STATISTICAL METHODS

FOR

ORDERED CATEGORICAL RESPONSES

AND

SPARSE CONTINGENCY TABLES

by

Jocelyn Ruth Dale

submitted

for the degree of

Doctor of Philosophy

and the

Diploma of Membership of Imperial College

June 1983

UNIVERSITY OF LONDON

Department of Mathematics

Imperial College of Science and Technology

London SW7 2BZ

Ma tahi ma rua ma toru ma wha

kue ea

Hōani E Pā E Mā

ko tēnei taku aroha

ABSTRACT

Techniques are examined for the statistical analysis of multinomial or product multinomial distributions in the form of contingency tables. Focus is on models for ordered responses and methods appropriate for tables with low cell means, i.e. sparse contingency tables. First, the literature on statistical models for ordered categorical responses and measures of association for bivariate ordered responses is reviewed. A family of global cross-ratio models for bivariate, ordered responses in the presence of a covariate is set up, with association between the response variables given as the cross-ratio of the 2×2 table of probabilities obtained by dichotomizing both marginal variables and the marginal cumulative probabilities specified in terms of generalized linear models. Interpretation of parameters and fit by maximum likelihood are discussed briefly and demonstrated in the analysis of datasets from medicine and zoology. An approach for the elimination of marginal parameters by maximizing a data-dependent cross-section of the likelihood is examined briefly. A score statistic to test for no association is shown to be equivalent to Spearman's rho. Next, asymptotic distributions of some statistics for testing the fit of a single sparse multinomial are reviewed, and their use illustrated by the analysis of a dataset from linguistics. The Pearson X^2 and likelihood ratio G^2 statistics for the fit of a sequence of independent, sparse multinomials to a dataset are shown to have normal distributions, as the number in the sequence approaches infinity. Simple sparseness conditions are given and results obtained for cell probabilities either known or dependent on a finite-dimensional parameter. Closeness of the exact distribution of X^2 to the normal is studied for finite samples using Edgeworth series and Pearson curve approaches. The skewness of X^2 is tabulated for various multinomials.

ACKNOWLEDGEMENTS

During the preparation of this thesis the author acknowledges the following:

Dr P.J. Brown, supervisor, for many helpful discussions;

Professor D.R. Cox for several constructive suggestions;

Dr P. McCullagh, who supervised my first year of research at Imperial College;

Dr R.L. Smith for helpful discussions on the work in §6.3.3;

Mr T.W. Dougall for the dataset of Table 3.3 and a useful correspondence;

Mrs. M. Robertson for accurate typing;

Various others who provided manuscripts, references, and critical comments;

Funding from a New Zealand National Research Advisory Council Postgraduate Research Fellowship and an Overseas Research Students' Fees Support Scheme award;

Colleagues, family and spouse for their solidarity.

TABLE OF CONTENTS

ABSTRACT 3

ACKNOWLEDGEMENTS 4

TABLE OF CONTENTS 5

LIST OF TABLES 8

LIST OF FIGURES 9

CHAPTER 0 INTRODUCTION 10

0.1 Preamble 10

0.2 Bivariate Discrete Ordered Responses 11

0.3 Goodness of Fit for Sparse Contingency Tables 13

0.4 Organization of Thesis 15

PART A STATISTICAL METHODS FOR ORDERED CATEGORICAL RESPONSES 16

NOTATION 17

CHAPTER 1 UNIVARIATE ORDERED CATEGORICAL RESPONSES: STATISTICAL MODELS 21

1.1 Introduction 21

1.2 Ordered Categorical Responses 21

1.3 Discrete Hazard Functions 24

1.4 Cumulative Probabilities 27

1.5 Cell Probabilities 30
1.5.1 Multinomial logistic and related models 30
1.5.2 Loglinear models and orthogonal polynomials 32

1.6 Discussion 34

CHAPTER 2 MEASURES OF ASSOCIATION FOR BIVARIATE ORDERED RESPONSES 36

2.1 Introduction: Measuring Association 36

2.2 Concordance 38

	2.3	Correlation	40
	2.4	Cross-Ratios	42
	2.5	Pooling Association for Several Response Tables	45
	2.6	The Mantel-Haenszel Test for Several 2×2 Tables	46
	2.7	A Score Statistic for r×c Tables	47
	2.8	Some Locally Most Powerful Test Statistics	49
	2.9	Discussion	52
CHAPTER 3		GLOBAL CROSS-RATIO MODELS	54
	3.1	Preamble	54
	3.2	Global Cross-Ratio Models	55
		3.2.1 General definition	55
		3.2.2 A parametric family of GCR models	57
	3.3	Fitting GCR Models	60
		3.3.1 Maximum likelihood estimation	60
		3.3.2 Closed-form starting estimates	61
	3.4	Examples	63
		3.4.1 Patients' postoperative pain level and medication requirements	63
		3.4.2 Colouring of first-winter male blackbirds	67
	3.5	Estimating the GCR in the Presence of Nuisance Parameters	74
		3.5.1 Introduction	74
		3.5.2 Coping with nuisance parameters: Anscombe's approach	75
		3.5.3 Empirical comparison of Anscombe and full likelihood methods	76
		3.5.4 Score test for no global association	84
	3.6	Summary	87
CHAPTER 4		DISCUSSION	88
	4.1	Introduction	88
	4.2	Local Association Models	89
	4.3	Comparison of Local and Global Association	93
		4.3.1 Introduction	93
		4.3.2 Lehmann's measures of dependence	93
		4.3.3 Effect of re-ordering the marginal variables	96
		4.3.4 Effect of splitting or combining marginal categories	98
		4.3.5 Interpretation as odds ratios	101
		4.3.6 Comparison with continuous distributions	103
	4.4	Conclusion	105

PART B GOODNESS-OF-FIT STATISTICS FOR SPARSE CONTINGENCY TABLES 107

NOTATION 108

CHAPTER 5 THE SPARSE MULTINOMIAL: DISTRIBUTION OF X^2 AND G^2 113

5.1 Sparseness 113

5.2 Known Cell Probabilities 114

5.3 G^2 for Direct Loglinear Models 118

5.4 Example: Measuring the Similarity of Two Poems 126

CHAPTER 6 SPARSE PRODUCT MULTINOMIALS: DISTRIBUTION OF X^2 AND G^2 131

6.1 Introduction 131

6.2 Completely Specified Distributions 132
6.2.1 Notation and preliminaries 132
6.2.2 Asymptotic distributions of X^2 and G^2 134

6.3 Parametric Case 139
6.3.1 Preliminaries 139
6.3.2 Regularity conditions for the maximum likelihood estimator of $\underset{\sim}{\theta}$ 140
6.3.3 Asymptotic distribution of X^2 and G^2 144
6.3.4 Example: Binary logistic regression 156

CHAPTER 7 DISCUSSION 160

7.1 Introduction 160

7.2 The Skewness of X^2 and the Behaviour of Upper Percentiles 161
7.2.1 Introduction 161
7.2.2 Closeness on the logit scale to normal percentiles 161
7.2.3 The family of Pearson curves 163
7.2.4 The Edgeworth series expansion 163
7.2.5 Calculation of the skewness of X^2 165

7.3 Further Aspects of Goodness of Fit 171

7.4 Concluding Remarks 174

APPENDIX I First Partial Derivatives of the Log Likelihood for GCR Models 175

APPENDIX II Hoadley's Conditions for Maximum Likelihood Estimators 177

REFERENCES 180

LIST OF TABLES

2.1 Comparison of measures of association 53

3.1 Number of patients, classified by type of operation, level of pain and medication requirements 64

3.2 GCR model fit to postoperative status dataset 65

3.3 Number of blackbirds, classified by colour score and period observed 68

3.4 GCR model fit to blackbird colouring dataset 70

3.5 True marginal probabilities for simulation study 77

3.6 Asymptotic relative efficiency of full likelihood to Anscombe estimate of Δ for generated datasets 79

3.7 Comparison of Anscombe and full likelihood estimators: simulated datasets 80

3.8 Comparison of Anscombe and full likelihood methods: published datasets 82

3.9 Likelihood-based confidence intervals for Δ: published datasets 83

4.1 Comparison of global and local association 106

5.1 The poem "Winter Retrospect": number of singly-occurring words in successive groups of twenty for both poems 127

5.2 Observed and expected counts under a Poisson model: dataset of Table 5.1 129

7.1 Percentage points which are close to normal percentiles 162

7.2 Pearson X^2 statistic for multinomial distributions 167

7.3 Pearson X^2 statistic: sparse multinomial distributions 169

LIST OF FIGURES

3.1 Fitted cutpoint position versus time period, for orbital ring colourings and upper mandible colourings: blackbird dataset, model (a) 72

3.2 Variation of $\tilde{\Delta}_A$ with sample size 78

6.1 Regions (stippled area) where $g_t(\underset{\sim}{\theta}) < 0$ for all t. 158

0

INTRODUCTION

0.1 Preamble

The statistical analysis of tables of counts is required in diverse disciplines including biology, sociology, psychology and demography. Although the data often have a continuous distribution, it is for reasons of economy or inaccuracy of measurement that they are recorded categorically and tabulated as counts. Other types of data are inherently discrete. In recent years there has been a surge of interest in fitting statistical models to contingency tables: for example loglinear models (Bishop, Fienberg and Holland, 1975) and analysis of deviance methods (Nelder & Wedderburn, 1972) have widespread use. This thesis gives two contributions to methods for the analysis of count data. They are outlined in §§0.2 and 0.3. The emphasis in the first is on building a statistical model for discrete bivariate responses (Chapters 1 to 4) and in the second (Chapters 5 to 7) on asymptotic distributions for standard goodness-of-fit statistics when expected cell frequencies remain small and the number of entries in the contingency table becomes large. In both cases, these developments are put in the context of the current literature.

0.2 Bivariate Discrete Ordered Responses

What makes a discrete, ordered response, and how can it be modelled? This question is dealt with in Chapter 1, where types of ordered response scale are discussed, along with various statistical models for them in the literature. For a bivariate response, it is necessary to model the relationship between the component responses, as well as their marginal behaviour. The current writing on measures of association for bivariate discrete, ordered responses is discussed in Chapter 2. The family of global cross-ratio (GCR) models defined in Chapter 3 is for bivariate responses dependent on a fixed covariate. Both sets of marginal cumulative probabilities depend on the covariate through a linear logistic model (McCullagh, 1980). Association is measured by global cross-ratios: the contingency table of probabilities for a given value of the covariate is collapsed into a 2×2 table by dichotomizing both component variables, summing over all the cells in each resulting quadrant and taking cross-ratio of the collapsed table - the GCR, ψ. The logarithm of ψ is allowed to depend parametrically on the position of the dichotomies and on the covariate. These two aspects of the bivariate response are unified into one statistical model, using an extension of models proposed by Pearson and Heron (1913) and by Plackett (1965) - models incorporating a surface of constant association in the continuous case. Examples of the fit of GCR models to datasets of blackbird colourings and post-operative status of patients are given in §3.4. When ψ is kept constant for all pairs of dichotomies, a simpler GCR model ensues. Anscombe's method (1981, Chapter 12) for estimating $\ln \psi$ is evaluated and compared with the maximum likelihood method, using published and simulated data. A locally most powerful test statistic for testing H_0: $\ln \psi = 0$

against H_1: $\ln \psi > 0$ is derived. The main competitors to GCR models are Goodman's (1979) local association models. A critical comparison of the two types of model, with emphasis on the association measures, is carried out in Chapter 4. It is concluded that the GCRs have application even when the boundary between adjacent marginal categories is poorly defined, whereas local association is applicable when all marginal categories are distinct entities and it is of interest to compare adjacent categories.

0.3 Goodness of Fit for Sparse Contingency Tables

In order to ascertain the overall fit of a statistical model to a contingency table of counts, one uses goodness-of-fit statistics, notably Pearson's X^2 and the likelihood ratio G^2. Assuming multinomial sampling and large means for each cell in a contingency table of fixed dimension, under the assumed model both G^2 and X^2 approach a χ^2 distribution as the number of observations tend to infinity - see for example Cramér (1946, 416-417). However, much of the count data collected has many cells with zero entries, suggesting small cell means. Thus an alternative asymptotic approach was called for. It is necessary to define properties of a contingency table which yield small cell means, and to develop an asymptotic theory that satisfies these conditions. The work of Morris (1975) and Koehler (unpublished) for a sequence of multinomial distributions with number of cells becoming large is discussed in Chapter 5. Both the X^2 and G^2 statistics, suitably standardized, have asymptotically normal distributions under the true model. Morris's results are for completely specified distributions: Koehler shows that they apply also for the likelihood ratio statistic to loglinear models which have closed-form maximum likelihood estimators. This chapter concludes with the analysis of a dataset concerning the similarity of two poems. In Chapter 6 we consider asymptotics for the product multinomial situation, with the number of multinomials becoming large while the cell means and dimension of each multinomial remain moderate. For the Pearson and likelihood ratio statistics asymptotic normality is established, both for known distributions, and under certain regularity conditions for cell probabilities dependent on a common finite-dimensional unknown $\underset{\sim}{\theta}$, estimated by maximum likelihood. The latter case is illustrated by a binary logistic model.

Approximations of upper moderate percentage points of the Pearson X^2 distribution to normal percentiles are studied in Chapter 7. Closeness is measured on the logit scale. Using Edgeworth series and Pearson curve approaches we conclude that for percentiles near the upper 5% point it is enough for the skewness to be small and positive.

0.4 Organization of the Thesis

The body of this thesis is presented in two sections, each self-contained in notation, development and conclusion. General introductions are given in §§0.2 and 0.3, and summaries of notation appear at the beginning of each part: the appendices and common list of references appear at the end.

A

STATISTICAL METHODS FOR ORDERED CATEGORICAL RESPONSES

NOTATION A

In this a summary of notation, brief definitions are given, followed by references to the first use in the text. Minor local notation is not cited. The order is alphabetical with capitals before lower case and Roman type preceding Greek type.

$A_{ij}, B_{ij}, C_{ij}, D_{ij}$	number of observations in quadrants $\{Z_1 \leq i,\ Z_2 \leq j\}, \ldots \{Z_1 > i,\ Z_2 > j\}$, §2.4
$A_{ij/k}, B_{ij/k}, C_{ij/k}, D_{ij/k}$	$\frac{1}{2}$ + (sum over quadrant) Y_{ijk}, §3.3.2
$F_{ij}, F_{ij}(x; \eta_x, \xi_x, \psi_x)$, $F_{ij}(x; \theta_x)$, $F_{ij}(\Delta; \eta, \xi)$	cumulative probability function for GCR model, §§4.3.3, 3.2.1, 3.2.1, 3.5.1.
I_j	$(t_{j-1}, t_j]$, §1.2
K	number of r×c tables in sequence, §2.1
M_H	one-sided Mantel-Haenszel statistic, §2.6
N_c, N_d, N_1, N_2	number of pairs of bivariate observations concordant, discordant, tied in the first co-ordinate, tied in the second co-ordinate, respectively, §2.2
P_c, P_d, P_1, P_2	proposition of pairs of bivariate observations concordant, discordant, tied in the first co-ordinate, tied in the second co-ordinate respectively, §2.2
$R_{i+}, R_{+j}; R_{i+k}, R_{+jk}$	row and column relative ranks, §§2.7, 3.5.4
R_j	number at risk of failure at t_j+, §1.3
$\bar{R}$	$\frac{1}{2}(1+n^{-1})$, §3.5.4

R_M	Mantel's score statistic, §2.7
S	sample space for Z, §1.2
S_j	number censored during I_j, §1.3
T	continuous random variable with cutpoints t_j such that $Z = z_j$ if and only if $t_{j-1} < T \leq t_j$, §1.2
T_q^{ij}	-1,0,1 according as $Z_{qi} <, =, > Z_{qj}$, q = 1,2, for pair (Z_{1i},Z_{2i}), (Z_{1j},Z_{2j}), §2.2
T_Δ	$(\partial\ell/\partial\Delta)(\Delta,\tilde{\eta},\tilde{\xi})\,\vert\,\Delta = 0$, §3.5.4
T_ψ	$= T_\Delta$, §2.8
U_j	indicator variable, §1.5.1
X	$= \underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m$, matrix of covariates, §3.3.1
Y_j	number of observations in interval I_j, §1.3
$Y, Y_1,\ldots,Y_m$	r×c table of counts, multinomially distributed, §2.1
Y_{ij}, Y_{ijk}	cell count, §2.1
Z, Z_1, Z_2	discrete, ordered random variable, with c,r,c categories respectively, §§1.2, 2.1
$\underset{\sim}{Z}$, $\underset{\sim}{Z}(\underset{\sim}{x})$	bivariate, discrete, ordered response, §§2.1, 4.2
a_{ik}	row category score, i = 1,...,r, §2.7
b_{jk}	column category score, j = 1,...,c, §2.7
c	number of rows, §2
f_i	polynomials orthogonal over category scores, §1.5.2
g, g_1, g_2	link function, (1.1)
h_j	discrete hazard function, §1.3
$\ell(\Delta,\underset{\sim}{\eta},\underset{\sim}{\xi};y)$, $\ell(\underset{\sim}{\theta};y_1,\ldots,y_m,X)$, $\ell(\psi,\underset{\sim}{\lambda};\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m)$, $\ell_A(\Delta;y)$	log likelihood function, (3.17), §§3.3.1, 2.8, (3.18)

n, n_k	number of observations in table Y, Y_k, §§1.3, 2.1, 3.3.1
p_j, $p_j(\underset{\sim}{x})$	probability, §1.2
t_j	j-th cutpoint of T, §1.2
$\Delta, \Delta_{ij}, \Delta_i$	log GCR parameter (3.4)-(3.7), §3.5.1
$\hat{\Delta}$	maximum likelihood estimator of Δ, §3.5.3
$\tilde{\Delta}_A$	Anscombe estimator of Δ, §3.5.2
$\Sigma_{ij}(\Delta)$	§3.5.4
α	Agresti's measure of concordance, §2.2; multiplier for marginal probability, §4.3.5
$\underset{\sim}{\alpha}_j, \overset{h}{\alpha}_j, \overset{c}{\underset{\sim}{\alpha}}_j, \alpha_{1i}, \alpha_{2j}$	cutpoint parameter, §§1.3, 1.4, 3.3.2
α_j	discrimination parameter, §1.5.1
α_i	coefficient of orthogonal polynomial, §1.5.2
$\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2$	regressor parameter for marginal cumulative probabilities, (3.8), (3.9)
γ	Kruskal's measure of concordance, (2.2); association parameter in local association model, (2.12), (2.13)
$\underset{\sim}{\gamma}$	regressor parameter for log GCR, (3.6)
$\gamma_j(\underset{\sim}{x})$	cumulative probability, §1.4
δ_j	log GCR parameter, (3.5)
$\underset{\sim}{\eta}, \underset{\sim}{\eta}_{\underset{\sim}{x}}$	marginal row probabilities, §§4.3.2, 3.2.1
$\tilde{\eta}_i$	$n^{-1} \sum_{q=1}^{i} Y_{q+}$, §3.5.2
θ_{ij}	local association parameter, §§2.4, 4.2
θ'_{ij}	$\pi_{ij}\pi_{rc}/(\pi_{ir}\pi_{rj})$, §4.2
θ^*_{ij}	local association parameter for permuted table, §4.3.3
κ_j	category parameter in rating model, §1.5.1
$\kappa_j(x)$	odds of cumulative probabilities §1.4

$\lambda_0(t)$	continuous-time baseline hazard rate, §1.3
$\hat{\underset{\sim}{\lambda}}_0$	maximum likelihood estimator of $\underset{\sim}{\lambda}$ when $\psi = \psi_0$, §2.8
μ	true value of response in rating model, §1.5.1
μ_i, ν_j	parameter in local association model II, §4.2
π_j	cell probability, $j = 1,\ldots,c$, §1.2
π_{ij}, $\pi_{ij/k}$	cell probability, $i = 1,\ldots,r$; $j = 1,\ldots,c$; $\Sigma_i\Sigma_j\pi_{ij/k} = 1$, §§2.1, 2.8, 4.2, 4.3.2
$\pi_{ij}(\underset{\sim}{x}_k;\underset{\sim}{\theta})$	cell probability, §3.3.1
ρ	correlation, §§2.4, 4.3.6
τ_j	threshold parameter, §1.5.1
τ	Kendall's measure of rank correlation, (2.1)
ϕ	parameter in local association model II, §4.2
ϕ_j	score parameter in rating model, §1.5.1
ψ, ψ_x	$(r-1)\times(c-1)$ matrix of GCR parameters, §3.2.1
ψ	measure of association, §2.1
$\psi, \psi(\underset{\sim}{x})$	scalar GCR, constant for all cutpoints, §2.4
ψ_{ij}, ψ_{ijx}	GCR parameter, §§2.4, 3.1, 4.3.3
ψ^*_{ij}	GCR parameter: permuted table, §4.3.3
$\hat{\psi}_A, \hat{\psi}_c, \tilde{\psi}_c, \hat{\psi}_W$	closed-form estimators of the GCR, §2.4
$\tilde{\xi}_j$	$n^{-1}\sum_{s=1}^{j} Y_{+s}$, $j = 1,\ldots,c-1$ §3.5.2
ξ_{jx}, ξ_j	cumulative column probability, $j = 1,\ldots,c-1$, §§3.2.1, 4.3.2.

1

UNIVARIATE ORDERED CATEGORICAL RESPONSES : STATISTICAL MODELS

1.1 Introduction

In this chapter we discuss different types of ordered categorical response. Statistical models presented in the literature are examined, in particular those which are in terms of discrete hazard functions (§1.3), cumulative probabilities (§1.4) and cell probabilities (§1.5). Most methods incorporate dependence on a covariate. Many of the data encountered in practice have a response variable which is ordered, and thus these techniques will prove beneficial both in their own right and for embedding in larger statistical analyses. Properties of the models presented in §§1.3-1.5 are summarized in §1.6.

1.2 Ordered Categorical Responses

What constitutes an ordered categorical response? In this subsection they are defined, and classified according to two criteria. Consider a random variable Z defined on a sample space S. Then Z is categorical or discrete if S is discrete. Further, Z is ordered if there is a total ordering on S, i.e. for all z_1 and z_2 elements of S, either $z_1 \leq z_2$ or $z_2 \leq z_1$. We consider Z to be the outcome of a single trial, and thus the elements of S are possible outcomes, or categories. Two helpful ways of classifying the scale of ordered categorical random variables are as answers to

(a) is the random variable inherently discrete, or is there an underlying continuous random variable which gives rise to it;

(b) is it qualitative or quantitative?

Formally, an ordered categorical random variable is inherently discrete if the categories are separate entities: points between adjacent categories do not make sense, e.g. the number of lambs born to a ewe in a season: 0, 1, 2, 3. By contrast, for a random variable which is discretized continuous there exists in principle a continuous random variable T such that $Z = z_j$ if and only if $t_{j-1} < T \leq t_j$. The points t_j are called cutpoints. Without loss of generality, assume that $t_{j-1} < z_j \leq t_j$ for all j. Thus the statement $Z = z_j$ is equivalent to $t_{j-1} < Z \leq t_j$. An example of a discretized continuous random variable is the proportion of a blackbird beak that is black: all, some, none - see §3.4.2. Reasons for discretization may be convenience, or inaccuracy of the measuring instrument. If the cutpoints are approximate, there may be a non-zero probability of misclassification. An ordered categorical random variable is quantitative if there is a metric, assigning distances between categories or between cutpoints. Then the element z_j, $j = 1,\ldots,c$ of S may be identified with $I_j = (t_{j-1}, t_j]$, $j = 1,c-1$, and $I_c = (t_{c-1}, \infty)$, i.e. $t_0 = -\infty$ and $t_c = \infty$, where there are c possible outcomes for Z. Otherwise, the random variable is qualitative, as in the example concerning the blackbird beak.

For quantitative discretized continuous random variables it is not always appropriate to assign a score to each category. An average value depends on the distribution of T on the continuous scale. It is more appropriate to define the categories in terms of the cutpoints t_j separating them, e.g. for income: less than £5000; £5000 to £10 000; greater than £10 000. Methods using numerical scores are appropriate for quantitative random variables, and best for inherently discrete scales. A qualitative, discrete random variable occurs in voting in

certain constituencies in the United Kingdom, where the vote may be cast for Labour, Liberal-Social Democrat (LSDP) or Conservative, if one considers that these three possible outcomes are indeed univariate ordered! An assignment of scores to these three categories presupposes measures of distance between adjacent categories. The implication then is that the distance between Labour and Conservative is the sum of the Labour-LSDP and LSDP-Conservative distances.

We now turn to ways of modelling these different types of response. In general, consider the modelling of a sequence of probabilities $p_j(\underset{\sim}{x})$, $0 < p_j(\underset{\sim}{x}) < 1$, as functions of parameters α_j, $\underset{\sim}{\beta}$, $j = 1,\ldots,c$, where $\underset{\sim}{x}$ is a fixed p-dimensional covariate. Now the right side of the relationship $p_j(\underset{\sim}{x}) = \alpha_j + \underset{\sim}{x}^T\underset{\sim}{\beta}$ will take values between zero and one only for certain values of α_j and $\underset{\sim}{\beta}$. Thus it is useful, if estimates are to be obtained using unconstrained optimization techniques, to transform the left side to another scale that has range $(-\infty,\infty)$. Let g be a function with an inverse which maps (0,1) onto the whole of $(-\infty,\infty)$, and set

$$g\{p_j(\underset{\sim}{x})\} = \alpha_j + \underset{\sim}{x}^T\underset{\sim}{\beta}. \qquad (1.1)$$

Then g is called a <u>link function</u>, since it connects the regression model to the probability itself, and equation (1.1) together with a sampling regime for observations according to the cell probabilities $\pi_j = \Pr[Z = z_j]$ gives rise to a <u>generalized linear model</u> for the p_j. This term is used in a more general sense than Nelder and Wedderburn (1972) used it: the relationship between the p_j and the π_j must be specified. We proceed to discuss generalized linear models for three types of p_j which usefully model ordered categorical responses: discrete hazard functions, $p_j = \pi_j/(\pi_1+\ldots+\pi_j)$; cumulative probabilities, $p_j = \pi_1+\ldots+\pi_j$; and the cell probabilities π_j.

1.3 Discrete Hazard Functions

When the categories in the ordered response have a sense of forward direction as in age or failure time, then it is appropriate to model the probability of an event in the j-th category, given that the process goes at least that far. In keeping with the survival literature, we will describe an event as a failure before or after time t_j, although the scale could be something else, e.g. distance travelled before an automobile tyre wears to 4 mm, or the number of attempts before learning a simple skill. For clarity, the model is expressed in terms of the underlying continuous random variable. Let T = time of failure, and suppose that the cutpoints $t_1 < \ldots < t_{c-1}$ divide the range of T into c contiguous failure-time categories I_j, as defined in §1.2. Then the discrete hazard function for the j-th category, $j = 1,\ldots,c-1$ is

$$h_j = \Pr[T \in I_j | T > t_{j-1}]$$
$$= \Pr[t_{j-1} < T \leq t_j]/\Pr[T > t_{j-1}].$$

This is the probability of an individual failing in the interval I_j, given its survival until t_{j-1}. The unconditional probability π_j of failure in I_j may be expressed in terms of the h_i, $i \leq j$ as follows:

$$\pi_j = \Pr[T \in I_j] = \prod_{i=1}^{j-1} \Pr[T > t_i | T > t_{i-1}]\Pr[T \in I_j | T > t_{j-1}]$$
$$= \prod_{i=1}^{j-1} (1-h_i)h_j.$$

Consider a sample of n independent observations with Y_j failing in interval I_j, $j = 1,\ldots,c$. Then the number at risk of failure at t^+_{j-1}, the beginning of I_j, is $R_j = n - Y_1 - \ldots - Y_{j-1}$. Thus

$$\Pr[Y_1 = y_1,\ldots,Y_c = y_c] = \binom{n}{y_1,\ldots,y_c} \prod_{j=1}^{c} \pi_j^{y_j}$$

$$= \prod_{j=1}^{c} \binom{r_j}{y_j} (1-h_j)^{r_j - y_j} h_j^{y_j} \quad . \tag{1.2}$$

Hence the distribution of $\underset{\sim}{Y} = (Y_1,\ldots,Y_c)^T$ is a product of c binomials, the j-th being $\text{bin}(R_j,h_j)$. In the literature the $\underset{\sim}{Y}$ are assumed dependent on a fixed covariate $\underset{\sim}{x}$ such that for a link function g(.)

$$g\{h_j(\underset{\sim}{x})\} = \alpha_j + \underset{\sim}{x}^T \underset{\sim}{\beta}, \qquad j = 1,\ldots,c-1 \ . \tag{1.3}$$

Then, conditional for each j on the risk set $R_j(\underset{\sim}{x})$, the likelihood is a product of independent binomial factors and $\underset{\sim}{\alpha} = (\alpha_1,\ldots,\alpha_{c-1})$ and $\underset{\sim}{\beta}$ may be estimated consistently by maximizing the resulting partial likelihood function with respect to $(\underset{\sim}{\alpha},\underset{\sim}{\beta})$ - see Cox and Hinkley (1974, Chapter 9) and Cox (1975).

Suppose that in addition to Y_j failing during I_j, S_j individuals are censored, i.e. they "disappear" during I_j without being known to fail, for $j = 1,\ldots,c-1$. Then $R_j = n - Y_1 - S_1 - \ldots - Y_{j-1} - S_{j-1}$. The bias introduced by discretizing the failure time distribution to assume that all censoring in I_j occurs at t^+_{j-1} concerned Thompson (1977). He put forward an argument to insert the factor $\{1 - h_j(\underset{\sim}{x})\}^{-s_j/2}$ in the partial likelihood function as a better representation of the censoring throughout the interval I_j. If very few censorings occur,

or if they are known to occur discretely at the beginning of each interval I_j, then such an adjustment would have little effect.

The logit link function $g(p) = \ln\{p/(1-p)\}$ has been discussed by Cox (1972, §6) and by Thompson (1977): Fienberg (1980) calls $h_j/(1-h_j) = \pi_j/(\pi_1+\ldots+\pi_{j-1})$ a <u>continuation ratio</u>. The complementary log-log link $g(p) = \ln\{-\ln(1-p)\}$ was used by Bartlett (1978) and Prentice and Gloeckler (1978). Kalbfleisch and Prentice (1973) obtained an expression for the α_j of (1.3) in terms of the baseline hazard function $\lambda_0(t)$ for the underlying continuous-time proportional hazards model

$$\alpha_j^h = \ln \int_{I_j} \lambda_0(t)\, dt, \qquad (1.4)$$

i.e. the hazard for $\underset{\sim}{\beta} = 0$.

Thus the baseline hazard rate is cumulated between observation times t_j.

Observe that (1.3) does not incorporate the failure times t_j or I_j except through the hazard functions h_j. Thus this family of statistical models is appropriate for qualitative responses. Reversal of the ordering of the categories in general changes the model. It may be expressed conditionally on the risk set as a product of binomial factors, and hence fitted by partial likelihood (Cox, 1975). If there is an appreciable amount of censoring and the discretization is coarse, modifications similar to those suggested by Thompson (1977) and others may be needed.

1.4 Cumulative Probabilities

Another natural way of exploiting the ordering in a discrete response $Z(\underset{\sim}{x})$ is to consider the cumulative probability $\gamma_j(\underset{\sim}{x}) = \Pr[Z(\underset{\sim}{x}) \leq t_j]$, $j = 1,\ldots,c$, with $-\infty = t_0 < t_1 < \ldots < t_c = +\infty$. Consider a link function $g(.)$ such that for a covariate, $\underset{\sim}{x}$

$$g\{\gamma_j(\underset{\sim}{x})\} = \alpha_j - \underset{\sim}{x}^T\underset{\sim}{\beta}, \qquad j = 1,\ldots,c-1 \ . \tag{1.5}$$

To keep the cell probabilities $\pi_j(\underset{\sim}{x}) = \gamma_j(\underset{\sim}{x}) - \gamma_{j-1}(\underset{\sim}{x})$ positive, let $\alpha_1 < \alpha_2 < \ldots < \alpha_{c-1}$. Equation 1.5 defines a generalized linear model for cumulative probabilities. It is discussed with several examples by McCullagh (1980). Useful link functions include the probit, logit and complementary log-log: the covariate $\underset{\sim}{x}$ may be continuous or discrete.

The linear logistic model, with link $g(p) = p/(1-p)$, is also known as the proportional odds model, since

$$\kappa_j(\underset{\sim}{x}) = \gamma_j(x)/\{1-\gamma_j(\underset{\sim}{x})\} = \kappa_j \exp(-\underset{\sim}{x}^T\underset{\sim}{\beta}),$$

and hence the ratio of odds on $Z \leq t_j$ at different values of the covariate do not depend on j:

$$\kappa_j(\underset{\sim}{x}_1)/\kappa_j(\underset{\sim}{x}_2) = \exp\{(\underset{\sim}{x}_2-\underset{\sim}{x}_1)^T\underset{\sim}{\beta}\} \ .$$

It is interesting to compare this model with the logit model for discrete hazard functions. Setting $\underset{\sim}{x} = \underset{\sim}{0}$ and the cutpoint parameters for the hazard model as Roman letters $A_1,\ldots,A_{c-1}$, by equating cell probabilities it can be shown that

$$\exp(A_j) = \frac{\exp(\alpha_j) - \exp(\alpha_{j-1})}{1 + \exp(\alpha_j)} , \qquad j = 1,\ldots,c-1 .$$

Thus the hazard parameter A_j depends not only on the spacing between the $\exp(\alpha_j)$, but also on the magnitude of $\exp(\alpha_j)$. A useful property of the logit model for cumulative probabilities is that it dichotomizes the response scale at successive cutpoints. Tukey (1977, Chapter 15) uses $\frac{1}{2}\ln\{\gamma_j/(1-\gamma_j)\}$, called folded logs or flogs in an exploratory manner. When c = 2 the linear logistic model reduces to the binary logistic model discussed by Cox (1970).

The complementary log-log link $g(p) = \ln\{-\ln(1-p)\}$ yields another formulation for the discretized proportional hazards model discussed in §1.3. The cumulative probabilities rather than the hazards are transformed, to yield cutpoint parameters

$$\alpha_j^c = \ln \int_0^{t_j} \lambda_0(u)\, du, \qquad j = 1,\ldots,c-1$$

where $\lambda_0(u)$ is the baseline hazard rate in the continuous-time model. The relationship to the discretized hazard cutpoint parameter is, by (1.4)

$$\exp(\alpha_j^c) = \sum_{q=1}^{j} \exp(\alpha_q^h) .$$

Thus both are discretized versions of the continuous-time proportional hazards model.

All the generalized models for cumulative probabilities have the property that if adjacent categories are combined the model is unchanged except for the loss of a cutpoint parameter. Further, the cutpoints enter the models only through the probabilities, and thus it is enough for the response under consideration to be qualitatively

ordered. Invariance of the model (except for a sign change for parameters) under reversing the orderings on the categories is satisfied for the logit and probit (cumulative normal) links: the complementary log-log link $g(p) = \ln\{-\ln(1-p)\}$ does not have the property $g(1-p) = -g(p)$.

To allow the scale of the response to change with $\underset{\sim}{x}$, McCullagh (1980, §6) proposes an extension to (1.5), replacing the right side by $\{\alpha_j - \underset{\sim}{x}^T \underset{\sim}{\beta}\}/\tau_x$. A similar effect is achieved by Farewell (1982), For the model (1.5) with complementary log-log link, he sets $\alpha_j - \alpha_{j-1} = \ddot{o}$, $j = 2, \ldots, c-1$, and postulates a gamma distribution for $\exp(\alpha_1)$.

1.5 Cell Probabilities

1.5.1 Multinomial logistic and related models

A class of models for ordered categorical responses expressed in terms of cell probabilities has been discussed by Andrich (1979). Threshold parameters τ_j, $0 = \tau_0 < \tau_1 < \ldots < \tau_c = 1$, and the "true" value of the quantity being measured μ are defined on a continuous scale. In order to obtain the discretized response, c-1 independent Bernouilli trials are performed to decide whether the true value μ satisfies $\mu \leq \tau_j$, $j = 1,\ldots,c-1$. The outcome is a vector $\underset{\sim}{U} = (U_1,\ldots,U_{c-1})^T$ of indicator variables, with $U_j = 1$ if $\mu \leq \tau_j$ and $U_j = 0$ otherwise. Then assuming a logistic distribution for each component of U,

$$\Pr[U_1 \leq u_1,\ldots,U_{c-1} \leq u_{c-1}] \propto \prod_{j=1}^{c-1} \exp\{\alpha_j(\mu-\tau_j)u_j\},$$

where $\alpha_j \geq 0$ is the discrimination at threshold τ_j. However, some of the possible outcomes $\underset{\sim}{U}$ are nonsensical, e.g. if $\underset{\sim}{U} = (0,1)$, then $\mu > \tau_1$ and $\mu < \tau_2$, a contradiction. Therefore we restrict consideration to the values of $\underset{\sim}{U}$ consistent with the ordering on the τ_j, i.e. those containing a sequence of ones followed by a sequence of zeros. There are only c such $\underset{\sim}{U}$. Thus, defining Z = number of zeros in $\underset{\sim}{U}$ yields

$$\Pr[Z = 0 \,|\, \underset{\sim}{\alpha},\mu,\underset{\sim}{\tau}] = 1/\gamma(\underset{\sim}{\alpha},\mu,\underset{\sim}{\tau})$$

$$\Pr[Z = j \,|\, \underset{\sim}{\alpha},\mu,\underset{\sim}{\tau}] = \exp\{\sum_{i=1}^{j} \alpha_i(\mu-\tau_i)\}/\gamma(\underset{\sim}{\alpha},\mu,\underset{\sim}{\tau}), \quad j = 1,\ldots,c-1, \tag{1.6}$$

where

$$\gamma(\underset{\sim}{\alpha},\mu,\underset{\sim}{\tau}) = 1 + \sum_{j=1}^{c-1} \exp\{\sum_{i=1}^{j} \alpha_i(\mu-\tau_i)\}.$$

For a clearer parameterization, let the category and score parameters respectively be

$$\kappa_j = -\sum_{i=1}^{j} \alpha_i \tau_i \quad \text{and} \quad \phi_j = \sum_{i=1}^{j} \alpha_i ,$$

with $\kappa_0 = \phi_0 = 0$, whence

$$\Pr[Z = j \,|\, \underset{\sim}{\kappa}, \mu, \underset{\sim}{\phi}] = \exp(\kappa_j + \phi_j \mu)/\gamma(\underset{\sim}{\kappa}, \mu, \underset{\sim}{\phi}) \tag{1.7}$$

where

$$\gamma(\underset{\sim}{\kappa}, \mu, \underset{\sim}{\phi}) = \sum_{j=0}^{c-1} \exp(\kappa_j + \phi_j \mu).$$

There is an ordering implicit in the ϕ_j and κ_j.

When the discrimination parameters α_j for all thresholds are equal, then

$$\Pr[Z = j \,|\, \underset{\sim}{\kappa}, \mu] \propto \exp(\kappa_j + j\mu) , \tag{1.8}$$

which is a multinomial logistic model, a member of the multiparameter exponential family and thus has desirable properties for estimation. In the educational testing context, when an experiment is designed so that questions of varying difficulty δ are given to subjects of differing ability β, then to model the response by the subjects to the questions, Andrich (1978) proposes substituting $\mu = \beta - \delta$ in (1.7). This is the rating model. It assumes that the relationship between ability of the subject and difficulty of the question have a linear relationship. Another model similar to (1.7) is discussed in unpublished work by Anderson. He considers $Z(\underset{\sim}{x})$, $\underset{\sim}{x}$ a fixed covariate.

