

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Massey University Library. Thesis Copyright Form

Title of thesis: A Review of some Models for the
Analysis of Contingency Tables

- (1) (a) I give permission for my thesis to be made available to readers in the Massey University Library under conditions determined by the Librarian.
- (b) ~~I do not wish my thesis to be made available to readers without my written consent for _____ months.~~
- (2) (a) I agree that my thesis, or a copy, may be sent to another institution under conditions determined by the Librarian.
- (b) ~~I do not wish my thesis, or a copy, to be sent to another institution without my written consent for _____ months.~~
- (3) (a) I agree that my thesis may be copied for Library use.
- (b) ~~I do not wish my thesis to be copied for Library use for _____ months.~~

Signed J.A. Anderson

Date 18/12/87

The copyright of this thesis belongs to the author. Readers must sign their name in the space below to show that they recognise this. They are asked to add their permanent address.

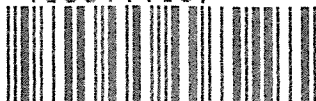
NAME AND ADDRESS

DATE

FOR
Reference Only

NOT TO BE REMOVED FROM THE LIBRARY

1088111067



A REVIEW OF SOME MODELS FOR THE ANALYSIS
OF CONTINGENCY TABLES

A thesis presented in partial fulfilment of the requirements
for the degree of Master of Arts in Statistics
at Massey University

Julie Anne Anderson

1987

ABSTRACT

Some models proposed for the analysis of contingency tables are reviewed and illustrated with examples.

These include standard loglinear models; models which are suitable for ordinal categorical variables such as ordinal loglinear, log-multiplicative and logit models, and models based on an underlying distribution for the response; and models for incomplete and square tables.

Estimation methods and inference are also discussed.

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr Doug Stirling, for his guidance and supervision of this project. Thanks also to Mrs Gail Haydock for the typing of this thesis. Finally, I wish to thank my husband Dean for his help, encouragement and support.

TABLE OF CONTENTS

	<u>Page</u>
Abstract	(i)
Acknowledgements	(ii)
Table of Contents	(iii)
Chapter I: Introduction	1
1.1 Categorical Variables	1
1.2 Two-Dimensional Tables	2
1.3 Three-Dimensional Tables	3
1.4 Sampling Models	4
1.4.1 Poisson	4
1.4.2 Multinomial	5
1.4.3 Product Multinomial	5
1.4.4 Equivalence of Results for Different Sampling Models	6
1.5 Response and Explanatory Variables	6
1.5.1 Three Responses	7
1.5.2 Two Responses	8
1.5.3 One Response	8
1.5.4 Types of Models that can be Fitted	8
1.6 Ordinal Categorical Data	9
1.6.1 Advantages of using Ordinal Methods	10
1.6.2 Odds Ratios for 2x2 Tables	11
1.6.2.1 Incidence of Colds Example	13
1.6.3 Odds Ratios for rxc Tables	13
1.6.3.1 Local Odds Ratios	14
1.6.3.2 Local-Global Odds Ratios	15
1.6.3.3 Global Odds Ratios	16
1.6.3.4 Dumping Severity Example	17

1.7	Estimation	19
1.8	Model testing	20
1.9	Structural and Sampling Zeroes	20
1.9.1	Monkey Example	20
1.10	Loglinear Models	23
1.10.1	Fitting Loglinear Models	24
1.11	Linear Models	24
1.11.1	Linear Models Specified as $A_{\pi} = X\beta$	24
1.11.2	Linear Models Specified in Terms of Constraints	25
1.11.3	Fitting linear models	26
1.12	Other Models	26

	<u>Page</u>
Chapter II: Nominal Loglinear Models	28
2.1 Two Dimensional Tables	28
2.1.1 Multiplicative Form of the Loglinear Model	28
2.1.1.1 Saturated Model	28
2.1.1.2 Independence Model	31
2.1.2 Additive Form	31
2.1.2.1 Saturated Model	31
2.1.2.2 Independence Model	33
2.1.2.2.1 Abortion Attitude Example	33
2.1.2.3 Goodness of Fit Statistics	35
2.1.2.3.1 Abortion Attitude Example	35
2.2 Three Dimensional Tables	36
2.2.1 Associations Between Three Variables	36
2.2.2 Hierarchial Models	38
2.2.3 Estimation	39
2.2.4 Model Testing	39
2.2.5 Abortion Attitude Example	40
2.2.6 Conditional Test Statistics	42
2.2.6.1 Partitioning Chi-Square	43
2.2.6.2 Abortion Attitude Example	43
2.2.6.3 University Admissions Example	43
2.3 Higher Order Contingency Tables	44
2.4 Other Loglinear Models	45

	<u>Page</u>
Chapter III: Some General Issues	46
3.1 Model Selection	46
3.1.1 Abortion Attitude Example	46
3.1.2 Residual Analysis	47
3.2 Interpretation	48
3.2.1 Abortion Attitude Example	48
3.3 Collapsing Tables	49
3.3.1 University Admissions Example	50

	<u>Page</u>
Chapter IV: Loglinear Models for Ordinal Variables	53
4.1 Disadvantages of Ignoring Ordinal Nature of Variables	53
4.1.1 Dumping Severity Example	53
4.1.1.1 Odds Ratios	54
4.1.1.2 Residual Analysis	54
4.2 Ordinal-Ordinal Tables	55
4.2.1 Dumping Severity Example	55
4.2.2 Linear by Linear Association Model	57
4.2.3 Estimation	58
4.2.4 Dumping Severity Example	59
4.2.5 Conditional Test of Independence	60
4.3 Ordinal-Nominal Tables	61
4.3.1 Estimation	63
4.3.2 Dumping Severity Example	63
4.3.3 Conditional Test of Independence	65
4.4 Higher Dimensions	67
4.4.1 Odds Ratios	68
4.4.2 All Variables Ordinal	68
4.4.3 Ordinal and Nominal Variables	69
4.4.3.1 One Nominal, Two Ordinal Variables	69
4.4.3.2 Two Nominal, One Ordinal Variable	70
4.4.3.3 Dumping Severity Example	70
4.4.4 Three Factor Interaction Models	74
4.4.4.1 Smoking Example	75

