IMPERIAL COLLEGE OF SCIENCE AND TECHNOLOGY

Department of Mathematics

ANALYSIS OF ORDERED CATEGORICAL DATA

by

Peter McCullagh

Thesis submitted for the degree of Doctor of Philosophy in the University of London and for the Diploma of Imperial College

August 1977

ABSTRACT

A general method for deriving models for ordered categorised response variables is described. This procedure relies on the existence of an underlying continuous variable which can be modelled and whose error distribution is known up to a few parameters. When the model is linear and the error distribution is logistic, the model for the category probabilities is cumulative logit linear or cumulative logit multiplicative depending on whether or not the error variance is constant. Some simple estimators of the parameters of interest are derived and simulation methods are used to examine their first and second order properties in small samples.

The relationship between the general model for ordered categories with arbitrary error distribution and asymptotically most powerful rank tests is described. The joint asymptotic distribution of a pair of rank tests under various hypotheses is derived and an example illustrates how these tests can be used jointly.

The logistic model for paired binary data is extended to the many category case. Some results concerning mixtures of binomial random variables are proved to help derive estimators of the parameter of interest in the presence of nuisance parameters. Two such estimators are described and simulation methods are used to investigate their properties. An example is given to show how these estimators can be used to compare several contingency tables.

The concepts of permutation invariance and palindromic invariance are introduced to differentiate between models for nominal categories and models which are suitable for ordered categories. Log linear models are shown to be suitable for nominal categories only. An

example is given of a palindromic hierarchy of models for square tables and Stuart's (1953) distance vision data are analysed using one of the models in this hierarchy.

ACKNOWLEDGEMENTS

I am grateful to my supervisor, Dr. A.C. Atkinson, and to Professor D.R. Cox for their constructive suggestions during the course of this work. I am indebted to the Health and Safety Executive for their generous financial support during the past three years. In particular, I would like to thank Mr. P.F. Collier, Dr. A.J. Fox and Dr. M. Greenberg who guided me through the morass of medical literature on the epidemiology of respiratory diseases. In addition, I would like to thank Dr. D.G. Clayton, Mr. A.G. Davies, Dr. M. Jacobsen, Mr. A.T. Jenkinson who suggested the name palindromic symmetry for one of the models, Dr. P.D. Oldham, Mr. C.E. Rossiter, and Mr. A.G. Tebbutt for their constructive comments at various stages.

This thesis was typed by Mrs. M.Robertson.

CONTENTS

TABLES

FIGURES

Chapter 1

INTRODUCTION

Ordered categorised data play an important role in a wide variety of areas where the measurements or observations are based on subjective assessments. Examples in psychology include the subjective assessment of one's mental attitude to work, sport etc. In human geographical studies the categories may be social classes which are usually considered to be ordered, or locations which are unordered. In the pharmaceutical industry, drugs for the alleviation of tension, depression, etc. have their effects measured on an ordered scale, e.g. no effect, slight improvement, marked improvement, complete recovery. In medicine and in the epidemiology of chest diseases, patients are assessed for the severity of disease on the basis of the doctor's assessment of the patient's radiograph. These assessments are usually on an ordered scale.

Despite the wide variety of applications there has been relatively little consideration given to the problem by statisticians. Some notable exceptions include Pearson (1901) who devised numerical scales for colours to help explain coat-colour inheritance in thoroughbred horses. Fisher (1963), pp.289-295, used an optimum scoring procedure in a designed experiment to investigate the reaction of blood samples tested with different sera. His procedure is optimal in the sense that the scores chosen maximise the multiple correlation coefficient, or equivalently they maximise the treatment sum of squares for a given total sum of squares. Various other optimal scoring procedures have appeared in the statistical-psychological literature since then. The main problem with Fisher's method is that the estimated scores may not be monotone and several non-linear programming procedures have

been devised to cope with this problem (Bradley et al. (1962), Nishisato (1975) and others). Other scoring methods in common use include integer scoring (Jacobsen, 1975) and scores derived from the normal distribution (Wise and Oldham, 1963). These scoring methods are not considered further in this thesis since (i) the scores are often difficult to interpret and (ii) the distributions of derived test statistics are very complicated.

The point of view taken in this thesis is that ordered categorised data arise from a partition of an underlying continuum. This was essentially Pearson's view and led him to develop his tetrachoric correlation coefficient for the 2×2 contingency table. Consideration of the underlying continuous variable obviates the problem of choosing category scores. Instead the parameters of interest are the category boundaries. These ideas led to the work of Ashford (1959a, 1959b), Snell (1964), Plackett (1965), Clayton (1974) and Simon (1974). Their work is discussed more fully in Chapters 2 and 3. For a brief review see Fienberg (1975).

Much of the work in this thesis is seen as a bridge between the standard methods for the analysis of binary data (Cox, 1970) and the usual linear models for the analysis of continuous data. The models suggested for ordered categories often parallel the more common log linear models which seem to dominate the modern statistical literature on contingency table analysis. In general, log linear models do not take account of category order, and are therefore more suited for analysing categories on a nominal scale.

The methods developed in this thesis concern problems where the dependent variable is on an ordered categorised scale. The explanatory variables may be continuous or on an unordered scale, i.e. a blocked

structure. A problem of interest to Bayesians arises when the dependent variable is continuous but there is prior information that the block effects are ordered. Thus, if there are three blocks whose effects are known to be ordered, we can draw strength from the observations in blocks 1 and 3 to make inferences about what might happen in block 2. The problem of ordered explanatory factors is not considered further in this thesis.

To construct a model for discrete observed data we first construct the model for the data as if the data were continuous, and use the associated model for the discrete data. Chapters 2 and 3 deal with the linear model including the two sample problem, regression, randomised block structures, etc. Some associated nonparametric tests are derived in Chapter 4. These include the Wilcoxon test, the sign test, etc. The remaining two chapters deal with the problem of matched pairs. This is particularly important in longitudinal studies of pneumoconiosis and other diseases where the same individuals are examined at intervals of a few years.

Each chapter contains an example to demonstrate the models and the problems involved in interpreting the parameters.

Chapter 2

THE TWO-SAMPLE PROBLEM

2.1 General model for ordered categories

The models for ordered categorised data described in this thesis can be considered as derived from models for continuous data. Thus, to derive a model for ordered categorised data we first express the model for the data as if it were continuous, and then use the associated model for categorised data. Suppose, for example, we have a model for the continuous random variable Y_i which states that Y_i has density $f(y - \mu_i)$ where f is known. The discrete categorised random variable X_i is defined by

$$X_i = j \text{ if } \theta_{j-1} \leq Y_i < \theta_j \qquad (j = 1,\ldots,k) \qquad (2.1.1)$$

where $\{\theta_i\}$ are a set of increasing real numbers, usually unknown. Hence, using (2.1.1), there is an associated model for the discrete random variable X_i. For convenience of notation we take $\theta_0 = -\infty$ and $\theta_k = \infty$, where k is the maximum value of the discrete variable x.

Equation (2.1.1) defines a censoring mechanism with unknown censoring points. It is clear that there is less information in the discrete variables $\{X_i\}$ than in the continuous variables $\{Y_i\}$. An attempt to quantify this information loss is made in 2.6. The $\{X_i\}$ can also be regarded as a partial ranking of the $\{Y_i\}$. Some optimum rank tests for partially and completely ranked data are derived in chapter 3 for testing hypotheses concerning location and scale.

For the two sample problem, a common model for continuous data

is that observations in the first sample have density $f(\frac{y-\mu_1}{\sigma})$ whereas observations in the second sample have density $f(\frac{y-\mu_2}{\sigma})$ where f is known but μ_1 μ_2 and σ, the common scale parameter, are unknown. All observations are assumed to be independent. It is clear from (2.1.1) that, if $\{\theta_j\}$ are unknown, there is no scale information contained in $\{x_i\}$. Similarly there is no information on absolute location. The best we can hope to estimate is a scale-free parameter such as $(\mu_2 - \mu_1)/\sigma$. It is convenient therefore, to set the scale parameter $\sigma = 1$ and to set $\mu_1 = 0$ or $\mu_1 + \mu_2 = 0$. It is possible to do valid tests of the hypothesis $H : \mu_2 = \mu_1$ using an estimate of the scale-free quantity $(\mu_2 - \mu_1)/\sigma$. The consequences of unequal scales within groups are considered in §2.7 and §3.4.

