooks for biologists, such as Bliss [1967], Snedecor and Cochran [1967], and Sokal and Rohlf [1969], focus on the analysis of such data in the special case of two-way cross-classifications, as do textbooks for social scientists, such as Blalock [1972]. More detailed treatments, by Maxwell [1961] and by Fleiss [1973], are also available. A review of the material presented in one or more of these books is adequate preparation for this presentation.

When we look at several categorical variables simultaneously, we say that they form a multidimensional contingency table, with each variable corresponding to one dimension of the table. Such tables present special problems of analysis and interpretation, and these problems have occupied a prominent place in statistical journals since the first article on testing in $2 \times 2 \times 2$ tables by Bartlett [1935].

Until recent years the statistical and computational techniques available for the analysis of cross-classified data were quite limited, and most researchers handled multidimensional cross-classifications by analyzing various two-dimensional marginal totals, that is, by examining the categorical variables two at a time. This practice has been encouraged by the wide availability of computer program packages that automatically produce chi-square statistics for all two-dimensional marginal totals of multidimensional tables. Although such an approach often gives great insight about the relationship among variables, it

(a) confuses the marginal relationship between a pair of categorical variables with the relationship when other variables are present,

(b) does not allow for the simultaneous examination of these pairwise relationships,

(c) ignores the possibility of three-factor and higher-order interactions among the variables.

My intention here is to present some of the recent work on the statistical analysis of cross-classified data using loglinear models, especially in the multidimensional situation. The models and methods that will be considered do not have the shortcomings mentioned above. All the techniques described will be illustrated by actual data. Readers interested in mathematical proofs should turn to the source articles or books cited.

I view this monograph as an introduction to a particular approach to the analysis of cross-classified categorical data. For more details on this approach, including mathematical proofs, various generalizations, and their ramifications, see Bishop, Fienberg, and Holland [1975] or Haberman [1974a, 1978]. Other presentations with differing contents or points of view include Cox [1970a], Gokhale and Kullback [1978], Goodman [1970, 1971b], Grizzle, Starmer, and Koch [1969], Ku, Varner, and Kullback [1971], Lancaster [1969], Lindsey [1973], and Plackett [1974]. Bock [1970, 1975] also discusses the analysis of cross-classified data, based on the notion of multinomial response relationships much like those considered here.

1.2 Forms of Multivariate Analysis

The analysis of cross-classified categorical data falls within the broader framework of multivariate analysis. A distinction will be made here between variables that are free to vary in response to controlled conditions—that is, *response* variables—and variables that are regarded as fixed, either as in experimentation or because the context of the data suggests they play a determining or causal role in the situation under study—that is, *explanatory* variables. Dempster [1971] notes that the distinction between response and explanatory variables need not be firm in a given situation, and in keeping with this view, in exploratory analyses, we often choose different sets of response variables for the same data set.

Of importance in describing various types of models and methods for multivariate analysis is the class of values assumed by the variables being examined. In many circumstances, we wish to distinguish among variables whose values are

(i) dichotomous (e.g., yes or no),
(ii) nonordered polytomous (e.g., five different detergents),
(iii) ordered polytomous (e.g., old, middle-aged, young),
(iv) integer-valued (e.g., nonnegative counts), or
(v) continuous (at least as an adequate approximation).

Variables with values of types (i) through (iv) are usually labeled *discrete*, although integer-valued variables might also be treated as if they were continuous. Here the term categorical will be used to refer primarily to types (i), (ii), and (iii), and the possibility of type (iv) will be ignored. Mixtures of categorical and continuous variables appear in many examples.

We can categorize classes of multivariate problems by the types of response and explanatory variables involved, as in the cross-classification of Table 1-1.

Table 1-1
Classes of Statistical Problems

		Explanatory Variables		
		Categorical	Continuous	Mixed
Response Variables	Categorical	(a)	(b)	(c)
	Continuous	(d)	(e)	(f)
	Mixed	?	?	?

The cells in the bottom row of this table all contain question marks in order to indicate the lack of generally accepted classes of multivariate models and methods designed to deal with situations involving mixtures of continuous and discrete response variables. Dempster [1973] has proposed a class of logit models that is of use here, but his approach has yet to see much application. The cells in the middle row correspond to problems dealt with by standard multivariate analysis, involving techniques such as
(d) analysis of variance,
(e) regression analysis,
(f) analysis of covariance (or regression analysis with some dummy variables).

The work on linear logistic response models by Walker and Duncan [1967], Cox [1970a], and Haberman [1974a] deals with problems for all three cells in the first row when there is a single dichotomous response variable, while the more recent results of Nerlove and Press [1973] handle multiple response variables. Linear logistic response models will be discussed to some extent in Chapter 6. Cell (a) of the table corresponds to cross-classified categorical data problems, and some of the most widely used models for their analysis will be described in the following chapters.