	<u>Page</u>
Chapter V: Log-Multiplicative Models	79
5.1 Estimation	80
5.2 Inference for Log-Multiplicative Models	81
5.3 Dumping Severity Example	81
5.4 Higher Dimensions	84

	<u>Page</u>
Chapter VI: Logit Models	85
6.1 Dichotomous Response	85
6.2 Polytomous Response	87
6.2.1 Adjacent Categories Logits	88
6.2.1.1 Ordinal-Ordinal	87
6.2.1.2 Ordinal-Nominal	89
6.2.2 Continuation Ratio Logits	89
6.2.3 Cumulative Logits	90
6.2.3.1 Heterogeneous Effects	90
6.2.3.1.1 Ordinal-Ordinal Tables	90
6.2.3.1.1.1 Dumping Severity Example	91
6.2.3.1.2 Ordinal-Nominal Tables	95
6.2.3.2 Homogenous Effects	95
6.2.3.2.1 Ordinal-Ordinal Tables	97
6.2.3.2.1.1 Dumping Severity Example	97
6.2.3.2.1.2 Conditional Test of Independence	99
6.2.3.2.2 Ordinal-Nominal Tables	99
6.2.3.2.2.1 Dumping Severity	100
6.2.3.2.2.2 Conditional Test of Independence	102
6.2.4 Cumulative Logit Models for Higher Dimensions	103
6.2.4.1 Homogeneous Linear Logit Effects	104
6.2.4.1.1 Dumping Severity Example	105
6.2.4.2 Higher Order Interaction Models	107
6.2.4.3 Heterogeneous Effects	107

Page

Chapter VII: Models Based on an Underlying Distribution for the Response	108
7.1 Distribution Functions	110
7.1.1 Normal	111
7.1.2 Logistic	111
7.1.3 Extreme Value	113
7.1.4 Estimation	115
7.1.5 Dumping Severity Example	118

	<u>Page</u>
Chapter VIII: Other Models	125
8.1 Mean Response Models	125
8.1.1 Dumping Severity Example	125
8.1.2 Three-Way Dumping Severity Example	126
8.2 Models for Incomplete Tables	127
8.2.1 Definitions	127
8.2.2 Quasi-Independence	127
8.2.3 Estimation	128
8.2.4 Monkey Example	128
8.2.5 Higher-Order Tables	129
8.2.5.1 Health Concerns Example	130
8.3 Models for Square Tables	132
8.3.1 Quasi-Independence	133
8.3.1.1 Social Mobility Example	134
8.3.2 Symmetry	136
8.3.2.1 Eye-Testing Example	136
8.3.3 Quasi-Symmetry	138
8.3.3.1 Eye-Testing Example	139
8.3.4 Marginal Homogeneity	139
8.3.4.1 Estimation	140
8.3.4.2 Eye-Testing Example	140
8.3.5 Multi-Dimensional Tables	142
8.4 Linear Models	142
8.4.1 Drug Example	142
8.5 Summary	145

Appendices

Appendix 1: Equivalence of MLEs under Poisson, Multinomial and Product-Multinomial Sampling Schemes	148
A.1.1 MLEs	148
A.1.2 Deviance	150
A.1.2.1 Multinomial	150
A.1.2.2 Poisson	150
A.1.2.3 Product-Multinomial	151
Appendix 2: Fitting Loglinear Models	152
A.2.1 Newton-Raphson Algorithm	152
A.2.1.1 Abortion Attitude Example	153
A.2.2 Interactive Proportional Fitting	154
A.2.2.1 Lizard Example	155
Appendix 3: Fitting Linear Models	159
A.3.1 Linear Models Specified in Terms of Constraints	159
A.3.1.1 Wedderburn's Algorithm for Finding MLEs of Generalized Linear Models Specified in Terms of Constraints	160
A.3.2 Linear Models Specified as $A\pi = X\beta$	165
A.3.2.1 Drugs Example	166
Appendix 4: Maximum Likelihood Equations for Loglinear Models for Three-Dimensional Tables	171
Appendix 5: Fitting Mean Response Models	174
A.5.1 Dumping Severity Example	174

	<u>Page</u>
Appendix 6: Quasi-Symmetry	178
A.6.1 One Dummy Variable	178
A.6.1.1 Eye-Testing Example	178
A.6.2 Two Dummy Variables	181
A.6.2.1 Eye-Testing Example	181
Appendix 7: Marginal Homogeneity	182
A.7.1 Method of Solving Simultaneous Equations	182
A.7.1.1 Eye-Testing Example	182
A.7.2 Generalized Linear Models Specified in Terms of Constraints	184
A.7.2.1 Eye-Testing Example	184
Appendix 8: Fitting Models Using Genstat	188
References	273

CHAPTER I: INTRODUCTION

1.1 Categorical Variables

This thesis discusses different types of models that can be used to describe categorical data. A categorical variable differs from a continuous variable in that rather than being able to take on a continuous range of values, it is only classified into a certain number of categories. An example would be marital status, which could have categories such as married, widowed, divorced, or "other". If we classify each member of a sample simultaneously on two or more categorical variables, then we can form a cross-classification table. For example, we might classify 1000 people by their marital status and age (where age has only been measured in categories) such as in Table 1.1.

Table 1.1: Cross-classification table of 1000 people by age and marital status

Age (years)	Marital Status				Total
	Married	Widowed	Divorced	Other	
< 25	100	10	10	180	300
25 - 40	200	50	100	50	400
> 40	120	75	80	25	300
TOTAL	420	135	190	255	1000

A cross-classification table is also referred to as a contingency table or cross-tabulation.

For some variables such as marital status and sex, the only sensible way to measure them is to classify them into categories. However, some other variables, such as age and income, can be measured on a

continuous scale, but it is often more convenient to simply categorize them.