There remains the question of choosing the density f which will usually not be known. For binary data, the logistic and probit or inverse normal response functions are most popular. In §2.2 it is argued that there is little difference between these two response functions, and since the logistic density is the simpler, it is chosen in preference to the normal density.

2.2 The Logistic Model for Ordered Categories

In most cases the underlying density, f, is unknown, and an appropriate density must be chosen. The usual choices are the logistic and normal densities since these have readily interpretable location and scale parameters. Tukey (1970, Ch.29) describes an alternative class of suitable distributions called "folded λ-powers" of which the logistic distribution is a special case. For binary data, methods based on the logistic and the normal distributions are called logit and probit analysis respectively. Some comparisons of logit and probit analyses are

given by Cox (1966), Chambers & Cox (1967) and Berkson (1951). These studies show that the two models are virtually indistinguishable. The logistic function is mathematically the simpler of the two and is usually preferred for the analysis of binary data. It is shown in this thesis that some properties of the logistic model for binary data can easily be extended to the many category situation.

The simplest model for the two sample problem is a direct analysis of the normal theory model.

$$f(y|\text{ sample 1}) = \exp(y + \tfrac{1}{2}\Delta)/\{1 + \exp(y + \tfrac{1}{2}\Delta)\}^2,$$

$$f(y|\text{ sample 2}) = \exp(y - \tfrac{1}{2}\Delta)/\{1 + \exp(y - \tfrac{1}{2}\Delta)\}^2.$$

The associated model for the discrete variable X is most easily expressed in terms of the cumulative probabilities γ_{1j}, γ_{2j} for $j = 1,\ldots,k-1$ by

$$\begin{aligned} \text{pr}(X \leq j|\text{ sample 1}) &= \gamma_{1j} = \exp(\theta_j + \tfrac{1}{2}\Delta)/\{1 + \exp(\theta_j + \tfrac{1}{2}\Delta)\}, \\ \text{pr}(X \leq j|\text{ sample 2}) &= \gamma_{2j} = \exp(\theta_j - \tfrac{1}{2}\Delta)/\{1 + \exp(\theta_j - \tfrac{1}{2}\Delta)\}. \end{aligned} \tag{2.1.1}$$

The model (2.2.1) is conveniently expressed in the cumulative logit linear form

$$\begin{aligned} \ln\{\gamma_{1j}/(1 - \gamma_{1j})\} &= \theta_j + \tfrac{1}{2}\Delta \\ \ln\{\gamma_{2j}/(1 - \gamma_{2j})\} &= \theta_j - \tfrac{1}{2}\Delta. \end{aligned} \tag{2.2.2}$$

An equivalent version of (2.2.2) is

$$\frac{\gamma_{1j}(1 - \gamma_{2j})}{\gamma_{2j}(1 - \gamma_{1j})} = e^{\Delta}, \qquad (2.2.3)$$

where e^{Δ} is the odds ratio of the event $X \leq j$ in the two populations.

The cumulative logit linear model for ordinal data as expressed in (2.2.2) is the direct analogue of the more common log linear models for nominal data which do not take account of the category order.

Snell (1964) considers a more general version of the model (2.1.1) where there are more than two samples. She uses approximate likelihood estimators using estimated category scores in addition to the exact maximum likelihood estimators. We consider this generalised model in Chapter 3. For a similar model based on the integrated normal response function see Ashford (1959a, 1959b) and Aitchison and Silvey (1957).

2.3 The Likelihood Equations

For a general contingency table the log likelihood is

$$\ell = \Sigma\, n_{ij} \ln(\pi_{ij}) + \text{const.} \qquad (2.3.1)$$

where the summation extends over all the indices, n_{ij} is the cell count for the (i,j) cell and $\{\pi_{ij}\}$ are the cell probabilities. For the logistic model described in §2.2 these cell probabilities are given by

$$\pi_{1j} = F(\theta_j + \tfrac{1}{2}\Delta) - F(\theta_{j-1} + \tfrac{1}{2}\Delta) = \gamma_{1,j} - \gamma_{1,j-1},$$

$$\pi_{2j} = F(\theta_j - \tfrac{1}{2}\Delta) - F(\theta_{j-1} - \tfrac{1}{2}\Delta) = \gamma_{2,j} - \gamma_{2,j-1},$$

where $F(x) = \exp(x)/(1 + \exp(x))$. The unknown parameters are $\{\theta_j\}$, $j = 1,\ldots,k-1$ and Δ.

A useful property of the logistic function is that its derivatives $F'(x)$, $F''(x)$ etc., are expressible as polynomials in $F(x)$. The first two derivatives are

$$F'(x) = f(x) = F(x)\{1 - F(x)\}$$

and

$$F''(x) = f'(x) = F(x)\{1 - F(x)\}\{1 - 2F(x)\}$$

The derivatives of the log likelihood (2.3.1) with respect to the unknown parameters are

$$\frac{\partial \ell(\Delta,\underline{\theta})}{\partial \Delta} = \frac{1}{2}\sum_{j=1}^{k} n_{1j}\,\pi_{1j}(1 - \gamma_{1j} - \gamma_{1j-1}) - \frac{1}{2}\sum_{j=1}^{k} n_{2j}\,\pi_{2j}(1 - \gamma_{2j} - \gamma_{2j-1}) \tag{2.3.2}$$

and

$$\frac{\partial \ell(\Delta,\underline{\theta})}{\partial \theta_j} = (n_{1j} - n_{1j+1})\gamma_{1j}(1 - \gamma_{1j}) + (n_{2j} - n_{2j+1})\gamma_{2j}(1 - \gamma_{2j}) \qquad (j = 1,\ldots,k-1)$$

In the above derivatives, π_{ij} and γ_{ij} are considered as functions of $\underline{\theta}$ and Δ. Except when $k = 2$, the likelihood equations cannot be solved analytically. They are, however, easy to solve iteratively using the estimates in §2.4 as starting values.

In many cases, interest centres on small to medium values of Δ. Formally, therefore, we construct a null hypothesis $H_0 : \Delta = 0$ which can be considered as the point of division between two qualitatively different possibilities, $\Delta > 0$ and $\Delta < 0$. In an experiment we may wish

to know roughly how many observations are necessary to discriminate between these two possibilities. To do this we need an estimate of the variance of $\hat{\Delta}$, the m.l.e. of Δ. In particular we need an estimate of this variance for small values of Δ.

The elements of the information matrix $I_{\Delta\theta}$, evaluated at $\Delta = 0$, are

$$-E\left(\frac{\partial^2 \ell}{\partial \Delta^2}\right) = \frac{n}{4} \Sigma \gamma_j (1 - \gamma_j)(\pi_j + \pi_{j+1})$$

$$-E\left(\frac{\partial^2 \ell}{\partial \Delta \partial \theta_j}\right) = \frac{1}{2}(n_1 - n_2)\gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}) \quad (j = 1,\ldots,k-1)$$

$$-E\left(\frac{\partial^2 \ell}{\partial \theta_j^2}\right) = n\gamma_j^2(1 - \gamma_j)^2\left(\frac{1}{\pi_j} + \frac{1}{\pi_{j+1}}\right) \quad (j = 1,\ldots,k-1)$$

and

$$-E\left(\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_{j+1}}\right) = -n\gamma_j\gamma_{j+1}(1 - \gamma_j)(1 - \gamma_{j+1})\frac{1}{\pi_{j+1}}, \quad (j = 1,\ldots,k-2)$$

where $\gamma_j = F(\theta_j)$, $\pi_j = \gamma_j - \gamma_{j-1}$, $n_1 = \sum_j n_{1j}$, $n_2 = \sum_j n_{2j}$ and $n_1 = n_1 + n_2$. All other second derivatives are zero.