Replacing $\mu = \underset{\sim}{x}^T\underset{\sim}{\beta}$ in (1.7) leads to a stereotype model with

$$\Pr[Z(\underset{\sim}{x}) = j \mid \underset{\sim}{\kappa}, \underset{\sim}{\beta}, \underset{\sim}{\phi}] \propto \exp(\kappa_j - \phi_j \underset{\sim}{x}^T\underset{\sim}{\beta}) \quad , \tag{1.9}$$

which is multinomial logistic.

All these models estimate the relative spacing of categories. Interpretation of the parameters, especially in the rating model, appears straightforward. In the constant discrimination case, the category labels contribute directly to the model. One can test for discrimination between the categories by relaxing the ordering restriction on the τ_j and hence the ϕ_j. In an example in Andrich (1979, §6.2), a discovery of threshold parameters out of order led him to combine the offending adjacent categories. This is similar to restricting the ϕ_j to being ordered. Clearly departures of this type from the model require special thought: it may mean that the response is not univariate ordered.

1.5.2 Loglinear models and orthogonal polynomials

When the ordered categorical response Z is quantitative and a score a_j is assigned to each response category, then a possible approach is to use loglinear models, i.e.

$$\ln \Pr[Z = a_j] = \beta_j a_j \quad , \qquad j = 1,\ldots,c.$$

For bivariate responses of this type, Nelder and Wedderburn (1972) and Haberman (1974) model the components as orthogonal polynomials in the scores. For the univariate response, let

$$\ln \Pr[Z = a_j] = \sum_{i=0}^{k} \alpha_i \, f_i(a_j), \qquad j = 1,\ldots,c \tag{1.10}$$

where $f_i(.)$ are orthogonal polynomials of degree i,

i.e. $$\sum_{j=1}^{c} f_i(a_j)\ f_{i'}(a_j) = 0, \qquad i \neq i' ,$$

and $k \leq c-1$. If the category scores are equally spaced, the required polynomials are given in Pearson and Hartley (1970): for the general case, an algorithm for generating the polynomials is given in Seber (1977, p.220). A polynomial of degree zero yields equiprobability; a polynomial of degree one postulates that the logarithms of the cell probabilities are linear in the scores a_j.

The application of orthogonal polynomials to log cell probabilities is analogous to the treatment of quantitative independent factors in the analysis of variance method for continuous responses. Their best use is in determining the degree of the overall polynomial description, i.e. the value of i for which $\alpha_j = 0$, $j \geq i$. Detailed interpretation of the model is probably best done in terms of the cell probabilities themselves.

1.6 Discussion

In §§1.3-1.5 we have examined various types of ordered categorical response, and ways of modelling them, with emphasis on describing the probability structure. In this final subsection we discuss the application of these models to the different situations defined in §1.2.

The generalized linear models for discrete hazard functions (§1.3) and the orthogonal polynomial methods (§1.5.2) are appropriate for inherently discrete responses: they treat the categories as separate, well-defined entities. While the presence of an underlying continuum helps in the understanding of the family of models for cumulative probabilities in §1.4, it is not necessary for their definition. Thus they are also applicable to inherently discrete scales if interest is in combining all the categories lower than a given level, for successive levels.

For response scales which are discretized continuous, consider the multinomial logistic models of §1.5.1 and the generalized linear models for cumulative probabilities (§1.4). Both are expressed in terms of cutpoints on a continuum. The parameter α_j in (1.6) is small if the j-th threshold is poorly defined. Further, the relative spacing between thresholds is fitted using the parameters τ_j. The models for cumulative probabilities (§1.4) are invariant (except for loss of a parameter) under the amalgamation of adjacent categories, a property not shared by the other models.

Most of the techniques discussed in this chapter are suitable for qualitative responses, since they do not require category or cutpoint scores. Ordering is imposed on the categories by conditioning (§1.3), or by ordering parameters (§§1.4, 1.5.1). Observe that under

reversal of the ordering on the response categories, the parameters are unchanged except for sign in the models of §1.5.1, and the cumulative probability models of §1.4 with symmetric link functions (e.g. logit, probit). The only methods specifically for quantitative responses - the equal discrimination model of §1.5.1 and the orthogonal polynomial model of §1.5.2 - require a score for each category. If adjacent categories are merged, that score needs to be adjusted.

Fitting all these models is done under the assumption for more than one observation of multinomial or Poisson sampling. For the discrete hazard models of §1.3, the partial likelihood methods of Cox (1972, 1975) enable the appropriate likelihood to be expressed as a product of c-1 independent binomial random variables. The loglinear models in (1.8) and (1.10), belonging to the exponential family, may be fitted using maximum likelihood, for example the statistical computing packages GLIM (Baker and Nelder, 1978) or Genstat (Alvey et al., 1977). The generalized linear models for cumulative probabilities may be fitted by maximum likelihood on McCullagh's interactive computer program PLUM (unpublished, 1979).

We have discussed a variety of statistical models for the discrete ordered response. Each is potentially useful - the examples in the literature bear this out. However the comments in this subsection are only appropriate as guidelines. A model in a particular situation is useful only when it sheds light on the problem in hand.

2

MEASURES OF ASSOCIATION FOR BIVARIATE ORDERED RESPONSES

2.1 Introduction: Measuring Association

The relationship between the components of a bivariate response is often interesting. In this chapter we review ways of describing association between two discrete, ordered responses. Three main types of association are covered: concordance (§2.2), correlation (§2.3) and cross-ratios (§2.4). In later subsections, tests for the same scalar association in several tables are examined.

Consider a random variable $\underset{\sim}{Z} = (Z_1, Z_2)$ with $\pi_{ij} = \Pr[Z_1 = i, Z_2 = j]$, $i = 1,\ldots,r$; $j = 1,\ldots,c$. Suppose that n independent observations occur, with the same distribution as $\underset{\sim}{Z}$. They are summarized in a table of counts Y with (i,j)-th entry Y_{ij}, and $\Sigma_i\Sigma_j Y_{ij} = n$. We denote a sequence of K such tables by $Y_1,\ldots,Y_k$. The k-th response table Y_k has a total of n_k observations, and its (i,j)-th cell has probability $\pi_{ij/k}$ and Y_{ijk} observations, $i = 1,\ldots,r$; $j = 1,\ldots,c$ with $\Sigma_i\Sigma_j\pi_{ij/k} = 1$. Y_k is generated by n_k independent random vectors $\underset{\sim}{Z}_k = (Z_{1k}, Z_{2k})$. A quantity obtained by summing over an index is written with the index replaced by a "plus" sign, thus: $\Sigma_i Y_{ijk} = Y_{+jk}$. An instance of a random variable is written in lower case, e.g. y_{+jk}. Any further distributional assumptions will be made as necessary.

We say that there is <u>positive association</u> between the components of Z if high ("more positive") values of Z_1 and Z_2 occur together: association is <u>negative</u> if high values of Z_1 tend to occur with low or more negative values of Z_2. These definitions apply even if the ordered

random variables Z_1 and Z_2 are qualitative (see §1.2). In an r×c contingency table, (r-1)×(c-1) degrees of freedom are available to describe the relationship between Z_1 and Z_2. However we will only consider association as a <u>scalar</u> ψ, measured or tested for by a scalar statistic. Useful measures of association take positive or negative values in line with the definition above, and in particular equal zero when Z_1 and Z_2 are independent. For example, the sample correlation provides this. Since the relationship between Z_1 and Z_2 is not necessarily described completely by a scalar, it is possible that $\psi = 0$ even if Z_1 and Z_2 are related in some way. Other valuable properties for measures of association, or at least a monotonic transformation of them, are interpretability, exploitation of the ordering of the marginal variables, robustness against minor changes in marginal distributions, and magnitude not dependent on sample size.

In the following subsections we proceed to discuss some measures of association given in the literature for the bivariate discrete, ordered response. Some of the statistics that we discuss are strictly <u>tests</u> for no association. For two statistical models which incorporate a measure of association ψ as a parameter, we present locally most powerful statistics for testing H_0: $\psi = \psi_0$ (no association) vs H_1: $\psi > \psi_0$. A discussion of measures of association when both responses are nominal is in Goodman and Kruskal (1959). For the case of one component response nominal and the other ordinal, see Agresti (1981).

2.2 Concordance

The basic notion of positive association is that both marginal variables tend to take large values together. Two bivariate observations $\underset{\sim}{z}_i$, i = 1,2 are said to be concordant if the slope of the line joining them is positive, and discordant if the slope is negative. This definition holds even if there is an ordering on possible values of the $\underset{\sim}{z}_i$, but they do not take quantitative values. For random variables taking only a few values, there is usually a non-zero probability that either component of the pair of observations has the same value. In an r×c table Y of responses as defined in §2.1, let N_c, N_d, N_1, N_2 be the number of pairs of observations which are concordant, discordant and tied in the first and second co-ordinate respectively. The population values P_q, q = c,d,1,2, denote the proportion of pairs in each classification. We now present measures of association based on them.

Kendall's (1948) measure τ of rank correlation

$$\hat{\tau} = \frac{(N_c - N_d)}{\sqrt{\{(N_c+N_d+N_1)(N_c+N_d+N_2)\}}} \qquad (2.1)$$

has been used widely for continuous data. A formula for $\mathrm{var}(N_c-N_d)$ for no association and ties (N_2,N_2 > 0) is given by Kendall [1948, equation 4.3]. The quantity τ can be interpreted as the correlation of the indicator variables T_q^{ij}, q = 1,2, which take values -1,0,1 according as $z_{qi}-z_{qj}$ is negative, zero or positive, q = 1,2, for all $\frac{1}{2}n(n-1)$ pairs of observations $\underset{\sim}{z}_i$, $\underset{\sim}{z}_j$ in the sample.

Another measure based on concordance is Goodman and Kruskal's γ [1954, §6]:

$$\hat{\gamma} = (N_c - N_d)/(N_c + N_d) \; . \qquad (2.2)$$

Only untied pairs of observations are considered, and $\hat{\gamma}$ gives the exceedance of concordant pairs over discordant pairs, as a proportion of the untied pairs of observations. An approximate variance is given in Goodman and Kruskal [1963, equation 3.5.7]. When there are no ties, $\gamma = \tau$.

More intuitive is Agresti's (1980) α:

$$\hat{\alpha} = N_c/N_d \tag{2.3}$$

which compares concordant and discordant pairs as a ratio. Now $\alpha = (1+\gamma)/(1-\gamma)$ and as $P_1, P_2 \to 0$, α becomes the odds on concordancy. For K response tables, Agresti gives a χ^2 test for equality of α in each table and presents as an overall estimate of α, a sum of the observed α in each table, the estimates weighted inversely in proportion to their estimated variance.

These three statistics do not take into account the degree of concordance between pairs. Slightly concordant and strongly concordant pairs of observations give equal contribution to these statistics. We now turn to measures of association which can take account of the "spacing" of the marginal categories.

2.3 Correlation

A natural way to measure association between continuous random variables is by means of the correlation coefficient. When a numerical score is associated with each marginal category of a discrete, ordered response, then Pearson's product-moment correlation of the scores may be calculated. If the marginal average ranks are used as scores, this measure is just Spearman's rank correlation for categorical data - see Stuart (1963). These cases are examined further in §§2.7 and 2.8. Clearly the correlation is dependent on the choice of score. Observe that Kendall's τ (§2.2) is the correlation of indicator variables, defined for each pair of observations.

Consider a bivariate contingency table of counts Y obtained from n independent observations $\underset{\sim}{Z}$. Assume that each $\underset{\sim}{Z}$ is an observation from a coarse discretization of an underlying bivariate normal distribution with correlation ρ. The correlation estimated from Y by the product-moment method has been found to be smaller in magnitude than that of the underlying continuous distribution. Lancaster and Hamdan (1964) noted it for Pearson and Lee's dataset on height of fathers and daughters: it was also shown by Olsson (1980) in a simulation study. A preferable method under these circumstances is the polychoric series estimate of ρ, introduced by Lancaster and Hamdan (1964). Both marginal distributions are expressed in terms of orthogonal step functions, and each step function is expanded as a Fourier series of Hermite-Chebyshev polynomials. This enables the noncentrality parameter of the asymptotic χ^2 distribution of the Pearson X^2 statistic for no association to be expanded via Hermite-Chebyshev polynomials as a power series in ρ, the correlation of the

underlying bivariate normal distribution. Equating the X^2 statistic for a table Y of n observations to its asymptotic expectation under H_1: $\rho \neq 0$ yields

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(Y_{ij}-n^{-1}Y_{i+}Y_{+j})^2}{n^{-1}Y_{i+}Y_{+j}} = (r-1)(c-1) + n\sum_{i=1}^{r-1}\sum_{j=1}^{c-1}\{\sum_{q=1}^{\infty} a_{iq}b_{jq}\rho^q\}^2 \qquad (2.4)$$

with a_{iq}, b_{jq} defined in terms of the marginal distributions, the orthogonal functions, and the bivariate normal distribution - Lancaster and Hamdan [1964, equation 17]. The polychoric series estimate of ρ is obtained by solving (2.4), with the series on the right terminated after only a few terms (e.g. q = 5 was used by Lancaster and Hamdan). This is a method of moments, since (2.4) equates the observed mean of X^2 for Y to the expected value.

Now the assumption of discretized bivariate normality is not always appropriate. It may be better to use a measure of association for bivariate, discrete responses, or a parameter from an appropriate discrete distribution, instead of this correlation estimate, which applies to an unobserved continuous distribution.

2.4 Cross-Ratios

For the 2×2 table, the cross-ratio $(\pi_{11}\pi_{22})/(\pi_{12}\pi_{21})$ is a measure of association which is invariant under multiplicative changes to a row or column. There are two interesting extensions of this to the ordered r×c table. Goodman (1979) has proposed the local association

$$\theta_{ij} = (\pi_{ij}\pi_{i+1,j+1})/(\pi_{i,j+1}\pi_{i+1,j}), \quad i = 1,\ldots,r-1;\ j = 1,\ldots,c-1.$$

If this is held equal to θ for all $(r-1)\times(c-1)$ possible subtables, then it gives rise to his uniform association model, a loglinear model [Goodman, 1979, equation 4.1a]. The parameter θ is estimated easily by maximum likelihood. Its interpretation in terms of odds on adjacent cell probabilities $\Pr[Z_1 = i \mid Z_2 = j]/\Pr[Z_1 = i+1 \mid Z_2 = j]$ vs. $\Pr[Z_1 = i \mid Z_2 = j+1]/\Pr[Z_1 = i+1 \mid Z_2 = j+1]$ applies best when both Z_1 and Z_2 are inherently discrete (§1.2).

The global cross-ratio, defined for each collapse of the table into four quadrants, is

$$\psi_{ij} = \Big(\sum_{a\le i}\sum_{b\le j}\pi_{ab}\sum_{a>i}\sum_{b>j}\pi_{ab}\Big)/\Big(\sum_{a\le i}\sum_{b>j}\pi_{ab}\sum_{a>i}\sum_{b\le j}\pi_{ab}\Big),$$

$i = 1,\ldots,r-1;\ j = 1,\ldots,c-1$. Closed-form estimation of the constant global cross-ratio $\psi = \psi_{ij}$ for all i,j has been considered by Clayton (1974, §7). Now in the notation of §2.1 let $\underset{\sim}{Z} = (Z_1, Z_2)$ be a bivariate, ordered response, and for each (i,j), $i = 1,\ldots,r$; $j = 1,\ldots,c$; divide the sample space into quadrants $\{Z_1 \le i,\ Z_2 \le j\}$, $\{Z_1 \le i,\ Z_2 > j\}$, $\{Z_1 > i,\ Z_2 \le j\}$, $\{Z_1 > i,\ Z_2 > j\}$. Consider n independent observations with the same distribution as $\underset{\sim}{Z}$, recorded in an r×c contingency table Y, and for each (i,j) let A_{ij}, B_{ij}, C_{ij}, D_{ij} be the number of observations falling in the above quadrants, in order. Clayton gives two estimators for ψ. Now

$$\ln \hat{\psi}_c = \frac{\sum_{i=1}^{r-1}\sum_{j=1}^{c-1} w_{ij} \ln\{(A_{ij}D_{ij})/(B_{ij}C_{ij})\}}{\sum_{i=1}^{r-1}\sum_{j=1}^{c-1} w_{ij}} , \qquad (2.5)$$

where $w_{ij} = n^{-5}(A_{i+1,c}-A_{i-1,c})(A_{r,j+1}-A_{r,j-1})A_{ic}C_{ic}A_{rj}B_{rj}$, $i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$, is a linear combination of observed log global cross-ratios. The weights w_{ij} are chosen to minimize the asymptotic variance of $\ln \hat{\psi}$ when $\psi = 1$. An extension to an r×c table of the Mantel-Haenszel statistic (1959, p.736) proposed by Clayton (1974) is

$$\tilde{\psi}_c = \frac{\sum_{i=1}^{r-1}\sum_{j=1}^{c-1} w'_{ij}A_{ij}D_{ij}}{\sum_{i=1}^{r-1}\sum_{j=1}^{c-1} w'_{ij}B_{ij}C_{ij}} ,$$

where the $w'_{ij} = n^{-3}(A_{i+1,c}-A_{i-1,c})(A_{r,j+1}-A_{r,j-1})$, $i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$, are such that when $\psi = 1$, var $\tilde{\psi}_c$ is minimized and

$$\text{var } \ln \hat{\psi}_c = \text{var } \ln \tilde{\psi}_c = \Big(\sum_{i=1}^{r-1}\sum_{j=1}^{c-1} w_{ij}\Big)^{-1} .$$

However this says nothing for the variance away from $\psi = 1$.

An estimator similar to $\ln \hat{\psi}_c$ is given by Anscombe (1981, Chapter 12). For each (i,j), let prime (') denote an increment of $\frac{1}{2}$ to the quadrant counts, e.g. $A'_{ij} = A_{ij} + \frac{1}{2}$. Then

$$\ln \hat{\psi}_A = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} w''_{ij} \ln\{(A'_{ij}D'_{ij})/(B'_{ij}C'_{ij})\} \qquad (2.6)$$

where $w''_{ij} \propto (A'^{-1}_{ij} + \ldots + D'^{-1}_{ij})^{-1}$, $i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$, and $\Sigma\Sigma w''_{ij} = 1$. This choice of w''_{ij} ignores correlation between the

terms of (2.6), but is shown by Anscombe to be approximately unbiased. Wahrendorf's (1980) closed-form estimator of $\ln \psi$ incorporates the correlations. Let C be the $(r-1)(c-1)\times(r-1)(c-1)$-dimensional variance-covariance matrix of the entries $\ln\{(A_{ij}D_{ij})/(B_{ij}C_{ij})\}$. Then

$$\ln \hat{\psi}_W = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} w^*_{ij} \ln\{(A_{ij}D_{ij})/(B_{ij}C_{ij})\} \qquad (2.7)$$

with w^*_{ij} the appropriate element of $\underset{\sim}{w}^* = C^{-1}\underset{\sim}{1}(\underset{\sim}{1}^T C^{-1}\underset{\sim}{1})^{-1}$. This statistic is superior to (2.5) and (2.6) since the weights are inversely proportional to the variance, for all ψ.

The parameter ψ may be embedded in statistical models - see Pearson and Heron (1913), Plackett (1965) and Chapter 3 of this thesis. The global cross-ratio may be interpreted as a ratio of odds on a cumulative probability: the odds on $\{Z_1 \leq i | Z_2 \leq j\}$ versus the odds on $\{Z_1 \leq i | Z_2 > j\}$, which is meaningful when Z_1 and Z_2 are discretized continuous (§1.2). A comparison of local and global association is set out in Chapter 4.

2.5 Pooling Association for Several Response Tables

We now turn to examining tests for common association in a sequence of bivariate tables of responses. The relationship of Mantel's score test for such tables to locally most powerful tests of zero association in local and global cross-ratio models is discussed in §2.8. Often the practitioner is presented with several two-way response contingency tables for which an overall evaluation of the association between the marginal responses is required. If the marginal distributions differ, then a misleading result may emerge if the tables are pooled and association evaluated for the resulting table. As recorded by Simpson (1951), the pooled association may even be of opposite sign to the associations in the contributing tables. The approach here is to combine the associations for all the contributing tables, in preference to combining the tables themselves.

2.6 The Mantel-Haenszel Test for Several 2×2 Tables

In a 2×2 contingency table, conditional on the marginal totals, association can be tested in terms of the distribution of the observation in cell (1,1). We adopt the notation of §2.1 for response tables Y_k, $k = 1,\ldots,K$ of dimension 2×2, and set the row and column totals at y_{i+k}, y_{+jk}, $i,j = 1,2$; $k = 1,\ldots,K$. If for each table there is no association between the margins, then Y_{11k} has a hypergeometric distribution, i.e.

$$\Pr[Y_{11k} = y] = \binom{y_{+1k}}{y}\binom{y_{+2k}}{y_{1+k}-y} \Big/ \binom{n_k}{y_{1+k}}, \quad y = 0,\ldots,\min(y_{1+k},y_{+1k}) .$$

The statistic proposed by Mantel and Haenszel (1959, p.735) for testing for a common association in the K tables is

$$M_H^2 = \frac{\{|\Sigma_k(Y_{11k}-y_{1+k}y_{+1k}/n_k)| - \frac{1}{2}\}^2}{\Sigma_k\, y_{1+k}y_{2+k}y_{+1k}y_{+2k}/\{n_k^2(n_k-1)\}} . \qquad (2.8)$$

The signed one-sided version, M_H, is also a useful test of association. It is easily shown by appealing to the Central Limit Theorem that as the number of tables becomes large, and under no association, $M_H^2 \sim \chi_1^2$. This test has been demonstrated by Birch (1964) to be uniformly most powerful against the alternative when the cross-product ratios are the same in each response table. The one-sided version may be used to test for say positive association. Asymptotically as the number of tables $K \to \infty$, M_H has the standard normal distribution. This result is helpful, as it enables the test to be performed on sparse contingency tables: it is not necessary that the expected frequency in each cell becomes large. Goodness-of-fit statistics for sparse contingency tables are dealt with in more generality in Part B.

2.7 A Score Statistic for r×c Tables

For bivariate categorical responses with more than two marginal categories, an obvious measure of association is the sample correlation, discussed briefly in §2.3. Fixed scores $a_{1k}<a_{2k}<\ldots<a_{rk}$ are assumed for the new categories and $b_{1k}<b_{2k}<\ldots<b_{ck}$, for the column categories of the k-th response table Y_k, $k = 1,\ldots,K$, and the usual product moment correlation taken, yielding Mantel's (1963) score statistic,

$$R_M = \frac{\sum_{k=1}^{K}\sum_{i=1}^{r}\sum_{j=1}^{c} Y_{ijk}\{a_{ik}-\bar{a}(Y_k)\}\{b_{jk}-\bar{b}(Y_k)\}}{\{\sum_{k=1}^{K}\sigma_a^2(Y_k)\sigma_b^2(Y_k)\}^{\frac{1}{2}}} \qquad (2.9)$$

where

$$\bar{a}(Y_k) = n_k^{-1}\sum_{i=1}^{r} y_{i+k}a_{ik} ,$$

$$\sigma_a^2(Y_k) = (n_k-1)^{-1}\sum_{i=1}^{r} y_{i+k}\{a_{ik}-\bar{a}(Y_k)\}^2$$

and $\bar{b}(Y_k)$ and $\sigma_b^2(Y_k)$ are defined analogously for all k. This statistic has power to detect alternatives in which the correlation is the same in each of the K response tables. In the 2×2 case, any choice of marginal scores for R_M which is constant for all k yields essentially M_H - see (2.8). However, there is no continuity correction of $\frac{1}{2}$ in R_M, since in general the scores are not integers.

When scores are not given, two possibilities arise:

(i) the relative ranks,

$$R_{i+k} = n_k^{-1}\{\sum_{q\leq i} Y_{q+k} + \tfrac{1}{2}(Y_{i+k}+1)\} , \qquad i = 1,\ldots,r;\ k = 1,\ldots,K \qquad (2.10)$$

for the rows, and a similar expression for columns.

(ii) equally spaced scores.

When K = 1 the former gives rise to Spearman's rank correlation, and the latter is similar to the statistic given by Yates (1948). However, Mantel (1963, p.696) warns against the indiscriminate use of relative ranks for scores when adjacent categories are clearly different, but have widely differing numbers of observations:

> 'In injury analysis, for example, fatal accidents may be relatively rare and will so be assigned a score only slightly higher than that for severe, or extremely severe, accidents. In contrast, with injuries of minor and moderate severity common, the difference in score between two such injuries will be large.'

Method (ii) of assigning the scores is acceptable if it is plausible that the scores <u>are</u> evenly spaced. Otherwise a method that requires no more than an ordering on the categories should be used, such as an extension of Wahrendorf's statistic $\hat{\psi}_W$, (2.7), for estimating the global cross-ratio, or one of the concordance statistics of §2.2.

2.8 Some Locally Most Powerful Test Statistics

Mantel's score statistic R_M defined in (2.9) may be used for testing the association parameters in certain statistical models for bivariate, discrete ordered responses. We will show that with suitably chosen marginal scores, R_M is equivalent to a locally most powerful test for these parameters. For $k = 1,\ldots,K$ consider n_k independent observations on the r×c contingency table Y_k with (i,j)-th cell probability $\pi_{ij/k}$ and count Y_{ijk}, $i = 1,\ldots,r$; $j = 1,\ldots,c$. The cell probabilities depend on an unknown $(\psi,\underset{\sim}{\lambda})$, where ψ is the association parameter and $\underset{\sim}{\lambda}$ is a vector of other parameters. Let $\ell(\psi,\underset{\sim}{\lambda};\ y_1,\ldots,y_K)$ be the log likelihood function for all K tables under this statistical model, and let $\hat{\underset{\sim}{\lambda}}_o$ be the maximum likelihood estimator of $\underset{\sim}{\lambda}$ under H_0: $\psi = \psi_0$ of no association. Then the locally most powerful test for ψ against $\psi > \psi_0$ is

$$LMP_0 = (\partial/\partial\psi)\ell(\psi,\underset{\sim}{\lambda};\ y_1,\ldots,y_K)\,|\,\psi = \psi_0,\ \underset{\sim}{\lambda} = \hat{\underset{\sim}{\lambda}}_0 \qquad (2.11)$$

- see Cox and Hinkley (1974, p.324).

Now Goodman's (1979) model for uniform local association satisfies

$$\ln \pi_{ij/k} = \alpha_i + \beta_j + ij\gamma, \qquad (2.12)$$

$i = 1,\ldots,r$; $j = 1,\ldots,c$; $k = 1,\ldots,K$; where $\Sigma\alpha_i = \Sigma\beta_j = 0$. The <u>local association</u> $\theta_{ij} = \gamma$. A straightforward extension, with row and column marginal scores $a_1<a_2<\ldots<a_r$ and $b_1<\ldots<b_r$ has

$$\ln \pi_{ij/k} = \alpha_i+\beta_j + a_ib_j\gamma \ , \qquad (2.13)$$

for all i,j,k, and the local association

$$\theta_{ij} = (a_{i+1}-a_i)(b_{j+1}-b_j)\gamma \ .$$

Then since the log likelihood $\ell(\gamma,\underset{\sim}{\alpha},\underset{\sim}{\beta};\ y_1,\ldots,y_K)$ up to a constant, equals

$$\sum_i y_{i++}\alpha_i + \sum_j y_{+j+}\beta_j + \sum_i \sum_j a_i b_j y_{ij+}\gamma$$

and

$$\partial\ell(\gamma,\underset{\sim}{\alpha},\underset{\sim}{\beta};\ y_1,\ldots,y_K)/\partial\gamma = \sum_i \sum_j a_i b_j y_{ij+}$$

for all values of $\gamma,\underset{\sim}{\alpha},\underset{\sim}{\beta}$, the locally most powerful test for $\gamma = 0$ against $\gamma > 0$ is

$$T_\gamma = \sum_i \sum_j a_i b_{ij} y_{ij+}$$

which is proportional to R_M for scores $\underset{\sim}{a}$, $\underset{\sim}{b}$, the same set for each response table y_k. This result emphasizes the dependence of the uniform local association model (2.12) on equally spaced marginal scores.

The constant global cross-ratio model on $y_1,\ldots,y_k$ has <u>global cross-ratio</u>

$$\psi = \frac{\sum_{a\le i}\sum_{b\le j} \pi_{ab/k} \sum_{a>i}\sum_{b>j} \pi_{ab/k}}{\sum_{a\le i}\sum_{b>j} \pi_{ab/k} \sum_{a>i}\sum_{b\le j} \pi_{ab/k}}$$

for all bivariate cutpoints (i,j) and all response tables k. The model is defined in detail in §3.2. We study here the model with a parameter for each row and column cumulative probability. Under H_0: $\psi = 1$, the maximum likelihood estimator of $\sum_{a\le i} \pi_{a+/k}$ is $\sum_{a\le i} Y_{a+k}/n_k$. Using this and a similar estimator for column

probabilities it can be shown (§3.5.4) that the locally most powerful statistic to test $\psi = 1$ against $H_1: \psi > 1$ is

$$T_\psi = 4 \sum_{k=1}^{K} \sum_{i=1}^{r} \sum_{j=1}^{c} Y_{ijk}(R_{i+k}-R_{++k})(R_{+jk}-R_{++k}),$$

where R_{i+k} is defined in (2.10), R_{+jk} is given similarly, and $R_{++k} = \frac{1}{2}(1+n_k^{-1})$. The statistic T_ψ is proportional to the correlation of the relative ranks, Spearman's rho if $K = 1$, or more generally Mantel's score statistic R_M with scores relative ranks.

2.9 Discussion

A review has been given of measures of association for discrete, bivariate, ordered responses in the form of two-way contingency tables - see Table 2.1. Most of the measures may be interpreted as correlations. Choice of a suitable measure should be based on its interpretability and the extent to which it incorporates available information on the marginal distributions. The application of a single measure to several response tables assumes that the same degree of association is present in each table - a hypothesis which should be tested. Very few distributional assumptions are necessary for application of the measures discussed in this chapter. The correlation estimated by the polychoric series method is that of an underlying bivariate normal distribution.

We have mentioned in §2.8 the relationship of some score statistics to locally most powerful tests for no association in local and global association models. However, the power of these statistics may well diminish as the association moves away from zero. Consider for example, behaviour of the correlation of relative ranks when a marginal distribution is not uniform (§2.7). The use of measures of association by themselves is but a first step in understanding the structure of bivariate data. In the following chapters we consider statistical models for discrete, bivariate responses, incorporating measures of association based on cross-ratios.

TABLE 2.1 Comparison of measures of association

Measure	Range*	Category scores required	Interpretation
concordance, §2.2			
τ	[-1,0,1]	no	correlation of indicator variables for concordance of pairs
γ	[-1,0,1]	no	
α	$(0,1,\infty)$	no	odds ratio
correlation			
polychoric, §2.3	[-1,0,1]	no	correlation: discretized bivariate normal
relative ranks §§2.3, 2.7, 2.8	[-1,0,1]	marginal relative ranks	correlation of scores
quantitative scores, §§2.3, 2.7, 2.8	[-1,0,1]	yes	
cross-ratios			
local, θ, §2.4	$(0,1,\infty)$	yes	odds ratio
global, ψ, §2.4	$(0,1,\infty)$	no	

*middle entry denotes values of parameter for no association.

3

GLOBAL CROSS-RATIO MODELS

3.1 Preamble

It is an advantage for the parameters in a statistical model to have an intuitive interpretation. With this aim in mind, the class of global cross-ratio models for bivariate discrete ordered responses $\underset{\sim}{Z}(\underset{\sim}{x}) = (Z_1(\underset{\sim}{x}), Z_2(\underset{\sim}{x}))$ is defined, where $\underset{\sim}{x}$ is a fixed covariate. Association between $Z_1(\underset{\sim}{x})$ and $Z_2(\underset{\sim}{x})$ is formulated as a global cross-ratio (GCR; §2.4) which may vary among cutpoints and values of the covariate $\underset{\sim}{x}$, with parameters linear in the logarithm of the GCR. The marginal distributions of $Z_1(\underset{\sim}{x})$ and $Z_2(\underset{\sim}{x})$ are expressed as generalized linear models of the cumulative probabilities (McCullagh, 1980 and §1.4 of this thesis). Discussion of GCR models with linear logistic margins concerns maximum likelihood fit under product multinomial sampling and the interpretation of parameters. Two datasets, on the postoperative condition of patients and the colouring of male blackbirds are analysed in §3.4 using these techniques. In §3.5 the problem of estimating the GCR is discussed when the marginal cumulative probabilities are unknown "nuisance" parameters. We examine a procedure of cross-sectional likelihood proposed by Anscombe (1981) for eliminating these marginal parameters, when the GCR takes the same value at all cutpoints and thus is a scalar. For published and simulated results, the method is compared with maximum likelihood. Point estimates and likelihood-based confidence intervals for the log GCR are similar for both approaches: asymptotic relative efficiency is near one for the datasets examined. A score statistic for testing the departure of the scalar GCR from its no-association value is derived in §3.5.4.

3.2 Global Cross-Ratio Models

3.2.1 General definition

Let $\underset{\sim}{Z}(\underset{\sim}{x}) = (Z_1(\underset{\sim}{x}), Z_2(\underset{\sim}{x}))$ be a discrete random vector, dependent on a fixed p-dimensional covariate $\underset{\sim}{x}$, quantitative or qualitative. Suppose that $Z_1(\underset{\sim}{x})$ takes the values $1,\ldots,r$ according to the cumulative probabilities $\eta_{ix} = \Pr[Z_1(\underset{\sim}{x}) \leq i]$, $i = 1,\ldots,r$ and similarly that the cumulative probability function for $Z_2(\underset{\sim}{x})$ is

$$\xi_{jx} = \Pr[Z_2(\underset{\sim}{x}) \leq j], \quad j = 1,\ldots,c, \quad \text{with } \eta_{ox} = \xi_{ox} = 0,$$

$$\eta_{rx} = \xi_{cx} = 1, \qquad \underset{\sim}{\eta}_x = (\eta_{1x},\ldots,\eta_{r-1,x})^T$$

and $\underset{\sim}{\xi}_x = (\xi_{1x},\ldots,\xi_{c-1,x})^T$. Restriction of Z_1 and Z_2 to these integer values does not reduce the generality of the model, since the bivariate cumulative probability function of $Z(\underset{\sim}{x})$ will depend on $Z_1(\underset{\sim}{x})$ and $Z_2(\underset{\sim}{x})$ only through $\underset{\sim}{\eta}_x$ and $\underset{\sim}{\xi}_x$ - see below. The cumulative probability function of $\underset{\sim}{Z}(\underset{\sim}{x})$ is $F_{ij}(\underset{\sim}{x};\underset{\sim}{\theta}_x) = \Pr[Z_1(\underset{\sim}{x}) \leq i, Z_2(\underset{\sim}{x}) \leq j]$, with $\underset{\sim}{\theta}_x = (\underset{\sim}{\eta}_x, \underset{\sim}{\xi}_x, \psi_x)$, where ψ_x is an $(r-1)\times(c-1)$ matrix of global cross-ratios to be defined. Now divide the space of possible values of $\underset{\sim}{Z}$ into four quadrants, $\{Z_1 \leq i, Z_2 \leq j\}$, $\{Z_1 \leq i, Z_2 > j\}$, $\{Z_1 > i, Z_2 \leq j\}$, $\{Z_1 > i, Z_2 > j\}$. This double dichotomy is said to occur at (bivariate) cutpoint (i,j). The $r\times c$ table of probabilities $\pi_{ij/x} = \Pr[Z_1(\underset{\sim}{x}) = i, Z_2(\underset{\sim}{x}) = j]$ is thus collapsed at cutpoint (i,j) to a 2×2 table.

Define the global cross-ratio (GCR) to be

$$\psi_{ijx} = \frac{\Pr[Z_1 \leq i,\ Z_2 \leq j]\Pr[Z_1 > i,\ Z_2 > j]}{\Pr[Z_1 > i,\ Z_2 \leq j]\Pr[Z_1 \leq i,\ Z_2 > j]}$$

$$= \frac{F_{ij}(\underset{\sim}{x};\underset{\sim}{\theta}_x)\{1-\eta_{ix}-\xi_{jx}+F_{ij}(\underset{\sim}{x};\underset{\sim}{\theta}_x)\}}{\{\xi_{jx}-F_{ij}(\underset{\sim}{x};\underset{\sim}{\theta}_x)\}\{\eta_{ix}-F_{ij}(\underset{\sim}{x};\underset{\sim}{\theta}_x)\}} . \qquad (3.1)$$

Now (3.1) may be solved for $F_{ij}(\underset{\sim}{x};\underset{\sim}{\theta}_x)$ in terms of $\underset{\sim}{\eta}_x$, $\underset{\sim}{\xi}_x$, $\underset{\sim}{\psi}_x$, yielding

$$F_{ij}(\underset{\sim}{x};\underset{\sim}{\eta}_x,\underset{\sim}{\xi}_x,\underset{\sim}{\psi}_x) = \begin{cases} \frac{1}{2}(\psi_{ijx}-1)^{-1}\{1+(\eta_{ix}+\xi_{jx})(\psi_{ijx}-1)-S(\eta_{ix},\xi_{jx},\psi_{ijx})\} & \\ \qquad \text{if } \psi_{ijx} \neq 1; & (3.2) \\ \eta_{ix}\xi_{jx} \qquad \text{if } \psi_{ijx} = 1, & (3.3) \end{cases}$$

$i = 1,\ldots,r-1;\ j = 1,\ldots,c-1$, where

$$S(\eta,\xi,\psi) = [\{1 + (\eta+\xi)(\psi-1)\}^2 + 4\psi(1-\psi)\eta\xi]^{\frac{1}{2}} .$$

It may be shown that F is a bona fide cumulative distribution provided that $F_{ijx}-F_{i-1,j,x}-F_{i,j-1,x}+F_{i-1,j-1,x} \geq 0$, where $F_{ijx} = F_{ij}(\underset{\sim}{x};\underset{\sim}{\theta}_x)$, $i = 1,\ldots,r;\ \ j = 1,\ldots,c$. This requires that the marginal cumulative probabilities satisfy

$$0 < \eta_{1x} < \eta_{2x} < \ldots < \eta_{r-1,x} < 1, \qquad 0 < \xi_{1x} < \ldots < \xi_{c-1,x} < 1$$

and that the ψ_{ijx} for adjacent cutpoints are not very different. When $\psi_{ijx} = \psi(\underset{\sim}{x})$, then $\psi(\underset{\sim}{x})$ is unrestricted - see Mardia (1970). If $\eta_{ix} = 1$ then $F_{ijx} = \xi_{jx}$, whatever the value of ψ_{ijx}. The distribution defined in generality by (3.2) and (3.3) is called a global cross-ratio model for $\underset{\sim}{Z}(\underset{\sim}{x})$. Now F_{ijx} depends on $\underset{\sim}{x}$ only through η_{ix}, ξ_{jx} and ψ_{jx}.

Moreover, the association parameters ψ_x are not dependent on the marginal parameters $\underset{\sim}{\eta}_x$ and $\underset{\sim}{\xi}_x$, since association is defined in terms of cross-ratios (§4.3.5). Hence a family of parametric models, hierarchical in both types of parameter, may be defined. For given $\underset{\sim}{x}$, the general GCR model of (3.2) is saturated - it has rc-1 parameters, equal to the number of degrees of freedom in the table. Thus for estimability it is necessary to reduce the dimension of the parameter space.

3.2.2 A parametric family of GCR models

Since the parameter space for GCR models is the product of the space for GCRs and for marginal cumulative probabilities when the GCRs do not depend on the cutpoints, these two sets of parameters are constrained separately in what follows.