1.2 Two-Dimensional Tables

Consider a two-way table of counts with the row variable, X , having r categories, and the column variable, Y , having c categories - thus there are r rows and c columns. We will denote the actual count in the i th row and j th column by n_{ij} , and the corresponding expected count under some model as m_{ij} . The row and column totals are:

$$n_{i+} = \sum_{j=1}^c n_{ij}$$

$$n_{+j} = \sum_{i=1}^r n_{ij}$$

The total number of observations is

$$\sum_i \sum_j n_{ij} = N$$

Assuming that neither category has fixed marginal totals, the probability that a given individual is classified into cell (i, j) is:

$$\pi_{ij} = P(X \text{ takes on level } i \text{ and } Y \text{ takes on level } j)$$

where

$$\pi_{ij} = \frac{m_{ij}}{N}$$

$$\sum_i \pi_{ij} = \pi_{+j}$$

$$\sum_j \pi_{ij} = \pi_{i+}$$

$$\sum_{ij} \pi_{ij} = 1$$

If X and Y are independent, then

$$\begin{aligned} \pi_{ij} &= P(X \text{ takes on level } i) \times P(Y \text{ takes on level } j) \\ &= \pi_{i+} \pi_{+j} \end{aligned}$$

Since the expected value of n_{ij} is

$$m_{ij} = N \pi_{ij}$$

then under the model of independence

$$m_{ij} = N \pi_{i+} \pi_{+j}$$

Later, we will discuss models that allow X and Y to be associated in some way. For these models the expected values depend on more than just the marginal probabilities.

1.3 Three-Dimensional Tables

We can extend the notation introduced in Section 1.2 to the case of three-way tables. A three-way table with variables X, Y and Z having r , c and ℓ categories respectively, will be said to have observed counts n_{ijk} with corresponding expected counts m_{ijk} and population probabilities π_{ijk} .

An example of a three-way table is Table 1.2 which classifies a sample of 1593 people by their age, religion and frequency of church attendance (Knoke and Burke, 1980, p.68).

Table 1.2: Effect of age and religion on church attendance

Religion	Age	Church Attendance			Total
		Low	Medium	High	
Non-Catholic	Young	322	124	141	587
	Old	250	152	194	596
Catholic	Young	88	45	106	239
	Old	28	24	119	171
TOTAL		688	345	560	1593

Later, we will formulate models that allow various types of association between the variables.

1.4 Sampling Models

There are three common sampling models that are used for the collection of cross-classified data. We will illustrate for the case of an $r \times c \times l$ table classified by variables X , Y and Z . These results can be easily generalized to tables of a different dimension.

1.4.1 Poisson

We observe a set of independent Poisson processes, one for each cell in the table over a fixed time period, with no prior knowledge of the total number of observations to be taken. The count n_{ijk} in each cell will have a Poisson distribution with mean m_{ijk} , i.e. the probability function for n_{ijk} has the form

$$f(n_{ijk}) = \frac{m_{ijk}^{n_{ijk}} e^{-m_{ijk}}}{n_{ijk}!}$$

The log likelihood function is

$$\log L(n_{ijk}) = \sum_{i,j,k} n_{ijk} \log m_{ijk} - \sum_{i,j,k} m_{ijk} - \sum_{i,j,k} n_{ijk}!$$

Since the cells contain counts having independent Poisson distributions, the total count in the table, N , has a Poisson distribution with mean

$$m_{+++} = \sum_{ijk} m_{ijk}$$

1.4.2 Multinomial

We take a fixed sample of size N and cross-classify each member of the sample according to the categorical variables. The cell counts $\{n_{ijk}\}$ will have the multinomial distribution specified by the sample size N and the $rc\ell$ population probabilities $\{\pi_{ijk}\}$. The probability of a particular set of cell counts $\{n_{ijk}\}$ that sum to N is the multinomial likelihood

$$L(n_{ijk}) = \frac{N!}{\prod_{i,j,k} n_{ijk}!} \prod_{i,j,k} \pi_{ijk}^{n_{ijk}}$$

The log likelihood is

$$\log L(n_{ijk}) = \sum_{i,j,k} n_{ijk} \log \pi_{ijk} + \log N! - \sum_{i,j,k} \log n_{ijk}!$$

The expected value of each n_{ijk} is $m_{ijk} = N\pi_{ijk}$.

1.4.3 Product Multinomial

For each combination of one or more categorical explanatory variables, we take a multinomial sample of fixed size which is classified by the remaining response variable(s). For example, suppose we fix the ℓ layer totals and take a sample of size n_{++k} for each k . Let $\pi_{ij(k)}$ be the probability of an observation falling into the i th category of X and the j th category of Y , given

that it falls into the k th category of Z (i.e. π_{ijk}/π_{++k}). The cell counts within the k th layer have the multinomial distribution specified by the sample size n_{++k} and the probabilities $\{\pi_{ijk} \mid k = 1, \dots, \ell\}$, and cell counts from different layers are independent. The cell counts in layer k have the probability function

$$L(n_{ij(k)}) = \frac{n_{++k}!}{\prod_{i,j} n_{ijk}!} \prod_{i,j} \pi_{ij(k)}^{n_{ijk}}$$

and the product of these from the ℓ layers gives the probability function for the whole table (the product multinomial likelihood)

$$L(n_{ijk}) = \prod_k \frac{n_{++k}!}{\prod_{i,j} n_{ijk}!} \prod_{i,j} \pi_{ij(k)}^{n_{ijk}}$$

The expected value of each n_{ijk} is $m_{ijk} = n_{++k} \pi_{ij(k)}$.

1.4.4 Equivalence of Results for Different Sampling Models

For the models that will be discussed in this thesis, the maximum likelihood estimates (MLEs) are the same for all sampling schemes. The one condition required is that a term corresponding to the fixed margin(s) in the product multinomial sampling scheme be included in the model (for more details see Appendix 1). Because of this equivalence, generally models will be phrased as though the sampling scheme was multinomial.

1.5 Response and Explanatory Variables

Each variable (i.e. margin) in a table can be thought of as either an explanatory variable (factor) which affects others, or as a response variable which depends on other factors.

For three-dimensional tables there are three possible combinations:

- (i) no explanatory, three response variables
- (ii) one explanatory, two response variables
- (iii) two explanatory, one response variable.

Examples of these three types of tables include Tables 1.2, 1.3 and 1.4.