The models used throughout this book rely upon a particular approach to the definition of interaction between or among variables in multidimensional contingency tables, based on cross-product ratios of expected cell values. As a result, the models are linear in the logarithms of the expected value scale; hence the label *loglinear* models. There are several analogies between interaction in these loglinear models and the notion of interaction in analysis-of-variance (ANOVA) models. These will be pointed out in the course of the discussion. The use of ANOVA-like notation is deceptive, however. In ANOVA models one tries to assess the *effects* of independent variables on a dependent variable and to partition overall variability. In contingency table analysis the ANOVA-like models are used to describe the structural relationship among the variables corresponding to the dimensions of the table. The distinction here is important, and the fact that many researchers have not understood it has led to considerable confusion.

When a distinction is made between explanatory variables and response variables, loglinear models can be converted into *logit* or *linear logistic* response models, in which one predicts log-odds quantities involving the dependent (or response) variables using a linear combination of effects due to the explanatory variables. There is a much closer analogy between these linear logistic models and the usual ANOVA or regression models. This point is discussed in considerable detail in Chapter 6.

1.3 Some Historical Background

The use of cross-classifications to summarize counted data clearly predates the early attempts of distinguished investigators such as Quetelet in the mid-nineteenth century to summarize the association between variables in a 2 × 2 table. Not until the turn of the century, however, did Pearson and Yule formulate the first major developments in the analysis of categorical data. Despite his proposal of the well-known chi-square test for independence for two-dimensional cross-classifications (see Pearson [1900a]), Karl Pearson preferred to view a cross-classification involving two or more polytomies as arising from a partition of a set of multivariate data, with an underlying continuum for each polytomy and a multivariate normal distribution for the "original" data. This view led Pearson [1900b] to develop his tetrachoric correlation coefficient for 2 × 2 tables and served as the basis for the approach adopted by many subsequent authors, such as Lancaster [1957] and Lancaster and Hamdan [1964]. This approach in some sense also led to Lancaster's method of partitioning chi-square, to which I shall return shortly. The most serious problems with Pearson's approach were (1) the complicated infinite series linking the tetrachoric correlation coefficient with the frequencies in a 2 × 2 table and (2) his insistence that it *always* made sense to assume an underlying continuum for a dichotomy or polytomy, even when the dichotomy of interest was dead–alive or employed–unemployed, and that it was reasonable to assume that the probability distribution over such a dead–alive continuum was normal.

Yule [1900], on the other hand, chose to view the categories of a cross-classification as fixed, and he set out to consider the structural relationship between or among the discrete variables represented by the cross-classification. This approach led him to consider various functions of the cross-product ratio, discussed here in Chapter 2. When there actually is an underlying continuum for each of two polytomies, the cross-product ratio for a 2 × 2 table resulting from a partitioning of the two variables simply is not a substitute for an estimate of the true correlation coefficient of the underlying continuum (see Plackett [1965] or Mosteller [1968]). Thus the methods proposed by Yule are not necessarily applicable in cases where the 2 × 2 table is simply a convenient summary device for continuous bivariate data and the original observations are in fact available.

The debate between Pearson and Yule was both lengthy and acrimonious (see, e.g., Pearson and Heron [1913]), and in some ways it has yet to be completely resolved, although the statistical literature of the past 25 years on this topic would indicate that Yule's position now dominates. In fact, Yule can be thought of as the founder of the loglinear model school of contingency table analysis, and most of the results in this book are an outgrowth of his pioneering work. However, the notions of Yule were not immediately generalized beyond the structure of two-dimensional tables. Thirty-five years passed before Bartlett [1935], as a result of a personal communication from R. A. Fisher, utilized Yule's cross-product ratio to define the concept of second-order interaction in a 2 × 2 × 2 contingency table (see Chapter 3).

While the multivariate generalizations of the Yule–Bartlett cross-product ratio or loglinear model approach were fermenting, the technique of standardization (see Bishop, Fienberg, and Holland [1975], Chapter 4, and Bunker et al. [1969]) to eliminate the effects of categorical covariates received considerable attention in the epidemiological literature. Standardization is basically a descriptive technique that has been made obsolete, for most of the purposes to which it has traditionally been put, by the ready availability of computer programs for loglinear model analysis of multidimensional contingency tables. Thus it is not discussed in detail in this book.