GCR parameters. In order to simplify the general matrix ψ_x of GCR parameters ψ_{ijx}, several options are open. The GCR may be set independent of $\underset{\sim}{x}$, or independent of the cutpoints but dependent on $\underset{\sim}{x}$. Since the ψ_{ijx} are defined on $(0,\infty)$, it is easier for estimation purposes to model $\Delta_{ijx} = \ln \psi_{ijx}$. Thus we have

$$\ln \psi_{ijx} = \Delta_{ij}; \tag{3.4}$$

$$\ln \psi_{ijx} = \Delta_i + \delta_j; \tag{3.5}$$

$$\ln \psi_{ijx} = \Delta + \underset{\sim}{\gamma}^T \underset{\sim}{x} \tag{3.6}$$

$$\ln \psi_{ijx} = \Delta \tag{3.7}$$

A hierarchy of models involves increasing successively the constraints on the parameters. For example (3.4), (3.5) and (3.7) form a hierarchy independent of $\underset{\sim}{x}$. The parameterizations of (3.6) and (3.7) define

constant GCR models, since given $\underset{\sim}{x}$ the GCR is constant for all cutpoints. A constant GCR model is given by Pearson and Heron (1913) and by Plackett (1965) for a bivariate response with specified marginal distributions, continuous or discrete. It is also named a contingency-type distribution by Mardia (1970). In the continuous case the density is called the surface of constant association. Models for trivariate and higher-dimensional responses may be defined in terms of a measure of association analogous to the GCR - see Anscombe (1981, 310-311).

Marginal cumulative probabilities. Generalized linear models for cumulative probabilities were discussed by McCullagh (1980) and in §1.4. Given link functions $g_1(.)$, $g_2(.)$ (§1.2), set

$$\begin{aligned} g_1(\eta_{ix}) &= \alpha_{1i} - \underset{\sim}{x}^T\underset{\sim}{\beta}_1, \qquad i = 1,\ldots,r-1; \\ g_2(\xi_{jx}) &= \alpha_{2j} - \underset{\sim}{x}^T\underset{\sim}{\beta}_2, \qquad j = 1,\ldots,c-1; \end{aligned} \tag{3.8}$$

where $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ are p-dimensional parameters which do not depend on $\underset{\sim}{x}$. For $g_q(p) = \ln\{p/(1-p)\}$, $q = 1,2$, (3.8) gives rise to the linear logistic models

$$\begin{aligned} \eta_{ix} &= \exp(\alpha_{1i}-\underset{\sim}{x}^T\underset{\sim}{\beta}_1)/\{1+\exp(\alpha_{1i}-\underset{\sim}{x}^T\underset{\sim}{\beta}_1)\}, \quad i = 1,\ldots,r-1 \\ \xi_{jx} &= \exp(\alpha_{2j}-\underset{\sim}{x}^T\underset{\sim}{\beta}_2)/\{1+\exp(\alpha_{2j}-\underset{\sim}{x}^T\underset{\sim}{\beta}_2\}, \quad j = 1,\ldots,c-1. \end{aligned} \tag{3.9}$$

A hierarchy of marginal parameterizations is set up by considering $\underset{\sim}{\eta}_x, \underset{\sim}{\xi}_x$, in full generality, then (3.9), then restricting $\underset{\sim}{\beta}_1 = \underset{\sim}{\beta}_2$, and finally $\underset{\sim}{\beta}_1 = \underset{\sim}{\beta}_2 = 0$. One may also postulate $\alpha_{1i}-\alpha_{1,i-1} = \alpha_1$, $i = 2,\ldots,r-1$, i.e. equal spacing of marginal cutpoints on the logit scale. Both the analyses carried out in §3.4 assume linear logistic margins.

A parametric family of GCR models. Choose link functions g_1, g_2 for modelling the margins, as in (3.8), and one of the expressions (3.4) to (3.7) for the GCR parameters. These together with (3.2) and (3.3) define a parametric family of GCR models. Such a family for a sequence of 2×2 tables with linear logistic margins was presented by Mantel and Brown (1973). It is equivalent to our (3.6) and (3.9). Those GCR models used in the analyses of §3.4 have the GCR constant over cutpoints as in (3.7) and linear logistic margins.

Interpretation. Now the GCR may be interpreted as an odds ratio of conditional events, the odds on $\{Z_1(\underset{\sim}{x}) \leq i | Z_2(\underset{\sim}{x}) \leq j\}$ versus the odds on $\{Z_1(\underset{\sim}{x}) \leq i | Z_2(\underset{\sim}{x}) > j\}$, where odds on event E = $\Pr(E)/\{1 - \Pr(E)\}$. Under (3.6) and (3.7), that odds is kept constant for all values of i and j. Observe that if j = c then the second odds becomes degenerate, since $\Pr[Z_2(x) > c] = 0$: the first is the unconditional odds on $\{Z_1(x) \leq i\}$. In its role as a marginal cumulative probability it is expressed by a similar model to that for the log GCR: the linear logistic. GCR models do not require marginal scores: they are essentially invariant under any monotonic transformation of the marginal response variables, and are defined even when only an ordering of the marginal categories is known. If adjacent marginal categories are combined, the GCR model for the new table has fewer parameters, and all remaining parameters have essentially the same interpretation as they had in the model for the original, expanded table.

3.3 Fitting GCR Models

3.3.1 Maximum likelihood estimation

A straightforward method of estimation with desirable asymptotic properties in our situation is maximum likelihood (Cox and Hinkley, 1974, §9.2). Since the maximum likelihood parameter estimates for most of the models in §3.2.2 are not available in closed form, a numerical approach has been adopted.

Consider n_k independent observations distributed according to the GCR family of models with covariate $\underset{\sim}{x}_k$, linear logistic margins (3.9) and GCR parameters independent of $\underset{\sim}{x}_k$, (3.4). Let Y_{ijk} be the number of observations in cell (i,j), with Y_k denoting the whole r×c table. Then the distribution of Y_k is multinomial on n_k observations and r×c cells. We will consider fitting this GCR model to a sequence of m independent tables $Y_1,\ldots,Y_m$, dependent on a matrix of covariates $X = (\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m)$. The log-likelihood function for $Y_1,\ldots,Y_m$, up to an additive function not dependent on $\underset{\sim}{\theta}$, is

$$\ell(\underset{\sim}{\theta};\, y_1,\ldots,y_m,X) = \sum_{k=1}^{m} y_{ijk} \ln \pi_{ij}(\underset{\sim}{x}_k;\underset{\sim}{\theta}) \ ,$$

where $\pi_{ij}(\underset{\sim}{x}_k;\underset{\sim}{\theta}) = \Pr\{Z_1(\underset{\sim}{x}_k)=i, Z_2(\underset{\sim}{x}_k) = j \,|\, \underset{\sim}{\theta}\}$, $\underset{\sim}{\theta} = (\underset{\sim}{\alpha}_1,\underset{\sim}{\alpha}_2,\underset{\sim}{\beta}_1,\underset{\sim}{\beta}_2,\Delta)$, and Δ is the $(r-1)\times(c-1)$ matrix of parameters Δ_{ij} in (3.4). Now ℓ is maximized using a quasi-Newton method from the NAG Subroutine Library (1978). The value of the log-likelihood function and its first derivatives with respect to $\underset{\sim}{\theta}$ are calculated algebraically (see Appendix I): the second derivatives are obtained numerically as the maximization proceeds. Our estimate of the variance-covariance matrix of the parameter estimates is minus the inverse of the matrix of second partial derivatives at the maximum, i.e. the inverse of the observed

Fisher information matrix. For a good estimate of the variance-covariance matrix, it was found necessary to maximize to within as small a tolerance as machine precision would allow.

3.3.2 Closed-form starting estimates

To commence the maximization routine, closed-form starting estimates are needed for the parameters. For the cutpoint parameters $\underset{\sim}{\alpha}_1$ and $\underset{\sim}{\alpha}_2$ in the marginal linear logistic models of (3.9), the approach of Cox (1970, equation 3.13) was used, treating row and column margins separately. Initial estimates for $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ were zero. We now outline the approach adopted to estimate log GCRs. For $i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$, define

$$\begin{aligned} A_{ij/k} &= \frac{1}{2} + \sum_{p\leq i}\sum_{q\leq j} Y_{pqk}; \\ B_{ij/k} &= \frac{1}{2} + \sum_{p\leq i}\sum_{q>j} Y_{pqk}; \\ C_{ij/k} &= \frac{1}{2} + \sum_{p>i}\sum_{q\leq j} Y_{pqk}; \\ D_{ij/k} &= \frac{1}{2} + \sum_{p>i}\sum_{q>j} Y_{pqk}\ . \end{aligned} \tag{3.10}$$

Then let

$$\tilde{\Delta}_{ij/k} = \ln\{(A_{ij/k}D_{ij/k})/(B_{ij/k}C_{ij/k})\} \tag{3.11}$$

and

$$w_{ij/k} = (A_{ij/k}^{-1} + B_{ij/k}^{-1} + C_{ij/k}^{-1} + D_{ij/k}^{-1})^{-1}\ .$$

It can be shown (Anscombe, 1981, p.313) that $\tilde{\Delta}_{ij/k}$ has approximate mean Δ_{ij} and variance $E(w_{ij/k}^{-1})$, for large n_k. Thus an approximately unbiased estimator of Δ_{ij} is

$$\tilde{\Delta}_{ij} = (\Sigma_k w_{ij/k}\tilde{\Delta}_{ij/k})/(\Sigma_k w_{ij/k}) \ . \tag{3.12}$$

Replacing $\frac{1}{2}$ by $\frac{1}{6}$ in (3.10) and multiplying the resulting estimator analogous to (3.11) by $\frac{1}{2}$ yields Tukey's double folded log or fflog for cutpoint (i,j) - see Tukey (1977, §15F). His method is used in an exploratory manner, without reference to a statistical model.

For estimating the Δ of (3.7), we adopt the approach of Anscombe (1981). Unfortunately the $\tilde{\Delta}_{ij/k}$ are correlated for the same k. Wahrendorf (1980) has shown this for estimators similar to (3.11). Despite this, the following approximately unbiased estimator has been found to work well as a starting estimate of Δ:

$$\tilde{\Delta} = (\Sigma n_k)^{-1}\Sigma n_k W_k^{-1}\{\Sigma_i\Sigma_j w_{ij/k}\tilde{\Delta}_{ij/k}\}, \tag{3.13}$$

where $W_k = \Sigma_i\Sigma_j w_{ij/k}$, $k = 1,\ldots,m$.

3.4 Examples

3.4.1 Patients' postoperative pain level and medication requirements

After surgery for duodenal ulcer, patients were classified according to the severity of the pain (none/slight/moderate) and how often they required medication (never/seldom/occasionally/regularly). Four different types of operation, classified VP, VA, VH and RA, had been performed in a randomised clinical trial, with approximately equal numbers undergoing each, a total of 1013 patients. Background to the trial is given in Price et al. (1970). The data from Williams and Grizzle (1972, Table 6) are presented in Table 3.1. We are interested in detecting

(a) differences in pain level and medication requirements between operations;

(b) whether medication use is associated with pain level.

The independence of patients receiving the same operation is assumed. Since both responses are categorical with ordered categories, it is plausible that a GCR model describes relationships (a) and (b). For simplicity we have fitted the GCR model with GCR e^{Δ} constant for all cutpoints and operations as in (3.7), and marginal distributions the linear logistic models of (3.9), one for pain level and one for frequency of medication. The covariate is $X = (I_3, \underset{\sim}{0}_3)$, with the rows indicating the operation. A summary of the maximum likelihood fit is given in Table 3.2. Overall, the fit is adequate, with $p > 0.05$ for the likelihood ratio G^2 statistic. Differences between operations are modelled by the parameters β_{qi}, $i = 1,2,3$; $q = 1,2$. None of these is statistically significant at the 5% level. Thus any differences among operations in medication requirements and pain levels are not detected statistically in this sample. The estimated log GCR $\hat{\Delta} = 2.76$ is

TABLE 3.1 Number of patients, classified by type of operation, level of pain and medication requirements

operation	pain level	medication requirements			
		never	seldom	occasionally	regularly
VP	none	170	7	8	0
	slight	18	5	8	3
	moderate	7	0	4	14
VA	none	170	7	5	2
	slight	22	7	8	1
	moderate	8	1	8	9
VH	none	176	8	5	2
	slight	26	6	5	5
	moderate	14	3	2	9
RA	none	181	6	6	2
	slight	17	12	7	3
	moderate	10	2	3	11

TABLE 3.2 GCR model fit to postoperative status dataset

Parameter*	Symbol	Estimate	Standard error
log ψ	Δ	2.76	0.17
pain N/S-S/M	$\alpha_{11}-\alpha_{12}$	-1.08	0.085
pain S/M	α_{12}	2.19	0.16
med N/S-S/O	$\alpha_{21}-\alpha_{22}$	-0.45	0.056
med S/O-O/R	$\alpha_{22}-\alpha_{23}$	-0.85	0.10
med O/R	α_{23}	2.66	0.18
pain VP-RA	β_{11}	-0.02	0.20
pain VA-RA	β_{12}	0.07	0.20
pain VH-RA	β_{13}	0.07	0.20
med VP-RA	β_{21}	0.08	0.21
med VA-RA	β_{22}	-0.03	0.21
med VH-RA	β_{23}	-0.18	0.22

Likelihood ratio G^2 statistic = 44.2 with 32 df

* Pain levels N = none; S = slight; M = moderate.

Medication requirements N = never; S = seldom; O = occasionally; R = regularly.

statistically very significant ($p > 0.001$), and hence we may conclude that there is a marked positive association between pain level and medication requirements in this sample. Let $\underset{\sim}{Z}_k = (Z_{1k}, Z_{2k})$ be the pain, medication outcome for a typical patient who had operation k. Then under the fitted model,

$$\frac{\Pr[Z_{2k} \leq j \mid Z_{1k} \leq i]}{\Pr[Z_{2k} > j \mid Z_{1k} \leq i]} = e^{2.76} \frac{\Pr[Z_{2k} \leq j \mid Z_{1k} > i]}{\Pr[Z_{2k} > j \mid Y_{1k} > i]} ,$$

$i = 1,2,3$; $j = 1,2,3,4$. For each operation k, level of pain i and medication requirement j, the odds that the medication level $\leq$ j given that the pain level $\leq$ i is expected to be about 16 times as high as the odds on level of medication $\leq$ j given that the pain level > i. An approximate 95% confidence level for the GCR is (11.3, 22.2).

Differences between adjacent cutpoint variables are strongly significant statistically, indicating that the categories for pain and medication are well defined. A cell-by-cell examination of residuals $(O - E)/\sqrt{E}$ suggests that the model fits well: the largest residual is 2.2, in cell (2,2) of the table for operation RA.

This analysis suggests a good fit of the constant GCR model, with a very strong relationship between postoperative pain level and medication requirements, for the patients in the study. The odds on less frequent medication requirements, given lower pain level, is estimated at between 11 and 22 times higher than more frequent medication levels, given low pain levels. There is no statistical evidence to support differences between types of operation in pain level or medication requirements.

3.4.2 Colouring of first-winter male blackbirds

Adult and immature male blackbirds Turdus merula can be differentiated by plumage characteristics. As a first-winter male blackbird matures, the beak and eye colourings change from black to yellow. These features were observed during eighteen ten-day periods from October 1, 1980 to the last week in March, 1981, when birds were caught, ringed and released as they entered winter roosts in Fife, Scotland. The following analysis, using a GCR model with linear logistic margins, examines possible associations between mandible and orbital ring colourings, and the relative rate of change in the two colourings.

All observations were recorded by one person, the upper mandible colourings (UM) on a scale 0 (totally black), 1 (less than 1/3 mandible yellow),..., 4 (totally yellow), and the colouring of the orbital ring around the right eye (OR) on a three-point scale. In order to reduce the number of zero entries in the table of data, we omitted from the analysis any time periods when there were observations on only one mandible or orbital ring category. Some upper mandible categories were merged for the same reason. Thus only data from periods 4, 5, 6, 8, 11, 12 and 14 were included, a total of 70 observations - see Table 3.3. It was considered unwise to combine time periods. The marginal distributions depend strongly on time, and thus in a combined table what appears to be association may be the effect of contributions from different times. There were occasional recaptures of the same bird in different time periods: nine birds were handled twice and one three times. Application of the GCR model with GCR constant over cutpoints (3.7) required the following assumptions:

TABLE 3.3 Number of blackbirds, classified by colour score and period observed

Orbital ring score	Period† 1* B	1* M	1* Y‡	2* B	2* M	2* Y	3* B	3* M	3* Y	4 B	4 M	4 Y	5 B	5 M	5 Y
0	1	0	0	2	0	0	5	0	0	9	3	0	7	2	0
1	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Orbital ring score	Period 6 B	6 M	6 Y	7* B	7* M	7* Y	8 B	8 M	8 Y	10* B	10* M	10* Y	11 B	11 M	11 Y
0	4	0	0	2	0	0	6	1	0	0	0	0	0	0	0
1	1	2	0	3	0	0	2	4	2	0	2	1	3	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Orbital ring score	Period 12 B	12 M	12 Y	14 B	14 M	14 Y	15* B	15* M	15* Y	17* B	17* M	17* Y	18* B	18* M	18* Y
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	1	2	0	1	3	2	1	1	0	0	0	1	0	0	1
2	0	0	1	0	1	6	0	0	0	0	0	0	0	0	0

* Omitted from analysis.

† Ten-day periods, from October 1, 1980; no data collected in period 9.

‡ Upper mandible category: B = 0; M = 1,2; Y = 3,4.

(i) that all sightings in a given period were independent and identically distributed;

(ii) that repeated sightings of the same bird in different periods were equivalent to sightings of different birds;

(iii) that there was a constant relationship throughout the study between orbital ring and upper mandible colourings.

Assumptions (i) and (iii) were acceptable, whereas (ii) may not strictly hold, as colourings are very unlikely to become blacker with time.

We fitted a GCR model with GCR constant over cutpoints and linear logistic margins dependent on the scalar covariate t, the time in ten-day periods. Two models were considered:

(a) allowing the marginal slopes β_1 and β_2 for orbital rings and upper mandibles to differ;

(b) constraining the slopes to be equal.

Parameter estimates and goodness-of-fit statistics under a maximum likelihood fit are presented in Table 3.4. Model (a) appears to fit the data very well: the goodness-of-fit statistics are both significant at the lower 5% level, under the asymptotic χ^2 distribution. Presumably this is partly because many cells have low or zero counts. The multinomial distribution may be incorrect. Examination of the Poisson standardized residuals suggests an adequate fit, with all residuals no larger in modulus than 2.2.

There is statistical evidence to suggest a difference between orbital rings and upper mandibles in rate of change of colouring with respect to time. In model (a) the standard error of the difference in marginal slopes is 0.099, and the likelihood ratio statistic for comparing

TABLE 3.4 GCR model fit to blackbird colouring dataset

Constant ψ, linear logistic for both margins, covariate time in ten-day periods

Parameter*	Symbol	Estimate	Standard error
log ψ	Δ	2.25	0.57
OR 0/1	α_{11}	3.75	0.76
OR 1/2	α_{12}	7.48	1.30
UM B/M	α_{21}	2.44	0.61
UM M/Y	α_{22}	4.26	0.77
OR slope	β_1	0.522	0.099
UM slope	β_2	0.288	0.068

Likelihood ratio G^2 = 33.9 with 49 d.f.

Pearson X^2 = 32.3 with 49 d.f.

Likelihood ratio for common slope = 6.3 with 1 d.f.

* OR = orbital ring; UM = upper mandible; a/b = cutpoint between categories a and b.

models (a) and (b) is statistically significant ($p \simeq 0.012$). Hence model (a) is preferred. The expected rate of change of colouring under model (a) is for orbital rings approximately 1.8 times that of upper mandibles, on the logit scale. We describe the fitted orbital ring slope, represented in Fig. 3.1, as follows. For any orbital ring colouring i, and any time period t,

$$\frac{\Pr[OR \leq i \text{ at period } t]}{\Pr[OR > i \text{ at period } t]} = e^{-0.522} \times \frac{\Pr[OR \leq i \text{ at period } t-1]}{\Pr[OR > i \text{ at period } t-1]} .$$

Now $\exp(-0.522) = 0.59$, so the odds on the orbital ring being "as black as or blacker than i' reduces by 0.59 each period. For upper mandible colouring, the expected reduction of odds is $\exp(-0.288) = 0.75$. The fitted probability of being in a particular marginal category in a given period can be read from Fig. 3.1. For example, following the dotted line (·····) for period 6:

$\Pr[UM = 1 \text{ or } 2] = \text{logit}^{-1}(2.53) - \text{logit}^{-1}(0.71) = 0.93 - 0.67 = 0.26$,

where $\text{logit}^{-1}(x) = e^x/(1 + e^x)$. The expected probability of having a given orbital ring and upper mandible colouring <u>simultaneously</u> in a given period is obtained by differencing the fitted bivariate cumulative probabilities. For any orbital ring score i and any upper mandible category j, given that $UM \leq j$, the odds on $OR \leq i$ are expected to be $\exp(2.25) \simeq 9.5$ times higher than the odds on $OR \leq i$ given $UM > j$. This odds ratio $\psi = e^{\Delta}$ has approximate 95% confidence interval (3.0, 30.1).

In summary, this statistical analysis suggests that observed orbital ring and upper mandible colourings were associated, in the first-winter blackbirds under study, with GCR estimated to be 9.5. The change in

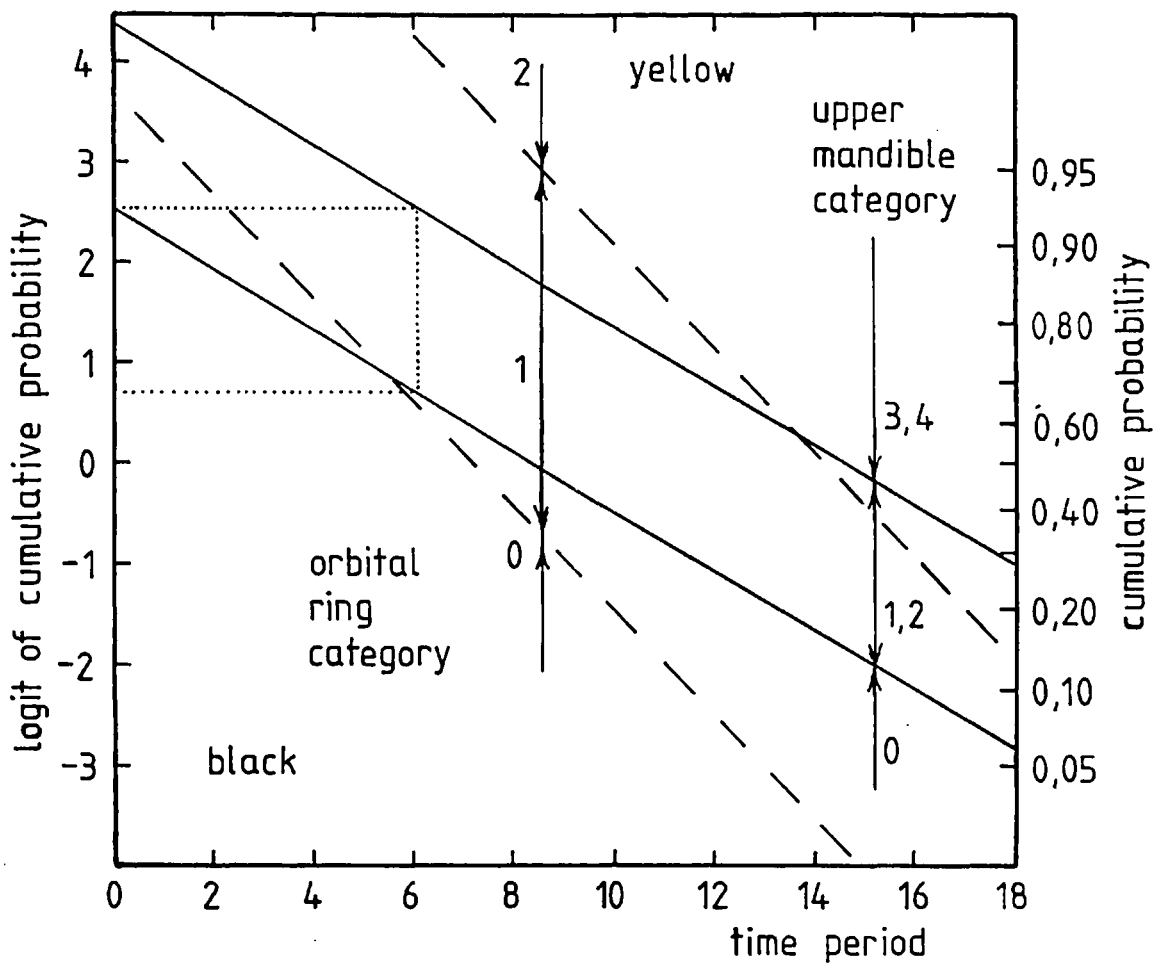


FIGURE 3.1 Fitted cutpoint position versus time period, for orbital ring colourings (————) and upper mandible colourings (— — —): blackbird dataset, model (a)

colouring over several months was about 1.8 times faster (on the logit scale) for orbital rings than for upper mandibles. With only 70 observations, these conclusions are tentative.

3.5 Estimating the GCR in the Presence of Nuisance Parameters

3.5.1 Introduction

Often only a subset of the parameters in a statistical model are of immediate concern. Consider a GCR model for a single r×c contingency table, with the global association parameter $\psi = e^{\Delta}$ a scalar as in (3.7), of interest. While the marginal cumulative probabilities $\underset{\sim}{\eta}$ and $\underset{\sim}{\xi}$ are unknown and require estimation, we do not wish to make inference about them. This situation is similar to that for two continuous random variables, when interest is in the correlation between them, rather than in their individual means. We investigate properties of a method proposed by Anscombe (1981) to deal with this problem. A score statistic is derived in §3.5.4 which is locally most powerful for testing $\Delta = 0$ against one-sided alternatives.

Consider a bivariate random variable $\underset{\sim}{Z} = (Z_1, Z_2)$ satisfying a GCR model as defined in (3.2), (3.3) and (3.7) for a single value of the covariate, i.e. the GCR is the same for all cutpoints. The cumulative distribution function for $\underset{\sim}{Z}$ is given by

$$F_{ij}(\Delta, \underset{\sim}{\eta}, \underset{\sim}{\xi}) = \begin{cases} \frac{1}{2}(e^{\Delta}-1)^{-1}[1+\{\eta_i+\xi_j\}(e^{\Delta}-1) - S(\Delta, \eta_i, \xi_j)] & \text{if } \Delta \neq 0; \quad (3.14) \\ \eta_i \xi_j & \text{if } \Delta = 0; \quad (3.15) \end{cases}$$

$i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$, $0 < \eta_1 < \ldots < \eta_r = 1$, $0 < \xi_1 < \ldots < \xi_c = 1$;

$$S(\Delta,\eta,\xi) = [\{1+(\eta+\xi)(e^{\Delta}-1)\}^2+4e^{\Delta}(1-e^{\Delta})\eta\xi]^{\frac{1}{2}}, \quad (3.16)$$

$0 < \eta < 1$ and $0 < \xi < 1$; $F_{ic}(\Delta, \underset{\sim}{\eta}, \underset{\sim}{\xi}) = \eta_i$, $i = 1,\ldots,r$; and $F_{rj}(\Delta, \underset{\sim}{\eta}, \underset{\sim}{\xi}) = \xi_j$, $j = 1,\ldots,c$. A random sample of size n from this distribution is tabulated in an r×c contingency table Y with (i,j)-th

entry Y_{ij} = no. of observations $\underset{\sim}{Z}$ with $Z_1 = i$, $Z_2 = j$. The table Y follows a multnomial distribution according to (3.14) and (3.15), with log-likelihood function, up to a constant not depending on parameters,

$$\ell(\Delta,\underset{\sim}{\eta},\underset{\sim}{\xi};y) = \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}\ln\{-S(\Delta,\eta_i,\xi_j) + S(\Delta,\eta_i,\xi_{j-1}) + S(\Delta,\eta_{i-1},\xi_j) - S(\Delta,\eta_{i-1},\xi_{j-1})\} - n\ln(e^{\Delta}-1) \qquad (3.17)$$

It is a function of r+c-1 parameters, all except one of which are "nuisance" parameters.

3.5.2 Coping with nuisance parameters: Anscombe's approach

We want to estimate Δ in such a way that the marginal parameters $\underset{\sim}{\eta}$ and $\underset{\sim}{\xi}$ are accounted for. Unfortunately the GCR models are not in the exponential family of distributions unless r = c = 2 or $\Delta = 0$. In fact the only sufficient statistic for $(\Delta,\underset{\sim}{\eta},\underset{\sim}{\xi})$ is the entire contingency table Y. However the statistic

$$\tilde{\eta}_i = n^{-1}\sum_{q=1}^{i} Y_{q+}, \qquad i = 1,\ldots,r-1$$

is marginally sufficient for η_i, i.e. it is sufficient with reference only to the distribution of the marginal totals $\underset{\sim}{Y}_{row} = (Y_{1+},\ldots,Y_{r+})$. Moreover, $\tilde{\eta}_i$ is the maximum likelihood estimator of η_i under the marginal likelihood of $\underset{\sim}{Y}_{row}$ and is thus consistent and efficient in this context. Similar properties hold for

$$\tilde{\xi}_j = n^{-1}\sum_{s=1}^{j} Y_{+s}, \qquad j = 1,\ldots,c-1$$

with respect to the column totals. When $\Delta = 0$, then by (3.15) the $\tilde{\underset{\sim}{\eta}}$ and $\tilde{\underset{\sim}{\xi}}$ are full maximum likelihood estimators. Anscombe's (1981) approach is to substitute $\tilde{\underset{\sim}{\eta}}$ and $\tilde{\underset{\sim}{\xi}}$ for $\underset{\sim}{\eta}$ and $\underset{\sim}{\xi}$ in the full log-likelihood function (3.17), and maximize it w.r.t. Δ :

$$\ell_A(\Delta;y) = \ell(\Delta,\tilde{\underset{\sim}{\eta}},\tilde{\underset{\sim}{\xi}};y). \tag{3.18}$$

Thus a data-dependent cross-section of ℓ is maximized. The value of Δ at the maximum $\tilde{\Delta}_A$, is called the Anscombe estimator of Δ. As the sample size n increases, and for all cell probabilities positive,

$$\tilde{\underset{\sim}{\eta}} \xrightarrow{p} \underset{\sim}{\eta} \qquad \text{and} \qquad \tilde{\underset{\sim}{\xi}} \xrightarrow{p} \underset{\sim}{\xi} ,$$

and thus $\tilde{\Delta}_A$ estimates Δ consistently [by Billingsley (1968), Theorem 5.1 concerning the convergence in probability of a continuous function of convergent sequences of random variables, since ℓ is continuous in the components of $\underset{\sim}{\eta}$ and $\underset{\sim}{\xi}$].

3.5.3 Empirical comparison of Anscombe and full likelihood methods

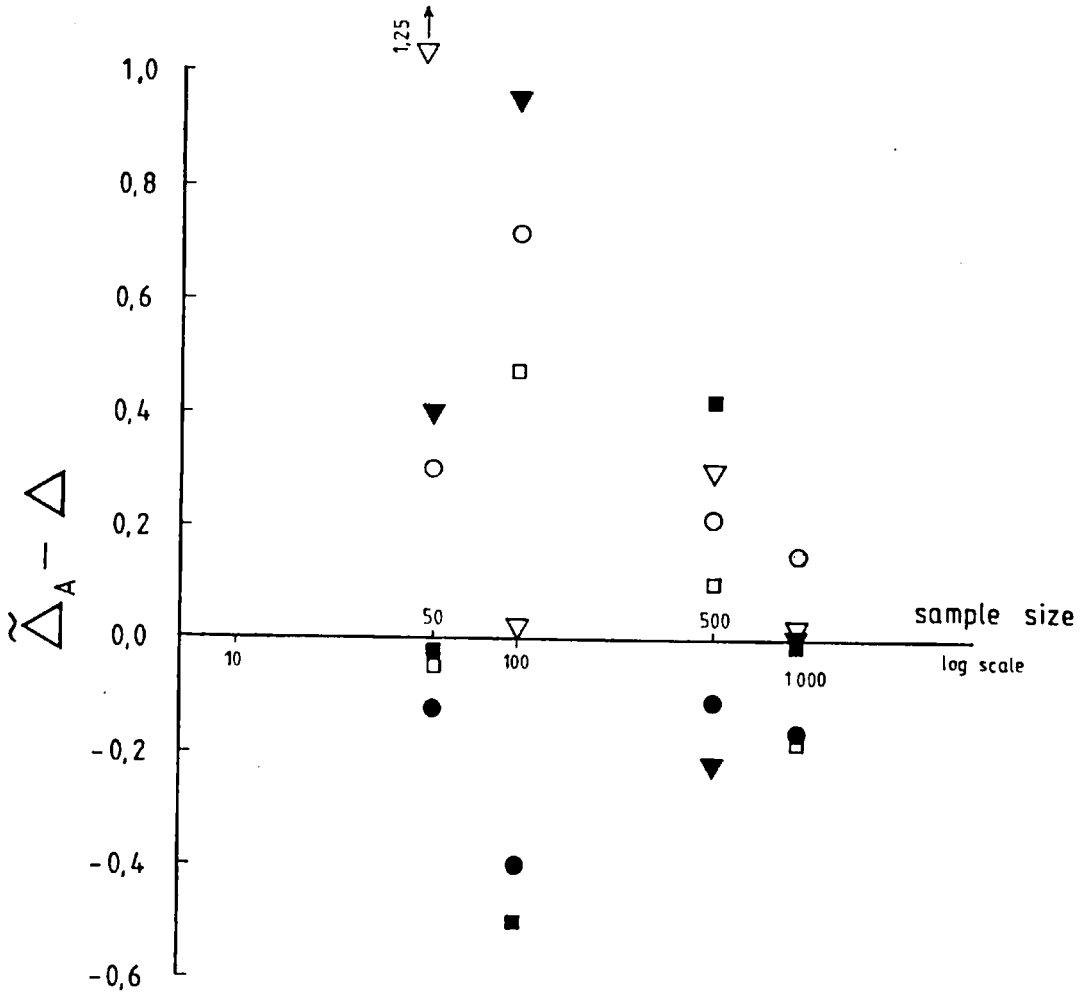
The appeal of Anscombe's cross-sectional approach is that Δ may be estimated by maximizing a function of one variable, instead of r+c-1 variables. However, it is at the cost of losing information on the relationship between Δ and the marginal parameters. We now examine the consistency of $\tilde{\Delta}_A$ and its asymptotic relative efficiency relative to the maximum likelihood estimator $\hat{\Delta}$, using data generated from several GCR distributions. Parameter estimates and likelihood-based confidence intervals for Δ using both methods are compared for published and simulated data.

Pseudo-random samples were taken from several GCR distributions on 3×4 tables, to compare Anscombe's method to the full likelihood method. The CDC Fortran random number generator RANF was used, the order of generation of samples randomized, and each new sequence of random numbers seeded by a number based on the run time of the previous run. To examine consistency and asymptotic relative efficiency of the two estimators, samples of size n = 50, 100, 500 and 1000 were generated randomly from GCR distributions. The marginal configurations shown in Table 3.5 were each used with Δ = 2 or 4, for all cutpoints, giving six "true distributions". We consider only positive Δ, since negative Δ is equivalent to reversing the order of either set of marginal probabilities.

TABLE 3.5 True marginal probabilities for simulation study

plotting symbol	marginal probabilities row				column
					0.10
o					0.45
					0.45
	0.1	0.3	0.3	0.3	
					0.10
∇					0.45
					0.45
	0.3	0.3	0.3	0.1	
					0.15
□					0.70
					0.15
	0.15	0.35	0.35	0.15	

Figure 3.2 shows, for the datasets generated, the variation of $\tilde{\Delta}_A - \Delta$ with sample size. As n increases, the absolute value of this quantity does decrease, indicating that $\tilde{\Delta}_A$ is a consistent estimator of Δ. The asymptotic relative efficiency of $\hat{\Delta}$ to $\tilde{\Delta}_A$, estimated using the ratio of observed Fisher information for the two methods, is given in Table 3.6.



KEY:

Symbols defined in Table 3.5
Open symbols: $\Delta = 2$
Solid symbols: $\Delta = 4$

FIGURE 3.2 Variation of $\tilde{\Delta}_A$ with sample size.

TABLE 3.6 Asymptotic relative efficiency of full likelihood to Anscombe estimate of Δ for generated datasets

marginal configuration*	sample size	$\Delta = 2$	$\Delta = 4$
o	50	0.94	0.90
	100	0.95	0.97
	500	0.95	0.97
	1000	0.96	0.96
▽	50	0.95	0.95
	100	0.96	0.96
	500	0.96	0.95
	1000	0.96	0.95
□	50	0.96	0.98
	100	0.99	0.99
	500	0.97	0.99
	1000	0.98	0.99

*defined in Table 3.5.

The ratio approaches one for larger sample sizes, pointing to an asymptotic reduction to zero of $\text{cov}(\hat{\underset{\sim}{\eta}}, \hat{\Delta})$ as n becomes large. However, this may not be so for smaller samples. Hence it was necessary to compare the distributions of $\tilde{\Delta}_A$ and $\hat{\Delta}$ for a small sample size.

A simulation study, comparing sample means and variances of the estimated Δ for the Anscombe and full likelihood methods, is summarized in Table 3.7. It suggests that $\tilde{\Delta}_A$ is significantly different ($\alpha = 0.05$) from $\hat{\Delta}$ for two of the ' □ ' marginal configurations - symmetric margins. There is little evidence to support different variances for the two estimators, except in the 100 observation simulation. With a sample size of 25 and correlation +0.8 between observed values of $\tilde{\Delta}_A$ and $\hat{\Delta}$, Morgan's test (1939) has power 0.14 for $\sigma^2_{\tilde{\Delta}_A} / \sigma^2_{\hat{\Delta}} = 0.89$. In fact the observed correlation is typically 1.000, suggesting a higher power than this. For the small range of GCR models for 3×4 tables sampled in this study, the Anscombe estimator of Δ does not appear to differ much from the full maximum likelihood estimator. An analysis summarized in

TABLE 3.7 Comparison of Anscombe and full likelihood estimators: simulated datasets.

true parameters		no.† obsns	mean of $\frac{1}{2}(\tilde{\Delta}_A+\hat{\Delta})$	mean of $(\tilde{\Delta}_A-\hat{\Delta})$	t_{mean} ‡	$\frac{\text{var}\tilde{\Delta}_A}{\text{var}\,\hat{\Delta}}$	t_{var} §
margins*	Δ						
□	2	200	2.10	-0.003	-2.75	0.92	-0.35
▽	2	200	2.04	-0.001	-1.14	0.90	-0.45
□	4	200	4.16	-0.002	-1.94	1.09	0.34
○	4	200	4.07	-0.002	-0.84	0.84	-0.81
□	4	100	4.28	-0.005	-2.29	0.70	-2.04

* defined in Table 3.5

† per simulated dataset, 20 such samples generated from each 'true distribution'

‡ t for paired comparisons

§ using Morgan's method (1939) for testing equality of normal variances in presence of correlation.

Table 3.8 of several examples from Anscombe (1981) and Wahrendorf (1980) suggests that the full and Anscombe likelihood estimators are close to each other even when the underlying model is not GCR.

Using the asymptotic result that

$$-2\{\ell(\hat{\Delta},\hat{\underset{\sim}{\eta}},\hat{\underset{\sim}{\xi}};y) - \ell(0,\hat{\underset{\sim}{\eta}},\hat{\underset{\sim}{\xi}};y)\} \sim \chi_1^2 ,$$

we let the <u>likelihood-based 100(1-α)% confidence interval</u> for Δ to be delimited by the roots of the equation

$$d(\Delta;y) = \ell(\Delta,\hat{\underset{\sim}{\eta}},\hat{\underset{\sim}{\xi}};y) - \ell(\hat{\Delta},\hat{\underset{\sim}{\eta}},\hat{\underset{\sim}{\xi}};y) = \frac{1}{2}\chi_1^2(\alpha). \qquad (3.19)$$

The analogous interval with respect to Anscombe's cross-sectional likelihood is obtained on replacing the full maximum likelihood estimators for $\underset{\sim}{\eta}$ and $\underset{\sim}{\xi}$ in (3.19) by their marginal counterparts $\tilde{\underset{\sim}{\eta}}$ and $\tilde{\underset{\sim}{\xi}}$. Table 3.9 compares these confidence intervals for two pseudorandom and two published datasets, each having constant GCR distributions. The values were obtained by graphical interpolation. The difference between the two methods is minimal even for relatively small sample sizes, at 95% and 99% confidence levels: the Anscombe intervals appear slightly shorter than their maximum likelihood counterparts, as is expected, and larger divergence occurs for the generated datasets at the limit farther from $\Delta = 0$. Nevertheless it appears that for 95% and 99% confidence intervals, conclusions would not be substantially affected if the Anscombe intervals were used in place of the more expensive maximum likelihood.