Table 1.3: Occupation (O), Education (E), and Aptitude (A)
of World War II volunteers

O1 (self employed, business)

O2 (self employed, professional)

	E1	E2	E3	E4		E1	E2	E3	E4
A1	42	55	22	3		1	2	8	19
A2	72	82	60	12		1	2	15	33
A3	90	106	85	25		2	5	25	83
A4	27	48	47	8		2	2	10	45
A5	8	18	19	5		0	0	12	19

O3 (teacher)

O4 (salary employed)

	E1	E2	E3	E4		E1	E2	E3	E4
A1	0	0	1	19		172	151	107	42
A2	0	3	3	60		208	198	206	92
A3	1	4	5	86		279	271	331	191
A4	0	0	2	36		99	126	179	97
A5	0	0	1	14		36	35	99	79

1.5.1 Three Responses

Type (i) tables are only rarely found in practice. However Table 1.3 can be thought of as one. The data, taken from Fienberg (1980, p. 45) refer to the classification of 4353 World War II volunteers into four occupational groups by four levels of education and five

levels of aptitude. Because of the sampling scheme and the way in which the individuals were classified (see Fienberg for further details), all three variables can be thought of as responses.

1.5.2 Two Responses

Table 1.4, taken from Fienberg (1980, p.27) is an example of the second type of table. The data refer to the perch heights and diameters of two different species of lizards. Species is an explanatory variable which affects the responses of height and diameter.

Table 1.4: Perch height and diameter of two species of lizards

Perch Diameter	Sagrei Species Perch height		Distichus Species Perch height	
	< 4.0"	> 4.0"	< 4.0"	> 4.0"
>4.75'	32	86	61	73
<4.75'	11	35	41	70

1.5.3 One Response

Type (iii) tables are the most common three-dimensional tables. An example is given in Table 1.2 which illustrates the effect of the explanatory variables, religion and age, on the response, frequency of church attendance.

1.5.4 Types of Models that can be Fitted

For type (i) tables only Poisson or multinomial sampling schemes are usually appropriate, whereas for types (ii) and (iii) we would also use a product-multinomial model in which the fixed marginal totals correspond to explanatory variables.

The distinction between explanatory and response variables certainly affects the interpretation of the results, but often does not affect the types of models that can be fitted. A sensible approach for the analysis of tables with one or more explanatory variables is to condition on the values of these margins, treating them as fixed even in those cases where they are not. We will discuss this approach more fully later.

1.6 Ordinal Categorical Data

When one or more of the variables in a cross-classification is measured on an ordinal scale, we can use models which take account of this to give more powerful tests of association and simpler, more incisive measures of this association than models which simply treat all the variables as nominal.

An illustration of an ordinal variable and the levels of its corresponding scale would be education which might be measured as primary school, high school, or tertiary education.

Other examples would be consumer rating of a new food product as dislike a lot, dislike, indifferent, like, like a lot; or measuring the softness of water as soft, medium or hard.

Ordinal scales commonly occur in many disciplines, such as the social sciences (e.g. for measuring attitudes and opinions), marketing (e.g. for preference scales), medicine (e.g. for describing severity of an injury, or degree of recovery from an illness). In many fields ordinal scales often result when discrete measurement is used with inherently continuous variables such as age, income or social status. Often it is possible to measure a variable perhaps even on a continuous scale, but much quicker and more convenient to simply measure it on an ordinal scale. For instance, the amount of sediment left on a filter pad may be simply classified as none, slight, moderate or excessive by comparing it to a photographic standard rather than drying it and precisely weighing it.

A categorical variable is referred to as "ordinal" rather than "interval" when there is a clear ordering of the categories but the absolute distances among them are unknown. For example, the variable

"education" is ordinal when measured with categories primary school, high school, university, but it is interval when measured with the integer values 0, 1, 2,... representing number of years of education.

An ordinal variable is quantitative because it corresponds to different quantities of a certain characteristic, while qualitative variables which are measured on a nominal scale have no such property. Examples of nominal variables are race, religion or marital status. The order of listing of the categories of a nominal variable is obviously unimportant.

1.6.1 Advantages of Using Ordinal Methods

Most of the well-known methods for analysing categorical data (such as the Pearson chi-squared test of independent or the common loglinear models discussed in Chapter II) treat all variables as nominal, i.e. the results are invariant to permutations of the categories of any of the variables.

Since ordinal variables are inherently quantitative, Agresti (1984) argues that their descriptive measures should be more like those for interval variables than those for nominal variables.

The advantages of using ordinal methods instead of the standard nominal procedures include:

1. Ordinal methods have greater power for detecting particular kinds of association;
2. Ordinal data description is based on measures that are similar to those (e.g. correlations, slopes) used in ordinary regression and analysis of variance for continuous variables;
3. Ordinal analyses can use a greater variety of models, most of which are more parsimonious and have simpler interpretations than the standard models for nominal variables.
4. Interesting ordinal models can be applied in settings where the standard nominal models are trivial or else have too many parameters to be tested for goodness of fit.

In Chapters IV to VII we will discuss particular classes of models that can be used to model ordinal categorical data. These include ordinal loglinear, log-multiplicative and logit models, as well as models based on underlying distributions for the response.

1.6.2 Odds Ratios for 2x2 Tables

The odds ratio is a measure that describes the degree of association in a 2x2 table - it is especially important in the study of ordinal models.

Consider the 2x2 population cross-classification with cell probabilities π_{ij} . Within row 1 the odds that variable 2 is in column 2 instead of column 1 is

$$\Omega_1 = \frac{\pi_{12}}{\pi_{11}}$$

Within row 2 the corresponding odds equals

$$\Omega_2 = \frac{\pi_{22}}{\pi_{21}}$$

Each Ω_i is nonnegative, with value greater than 1.0 if column 2 is more likely than column 1.

The ratio of these odds

$$\begin{aligned} \theta = \frac{\Omega_2}{\Omega_1} &= \frac{\pi_{22}/\pi_{21}}{\pi_{12}/\pi_{11}} \\ &= \frac{\pi_{11} \pi_{22}}{\pi_{21} \pi_{12}} \end{aligned}$$

is the odds ratio. It is sometimes called the cross product ratio, since it is the ratio of the products $\pi_{11} \pi_{22}$ and $\pi_{12} \pi_{21}$ of proportions from cells that are diagonally opposite.