The asymptotic variance of $\hat{\Delta}$ is given by the (1,1) element of $I_{\Delta\underset{\sim}{\theta}}^{-1}$. To evaluate this element, partition $I_{\Delta\underset{\sim}{\theta}}$ as follows.

$$I_{\Delta\underset{\sim}{\theta}} = \begin{pmatrix} c & \underset{\sim}{d}^T \\ \underset{\sim}{d} & \underset{\sim}{J} \end{pmatrix} \begin{matrix} 1 \\ (k-1) \end{matrix},$$

where $c = -E\left(\frac{\partial^2 \ell}{\partial \Delta^2}\right)$ is a scalar, $\underset{\sim}{d} = -E\left(\frac{\partial^2 \ell}{\partial \Delta \partial \underset{\sim}{\theta}}\right)$ and $J = -E\left(\frac{\partial^2 \ell}{\partial \theta \partial \theta^T}\right)$, all being evaluated at $\Delta = 0$. The matrix $\underset{\sim}{J}$ is in fact a symmetric Jacobi matrix which is more readily recognisable in its standard form $\underset{\sim}{J}_1$ where

$$\underset{\sim}{J} = n \underset{\sim}{D} \underset{\sim}{J}_1 \underset{\sim}{D}$$

where $\underset{\sim}{D} = \text{diag}\{\gamma_i(1 - \gamma_i)\}$ and $\underset{\sim}{J}_1$ is tri-diagonal.

$$\underset{\sim}{J}_1 = \begin{Bmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_2} & -\frac{1}{\pi_2} & 0 & & & \\ -\frac{1}{\pi_2} & \frac{1}{\pi_2} + \frac{1}{\pi_3} & -\frac{1}{\pi_3} & & 0 & \\ 0 & -\frac{1}{\pi_3} & \frac{1}{\pi_3} + \frac{1}{\pi_4} & \ddots & & \\ & & \ddots & \ddots & -\frac{1}{\pi_{k-1}} & \\ & 0 & & -\frac{1}{\pi_{k-1}} & \frac{1}{\pi_{k-1}} + \frac{1}{\pi_k} & \end{Bmatrix}$$

The inverse of $\underset{\sim}{J}_1$ is a Green's matrix $\underset{\sim}{G}_1$ (Karlin, 1968, pp.112-115). A similar pair of matrices is considered by Roy & Sarhan (1956). The elements $g_{ij}^{(1)}$ of $\underset{\sim}{G}_1$ are

$$g_{ij}^{(1)} = \begin{cases} \gamma_i(1 - \gamma_j) & (i \leq j) \\ \gamma_j(1 - \gamma_i) & (i \geq j), \end{cases}$$

so that the elements $\{g_{ij}\}$ of $\underset{\sim}{G}(= \underset{\sim}{J}^{-1})$ are

$$g_{ij} = \begin{cases} 1/\{n\gamma_j(1 - \gamma_i)\} & (i \leq j) \\ 1/\{n\gamma_i(1 - \gamma_j)\} & (i \geq j). \end{cases}$$

After considerable algebraic manipulation it can be shown that the asymptotic variance of $\hat{\Delta}$ is given by

$$\text{var}(\hat{\Delta}) = [\frac{n_1 n_2}{n} \sum_{j=1}^{k-1} \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1})]^{-1} + O(\Delta). \qquad (2.3.3)$$

Some alternative expressions for (2.3.3) are given in §2.4. Consideration of terms in Δ shows that the coefficient of Δ in the expression for $\text{var}(\hat{\Delta})$ depends on $n_1 - n_2$. Thus, when the experiment is balanced ($n_1 = n_2$), (2.3.3) is correct to first order in Δ, as can be seen by considering the symmetry of the problem.

Some alternatives to $\hat{\Delta}$ are considered in §2.4. These are shown to be fully efficient when the true value of Δ is small.

2.4 Two Simple Odds Ratio Estimators of Δ

In this section two estimators of Δ which can be computed directly from the data, are considered. These are shown to be approximately unbiased and asymptotically fully efficient when Δ is small. They can be used as estimators in their own right or they can be used as starting values in an iterative procedure to find the maximum likelihood estimators.

The third formulation of the logistic model (2.3.3) is in terms of the odds ratio

$$\frac{\gamma_{1j}(1 - \gamma_{2j})}{\gamma_{2j}(1 - \gamma_{1j})} = e^{\Delta} \qquad (j = 1,\ldots,k-1.)$$

This formulation suggests an estimator based on a weighted combination of the k - 1 sample odds ratios. The two estimators considered are

$$\tilde{\Delta} = \sum_{j=1}^{k-1} \tilde{w}_j \ln\left\{\frac{c_{1j}(1 - c_{2j})}{c_{2j}(1 - c_{1j})}\right\} \qquad (2.4.1)$$

with $\Sigma\tilde{w}_j = 1$, and

$$\Delta^* = \ln\left\{\frac{\sum_{j=1}^{k-1} w_j^* R_{1j}(n_2 - R_{2j})}{\sum_{j=1}^{k-1} w_j^* R_{2j}(n_1 - R_{1j})}\right\} \qquad (2.4.2)$$

where c_{1j}, c_{2j} are sample estimates of γ_{1j}, γ_{2j}; R_{1j} and R_{2j} are the cumulative sums within each group and $\tilde{w}_j$ and w_j^* are weights chosen to minimise var($\tilde{\Delta}$), var(Δ^*).

Clayton (1974) considers these estimators with γ_{1j}, γ_{2j} estimated by

$$c_{ij} = R_{ij}/n_i$$

which is the sample cumulative frequency. In this form, both $\tilde{\Delta}$ and Δ^* can be infinite. To ensure that the estimators remain finite we take instead

$$c_{ij} = (R_{ij} + \tfrac{1}{2})/(n_i + 1) \qquad (2.4.3)$$

for $\tilde{\Delta}$ and add 1/2 to both numerator and denominator of (2.4.2). The asymptotically optimum weights are

$$\tilde{w}_j \propto \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}) \qquad (2.4.4)$$

and $$w_j^* = \pi_j + \pi_{j+1} .$$

These are also estimated from the data. Thus the estimator c_j of γ_j is

$$c_j = (R_{\cdot j} + \tfrac{1}{2})/(n_\cdot + 1)$$

where the subscript . indicates summation over the relevant index, and π_j is estimated by $c_j - c_{j-1}$. Clayton (1974) showed that, with the weights (2.4.4), the asymptotic variance of both estimators is

$$\text{var}(\Delta^*) = \text{var}(\tilde{\Delta}) = \left[\frac{n_1 n_2}{n} \sum_{j=1}^{k-1} \gamma_j(1-\gamma_j)(\pi_j+\pi_{j+1})\right]^{-1} \qquad (2.4.5)$$

for small Δ. We note that (2.4.5) is the same as $\text{var}(\hat{\Delta})$, the asymptotic variance of the m.l.e. of Δ given in (2.3.3). Hence both $\tilde{\Delta}$ and Δ^* are asymptotically fully efficient when Δ is small.

Since $\tilde{\Delta}$ and Δ^* can be calculated non-iteratively they can be computed on programmable hand or desk calculators. Programs are available for calculating Δ^* on the SR56, SR52, HP25 and HP65 pocket calculators. These programs also calculate the variance by estimating the parameters of (2.4.5).

There are many equivalent versions of the expression (2.4.5). These arise because of the relation $\gamma_j - \gamma_{j-1} = \pi_j$. Some equivalent forms are:

$$\text{(i)} \quad \sum_{j=1}^{k-1} \gamma_j(1-\gamma_j)(\pi_j+\pi_{j+1}),$$

$$\text{(ii)} \quad \sum_{j=1}^{k} \pi_j(1-\gamma_j-\gamma_{j-1})^2,$$

$$\text{(iii)} \quad \sum_{j=1}^{k-1} \gamma_j\gamma_{j+1}\pi_{j+1}, \qquad (2.4.6)$$

$$\text{(iv)} \quad \sum_{j=1}^{k-1} (1-\gamma_j)(1-\gamma_{j-1})\pi_j,$$

$$\text{(v)} \quad 1/3 - \frac{1}{3}\sum_{j=1}^{k} \pi_j^3 .$$

As before, γ_0 and γ_k are defined to be 0 and 1 respectively.

A typical proof of the equality of expressions (i)-(iv) involves expanding and rearranging terms. We give here a proof of (v) = (iii), which is a little different.