TABLE 3.8 Comparison of Anscombe and full likelihood methods: published datasets

dataset	no. obsns	method‡	G^2§	df	$\hat{\Delta}$	parameter estimates							
						column cumulative probabilities/row cumulative probabilities							
Gilby* (1911, p.106)	1725	A	17.14	14	-1.002	0.369	0.804	0.958	0.075	0.202	0.438	0.748	0.966
		M	17.10		-1.004	0.367	0.802	0.958	0.075	0.203	0.439	0.748	0.966
Kendall and Stuart† (1977, p.29)	898	A	9.42	8	0.297	0.205	0.669	0.918	0.406	0.747	0.908		
		M	9.41		0.297	0.205	0.669	0.918	0.406	0.747	0.908		
Kiser and Whelpton* (1950, p.402)	1438	A	56.21	14	0.663	0.280	0.423	0.736	0.166	0.506	0.589	0.750	0.848
		M	56.20		0.665	0.281	0.422	0.735	0.167	0.508	0.592	0.752	0.850
Stuart*(1953)	3242	A	303.91	8	3.269	0.325	0.566	0.842	0.324	0.568	0.852		
		M	298.29		3.298	0.330	0.566	0.838	0.331	0.567	0.847		

* in Anscombe (1981, Chapter 12)

† in Wahrendorf (1980)

‡ A = Anscombe "cross-sectional" likelihood; M = full maximum likelihood

§ likelihood ratio goodness-of-fit statistic

TABLE 3.9 Likelihood-based confidence intervals for Δ: published datasets

dataset	no. obsns	method†	$\hat{\Delta}$	95%		99%	
				lower	upper	lower	upper
margins ∇*	100	A	2.04	1.51	2.56	1.36	2.72
$\Delta = 2$		M	2.05	1.51	2.58	1.36	2.76
margins $\square$*	100	A	3.94	3.01	4.98	2.74	5.38
$\Delta = 4$		M	3.95	3.01	5.00	2.72	5.42
Gilby (1911)	1725	A	-1.00	-1.16	-0.85	-1.21	-0.79
		M	-1.00	-1.17	-0.85	-1.22	-0.79
Kendall and Stuart (1963)	458	A	0.30	0.08	0.52	0.01	0.58
		M	0.30	0.08	0.52	0.01	0.58

* simulated dataset with true margins as shown in Table 3.7.

† A = Anscombe "cross-sectional" likelihood; M = full maximum likelihood.

3.5.4 Score test for no global association

Consider an r×c table of counts Y distributed as a multinomial with total n and probability structure given by (3.14) and (3.15). We derive for Y the locally most powerful statistic

$$T_\Delta = \partial \ell(\Delta, \tilde{\eta}, \tilde{\xi}; y)/\partial\Delta \,|\, \Delta = 0 \qquad (3.20)$$

to test H_0: $\Delta = 0$ against the alternative H_1: $\Delta > 0$ - see Cox and Hinkley (1974, p.324). Now T_Δ is the slope at $\Delta = 0$ of Anscombe's cross-sectional likelihood (3.18). Define

$$\pi_{ij}(\Delta) = \Pr[Z_1 = i,\ Z_2 = j] = \tfrac{1}{2}\Sigma_{ij}(\Delta)/(e^{\Delta}-1)$$

where $\Sigma_{ij}(\Delta) = -S(\Delta,\eta_i,\xi_j) + S(\Delta,\eta_{i-1},\xi_j) + S(\Delta,\eta_i,\xi_{j-1}) - S(\Delta,\eta_{i-1},\xi_{j-1})$,

with the $S(\Delta,.,.)$ as in (3.16). As $\Delta \to 0$, $S(\Delta,\eta,\xi) \to 1$ for all $\eta,\xi \in (0,1)$, and thus by the continuity of the logarithm and l'Hôpital's rule, it can be shown that in the limit as $\Delta \to 0$

$$\ln \pi_{ij}(\Delta) = -\Delta + \ln\{\Sigma'_{ij}(\Delta)\} - \ln 2$$

$$\frac{\partial}{\partial\Delta}\{\ln \pi_{ij}(\Delta)\} = \tfrac{1}{2}\{\Sigma''_{ij}(\Delta)/\Sigma'_{ij}(\Delta) - 1\}, \qquad (3.21)$$

where prime (') denotes partial differentiation w.r.t. Δ. It is straightforward to prove that

$$S'(0,\eta,\xi) = \eta + \xi - 2\eta\xi$$

$$S''(0,\eta,\xi) = -4\eta^2\xi^2 + 4\eta\xi(\eta+\xi) - 6\eta\xi + \eta + \xi,$$

where $\eta,\xi \in (0,1)$, whence

$$\Sigma'_{ij}(0) = 2(\eta_i-\eta_{i-1})(\xi_j-\xi_{j-1});$$

$$\Sigma''_{ij}(0) = 2(\eta_i-\eta_{i-1})(\xi_j-\xi_{j-1})\{2(\eta_i+\eta_{i-1}-1)(\xi_j+\xi_{j-1}-1) + 1\},$$

yielding from (3.21)

$$\lim(\partial/\partial\Delta)\ln \pi_{ij}(\Delta) = (\eta_i+\eta_{i-1}-1)(\xi_j+\xi_{j-1}-1) \qquad (3.22)$$

as $\Delta \to 0$. Now setting

$$\eta_{io} = n^{-1} \sum_{q\le i} Y_{q+} \quad , \quad i = 1,\ldots,r-1;$$

$$\xi_{jo} = n^{-1} \sum_{s\le j} Y_{+s} \quad , \quad j = 1,\ldots,c-1,$$

the maximum likelihood estimators of $\underset{\sim}{\eta}$ and $\underset{\sim}{\xi}$ under H_0: $\Delta = 0$, it follows from (3.20) and (3.22) that

$$T_\Delta = n^{-2} \sum_{i=1}^{r}\sum_{j=1}^{c} Y_{ij}\{2\sum_{q=1}^{i-1} Y_{q+} + Y_{i+} - n\}\{2\sum_{s=1}^{j-1} Y_{+s} + Y_{+j} - n\}$$

$$= 4 \sum_{i=1}^{r}\sum_{j=1}^{c} Y_{ij}(R_{i+}-\bar{R})(R_{+j}-\bar{R})$$

where $nR_{i+} = \sum_{q=1}^{i-1} Y_{q+} + \frac{1}{2}(Y_{i+}+1)$, $i = 1,\ldots,r$, $\bar{R} = \frac{1}{2}(1+n^{-1})$, and the R_{+j} are defined analogously: R_{i+} and R_{+j} are the average relative ranks for row i and column j respectively - c.f. (2.10). Hence the score statistic T_Δ is equivalent to Spearman's rho for tied ranks (Stuart, 1963): it is a linear function of the product-moment correlation of the marginal relative ranks.

For testing H_0 in the several-table case with a common Δ and varying marginal parameters, the locally most powerful statistic based on (3.20) is equivalent to the sum of several statistics of the above form, with within-table marginal ranks replacing the R_{i+} and R_{+j}. This statistic has a χ_1^2 distribution under H_0: $\Delta = 0$, provided that there are enough tables to invoke the Central Limit Theorem. It is

appropriate even for tables with many zero entries, since the asymptotic assumptions require that the number of tables becomes large, instead of the number of observations per cell. Asymptotic distributions of goodness-of-fit statistics for such tables are discussed in Part B of this thesis.

3.6 Summary

We have presented a flexible class of models for bivariate, discrete, ordered responses. GCR models exploit the ordering of the marginal response variables, since the association between them is defined in terms of probabilities of quadrants.

The GCR can be interpreted as a ratio of odds on conditional events. It may vary with a covariate, or with cutpoints, or both - see (3.4) to (3.7). For marginal distributions taking the linear logistic model, each marginal regression parameter may be considered to be a log odds on cumulative probabilities. Many other parameterizations of the marginal cumulative probabilities are possible, including the generalized linear models of McCullagh (1980) - see §1.4. The fit of GCR models was discussed in §3.3 and demonstrated in §3.4.

An approach was suggested by Anscombe (1981) for estimating the GCR of a constant GCR model for a single r×c contingency table by maximizing a cross-section of the likelihood function (§3.5). For the situations studied, the estimator of the log GCR behaved in general similarly to the full maximum likelihood estimator. A locally most powerful score statistic for testing for no association was shown in §3.5.4 to be equivalent to Spearman's rho. Properties of GCR models are discussed and compared to those of local association models (Goodman, 1979) in Chapter 4.

4

DISCUSSION

4.1 Introduction

Having described global cross-ratio (GCR) models in Chapter 3, we aim in this chapter to put them in the context of the statistical literature. The local association models presented by Goodman (1979) also apply to two-way contingency tables with both margins ordered. These models are defined in §4.2. Their properties are compared with those of GCR models in §4.3, with emphasis on the association aspects. It transpires that local association is most appropriate when both components of the response variable are inherently discrete (§1.2) and it is meaningful to compare adjacent categories. In contrast, the definition of GCRs best lends their application to bivariate responses which are discretized continuous (§1.2). A summary of the arguments in §4.3 appears in Table 4.1.

4.2 Local Association Models

Local association models for two-way contingency tables were introduced by Goodman (1979) for tables with row and column categories ordered. Consider a bivariate response $\underset{\sim}{Z} = (Z_1, Z_2)$, and let $\pi_{ij} = \Pr[Z_1 = i, Z_2 = j]$, $i = 1,\ldots,r$; $j = 1,\ldots,c$, with $\Sigma\Sigma\pi_{ij} = 1$. Suppose that n independent observations are taken from this distribution. Then the results may be summarized in an r×c contingency table Y, with Y_{ij}, the (i,j)-th entry, being the number of observations for which $Z_1 = i$ and $Z_2 = j$. For convenience we have labelled the marginal categories with integers: this does not assume anything about the spacing of the categories or the marginal distributions. The local association at cell (i,j) is given by

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \quad , \qquad i = 1,\ldots,r-1; \quad j = 1,\ldots,c-1.$$

$$= \frac{\{\pi_{ij}/(\pi_{ij}+\pi_{i,j+1})\}/\{\pi_{i,j+1}/(\pi_{ij}+\pi_{i,j+1})\}}{\{\pi_{i+1,j}/(\pi_{i+1,j}+\pi_{i+1,j+1})\}/\{\pi_{i+1,j+1}/(\pi_{i+1,j}+\pi_{i+1,j+1})\}} \qquad (4.1)$$

- see also §2.4.

Thus conditional on $\underset{\sim}{Z} \in \{(i,j),(i,j+1),(i+1,j),(i+1,j+1)\}$, θ_{ij} is the odds on $\{Z_2 = j | Z_1 = i\}$ versus the odds on $\{Z_2 = j | Z_1 = i+1\}$, a "local odds ratio". If the θ_{ij} are allowed to vary freely for all i and j, then the θ_{ij} together with the marginal probabilities $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ completely specify the cell probabilities. With this much generality it is not necessary for the marginal categories to be ordered, since the local association for cell (i,j) can be expressed as a ratio of association θ^{*}_{ij} of (i,j) with respect to any cell. For example, consider the cell (r,c), and define

$$\theta'_{ij} = \pi_{ij}\pi_{rc}/\{\pi_{ic}\pi_{rj}\} . \tag{4.2}$$

Then

$$\theta_{ij} = \frac{\theta'_{ij}\theta'_{i+1,j+1}}{\theta'_{i,j+1}\theta'_{i+1,j}} \tag{4.3}$$

and conversely

$$\theta'_{ij} = \prod_{q=i}^{r-1}\prod_{s=j}^{c-1}\theta_{qs}, \qquad i = 1,\ldots,r-1; \quad j = 1,\ldots,c-1, \tag{4.4}$$

showing the equivalence of the $\{\theta_{ij}\}$ and $\{\theta'_{ij}\}$.

Intuitively, the local association measure is more useful for data with ordered categories. If $\theta_{ij} > 1$ for all (i,j), this indicates that observations falling in rows i or i+1 and in columns j or j+1 are more likely to occur in the diagonal cells of this 2×2 submatrix i.e. with both variables "high" or both "low". We now present parameterizations for the cell probabilities which yield restrictions on the θ_{ij}. Let

$$\pi_{ij} = \alpha_i\beta_j \exp(\phi\mu_i\nu_j), \qquad i = 1,\ldots,r; \quad j = 1,\ldots,r;$$

where suitable identifiability constraints are imposed on the parameters, e.g.

$$\prod\alpha_i = \prod\beta_j = 1 ; \tag{4.5}$$

$$\Sigma\mu_i = \Sigma\nu_j = 0 ; \tag{4.6}$$

$$\Sigma\mu_i^2 = \Sigma\nu_j^2 = 1. \tag{4.7}$$

This is Model II of Goodman (1979). It is loglinear only if the $\{\mu_i\}$ and $\{\nu_j\}$ are known. The local association θ_{ij}

$$\ln\theta_{ij} = \phi(\mu_{i+1}-\mu_i)(\nu_{j+1}-\nu_j), \qquad i = 1,\ldots,r-1; \tag{4.8}$$

$j = 1,\ldots,c-1$. For the $\ln \theta_{ij}$ to be of the same sign for all (i,j), it is enough that the $\{\mu_i\}$ and $\{\nu_j\}$ are monotonic sequences. Departures of the fitted scores from monotonicity suggest either that the imposed ordering does not apply to the association, or that adjacent categories overlap.

If slightly different constraints are put on the $\{\mu_i\}$ and $\{\nu_j\}$, i.e.

$$\Sigma\pi_{i+}\mu_i = \Sigma\pi_{+j}\nu_j = 0$$

$$\text{and} \quad \Sigma\pi_{i+}\mu_i^2 = \Sigma\pi_{+j}\nu_j^2 = 1, \tag{4.9}$$

then Goodman (1981b) shows that the scores obtained under maximum likelihood and simple conditions are similar to those for the principal canonical correlation in canonical correlation analysis. Such an analysis which is not model-based is described for example in Kendall and Stuart (1967, p.568). Goodman suggests an ordering of the row and column categories based on the ordering of the fitted scores from either analysis. It would surely be fairer to one's prior assumptions to postulate an ordering <u>before</u> carrying out the analysis and at worst constrain adjacent scores to be equal. This is what McCullagh (1980) does for cutpoint parameters for the discrete singly-ordered response- see also §1.4.

Another perhaps more intuitive parameterization is Goodman's (1979) Model I, where the cell probabilities satisfy

$$\pi_{ij} = \theta\alpha_i\beta_j\gamma_i^{j}\delta_j^{i} \tag{4.10}$$

with the superscripts denoting powers, under suitable identifiability constraints, for example:

$$\prod \alpha_i = \prod \beta_j = \prod \gamma_i = \prod \delta_j = 1$$

and

$$\Sigma(\ln \gamma_i)^2 = \Sigma(\ln \delta_j)^2 = 1. \tag{4.11}$$

Equation 4.10 defines a loglinear model, and the local association

$$\theta_{ij} = (\gamma_{i+1}/\gamma_i)(\delta_{j+1}/\delta_j), \tag{4.12}$$

is linear in the logarithms of the $\{\gamma_i\}$ and $\{\delta_j\}$.

Constraints on the parameters of models I and II reduce dependence of θ_{ij} to the row variable or the column variable, or to be constant for all cells. Row-effect association models are defined by δ_{j+1}/δ_j = constant in (4.12) for model I or $\nu_{j+1}-\nu_j$ = constant in (4.8) for model II. Thus with respect to local association there is no ordering of rows imposed, but column effects on the $\ln \theta_{ij}$ are linear for model I and multiplicative for model II. Under the uniform association model, with $\theta_{ij} = \theta$ for all (i,j), models I and II are equivalent.

4.3 Comparison of Local and Global Association Models

4.3.1 Introduction

We have discussed measures of association in general in Chapter 2. In this subsection we compare models for local and global association, as defined in §§4.2 and 3.2 respectively. The emphasis is on the association aspects of the models. Since cross-ratios are invariant under reweighting of rows or columns (§4.3.5 and Bishop, Fienberg and Holland, 1975, p.14), we lose little by focusing attention on the cross-ratios themselves. Firstly the models are related in terms of Lehmann's (1966) measures of dependence. Invariance of the models under re-ordering of the marginal categories is studied. We examine the effects on both types of model of poorly-defined boundaries between marginal categories. Interpretation of parameters is considered. The closeness of constant or uniform association models to the discretized bivariate normal is discussed as in the literature. These arguments are summarized in Table 4.1.

4.3.2 Lehmann's measures of dependence

Several concepts of dependence between pairs of random variables $\underset{\sim}{Z} = (Z_1, Z_2)$ were introduced by Lehmann (1966), enabling different types of association to be classified in terms of their "strength". The first type of dependence, positive quadrant dependence, holds if

$$\Pr[Z_1 \leq z_1,\ Z_2 \leq z_2] \geq \Pr[Z_1 \leq z_1]\Pr[Z_2 \leq z_2] \qquad (4.13)$$

for all z_1, z_2. This inequality asserts that the "direction" of departures from independence of the cumulative distribution function is the same throughout the bivariate sample space. If (4.13) is satisfied, Lehmann shows that Kendall's τ and Spearman's rho are both positive.

Now the GCR model with $\psi_{ij} > 1$ satisfies (4.13) at $\underset{\sim}{Z} = (i,j)$, for if we let $\eta_i = \Pr[Z_1 \leq i]$, $\xi_j = \Pr[Z_2 \leq j]$ and $F_{ij} = \Pr[Z_1 \leq i, Z_2 \leq j]$ with $\eta_i, \xi_j, F_{ij} \in (0,1)$, then by (3.1)

$$\frac{F_{ij}(1-\eta_i-\xi_j+F_{ij})}{(\eta_i-F_{ij})(\xi_j-F_{ij})} = \psi_{ij} > 1,$$

whence it follows easily that $F_{ij} > \eta_i \xi_j$ at cutpoint (i,j). Thus GCR models with all GCR's > 1 have positive quadrant dependence.

A stronger condition is for concordant pairs of observations (a,b) and (a',b') (i.e. $a-a'$ and $b-b'$ of the same sign - see §2.2). Let $\pi_{ab} = \Pr[Z_1 = a, Z_2 = b]$, $a = 1,\ldots,r$; $b = 1,\ldots,c$. Then the bivariate distribution of $\underset{\sim}{Z}$ defined by the $\{\pi_{ab}\}$ shows <u>positive likelihood ratio dependence</u> if

$$\pi_{ab}\pi_{a'b'} \geq \pi_{ab'}\pi_{a'b} \tag{4.14}$$

for (a,b) and (a',b') concordant.

It is easy to show that local association models with all $\theta_{ij} > 1$ satisfy this property, for as in (4.4)

$$\frac{\pi_{ab}\pi_{a'b'}}{\pi_{ab'}\pi_{a'b}} = \prod_{i=a}^{a'-1} \prod_{j=b}^{b'-1} \theta_{ij} > 1,$$

assuming without loss of generality that $a < a'$ and $b < b'$. To show that positive likelihood ratio dependence implies positive quadrant dependence, observe by (4.14) for $a = 1,\ldots,r-1$; $b = 1,\ldots,c-1$ that

$$\pi_{ab}\pi_{rc} - \pi_{ac}\pi_{br} \geq 0.$$

Summing the left side over a = 1,...,i; b = 1,...,j yields

$$\pi_{rc}\Pr[Z_1 \leq i,\ Z_2 \leq j] - \Pr[Z_1 \leq i]\Pr[Z_2 \leq j] \geq 0.$$

Since $0 < \pi_{rc} < 1$, (4.13) is established.

The following counter-example illustrates that there are global cross-ratio models with all global cross-ratios greater than one which do not satisfy positive likelihood ratio dependence. Consider the 2×3 table of cell probabilities:

2/11	1/11	2/11
1/11	2/11	3/11

Then

$$\psi_{11} = 2\times5/(1\times3) = 10/3 > 1;$$
$$\psi_{12} = 3\times3/(3\times2) = 3/2 > 1;$$
$$\theta_{11} = 2\times2/(1\times1) = 4 > 1;$$
$$\theta_{12} = 1\times3/(2\times2) = 3/4 < 1.$$

Thus a local departure from positive likelihood-ratio dependence does not necessarily violate positive quadrant dependence. For applications where choice between adjacent categories is subject to some error but interest is in cumulative effects, the preferable measure of association is the GCR. This problem may occur when the boundary between adjacent categories is poorly defined - see §4.3.4.

4.3.3 Effect of re-ordering the marginal variables

How much is the form of a model for bivariate, discrete, ordered responses affected by changing the ordering imposed on the row and column categories? We investigate the effect on the association parameters ψ and θ for GCR and local association models.

Consider a saturated model for an r×c contingency table of probabilities, i.e. one in which there are rc - 1 identifiable parameters. Then the model may be specified in terms of

(i) cell probabilities π_{ij}, $i = 1,\ldots,r$; $j = 1,\ldots,c$;

(ii) local cross-ratios θ_{ij}, $i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$ and $\pi_{11},\pi_{21},\ldots,\pi_{r1}$; $\pi_{12},\pi_{13},\ldots,\pi_{1,c-1}$ (don't need to specify $\theta_{r-1,c-1}$); or

(iii) global cross ratios ψ_{ij}, $i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$ and the marginal cumulative probabilities $\eta_i = \Pr[Z_1 \leq i]$, $i = 1,\ldots,r-1$; $\xi_j = \Pr[Z_2 \leq j]$, $j = 1,\ldots,c-1$, with restrictions on $\underset{\sim}{\eta},\underset{\sim}{\xi},\psi$ to keep the cell probabilities positive.

These three methods are equivalent, for (i) implies (ii) by definition of the $\{\theta_{ij}\}$; (i) implies (iii) by definition of the $\{\psi_{ij}\}$, $\{\eta_i\}$ and $\{\xi_j\}$; (ii) implies (i) by the algorithm

$$\pi_{ij} = \frac{\theta_{i-1,j-1}\pi_{i-1,j-1}}{\pi_{i-1,j}\pi_{i,j-1}}, \qquad i = 2,\ldots,r;$$

$j = 2,\ldots,c$; $\{\pi_{1j}\}$ and $\{\pi_{i1}\}$ are prespecified; (ii) follows from (iii) since by (3.2) and (3.3) the cumulative probabilities $F_{ij} = \Pr[Z_1 \leq i, Z_2 \leq j]$ may be expressed in terms of η_i, ξ_j and ψ_{ij}, and the cell probabilities are obtained from

$$\pi_{ij} = F_{ij} - F_{i-1,j} - F_{i,j-1} + F_{i-1,j-1},$$

where $F_{00} = F_{i0} = F_{0j} = 0$, $i = 1,\ldots,r$; $j = 1,\ldots,c$, with $\pi_{ij} \geq 0$ for suitable values of $\underset{\sim}{\eta}$, $\underset{\sim}{\xi}$ and ψ. Thus in this fully parameterized case, it is always possible to define a global or local association model. However, any ordering imposed on the association parameters may be violated if the ordering of the row or column variables is changed arbitrarily. For example consider a 2×3 table in which the columns are in the order (3,1,2) with respect to a table $[\pi_{ij}]$ with local cross-ratios greater than one. The permuted table of probabilities is

π_{13}	π_{11}	π_{12}
π_{23}	π_{21}	π_{22}

.

We denote local and global association parameters with respect to the unpermuted table by $\theta_{ij}, \psi_{ij} > 1$, $j = 1,2$, and attach an asterisk (*) for the permuted table shown. Thus

$$\theta^*_{11} = \frac{\pi_{13}\pi_{21}}{\pi_{23}\pi_{11}} = \frac{\pi_{12}\pi_{21}}{\pi_{11}\pi_{22}} \cdot \frac{\pi_{22}\pi_{13}}{\pi_{12}\pi_{23}}$$

$$= \theta_{11}^{-1}\,\theta_{12}^{-1} < 1;$$

$$\theta^*_{12} = \theta_{11} > 1.$$

Hence the permuted table does not show positive likelihood ratio dependence. Similarly

$$\psi^*_{11} = \frac{\pi_{13}(\pi_{21}+\pi_{22})}{\pi_{23}(\pi_{11}+\pi_{12})} = \psi_{12}^{-1} < 1;$$

$$\psi^*_{12} = \frac{(\pi_{13}+\pi_{11})\pi_{22}}{(\pi_{23}+\pi_{21})\pi_{12}} \quad .$$

Note that ψ_{12}^{*} is not even derived from a double dichotomy of the original table.

If the ordering on the row or column variables is reversed, this has the result of inverting each cross-ratio, whether local or global. When both marginal orderings are reversed, it follows that the cross-ratios remain unchanged. The effect of reversing the ordering of a marginal variable on marginal parameters depends on their invariance properties. For the general parameterization of Goodman (1979) and GCR models with marginal cumulative probabilities under generalized linear models with symmetric link functions (§1.4), the marginal parameter structures are left essentially the same except for a sign change.

4.3.4 Effect of splitting or combining marginal categories

When the "dividing line" between two adjacent ordered marginal categories is not clear-cut, irregularities may be introduced into a contingency table whose local and global cross-ratios would otherwise all be greater than one. We discuss the effect on the table's cross-ratios of introducing a single marginal cutpoint which is ill-defined. We show that, because of the cumulative nature of global cross-ratios, they are distorted much less than the local cross-ratios. Thus it is preferable to use GCRs in the presence of ill-defined cutpoints: local cross-ratios are better employed in detecting such cutpoints. When it is a question of combining adjacent categories, all remaining GCRs are left unchanged: only local association parameters in the region of the merger will change.

An ill-defined marginal cutpoint. Consider the bivariate response $\underset{\sim}{Z}$ which gives rise to $\pi_{ij} = \Pr[Z_1 = i, Z_2 = j]$. Suppose that there is a continuous random variable T and cutpoints $-\infty = t_0 < t_1 < t_2 < \ldots < t_r = \infty$ such that $t_{i-1} < T < t_i$ if and only if $Z_1 = i$, $i = 1,\ldots,r-1$. Consider a random variable τ uniformly distributed on (t_{i-1}, t_i) independently of Z_1 and Z_2, and let $Z_1 = i'$ if $t_{i-1} < Z_1 \leq \tau$ and $Z_1 = i''$ if $\tau < Z_1 \leq t_i$. Thus τ is an ill-defined cutpoint between t_{i-1} and t_i, and yields an extra row in the contingency table, with row categories $1,\ldots,i-1, i', i'' \equiv i, i+1,\ldots,r$. Since τ is independent of $\underset{\sim}{Z}$, it follows that $\pi_{i''j} = \pi_{i'j} = \frac{1}{2}\pi_{ij}$ for all j, and given that an observation lies in row i, it occurs in i' or i'' with probability $\frac{1}{2}$. Cross-ratios from the new table are given an asterisk superscript (*). Clearly local cross-ratios not involving i' or i'' are unchanged. For $j = 1,\ldots,c-1$

$$\theta^*_{i-1,j} = \frac{\pi_{i-1,j}\pi_{i',j+1}}{\pi_{i-1,j+1}\pi_{i',j}} = \frac{\pi_{i-1,j}\,\frac{1}{2}\pi_{i,j+1}}{\pi_{i-1,j+1}\,\frac{1}{2}\pi_{ij}} = \theta_{i-1,j};$$

$$\theta^*_{i',j} = \frac{\pi_{i'j}\pi_{i'',j+1}}{\pi_{i',j+1}\pi_{i'',j}} = \frac{\frac{1}{2}\pi_{ij}\,\frac{1}{2}\pi_{i,j+1}}{\frac{1}{2}\pi_{i,j+1}\,\frac{1}{2}\pi_{ij}} = 1;$$

$$\theta^*_{i'',j} = \theta_{i,j}.$$

Local cross-ratios involving both rows i' and i'' equal one. This also occurs if $\pi_{i'j} : \pi_{i''j}$ is a fixed ratio for all j. All other local cross-ratios are left unaltered. Similarly for global cross ratios, all except the $\psi_{i',j}$, $j = 1,\ldots,c-1$ are unchanged. Suppose that the 2×2 matrix of probabilities obtained from the double dichotomy at $(i,j) \equiv (i'',j)$ is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

Let $\varepsilon = \Sigma_{q \leq j} \pi_{i'q}$ and $\delta = \pi_{i'+} - \varepsilon$. Then

$$\psi_{i'j} = \psi_{ij} \frac{(1-\varepsilon/A)(1+\delta/D)}{(1-\delta/B)(1+\varepsilon/C)} \quad . \tag{4.15}$$

Thus

$$\begin{aligned} \ln \psi_{i'j} = \ln \psi_{ij} &- \varepsilon/A - \tfrac{1}{2}\varepsilon^2 A^{-2} + \delta/D - \tfrac{1}{2}\delta^2 D^{-2} \\ &+ \delta/B + \tfrac{1}{2}\delta/B^2 - \varepsilon/C + \tfrac{1}{2}\varepsilon^2/C^2 \\ &+ o(\varepsilon^2/A^2,\ \delta^2/D^2,\ \delta^2/B^2,\ \varepsilon^2/C^2). \end{aligned}$$

Thus if A,B,C,D are not too small and ε and δ are of similar magnitude, there is not much change to the global cross-ratio in row i'. If there is a possibility of an ill-defined cutpoint, then local cross-ratios are better for pin-pointing it. Alternatively, when the aim is to describe association despite an ill-defined cutpoint, the above arguments suggest that GCRs are preferable.

In the case of several ill-defined cutpoints the approach would be either to model the measurement error, or to have only a few marginal categories. The former method was adopted by Farewell (1982) in the context of discrete hazard functions for univariate responses - see §1.3. Andrich (1978, 1979) and Anderson (unpublished) include a discrimination parameter for each cutpoint in their multinomial logistic models (§1.5.1).

Merging adjacent marginal categories. Suppose that row categories i-1 and i of the r×c table of probabilities are combined, and call the new category i'. Then the resulting table of probabilities has (r-1)×c entries, and only (r-2)×(c-1) cross-ratios. By definition, the remaining GCRs are identical to those in the original table. However, 2(c-1) of the local cross-ratios change - the "locality" of these

parameters has widened. Thus the uniform local association models of (2.12) and (2.13) are affected by such a move: it is necessary to adjust the row score of the new larger category, presumably to $\frac{1}{2}(a_i+a_{i-1})$. Since the merging of adjacent marginal categories especially at the end of the range is common practice, these effects are important.

4.3.5 Interpretation as odds ratios

An advantage of using cross-ratios to measure association is that they may be expressed as odds ratios. In the 2×2 case, where $\pi_{ij} = \Pr[Z_1 = i, Z_2 = j]$, $i,j = 1,2$, the cross ratio $\psi = \pi_{11}\pi_{22}/(\pi_{12}\pi_{21})$ may be written

$$= \frac{\pi_{11}/\pi_{1+}}{\pi_{12}/\pi_{1+}} \cdot \frac{\pi_{22}/\pi_{2+}}{\pi_{21}/\pi_{2+}}$$

$$= \frac{\Pr[Z_2 = 1|Z_1 = 1]}{\Pr[Z_2 = 2|Z_1 = 1]} \cdot \frac{\Pr[Z_2 = 2|Z_1 = 2]}{\Pr[Z_2 = 1|Z_1 = 2]}$$

$$= \frac{\text{odds}\{Z_2 = 1|Z_1 = 1\}}{\text{odds}\{Z_2 = 1|Z_1 = 2\}}, \qquad (4.16)$$

where odds (E) = $\Pr[E]/\{1 - \Pr[E]\}$. Thus ψ is an odds ratio of conditional probabilities. By symmetry ψ is equal to a similar expression to (4.16) with the roles of Z_1 and Z_2 reversed:

$$\psi = \text{odds}\{Z_1 = 1|Z_2 = 1\}/\ \text{odds}\{Z_1 = 1|Z_2 = 2\}.$$

If $\Pr[Z_1 = 1|Z_2 = 1]$ and $\Pr[Z_1 = 1|Z_2 = 2]$ are small, ψ approximates well to the relative risk $(\pi_{11}/\pi_{+1})/(\pi_{12}/\pi_{+2})$ - see for example Mantel and Haenszel (1959, p.730). Now odds ratios are invariant under changes

to the marginal distribution, for suppose that π_{1+} becomes $\alpha\pi_{1+}$ and π_{2+} becomes $(1-\alpha)\pi_{2+}$, $0 < \alpha < 1$, then $\pi_{1j} \to \alpha\pi_{1j}$, $\pi_{2j} \to (1-\alpha)\pi_{2j}$, but

$$\psi \to \frac{\alpha\pi_{11}(1-\alpha)\pi_{22}}{\alpha\pi_{12}(1-\alpha)\pi_{21}} = \psi .$$

The local cross ratio θ_{ij} for an r×c contingency table (§4.2) may be expressed, conditional on $Z_1 = i$ or $i+1$ and $Z_2 = j$ or $j+1$, as an odds ratio:

$$\theta_{ij} = \text{odds}\{Z_1 = i \mid Z_2 = j,\ Z_1 = i \text{ or } i+1\}/\text{odds}\{Z_1 = i \mid Z_2 = j+1,\ Z_1 = i \text{ or } i+1\},$$

$i = 1,\ldots,r-1$; $j = 1,\ldots,c-1$. This is meaningful if the main interest is in 2×2 subtables of adjacent cells. For example, change in political allegiance could well be only to the adjacent party.

In contrast, the GCR ψ_{ij} may be expressed as an odds ratio of cumulative events:

$$\psi_{ij} = \text{odds}\{Z_1 \le i \mid Z_2 \le j\}/\text{odds}\{Z_1 \le i \mid Z_2 > j\}$$
$$= \text{odds}\{Z_2 \le j \mid Z_1 \le i\}/\text{odds}\{Z_2 \le j \mid Z_1 > i\}.$$

Thus the odds on a marginal variable not exceeding a value are compared for the other marginal variable below and above a cutpoint, respectively. This measure is of interest if both marginal variables may be dichotomized at successive cutpoints, as for example when there is an underlying continuous scale.

4.3.6 Comparison with continuous distributions

How close are association models to some bivariate continuous counterparts? We examine especially the constant association models, where the association parameter ψ or θ has the same value for all cutpoints, and their relationship to the discretized bivariate normal. This subject was discussed by Goodman (1981a) who pointed out the similarity of form of the standard bivariate normal density at $(X_1 = x_1, X_2 = x_2)$ which has a factor $\exp\{-\frac{1}{2}x_1x_2\rho/(1-\rho^2)\}$ in the density, and local association model II $\pi_{ij} = \alpha_i\beta_j \exp(\phi\mu_i\nu_j)$. When the local association scores $\{\mu_i\}$ and $\{\nu_j\}$ are equally spaced with identifiability conditions (4.9), and the margins transformed to discretized normal probabilities, then one may equate $\rho/(1-\rho^2) \simeq \phi$ and compare the local association model cell probabilities with those for the bivariate normal, discretized at the corresponding equally spaced cutpoints.

GCR models are traditionally related to the bivariate normal by using the continuous version with unit normal margins. The densities at (0,0) are equated, suggesting that $\rho = (\psi-1)/(\psi+1)$ - see Plackett (1965), and Anscombe (1981, p.306). All three distributions in their discretized form are compared for $\rho = 0.5$ with spacing 0.5 between marginal categories - see Goodman (1981a, Table 1). Cell probabilities for the uniform local association model are generally closer to the bivariate normal than the corresponding GCR model probabilities. However this is a special example in several respects. Firstly, the marginal cutpoints are equally spaced along a normal density. This suggests that marginal probabilities are smaller at either end of the scale and become gradually larger and smaller again, i.e. for some i, $\pi_{1+} \leq \pi_{2+} \leq \cdots \leq \pi_{i+}$; $\pi_{i+} \geq \pi_{i+1,+} \geq \cdots \geq \pi_{r+}$. Indeed, Goodman has

pointed out that if the marginal cutpoint spacings for the discretized bivariate normal vary excessively, it is no longer approximated well by a uniform local association model. Secondly the assumption of a normal surface may be incorrect in practice. If the aim is to estimate the normal correlation, the method of polychoric correlation, introduced by Lancaster and Hamdan (1964) and discussed in §2.3, leads directly to an estimate of ρ. Observe that the continuous version of the global association model is available, and is not restricted to normal margins - see Plackett (1965) and Mardia (1970, §8).

4.4 Conclusion

In this chapter we have sought to compare local and global association models for an r×c contingency table. The main thread of argument is outlined in Table 4.1. Local association models with all associations exceeding unity satisfy a stronger form of dependence than corresponding global association models. Thus an ill-defined marginal cutpoint can affect nearby values of local association drastically, while changing global associations only slightly. If a marginal variable is re-ordered arbitrarily, the dependence structure of both models may be disrupted. However, reversal of the ordering merely inverts all the association parameters: reversing the order of both margins leaves all the association parameters unchanged. The interpretation of the global cross-ratio as an odds ratio of conditional cumulative events is more intuitive than the local odds interpretation of the local association parameter. For special cases, the local association model is a good approximation to a discretized bivariate normal distribution.

How do the two models compare in practical use? Time, and the types of application encountered will tell. The local association models appear best suited to situations where the marginal categories are inherently discrete, and where relationships between adjacent categories are important. In contrast, GCR models are preferable for marginal scales which have an underlying continuum. They can cope with ill-defined categories, and the interpretation of the cross-ratios as conditional odds ratios of cumulative probabilities is appealing.

TABLE 4.1 Comparison of global and local association

criterion	sub-section of §4.3	performance	
		global association ψ_{ij}, §3.2	local association θ_{ij}, §4.2
positive quadrant dependence	2	satisfied if all $\psi_{ij} \geq 1$	satisfied if all $\theta_{ij} \geq 1$
positive lik. ratio dependence	2	not necessarily satisfied	satisfied if all $\theta_{ij} \geq 1$
reordering marginal categories	3		
-arbitrary		ψ_{ij} loses double dichotomy interpretation	local association model, but sign ln θ_{ij} may change
-reverse order		$\psi^*_{ij} = \psi^{-1}_{ij}$	$\theta^*_{ij} = \psi^{-1}_{ij}$
splitting/merging adjacent marginal categories	4	little/no change in ψ_{ij}	drastic changes possible only for θ_{ij} near split/merge
odds ratio of events	5	odds$\{Z_1 \leq i \mid Z_2 \leq j\}$	odds$\{Z_1 = i \mid Z_2 = j,$ $Z_1 = i$ or $j\}$
discretization, for constant association	6	constant (global) association model, margins specified	discretized bivariate normal, equally spaced marginal cutpoints

B

GOODNESS-OF-FIT STATISTICS FOR SPARSE CONTINGENCY TABLES

NOTATION B

A summary of the notation used in Part B is presented in this section. Brief definitions are given followed by references to the definition given in the text. Where a theorem is given as a reference, the symbol in question was defined in the statement or proof of the theorem. Minor local notations are not listed.