During the past 25 years, the statistical literature on the analysis of categorical data has focused primarily on three classes of parametric models: (1) loglinear models, (2) additive models, and (3) models resulting from partitioning chi-square, which may be viewed as a combination of multiplicative and additive. This last class of models, which is usually associated with the work of Lancaster, is much misunderstood, as Darroch [1974, 1976] has recently noted. In addition, there has been a related literature on measures of association (e.g., Goodman and Kruskal [1954, 1959, 1963, 1972]). Although different groups of authors use different methods of estimation (maximum likelihood, minimum modified chi-square, or minimum discrimination information), almost all of the recent literature can be traced back either to the 1951 paper of Lancaster or to the work of S. N. Roy and his students at North Carolina in the mid-1950s (e.g., Roy and Kastenbaum [1956], Roy and Mitra [1956]). It is interesting that Roy's students developed his ideas using both the minimum modified chi-square approach (e.g., Bhapkar and Koch [1968a, b], Grizzle, Starmer, and Koch [1969]) and the method of maximum likelihood (e.g., Bock [1970], Kastenbaum [1974]).

The major advances in the literature on multidimensional contingency tables in the 1960s grew out of Roy and Kastenbaum's work and papers by Birch [1963], Darroch [1962], Good [1963], and Goodman [1963, 1964]. These advances coincided with the emergence of interest in and the availability of high-speed computers, and this work received substantial impetus from several large-scale data analysis projects. Much of the recent literature on loglinear models can be linked directly to the National Halothane Study (see Bunker et al. [1969], Bishop, Fienberg, and Holland [1975], Mosteller [1968]), while problems in the Framingham Study led to work on linear logistic models involving both categorical and continuous predictor variables (e.g., Cornfield [1962], Truett, Cornfield, and Kannel [1967]). The Framingham study work paralleled work on linear logistic models by Cox [1966] and had ties to the earlier contributions of Berkson [1944, 1946].

A fairly complete bibliography for the statistical literature on contingency tables through 1974 is given by Killion and Zahn [1976].

1.4 A Medical Example

Table 1-2 presents data compiled by Cornfield [1962] from the Framingham longitudinal study of coronary heart disease (see Dawber, Kannel, and Lyell [1963] for a detailed description). Variable 1 is a binary response variable indicating the presence or absence of coronary heart disease, while variable 2

Table 1-2
Data from the Framingham Longitudinal Study of Coronary Heart Disease (Cornfield [1962])

Coronary Heart Disease	Serum Cholesterol (mg/100 cc)	Systolic Blood Pressure (mm Hg) < 127	127–146	147–166	167+
Present	< 200	2	3	3	4
	200–219	3	2	0	3
	220–259	8	11	6	6
	≥ 260	7	12	11	11
Absent	< 200	117	121	47	22
	200–219	85	98	43	20
	220–259	119	209	68	43
	≥ 260	67	99	46	33

(serum cholesterol at four levels) and variable 3 (blood pressure at four levels) are explanatory. The data as displayed in Table 1-2 form a 2 × 4 × 4 three-dimensional contingency table.

This example is typical of those encountered in medical contexts. Although serum cholesterol and blood pressure might well be viewed as continuous variables, the values of these variables have been broken up into four categories each, corresponding to different levels of a priori perceived risk of coronary heart disease. An alternative to this approach would be to treat the variables as continuous and to use a regression-like logistic response model that expresses the dependency of coronary heart disease in a smooth and simple fashion.

2
Two-Dimensional Tables

2.1 Two Binomials

We often wish to compare the relative frequency of occurrence of some characteristic for two groups. In a review of the evidence regarding the therapeutic value of ascorbic acid (vitamin C) for treating the common cold, Pauling [1971] describes a 1961 French study involving 279 skiers during two periods of 5–7 days. The study was double-blind with one group of 140 subject receiving a placebo while a second group of 139 received 1 gram of ascorbic acid per day. Of interest is the relative occurrence of colds for the two groups, and Table 2-1 contains Pauling's reconstruction of these data.

If P_1 is the probability of a member of the placebo group contracting a cold and P_2 is the corresponding probability for the ascorbic acid group, then we are interested in testing the hypothesis that $P_1 = P_2$. The observed numbers of colds in the two groups, $x_{11} = 31$ and $x_{21} = 17$ respectively, are observations on independent binomial variates with probabilities of success P_1 and P_2 and sample sizes $n_1 = 140$ and $n_2 = 139$. The difference in observed proportions,

$$\bar{P}_1 - \bar{P}_2 = \frac{x_{11}}{n_1} - \frac{x_{21}}{n_2},$$

has mean $P_1 - P_2$ and variance

$$\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}.$$

Table 2-1
Incidence of Common Colds in a Double-Blind Study Involving 279 French Skiers (Pauling [1971])

(a) Observed values

		Cold	No Cold	Totals
Treatment	Placebo	31	109	140
	Ascorbic Acid	17	122	139
	Totals	48	231	279