To prove (v) = (iii) we proceed as follows.

$$1 \equiv \left(\sum_{j=1}^{k} \pi_j\right)^3 = \sum_{j=1}^{k} \pi_j \left\{\sum_{j=1}^{k} \pi_j^2 + 2 \sum_{i\,j} \pi_i \pi_j\right\}$$

$$= \sum_{j=1}^{k} \pi_j^3 + \sum_{i \neq j}^{k} \pi_i \pi_j^2 + 2 \sum_{i \neq j}^{k} \pi_i \pi_j \gamma_{j-1} + 2 \sum_{j=1}^{k} \gamma_{j-1} \pi_j^2$$

$$= \Sigma \pi_j^3 + 3 \sum_{j=1}^{k} \gamma_{j-1} \pi_j^2 + \sum_{j=1}^{k} (1 - \gamma_j) \pi_j^2 + 2 \sum_{j=1}^{k} \gamma_{j-1}^2 \pi_j$$

$$+ 2 \sum_{j=1}^{k} (1 - \gamma_j) \gamma_{j-1} \pi_j \, .$$

Hence

$$1 - \sum_{j=1}^{k} \pi_j^3 = 3 \sum_{j=1}^{k} \gamma_{j-1} \pi_j^2 + 3 \sum_{j=1}^{k} \gamma_{j-1}^2 \pi_j$$

$$= 3 \sum_{j=1}^{k} \gamma_j \gamma_{j-1} \pi_j$$

$$= 3 \sum_{j=1}^{k-1} \gamma_j \gamma_{j+1} \pi_{j+1} \, .$$

Expressions (2.4.4) are important as they arise in various guises throughout this thesis where the logistic distribution is used.

Note that for binary data ($k = 2$), $\tilde{\Delta} = \Delta^*$ and they are both approximately equal to the maximum likelihood estimator $\hat{\Delta}$. For binary data

$$\hat{\Delta} = \ln \left(\frac{n_{11} n_{22}}{n_{12} n_{21}}\right)$$

which is the log of the cross ratio. Various modifications of $\hat{\Delta}$ have been suggested to cope with zero frequencies. The most common modification is

$$\tilde{\Delta} = \ln\left\{\frac{(n_{11}+\frac{1}{2})(n_{22}+\frac{1}{2})}{(n_{12}+\frac{1}{2})(n_{21}+\frac{1}{2})}\right\}$$

which is unbiased except for terms of $O(n^{-2})$. For a discussion of these modifications see Gart and Zweifel (1967) and Plackett (1974, pp.38-40).

2.5 Bias of Odds Ratio Estimators

It is clear that the estimator $\tilde{\Delta}$ is asymptotically unbiased since it is a weighted sum of asymptotically unbiased estimators of Δ. However Δ^* is not of this form, but is analogous to the Mantel-Haenszel method for combining odds ratios from several 2×2 tables (Mantel and Haenszel, 1959). We now prove a general result which shows that this type of estimator is asymptotically unbiased under fairly mild restrictions.

Let X_j/Y_j $(j = 1,\ldots,k)$ be a sequence of ratio estimators of a parameter μ such that, asymptotically, $E(\frac{X_r}{Y_r}) = \mu < \infty$, $Y_r \geq 0$ and $\mathrm{Var}(\frac{X_r}{Y_r}) = O(n^{-1})$. Let w_j $(j = 1,\ldots,k)$ be a set of fixed positive weights. Then

$$\psi = \sum_{j=1}^{k} w_j X_j \Big/ \sum_{j=1}^{k} w_j Y_j$$

is asymptotically $(n \to \infty)$ unbiased for μ.

Proof:

$$E(\psi) = E\{\sum_{j=1}^{k} w_j X_j \Big/ \sum_{j=1}^{k} w_j Y_j\} = E\{\sum_{j=1}^{k} w_j^* X_j/Y_j\}$$

where $w_j^* = w_j Y_j / \Sigma w_j Y_j$. We need to show that the covariance of w_j^* and the ratio X_j/Y_j is weak enough for the bias to tend towards zero. Expanding further we get

$$E(\psi) = \mu + \Sigma_{i,j} \operatorname{cov}(w_i^*, X_j/Y_j)$$

$$= \mu + \Sigma_{i,j} \rho_{ij} [\operatorname{var}(w_i^*) \operatorname{var}(X_j/Y_j)]^{1/2}$$

where ρ_{ij} is the correlation of w_i^* and X_j/Y_j. Since w_j^* is bounded between 0 and 1 it follows that $\operatorname{var}(w_j^*)$ is bounded between 0 and 1/4. Since, by assumption, $\operatorname{var}(X_j/Y_j) = O(\frac{1}{n})$ the double sum is of the order $n^{-1/2}$ and tends to zero. Hence ψ is asymptotically unbiased for μ.

It is easy to see, however, that Mantel-Haenszel type estimators are biased in small samples since the weights w_j^* are negatively correlated with the ratio X_j/Y_j. In the above proof, it is sufficient that $X_j/Y_j = \mu + O_p(1)$. Then it follows that $\psi = \mu + O_p(1)$, although the moments of ψ may be infinite.

From the above result it follows that Δ^* is asymptotically unbiased. Some simulation results are given in §2.8 to demonstrate its behaviour in small samples.

2.6 Loss of Information Due to Grouping

One way of measuring the relative efficiency of two experimental designs is to use the ratio of the number of observations required under the different designs to achieve the same precision of estimation. The absolute efficiency can then be defined by comparing a particular design with the best possible design. This idea is the basis of such statistical measures as Pitman efficiency. A measure of inefficiency

is therefore a measure of the information lost through using a sub-optimal design.

For the two sample problem we consider Δ to be the sole parameter of interest. We may be interested in estimation or testing the hypothesis $H_0 : \Delta = 0$. From (2.4.3) or (2.3.3) the variance of an estimator of Δ is inversely proportional to

$$\sum_{j=1}^{k-1} \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}) \ . \qquad (2.6.1)$$

We consider alternative designs where (a) the category boundaries $\{\theta_j\}$ can be chosen freely but the number of categories k is fixed, and (b) both $\underset{\sim}{\theta}$ and k are allowed to vary. The 'best' design when the category boundaries can be chosen is one which maximises (2.6.1). It is easily shown that the design which achieves this maximum has $\pi_j = 1/k \ (j = 1,\ldots,k)$, or equal cell probabilities and the maximum value of (2.6.1) is $\frac{1}{3}(k^2 - 1)/k^2$. The ratio of the number of observations necessary to achieve a given accuracy is inversely proportional to the ratio of the variances, so that the asymptotic efficiency of a design with category boundaries $\theta_1,\ldots\theta_{k-1}$ relative to the best design with k categories is

$$\{1 - \sum_{j=1}^{k} \pi_j^3\} \ \frac{k^2}{(k^2 - 1)} \qquad (2.6.2)$$

where $\pi_j = F(\theta_j) - F(\theta_{j-1})$.

When the alternative designs have an arbitrary number of categories or when the continuous variables are observable, the asymptotic efficiency of a given design relative to the optimum is $1 - \sum_{j=1}^{k} \pi_j^3$ which can be obtained from (2.6.2) by putting $k = \infty$.

In radiological data, for example, typical category frequencies are (0.9, 0.04, 0.04, 0.02). In principle, though extremely difficult in practice, the category boundaries could be redefined to give a more uniform set of frequencies (1/4, 1/4, 1/4, 1/4). The efficiency of the former design relative to the latter is approximately 0.3. Thus, for every 3 observations required under the re-defined scheme, 10 are required under the old scheme. In fact the equal frequency scheme for four categories is almost as efficient (0.94) as a scheme which has no censoring mechanism at all.

The importance of expression (2.4.6) is that they measure the efficiency of the censoring scheme. In fact they are direct generalisations of the binomial variance formula $p(1 - p)$. We note that for well chosen values of the censoring points $\{\theta_j\}$ and hence $\{\pi_j\}$ the asymptotic efficiency relative to the optimum under (b) is high even when k is small. The efficiency is $(k^2 - 1)/k^2$ (= .75, .89, .94, .96...) for the first few values of k.

2.7 Adequacy of Model

The model described in this chapter makes two assumptions which are (a) a logistic error distribution and (b) equal variances within each sample. Assumption (a) may break down, but Chambers & Cox (1967) have shown that the difference between a logistic and a normal error function is very difficult to detect even in fairly large samples. Thus it is likely that only gross departures from (a) can be detected. We concentrate on models which allow departures from assumption (b).

Two approaches are possible. The first involves testing for heteroscedasticity using one of the non-parametric tests in Chapter 4, and then fitting the logistic model if there is no evidence of heteroscedasticity.