A_{jk}, A_k	marginal configuration for contingency table, §5.3.
B_{jk}, B_j	marginal configuration for contingency table, §5.3.
$D_i(\underset{\sim}{\theta})$	p×r matrix with (s,j)-th entry $p_{ij}^{-1}(\underset{\sim}{\theta})\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s$, Theorem 6.3.
G^2, $G^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k)$, G_k^2, $G_k^2(\underset{\sim}{\theta})$	likelihood ratio goodness-of-fit statistic, §§5.1, 6.2.1, 6.3.3.
I, I_k	index set or configuration for contingency table, §5.3.
$I(x,\lambda)$	Poisson information kernel, §5.2
M	upper bound on elements of $\underset{\sim}{p}_i'(\underset{\sim}{\theta})$, condition 8, §6.3.2.
M'	upper bound on elements of $p_i''(\underset{\sim}{\theta})$, condition 9, §6.3.2.
N	upper bound on n_i, Theorem 6.2, condition 3(ii), §6.3.2.
$O_p(1)$, $o_p(1)$	stochastic order, §6.3.3
Poi	Poisson, §5.2

Q_i	$Z_{gi}-m_{gi}$, Theorem 6.2.
$Q_i(\underset{\sim}{\theta})$	i-th contribution to $G_k^2(\underset{\sim}{\theta})$, Theorem 6.4.
R	the real line, $(-\infty,\infty)$.
$R_i(\underset{\sim}{\theta})$	$\ell_i(\underset{\sim}{\theta}) - \ell_i(\underset{\sim}{\theta}_o)$, Proposition 6.1.
$R_i(\underset{\sim}{\theta},\rho)$	$\sup\{R_i(\underset{\sim}{\theta}): \|\underset{\sim}{\theta}-\underset{\sim}{\theta}_o\| \leq \rho\}$, Proposition 6.1.
$V_i(\rho)$	$\sup\{R_i(\underset{\sim}{\theta}): \|\underset{\sim}{\theta}\| > \rho\}$, Proposition 6.1.
$V_i(\underset{\sim}{\theta})$	variance-covariance matrix of $(\underset{\sim i}{Y}^{2T}, \underset{\sim i}{Y}^T)^T$, Theorem 6.3.
$W_i(\underset{\sim}{\theta})$	$\underset{\sim i}{a}^T(\underset{\sim}{\theta})\underset{\sim i}{Y}^2 + \underset{\sim i}{b}^T(\underset{\sim}{\theta})\underset{\sim i}{Y} - n_i$, Theorem 6.3.
X^2, $X^2(\underset{\sim k}{Y};\underset{\sim k}{p})$, X_k^2, $X_k^2(\underset{\sim}{\theta})$	Pearson goodness-of-fit statistic, Lemma 6.2, §§5.1, 6.2.2, 6.2.3.
X_i	sequence of independent random variables, Lemmas 6.4, 6.6.
$\underset{\sim k}{X}$	k-vector of independent random variables $X_{ik} \sim \text{Poi}(\mu_{ik})$, §5.2.
$\underset{\sim}{Y}$, $\underset{\sim i}{Y}$	multinomially distributed r-vector, §§6.2.1, 6.2.2, 6.3.1.
$\underset{\sim}{Y}^2$	$(Y_1^2,\ldots,Y_r^2)^T$, Theorem 6.3.
$Y(\underset{\sim}{i},I)$	observed count in cell $\underset{\sim}{i}$ of contingency table indexed by I, §5.3.
$\underset{\sim}{Y}(I)$	vector of counts in contingency table indexed by I, §5.3.
$\underset{\sim k}{Y}$	k-vector of observed counts, §5.2.
Z_{gi}	i-th term of G_k^2, §6.2.2.
$Z_i(\underset{\sim}{\theta})$	i-th term of $X_k^2(\underset{\sim}{\theta})$, Theorem 6.3.
Z_{xi}	i-th term of X_k^2, §6.2.2.
$a_{ij}(\underset{\sim}{\theta})$	$\{n_i p_{ij}(\underset{\sim}{\theta})\}^{-1}$, Theorem 6.3.
$a_k(\underset{\sim}{\theta})$	$w_{(k)}(\underset{\sim}{\theta}_o)/w_{(k)}(\underset{\sim}{\theta})$, Theorem 6.4.

$\underset{\sim}{b}_i(\underset{\sim}{\theta})$	$\underset{\sim}{\mu}_k^T(\underset{\sim}{\theta})\Gamma^{-1}(\underset{\sim}{\theta})\underset{\sim}{D}_i(\underset{\sim}{\theta})$, Theorem 6.4.
bin	binomial, §6.3.4.
$b_k(\underset{\sim}{\theta})$	$\{g_k(\underset{\sim}{\theta}_o) - g_k(\underset{\sim}{\theta})\}/w_{(k)}(\underset{\sim}{\theta}_o)$, Theorem 6.4.
$f_k(\underset{\sim}{\theta})$	sequence of functions continuous in $\underset{\sim}{\theta}$, uniformly in k, Lemma 6.5.
$g_k(\underset{\sim}{\theta})$	$E[G_k^2(\underset{\sim}{\theta}_1)\|\underset{\sim}{\theta}_1 = \underset{\sim}{\theta}]$, Theorem 6.4.
k_α^*	$\Phi^{-1}(1-\alpha)$, §7.2.4.
$\ell(x;\lambda)$	Poisson log likelihood, §5.2.
$\ell_i(\underset{\sim}{\theta};y_i)$, $\ell_i(\underset{\sim}{\theta})$	contribution to log-likelihood for $\underset{\sim}{Y}_i \sim mn_r(n_i,p_i(\underset{\sim}{\theta}))$, §6.3.1.
m_{gi}	EZ_{gi}, §6.2.2.
$m_{g(k)}$	asymptotic mean for G^2, §5.2.
m_k'	exact mean for X^2, §5.2.
$mn_k(n,\underset{\sim}{p})$	multinomial distribution on a k-vector $\underset{\sim}{Y}$ with cell probabilities $\underset{\sim}{p}$ and $n = \underset{\sim}{1}_k^T\underset{\sim}{Y}$, §§5.2, 6.2.1.
$m_{x(k)}$	asymptotic mean for X^2, §5.2; k(r-1), §6.2.2.
n, n_i, n_k, n_t	sample size of (1-th, k-th) multinomial; t-th binomial sample, §§6.2, 6.3.1, 5.1; 6.3.4.
n(I)	total number of observations in table $\underset{\sim}{Y}(I)$, §5.3.
$\underset{\sim}{p}$, $\underset{\sim}{p}_i$, $\underset{\sim}{p}_i(\underset{\sim}{\theta})$	r-vector of cell probabilities, §§6.2.2, 6.3.1.
$\underset{\sim}{p}(\underset{\sim}{i},I)$	cell probability in cell i of contingency table indexed by I, §5.3.
$\underset{\sim}{p}(I)$	vector of cell probabilities in contingency table indexed by I, §5.3.
$\underset{\sim}{p}_k$	k-vector of cell probabilities, §5.1.
$p_t(\underset{\sim}{\theta})$	binomial probability, §6.3.4.
$q_k(\underset{\sim}{\theta})$	$v_{(k)}(\underset{\sim}{\theta}_o)/v_{(k)}(\underset{\sim}{\theta})$, Theorem 6.3.
r	number of cells per multinomial, §6.2.1.

t_α upper 100(1-α)% point of a distribution with zero mean and variance one, §7.2.1.

v_{gi}^2 var Z_{gi}, §6.2.2.

$v_{g(k)}^2$ asymptotic variance for G^2, §5.2.

$v_i^2(\underset{\sim}{\theta})$ var $W_i(\underset{\sim}{\theta})$, Theorem 6.3.

$v_k'^2$ exact variance for X^2, §5.2.

v_{xi}^2 var Z_{xi}, §6.2.2.

$v_{x(k)}^2$ asymptotic variance for X^2, §5.2.

$w_i^2(\underset{\sim}{\theta})$ var $Q_i(\underset{\sim}{\theta})$, Theorem 6.4.

$\mathcal{A}$, $\mathcal{B}$ set of configurations, all marginal to the same contingency table, §5.3.

$\Gamma(\underset{\sim}{\theta})$ $\lim k^{-1}\Gamma_{(k)}(\underset{\sim}{\theta})$, §6.3.1.

$\Gamma_i(\underset{\sim}{\theta})$ expected Fisher information for $\underset{\sim}{Y}_i$, §6.3.1.

Θ parameter space, §6.3.1.

$\Phi(.)$ standard normal cumulative probability distribution function, §7.2.2.

γ_1, γ_2 skewness, kurtosis of distribution with mean zero and variance one, §§7.2.1, 7.2.4.

δ lower bound on: elements of $\underset{\sim}{p}$, Lemma 6.3; $n_i p_i$, Theorem 6.1; $\underset{\sim}{p}_i$, Theorem 6.2; $\underset{\sim}{p}_i(\underset{\sim}{\theta})$, condition 3(i), §6.3.2.

δ' $|p_{ij}^{-1} - r^{-1}| > \delta'$, some j, if $n_i = 1$, Lemma 6.3, Theorems 6.1, 6.2, condition 10, §6.3.2.

$\underset{\sim}{\theta}$, $\hat{\underset{\sim}{\theta}}_k$ p-dimensional unknown; its maximum likelihood estimator based on $\underset{\sim}{Y}_1,\ldots,\underset{\sim}{Y}_k$, §6.3.1.

$\underset{\sim}{\theta}_o$ "true" value of $\underset{\sim}{\theta}$, §6.3.1.

μ_i EX_i, Lemma 6.4.

$\mu(\underset{\sim}{i},I)$, $\underset{\sim}{\mu}(I)$	$\underset{\sim}{i}$-th cell mean; vector of cell means in contingency table indexed by I, §5.3
μ_{ik}	$n_i p_{ik}$, §5.1.
$\underset{\sim}{\mu}_k$	k-vector of cell means μ_{ik}, §5.1.
$\underset{\sim}{\mu}_k(\underset{\sim}{\theta})$	$k^{-1}E[\underset{\sim}{X}_k^{2\prime}(\underset{\sim}{\theta})\,\vert\,\underset{\sim}{\theta}]$, Theorem 6.3.
$\underset{\sim}{1}_q$	the q-dimensional column vector of ones
' (prime)	partial derivative w.r.t. $\underset{\sim}{\theta}$, §6.3.1.
bracketed subscript	$a_{(s)} = \sum_{i=1}^{s} a_i$, §6.2.1.
$\#(A_j)$, $\#(I)$	number of cells in configuration A_j; in contingency table indexed by I, §5.3.
$\cap$	$A \cap B$, A,B configurations: the set of indices obtained by summing over all indices summed over in either A or B, §5.3.
$\emptyset$	The null configuration, denoting summation over all indices, §5.3.

5

THE SPARSE MULTINOMIAL : DISTRIBUTION OF X^2 AND G^2

5.1 Sparseness

It is often interesting to test the fit of data to a multinomial distribution. Let

$$\underset{\sim}{Y}_k = (Y_{1k},\ldots,Y_{kk})^T$$

and suppose that we wish to test whether $\underset{\sim}{Y}_k \sim mn_k\ (n_k, \underset{\sim}{p}_k)$, where $1_{\sim k}^T \underset{\sim}{p}_k = 1$ and $1_{\sim k}^T \underset{\sim}{Y}_k = n_k$. Two widely used statistics are Pearson's X^2 (1900)

$$X^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k) = \sum_{i=1}^{k} (Y_{ik} - n_k p_{ik})^2/(n_k p_{ik}) \qquad (5.1)$$

and the likelihood ratio statistic (Neyman and Pearson, 1928)

$$G^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k) = 2 \sum_{i=1}^{k} Y_{ik} \ln \{Y_{ij}/(n_k p_{ik})\}. \qquad (5.2)$$

It is well known that for fixed k and $\underset{\sim}{p}_k$ and under the hypothesized distribution, both X^2 and G^2 approximate a χ^2_{k-1} distribution as n_k becomes large. This implies that each cell mean $\mu_{ik} = n_k p_{ik}$ becomes large in the limit. Cochran (1954) gives a working rule that for the X^2 statistic to satisfy this assumption it is enough that $\mu_{ik} \geq 1$ for all i and $\mu_{ik} \geq 5$ for at least 80% of the i. Koehler and Larntz (1980) observe that the chi-squared approximation for G^2 is adequate if n_k/k lies between 0.5 and 1 and the p_{ik} are equal. But what if conditions of this sort are not met? For example, consider a study in which several discrete variables are recorded

for each subject. Suppose that very few respondents have the same attributes, and thus many of the cells in the contingency table of observations contain low counts or zeros. The distribution of such a table has two main characteristics: a large number of cells k, and a high proportion of small cell means μ_{ik}. A contingency table of counts with a large number of zero or low count cells is said to be sparse: conditions on the distribution of a contingency table which keep the number of cells k large and most of the cell means μ_{ik} small are called sparseness conditions.

Since the type of data described above occur frequently in practice, it is important to study the behaviour of goodness-of-fit statistics applied to sparse contingency tables. In §5.2 we discuss an approach by Morris (1975) which gives sparseness conditions for a multinomial distribution to ensure the asymptotic normality of X^2 and G^2 when the sequence of probability vectors $\underset{\sim}{p}_k$ is known. An extension due to Koehler (unpublished) for loglinear models whose cell means may be estimated in closed form by maximum likelihood is studied in §5.3. These methods are illustrated in §5.4 by an analysis of a dataset on the similarity of two poems.

5.2 Known Cell Probabilities

Consider the case where the multinomial distribution of $\underset{\sim}{Y}_k$ is completely specified for all k. Since its elements Y_{ik} are correlated, the terms in the goodness-of-fit statistics do not in general form a sequence of independent random variables. Morris (1975, Lemma 2.2) sidesteps this problem by a conditioning argument. He shows that the sum of a sequence of certain functions f_{ik} of independent sparse random

variables $X_{ik} \sim \text{Poi}(\mu_{ik})$, have the same asymptotic distribution, whether or not they satisfy $\Sigma_i(X_{ik}-\mu_{ik}) = 0$. An important condition on the f_{ik} is that $\text{cov}(f_{ik}(X_{ik}),X_{ik}) = 0$, $i = 1,\ldots,k$. Hence the normalizing constants in Theorem 5.1 for a single sparse multinomial are based on independent Poisson random variables whose means equal the expected cell frequencies of the multinomial. Morris also considers the distributions of X^2 and G^2 under certain alternative hypotheses: in this subsection we simply present his result for the null hypothesis. Firstly we define the normalizing constants for the goodness-of-fit statistics.

For each k let $\underset{\sim}{X}_k$ be a vector of independent random variables with $X_{ik} \sim \text{Poi}(\mu_{ik})$ where $\mu_{ik} = n_k p_{ik} = EY_{ik}$, $i = 1,\ldots,k$. Then set

$$m_{x(k)} = k,$$
$$v^2_{x(k)} = 2k + n_k^{-1}\left(\sum_{i=1}^{k} p_{ik}^{-1} - k^2\right).$$

These are the asymptotic mean and variance required for the X^2 statistic. Let the log likelihood for $X \sim \text{Poi}(\lambda)$ be

$$\ell(x;\lambda) = x - x \ln \lambda \qquad \text{if } x > 0$$
$$= -\lambda \qquad \text{if } x = 0.$$

Now the Kullback-Liebler Poisson information kernel for X is

$$I(x,\lambda) = \ell(x;x) - \ell(x;\lambda)$$
$$= x \ln (x/\lambda) - x + \lambda \qquad \text{if } x > 0$$
$$= \lambda \qquad \text{if } x = 0.$$

Then the normalizing constants required for the G^2 statistic are

$$m_{g(k)} = 2 \sum_{i=1}^{k} EI(X_{ik},\mu_{ik}) \tag{5.3}$$

$$v^2_{g(k)} = 4 \sum_{i=1}^{k} \text{var}[I(X_{ik},\mu_{ik}) - X_{ik} n_k^{-1} \sum_{j=1}^{k} \text{cov}\{I(X_{jk},\mu_{jk}),X_{ik}\}]. \tag{5.4}$$

The approach used by Morris in proving the following theorem enables the condition $1_{\sim k}^T Y_{\sim k} = n_k$ to be lifted as k becomes large. The problem is reduced to a central limit problem for a sum of independent random variables.

Theorem 5.1 [Asymptotic normality of X^2, G^2: known sparse multinomial] For $k = 1,2,\ldots$ define $Y_{\sim k} \sim mn_k(n_k,p_{\sim k})$ with $p_{ik} > 0$ known and $\mu_{ik} = n_k p_{ik}$, $i = 1,\ldots,k$. Suppose that the random vector $X_{\sim k}$ has independent components $X_{ik} \sim Poi(\mu_{ik})$, $i = 1,\ldots,k$. Let the following conditions hold:

(i) $\max_{1\le i\le k} p_{ik} = o(1)$ as $k \to \infty$;

(ii) there exists an $\varepsilon > 0$ such that $n_k p_{ik} \ge \varepsilon$ for $i = 1,\ldots,k$ and $k = 1,2,\ldots$.

Then in the limit as $k \to \infty$

(a) $\mathcal{L}[\{X^2(Y_{\sim k};p_{\sim k}) - m_{x(k)}\}/v_{x(k)}] \to N(0,1)$;

(b) $\mathcal{L}[\{G^2(Y_{\sim k};p_{\sim k}) - m_{g(k)}\}/v_{g(k)}] \to N(0,1)$.

Proof: Morris [1975, Theorems 5.1 and 5.2].

Sparseness conditions (i) and (ii) permit the cell means to remain moderate, and prohibit any cell or finite subset of cells from containing too much probability in the limit. Observe that the limiting distributions

of X^2 and G^2 both depend upon the vector of cell probabilities $\underset{\sim}{p}_k$, in contrast to the result for the traditional fixed-k asymptotics. Further, the X^2 statistic has in general a different distribution to that of G^2. Koehler and Larntz (1980) explain this in terms of the minimum contribution of small observed counts to the statistics:

> 'For a cell with an expected frequency larger than one, an observed count of zero or one makes a larger minimum contribution to G_k^2 than X_k^2. Consequently, when most expected cell frequencies are in the range of 1.0 to 5.0, the first two moments for G_k^2 are larger than for X_k^2. The contribution to X_k^2 for a nonzero count can be quite large when the expected frequency is less than one, however, and the first two moments for X_k^2 are larger than the corresponding moments for G_k^2 when a sufficient number of expected frequencies are less than one.'

They also observe that a change from normalizing constants based on independent Poisson distributions to the exact moments based on the multinomial distribution does not affect the asymptotic normality of the X^2 statistic. It is illustrated as follows. Let m_k' and $v_k'^2$ denote the exact mean and variance of $X^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k)$, as given in Lemma 6.2(a) and (b), and let $m_k = m_{x(k)}$, $v_k = v_{x(k)}$. Then

$$\{X^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k) - m_k'\}/v_k' = \{(X^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k) - m_k)/v_k\}(v_k/v_k') + (m_k - m_k')/v_k'.$$

Now $v_k^2 - v_k'^2 = 2\{1 + (k-1)/n_k\}$.

Since $k^{-1} = \min_{\underset{\sim}{p}_k} (\max_{1\le i\le k} p_{ik})$,

it follows by condition (i) of Theorem 5.1 that $k^{-1}n_k \ge \varepsilon$, i.e. $kn_k^{-1} \le \varepsilon^{-1}$, implying that $v_k^2 - v_k'^2 = O(1)$ as $k \to \infty$, and thus that $v_k/v_k' \to K$, fixed.

Moreover, $m_k - m_k' = 1$, and by Lemma 6.2(b) $v_k'^2 \geq 2(k-1)$ as $k \to \infty$. Thus

$$(X^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k)-m_k')/v_k' = [\{X^2(\underset{\sim}{Y}_k;\underset{\sim}{p}_k)-m_k\}/v_k]\{K + o(1)\}+ O(k^{-\frac{1}{2}}) \qquad \text{as } k \to \infty,$$

whence result.

5.3 G^2 for Direct Loglinear Models

A large class of statistical models for contingency tables are the loglinear models in which the logarithms of cell probabilities are expressed as a linear combination of parameters. In this subsection only direct loglinear models, i.e. those with explicit maximum likelihood estimates for cell means, are considered. We examine an unpublished result of Koehler for the asymptotic normality of G^2 when observed and fitted values are compared under multinomial sampling. This is of interest since the dimension of the parameter space approaches infinity as the total number of cells becomes large, and the statistic is formulated using estimated parameters. The special case of a two-way table with the number of row and column categories becoming large and no association between the margins is mentioned by Fienberg (1979, §6). We present Koehler's general result of the asymptotic normality of G^2 for sparse contingency tables when a fixed number of marginal factors expand to infinitely many categories. Koehler also gives conditions under which the normalizing constants for G^2 may be replaced by maximum likelihood estimates. These conditions require that at least some of the cell means become large in the limit. Unfortunately the methods of Chapter 6 do not appear to be applicable here, since they are for finite-dimensional parameter spaces.

The contingency table structures discussed in this subsection may be expressed as cross-classifications of factors. We consider a sequence of contingency tables, the k-th of which has index set $I_k = \{(i_1,\ldots,i_q),\ i_\ell = 1,\ldots,r_\ell(k),\quad \ell = 1,\ldots,q\}$, with q factors. A marginal subtable is obtained from a larger table by summing over some of its factors. For example, given a two-way table with index set $I_k = \{(i,j)\colon i = 1,\ldots,r_k;\ j = 1,\ldots,c_k\}$, the row marginal table consists of the row totals of the original table, and is indexed by $A_k = \{(i,+)\colon i = 1,\ldots,r_k\}$, where the symbol "+" denotes summation over the index it replaces. In general, cell probabilities, counts and means are denoted by $p(\underset{\sim}{i},I)$, $Y(\underset{\sim}{i},I)$ and $\mu(\underset{\sim}{i},I)$ for each $\underset{\sim}{i} \in I$, and the entire table by $\underset{\sim}{p}(I)$, $\underset{\sim}{Y}(I)$ and $\underset{\sim}{\mu}(I)$, respectively. The total number of observations is given by

$$n(I) = \sum_{i \in I} Y(\underset{\sim}{i},I).$$

Cell probabilities, counts and means for marginal tables indexed by A_j are $p(\underset{\sim}{i},A_j)$, $Y(\underset{\sim}{i},A_j)$ and $\mu(\underset{\sim}{i},A_j)$, where $\underset{\sim}{i}$ is an index to the marginal table which corresponds to $\underset{\sim}{i} \in I$ in all components which are not "+". We sometimes refer to the index set A_j as a configuration, with $\#(A_j)$ cells.

Formally, a loglinear model for a contingency table indexed by I is direct if there exist sets of configurations $\mathcal{A} = \{A_1,\ldots,A_r\}$ and $\mathcal{B} = \{B_1,\ldots,B_{r-1}\}$ with each B_j properly marginal to at least one A_i such that for each $\underset{\sim}{i} \in I$

$$p(\underset{\sim}{i},I) = \frac{\prod_{j=1}^{r} p(\underset{\sim}{i},A_j)}{\prod_{q=1}^{r-1} p(\underset{\sim}{i},B_q)} . \qquad (5.5)$$

An algorithm for checking whether a loglinear model is direct is given by Bishop, Fienberg and Holland (1975, p.76). For each cell $\underset{\sim}{i} \in I$ the maximum likelihood estimator of the cell frequency $Y(\underset{\sim}{i},I)$ is given by

$$\hat{Y}(\underset{\sim}{i},I) = \{\prod_{j=1}^{r} Y(\underset{\sim}{i},A_j)\} / \{\prod_{q=1}^{r-1} Y(\underset{\sim}{i},B_q)\} \quad . \tag{5.6}$$

To obtain the correct units for $\hat{Y}(\underset{\sim}{i},I)$, it is necessary to have one more term in the numerator than in the denominator of (5.6). This may be achieved by setting some of the B_j to be the null configuration ϕ, with $Y(\underset{\sim}{i},\phi) = n(I)$. Consider for $k = 1,2,\ldots$ the contingency table of counts $\underset{\sim}{Y}(I_k)$ satisfying a multinomial distribution of dimension $\#(I_k)$ with cell probabilities $\underset{\sim}{p}(I_k)$ and total $n(I_k)$. Suppose that there are two sets of configurations $\mathcal{A}_k = \{A_{1k},\ldots,A_{rk}\}$ and $\mathcal{B}_k = \{B_{ik},\ldots,B_{r-1,k}\}$ which define a direct loglinear model for $\underset{\sim}{Y}(I_k)$, with cell means

$$\mu(\underset{\sim}{i},I_k) = \{\prod_{j=1}^{r} \mu(\underset{\sim}{i},A_{jk})\} / \{\prod_{q=1}^{r-1} \mu(\underset{\sim}{i},B_{qk})\}. \tag{5.7}$$

For each $\underset{\sim}{i} \in I_k$, define $\hat{p}(\underset{\sim}{i},I_k) = \hat{Y}(\underset{\sim}{i},I_k)/n_k$. Then, by observing that

$$\frac{Y(\underset{\sim}{i},I_k)}{\hat{Y}(\underset{\sim}{i},I_k)} = \frac{Y(\underset{\sim}{i},I_k) \prod_{j=1}^{r} \mu(\underset{\sim}{i}_j,A_{jk}) \prod_{q=1}^{r-1} Y(\underset{\sim}{i}_q,B_{qk})}{\mu(\underset{\sim}{i},I_k) \prod_{j=1}^{r} Y(\underset{\sim}{i}_j,A_{jk}) \prod_{q=1}^{r-1} \mu(\underset{\sim}{i}_q,B_{qk})} \quad ,$$

the likelihood ratio G^2 statistic for comparing observed counts to direct maximum likelihood estimates under (5.7) may be written

$$G^2(\underset{\sim}{Y}(I_k);\hat{\underset{\sim}{p}}(I_k)) = 2 \sum_{\underset{\sim}{i}\in I_k} Y(\underset{\sim}{i},I_k)\ln\{Y(\underset{\sim}{i},I_k)/\mu(\underset{\sim}{i},I_k)\}$$

$$-2 \sum_{j=1}^{r} \sum_{\underset{\sim}{i}\in A_{jk}} Y(\underset{\sim}{i},A_{jk})\ln\{Y(\underset{\sim}{i},A_{jk})/\mu(\underset{\sim}{i},A_{jk})\}$$

$$+ 2 \sum_{q=1}^{r-1} \sum_{\underset{\sim}{i}\in B_{qk}} Y(\underset{\sim}{i}_q,B_{qk})\ln\{Y(\underset{\sim}{i},B_{qk})/\mu(\underset{\sim}{i},B_{qk})\}.$$

$$= G^2(\underset{\sim}{Y}(I_k);\underset{\sim}{p}(I_k)) - \sum_{j=1}^{r} G^2(\underset{\sim}{Y}(A_{jk});\underset{\sim}{p}(A_{jk}))$$

$$+ \sum_{q=1}^{r-1} G^2(\underset{\sim}{Y}(B_{qk});\underset{\sim}{p}(B_{qk})). \qquad (5.8)$$

Now for k = 1,2,... consider a table $\underset{\sim}{X}(I_k)$ of independent random variables $X(\underset{\sim}{i},I_k) \sim Poi(\mu(\underset{\sim}{i},I_k))$ for each $\underset{\sim}{i} \in I_k$. Then by analogy with the statement of Theorem 5.1(a), the asymptotic mean for $G^2(\underset{\sim}{Y}(I_k);\hat{\underset{\sim}{p}}(I_k))$ is

$$m(I_k) = 2 \sum_{\underset{\sim}{i}\in I_k} EI(X(\underset{\sim}{i},I_k);\mu(\underset{\sim}{i},I_k))$$

$$-2 \sum_{j=1}^{r} \sum_{\underset{\sim}{i}\in A_{jk}} EI(X(\underset{\sim}{i},A_{jk});\mu(\underset{\sim}{i},A_{jk}))$$

$$+2 \sum_{q=1}^{r-1} \sum_{\underset{\sim}{i}\in B_{qk}} EI(X(\underset{\sim}{i},B_{qk});\mu(\underset{\sim}{i},B_{qk})), \qquad (5.9)$$

and the asymptotic variance $v^2(I_k)$ is obtained from (5.4) by substituting $X_{ik} = X(\underset{\sim}{i},I_k)$, $\mu_{ik} = \mu(\underset{\sim}{i},I_k)$ and $n_k = n(I_k)$, and summing over $\underset{\sim}{i} \in I_k$. Thus Koehler's results, which we state and prove below, follow because $G^2(\underset{\sim}{Y}(I_k);\hat{\underset{\sim}{p}}(I_k))$ may be split into terms involving true parameters rather than estimates. For simplicity we firstly state the result for the two-way table with both margins expanding, under no association - see Fienberg (1979, §6).

Theorem 5.2 [Asymptotic normality of G^2: two-way contingency table, both margins expanding, no association].

For $k = 1,2,\ldots$ consider an $r_k \times c_k$ contingency table of counts $\underset{\sim}{Y}(I_k)$ where $I_k = \{(i,j): i = 1,\ldots,r_k;\ j = 1,\ldots,c_k\}$ with row and column marginal configurations given by $A_{1k} = \{(i,+): i = 1,\ldots,r_k\}$ and $A_{2k} = \{(+,j) : j = 1,\ldots,c_k\}$. Suppose that a multinomial model on $n(I_k)$ observations is satisfied, with cell probabilities $\underset{\sim}{p}(I_k)$ such that for each $\underset{\sim}{i} \in I_k$

$$p(\underset{\sim}{i},I_k) = p(\underset{\sim}{i},A_{1k})\ p(\underset{\sim}{i},A_{2k}) .$$

Let the following conditions hold as $k \to \infty$:

(i) $r_k \to \infty$ and $c_k \to \infty$;

(ii) $\max_{\underset{\sim}{i} \in A_{jk}} p(\underset{\sim}{i},A_{jk}) = o(1), \qquad j = 1,2;$

(iii) there exists a fixed $\varepsilon > 0$ such that $\mu(\underset{\sim}{i},I_k) > \varepsilon$ for all $\underset{\sim}{i} \in I_k$ and all k.

Then in the limit as $k \to \infty$

$$\mathcal{L}[\{G^2(\underset{\sim}{Y}(I_k);\hat{\underset{\sim}{p}}(I_k)) - m(I_k)\}/v(I_k)] \to N(0,1).$$

Proof: Koehler, Theorem 4.1, and by Theorem 5.3.

We now present a more general theorem which allows several but not necessarily all of the factors in a contingency table to expand to infinitely many categories in the limit. An intersection ($\cap$) of two configurations is defined by summing over each factor which is summed over in either set. If all factors are summed over, this defines the "grand mean" configuration $\emptyset$.

Theorem 5.3 [Asymptotic normality of G^2: direct loglinear model, several expanding factors].

For $k = 1,2,\ldots$ let $\underset{\sim}{Y}(I_k)$ be a contingency table of counts with fixed total $n_k = n(I_k)$ and cell probabilities $\underset{\sim}{p}(I_k)$ under multinomial sampling. Let U_k be the configuration obtained by summing over all factors in I_k which remain bounded, and suppose that there exist sets of indices $\mathcal{A}_k = \{A_{1k},\ldots,A_{rk}\}$ and $\mathcal{B}_k = \{B_{1k},\ldots,B_{r-1,k}\}$ defining a direct loglinear model:

$$\mu(\underset{\sim}{i},I_k) = \frac{\prod_{j=1}^{r} \mu(\underset{\sim}{i},A_{jk})}{\prod_{q=1}^{r-1} \mu(\underset{\sim}{i},B_{qk})} \quad \text{for all } \underset{\sim}{i} \in I_k.$$

Let $n_{jk} = \#(A_{jk})$, $j = 1,\ldots,r$. Suppose that as $k \to \infty$, $n_k \to \infty$ and

(i) $n_{jk}/n_k = o(1)$, $j = 1,\ldots,r$;

(ii) $\max_{\underset{\sim}{i}\in C_k} p(\underset{\sim}{i},C_k) = o(1)$ for all configurations C_k such that $C_k \cap U_k \neq \emptyset$.

(iii) there exists a fixed $\varepsilon > 0$ such that $\mu(\underset{\sim}{i},I_k) > \varepsilon$ for all $\underset{\sim}{i} \in I_k$ and for all k.

Then in the limit as $k \to \infty$

$$\mathcal{L}[\{G^2(\underset{\sim}{Y}(I_k); \underset{\sim}{\hat{p}}(I_k)) - m(I_k)\}/v(I_k)] \to N(0,1).$$

Proof: [Koehler, Theorem 4.2].

The proof of this theorem depends on separating out the factors of I_k which have a bounded number of categories as $k \to \infty$. Now by (5.8) and (5.9)

$$\{G^2(\underset{\sim}{Y}(I_k);\hat{\underset{\sim}{p}}(I_k)) - m(I_k)\}/v(I_k) = \{G^2(Y(I_k);p(I_k)) - m(I_k)\}/v(I_k)$$

(5.10)

$$- \sum_{\substack{j: \\ A_{jk} \cap U_k \neq \emptyset}} \frac{v(A_{jk})}{v(I_k)} [\{G^2(\underset{\sim}{Y}(A_{jk});\underset{\sim}{p}(A_{jk})) - m(A_{jk})\}/v(A_{jk})] \quad (5.11)$$

$$- \sum_{\substack{j: \\ A_{jk} \cap U_k = \emptyset}} \{G^2(\underset{\sim}{Y}(A_{jk});\underset{\sim}{p}(A_{jk})) - m(A_{jk})\}/v(I_k) \quad (5.12)$$

$$+ \sum_{\substack{j: \\ B_{jk} \cap U_k \neq \emptyset}} \frac{v(B_{jk})}{v(I_k)} [\{G^2(\underset{\sim}{Y}(B_{jk});\underset{\sim}{p}(B_{jk})) - m(B_{jk})\}/v(B_{jk})] \quad (5.13)$$

$$+ \sum_{\substack{j: \\ B_{jk} \cap U_k = \emptyset}} \{G^2(\underset{\sim}{Y}(B_{jk});\underset{\sim}{p}(B_{jk})) - m(B_{jk})\}/v(I_k). \quad (5.14)$$

By Theorem 5.1(b) the first term has a limiting standard normal distribution. In (5.11) and (5.13), for each configuration $\#(A_{jk} \cap U_k) \to \infty$ and $\#(B_{jk} \cap U_k) \to \infty$, since each index in U_k not summed over has infinitely many possible values in the limit, by definition of U_k. Thus and by (ii), (iii) and Theorem 5.1(b), the terms in square brackets become standard normal as $k \to \infty$. Now for $C_k \in \mathcal{A}_k \cup \mathcal{B}_k$ with $C_k \cap U_k \neq \emptyset$

$$\frac{v(C_k)}{v(I_k)} = O(v(C_k)/k^{\frac{1}{2}}) = o(1) \qquad \text{as } k \to \infty$$

since $v(I_k) = O(k^{\frac{1}{2}})$ exactly, by (5.47) of Morris (1975). Thus (5.11) and (5.13) are $o_p(1)$ as $k \to \infty$. For all configurations $D_k \in \mathcal{A}_k \cup \mathcal{B}_k$ with $D_k \cap U_k = \emptyset$, $\#(D_k)$ is finite for all k. Thus we are in the

fixed-dimension situation. Now $\mu(\underset{\sim}{i},D_k) \to \infty$ for all $\underset{\sim}{i} \in D_k$ since each μ is defined in the limit as the sum of infinitely many cell means, each bounded below by ε, by (ii) and the definition of D_k. Hence as $k \to \infty$ $G^2(\underset{\sim}{Y}(D_k);\underset{\sim}{p}(A_j))$ has a limiting χ^2 distribution with $\#(D_k)-1$ degrees of freedom, and thus is bounded in probability. Since $v(I_k) = O(k^{\frac{1}{2}})$ exactly, (5.12) and (5.14) are $o_p(1)$ as $k \to \infty$. This completes the proof.

It is clear that Theorem 5.2 follows from Theorem 5.3, with $\mathcal{A}_k = \{A_{1k},A_{2k}\}$ and $\mathcal{B}_k = B_{1k}$ where $B_{1k} = \emptyset$, the "grand mean" configuration. Koehler goes on to prove a further result in which the number of factors with unbounded dimensionality may be unbounded in the limit. This is conceptually complicated, but is also a good approximation to what happens in some studies in practice: as the number of subjects in a study increases, more variables per subject are measured.

The main problem with the work discussed in this subsection is that the centring constants for the asymptotic distribution of G^2 depend on the true cell probabilities, although the G^2 statistic itself doesn't. Koehler proposed conditions to enable maximum likelihood estimates to be substituted in the centring constants for the true probabilities: they required that at least some of the cell means become larger in the limit. Another approach for a general finite-dimensional parameterization under product multinomial sampling is given in Chapter 6. The results of this present chapter are illustrated in the next subsection by the analysis of a dataset for comparing two poems.

5.4 Example: Measuring the Similarity of Two Poems

An interesting sparse dataset, presented by Eckler (1973), is derived from two poems each made from the same set of words, and same line and stanza length - vocabularyclept poems (Lindon and Bergerson, 1969). A poem was written by Bergerson and an alphabetic list of words and layout of the poem were published. The challenge was to produce a poem which satisfied these conditions. As it happened, the original title, "Winter Retrospect", was also suitable for Lindon's reconstruction.

The problem suggested by Eckler is to measure whether the two poems are independent. There are several ways that this could be done in principle, depending on the sense in which independence is understood. For example if sentence structure were important, we could examine the number of clauses per sentence: if independence of meaning, we could compare locations of "associated" words in the two poems.

We follow Eckler's approach, and examine the position of words which occur once in each poem. There are 162 such words, out of a total of 480. Each poem has been divided into 24 groups of 20 words. Entry (i,j) of Table 5.1, i,j = 1,...,24, is the number of singly-occurring words which fall in group i of Bergerson's poem and group j of Lindon's poem. Let p_{ij} = the probability of an observation in group (i,j), i,j = 1,...,24. We will consider the probability structure of Table 5.1, assuming multinomial sampling.

Suppose that one poem is a complete randomization with respect to the other. Then

$$H_o : p_{ij} = \frac{1}{576}, \qquad i,j = 1,\ldots,24$$

holds. If at the other extreme Lindon's poem is a replica of Bergerson's,

TABLE 5.1 The poem "Winter Retrospect": number of singly-occurring words in successive groups of twenty for both poems

Bergerson's poem: twenty-word group	Lindon's poem: twenty-word group																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	2	0	1	0	0	1	0	0	0	1	1	1	0	0	0	0	0	2	1	0	0	0	0	0
2	1	1	1	2	0	0	0	0	0	0	0	1	0	0	2	1	0	0	0	0	0	1	1	0
3	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
5	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	1	0	0	0	1	0	1	0
6	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	1	0
7	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	4	1	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	3	1	0	0	1
9	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	1	0	1	0
10	0	0	0	0	0	1	1	0	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0
11	0	0	1	0	0	0	4	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
12	0	0	0	0	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0	1	1	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	1	2	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1
15	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	1	0	2	0	0	0	1	0	0	0	0	0	0	0	1	0	0
17	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1
18	1	0	0	0	0	2	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0
19	0	2	0	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0
20	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0
21	0	1	0	0	1	0	0	0	0	0	1	0	0	0	2	0	0	1	0	0	0	1	0	1
22	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0
23	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
24	0	0	1	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	1	0	0	0	0

the situation is $H_1 : p_{ij} = \delta_{ij} p_i'$, $i,j = 1,\ldots,24$, with $\Sigma p_i' = 1$. In general $p_i' \neq 1/24$, since the singly-occurring words do not occur uniformly throughout Bergerson's poem. An intermediate situation $H_2 : p_{ij} = p_i q_j$, $i,j = 1,\ldots,24$ with $\Sigma p_i = \Sigma q_j = 1$ allows for different distributions within the poems of uniquely occurring words, without sacrificing independence in use of these words by the two authors. It appears intuitively that neither H_0 nor H_1 are appropriate to our dataset: H_2 is a more general "null hypothesis" that may be satisfied. We will examine the appropriateness of H_0 and H_2, using the Pearson X^2 goodness-of-fit statistic of §5.2 for a single, sparse multinomial. We denote by Z_x the X^2 statistic, standardized by the constants $m_{x(k)}$ and $v^2_{x(k)}$ of (5.3) and (5.4).