Each odds Ω_i can be expressed as

$$\begin{aligned}\Omega_i &= \frac{\pi_{i2}/\pi_{i+}}{\pi_{i1}/\pi_{i+}} \\ &= \frac{\pi_{2(i)}}{\pi_{1(i)}}\end{aligned}$$

so

$$\theta = \frac{\pi_{2(2)}/\pi_{1(2)}}{\pi_{2(1)}/\pi_{1(1)}}$$

The row and column variables are independent if and only if $\Omega_1 = \Omega_2$ (and so $\theta = 1.0$). If $1 < \theta < \infty$, then individuals in row 2 are more likely to be in column 2 than are individuals in row 1, i.e.

$\pi_{2(2)} > \pi_{2(1)}$. If $0 < \theta < 1$, individuals in row 2 are less likely to be in column 2 than are individuals in row 1, i.e. $\pi_{2(2)} < \pi_{2(1)}$.

For sample cell frequencies $\{x_{ij}\}$, a sample analog of θ is

$$\hat{\theta} = \frac{n_{11} n_{22}}{n_{21} n_{12}}$$

The value of θ does not change if both cell frequencies within any row are multiplied by a nonzero constant, or if both cell frequencies within any column are multiplied by a constant. So $\hat{\theta}$ estimates the same characteristic (θ) even if disproportionately large or small samples are selected from the various marginal categories of a variable. In particular, it estimates the same characteristic regardless of whether sampling is full multinomial or independent multinomial. It also takes the same value if the orientation of the table is reversed so that the rows become the columns and the columns become the rows.

If the order of the rows or the order of the columns is reversed, the new value of θ is simply the inverse of the original value. So two values of θ that are the inverse of one another (such as 3 and $1/3$) represent the same degree of association, but in opposite directions.

The odds ratio is a multiplicative function of the cell proportions. Its logarithm is an additive function, i.e.

$\log \theta = \log \pi_{11} - \log \pi_{12} = \log \pi_{21} + \log \pi_{22}$ and may equal any real number. The log odds ratio is symmetric about the independence value of 0.0 in the sense that a reversal of the two rows or the two columns results in a change of its sign.

1.6.2.1 Incidence of Colds Example

Pauling (1971) describes a double-blind study to evaluate the effect of ascorbic acid (vitamin C) on the common cold. One group of 140 skiers received a placebo, while a second of 139 received 1 g of ascorbic acid per day. The incidence of colds was recorded and is shown in Table 1.5.

Table 1.5: Incidence of common colds

Treatment	No Cold	Cold
Ascorbic acid	122	17
Placebo	109	31

The odds of catching a cold for the ascorbic acid group are $17/122 = .14$, while the odds for the placebo group are $31/109 = 0.28$. The ratio of these odds is $0.28/0.14 = (122 \times 31)/(109 \times 17) = 2.04$. This means that the odds of catching a cold were 2.04 times higher for the placebo group than for the ascorbic acid group. This odds ratio is significantly higher than 1.0, so it is plausible that administration of vitamin C helped to prevent the occurrence of colds.

1.6.3 Odds Ratio for rxc Tables

For the general rxc table odds ratios can be formed using each of $\binom{r}{2} = r(r-1)/2$ pairs of rows in combination with each of the $\binom{c}{2} = c(c-1)/2$ pairs of columns. For rows a and b and columns c and d, the odds ratio $(\pi_{ac} \pi_{bd})/(\pi_{bc} \pi_{ad})$ uses four cells occurring in a

rectangular pattern (see Figure 1), and there are $\binom{r}{2} \binom{c}{2}$ odds ratios of this type. The independence of the two variables is equivalent to the condition that all these population odds ratios equal 1.0.

However, there is much redundant information when the entire set of these odds ratios is used to characterize the association in a table.

1.6.3.1 Local Odds Ratios

A basic set of $(r-1)(c-1)$ odds ratios is

$$\theta_{ij} = \frac{\pi_{ij} \pi_{i+1,j+1}}{\pi_{i,j+1} \pi_{i+1,j}}, \quad i=1, \dots, r-1, \\ j=1, \dots, c-1$$

Figure 1.1: General Odds Ratio $\pi_{ac} \pi_{bd} / \pi_{bc} \pi_{ad}$

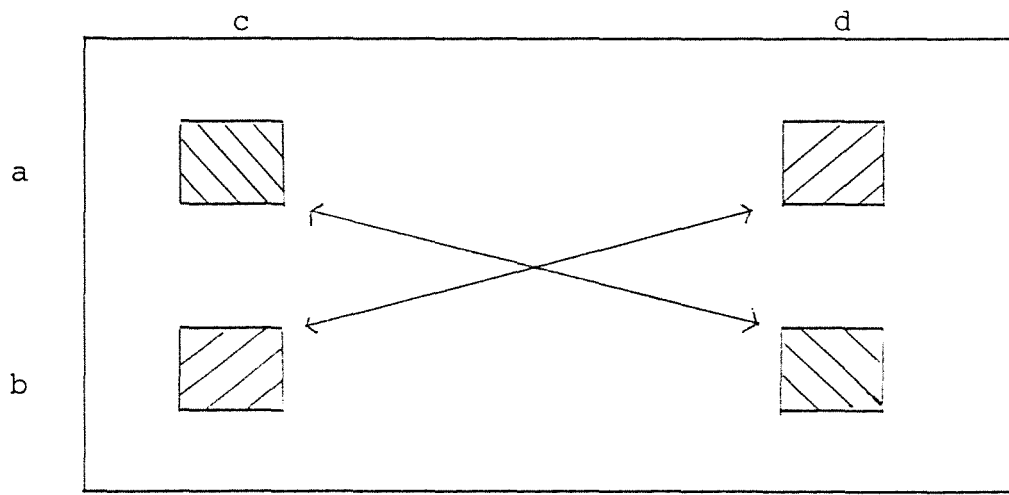
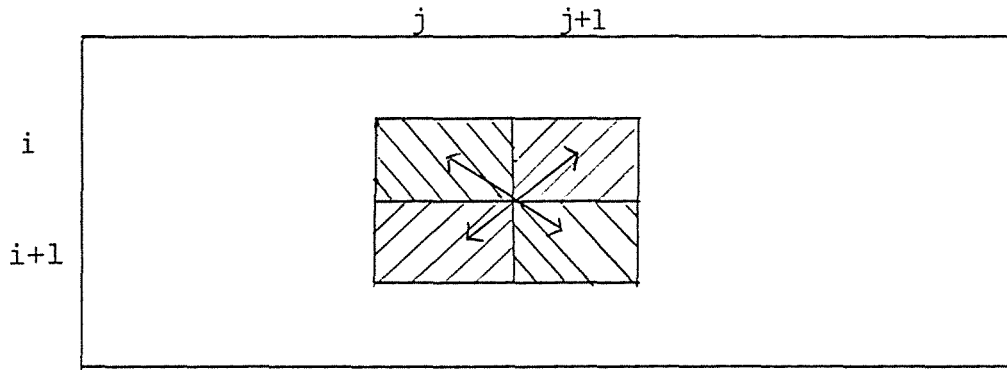