The second approach is to fit a more general model which makes allowance for heteroscedasticity. The more general model can be written in cumulative logit linear form as

$$\begin{aligned} \ln\{\gamma_{1j}/(1-\gamma_{1j})\} &= \tau_1(\theta_j + \tfrac{1}{2}\Delta) \\ \ln\{\gamma_{2j}/(1-\gamma_{2j})\} &= \tau_2(\theta_j - \tfrac{1}{2}\Delta) \end{aligned} \qquad (2.7.1)$$

where, without loss of generality, $\tau_1 = 1$. Interpretation of the parameter Δ in model (2.7.1) is complicated by the unequal within groups variance, but if τ_2 is close to 1, it is reasonable to interpret Δ as an approximate log odds ratio.

When all the parameters are estimated by maximum likelihood, or by any other suitable method, the goodness of fit of (2.7.1) can be tested by a X^2 statistic on $k - 3$ degrees of freedom. The homoscedastic model (2.2.2) can be tested by a X^2 statistic on $k - 2$ degrees of freedom.

It is impossible to give general guidelines about how to proceed when neither model fits the data. In the example of §3.4, consideration of the residuals indicates that one of the groups is an 'outlier'.

2.8 Simulation Results for the Estimators $\tilde{\Delta}$, Δ^*.

We present briefly some simulation results for the estimators $\tilde{\Delta}$ and Δ^*. These are not intended to be in any way comprehensive, but it is hoped that they give an indication of bias and variability of Δ^* and $\tilde{\Delta}$ in medium sized samples.

Four category data was chosen with equal sample sizes and roughly equal frequencies for each category. A slight adjustment of the results is necessary since the data were generated from a logistic distribution

with unit variance as opposed to the standard logistic distribution which has variance $\pi^2/3$. A displacement δ on this scale is equivalent to $\Delta = \delta\pi/\sqrt{3} = 1.814\ \delta$. The estimators $\tilde{\delta}$ and δ^* are scaled versions of $\tilde{\Delta}$ and Δ^* respectively. Some general conclusions are as follows.

(a) Both Δ^* and $\tilde{\Delta}$ are biased towards the origin in small samples.

(b) The bias of Δ^* is approximately twice that of $\tilde{\Delta}$.

(c) The bias is negligible when $n_1,\ n_2 \geq 100$ and is of the order of 5% when $n_1 = n_2 = 20$.

The conclusion (c) will depend on the number of categories, so that the bias could be larger than 5% when $n_1 = n_2 = 20$, if the twenty observations were divided into, say, 10 categories. When $n_1 \neq n_2$ the bias could be expected to be of the order $\frac{1}{n_1} + \frac{1}{n_2}$, so for the bias to be small we need both n_1 and n_2 large.

The simulation results for small samples ($n_1 = n_2 = 20$) are given in table 2.1 and the corresponding results for large samples in table 2.2. For both sample sizes the entries in columns 2 and 4 are the mean of 1000 repetitions at each value of δ. Columns 3 and 5 give the standard deviation of the 1000 repetitions.

The standard deviation as estimated by (2.4.5) was relatively constant but increased from 0.314 for $\delta = 0$ to 0.323 for $\delta = 1.9$ for the small samples. Thus (2.4.5) is a slight underestimate for larger values of Δ. Similarly, for the large samples the estimate of Std $(\tilde{\Delta})$ (2.4.5) increased from 0.142 to 0.145 over the same range, and is smaller than the true standard deviation when δ is large.

Table 2.1

Small sample simulation results for $\tilde{\Delta}$ and Δ^*; $k = 4$, $n_1 = n_2 = 20$, 1000 repetitions for each value of δ.

δ	$\bar{\tilde{\delta}}$ +	Std $(\tilde{\delta})$	$\bar{\delta}^*$ +	Std (δ^*)
0	-.004	.308	-.003	.295
.1	.111	.319	.107	.306
.2	.204	.318	.195	.304
.3	.283	.326	.273	.313
.4	.404	.313	.389	.299
.5	.491	.325	.473	.314
.6	.600	.335	.578	.320
.7	.677	.325	.654	.312
.8	.783	.333	.752	.316
.9	.894	.348	.861	.331
1.0	.980	.349	.944	.328
1.1	1.075	.347	1.034	.324
1.2	1.191	.355	1.147	.333
1.3	1.260	.370	1.215	.347
1.4	1.355	.367	1.305	.342
1.5	1.437	.383	1.382	.353
1.6	1.541	.371	1.484	.340
1.7	1.603	.375	1.541	.345
1.8	1.687	.384	1.617	.342
1.9	1.788	.404	1.704	.347

$$\delta = \Delta\sqrt{3}/\pi$$
$$\tilde{\delta} = \tilde{\Delta}\sqrt{3}/\pi$$
$$\delta^* = \Delta\sqrt{3}/\pi$$

+Average of 1000 estimates of δ.

Table 2.2

Large sample simulation results for $\tilde{\Delta}$ and Δ^*; $k = 4$, $n_1 = n_2 = 100$
1000 repetitions for each value of δ.

δ	$\bar{\tilde{\delta}}$ [+]	Std($\tilde{\delta}$)	$\bar{\delta}^*$ [+]	Std(δ^*)
0	-.010	.144	-.010	.143
.1	.103	.148	.103	.147
.2	.193	.142	.191	.141
.3	.298	.143	.295	.141
.4	.394	.137	.391	.135
.5	.488	.142	.484	.141
.6	.590	.147	.584	.145
.7	.693	.155	.686	.152
.8	.804	.156	.796	.154
.9	.900	.161	.891	.159
1.0	.996	.164	.986	.161
1.3	1.294	.179	1.280	.175
1.6	1.601	.191	1.583	.186
1.9	1.880	.207	1.859	.201

$$\delta = \Delta\sqrt{3}/\pi$$

$$\tilde{\delta} = \tilde{\Delta}\sqrt{3}/\pi$$

$$\delta^* = \Delta^*\sqrt{3}/\pi$$

[+]Average of 1000 estimates of δ.

FIG. 2.1

Graph of simulation results for unpaired estimators $\tilde{\Delta}, \Delta^*$ in small samples. (see table 2.1)