Under the equiprobable multinomial model H_0 the statistic $z_x = 1.93$ ($p \simeq 0.03$) and thus H_0 would be rejected at the upper 5% level, by Theorem 5.1(a). Now the skewness of X^2 for this model is 0.33, so by the discussion in Chapter 7, the upper 5% point of Z_x is "close" to the normal upper 5% percentile, i.e. the asymptotic result holds approximately. Under H_2, the exact values of row and column probabilities are unknown. Estimating them by maximum likelihood estimates in Z_x yields $z_x = 0.33$ ($p \simeq 0.37$, assuming $Z_x \sim N(0,1)$). We are making strong assumptions about the asymptotic distribution of X^2: for the G^2 statistic, this approach was valid under conditions which required that some cell probabilities remained large - see the discussion after Theorem 5.3, and Koehler (unpublished, §5). An examination of the residuals for both models points to a poor fit of the counts of 4, for both models, i.e. cells (7,21) and (11,7), which correspond to the words (crude, oven, loaf, bakes) and (head, hand, elbow, quilt). The residuals are reduced from 7.01

for the fit under H_0 to 4.62 and 6.22 for H_2. These two groups of four words were used in close proximity by both poets, suggesting that they were also associated in meaning in both poems.

Eckler (1973) carried out an informal analysis of Table 5.1. He proposed independent Poisson models for each cell of the table, with each singly-occurring word falling in a cell at mean rate 162/576 = 0.28 in each cell. Since the departure of observed from expected counts was appreciable only for the counts of 4 (see Table 5.2), his conclusions were similar to ours.

TABLE 5.2. Observed and expected counts under a Poisson model: dataset of Table 5.1

Count per cell	Number of cells Observed	Expected
0	439	435.0
1	117	112.0
2	17	17.3
3	1	1.64
4	2	0.12
Total	576	576.06

We slightly prefer the multinomial model, since the number of singly-occurring words is fixed by the first poem. However, the asymptotic distribution of X^2 for a large, sparse contingency table with each cell distributed independently as a Poisson is normal, with the same centring constants as under conditions (i) and (ii) of Theorem 5.1, by the Liapounov Central Limit Theorem [Morris, 1975, in the proof of his Lemma 2.2]. Thus the standardized X^2 for

independent Poissons is equal to that for the sparse multinomial, given that cell means and observations agree.

Eckler did similar analyses based on subdivision of the poems into groups of 40 and 80 words. In the 80-word analysis, the same 14-word cluster was used by both poets - in fact both poems commenced with a similar idea. An analysis by Bishop, Fienberg and Holland (1975, 340-342) of Table 5.1 is essentially the same as that of Eckler. Their claim that both poems have 24 lines is false - they have 57 lines each. For the text of the poems, see the original article by Lindon and Bergerson (1969) or Eckler (1973).

6

SPARSE PRODUCT MULTINOMIALS : DISTRIBUTION OF X^2 AND G^2

6.1 Introduction

Asymptotic distributions of goodness-of-fit statistics for sparse product multinomial distributions are studied in this chapter. Each contributing multinomial is of fixed dimension, but the number of these distributions in the product becomes large in the limit. This regime has an effect similar to the single sparse multinomial of Chapter 5 on cell means: the covariance structure is different. In §6.2, the asymptotic normality of the X^2 and G^2 statistics is established for a product of completely specified multinomials, under mild sparseness conditions. The approach is generalized in §6.3 to allow cell probabilities to depend on a finite-dimensional unknown which is estimated by maximum likelihood. Regularity conditions for the parameterization are based on those of Hoadley (1971). Both statistics are shown to be asymptotically normal, with asymptotic variances different from those for the completely specified case. A binary logistic model which satisfies the criteria for §6.3.3 is presented in §6.3.4.

6.2 Completely Specified Distributions

6.2.1 Notation and preliminaries

Let $\{\underset{\sim}{Y}_i\}_{i=1}$ be a sequence of independent, discrete random variables, where $\underset{\sim}{Y}_i = (Y_{i1},\ldots,Y_{ir})^T$, $r \geq 2$, with $\underset{\sim}{1}^T\underset{\sim}{Y}_i = n_i$, $i = 1,\ldots,k$. Consider the multinomial distributions $mn_r(n_i,\underset{\sim}{p}_i)$, with $\underset{\sim}{p}_i = (p_{i1},\ldots,p_{ir})^T$ such that $p_{ij} > 0$, $j = 1,\ldots,r$, and $\underset{\sim}{1}^T\underset{\sim}{p}_i = 1$, $i = 1,\ldots,k$. The distributions are completely specified. The Pearson statistic for measuring the fit of the $\{\underset{\sim}{Y}_i\}$ to the given multinomial distribution is

$$X_k^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} (Y_{ij} - n_ip_{ij})^2/(n_ip_{ij}). \tag{6.1}$$

Similarly we define the likelihood ratio statistic for testing the fit of the $\{\underset{\sim}{Y}_i\}$ to the same sequence of distributions to be

$$G_k^2 = 2\sum_{i=1}^{k}\sum_{j=1}^{r} Y_{ij}\ln\{Y_{ij}/(n_ip_{ij})\}.$$

Interest is in the limiting distribution of these statistics as $k \to \infty$, while the cell means n_ip_{ij} remain bounded. Limits of summation Σ may be omitted if these limits are obvious. The sum from 1 to k of a_i is denoted by $a_{(k)}$. Firstly we require some preliminary results.

Lemma 6.1

Let $p_i > 0$, $i = 1,\ldots,r$, satisfy $\Sigma p_i = 1$.
Then $\Sigma p_i^{-1} \geq r^2$, with equality if and only if all $p_i = r^{-1}$.

Proof. By the Lagrange Identity [Rao, 1973, p.55],

$$0 \leq \frac{1}{2}\sum_{i\neq j}\sum\{\sqrt{(p_i/p_j)} - \sqrt{(p_j/p_i)}\}^2 = \Sigma p_i^{-1}\Sigma p_j - (\Sigma p_i^{-\frac{1}{2}+\frac{1}{2}})^2 = \Sigma p_i^{-1} - r^2,$$

which is equal to zero if and only if $p_i = r^{-1}$, $i = 1,\ldots,r$.

Lemma 6.2

Let $\underset{\sim}{Y} = (Y_1,\ldots,Y_r)^T \sim mn_r(n,\underset{\sim}{p})$ with $p_j > 0$, $j = 1,\ldots,r$ and $\underset{\sim}{1}_r^T\underset{\sim}{p} = 1$. Define $X^2 = \Sigma_j(Y_j - np_j)^2/(np_j)$. Then

(a) $E\,X^2 = r - 1$;

(b) $\text{var}\,X^2 = 2(1 - n^{-1})(r - 1) + n^{-1}(\Sigma p_j^{-1} - r^2)$;

(c) $E[(X^2 - EX^2)^3] = 8(r - 1) + 2n^{-1}\{(2 - 3r)(3r + 8) + 11\Sigma p_j^{-1}\}$
$+ n^{-2}\{2(r^3 + 9r^2 + 14r - 12) - (3r + 22)\Sigma p_j^{-1} + \Sigma p_j^{-2}\}$;

(d) $E\{(X^2 - EX^2)^4\} = 12(r - 1)(r + 3) + 12n^{-1}\{(r + 31)\Sigma p_j^{-1}$
$- (r^3 + 25r^2 + 44r - 38)\} + n^{-2}\{112\Sigma p_j^{-2} + 3(\Sigma p_j^{-1})^2$
$- 2(3r^2 + 118r + 658)\Sigma p_j^{-1} + 3(r^4 + 44r^3 + 328r^2 + 488r - 380)\}$
$+ n^{-3}\{\Sigma p_j^{-3} - 4(r + 28)\Sigma p_j^{-2} - 3(\Sigma p_j^{-1})^2 + 4(3r^2 + 56r + 236)\Sigma p_j^{-1}$
$- 6(r^4 + 20r^3 + 116r^2 + 160r - 120)\}$.

Proof. See Haldane (1937).

Lemma 6.3

Consider $\underset{\sim}{p} = (p_1,\ldots,p_r)^T$, $r \geq 2$, with $\underset{\sim}{p}^T\underset{\sim}{1}_r = 1$, and suppose that there exist constants δ and δ' such that

(i) $p_j \geq \delta > 0$, $j = 1,\ldots,r$;

(ii) there is a j ($j = 1,\ldots,r$) such that $|p_j - r^{-1}| \geq \delta' > 0$.

Then there exist constants ε and ε' such that

(a) $\Sigma p_j^{-1} - r^2 \geq \varepsilon > 0$;

(b) $\Sigma_j p_j\{\ln p_j - (\Sigma_q p_q \ln p_q)\}^2 \geq \varepsilon' > 0$.

Proof. Let $S = \{(p_1,\ldots,p_r)$: hypotheses (i) and (ii) hold$\}$. Then the expressions in (a) and (b) are continuous, real-valued functions of $\underset{\sim}{p}$ defined on S, which is a closed, bounded subset of R^r. By Royden [1968, Chapter 2, Proposition 17], the minimum of both these functions is attained in S. By (ii) and Lemma 6.1, (a) is positive

for all values of $\underset{\sim}{p}$ in S. Observe that (b) is a variance which is positive unless all p_j are equal. Hence the required ε and ε' exist.

6.2.2 Asymptotic distributions of X^2 and G^2

We present conditions under which the X^2 and G^2 statistics approach normality as the number of independent multinomials k in the sequence $\{Y_i\}_{i=1}$ increases. The requirements for the asymptotic normality of X^2 are slightly weaker than those for G^2, since the first few moments of each component of X^2 may be kept finite by bounding the cell means uniformly away from zero. Firstly we set up some notation. Consider a sequence of independent random variables $\underset{\sim}{Y}_1,\ldots,\underset{\sim}{Y}_k$ defined as in §6.2.1 and indexed by i = 1,...,k. Let

$$Z_{xi} = \sum_{j=1}^{r} (Y_{ij} - n_i p_{ij})^2/(n_i p_{ij})$$

and $v_{xi}^2 = \text{var } Z_{xi}$, as in Lemma 6.2(b). Then

$$X_k^2 = \sum_{i=1}^{k} Z_{xi},$$

with mean $m_{x(k)} = k(r - 1)$ and variance $v_{x(k)}^2 = \Sigma_i v_{xi}^2$. Similarly define

$$Z_{gi} = 2 \sum_{j=1}^{r} Y_{ij} \ln\{Y_{ij}/(n_i p_{ij})\},$$

$m_{gi} = EZ_{gi}$ and $v_{gi}^2 = \text{var } Z_{gi}$. Then

$$G_k^2 = \sum_{i=1}^{k} Z_{gi},$$

with mean $m_{g(k)} = \Sigma_i m_{gi}$ and variance $v_{g(k)}^2 = \Sigma_i v_{gi}^2$. All the constants $(n_i, \underset{\sim}{p}_i)$, i = 1,...,k, are assumed known.

Theorem 6.1. [Asymptotic normality of X^2]

Let $\underset{\sim}{Y}_i \sim mn_r(n_i, \underset{\sim}{p}_i)$, $i = 1,\ldots,k$, be a sequence of independent random variables. Suppose that there exist constants δ and δ' such that

(i) $n_i p_{ij} > \delta > 0$ for all $j = 1,\ldots,r$; $i = 1,\ldots,k$;

(ii) if $n_i = 1$, there exists a j ($j = 1,\ldots,r$) for which

$$|p_{ij} - r^{-1}| > \delta' > 0.$$

Let Z_{xi}, X_k^2 and $v_{x(k)}^2$ be defined as above. Then in the limit as $k \to \infty$,

$$\mathcal{L}[\{X_k^2 - k(r-1)\}/v_{x(k)}] \to N(0,1).$$

Proof. We denote the fourth central moment of Z_{xi} and X_k^2 by μ_{4i} and $\mu_{4(k)}$ respectively. Now by Lemma 6.2(b), and also by (ii) and Lemma 6.3(a) if $n_i = 1$, for all i there exists a constant $m(r,\delta)$ such that $v_{xi}^2 \geq m(r,\delta) > 0$. By (i),

$$0 \leq n_i^{-a} \Sigma_j p_{ij}^{-a} < r\,\delta^{-a}, \quad a = 1,2,\ldots .$$

Then, substituting into Lemma 6.2(d), there is a constant $M(r,\delta)$ such that for all i, $0 \leq \mu_{4i} < M(r,\delta)$. Thus

$$0 \leq \frac{\mu_{4(k)}}{(v_{x(k)}^2)^2} = \frac{\Sigma_i \mu_{4i}}{(\Sigma v_{xi}^2)^2} < k^{-1} \frac{M(r,\delta)}{m^2(r,\delta)} \to 0 \quad \text{as } k \to \infty.$$

The result follows by Liapounov's Central Limit Theorem [Loève, 1963, §20, B(ii), p.275] for the fourth absolute moment.

Theorem 6.2. [Asymptotic normality of G^2]

Let $\underset{\sim}{Y}_i \sim mn_r(n_i, \underset{\sim}{p}_i)$ be a sequence of independent random variables with Z_{gi}, G_k^2, $m_{g(k)}$ and $v_{g(k)}^2$ defined as above, $i = 1,\ldots,k$. Suppose that there exist constants N, δ, δ' such that

(i) $1 \leq n_i \leq N < \infty$;

(ii) $p_{ij} > \delta > 0$;

(iii) if $n_i = 1$ then there exists a j $(j = 1,\ldots,r)$ for which

$$|p_{ij} - r^{-1}| > \delta' > 0.$$

Then in the limit as $k \to \infty$

$$\mathcal{L}[\{G_k^2 - m_{g(k)}\}/v_{g(k)}] \to N(0,1).$$

Proof. For $i = 1,\ldots,k$ let $Q_i = Z_{gi} - m_{gi}$. In order to establish the asymptotic normality of G_k^2, we will invoke the Bounded Liapounov Theorem which requires that the Q_i are bounded uniformly in modulus and that their sum has variance approaching infinity in the limit.

For $i = 1,\ldots,k$, $\frac{1}{2}Z_{gi} = \Sigma_j Y_{ij} \ln Y_{ij} - n_i \ln n_i - \Sigma_j Y_{ij} \ln p_{ij}$. Now $\Sigma_j Y_{ij} \ln Y_{ij} \leq \Sigma_j Y_{ij}^2 \leq \max_q Y_{iq} \Sigma_j Y_{ij} \leq N^2$. Thus

$$\tfrac{1}{2}|Z_{gi}| \leq N^2 + N \ln N + N \ln \delta^{-1} \equiv \frac{1}{2} M, \tag{6.2}$$

say. Then $|Q_i| \leq |Z_{gi}| + m_{gi} \leq 2M$, establishing uniform boundedness.

Now if $n_i = 1$ then by (ii), (iii) and Lemma 6.3(b) there exists for all such i a constant τ such that $\operatorname{var} Q_i = v_{gi}^2 > \tau > 0$. If $n_i \geq 2$ then we must show that v_{gi}^2 has a positive lower bound, even if all the p_{ij} are equal, $j = 1,\ldots,r$. Without loss of generality, assume that $p_{i1} \leq p_{i2} \leq \ldots \leq p_{ir}$. Now let $S_i = \{\underset{\sim}{s} = (s_1,\ldots,s_r)\colon s_j \geq 0$ an integer and $\Sigma s_j = n_i\}$. For $\underset{\sim}{s} \in S_i$, put $t_i(\underset{\sim}{s}) = \Sigma_j s_j \ln\{s_j/(n_i p_{ij})\}$, and $p_i(\underset{\sim}{s}) = n_i!/(s_1!\ldots s_r!)p_{i1}^{s_1}\ldots p_{ir}^{s_r}$. Then

$$EZ_{gi} = 2 \sum_{\underset{\sim}{s} \in S_i} p_i(\underset{\sim}{s})\, t_i(\underset{\sim}{s}).$$

Let $\underset{\sim}{s}_1 = (n_i, \underset{\sim}{0}^T_{r-1})$ and $\underset{\sim}{s}_2 = (n_i-1, 1, \underset{\sim}{0}^T_{r-2})$. Then $t_i(\underset{\sim}{s}_1) = -n_i \ln p_{i1}$ and $t_i(\underset{\sim}{s}_2) = (n_i-1)\ln\{(n_i-1)/(n_i p_{i1})\} - \ln(n_i p_{i2})$, implying that

$$t_i(\underset{\sim}{s}_1) - t_i(\underset{\sim}{s}_2) = \ln[n_i\{n_i/(n_i - 1)\}^{n_i-1} p_{i2}/p_{i1}] > 0.$$

Thus since Z_{gi} takes at least two distinct values with positive probability, there exists, for all i with $n_i \geq 2$, a τ' such that var $Z_{gi} \geq \tau' > 0$. Hence for all i,

$$v^2_{gi} \geq \min(\tau, \tau') \equiv \tau'' > 0, \tag{6.3}$$

implying that $v_{g(k)} \geq (k\tau'')^{\frac{1}{2}} \to \infty$ as $k \to \infty$. The result follows by the Bounded Liapounov Theorem [Loève, 1963, p.277].

Notes.

1. Design zeros ($p_{ij} = 0$ for some j) may be accommodated by deleting the cells with $p_{ij} = 0$ and adjusting r and the expressions for mean and variance of X^2 and G^2, provided that the new $r \geq 2$.

2. Suppose that $n_i = 1$ and the conditions of Theorem 6.1(ii) and Theorem 6.2(iii) do not hold: all cell probabilities are equal. Then $Z_{xi} = r - 1$ and $Z_{gi} = 2 \ln r$, i.e. the contribution to both goodness-of-fit statistics is constant, whichever of the Y_{ij} equals one. If $n_i > 1$ for all except finitely many i, then these conditions are not important.

3. The asymptotic normality of X^2 and G^2 does not imply their asymptotic equivalence. The argument used by Koehler and Larntz (1980) and quoted in §5.2 for explaining the different asymptotic means of X^2 and G^2 for a single sparse multinomial applies here, since the terms involved are similar. In particular, both asymptotic means are functions of the probabilities $\underset{\sim}{p}_i$, and for cell frequencies less than one, the asymptotic mean of X^2 is larger than that of G^2.

The same is likely to be true for the asymptotic variances, although the covariance structure of the product multinomial is more complicated (its variance-covariance matrix is block diagonal) than that of the sequence of independent Poisson random variables used in the centring constants of Theorem 5.1.

6.3 Parametric Case

6.3.1 Preliminaries

We study in this subsection a sequence of k independent multinomial random variables, each dependent on a common unknown $\underset{\sim}{\theta}$ which is estimated by maximum likelihood. Under ten conditions, the distributions of the Pearson X^2 and likelihood ratio G^2 statistics are shown to approach normality as the number of multinomials k becomes large. The statistics have the same asymptotic means as their counterparts when $\underset{\sim}{\theta}$ is known, but different asymptotic variances. For $i = 1,\ldots,k$, let $\underset{\sim i}{Y}^T = (Y_{i1},\ldots,Y_{ir})^T$, $r \geq 2$, be a sequence of random vectors distributed as $mn_r(n_i, \underset{\sim}{p}_i(\underset{\sim}{\theta}))$, where $\underset{\sim}{p}_i(\underset{\sim}{\theta}) = (p_{i1}(\underset{\sim}{\theta}),\ldots,p_{ir}(\underset{\sim}{\theta}))^T$, $\underset{\sim r}{1}^T \underset{\sim}{p}_i(\underset{\sim}{\theta}) = 1$ and $\underset{\sim}{\theta} \in \Theta$, a subset of R^p. The contribution of $\underset{\sim i}{Y}$ to the log likelihood function up to an additive constant independent of $\underset{\sim}{\theta}$ is

$$\ell_i(\underset{\sim}{\theta}) = \sum_{j=1}^{r} y_{ij} \ln p_{ij}(\underset{\sim}{\theta}) ,$$

for $i = 1,\ldots,k$. For notational simplicity we suppress dependence on $\underset{\sim i}{Y}$. The partial derivative of a function with respect to $\underset{\sim}{\theta}$ is denoted by a prime ('). Thus $\underset{\sim i}{\ell}'(\underset{\sim}{\theta}) = \partial \ell_i(\underset{\sim}{\theta})/\partial \underset{\sim}{\theta}$, a p-dimensional column vector whose s-th component is $\partial \ell_i(\underset{\sim}{\theta})/\partial \theta_s$, $s = 1,\ldots,p$. Let

$$\Gamma_i(\underset{\sim}{\theta}) = E[\underset{\sim i}{\ell}'(\underset{\sim}{\theta})\underset{\sim i}{\ell}'^T(\underset{\sim}{\theta}) \,|\, \underset{\sim}{\theta}]$$

and

$$\bar{\Gamma}_k(\underset{\sim}{\theta}) = k^{-1} \sum_{i=1}^{k} \Gamma_i(\underset{\sim}{\theta}).$$

6.3.2 Regularity conditions for the maximum likelihood estimator of $\underset{\sim}{\theta}$

We present conditions on the multinomial probabilities and log likelihood function which ensure the consistency and asymptotic normality of the maximum likelihood estimator $\hat{\underset{\sim}{\theta}}_k$. These are based on Hoadley's (1971) conditions, which apply in a more general setting. Several of the conditions given are stronger than necessary, but are relatively easy to check. They require essentially that as the number of multinomials k approaches infinity, all cell probabilities are differentiable twice with respect to $\underset{\sim}{\theta}$ uniformly in i and bounded uniformly away from zero; the number of observations n_i per multinomial is bounded uniformly in i; the expected log likelihood function is maximized only at the true value $\underset{\sim}{\theta}_o$; the expected contribution to the log-likelihood is on average negative for $\| \underset{\sim}{\theta} \|$ large enough; the order of differentiation and summation is unimportant for the first two derivatives of $\ell_i(\underset{\sim}{\theta})$, i = 1,2,... with respect to $\underset{\sim}{\theta}$, and hence also for the cell probabilities; the asymptotic expected Fisher information matrix for $\hat{\underset{\sim}{\theta}}$ is positive definite; the first and second derivatives of the cell probabilities with respect to $\underset{\sim}{\theta}$ are bounded uniformly in i for $\underset{\sim}{\theta}$ close enough to $\underset{\sim}{\theta}_o$. The final condition ensures that X^2 and G^2 have non-constant contributions from each $\underset{\sim}{Y}_i$ with $n_i = 1$.

Consider the following conditions on the sequence of multinomial distributions defined in §6.3.1:

1. The parameter space Θ is a closed subset of R^p, with $\underset{\sim}{\theta}_o$ an interior point of Θ.

2. The $p_{ij}(\underset{\sim}{\theta})$, i = 1,...,k; j = 1,...,r are continuous and twice differentiable with respect to $\underset{\sim}{\theta}$, uniformly in i.

3. There exist positive constants δ and N such that

   (i) $p_{ij}(\underset{\sim}{\theta}) > \delta > 0$ for all i = 1,...,k; j = 1,...,r;

   (ii) $1 < n_i < N$ for all i = 1,...,k.

4. (i) $\lim_{k\to\infty} k^{-1} \sum_{i=1}^{k} n_i \sum_{j=1}^{r} p_{ij}(\underset{\sim}{\theta}_o)\ln\{p_{ij}(\underset{\sim}{\theta})/p_{ij}(\underset{\sim}{\theta}_o)\} < 0$ for $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}_o$;

(ii) There exists a $\rho > 0$ such that

$$\lim_{k\to\infty} k^{-1} \sum_{i=1}^{k} n_i \sup_{\|\underset{\sim}{\theta}\|>\rho} \sum_{j=1}^{r} p_{ij}(\underset{\sim}{\theta}_o)\ln\{p_{ij}(\underset{\sim}{\theta})/p_{ij}(\underset{\sim}{\theta}_o)\} < 0$$

5. $E[\underset{\sim i}{\ell}'(\underset{\sim}{\theta})\,|\,\underset{\sim}{\theta}] = \underset{\sim}{0}, \quad i = 1,2,\ldots$.

6. $\Gamma_i(\underset{\sim}{\theta}) \equiv E[\underset{\sim i}{\ell}'(\underset{\sim}{\theta})\underset{\sim i}{\ell}'^{T}(\underset{\sim}{\theta})\,|\,\underset{\sim}{\theta}] = -E[\underset{\sim i}{\ell}''(\underset{\sim}{\theta})\,|\,\underset{\sim}{\theta}], \quad i = 1,2,\ldots$.

7. Let $\bar{\Gamma}_k(\underset{\sim}{\theta})$ be as defined in §6.3.1. Then as $k \to \infty$, $\bar{\Gamma}_k(\underset{\sim}{\theta}) \to \Gamma(\underset{\sim}{\theta})$, positive definite.

8. There exist $M > 0$ and $\varepsilon > 0$ such that $|\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s| < M$ for all $i = 1,\ldots,k$; $j = 1,\ldots,r$; $s = 1,\ldots,p$, and for all $\underset{\sim}{\theta}$ such that $\|\underset{\sim}{\theta} - \underset{\sim}{\theta}_o\| < \varepsilon$.

9. There exist $\varepsilon' > 0$ and $M' > 0$ such that $|\partial^2 p_{ij}(\underset{\sim}{\theta})/\partial\theta_q\partial\theta_s| < M'$ for all $\underset{\sim}{\theta} \in \{\underset{\sim}{\theta}: \|\underset{\sim}{\theta}-\underset{\sim}{\theta}_o\| < \varepsilon'\}$ and for all $i = 1,\ldots,k$; $j = 1,\ldots,r$; $q,s = 1,\ldots,p$.

10. There exists a $\delta' > 0$ such that for all i with $n_i = 1$ $|p_{ij}(\underset{\sim}{\theta}) - r^{-1}| > \delta'$ for some $j = 1,\ldots,r$.

Proposition 6.1

Let $\hat{\underset{\sim}{\theta}}_k$ be the maximum likelihood estimator of $\underset{\sim}{\theta}$ based on $\underset{\sim 1}{Y},\ldots,\underset{\sim k}{Y}$, with $\underset{\sim i}{Y} \sim mn_r(n_i,\underset{\sim i}{p}(\underset{\sim}{\theta}_o))$, and conditions 1 to 9 satisfied. Then as $k \to \infty$,

(a) $\hat{\underset{\sim}{\theta}}_k \xrightarrow{p} \underset{\sim}{\theta}_o$;

(b) $\mathcal{L}\{k^{\frac{1}{2}}(\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o)\} \to N(\underset{\sim}{0},\Gamma^{-1}(\underset{\sim}{\theta}_o))$.

Proof. We establish Hoadley's (1971) conditions for consistency (C) and asymptotic normality (N) of $\hat{\underset{\sim}{\theta}}_k$, listed in Appendix II. Those assumptions with a prime (') are stronger versions, which are more easily verified. We use the notation defined in Appendix II.

C1 follows from condition 1: C2 from conditions 2, 3(ii) and the definition of the multinomial probability function.

C3': By condition 3, $R_i(\underset{\sim}{\theta}) \equiv \ell_i(\underset{\sim}{\theta}) - \ell_i(\underset{\sim}{\theta}_o)$ satisfies $|R_i(\underset{\sim}{\theta})| \leq 2Nr \ln\{(1-\delta)/\delta\}$.

Thus all moments of $R_i(\underset{\sim}{\theta},\rho) \equiv \sup\{R_i(t): \|\underset{\sim}{t}-\underset{\sim}{\theta}\| \leq \rho\}$, and $V_i(r) \equiv \sup\{R_i(\underset{\sim}{\theta}): \|\underset{\sim}{\theta}\| > r\}$ are bounded above, with the bound not depending on i. In particular, this is true for the second moment about zero, whence the results.

C4': Let $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}_o$. Then

$$z_k(\underset{\sim}{\theta}) \equiv \sum_{i=1}^{k} n_i \sum_{j=1}^{r} p_{ij}(\underset{\sim}{\theta}_o)\ln\{p_{ij}(\underset{\sim}{\theta})/p_{ij}(\underset{\sim}{\theta}_o)\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{r} E[Y_{ij} \ln\{p_{ij}(\underset{\sim}{\theta})/p_{ij}(\underset{\sim}{\theta}_o)\}|\underset{\sim}{\theta}_o]$$

$$= \sum_{i=1}^{k} E\left[\{\ell_i(\underset{\sim}{\theta}) - \ell_i(\underset{\sim}{\theta}_o)\}|\underset{\sim}{\theta}_o\right].$$

Now by condition 4(i), in the limit as $k \to \infty$

$$k^{-1} z_k(\underset{\sim}{\theta}) < 0,$$

establishing C4'(i).

C4'(ii): By condition 4(ii) there exists a $\rho > 0$ such that the following expression is negative, as $k \to \infty$:

$$k^{-1} \sum_{i=1}^{k} n_i \sup_{\|\underset{\sim}{\theta}\| > \rho} \sum_{j=1}^{r} p_{ij}(\underset{\sim}{\theta}_o) \ln\{p_{ij}(\underset{\sim}{\theta})/p_{ij}(\underset{\sim}{\theta}_o)\}$$

$$= k^{-1} \sum_{i=1}^{k} \sup_{\|\underset{\sim}{\theta}\| > \rho} \sum_{j=1}^{r} E[Y_{ij} | \underset{\sim}{\theta}_o] \ln\{p_{ij}(\underset{\sim}{\theta})/p_{ij}(\underset{\sim}{\theta}_o)\}$$

$$= k^{-1} \sum_{i=1}^{k} E\Big[\sup_{\|\underset{\sim}{\theta}\| > \rho} \sum_{j=1}^{r} Y_{ij} \ln\{p_{ij}(\underset{\sim}{\theta})/p_{ij}(\underset{\sim}{\theta}_o)\} \Big| \underset{\sim}{\theta}_o \Big]$$

since the expectation for each i is a linear combination of a finite number of terms which do not depend on $\underset{\sim}{\theta}$ except through $\ln\{p_{ij}(\underset{\sim}{\theta})\}$.

$$= k^{-1} \sum_{i=1}^{k} E[V_i(\rho) | \underset{\sim}{\theta}_o]$$

$$= \bar{v}_k(\rho), \quad \text{whence C4'(ii).}$$

C5: By definition, $R_i(\underset{\sim}{\theta},\rho)$ and $V_i(\rho)$ are suprema of the function $R_i(\underset{\sim}{\theta})$. Now $R_i(\underset{\sim}{\theta})$ is a differentiable function of $\underset{\sim}{Y}_i$, and is of finite range by condition 3. Thus condition C5 is satisfied.

N1 follows from condition 1 if we restrict ourselves to true values of $\underset{\sim}{\theta}$.

N2 is true by Hoadley [1971, Theorem 1], since conditions C1 to C5 hold.

N3 and N4 are established by condition 2 and the definition of the multinomial log likelihood.

N5, N6 and N7 are restatements of conditions 5, 6 and 7 for the product multinomial.

$$N8': \left|\partial \ell_i(\underset{\sim}{\theta}_o)/\partial\theta_q\right| = \left|\sum_{j=1}^{r} Y_{ij} p_{ij}^{-1}(\underset{\sim}{\theta}_o)\partial p_{ij}(\underset{\sim}{\theta}_o)/\partial\theta_q\right|$$

$$\leq rN\delta^{-1}M$$

by conditions 3 and 8. Thus N8' is satisfied for $K = (rNM/\delta)^3$.

$$N9: \quad \left|\partial^2 \ell_i(\underset{\sim}{\theta})/\partial\theta_q\partial\theta_s\right| = \left|\sum_{j=1}^{r} Y_{ij}[p_{ij}^{-1}(\underset{\sim}{\theta})\partial^2 p_{ij}(\underset{\sim}{\theta})/\partial\theta_q\partial\theta_s - p_{ij}^{-2}(\underset{\sim}{\theta})\{\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_q\}\{\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s\}]\right|$$

$$\leq rN(\delta^{-1}M' + \delta^{-2}M^2)$$

when $\|\underset{\sim}{\theta}-\underset{\sim}{\theta}_o\| < \min(\varepsilon,\varepsilon')$, by conditions 3, 8 and 9. This is a stronger result than N9 since the bound on the elements of ℓ_i'' is uniform. Thus N9 follows easily.

Hence by conditions C1, C2, C3' and C4', (a) follows by Hoadley's Theorem 1 (1971). Similarly (b) follows from N1 to N7, N8' and N9 by his Theorem 2.

6.3.3 Asymptotic distribution of X^2 and G^2

In this subsection we establish the asymptotic normality of the goodness-of-fit statistics X^2 and G^2 in the parametric case, when $\underset{\sim}{\theta}$ is estimated by maximum likelihood. As in §6.2, the number of independent multinomials k in the product increases to infinity, but the number of observations n_i per multinomial is kept bounded. Conditions 1 to 9 of §6.3.2 are sufficient for the consistency and asymptotic normality of $\hat{\underset{\sim}{\theta}}_k$: condition 10 prevents constant contributions to the statistics by constraining the probabilities of a contribution to be unequal when its sample size $n_i = 1$. The Pearson X^2 and likelihood ratio G^2 statistics

$X_k^2(\underset{\sim}{\theta})$ and $G_k^2(\underset{\sim}{\theta})$ for testing the fit of the random variables $\underset{\sim}{Y}_i$ to the sequence of independent multinomials $mn_r(n_i, \underset{\sim}{p}_i(\underset{\sim}{\theta}))$ $i = 1,\ldots,k$, are defined as in §6.2.2, with the $\underset{\sim}{p}_i$ replaced by $\underset{\sim}{p}_i(\underset{\sim}{\theta})$ and expectations taken for fixed $\underset{\sim}{\theta}$. Before proving the theorems we require some definitions and preliminary results.

Definition: O_p, o_p notation for stochastic order of sequences.
Let $\{X_k\}$ be a sequence of random variables and $\{a_k\}$ a sequence of constants. Then $X_k = O_p(a_k)$ if there exists for every $\eta > 0$ a constant $M(\eta) \geq 0$ and an integer $K(\eta)$ such that for all $k \geq K(\eta)$

$$\Pr[|X_k| \leq a_k M(\eta)] \geq 1 - \eta.$$

$X_k = o_p(a_k)$ if for all $\varepsilon > 0$ and $\eta > 0$ there exists an integer $K(\varepsilon,\eta)$ such that for all $k \geq K(\varepsilon,\eta)$

$$\Pr[|X_k| < a_k \varepsilon] \geq 1 - \eta.$$

This notation was introduced by Mann and Wald (1943), and is discussed by Bishop, Fienberg and Holland (1975, §14.4). Usually $a_k = k^\alpha$, α real.

Lemma 6.4

Let X_i, $i = 1,2,\ldots$ be a sequence of independent random variables, bounded uniformly in modulus, and with mean μ_i. Let $\bar{X}_k = k^{-1}(X_1+\ldots+X_k)$ with an analogous expression for $\bar{\mu}_k$. Then in the limit as $k \to \infty$

(a) $\bar{X}_k - \bar{\mu}_k = O_p(k^{-\frac{1}{2}})$;

(b) $\bar{X}_k - \bar{\mu}_k = o_p(1)$.

Proof. Since the X_i are uniformly bounded, there exists for all i an $M > 0$ such that $\text{var } X_i = \sigma_i^2 \leq M^2$. Thus $\text{var } \bar{X}_k = \bar{\sigma}_k^2 \leq k^{-1}M^2$. Then by Chebyshev's Inequality [Rao, 1973, 2b.2.6],

$$\Pr[|\bar{X}_k - \bar{\mu}_k| < \eta^{-\frac{1}{2}} M k^{-\frac{1}{2}}] \geq \Pr[|\bar{X}_k - \bar{\mu}_k| < \bar{\sigma}_k \eta^{-\frac{1}{2}}] > 1-\eta$$

for all $\eta > 0$, whence (a), with $M(\eta) = \eta^{-\frac{1}{2}} M$, and (b) with $\varepsilon = \eta^{-\frac{1}{2}} M k^{-\frac{1}{2}}$.

Lemma 6.5

Let $\{f_k(\underset{\sim}{\theta})\}$ be a sequence of real-valued functions continuous in $\underset{\sim}{\theta}$ uniformly in k, and suppose that the following hold:

(i) $\lim_{k\to\infty} f_k(\underset{\sim}{\theta}_o) = f_o$;

(ii) there is a sequence of statistics $\underset{\sim}{T}_k$ with $\underset{\sim}{T}_k \xrightarrow{p} \underset{\sim}{\theta}_o$.

Then $f_k(\underset{\sim}{T}_k) \xrightarrow{p} f_o$.

Proof. By the continuity of the f_k uniformly in k, there exists $M > 0$ such that for all k

$$|f_k(\underset{\sim}{\theta}) - f_k(\underset{\sim}{\theta}_o)| \leq M||\underset{\sim}{\theta} - \underset{\sim}{\theta}_o|| \qquad \text{as} \quad \underset{\sim}{\theta} \to \underset{\sim}{\theta}_o. \tag{6.4}$$

Now

$$|f_k(\underset{\sim}{T}_k) - f_o| \leq |f_k(\underset{\sim}{T}_k) - f_k(\underset{\sim}{\theta}_o)| + |f_k(\underset{\sim}{\theta}_o) - f_o| \quad .$$

The first term on the right $\xrightarrow{p} 0$ by (6.4) and (ii); the second term $\to 0$ by (i), establishing the required result. [See also Billingsley, 1968, Theorem 5.5].

Lemma 6.6

Consider two sequences X_k and Y_k of random variables such that

(i) $X_k \xrightarrow{\mathcal{D}} X$;

(ii) $X_k - Y_k = o_p(1)$

as $k \to \infty$, with X a real-valued random variable. Then

$$Y_k \xrightarrow{\mathcal{D}} X.$$

Proof. See Billingsley [1968, Theorem 4.1] for a proof of a more general result.

Theorem 6.3 [Asymptotic normality of $X_k^2(\hat{\underset{\sim}{\theta}}_k)$].

Consider the sequence of multinomial random variables $\underset{\sim}{Y}_i$, $i = 1,2,\ldots$ defined in §6.3.1 and suppose that conditions 1 to 10 hold. Let $\hat{\underset{\sim}{\theta}}_k$ be the maximum likelihood estimator of $\underset{\sim}{\theta}$;

$$\underset{\sim}{\mu}_k(\underset{\sim}{\theta}) = k^{-1}E[\underset{\sim}{X}_k^{2'}(\underset{\sim}{\theta})|\underset{\sim}{\theta}];$$

$$a_{ij}(\underset{\sim}{\theta}) = \{n_i p_{ij}(\underset{\sim}{\theta})\}^{-1}, \quad j = 1,\ldots,r; \quad \underset{\sim}{a}_i(\underset{\sim}{\theta}) = (a_{i1}(\underset{\sim}{\theta}),\ldots,a_{ir}(\underset{\sim}{\theta}))^T;$$

$D_i(\underset{\sim}{\theta})$ the $p \times r$ matrix with (s,j)-th entry

$$p_{ij}^{-1}(\underset{\sim}{\theta})\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s, \quad s = 1,\ldots,p; \quad j = 1,\ldots,r;$$

$$\underset{\sim}{b}_i^T(\underset{\sim}{\theta}) = \underset{\sim}{\mu}_k^T(\underset{\sim}{\theta})\Gamma^{-1}(\underset{\sim}{\theta})D_i(\underset{\sim}{\theta});$$

$\underset{\sim}{Y}_i^2 = (Y_{i1}^2,\ldots,Y_{ir}^2)^T$, and $V_i(\underset{\sim}{\theta})$ the variance-covariance matrix of the $2r$-vector $(\underset{\sim}{Y}_i^{2T}, \underset{\sim}{Y}_i^T)^T$, $i = 1,\ldots k$; $k = 1,\ldots$. Then under H_o: $\underset{\sim}{\theta} = \underset{\sim}{\theta}_o$, and as $k \to \infty$,

$$\mathcal{L}\left[\frac{X_k^2(\hat{\underset{\sim}{\theta}}_k) - k(r-1)}{\{\sum_{i=1}^{k}(\underset{\sim}{a}_i^T(\hat{\underset{\sim}{\theta}}_k),\underset{\sim}{b}_i^T(\hat{\underset{\sim}{\theta}}_k))V_i(\hat{\underset{\sim}{\theta}}_k)(\underset{\sim}{a}_i^T(\hat{\underset{\sim}{\theta}}_k),\underset{\sim}{b}_i^T(\hat{\underset{\sim}{\theta}}_k))^T\}^{\frac{1}{2}}}\right] \to N(0,1).$$

Proof. For $i = 1,\ldots,k$, let

$$W_i(\underset{\sim}{\theta}) = \underset{\sim}{a}_i^T(\underset{\sim}{\theta})\underset{\sim}{Y}_i^2 + \underset{\sim}{b}_i^T(\underset{\sim}{\theta})\underset{\sim}{Y}_i - n_i,$$

with $v_i^2(\underset{\sim}{\theta}) = \mathrm{var}[W_i(\underset{\sim}{\theta})|\underset{\sim}{\theta}]$, and adopt the summation convention over subscripts in parentheses. We prove that

$$\{X_k{}^2(\hat{\underset{\sim}{\theta}}_k) - W_{(k)}(\underset{\sim}{\theta}_o)\}/v_{(k)}(\underset{\sim}{\theta}_o) = o_p(1) \tag{6.5}$$

as $k \to \infty$, and $EX_k^2(\underset{\sim}{\theta}_o) = EW_{(k)}(\underset{\sim}{\theta}_o)$. Then by Lemma 6.6 for a sequence of

pairs of random variables whose difference converges in probability to zero, and by the Central Limit Theorem, we establish

$$\mathcal{L}[\{X_k^2(\hat{\underset{\sim}{\theta}}_k) - k(r-1)\}/v_{(k)}(\underset{\sim}{\theta}_o)] \to N(0,1) \tag{6.6}$$

as $k \to \infty$. By appealing to Slutsky's Theorem it is shown that $v_{(k)}(\underset{\sim}{\theta}_o)$ may be replaced by $v_{(k)}(\hat{\underset{\sim}{\theta}}_k)$, whence the required result.