Figure 1.2: Local Odds Ratio θ_{ij}



This basic set determines all $\binom{r}{2} \binom{c}{2}$ odds ratios that can be formed from pairs of rows and pairs of columns. Independence of the two variables is therefore also equivalent to the condition that the odds ratios in the basic set are equal to one.

These odds ratios are formed using cells in adjacent rows and adjacent columns, as illustrated in Figure 1.2. Their volumes describe the relative magnitude of "local" associations in the table, so they are called local odds ratios.

1.6.3.2 Local-Global Odds Ratios

Another family of odds ratios is

$$\theta'_{ij} = \frac{\left(\sum_{b \leq j} \pi_{ib} \right) \left(\sum_{b > j} \pi_{i+1,b} \right)}{\left(\sum_{b > j} \pi_{ib} \right) \left(\sum_{b \leq j} \pi_{i+1,b} \right)}$$

$$i = 1, \dots, r-1,$$

$$j = 1, \dots, c-1$$

These odds ratios are local in the row variable but "global" in the column variable, since all c categories of the column variable are used in each odds ratio (see Figure 1.3). They are particularly meaningful when a distinction is made between response and explanatory variables.

1.6.3.3 Global Odds Ratios

A third family of odds ratios is

$$\theta''_{ij} = \frac{\left(\sum_{a \leq j} \sum_{b \leq j} \pi_{ab} \right) \left(\sum_{a > i} \sum_{b > j} \pi_{ab} \right)}{\left(\sum_{a \leq i} \sum_{b > j} \pi_{ab} \right) \left(\sum_{a > i} \sum_{b \leq j} \pi_{ab} \right)}$$

These measures are the regular odds ratios computed for the 2x2 tables corresponding to the $(r-1)$ $(c-1)$ ways of collapsing the row and column classification into dichotomies. They treat row and column variables alike and describe associations that are global in both variables (see Figure 1.4).

Figure 1.3: Local-Global Odds Ratio θ'_{ij}

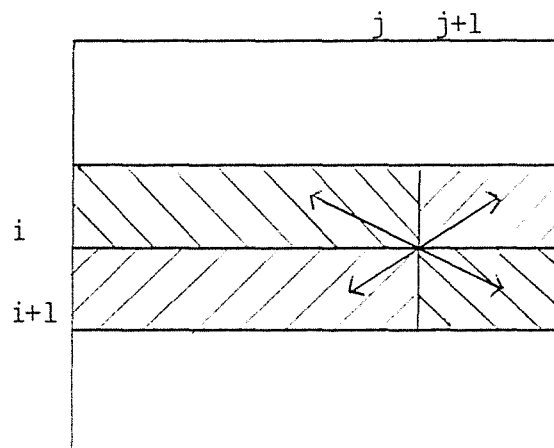
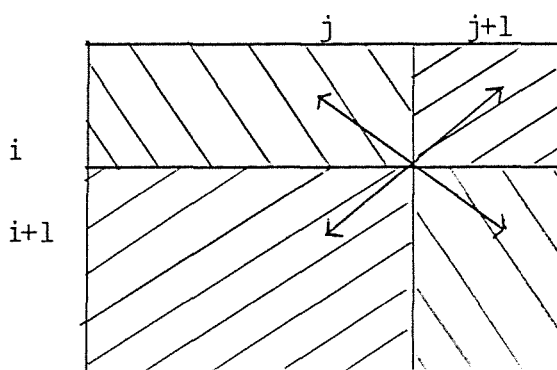


Figure 1.4: Global Odds Ratio θ''_{ij}



For local, local-global and global odds ratios, independence is equivalent to all log odds ratios equalling zero. An association described by one of these measures is referred to as "positive" or "negative" according to the sign of the log odds ratio.

If all $\log \theta_{ij} > 0$, then all $\log \theta'_{ij} > 0$. If all $\log \theta'_{ij} > 0$, then all $\log \theta''_{ij} > 0$. The converses of these statements are not true (Agresti, 1984). The condition that all local log odds ratios be positive is therefore the most stringent of three possible definitions for "uniformly positive association".

The less localized the odds ratio, the more precise its sample value tends to be as an estimation of its population value, since the standard error involves the inverses of larger sample totals. So if all the $\{\theta_{ij}\}$ are approximately equal, if the $\{\theta'_{ij}\}$ are approximately equal, and if the $\{\theta''_{ij}\}$ are approximately equal, the sample estimates of the third set will tend to be smoothest.

1.6.3.4 Dumping Severity Example

We will illustrate these three types of odds ratios for ordinal variables using the data in Table 1.6, from Grizzle, Starmer and Koch (1969). The data refer to a comparison of four different operations for treating duodenal ulcer patients. The operations

correspond to removal of various amounts of the stomach. Operation A is drainage and vagotomy, B is 25% resection and vagotomy, C is 50% resection and vagotomy, and D is 75% resection. The categories of operation are ordered, with A being the least severe operation and D corresponding to the greatest removal of stomach. The variable "dumping severity" describes the extent of a possible undesirable side effect of the operation. The categories of this variable are also ordered, with the response "none" representing the most desirable result.