(b) Expected values under independence

		Cold	No Cold	Totals
Treatment	Placebo	24.1	115.9	140
	Ascorbic Acid	23.9	115.1	139
	Totals	48	231	279

If $P_1 = P_2$, then we could estimate the common value by

$$\bar{P} = \frac{\text{total no. of colds}}{n_1 + n_2} \tag{2.1}$$

and the estimated variance of $\bar{P}_1 - \bar{P}_2$ by

$$\bar{P}(1 - \bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{2.2}$$

Assuming that the hypothesis $P_1 = P_2$ is correct, a reasonable test can be based on the approximate normality of the standardized deviate

$$z = \frac{\bar{P}_1 - \bar{P}_2}{\sqrt{\bar{P}(1 - \bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{2.3}$$

For our example we get

$$z = \frac{\frac{31}{140} - \frac{17}{139}}{\sqrt{\frac{48}{279} \times \frac{231}{279} \times \left(\frac{1}{140} + \frac{1}{139}\right)}} = 2.19,$$

a value that is significant at the 0.05 level. If we take these data at face value, then we would conclude that the proportion of colds in the vitamin C group is smaller than that in the placebo group. This study, however, has a variety of severe shortcomings (e.g., the method of allocation is not specified and the evaluation of symptoms was largely subjective). For a further discussion of these data, and for a general review of the studies examining the efficacy of vitamin C as a treatment for the common cold up to 1974, see Dykes and Meier [1975].

As an alternative to using the normal approximation to the two-sample binomial problem, we could use the Pearson chi-square statistic (see Pearson [1900a]),

$$X^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}, \tag{2.4}$$

where the summation is over all four cells in Table 2-1. We obtain the expected values by estimating $P_1 = P_2 = P$ (the null value) as $\bar{P} = 48/279$; that is, we multiply the two sample sizes n_1 and n_2 by $\bar{P}$, obtaining the expected values for the (1, 1) and (2, 1) cells, and then get the other two expected values by subtraction. Table 2-lb shows these expected values, and on substituting the observed and expected values in expression (2.4) we get $X^2 = 4.81$, a value that may be referred to a χ^2 distribution with 1 d.f. (degree of freedom). A large value of X^2 corresponds to a value in the right-hand tail of the χ^2 distribution and is indicative of a poor fit. Rather than using the χ^2 table we note that the square root of 4.81 is 2.19, the value of our z-statistic computed earlier. Some elementary algebra shows that, in general, $z^2 = X^2$. If we set $x_{12} = n_1 - x_{11}$ and $x_{22} = n_2 - x_{21}$, then

$$z^2 = \frac{\left(\frac{x_{11}}{n_1} - \frac{x_{21}}{n_2}\right)^2}{\left(\frac{x_{11}+x_{21}}{n_1+n_2}\right)\left(\frac{x_{12}+x_{22}}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{2.5}$$

$$= \frac{[x_{11}(n_2 - x_{21}) - x_{21}(n_1 - x_{11})]^2(n_1+n_2)}{(x_{11}+x_{21})(x_{12}+x_{22})n_1n_2}$$

and

$$X^2 = \frac{\left[x_{11} - n_1\left(\frac{x_{11}+x_{21}}{n_1+n_2}\right)\right]^2}{n_1\left(\frac{x_{11}+x_{21}}{n_1+n_2}\right)} + \frac{\left[x_{12} - n_1\left(\frac{x_{12}+x_{22}}{n_1+n_2}\right)\right]^2}{n_1\left(\frac{x_{12}+x_{22}}{n_1+n_2}\right)}$$

$$+ \frac{\left[x_{21} - n_2\left(\frac{x_{11}+x_{21}}{n_1+n_2}\right)\right]^2}{n_2\left(\frac{x_{11}+x_{21}}{n_1+n_2}\right)} + \frac{\left[x_{22} - n_2\left(\frac{x_{12}+x_{22}}{n_1+n_2}\right)\right]^2}{n_2\left(\frac{x_{12}+x_{22}}{n_1+n_2}\right)} \tag{2.6}$$

$$= \frac{[x_{11}(n_2 - x_{21}) - x_{21}(n_1 - x_{11})]^2(n_1+n_2)}{(x_{11}+x_{21})(x_{12}+x_{22})n_1n_2}.$$

The use of the statistic X^2 is also appropriate for testing for independence in 2 × 2 tables, as noted in the next section.

Throughout this book we use the Greek quantity χ^2 to refer to the chi-square family of probability distributions, and the Roman quantity X^2 to refer to the Pearson goodness-of-fit test statistic given in general by expression (2.4).

2.2 The Model of Independence

We have just examined a 2 × 2 table formed by considering the counts generated from two binomial variates. For this table the row totals were fixed