For $i = 1,\ldots,k$ let

$$Z_i(\underset{\sim}{\theta}) = \sum_{j=1}^{r} \{Y_{ij} - n_i p_{ij}(\underset{\sim}{\theta})\}^2/\{n_i p_{ij}(\underset{\sim}{\theta})\}.$$

$$= \sum_{j=1}^{r} Y_{ij}^2/\{n_i p_{ij}(\underset{\sim}{\theta})\} - n_i .$$

Then, for $s = 1,\ldots,p$ and $i = 1,\ldots,k$

$$\partial Z_i(\underset{\sim}{\theta})/\partial\theta_s = -n_i^{-1} \sum_{j=1}^{r} Y_{ij}^2\, p_{ij}^{-2}(\underset{\sim}{\theta})\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s . \tag{6.7}$$

Thus and by conditions 2 and 3 the $Z_i(\underset{\sim}{\theta})$ are a sequence of functions continuous and differentiable with respect to $\underset{\sim}{\theta}$, uniformly in i. Hence $k^{-1}X_k^2(\underset{\sim}{\theta}) = k^{-1}Z_{(k)}(\underset{\sim}{\theta})$ is continuous and differentiable w.r.t. $\underset{\sim}{\theta}$, uniformly in k. Therefore it may be expanded in Young's form of a multivariate Taylor series expansion about $\underset{\sim}{\theta}_o$ [cf. Serfling, 1980, Theorem 1.12.1C] with $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}_k$:

$$k^{-1}X_k^2(\hat{\underset{\sim}{\theta}}_k) = k^{-1}X_k^2(\underset{\sim}{\theta}_o) + k^{-1}(\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o)^T X_k^{2\prime}(\underset{\sim}{\theta}_o) + o(\| \hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o \|). \tag{6.8}$$

Now by (6.7) and conditions 3 and 8

$$|\partial Z_i(\underset{\sim}{\theta})/\partial\theta_s| < rN^2\delta^{-2}M, \tag{6.9}$$

for $i = 1,\ldots,k$; $s = 1,\ldots,p$ and $\underset{\sim}{\theta}$ close enough to $\underset{\sim}{\theta}_o$. Thus by Lemma 6.4(a)

$$k^{-1}\underset{\sim}{X}^2_k{}'(\underset{\sim}{\theta}_o) = \underset{\sim}{\mu}_k(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}}). \tag{6.10}$$

We next show that

$$\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o = k^{-1}\Gamma^{-1}(\underset{\sim}{\theta}_o)\underset{\sim}{\ell}'_{(k)}(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}}). \tag{6.11}$$

By condition 2 and the definition of the log likelihood function $\ell_i(\underset{\sim}{\theta})$, $k^{-1}\ell_{(k)}(\underset{\sim}{\theta})$ is continuous and twice differentiable w.r.t. $\underset{\sim}{\theta}$. Further, by conditions 3 and 8

$$|\partial\ell_i(\underset{\sim}{\theta})/\partial\theta_s| = |\sum_{j=1}^{r} Y_{ij}p_{ij}^{-1}(\underset{\sim}{\theta})\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s)| \leq rN\delta^{-1}M \tag{6.12}$$

for $\underset{\sim}{\theta}$ close enough to $\underset{\sim}{\theta}_o$, $i = 1,\ldots,k$; $s = 1,\ldots,p$. Hence $k^{-1}\underset{\sim}{\ell}'_{(k)}(\underset{\sim}{\theta})$ is continuous and differentiable w.r.t. $\underset{\sim}{\theta}$ uniformly in k, and on substituting $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}_k$ in Young's form of a multivariate Taylor series expansion about $\underset{\sim}{\theta}_o$,

$$k^{-1}\underset{\sim}{\ell}'_{(k)}(\hat{\underset{\sim}{\theta}}_k) = k^{-1}\underset{\sim}{\ell}'_{(k)}(\underset{\sim}{\theta}_o) + k^{-1}\ell''^{T}_{(k)}(\underset{\sim}{\theta}_o)(\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o) + o(\|\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o\|). \tag{6.13}$$

For $i = 1,\ldots,k$; $j = 1,\ldots,r$; $q,s = 1,\ldots,p$ let

$$\lambda_{ij}^{(qs)} = p_{ij}^{-1}(\underset{\sim}{\theta})\partial^2 p_{ij}(\underset{\sim}{\theta})/(\partial\theta_q\partial\theta_s) - p_{ij}^{-2}(\underset{\sim}{\theta})\{\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_q\}\{\partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s\}|\underset{\sim}{\theta} = \underset{\sim}{\theta}_o.$$

Then it can be shown by conditions 3, 8 and 9 for $\underset{\sim}{\theta} = \underset{\sim}{\theta}_o$ that

$$|\partial^2\ell_i(\underset{\sim}{\theta})/(\partial\theta_q\partial\theta_s)| = |\sum_{j=1}^{r} Y_{ij}\lambda_{ij}^{(qs)}| \leq rN(\delta^{-1}M' + \delta^{-2}M^2). \tag{6.14}$$

Thus by Lemma 6.4(a) and conditions 6 and 7

$$k^{-1}\ell''_{(k)}(\underset{\sim}{\theta}_o) = -\Gamma(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}}) . \tag{6.15}$$

Now by condition 5, (6.12) and Lemma 6.4(b),

$$k^{-1}\underset{\sim}{\ell}'_{(k)}(\underset{\sim}{\theta}_o) = o_p(1) . \tag{6.16}$$

By Proposition 6.1(b) and Chebyshev's Inequality [Rao, 1973, 2b.2.6],

$$\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o = o_p(k^{-\frac{1}{2}}) , \tag{6.17}$$

whence

$$o(\| \hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o \|) = o_p(k^{-\frac{1}{2}}). \tag{6.18}$$

The left side of (6.13) equals zero by definition of the maximum likelihood estimator $\hat{\underset{\sim}{\theta}}_k$. Thus, applying condition 7, (6.15), and (6.18) to (6.13) yields

$$\underset{\sim}{0} = k^{-1}\underset{\sim}{\ell}'_{(k)}(\underset{\sim}{\theta}_o) + \{-\Gamma(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}})\}^T(\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}}). \tag{6.19}$$

Observe that for large enough k, $o_p(k^{-\frac{1}{2}}) < 1$ with probability $1-\varepsilon$, for $\varepsilon > 0$ as small as desired. Thus, on pre-multiplying (6.19) by $\{-\Gamma(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}})\}^{-1}$ and by (6.16), (6.11) is obtained.

We now proceed to show (6.5). Substituting (6.11), (6.10) and (6.18) into (6.8), and applying (6.16) we obtain

$$\begin{aligned} k^{-1}X_k^2(\hat{\underset{\sim}{\theta}}_k) &= k^{-1}X_k^2(\underset{\sim}{\theta}_o) + k^{-1}\underset{\sim}{\ell}'^T_{(k)}(\underset{\sim}{\theta}_o)\Gamma^{-1}(\underset{\sim}{\theta}_o)\underset{\sim}{\mu}_k(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}}) \\ &= k^{-1}W_{(k)}(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}}) \end{aligned} \tag{6.20}$$

by definition of $\ell'_{(k)}(\underset{\sim}{\theta}_o)$. Since by condition 5 $Ek^{-1}\ell'_{(k)}(\underset{\sim}{\theta}_o) = \underset{\sim}{0}$, it follows that

$$EX_k^2(\underset{\sim}{\theta}_o) = EW_{(k)}(\underset{\sim}{\theta}_o) = k(r-1). \tag{6.21}$$

We next establish that the $v_i^2(\underset{\sim}{\theta}) = \text{var } W_i(\underset{\sim}{\theta})$, $i = 1,\ldots,$ have a positive lower bound, uniformly in i. Now by conditions 3 and 10 and by Lemma 6.2(b),

$$\text{var } Z_i(\underset{\sim}{\theta}) \geq \begin{cases} (\delta')^{-1} & \text{if } n_i = 1 \\ r-1 & \text{if } n_i > 1. \end{cases}$$

Further, $W_i(\underset{\sim}{\theta}) = \underset{\sim i}{a}^T(\theta)\underset{\sim i}{Y}_i^2 + \underset{\sim i}{b}^T(\theta)\underset{\sim i}{Y}_i - n_i$ is not a constant, since by condition 3 the coefficients of the Y_{ij}^2 are non-zero. Thus there exists for all i and for all $\underset{\sim}{\theta} \in \Theta$ a $\tau > 0$ such that $v_i^2(\underset{\sim}{\theta}) > \tau > 0$. This implies that

$$v_{(k)}(\underset{\sim}{\theta}) = \{\sum_{i=1}^{k} v_i^2(\underset{\sim}{\theta})\}^{\frac{1}{2}} > (\tau k)^{\frac{1}{2}} \to \infty \qquad \text{as } k \to \infty. \tag{6.22}$$

To establish the boundedness of the terms in $W_{(k)}(\underset{\sim}{\theta})$ uniformly in k, observe that by conditions 3 and 8 and by (6.9)

$$|\underset{\sim i}{a}^T(\underset{\sim}{\theta})\underset{\sim i}{Y}_i^2 + \underset{\sim i}{b}^T(\underset{\sim}{\theta})\underset{\sim i}{Y}_i - n_i| \leq r\delta^{-1}N^2 + r^2N^3\delta^{-3}M^2\underset{\sim p}{1}^T\Gamma^{-1}(\underset{\sim o}{\theta})\underset{\sim p}{1} + N$$

for $\underset{\sim}{\theta}$ close enough to $\underset{\sim o}{\theta}$. Thus, $W_{(k)}(\underset{\sim o}{\theta}) - r(k-1)$ is a sum of independent, uniformly bounded random variables with zero mean, by (6.21). Its standard deviation $v_{(k)}(\underset{\sim o}{\theta}) \to \infty$ as $k \to \infty$ by (6.22). Therefore by the Bounded Liapounov Central Limit Theorem [Loève, 1963, p.277],

$$\mathcal{L}[\{W_{(k)}(\underset{\sim o}{\theta}) - k(r-1)\}/v_{(k)}(\underset{\sim o}{\theta})] \to N(0,1) \quad \text{as } k \to \infty. \tag{6.23}$$

Then, on dividing (6.20) by $k^{-\frac{1}{2}}v_{(k)}(\underset{\sim}{\theta}_o)$, which by (6.22) is a function bounded uniformly away from zero, it follows that

$$\{X_k^2(\hat{\underset{\sim}{\theta}}_k) - k(r-1)\}/v_{(k)}(\underset{\sim}{\theta}_o) - \{W_{(k)} - k(r-1)\}/v_{(k)}(\underset{\sim}{\theta}_o) = o_p(1),$$

establishing (6.5). Hence by (6.23) and by Lemma 6.6 [cf. Billingsley, 1968, Theorem 4.1], (6.6) is established.

Let $q_k(\underset{\sim}{\theta}) = v_{(k)}(\underset{\sim}{\theta}_o)/v_{(k)}(\underset{\sim}{\theta})$. We must show that this is a sequence continuous in $\underset{\sim}{\theta}$ uniformly in k. Now by conditions 2 and 3 the $k^{-1}W_{(k)}(\underset{\sim}{\theta})$ are continuous in $\underset{\sim}{\theta}$ uniformly in k. Thus $k^{-\frac{1}{2}}v_{(k)}(\underset{\sim}{\theta})$ has the same property. Further, by (6.22) the sequence $k^{-\frac{1}{2}}v_{(k)}(\underset{\sim}{\theta})$ is bounded uniformly in k away from 0. It follows that $q_k(\underset{\sim}{\theta})$ is continuous in $\underset{\sim}{\theta}$ uniformly in k, and $q_k(\underset{\sim}{\theta}_o) = 1$ for all k. Thus by Lemma 6.5 and Proposition 6.1(a), $q_k(\hat{\underset{\sim}{\theta}}_k) \xrightarrow{P} 1$.

The desired result is obtained by applying Slutsky's Theorem [Serfling, 1980, Theorem 1.5.4(ii)] and (6.6) to

$$q_k(\hat{\underset{\sim}{\theta}}_k)\{X_k^2(\hat{\underset{\sim}{\theta}}_k) - r(k-1)\}/v_{(k)}(\underset{\sim}{\theta}_o).$$

Note. The variance-covariance matrices $V_i(\underset{\sim}{\theta})$ are not of full rank because of the constraint $\Sigma_j Y_{ij} = n_i$, which implies that for any random variable X and fixed j

$$\operatorname{cov}(Y_{ij},X) = -\sum_{q\neq j} \operatorname{cov}(Y_{iq},X),$$

i.e. the rank is at most 2r - 1. If r = 2 then

$$\operatorname{cov}(Y_{i2}^2,X) = \operatorname{cov}(Y_{i1}^2,X) - 2n_i\operatorname{cov}(Y_{i1},X),$$

reducing $\operatorname{rank}(V_i(\underset{\sim}{\theta}))$ to 2r - 2 = 2. However, more importantly, the $v_i(\underset{\sim}{\theta})$ have been shown to be uniformly bounded away from zero.

Theorem 6.4 [Asymptotic normality of $G_k^2(\hat{\underset{\sim}{\theta}}_k)$].

Let $\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, \ldots$ be a sequence of independent random variables with $\underset{\sim}{Y}_i \sim mn_r(n_i, \underset{\sim}{p}_i(\theta))$, as in §6.3.1. Suppose that conditions 1 to 10 of §6.3.2 hold, and let

$$w_i^2(\underset{\sim}{\theta}) = \mathrm{var}[2 \sum_{j=1}^{r} Y_{ij} \ln[Y_{ij}/\{n_i p_{ij}(\underset{\sim}{\theta})\}] \,|\, \underset{\sim}{\theta}] ,$$

$i = 1,2,\ldots$. Then under H_o: $\underset{\sim}{\theta} = \underset{\sim}{\theta}_o$ and as $k \to \infty$

$$\mathcal{L}\{G_k^2(\hat{\underset{\sim}{\theta}}_k) - E[G_k^2(\underset{\sim}{\theta}) \,|\, \underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}_k]\}/w_{(k)}(\hat{\underset{\sim}{\theta}}_k)] \to N(0,1).$$

Proof. By arguments concerning continuity and differentiability w.r.t. $\underset{\sim}{\theta}$ of $G_k^2(\underset{\sim}{\theta})$ uniformly in k, we show that

$$\{G_k^2(\hat{\underset{\sim}{\theta}}_k) - G_k^2(\underset{\sim}{\theta}_o)\}/w_{(k)}(\underset{\sim}{\theta}_o) = o_p(1) \qquad (6.24)$$

as $k \to \infty$. Then by Theorem 6.2, and Lemma 6.6 for a sequence of differences of random variables which tend in probability to zero as $k \to \infty$, it follows that

$$\mathcal{L}[\{G_k^2(\hat{\underset{\sim}{\theta}}_k) - E[G_k^2(\underset{\sim}{\theta}) \,|\, \underset{\sim}{\theta} = \underset{\sim}{\theta}_o]\}/w_{(k)}(\underset{\sim}{\theta}_o)] \to N(0,1) \qquad (6.25)$$

as $k \to \infty$. Finally, by Slutsky's theorem we establish that $\underset{\sim}{\theta}_o$ may be replaced by $\hat{\underset{\sim}{\theta}}_k$ in (6.25), whence the desired result.

For $i = 1,2,\ldots$ let

$$Q_i(\underset{\sim}{\theta}) = 2 \sum_{j=1}^{r} Y_{ij} \ln[Y_{ij}/\{n_i p_{ij}(\underset{\sim}{\theta})\}] \quad .$$

Then, adopting the summation notation for subscripts in parentheses, $Q_{(k)}(\underset{\sim}{\theta}) = G_k^2(\underset{\sim}{\theta})$. For $s = 1,\ldots,p$

$$\partial Q_i(\underset{\sim}{\theta})/\partial\theta_s = -2 \sum_{j=1}^{r} Y_{ij} p_{ij}^{-1}(\underset{\sim}{\theta}) \partial p_{ij}(\underset{\sim}{\theta})/\partial\theta_s$$

$$= -2\partial\ell_i(\underset{\sim}{\theta})/\partial\theta_s . \qquad (6.26)$$

Thus and by (6.12), $k^{-1}G_k^{2\prime}(\underset{\sim}{\theta})$ is bounded in modulus uniformly in k for $\underset{\sim}{\theta}$ close enough to $\underset{\sim}{\theta}_o$. Further, by conditions 2 and 3, $k^{-1}G_k^2(\underset{\sim}{\theta})$ is continuous and differentiable w.r.t. $\underset{\sim}{\theta}$ uniformly in k. Therefore, putting $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}_k$ in Young's form of the Taylor series expansion about $\underset{\sim}{\theta}_o$ [cf. Serfling, 1980, Theorem 1.12.1C],

$$k^{-1}G_k^2(\hat{\underset{\sim}{\theta}}_k) = k^{-1}G_k^2(\underset{\sim}{\theta}_o) + k^{-1}(\hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o)^T G_k^{2\prime}(\underset{\sim}{\theta}_o) + o(\| \hat{\underset{\sim}{\theta}}_k - \underset{\sim}{\theta}_o \|) \quad . \tag{6.27}$$

Then by (6.26), (6.17), (6.16) and (6.18) it follows that

$$k^{-1}G_k^2(\hat{\underset{\sim}{\theta}}_k) = k^{-1}G_k^2(\underset{\sim}{\theta}_o) + o_p(k^{-\frac{1}{2}}). \tag{6.28}$$

By (6.3) there exists a positive lower bound on the $w_i^2(\underset{\sim}{\theta})$ which is uniform in i since when $\underset{\sim}{\theta}$ is fixed $G_k^2(\underset{\sim}{\theta})$ satisfies the conditions of Theorem 6.2, namely conditions 3 and 10. Thus there exists $\tau > 0$ such that for all k

$$k^{-\frac{1}{2}}w_{(k)}(\underset{\sim}{\theta}) > \tau > 0. \tag{6.29}$$

On dividing (6.28) by (6.29), (6.24) is established. Further,

$$\{G_k^2(\hat{\underset{\sim}{\theta}}_k) - E[G_k^2(\underset{\sim}{\theta})|\underset{\sim}{\theta} = \underset{\sim}{\theta}_o]\}/w_{(k)}(\underset{\sim}{\theta}_o) - \{G_k^2(\underset{\sim}{\theta}_o) - E[G_k^2(\underset{\sim}{\theta})|\underset{\sim}{\theta} = \underset{\sim}{\theta}_o]\}/w_{(k)}(\underset{\sim}{\theta}_o) = o_p(1). \tag{6.30}$$

Therefore by Theorem 6.2 applied to $G_k^2(\underset{\sim}{\theta}_o)$ and by Lemma 6.6 [cf. Billingsley, 1968, Theorem 4.1], it follows that (6.25) holds.

For fixed $\underset{\sim}{\theta}_o$, let $a_k(\underset{\sim}{\theta}) = w_{(k)}(\underset{\sim}{\theta}_o)/w_{(k)}(\underset{\sim}{\theta})$, $g_k(\underset{\sim}{\theta}) = E[G_k^2(\underset{\sim}{\theta}_1)|\underset{\sim}{\theta}_1 = \underset{\sim}{\theta}]$ and $b_k(\underset{\sim}{\theta}) = \{g_k(\underset{\sim}{\theta}_o) - g_k(\underset{\sim}{\theta})\}/w_{(k)}(\underset{\sim}{\theta}_o)$, $k = 1,2,\ldots$. Then

$$a_k(\underset{\sim}{\theta}_o) = 1 \quad \text{and} \quad b_k(\underset{\sim}{\theta}_o) = 0 \text{ for all } k. \tag{6.31}$$

Observe that

$$\{G_k^2(\hat{\underset{\sim}{\theta}}_k) - g_k(\hat{\underset{\sim}{\theta}}_k)\}/w_{(k)}(\hat{\underset{\sim}{\theta}}_k) = a_k(\hat{\underset{\sim}{\theta}}_k)[\{G_k^2(\hat{\underset{\sim}{\theta}}_k)-g_k(\underset{\sim}{\theta}_o)\}/w_{(k)}(\underset{\sim}{\theta}_o)+b_k(\hat{\underset{\sim}{\theta}}_k)]. \tag{6.32}$$

We next show convergence in probability of $a_k(\hat{\underset{\sim}{\theta}}_k)$ and $b_k(\hat{\underset{\sim}{\theta}}_k)$. Now $a_k(\underset{\sim}{\theta}) = k^{-\frac{1}{2}}w_{(k)}(\underset{\sim}{\theta}_o)/\{k^{-\frac{1}{2}}w_{(k)}(\underset{\sim}{\theta})\}$. The denominator is bounded below uniformly in k by (6.29): the numerator is bounded above uniformly in k since by (6.2) the $Q_i(\underset{\sim}{\theta})$ are bounded in modulus uniformly in k for $\underset{\sim}{\theta}$ fixed, with the bound not depending on $\underset{\sim}{\theta}$, and the same property holds for central moments. By conditions 2 and 3 and by definition, the $Q_i(\underset{\sim}{\theta})$ and hence the $\{k^{-\frac{1}{2}}w_{(k)}(\underset{\sim}{\theta})\}$ are continuous w.r.t. $\underset{\sim}{\theta}$ uniformly in i. Thus the $a_k(\underset{\sim}{\theta})$ are continuous in $\underset{\sim}{\theta}$ uniformly in k. Therefore by (6.31), Proposition 6.1(a) and Lemma 6.5, $a_k(\hat{\underset{\sim}{\theta}}_k) \xrightarrow{p} 1$ as $k \to \infty$.

Now $k^{-1}G_k^2(\underset{\sim}{\theta})$ has already been shown to be continuous and differentiable w.r.t. $\underset{\sim}{\theta}$, uniformly in k. So it follows that its expectation given $\underset{\sim}{\theta}$, $k^{-1}g_k(\underset{\sim}{\theta})$, has the same properties. Thus by Young's form of the multivariate Taylor series expansion about $\underset{\sim}{\theta} = \underset{\sim}{\theta}_o$, with $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}_k$,

$$k^{-1}\{g_k(\hat{\underset{\sim}{\theta}}_k) - g_k(\underset{\sim}{\theta}_o)\} = k^{-1}(\hat{\underset{\sim}{\theta}}_k-\underset{\sim}{\theta}_o)^T g_k'(\underset{\sim}{\theta}_o) + o(\| \hat{\underset{\sim}{\theta}}_k-\underset{\sim}{\theta}_o \|)$$

$$= o_p(k^{-\frac{1}{2}}) \tag{6.33}$$

by condition 5 and (6.18). Dividing (6.33) by (6.29) at $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}_k$ yields $b_k(\hat{\underset{\sim}{\theta}}_k) = o_p(1)$ as $k \to \infty$. By (6.25) and two applications of Slutsky's Theorem [Serfling, 1980, 1.5.4(i)(ii)] to the right side of (6.32), the desired result is obtained.

6.3.4 Example: Binary logistic regression

Consider a sequence Y_t, $t = 1,\ldots,T$, of independent random variables, with $Y_t \sim \text{bin}(n_t, p_t(\underset{\sim}{\theta}))$, with $n_t \leq N$, $\underset{\sim}{\theta} = (\alpha,\beta)$ and

$$p_t(\underset{\sim}{\theta}) = \exp(\alpha - \beta t/T)/\{1 + \exp(\alpha - \beta t/T)\}$$

where $(\alpha,\beta) \in \Theta = \{(\alpha,\beta): -A \leq \alpha \leq -a,\ 0 \leq \beta \leq B\}$, where A,B are large positive numbers and $a > 0$ is small. The true value of $\underset{\sim}{\theta}$ is denoted by $\underset{\sim}{\theta}_o = (\alpha_o,\beta_o)$, an interior point of Θ. The likelihood function for the sequence $Y_1,\ldots,Y_T$ is proportional to

$$\prod_{t=1}^{T} \{\exp(\alpha - \beta t/T)\}^{y_t} \{1 + \exp(\alpha - \beta t/T)\}^{-n_t} .$$

The parameterization is important: the coefficient of β must not become large in modulus as T approaches infinity, or the property of differentiability in $\underset{\sim}{\theta}$ uniformly in t of the $\{p_t(\underset{\sim}{\theta})\}$, required by condition 2, would be violated. By condition 4 (ii), for $(\alpha,\beta) \neq (\alpha_o,\beta_o)$, and $\|(\alpha,\beta)\|$ large enough, $\alpha - (t/T)\beta$ must be bounded away from $\alpha_o - (t/T)\beta_o$, for a proportion of the t, hence the definition of Θ. Other definitions are possible, e.g. $\Theta = \{(\alpha,\beta): a \leq \alpha \leq A,\ -B \leq \beta \leq 0\}$. The specification of $a > 0$ is needed if $n_t = 1$: if $n_t > 1$ for all t, then $a = 0$ suffices.

We now verify the conditions given in §6.3.2 under which the X^2 and G^2 statistics of §6.3.3 are asymptotically normal.

Condition 1 is satisfied by definition of Θ.

Condition 2: it is enough to examine $p_t(\underset{\sim}{\theta}) \equiv p_{t1}(\underset{\sim}{\theta})$, since $p_{t2}(\underset{\sim}{\theta}) \equiv 1 - p_t(\underset{\sim}{\theta})$ has the same properties of continuity and differentiability. Now $p_t(\underset{\sim}{\theta})$ is continuous w.r.t. $\underset{\sim}{\theta}$, uniformly in t: it is continuous and uniformly bounded. Observe that

$$p_t'(\underset{\sim}{\theta}) = p_t(\underset{\sim}{\theta})\{1 - p_t(\underset{\sim}{\theta})\}(1, -t/T)^T; \tag{6.34}$$

$$p_t''(\underset{\sim}{\theta}) = p_t(\underset{\sim}{\theta})\{1 - p_t(\underset{\sim}{\theta})\}\{1 - 2p_t(\underset{\sim}{\theta})\} \begin{bmatrix} 1 & -t/T \\ -t/T & t^2/T^2 \end{bmatrix}; \tag{6.35}$$

which are both continuous w.r.t. $\underset{\sim}{\theta}$, uniformly in t, as required.

Condition 3(i): The minimum over $\underset{\sim}{\theta}$ and t of $1-p_t(\underset{\sim}{\theta})$ occurs when $\alpha - \beta t/T$ is maximized, i.e. when $\alpha = -a$ and $\beta = 0$ for all t. This yields $1-p_t(\underset{\sim}{\theta}) = (1+e^{-a})^{-1} > 0$. Likewise the minimum of $p_t(\underset{\sim}{\theta})$ occurs when $\alpha = -A$, $\beta = B$ and $t = T$, giving $p_t(\underset{\sim}{\theta}) = \dfrac{\exp(-A-B)}{1 + \exp(-A-B)} > 0$ since $-A-B$ is finite. Further, a, A and B do not depend on $\underset{\sim}{\theta}$ or t.

Condition 3(ii) is satisfied by definition.

Condition 4(i): Let $g_t(\underset{\sim}{\theta})$ be defined such that

$$n_t^{-1} g_t(\underset{\sim}{\theta}) = p_t(\underset{\sim o}{\theta})\ln\{p_t(\underset{\sim}{\theta})/p_t(\underset{\sim o}{\theta})\} + \{1-p_t(\underset{\sim o}{\theta})\}\ln\{(1-p_t(\underset{\sim}{\theta}))/(1-p_t(\underset{\sim o}{\theta})\},$$

$t = 1,\ldots,T$. We must show that $T^{-1}g_{(T)}(\underset{\sim}{\theta}) < 0$ as $T \to \infty$, for $\underset{\sim}{\theta} \neq \underset{\sim o}{\theta}$. Fix $\underset{\sim}{\theta} \neq \underset{\sim o}{\theta}$. Then $g_t(\underset{\sim}{\theta}) = 0$ if and only if $\alpha - \beta t/T = \alpha_o - \beta_o t/T$. This can happen for at most one $t = t_1$, say. Hence all g_t except this one are negative, and thus their average, $T^{-1}g_{(T)}(\underset{\sim}{\theta}) < 0$ for all T. In particular, $\lim T^{-1}g_{(T)}(\underset{\sim}{\theta}) < 0$ as $T \to \infty$.

Condition 4(ii): We must show that there exists a $\rho > 0$ for which

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \sup_{\|\underset{\sim}{\theta}\|>\rho} g_t(\underset{\sim}{\theta}) < 0.$$

It is enough that a proportion less than one of the t satisfy

$$\sup_{\|\underset{\sim}{\theta}\|>\rho} g_t(\underset{\sim}{\theta}) = 0. \tag{6.36}$$

This occurs if there is a $\underset{\sim}{\theta} \in \Theta$ with $\| \underset{\sim}{\theta} \| > \rho$ and $p_t(\underset{\sim}{\theta}) = p_t(\underset{\sim}{\theta}_o)$, implying that

$$\alpha = \beta t/T + \alpha_o - \beta_o t/T \ .$$

Such a $\underset{\sim}{\theta}$ lies on the line in Fig. 6.1 through O' with slope t/T.

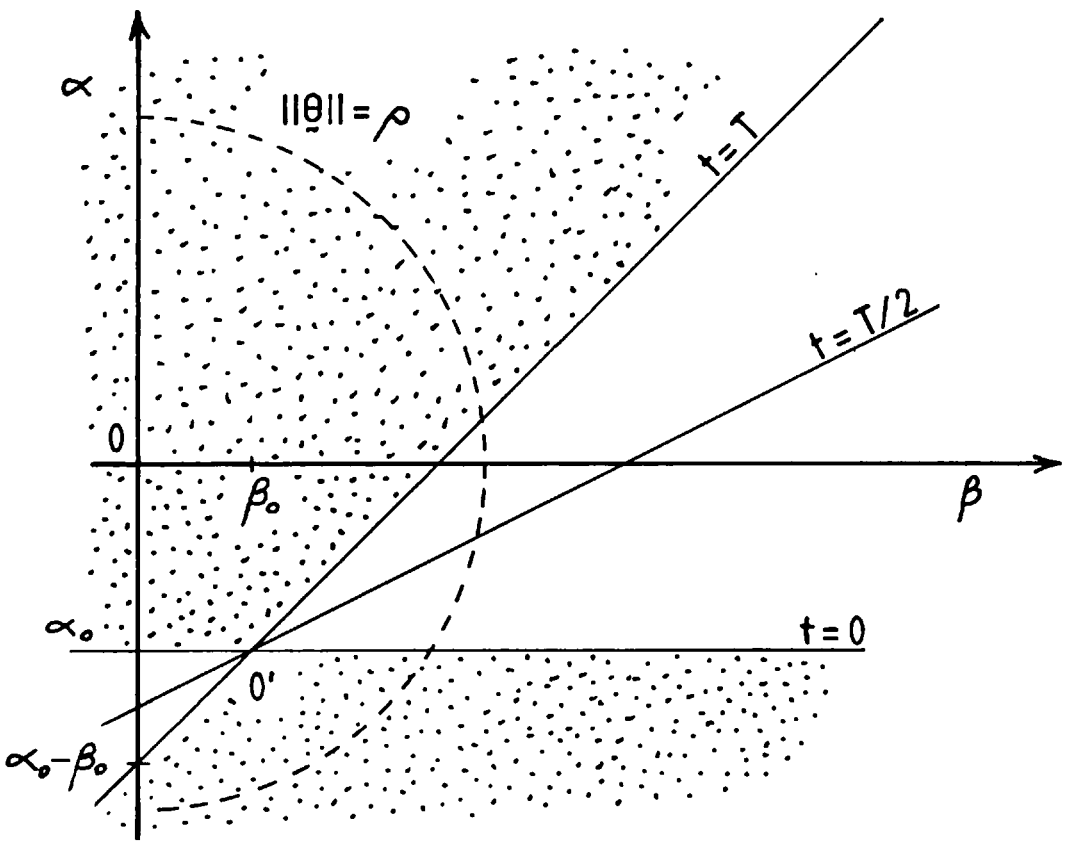


FIGURE 6.1 Regions (stippled area) where $g_t(\underset{\sim}{\theta}) < 0$ for all t.

Equation 6.36 is not satisfied for all t if $\rho > (\alpha_o^2 + B^2)^{\frac{1}{2}}$.

Conditions 5 and 6 are satisfied since the interchange of the order of summation and differentiation does not affect the binary logistic log likelihood.

Condition 7: Writing $p_t = p_t(\underset{\sim}{\theta})$,

$$\ell_t(\underset{\sim}{\theta}) = y_t \ln\{p_t/(1-p_t)\} + n_t \ln(1-p_t);$$

$$\frac{\partial \ell_t(\underset{\sim}{\theta})}{\partial p_t} = \frac{y_t}{p_t(1-p_t)} - \frac{n_t}{1-p_t} \ .$$

By (6.34) and condition 6,

$$\Gamma_T(\underset{\sim}{\theta}) = \begin{bmatrix} T^{-1}\sum_{t=1}^{T} n_t p_t(1-p_t) & -T^{-2}\sum_{t=1}^{T} n_t p_t(1-p_t)t \\ -T^{-2}\sum_{t=1}^{T} n_t p_t(1-p_t)t & T^{-3}\sum_{t=1}^{T} n_t p_t(1-p_t)t^2 \end{bmatrix}.$$

Each element in this matrix has a finite limit, by the Bounded Convergence Theorem. The limiting matrix has a positive determinant by a version of the Cauchy-Schwarz inequality for weighted sums. Thus, being a symmetric 2×2 matrix, it is positive definite.

Condition 8: Now from (6.34), using $|.|$ for absolute values,

$$|\underset{\sim}{p}_t'(\underset{\sim}{\theta})| \leq \frac{1}{4}(1,\ t/T)^T \leq \frac{1}{4}(1,1)^T \quad ,$$

for all $\underset{\sim}{\theta} \in \Theta$, as required.

Condition 9: Similarly, by (6.35)

$$|p_t''(\underset{\sim}{\theta})| \leq \frac{1}{4}\begin{bmatrix} 1 & t/T \\ t/T & t^2/T^2 \end{bmatrix} \leq \frac{1}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

for all $\underset{\sim}{\theta} \in \Theta$.

Condition 10: This condition implies that $p_t(\theta) \neq \frac{1}{2}$, whence $\alpha - \beta t/T \neq 0$. But $\alpha - \beta t/T < 0$ with upper bound $-a$. This gives

$$|p_t(\underset{\sim}{\theta}) - \tfrac{1}{2}| = |\{1 - p_t(\underset{\sim}{\theta})\} - \tfrac{1}{2}| \geq |e^{-a}/(1+e^{-a}) - \tfrac{1}{2}|$$

$$= \frac{1 - e^{-a}}{2(1+e^{-a})}$$

$$= d(a) > 0.$$

To satisfy condition 10, choose δ' such that $0 < \delta' < d(a)$.

7

DISCUSSION

7.1 Introduction

In this chapter we conclude our study of goodness-of-fit statistics for sparse contingency tables. The closeness of moderate, upper percentiles of the X^2 statistic for the product multinomials of §6.2 to the corresponding normal percentage point is examined in §7.2 We show that it is enough, under a criterion of closeness on the logit scale for percentiles near the upper 5% point, that the skewness γ_1 of X^2 satisfies $0 \leq \gamma_1 \leq 0.5$. Product multinomials which comply are tabulated. The discussion is rounded off in §7.3 with reference to power of the statistics; residuals; and goodness-of-fit statistics for comparing models.

7.2 The skewness of X^2 and Behaviour of Upper Percentiles

7.2.1 Introduction

When is a moderate, upper percentile of the X^2 statistic of Theorem 6.1 for sparse product multinomials close to its counterpart from the standard normal distribution? In this sub-section we propose measuring closeness on the logit scale, and apply it to Pearson curves and the Edgeworth series expansion. For the upper 5% point of the standardized X^2 to be close to 1.645, we need the skewness $\gamma_1 \leq 0.5$. Values of γ_1 are tabulated for various multinomials, as a guide for the use in practice of the standardized X^2.

Similar calculations for the G^2 statistic involve the summation of complicated series. However, the measures of closeness to normality that we use in this section for moderate, upper percentage points of X^2 apply equally to the G^2 statistic. They also apply to X^2 for the single sparse multinomial of §5.2, standardized by its exact mean and variance.

7.2.2 Closeness on the logit scale to normal percentiles

Our interest is in the closeness of a moderate upper percentage point of the Pearson X^2 distribution to the corresponding normal percentile. The upper $100\alpha\%$ point t_α of a distribution with mean zero and unit variance is approximately normal if

$$|\text{logit}\{\Phi(t_\alpha)\} - \text{logit}(1-\alpha)| \leq d, \qquad (7.1)$$

where $\Phi(.)$ is the standard normal cumulative distribution function, $\text{logit}(p) = \ln\{p/(1-p)\}$ and d is small. When a percentage point t_α satisfies (7.1) for $d = 0.35$ then we say that it is close to normal: if it satisfies (7.1) for $d = 0.70$ then it is deemed to be fairly close to normal. Closeness in terms of normal probabilities is shown in Table 7.1.

TABLE 7.1 Percentage points which are close to normal percentiles

closeness criterion, d (logit scale)		nominal upper 100α% point	acceptable range of upper % points*(nominal %)	
"close"	±0.35	1.64 (5 %)	1.48 (7.0%)†	1.80 (3.6%)
		2.33 (1 %)	2.20 (1.4%)	2.46 (0.7%)
"fairly close"	±0.70	1.64 (5 %)	1.31 (9.6%)	1.95 (2.5%)
		2.33 (1 %)	2.05 (2.0%)	2.58 (0.5%)

* Calculated from (7.1) to be $\Phi^{-1}[\text{logit}^{-1}\{\text{logit}(1-\alpha) \pm d\}]$.

† Corresponding upper percentage level for standard normal distribution.

7.2.3 The family of Pearson curves

Members of Pearson's family of frequency distributions for continuous random variables are unimodal, and have zero slope when the density is zero. They are standardized to have mean zero and variance one. A given value of the skewness γ_1 and of the kurtosis γ_2 defines a Pearson curve. The member with $\gamma_1 = \gamma_2 = 0$ is the standard normal distribution. For a detailed discussion, see Elderton and Johnson (1969) or Kendall and Stuart (1977, p.159ff).