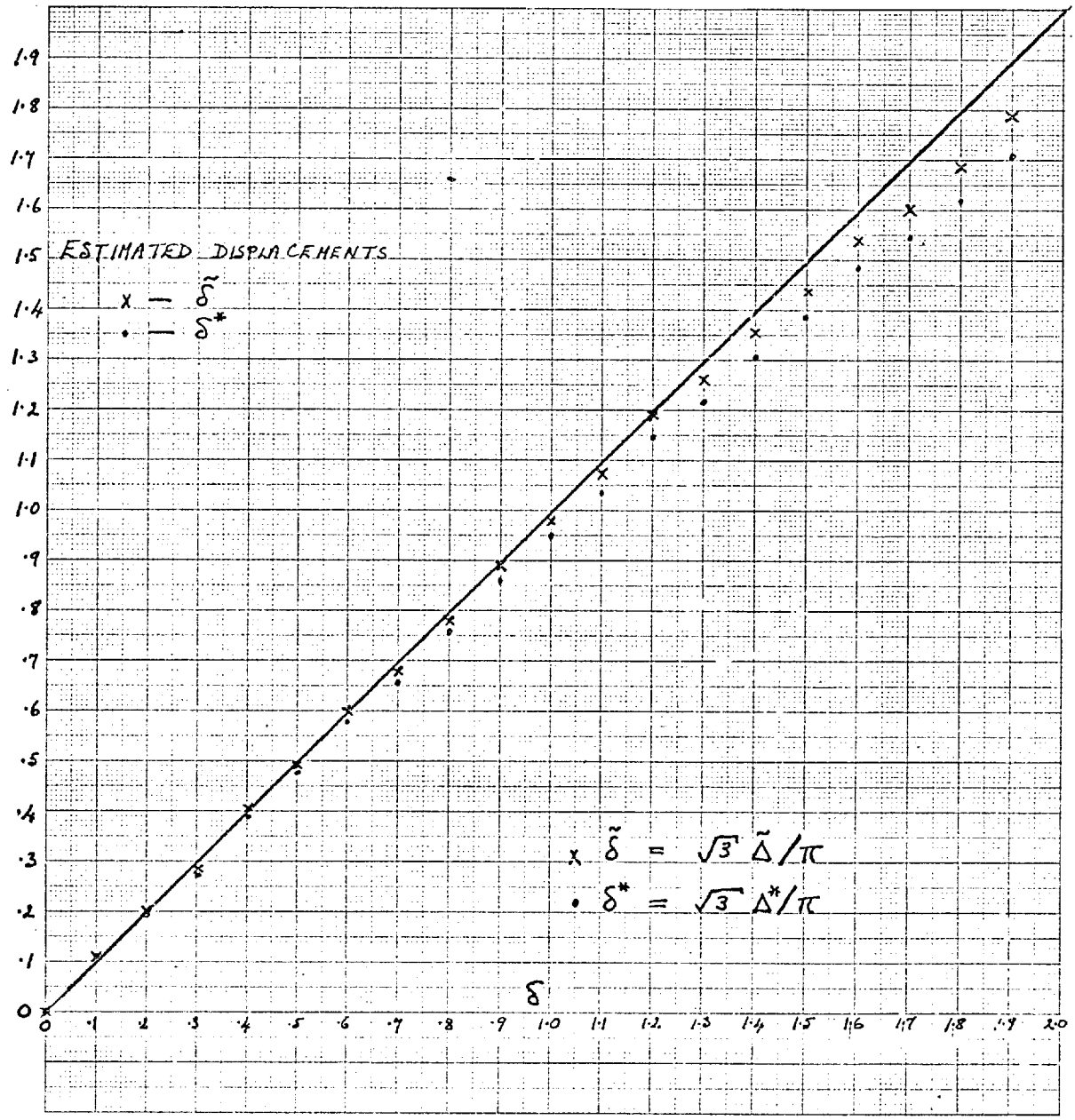
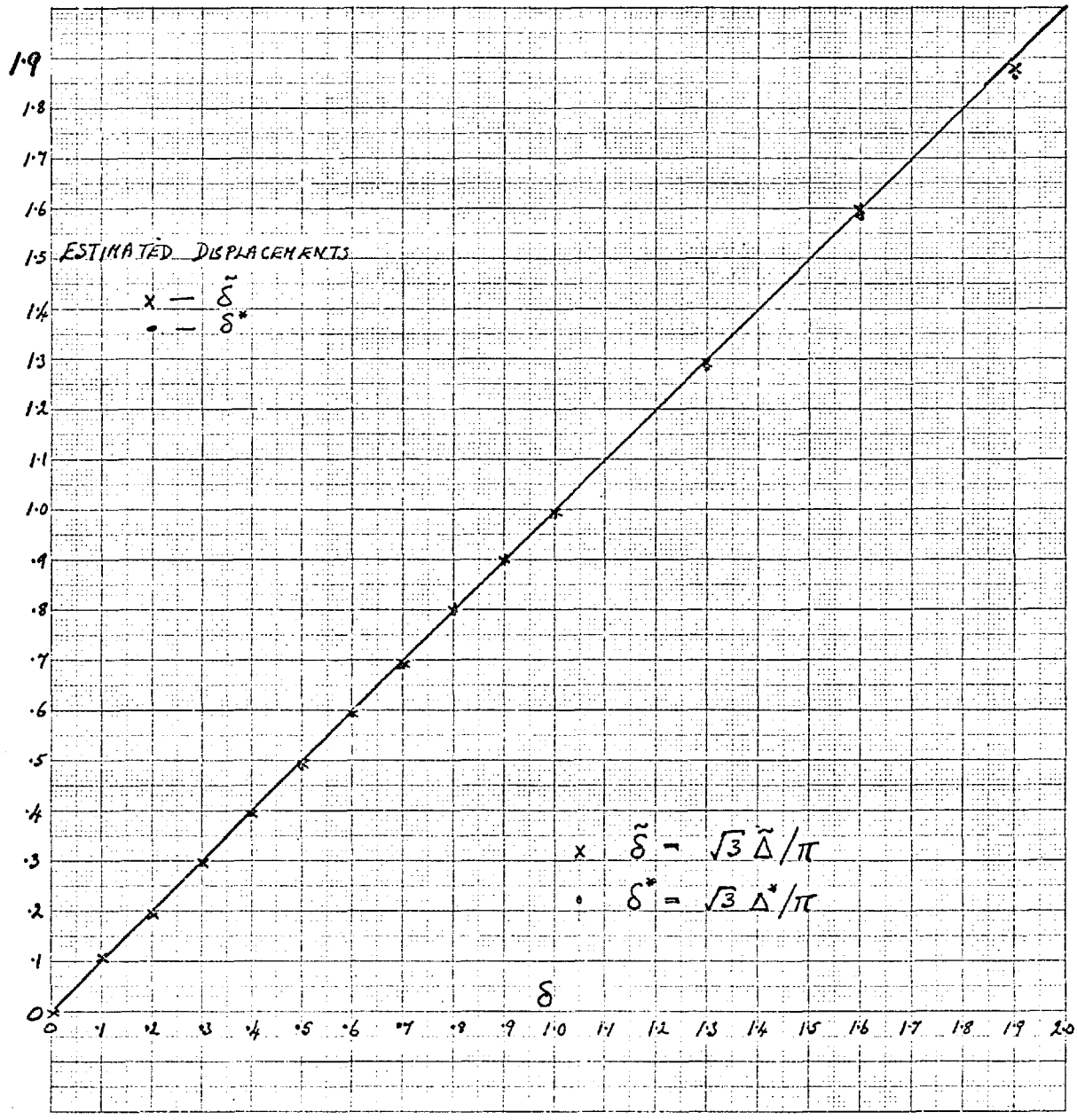


FIG. 2.2

Graph of simulation results for unpaired estimators $\tilde{\Delta}, \Delta^*$ in large samples. (see table 2.2)



2.9 Example

The data in table 2.3 from Holmes & Williams (1954) are a comparison of the tonsil sizes of carriers and non-carriers of Streptococcus Pyrogenes. Tonsil size is measured on an ordered three-category scale.

Table 2.3

Tonsil size of carriers and non-carriers of Streptococcus pyrogenes

Tonsil size	carriers	non-carriers	total
normal	19	497	516
enlarged	29	560	589
greatly enlarged	24	269	293
TOTAL	72	1326	1398

From Holmes & Williams (1954).

The estimates Δ^* and $\tilde{\Delta}$ are .580 and .565 respectively with estimated standard deviation .225, thus indicating that tonsil sizes in carriers are larger than in non-carriers.

To check the adequacy of the cumulative logit linear model (2.2.2) the complete model was fitted by maximum likelihood. The inverse matrix of second derivatives at the maximum gives an estimate for $\mathrm{var}(\hat{\Delta})$ which can be compared with the estimated variance of $\tilde{\Delta}$ and Δ^*. The m.l.e.'s of the parameters with their standard deviations and covariance matrix are

$$\hat{\Delta} = .603 \pm .226$$

$$\hat{\theta}_1 = -.810 \pm .118 \,, \qquad \underset{\sim}{V} = \begin{pmatrix} .051 & & \\ -.024 & .014 & \\ -.023 & .012 & .015 \end{pmatrix} .$$

$$\hat{\theta}_2 = 1.061 \pm .122$$

With the above estimates the X^2 goodness of fit statistic on one degree of freedom is 0.30, indicating a good fit. For alternative analyses of this data see Clayton (1974) and Armitage (1971) who uses a method based on partitioning the total X^2 goodness of fit statistic. A further analysis of the same data is given in §4.

Chapter 3

THE GENERAL CUMULATIVE LOGIT LINEAR MODEL

3.1 Introduction

The linear logistic model (2.2.2) for the two sample problem has an obvious extension to the many sample situation. The subscript i is used to denote the sample and j to denote the category. We consider first a saturated model where there are r row parameters $\{\alpha_i\}$ to explain the differences between the r samples. An unsaturated model is one where the r row effects are explained by $s < r$ parameters. The saturated linear logistic model can be written

$$\ln\{\gamma_{ij}/(1-\gamma_{ij})\} = \alpha_i + \theta_j \qquad (3.1.1)$$

$$(i = 1,\ldots,r;\ j = 1,\ldots,k-1)$$

where γ_{ij} is the theoretical cumulative frequency in the i^{th} row. To avoid redundancies in the parameters it is convenient to impose an estimability condition in (3.1.1) such as $\sum_{i=1}^{r} \alpha_i = 0$. The parameters θ_j are the category boundaries on the logit scale. Ashford (1959a, 1959b) considers a similar model with an integrated normal response function.