Table 1.6: Dumping severity and operation

Operation	Dumping Severity			Total
	None	Slight	Moderate	
A	61	28	7	96
B	68	23	13	104
C	58	40	12	110
D	53	38	16	107
TOTAL	240	129	48	417

Table 1.7 contains the sample values $\{\hat{\theta}_{ij}\}$, $\{\hat{\theta}'_{ij}\}$ and $\{\hat{\theta}''_{ij}\}$ of the ordinal odds ratios.

To illustrate the calculation of the values in Table 1.7:

$$\hat{\theta}_{12} = \frac{28 \times 13}{23 \times 7} = 2.26$$

$$\hat{\theta}'_{12} = \frac{(61+28) \times 13}{(68+23) \times 7} = 1.82$$

$$\hat{\theta}''_{12} = \frac{(61+28) \times (13 \times 12 \times 16)}{(68+23+58+40+53+38) \times 7} = 1.86$$

Table 1.7: Values of Ordinal Odds Ratios for Dumping Severity Data

	$\hat{\theta}_{ij}$		$\hat{\theta}'_{ij}$		$\hat{\theta}''_{ij}$	
	j 1	2	1	2	1	2
i 1	0.74	2.26	0.92	1.82	1.38	1.86
2	2.04	0.53	1.69	0.86	1.74	1.33
3	1.04	1.40	1.14	1.44	1.55	1.53

The value of $\hat{\theta}_{12}$ means that the estimated odds that dumping is moderate instead of slight is 2.26 times higher for operation B than for A.

The value of $\hat{\theta}'_{12}$ means that the estimate odds that dumping is moderate instead of none or slight is 1.82 times higher for operation B than A.

The value of $\hat{\theta}''_{12}$ means that the estimated odds that dumping is moderate instead of none or slight is 1.86 times higher when some stomach is removed (operations B, C, D) than when none is removed (A).

All three sets of measures indicate a generally positive association, though the $\{\hat{\theta}''_{ij}\}$ show the most consistency.

1.7 Estimation

For all the models discussed in this thesis, the parameters and expected cell counts are estimated by the method of Maximum Likelihood (Lindgren, 1976, p.269).

This well-known statistical principle gives parameter estimates with certain known properties (e.g. asymptotic efficiency, consistency, asymptotic normality with known parameters, etc.) as well as giving

rise to powerful likelihood-ratio tests which can be used to test whether specific models fitted are feasible.

1.8 Model Testing

To test the goodness-of-fit of the various models, we can use either of the following two statistics:

$$\chi^2 = \sum_i \frac{(n_i - m_i)^2}{m_i}$$

$$G^2 = 2 \sum_i n_i \log \frac{n_i}{m_i}$$

which are asymptotically equivalent. Under the null hypothesis, both χ^2 and G^2 are asymptotically distributed as chi-square. χ^2 is the Pearson chi-square statistic (Pearson, 1900), and G^2 is a likelihood-ratio (LR) statistic, known as the "deviance" in the terminology of generalized linear models. Although both tests usually lead to very similar conclusions, we will use G^2 as the LR statistic is much more useful in testing significance of model terms.

1.9 Structural and Sampling Zones

Zero entries in contingency tables are of two types - structural and sampling zeroes. Structural (fixed) zeros occur when it is impossible to observe values for certain combinations of the variables, e.g. males who have had a hysterectomy. Sampling (random) zeroes are due to sampling variation and the relatively small size of the sample when compared with the large number of cells; they disappear when the sample size is increased sufficiently.

When structural zeros occur in a table, it is still possible to analyse the data using models which will be discussed in Section 8.2.

When sampling zeroes are scattered haphazardly throughout the table, there are usually no problems - the appropriate models are fitted in the normal manner.

However, sometimes the zero entries are placed in such a way that when computing estimated values to satisfy the constraints, if one zero entry is given a positive value, then another must be given a negative value. If the extra constraint that all estimated values must be non-negative is applied, then these entries will have estimated values of zero. Table 1.9 gives an example of such a case:

Table 1.9: A table with two-dimensional marginal total equal to zero

	Y1	Y2		Y1	Y2
X1	0	5		6	10
X2	0	12		5	8

The n_{+11} marginal total is zero. Thus, any model which requires this marginal total to be fitted must necessarily estimate the (1,1,1) and (2,1,1) cells as zero.

It is this circumstance which has given risen to much debate about the "correct" degrees of freedom applying to the deviance in such a case.

There are three views stated in the literature. The first and most widely stated view is that in order to test the goodness-of-fit of a model that uses a set of observed marginal totals with at least one zero entry, the degrees of freedom associated with the test statistic must be reduced (Bishop et al, 1975; Fienberg, 1980; Brown and Fuchs, 1983 and 1984; Aston and Wilson, 1984). This means that if an observed marginal entry is zero, then both the observed and estimated entries for all cells included in that total must be zero, and so the fit of the model for those cells is known to be perfect. As a result, the degrees of freedom associated with the fit of the zero cell values must be deleted. The formula for the degrees of freedom is given as

$$df = (n_C - n_Z) - (n_P - n_N)$$

where

n_c = number of cells in the table,

n_z = number of cells with estimated values equal to zero,

n_p = number of parameters specified in the model,

n_n = number of parameters that cannot be estimated because of zero marginal totals.

However, in a recent paper, Baker et al (1985) have asserted that such a treatment is incorrect. They state that if a zero occurs in a margin that was fixed prior to the experiment, then by definition the cells in the table contributing to that margin are structural zeros (and are weighted out of all analyses of the table). Therefore, by extension, if a table is analysed "conditional on a margin that was not actually fixed in the experiment, then cells in the table that were not structural zeroes in the experiment will become structural zeroes in the analysis if they contribute to a zero cell in the conditioning margin". However, after having dealt with these "structural" zeroes, if any other zero cells remain which contribute to a margin which is not conditioned on, then no adjustment whatsoever is to be made to the degrees of freedom.

The third view is that of Stirling (1986) who asserts that both the previous two methods are incorrect. He states that to obtain the correct degrees of freedom for any model, one should always use the formula

df = difference between the number of estimable parameters for the model in question and for the saturated model.