"Closeness" of the upper 5% point of a Pearson curve to the corresponding normal percentile depends largely on γ_1: it is satisfied if $0 \leq \gamma_1 \leq 0.5$ and $0 \leq \gamma_2 \leq 6.0$. For the upper 1% point, the values of both γ_1 and γ_2 are important in determining closeness - see Pearson and Hartley (1972, Table 32). Thus, for the upper 5% point to be "close" to its normal counterpart we need $0 \leq \gamma_1 \leq 0.5$.

7.2.4 The Edgeworth series expansion

Consider a sequence $Y_1,\ldots,Y_n$ of independent, identically distributed, continuous random variables, each with mean zero and variance one. Define γ_j to be the (j+2)-th cumulant of Y_1, and let $Y = \Sigma_{i=1}^{n} Y_i$. The skewness and kurtosis for distributions standardized in this way are γ_1 and γ_2 respectively. Now the (j+2)-th cumulant of Y is equal to $n^{-\frac{1}{2}}\gamma_j$. This gives a motivation for considering the Edgeworth series expansion of the cumulative distribution function of Y - see Cox and Hinkley (1974, p.464, equation 3). Its upper $100\alpha\%$ point may be expressed via the Fisher-Cornish inversion as

$$k_\alpha^* + \frac{\gamma_1}{6\sqrt{n}}(k_\alpha^{*2}-1) + \frac{\gamma_2}{24n}(k_\alpha^{*3}-3k_\alpha^*) - \frac{\gamma_1^2}{36n}(2k_\alpha^{*3}-5k_\alpha^*) + o(n^{-1}), \qquad (7.2)$$

where $\Phi(k_\alpha^*) = 1 - \alpha$. Evaluation of these first four terms leads to approximate percentage points for distributions with given γ_1 and γ_2. There is a fuller discussion of these curves in Johnson and Kotz (1970, §12.4.2). The estimated cumulative distribution function is monotonic only for small values of γ_1 and γ_2 - see the Figure in Barton and Dennis (1952). But this is not important here, since we are estimating percentiles, not densities.

The skewness γ_1 of Pearson's X^2 for a multinomial is usually positive - see Lemma 6.2(c) and Tables 7.2(a) and 7.3(a). Further, positive γ_1 leads to upper tail probabilities higher than those of the normal, which is the asymptotic distribution of Pearson X^2. Thus we wish to find a $\gamma_1 > 0$ below which departures from normality are slight However, for any finite dataset, the distribution of X^2 is discrete and so the appropriate value of γ_1 would apply to a smooth approximation to the distribution of X^2. Indications are that only a little smoothing would be necessary. A single continuity correction is unsuitable here, because the points at which X^2 occurs with positive probability are not equally spaced.

Now the term in γ_2 vanishes from (7.2) if $k_\alpha^* = \sqrt{3} = 1.732$. For the normal distribution this is the upper 4.2% point. We set $n = 1$ and $k_\alpha^* = 3$ in (7.2) and adopt the "closeness" criterion of 0.35 on the logit scale to the normal upper 4.2% point. Now $\text{logit}(1 - 0.042) = 3.13$, $\text{logit}^{-1}(3.48) = 0.97$ and $\Phi^{-1}(0.970) = 1.881$. Thus

$$\sqrt{3} + \gamma_1/3 - (\sqrt{3}/36)\gamma_1^2 \leq 1.881,$$

i.e. $\gamma_1 \leq 0.48$ or $\gamma_1 \geq 6.45$. We are interested in an <u>upper</u> bound on γ_1 and so choose $\gamma_1 \leq 0.48$. This result is qualitatively the same as that obtained in §7.2.3 using Pearson curves.

7.2.5 Calculation of the skewness of X^2

From the discussions in §§7.2.2, 7.2.3 and 7.2.4 it follows for random variables where γ_j is assumed zero ($j \geq 3$) that if $\gamma_1 \leq 0.5$, then the percentage points near to the upper 5% point are close in the sense of (7.1) to their normal counterparts. The X^2 distribution for a multinomial is discrete: we will assume that the multinomial has enough cells and observations so that the cumulative distribution function of X^2 is well approximated by a smooth, increasing function. We calculate the skewness γ_1 of X^2 for single multinomials with various numbers of cells, r; numbers of observations, n; and probability vectors, $\underset{\sim}{p}$. A one-parameter subspace of the probability space is studied, with ℓ probabilities taking the same value $p \leq r^{-1}$ and $r - \ell$ taking the same larger value $(1 - \ell p)/(r - \ell)$. Now $\gamma_1 = E[(X^2 - EX^2)^3]\{\text{var } X^2\}^{-3/2}$, with mean, variance and third central moment of X^2 given in Lemma 6.2.

We now use the concept of replication as a way of judging when the asymptotic result of Theorem 1 holds approximately, at the upper 5% point. Consider k independent random vectors $\underset{\sim}{Y}_i \sim mn_r(n,\underset{\sim}{p})$, $i = 1,\ldots,k$. Let $X_1^2(\underset{\sim}{Y}_i)$ be the Pearson X^2 statistic for the goodness of fit of $\underset{\sim}{Y}_i$ to the given multinomial distribution, and set its skewness to be γ_1. Then the skewness of

$$X_k^2 = \sum_{i=1}^{k} X_1^2(\underset{\sim}{Y}_i)$$

is $\gamma_1(k) = k^{-\frac{1}{2}}\gamma_1$. We evaluate k_0, the number of replicates required so that $\gamma_1(k_0) \leq 0.5$, i.e.

$$k_0 = \begin{cases} \min(4^2\gamma_1,\ 1) & \text{if } \gamma_1 > 0; \\ 1 & \text{otherwise.} \end{cases}$$

This replication approach mimics what may happen in practice: the same X^2 statistic would be observed if the elements of $\underset{\sim}{p}$ were permuted for each $\underset{\sim}{Y}_i$. We ignore the possibility of collapsing by the sufficiency argument to $\Sigma\underset{\sim}{Y}_i$.

The skewness of X^2 is tabulated for various multinomial distributions in Tables 7.2(a) and 7.3(a). The entries for sample size $n = \infty$ are calculated from the skewness of a χ^2_{r-1} distribution, $\gamma_1 = 2\sqrt{2}(r-1)^{-\frac{1}{2}}$. Tables 7.2(b) and 7.3(b) give the number of replications k required to reduce the skewness of the replicated sample to 0.5 from that shown in the corresponding entry in Table (a). We make the following general observations:

1. γ_1 tends to be higher for small values of r.

2. γ_1 is higher when the smallest probability $p = 0.01$ than when $p = 0.1$ (Table 7.2).

3. γ_1 appears to attain a minimum when all cell probabilities are equal, except when $n = 1$.

4. For multinomials with many cells (Table 7.2), γ_1 appears unaffected (to our accuracy) by changing cell probabilities, and γ_1 becomes negligible for $n \geq 100$, and $r \geq 100$, for the probability regimes studied. The choice of $p = r^{-1} - r^{-2}$ is to satisfy Morris's (1975) criterion for a single, sparse multinomial, including $\max_{1 \leq i \leq r} p_i \to 0$ as the number of cells $r \to \infty$.

5. If $n = 1$ and $p = r^{-1}$, then X^2 is a constant, hence $\gamma_1 \to \infty$.

The recommendations below for when to use the standardized asymptotic approximation to X^2 at the upper 5% point are based on Tables 7.3(a) and 7.3(b). In the equiprobable case, a product of $k \geq 9$ multinomials each with $r \geq 5$ and $n \geq 5$ satisfies our closeness criterion. The unsymmetric cases studied with $r \geq 5$, $n \leq 10$ and $8 \leq k \leq 79$ also suffice.

TABLE 7.2 Pearson X^2 statistic for multinomial distributions

(a) Skewness γ_1

No. cells, r		2	5		10	
No. small probs, ℓ sample size, n		1	1	4	1	9
small prob.						
p	1	2.67	2.67	0.41	∞	∞
	2	5.32	2.43	2.55	2.67	2.67
	5	5.88	1.82	2.73	1.48	1.48
0.1	10	5.11	1.61	2.30	1.19	1.19
	10^2	3.17	1.43	1.53	0.97	0.97
	10^3	2.86	1.42	1.43	0.95	0.95
	10^4	2.83	1.41	1.42	0.94	0.94
	∞	(2.83)	(1.41)	(1.41)	(0.94)	(0.94)
	1	9.85	9.85	4.69	9.85	2.87
	2	8.29	7.61	4.00	6.65	2.50
	5	7.50	5.62	3.68	3.87	2.38
0.01	10	7.49	4.45	3.72	2.60	2.45
	10^2	5.89	1.99	2.93	1.13	1.93
	10^3	3.34	1.48	1.66	0.96	1.10
	10^4	2.88	1.42	1.44	0.94	0.96
	∞	(2.83)	(1.41)	(1.41)	(0.94)	(0.94)
	1	∞	∞		∞	
	2	0.00	1.50		2.67	
equal	5	1.90	1.42		1.48	
cell	10	2.39	1.42		1.19	
probs	10^2	2.79	1.41		0.97	
	10^3	2.82	1.41		0.95	
	10^4	2.83	1.41		0.94	
	∞	(2.83)	(1.41)		(0.94)	

/contd

TABLE 7.2 continued

(b) Number of replications k to satisfy $\gamma_1 \leq 0.5$

No. cells, r		2	5		10	
No. small probs, ℓ		1	1	4	1	9
sample size, n						
small prob. p	1	28	28	1	∞	∞
	2	113	24	26	28	28
0.1	5	138	13	30	9	9
	10	105	10	21	6	6
	10^2	40	8	9	4	4
	10^3	33	8	8	4	4
	10^4	32	8	8	4	4
	1	388	388	88	388	33
	2	275	232	64	177	25
0.01	5	225	126	54	60	23
	10	224	79	55	27	24
	10^2	139	16	34	5	15
	10^3	45	9	11	4	5
	10^4	33	8	8	4	4
	1	∞	∞		∞	
equal	2	1	9		28	
cell	5	14	8		9	
probs	10	23	8		6	
	10^2	31	8		4	
	10^3	32	8		4	
	10^4	32	8		4	

Note: For each value of p, entries below the solid line (——) require at least 100 observations in total to satisfy $\gamma_1 \leq 0.5$: those below the dashed line (- - - - -) require at least 1000 observations.

TABLE 7.3 Pearson X^2 statistic: sparse multinomial distributions

(a) Skewness γ_1

No. cells, r		10			100			1000		
No. small probs, ℓ No. observations, n		1	2	8	1	20	80	1	200	800
small prob $p = r^{-1}-r^{-2}$	1	2.87	1.67	-0.98	9.90	1.52	-1.44	31.51	1.50	-1.49
	10	1.19	1.20	1.25	1.71	1.71	1.71	4.78	4.78	4.78
	10^2	0.97	0.97	0.97	0.42	0.42	0.42	0.54	0.54	0.54
	10^3	0.95	0.95	0.95	0.30	0.30	0.30	0.13	0.13	0.13
	10^4	0.94	0.94	0.94	0.29	0.29	0.29	0.09	0.09	0.09
equal cell probs	1		∞			∞			∞	
	10		1.19			1.71			4.78	
	10^2		0.97			0.42			0.54	
	10^3		0.95			0.30			0.13	
	10^4		0.94			0.29			0.09	

(b) Number of replications k to satisfy $\gamma_1 \leq 0.5$

No. cells, r		10			100			1000		
No. small probs, ℓ No. observations, n		1	2	8	1	20	80	1	200	800
small prob $p = r^{-1}-r^{-2}$	1	33	11	1	392	9	1	3972	9	1
	10	6	6	6	12	12	12	91	91	91
	10^2	4	4	4	1	1	1	1	1	1
	10^3	4	4	4	1	1	1	1	1	1
	10^4	4	4	4	1	1	1	1	1	1
equal cell probs	1		∞			∞			∞	
	10		6			12			91	
	10^2		4			1			1	
	10^3		4			1			1	
	10^4		4			1			1	

Single sparse multinomials with $r \geq 100$ and $n \geq 100$ have small enough skewness, for all of the probability regimes studied. However, for any given unsymmetric hypothesis, one should calculate the skewness directly, as this study has not covered all possibilities.

7.3 Further Aspects of Goodness of Fit

Goodness-of-fit statistics for sparse contingency tables have been studied in terms of their asymptotic distribution (Chapters 5 and 6) and small-sample behaviour (§7.2). In this subsection we examine briefly three other aspects: power of the statistics; residuals; and statistics for comparing the fit of competing models.

When there is a departure from the hypothesized model, it is important that it is detected. In general a given goodness-of-fit statistic may be powerful against some alternative hypotheses, but not others. For the single, sparse multinomial with r equiprobable cells, Stein (Hoeffding, 1965, p.372) has shown the Pearson X^2 statistic to be much more powerful than G^2 against near alternatives $\underset{\sim}{p}$ satisfying $\max_i |p_i - r^{-1}| = o(r^{-1})$ for moderate significance levels, with n/r moderate and r large. This was confirmed somewhat in a Monte Carlo study reported by Koehler and Larntz (1980, §4). For near alternatives to a symmetrical hypothesis (i.e. all cell probabilities equal), X^2 dominated G^2 slightly. When few true probabilities were large and the rest near zero they found X^2 was much more powerful than G^2. The reverse occurred when most of the cell probabilities were equal, but a few close to zero. However, X^2 dominated G^2 for a large part of the simplex, for the number of cells k moderately large. They found no simple conclusion for power against unsymmetrical null hypotheses, since either X^2 or G^2 could dominate. It is hoped that similar results hold in the product multinomial situation.

The examination of residuals is an effective way of finding where in the contingency table departures from the fitted statistical model occur. Both X^2 and G^2 may be written as a sum over all cells of squared residuals R, where $R_X = (Y - \hat{Y})/\hat{Y}$ for X^2, and

$R_g = \pm\sqrt{2}\{Y \ln|Y/\hat{Y}|\}^{\frac{1}{2}}$ for G^2, attaching a negative sign to R_g if $Y < \hat{Y}$, where Y is the observed count in a cell and $\hat{Y}$ is the fitted or expected count, according to the hypothesized model. These two types of residual behave differently when Y and $\hat{Y}$ are small:

Residual	Value of residual	
	Y = 0	Y = 1
$(Y - \hat{Y})/\hat{Y}$	$-\sqrt{\hat{Y}}$	$(1 - \hat{Y})/\sqrt{\hat{Y}}$
$\pm\sqrt{2}\{Y \ln (Y/\hat{Y})\}^{\frac{1}{2}}$	0	$-\sqrt{2}(\ln \hat{Y})^{\frac{1}{2}}$

In particular, R_g has the same zero contribution to G^2 for all cells with zero counts, while R_x has a negative contribution. When Y = 1, both residuals are positive for $\hat{Y} < 1$, with $R_g < R_x$ for $\hat{Y} < 0.199$ and $R_g > R_x$ for $0.199 < \hat{Y} < 1$. Thus it is to be expected, as discussed by Koehler and Larntz (1980) and in §§5.2 and 6.2.2, that the asymptotic distributions of X^2 and G^2 differ for small expected cell probabilities. It must be remembered that residuals are correlated: a departure from the model in one cell cannot necessarily be isolated, especially when the number of cells in the multinomial is small.

We have studied statistics for measuring the fit of a statistical model to the data. It is also interesting to compare the fit of two possible models. Haberman (1977) discussed the asymptotic distribution of X^2 and G^2 for comparing the maximum likelihood fit of loglinear models, one a special case of the other. He represented each loglinear model in terms of a linear manifold $\mathcal{M}_t$ which contained the vector of length k of true log cell means. The sampling scheme was represented by a linear manifold $\mathcal{N}_k \subset \mathcal{M}_k$ in which the required restrictions on marginal configurations were satisfied. Under conditions for the larger model on elements of the orthogonal complement $\mathcal{B}_k$ of $\mathcal{N}_k$

with respect to $\mathcal{M}_k$ and on an inner product weighted by the true cell means, he showed that X^2 and G^2 were asymptotically χ^2_q where $q > 0$ was the limit of the difference in dimension of the $\mathcal{S}_k$ for the two models, as $k \to \infty$, and that the difference $G^2 - X^2$ converged in probability to zero. It is easily seen to be so for direct loglinear models. If the smaller model differs from the larger model by a parameter of finite dimension q, then the G^2 statistic for comparing the two models may be expressed in terms of the configurations which are in one model but not the other. Since their dimension remains finite, the number of observations per cell becomes large in the limit and the classical asymptotics may be applied, even if the table of counts from which the configuration of interest was derived is sparse. Thus both the G^2 and X^2 statistics for comparing the models have a χ^2_q distribution in the limit - see Bishop, Fienberg and Holland (1975, §4.2.2).

7.4 Concluding Remarks

The asymptotic framework discussed in Chapters 5 and 6 enables the fit of contingency tables with many observed zeros to be ascertained using the Pearson X^2 and likelihood ratio G^2 statistics. Conditions on cell means bound them uniformly away from zero and infinity as the number of cells becomes large. The second bound is achieved in Chapter 5 for the single sparse multinomial by requiring the largest cell probability to approach zero in the limit: for product multinomials, the requirement in Theorems 6.2 to 6.4 is that the number of observations per multinomial is bounded uniformly. Application of the asymptotic distributions at the upper 5% level to test the fit of known distributions appears justified by §7.2 for the single sparse multinomials studied when the number of cells and observations exceed 100. For product multinomials under a null hypothesis of equal cell probabilities, it is enough that there are more than five cells and observations per multinomial, and at least nine independent multinomials in the product. The skewness of X^2 should be calculated for unsymmetric hypotheses, to determine whether the asymptotic approximation is appropriate. Further research is needed to establish the asymptotic distribution of statistics for comparing models, in a more general context to that studied by Haberman (1977).

APPENDIX I

First Partial Derivatives of the Log Likelihood for GCR Models

Consider the GCR model defined for table Y_k consisting of n_k observations at x_k according to (3.4) and (3.9), with $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \Delta)$, ($\Delta_{ij} = \ln \psi_{ij}$), and marginal distributions linear logistic. The cell probabilities are given by

$$\pi_{ij/k} \equiv \pi_{ij}(x_k;\theta) = F_{ijk} - F_{i-1,j,k} - F_{i,j-1,k} + F_{i-1,j-1,k},$$

where $F_{ijk} = F(i,j,x_k;\theta)$, $i = 1,\ldots,r$; $j = 1,\ldots,c$. Let $\eta_{ik} = F_1(i,x_k;\theta)$, $i = 1,\ldots,r-1$ and $\xi_{jk} = F_2(j,x_k;\theta)$, $j = 1,\ldots,c-1$. Then the contribution to the log likelihood function for the global cross-ratio model at x_k is, up to an arbitrary constant,

$$\ell_k(y_k,x_k;\theta) = \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ijk} \ln \pi_{ij}(x_k;\theta).$$

The full likelihood for all m contributions is obtained by summing the ℓ_k:

$$\ell(\theta;y_1,\ldots,y_m,x_1,\ldots,x_m) = \sum_{k=1}^{m} \ell_k(\theta;y_k,x_k) . \qquad (A.1)$$

Now from the definition of F and its marginal distributions given in (3.2)

$$\begin{aligned}
\partial\ell_k/\partial F_{ijk} &= y_{ijk}/\pi_{ij/k} - y_{i+1,j,k}/\pi_{i+1,j/k} - y_{i,j+1,k}/\pi_{i,j+1/k} \\
&\quad + y_{i+1,j+1,k}/\pi_{i+1,j+1/k}, \qquad i = 1,\ldots,r-1; \quad j = 1,\ldots,c-1; \\
&= y_{rjk}/\pi_{rj/k} - y_{r,j+1,k}/\pi_{r,j+1/k}, \quad i = r; \; j = 1,\ldots,c-1; \\
&= y_{ick}/\pi_{ic/k} - y_{i+1,c,k}/\pi_{i+1,c/k}, \quad i = 1,\ldots,r-1; \; j = c; \\
&= y_{rck}/\pi_{rc/k}, \qquad i = r; \quad j = c; \qquad (A.2)
\end{aligned}$$

and

$$\frac{\partial \ell_k}{\partial \eta_{ik}} = \frac{\partial \ell_k}{\partial F_{ick}} + \sum_{j=1}^{c-1} \frac{\partial \ell_k}{\partial F_{ijk}} \cdot \frac{\partial F_{ijk}}{\partial \eta_{ik}} \quad . \tag{A.3}$$

Now it is easily shown by (3.9) that

$$\partial F_{ijk}/\partial \eta_{ik} = \frac{1}{2}[1-S_{ijk}^{-1}\{1 + \exp(\Delta_{ij})(\eta_{ik}-\xi_{jk}) - \eta_{ik} - \xi_{jk}\}] \tag{A.4}$$

and by (3.4) that

$$\partial F_{ijk}/\partial \Delta_{ij} = \exp(\Delta_{ij})\, 2\{\exp(\Delta_{ij})-1\}^{2} S_{ijk}^{-1}]^{-1} \times [1-S_{ijk}+\{\exp(\Delta_{ij})-1\}(\eta_{ik}+\xi_{jk}-2\eta_{ik}\xi_{jk})] \tag{A.5}$$

where

$$S_{ijk} = ([1+\{\exp(\Delta_{ij})-1\}(\eta_{ik}+\xi_{jk})]^{2}+4\exp(\Delta_{ij})\{1-\exp(\Delta_{ij})\}\eta_{ik}\xi_{jk})^{\frac{1}{2}},$$
$$i = 1,\ldots,r-1;\ j = 1,\ldots,c-1.$$

By the chain rule for differentiation, the required partial derivatives are

$$\partial \ell_k/\partial \alpha_{1i} = \partial \ell_k/\partial \eta_{ik} \cdot \eta_{ik}(1-\eta_{ik}) \quad , \quad i = 1,\ldots,r-1; \tag{A.6}$$

$$\partial \ell_k/\partial \underset{\sim}{\beta}_1 = -\underset{\sim}{x}_k \sum_{i=1}^{r-1} \partial \ell_k/\partial \alpha_{1i} \quad , \tag{A.7}$$

and

$$\frac{\partial \ell_k}{\partial \Delta_{ij}} = \sum_{q=1}^{r-1} \sum_{s=1}^{c-1} \frac{\partial \ell_k}{\partial F_{qsk}} \frac{\partial F_{ijk}}{\partial \Delta_{ij}} \qquad \begin{matrix} i = 1,\ldots,r-1; \\ j = 1,\ldots,c-1. \end{matrix} \tag{A.8}$$

These expressions are obtained in full by substitution from (A.2)-(A.5); The derivatives of the likelihood for all m contributions are obtained by summation of (A.6)-(A.8) over k = 1,...,m. For the GCR model with Δ constant for all cutpoints (i,j) as in (3.7),

$$\partial \ell_k/\partial \Delta = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \partial \ell_k/\partial \Delta_{ij}.$$

Similar expressions to (A.6) and (A.7) are obtained for $\partial \ell_k/\partial \alpha_{2j}$ (j = 1,...,c-1) and $\partial \ell_k/\partial \underset{\sim}{\beta}_2$, respectively.

APPENDIX II

Hoadley's Conditions for Maximum Likelihood Estimators

We present Hoadley's (1971) conditions for consistency and asymptotic normality of the maximum likelihood estimate $\hat{\underset{\sim}{\theta}}$ for independent non-identically distributed random variables which depend on a finite-dimensional unknown $\underset{\sim}{\theta}$. They are used for establishing the conditions in §6.3.2 for consistency and asymptotic normality of the maximum likelihood estimator, for product multinomials satisfying sparseness conditions as the number of cells k becomes large. Hoadley also uses some stronger, more easily proved conditions, marked with a prime ('). Before stating Hoadley's conditions, we set up some notation.

For $i = 1,\ldots,k$, let $\underset{\sim}{Y}_i = (Y_{i1},\ldots,Y_{ir})^T$ $(r \geq 2)$ be a sequence of independent random vectors defined on a probability space $(\Omega, \mathcal{F}, P_\theta)$ and taking values in a measure space $(\mathcal{Y}, \mathcal{A}, \mu)$, where $\underset{\sim}{\theta} \in \Theta$ is a subspace of R^p. The probability function is given as $f_i(\underset{\sim}{y}\,|\underset{\sim}{\theta}) = P[\underset{\sim}{Y}_i = \underset{\sim}{y}|\underset{\sim}{\theta}]$. Let

$$R_i(\underset{\sim}{\theta}) = \ln[f_i(\underset{\sim}{Y}_i|\underset{\sim}{\theta})/f_i(\underset{\sim}{Y}_i|\underset{\sim}{\theta}_0)] \qquad \text{if } f_i(\underset{\sim}{Y}_i|\underset{\sim}{\theta}_0) > 0$$

$$= 0 \qquad \text{otherwise;}$$

$$R_i(\underset{\sim}{\theta},\rho) = \sup\{R_i(\underset{\sim}{t}): \|\underset{\sim}{t} - \underset{\sim}{\theta}\| \leq \rho\};$$

$$V_i(r) = \sup\{R_i(\underset{\sim}{\theta}): \|\underset{\sim}{\theta}\| > r\};$$

where $\|\cdot\|$ is the usual Euclidean norm on R^p. For any random variable X let

$$X^{(B)} = X \qquad \text{if } X \geq -B$$

$$= B \qquad \text{if } X < -B, \quad \text{where } B \geq 0.$$

We denote expectations for $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$ by lower case: $ER_i(\underset{\sim}{\theta},\rho) = r_i(\underset{\sim}{\theta},\rho)$, $EV_i^{(B)}(r) = v_i^{(B)}(r)$, and so on. If the value for $\underset{\sim}{\theta}$ under which expectation is taken is not stated, it is assumed to be $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$. The arithmetic mean of the quantities, a_i, $i = 1,\ldots,k$ is $\bar{a}_k = k^{-1}\Sigma_i a_i$.

Let $\ell_i(\underset{\sim}{y},\underset{\sim}{\theta}) = \ln f_i(\underset{\sim}{y}|\underset{\sim}{\theta})$, $i = 1,\ldots,k$. We define $\underset{\sim}{\ell}_i'(\underset{\sim}{y}|\underset{\sim}{\theta})$ to be the vector whose j-th component is $\ell_{i,j}'(\underset{\sim}{y},\underset{\sim}{\theta}) = (\partial/\partial\theta_j)\ell_i(\underset{\sim}{y},\underset{\sim}{\theta})$, $j = 1,\ldots,p$. Similarly we set $\ell_i''(\underset{\sim}{y},\underset{\sim}{\theta})$ to be the $p\times p$ matrix with (j,ℓ)-th entry $\ell_{i,j\ell}''(\underset{\sim}{y},\underset{\sim}{\theta}) = (\partial^2/\partial\theta_j\partial\theta_\ell)\ell_i(\underset{\sim}{y},\underset{\sim}{\theta})$.

Consistency. Let the following conditions hold:

C1: Θ is a closed subset of R^p;

C2: $f_i(\underset{\sim}{Y}_i|\underset{\sim}{\theta})$ is an upper semicontinuous function of $\underset{\sim}{\theta}$, uniformly in i, almost surely with respect to the probability measure P;

C3: there exist $\rho^* = \rho^*(\underset{\sim}{\theta}) > 0$, $\delta > 0$, $K > 0$ and $r > 0$ satisfying

(i) $E[R_i^{(O)}(\underset{\sim}{\theta},\rho)]^{1+\delta} \leq K, \quad 0 \leq \rho \leq \rho^*$;

(ii) $E[V_i^{(O)}(r)]^{1+\delta} \leq K$;

C4: there exists $B > 0$ and $r > 0$ such that

(i) $\bar{r}^{(B)}(\underset{\sim}{\theta}) = \limsup \bar{r}_k^{(B)}(\underset{\sim}{\theta}) < 0, \quad \underset{\sim}{\theta} \neq \underset{\sim}{\theta}_0$;

(ii) $\limsup \bar{v}_k^{(B)}(r) < 0$;

C5: $R_i(\underset{\sim}{\theta},\rho)$ and $V_i(r)$ are measurable functions of $\underset{\sim}{Y}_i$, $i = 1,\ldots,K$.

Some stronger, more easily proved versions given by Hoadley are:

C3': there exists $\rho^* = \rho^*(\underset{\sim}{\theta}) > 0$, $r > 0$ and $K > 0$ such that

(i) $E[R_i(\underset{\sim}{\theta},\rho)]^2 \leq K, \quad 0 \leq \rho \leq \rho^*$;

(ii) $E[V_i(r)]^2 \leq K$.

C4': (i) $\lim_{k\to\infty} \bar{r}_k(\underset{\sim}{\theta}) < 0 \quad$ for $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}_0$;

(ii) there exists $\rho > 0$ such that

$$\lim_{k\to\infty} \bar{v}_k(\rho) < 0.$$

Normality. Consider the following conditions:

N1: Θ is an open subset of R^p;

N2: $\hat{\underset{\sim}{\theta}}_k$ converges in probability to $\underset{\sim}{\theta}_o$;

N3: $\ell_i'(Y_{\sim i}, \underset{\sim}{\theta})$ and $\ell_i''(Y_{\sim i}, \underset{\sim}{\theta})$ exist almost surely with respect to the probability measure P, $i = 1,\ldots,k$;

N4: $\ell_i''(Y_{\sim i}, \underset{\sim}{\theta})$ is a continuous function of $\underset{\sim}{\theta}$, uniformly in i, almost surely with respect to P and is a measurable function of $Y_{\sim i}$;

N5: $E[\ell_i'(Y_{\sim i}, \underset{\sim}{\theta}) \mid \underset{\sim}{\theta}] = \underset{\sim}{0}$, $i = 1,\ldots,k$;

N6: $\Gamma_i(\underset{\sim}{\theta}) = E[\ell_i'(Y_{\sim i}, \underset{\sim}{\theta})\ell_i'^{T}(Y_{\sim i}, \underset{\sim}{\theta}) \mid \underset{\sim}{\theta}] = -E\ \ell_i''(Y_{\sim i}, \underset{\sim}{\theta}) \mid \underset{\sim}{\theta}$, $i = 1,\ldots,k$;

N7: $\bar{\Gamma}_k(\underset{\sim}{\theta}) = k^{-1}\Gamma_{(k)}(\underset{\sim}{\theta}) \to \Gamma(\underset{\sim}{\theta})$, positive definite;

N8: there exists a $\delta > 0$ such that $k^{-1-\delta/2}\Sigma_i E\left|\underset{\sim}{\lambda}^T \ell_i'(Y_{\sim i}, \underset{\sim}{\theta}_o)\right|^{2+\delta} \to 0$ for all $\underset{\sim}{\lambda} \in R^p$;

N9: there exist $\varepsilon > 0$, $\delta > 0$, $K > 0$ and random variables $B_{i,j\ell}(Y_{\sim i})$ satisfying

(i) $\sup\{\left|\ell''_{i,j\ell}(Y_{\sim i}, \underset{\sim}{t})\right| : \left\| \underset{\sim}{t} - \underset{\sim}{\theta}_o \right\| \leq \varepsilon\} \leq B_{i,j\ell}(Y_{\sim i})$;

(ii) $E\left|B_{i,j\ell}(Y_{\sim i})\right|^{1+\delta} \leq K$.

A stronger, more easily proved condition provided by Hoadley is, for $i = 1,\ldots,k$; $q = 1,\ldots,p$:

N8': there exists $K > 0$ such that

$$E\left|\ell'_{i,q}(Y_{\sim i}, \underset{\sim}{\theta}_o)\right|^3 \leq K.$$

REFERENCES

Agresti, A. (1980). Generalized odds ratios for ordinal data. Biometrics 36, 59-67.

Agresti, A. (1981). Measures of nominal-ordinal association. J. Amer. Statist. Ass. 76, 524-529.

Alvey, N.G. et al. (1977). Genstat, a general statistical program. Harpenden, Herts., United Kingdom: Rothamsted Experimental Station.

Anderson, J.A. (1982). Regression modelling for ordered categorical variables. Unpublished manuscript.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika 43, 561-573.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. Biometrics 35, 403-415.

Anscombe, F.J. (1981). Computing in statistical science through APL. New York: Springer-Verlag.

Baker, R.J. and Nelder, J.A. (1978). The GLIM system manual, Release 3. Oxford: Numerical Algorithms Group.

Bartlett, N.R. (1978). A survival model for a wood preservative trial. Biometrics 34, 673-679.

Barton, D.E. and Dennis, K.E. (1952). The conditions under which Gram-Charlier and Edgeworth curves are positive definite and unimodal. Biometrika 39, 425-427.

Billingsley, P. (1968). Convergence of probability measures. New York: Wiley.

Birch, M.W. (1964). The detection of partial association. I: the 2×2 case. J. Roy. Statist. Soc. B 26, 313-324.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, Massachusetts: MIT Press.

Clayton, D.G. (1974). Some odds-ratio statistics for the analysis of ordered categorical data. Biometrika 61, 525-531.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests. Biometrics 10, 417-451.

Cox, D.R. (1970). The analysis of binary data. London: Chapman and Hall.

Cox, D.R. (1972). Regression models and life tables (with Discussion). J. Roy. Statist. Soc. B 34, 187-220.

Cox, D.R. (1975). Partial likelihood. Biometrika 62, 269-276.

Cox, D.R. and Hinkley, D.V. (1974). Theoretical statistics. London: Chapman and Hall.

Cramér, H. (1946). Mathematical methods of statistics. Princeton: Princeton University Press.

Eckler, A.R. (1973). The similarity of two poems. In Statistics by example: finding models, edited by F. Mosteller et al., 75-87.

Elderton, W.P. and Johnson, N.L. (1969). Systems of frequency curves. Cambridge: University Press.

Farewell, V.T. (1982). A note on regression analysis of ordinal data with variability of classification. Biometrika 69, 533-538.

Fienberg, S.E. (1979). The use of chi-squared statistics for categorical data problems. J. Roy. Statist. Soc. B 41, 54-64.

Fienberg, S.E. (1980). The analysis of cross-classified categorical data, Second edition. Cambridge, Massachusetts: MIT Press.

Gilby, W.H. (1911). On the significance of the teacher's appreciation of general intelligence. Biometrika 8, 94-108.

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. J. Amer. Statist. Ass. 74, 537-552.

Goodman, L.A. (1981a). Association models and the bivariate normal for contingency tables with ordered categories. Biometrika 68, 347-355.

Goodman, L.A. (1981b). Association models and the canonical correlation in the analysis of cross-classifications having ordered categories. J. Amer. Statist. Ass. 76, 320-334.

Goodman, L.A. and Kruskal, W.H. (1954). Measures of association for cross classifications. J. Amer. Statist. Ass. 49, 732-764.

Goodman, L.A. and Kruskal, W.H. (1959). Measures of association for cross classifications. II: Further discussion and references. J. Amer. Statist. Ass. 54, 123-163.

Goodman, L.A. and Kruskal, W.H. (1963). Measures of association for cross classifications. III: Approximate sampling theory. J. Amer. Statist. Ass. 58, 310-364.

Haberman, S.J. (1974). Log-linear models for frequency tables with ordered classifications. Biometrics 30, 589-600.

Haberman, S.J. (1977). Log-linear models and frequency tables with small expected cell counts. Ann. Statist. 5, 1148-1169.

Haldane, J.B.S. (1937). The exact value of the moments of the distribution of χ^2, used as a test of goodness of fit, when expectations are small. Biometrika 29, 133-143.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimates for the independent not identically distributed case. Ann. Math. Statist. 42, 1977-1991.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. Ann. Math. Statist. 36, 369-401.

Johnson, N.L. and Kotz, S. (1970). Distributions in statistics: Continuous univariate distributions, I. Boston: Houghton Mifflin.

Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267-278.

Kendall, M.G. (1948). Rank correlation methods. London: Griffin.

Kendall, M.G. and Stuart, A. (1967). The advanced theory of statistics, Volume II, Second edition. London: Griffin.

Kendall, M.G. and Stuart, A. (1977). The advanced theory of statistics, Volume I, Fourth edition. London: Griffin.

Kiser, C.V. and Whelpton, P.K. (1950). Fertility planning and fertility rates by socio-economic status. In Social and psychological factors affecting fertility, Volume 2, by P.K. Whelpton and C.V. Kiser (Eds). Millbank Memorial Fund.

Koehler, K. ( ). Distribution of the likelihood ratio goodness-of-fit statistic for contingency tables with small expected frequencies. Unpublished manuscript.

Koehler, K.J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. J. Amer. Statist. Ass. 75, 336-344.

Lancaster, H.O. and Hamdan, M.A. (1964). Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters. Psychometrika 29, 383-391.

Lehmann, E.L. (1966). Some concepts of dependence. Ann. Math. Statist. 37, 1137-1153.

Lindon, J.A. and Bergerson, H.W. (1969). The vocabularyclept poem, no. 1. Word Ways, J. Recreational Linguistics 2, 85-89.

Loève, M. (1963). Probability theory, Third edition. Princeton, New Jersey: Van Nostrand.

McCullagh, P. (1980). Regression models for ordinal data (with Discussion). J. Roy. Statist. Soc. B 42, 109-142.

Mann, H.B. and Wald, A. (1943). On stochastic limit and order relationships. Ann. Math. Statist. 14, 217-226.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. J. Amer. Statist. Ass. 58, 690-700.

Mantel, N. and Brown, C. (1973). A logistic reanalysis of Ashford and Sowden's data on respiratory symptoms in British coal miners. Biometrics 29, 649-665.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst. 22, 719-748.

Mardia, K.V. (1970). Families of bivariate distributions. London: Griffin.

Morgan, W.A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. Biometrika 31, 13-19.

Morris, C. (1975). Central limit theorems for multinomial sums. Ann. Statist. 3, 165-188.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. J. Roy. Statist. Soc. A 135, 370-384.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20-A, 175-247, 264-299.

Numerical Algorithms Group (1978). NAG FORTRAN library manual, Mark 7. Oxford: Numerical Algorithms Group, Ltd.

Olsson, U. (1980). Measuring association in ordered 2-way contingency tables. J. Marketing Res. 17, 391-394.

Pearson, E.S. and Hartley, H.O. (1970). Biometrika tables for statisticians, Volume I, Third edition. Cambridge: University Press.

Pearson, E.S. and Hartley, H.O. (1972). Biometrika tables for statisticians, Volume II. Cambridge: University Press.

Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. 50, 157-175.

Pearson, K. and Heron, D. (1913). On theories of association. Biometrika 9, 159-315.

Plackett, R.L. (1965). A class of bivariate distributions. J. Amer. Statist. Ass. 60, 516-522.

Prentice, R.L. and Gloeckler, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34, 57-67.

Price, W.E., Grizzle, J.E., Postlethwait, R.W., Johnson, W.D. and Grabicki, P. (1970). Results of operation for duodenal ulcer. Surgery, Gynecology and Obstetrics 131, 233-244.

Rao, C.R. (1973). Linear statistical inference and its applications, Second edition. New York: Wiley.

Royden, H.L. (1968). Real analysis, Second edition. New York: Macmillan.

Seber, G.A.F. (1977). Linear regression analysis. New York: Wiley.

Serfling, R.J. (1980). Approximation theorems of mathematical statistics. New York: Wiley.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. J. Roy. Statist. Soc. B 13, 238-241.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. Biometrika 40, 105-110.

Stuart, A. (1963). Calculation of Spearman's rho for ordered two-way classifications. Amer. Statistician 17:4, 23-24

Thompson, W.A. (1977). On the treatment of grouped observations in life studies. Biometrics 33, 463-470.

Tukey, J.W. (1977). Exploratory data analysis. Reading, Massachusetts: Addison-Wesley.

Wahrendorf, J. (1980). Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions. Biometrika 67, 15-21.

Williams, O.D. and Grizzle, J.E. (1972). Analysis of contingency tables having ordered response categories. J. Amer. Statist. Ass. 67, 55-63.

Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characters. Biometrika 35, 176-181.