We may wish to explain the differences between rows in terms of some explanatory variables $\underset{\sim}{x}_i = (x_{i1}\ldots x_{is})$, with $s < r$. The linear logistic model is

$$\ln\{\gamma_{ij}/(1-\gamma_{ij})\} = \underset{\sim}{\beta}^T \underset{\sim}{x}_i + \theta_j \qquad (3.1.2)$$

where $\underset{\sim}{\beta} = (\beta_1 \ldots \beta_s)^T$ are unknown parameters. Since $s < r$ no estimability constraints are necessary provided that the $s \times s$ matrix $\sum_{i=1}^{r} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})^T$, where $\bar{\underset{\sim}{x}} = \frac{1}{r} \sum_{i=1}^{r} \underset{\sim}{x}_i$, is nonsingular. Equivalently, the matrix $\underset{\sim}{X} = (\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_r, \underset{\sim}{1})$ must be of full rank s where 1 is the unit vector of length s.

Simon (1974) considers the saturated model (3.1.1) and suggests an iterative procedure for estimating the parameters. In §3.2 a generalised empirical logistic transform is used to estimate the parameters $\{\alpha_i\}$ in (3.1.1) or alternatively the parameter β in (3.1.2). This transform is the analogue of the empirical logistic transform for binary data. These estimates can then be used as starting values in an iterative procedure for obtaining maximum likelihood estimates of the parameters.

3.2 A Generalised Empirical Logistic Transform

The empirical logistic transform (Cox, 1970, p.78ff) is useful for analysing binary data since it transforms the frequency of success R_i in n_i trials to a variable Z_i' which is approximately normally distributed with simple mean and variance V_i' which can be estimated. There are two alternative definitions of the empirical logistic transform but the difference between them is relatively unimportant and we consider only the definition:

$$Z_i' = \ln\{(R_i + \tfrac{1}{2})/(n_i - R_i + \tfrac{1}{2})\} \qquad (3.2.1)$$

$$V_i' = \frac{(n_i + 1)(n_i + 2)}{n_i(R_i + 1)(n_i - R_i + 1)} \; . \qquad (3.2.2)$$

We note that for large values of n_i and R_i, V_i' is approximately $(n_i p_i q_i)^{-1}$ where p_i is the probability of success in the i^{th} group. The analysis now proceeds using the transformed variables Z_i' which are assumed to be approximately normally distributed with known variance V_i'. The model can be fitted using iterative weighted least squares and the scaled residuals tested for normality (Cox, 1970, pp.81-83).

We now examine a generalisation of (3.2.1) and (3.2.2) to multi-category data. This generalisation is an extension of the estimator $\tilde{\Delta}$ of hapter 2. The alternative estimator, Δ^* does not have an analogue when there are more than two groups.

The estimator $\tilde{\Delta}_{12}$ of the logistic difference between the first two groups can be written as

$$\tilde{\Delta}_{12} = \sum_{j=1}^{k-1} \tilde{w}_j \ln\left\{\frac{R_{1j} + \frac{1}{2}}{n_1 - R_{1j} + \frac{1}{2}}\right\} - \sum_{j=1}^{k-1} \tilde{w}_j \ln\left\{\frac{R_{2j} + \frac{1}{2}}{n_2 - R_{2j} + \frac{1}{2}}\right\},$$

where R_{ij} is the cumulative sum of the observations in the i^{th} group. Note that this is the difference between a function of the observations in group 1 and the same function of the observations in group 2. This contrasts with Δ^*_{12} which can be written as

$$\Delta^*_{12} = \ln\left\{\frac{1}{2} + \sum_{j=1}^{k-1} w_j^* R_{1j}(n_2 - R_{2j})\right\} - \ln\left\{\frac{1}{2} + \sum_{j=1}^{k-1} w_j^* R_{2j}(n_1 - R_{1j})\right\}.$$

Thus Δ^*_{12} cannot be expressed as the difference between a function of the observations in group 1 and the same function of the observations in group 2. One important consequence of this is that the additivity property of the parameters

$$\Delta_{12} + \Delta_{23} = \Delta_{13}$$

is satisfied by $\tilde{\Delta}$ but not by Δ^*.

The transformation

$$Z_i = \sum_{j=1}^{k-1} \tilde{w}_j \ln\left\{\frac{R_{ij} + \frac{1}{2}}{n_i - R_{ij} + \frac{1}{2}}\right\} \qquad (3.2.3)$$

is called the generalised empirical logistic transform of the observations in the i^{th} group. The weights w_j in (3.2.3) are estimated from the column cumulative totals $\{R._j\}$.

$$\tilde{w}_j = (R._j + \tfrac{1}{2})(n. - R._j + \tfrac{1}{2})(R._{j+1} - R._j) \qquad (3.2.4)$$

$$\text{with} \quad \sum_{j=1}^{k-1} w_j = 1.$$

For large values of n_i the transformed variables Z_i can be treated as independent and approximately normally distributed with variance V_i estimated by

$$V_i = \left[n_i \sum_{j=1}^{k-1} c_j(1 - c_j)(c_{j+1} - c_{j-1})\right]^{-1} \qquad (3.2.5)$$

where $c_j = R._j/n..$. Some expressions which are equivalent to (3.2.5) but may be easier to calculate, are given in (2.4.6). When the differences between the various groups are large and the group totals n_i are large, an improved estimate of V_i is given by

$$V_i'' = \left[n_i^{-1} \sum_{j=1}^{k-1} \tilde{w}_j^2 (R_{ij} + 1)(n_i - R_{ij} + 1) + 2n_i^{-1} \sum_{j<\ell}^{k-1} \sum \tilde{w}_j \tilde{w}_\ell (R_{ij} + 1)\right.$$

$$\left.(n_i - R_{i\ell} + 1)\right]^{-1} \qquad (3.2.6)$$

Under the saturated model (3.1.1) the derived model for the transformed variables Z_i is

$$E(Z_i) = \mu + \alpha_i \quad (i = 1,\ldots,r) \tag{3.2.7}$$

where $\mu = \Sigma \tilde{w}_j \theta_j$ is considered to be a nuisance parameter. The unsaturated model (3.1.2) becomes

$$E(Z_i) = \mu + \underset{\sim}{\beta}^T \underset{\sim}{x}_i \quad (i = 1,\ldots,r).$$

The scaled residuals under the unsaturated model

$$(Z_i - \hat{Z}_i)/\sqrt{v_i} \quad (i = 1,\ldots,r)$$

can be tested for standard normality, outliers etc. using graphical or other methods. (Under the saturated model (3.2.7) the residuals are all zero.) Some loss of efficiency is to be expected when the transformed variables are used. The method should be used only when the number of categories is fairly small, typically 5 or less, and the cell counts are large, typically 5 or more. In such cases, the ease with which the transformed variables can be handled often outweighs the small loss of efficiency incurred.

A further important condition for the applicability of the generalised empirical logistic transform is that the scales within groups should be equal. In the example of §3.4 this condition is not satisfied and hence an alternative method of analysis described in §3.3 is used.

3.3 Adequacy of Model

We consider first tests for the adequacy of the saturated model (3.1.1) and the unsaturated model (3.1.2) when all the parameters are estimated by maximum likelihood or by any other asymptotically efficient method such as minimum chi-squared. The saturated model has r + k-1 parameters with one linear constraint and there are r linear constraints on the cell probabilities since

$$\sum_{j=1}^{k} \pi_{ij} = 1 \qquad (i = 1,\ldots,r.)$$

This leaves (r - 1)(k - 2) degrees of freedom to test for goodness of fit. For the unsaturated model (3.1.2) the equivalent degrees of freedom for goodness of fit is (r - 1)(k - 1) - s.

If there is evidence of inhomogeneity of variance such as the presence of patterns of large residuals under the linear model, then the more general multiplicative model

$$\ln\{\gamma_{ij}/(1 - \gamma_{ij})\} = \tau_i(\theta_j + \alpha_i) \tag{3.3.1}$$

or

$$\ln\{\gamma_{ij}/(1 - \gamma_{ij}) = \tau_i(\theta_j + \beta^T \underset{\sim}{x}_i) \tag{3.3.2}$$

is appropriate. It is necessary to impose a constraint on the $\{\tau_i\}$ such as $\tau_1 = 1$ or $\prod_{i=1}^{r} \tau_i = 1$.