He explains that if this formula is used, then "it makes no difference whether or not structural zeroes are kept in the data, whether the margins are classified as responses or explanatory variables, or whether log-linear or logistic models are used when there is a single binary response". However, to correctly apply the formula, "we must correctly identify all estimable parameters. This has been incorrectly done by some previous authors".

It can therefore be seen that the literature on methods for sparse contingency tables is still controversial.

1.9.1 Monkey Example

An example of a table containing both sampling and structural zeroes is Table 1.8, which is taken in a slightly modified form from Fienberg (1980, p.146). The table gives the distribution of genital display among six squirrel monkeys (labelled R to W). For each display there is an active and passive participant, but a monkey never displays towards himself. Thus the dashes in the table indicate structural zeroes. There are also several sampling zeroes such as in cell (1,6) where there is no a priori reason to suppose that the event is impossible. We will assume that the opportunity was not available to observe monkey T as an active participant.

Table 1.8: Genital display in a colony of squirrel monkeys

Active Participant	Passive Participant					
	R	S	T	U	V	W
R	-	1	5	8	9	0
S	29	-	14	46	4	0
U	2	3	1	-	38	2
V	0	0	0	0	-	1
W	9	25	4	6	13	-

1.10 Loglinear Models

Let $\underline{n}' = (n_1, \dots, n_I)$ and $\underline{m}' = (m_1, \dots, m_I)$ denote the observed and expected counts for the I cells in the table. For simplicity, we will use a single index, though the table may be multi-dimensional.

Loglinear models have the form

$$\log m_i = \underset{\sim}{x}'_i \underset{\sim}{\beta}$$

where $\underline{\beta}$ is a $p \times 1$ vector of parameters and \underline{x}'_i is a row vector of known constants, the choice of which depends on what kind of association one wishes to model. In the nomenclature of analysis of variance \underline{x}'_i is the i th row of the $I \times p$ design matrix X , i.e.

$$\log \underline{m} = X \underline{\beta}$$

Many of the models discussed in this thesis are simply special cases of the more general category of loglinear models. They can be used to model many kinds of association, and so are probably the most common models used in practice for contingency tables.

1.10.1 Fitting Loglinear Models

Loglinear models can be fitted quite easily using either the Newton-Raphson algorithm or the Iterative Proportional Fitting algorithm. These are discussed in Appendix 2.

1.11 Linear Models

Linear models relate the expected cell count to a linear function of parameters. The two common methods of specifying these models are:

- (i) directly, in a form such as $A \underline{\pi} = X \underline{\beta}$, or
- (ii) indirectly, in terms of constraints.

They are not as commonly used as loglinear models, as they usually specify fairly unusual kinds of relationships between the variables of a contingency table. Nevertheless, they form a powerful and useful class of models which can be used to test specific hypotheses that could not normally be tested using loglinear models.

1.11.1 Linear models specified as $A \underline{\pi} = X \underline{\beta}$

Consider the cell counts in a contingency table as making up an $I \times 1$ vector \underline{m} . The cell probabilities corresponding to these counts make up an $I \times 1$ vector $\underline{\pi}$. The vector $\underline{\pi}$ may correspond to

- (i) a Poisson or single multinomial distribution, or to
- (ii) a product-multinomial distribution.

In the former case $\sum \pi_i = 1$, while in the product-multinomial case the set of I cells is comprised of several subsets, each of which corresponds to a separate multinomial sample, and the sum of the elements of $\underline{\pi}$ over each subset is unity.

If $\underline{\beta}$ is a $K \times 1$ vector of unknown parameters, A is a known $J \times I$ matrix with linearly independent rows, and X is a known $J \times K$ matrix, with linearly independent columns, with $I > J > K$, then we can write the expected cell probabilities in terms of a linear function of the model parameters as

$$A\underline{\pi} = X\underline{\beta}.$$

Haber (1985) discusses linear models which are formulated in this way.

1.11.2 Linear models specified in terms of constraints

Suppose we have some hypothesis about the cell counts which can be specified in terms of E constraints. We can write the constraints as

$$F\underline{\pi} = \underline{0}$$

where F is an $E \times I$ matrix with E linearly independent rows. Further constraints are imposed by the sampling design. These constraints guarantee that the sum of the probabilities within each sample will be equal to one (or equivalently that the sum of the counts within each sample will be equal to the correct marginal total). They can be written

$$D'\underline{\pi} = \underline{1}_S$$

where S is the number of samples ($S > 1$) and $D = \{d_{is}\}$ is the $I \times S$ matrix defined by

$$d_{is} = \begin{cases} 1 & \text{if cell } i \text{ belongs to sample } s \\ 0 & \text{otherwise} \end{cases}$$

In terms of the cell counts the constraints can be written

$$D' R \underline{m} = \underline{1}_S$$

where R is the $I \times I$ diagonal matrix with diagonal elements $r_{ii} = 1/m_{i+}$, where m_{i+} is the marginal total of the i th sample ($i=1, \dots, S$), and off-diagonal elements zero.

Thus, the constraints on the cell counts can be written as

$$L \underline{m} = \underline{c}$$

where L' is the $I \times (E+S)$ matrix $L' = (F' : R'D)$ and \underline{c}' is the $1 \times (E+S)$ vector consisting of E zeroes and S ones, i.e.

$$\underline{c}' = (\underline{0}_E' : \underline{1}_S').$$

A vector \underline{a} which satisfies these constraints is

$$\underline{a} = R^{-1} \underline{a}^*$$

where R^{-1} is the $I \times I$ diagonal matrix with diagonal elements $r_{ii} = m_{i+}$, and \underline{a}^* is the $I \times 1$ vector with i th element $a_i = 1/B_i$ where B_i is the number of cells in the i th sample.

1.11.3 Fitting Linear Models

Linear models specified in terms of constraints can be easily fitted using the algorithm of Wedderburn (1974). Details of this are given in Appendix 3.

Linear models specified as $A \underline{\pi} = X \underline{\beta}$ can be most easily fitted by reformulating in terms of constraints so that we can then apply Wedderburn's algorithm. Appendix 3 gives further details.

1.12 Other Models

As mentioned previously, most models discussed in this thesis are either linear or loglinear, and so can be fitted using the general

algorithms appropriate for these. Where a model does not fall into one of these two classes, details of estimation methods will be given separately.