The multiplicative models (3.3.1) and (3.3.2) each have an extra r - 1 parameters so there are (r - 1)(k - 3) degrees of freedom left to test the fit of the saturated model (3.3.1) and (r - 1)(k - 2) - s to test the fit of the unsaturated model (3.3.2).

It is impossible to give general guidelines on how to proceed if the multiplicative models do not fit. Examination of residuals often helps to identify outliers. These can then be discarded if they are errors or they may be of particular interest precisely because they are outliers and should be investigated further.

An example with unequal scales in each group is given in §3.4.

3.4 Example

This example from Bradley, Katti and Coons (1962) is a 5-treatment experiment where the observations are on a 5-category ordered scale. The categories represent the subjective responses of individuals in a food-testing experiment and range from terrible (category 1) to excellent (category 5). Bradley et al. analysed this data using an optimum scoring technique. Snell (1964) analysed the same data using a model similar to the linear logistic model (3.1.1), but not allowing for differences in scales.

A fairly general Fortran program was written to fit the models (3.1.1) and (3.1.2) with options for restricting the scale parameters or restricting the block parameters $\{\alpha_i\}$. The saturated model (3.1.1) is appropriate for this example since there is no suggestion that the five blocks (treatments) can be explained in fewer than five parameters.

Table 3.1 gives the data. The fitted values together with the residuals under the linear logistic model, $\ln\{\gamma_{ij}/(1 - \gamma_{ij})\} = \theta_j + \alpha_i$, are given in table 3.2.

The parameter estimates under the linear model are

$$\hat{\underset{\sim}{\theta}} = (-1.5694, -.5176, .3688, 2.6646) \text{ and}$$

$$\hat{\underset{\sim}{\alpha}} = (-.0464, -.5090, 1.0934, .5281, -1.0660).$$

Table 3.1

Response frequency under different treatments in a taste testing experiment

	Response category					
	terrible	poor	fair	good	excellent	
Treatment	1	2	3	4	5	Total
1	9	5	9	13	4	40
2	7	3	10	20	4	44
3	14	13	6	7	0	40
4	11	15	3	5	8	42
5	0	2	10	30	2	44

Source: Bradley et al. (1962).

Table 3.2

Fitted values and residuals* under linear logistic model

	Response category				
Treatment	1	2	3	4	5
1	6.63 (2.37)	7.87 (-2.87)	8.69 (.31)	14.08 (-1.08)	2.72 (1.28)
2	4.89 (2.11)	5.71 (-3.71)	8.86 (1.14)	18.97 (1.03)	4.57 (-.57)
3	15.33 (-1.33)	10.28 (2.72)	6.87 (-.87)	6.61 (.39)	.91 (-.91)
4	10.96 (0.04)	10.15 (4.85)	8.72 (-5.72)	10.51 (-5.51)	1.66 (6.34)
5	2.94 (-2.94)	4.55 (-2.55)	7.13 (2.87)	21.97 (8.03)	7.40 (-5.40)

$X^2 = 53.4$, $G^2 = 50.4$, both on 12 d.f.

*Residuals are differences between observed and fitted frequencies.

Table 3.3

Fitted values and residuals* under multiplicative logistic model

Treatment	Response category 1	2	3	4	5
1	7.86 (1.14)	7.88 (-2.88)	7.35 (1.65)	12.28 (.72)	4.62 (-.62)
2	4.64 (2.36)	7.41 (-4.41)	8.98 (1.02)	17.56 (2.44)	5.41 (-1.41)
3	13.79 (.21)	12.91 (.09)	7.22 (-1.22)	5.42 (1.58)	.66 (-.66)
4	13.74 (-2.74)	7.45 (7.55)	5.75 (-2.75)	9.54 (-4.54)	5.53 (2.47)
5	0.24 (-.24)	2.22 (-.22)	9.29 (.71)	30.53 (-.53)	1.73 (.27)

$X^2 = 20.8$, $G^2 = 23.2$, both on 8 d.f.

Table 3.4

Fitted values and residuals* under multiplicative logistic model, omitting treatment 4.

Treatment	Response category 1	2	3	4	5
1	8.91 (.09)	5.73 (-.73)	7.89 (1.11)	13.61 (-.61)	3.86 (.14)
2	5.76 (1.24)	5.39 (-2.39)	9.23 (.77)	19.04 (.96)	4.58 (-.58)
3	14.57 (-.57)	10.66 (2.34)	9.05 (-3.05)	5.43 (1.57)	.30 (-.30)
5	0.58 (-.58)	1.97 (.03)	9.01 (.99)	30.86 (-.86)	1.57 (.43)

$X^2 = 4.9$, $G^2 = 7.9$, both on 6 d.f.

*Residuals are differences between observed and fitted frequencies.

A large positive value for $\hat{\alpha}_i$ indicates a bias towards the lower categories and conversely a large negative value indicates a bias towards the higher categories. The residuals in table 3.2 indicate that the fit is not very good in blocks 4 and 5. To judge the significance of these residuals we can calculate the usual X^2 goodness of fit statistic, but the likelihood ratio statistic, G^2, is just as convenient and has the same asymptotic distribution as X^2. The values of the two statistics are $X^2 = 53.35$ and $G^2 = 50.36$, each on 12 degrees of freedom. Although the cell counts are not very large, it is clear that the linear model does not fit very well.

There is some evidence in the data of table 3.1, of unequal scales in each block. Consequently we try the multiplicative model $\ln\{\gamma_{ij}/(1-\gamma_{ij})\} = \tau_i(\theta_j + \alpha_i)$ with $\Sigma\alpha_i = 0$ and $\tau_1 = 1$. The fitted values and the new residuals are given in table 3.3. The parameter estimates under the multiplicative model are

$$\hat{\underset{\sim}{\theta}} = (-1.3514, -.3757, .3687, 2.0924)$$

$$\hat{\underset{\sim}{\alpha}} = (-.0567, -.4430, .8838, .3986, -.7826)$$

and $$\hat{\underset{\sim}{\tau}} = (1.0000, 1.1914, 1.3721, .7572, 2.4412).$$

The goodness of fit statistics are $X^2 = 20.84$ and $G^2 = 23.21$ on 8 degrees of freedom. The difference 53.35-20.84 = 32.51 or 50.36 - 23.31 = 27.15 on 4 degrees of freedom can be used as an approximate test for equality of scales. It is thus clear that the within groups variances are unequal.

The multiplicative model is a considerable improvement over the linear model, but there are still some large residuals especially in block 4. Furthermore the goodness of fit statistics X^2 and G^2 are larger than we might expect. (The 5% point for X^2 on 8 degrees of freedom is 15.5) The multiplicative model was, therefore, re-fitted with block 4

omitted. The fitted values, together with the new residuals, are given in table 3.4. The new parameter estimates are

$$\hat{\underset{\sim}{\theta}} = (-1.3039, -.6034, .2007, 2.1834),$$

$$\hat{\underset{\sim}{\alpha}} = (.0534, -.3279, .9465, -.6729)$$

and $\hat{\underset{\sim}{\tau}} = (1.0000, 1.1603, 1.5597, 2.1834).$

It should be noted that τ_i is inversely proportional to the standard deviation in the i^{th} group. Hence large values of $\hat{\tau}_i$ indicate little scatter while small values indicate more scatter (relative to the first group since $\hat{\tau}_1 \equiv 1$).

The goodness of fit statistics for table 3.4 are $X^2 = 4.9$ and $G^2 = 7.9$, both on 6 degrees of freedom. These indicate a satisfactory fit. There is, therefore, considerable evidence that the observations in block 4 are 'outliers' and do not conform to the pattern of the remaining 4 blocks. This discrepancy was noted by Snell (1964).

To compare the average responses $\hat{\alpha}_i$ we need a rough estimate of their variability. An estimate of the variance matrix can be obtained from the matrix of second derivatives of the log likelihood. However, the variance estimate (3.2.5) of the generalised empirical logistic transform gives an indication of the variability of $\{\hat{\alpha}_i\}$ when the scale parameters $\{\beta_i\}$ are approximately equal. Thus $\text{var}(\hat{\alpha}_i) \sim 0.3/n_i \sim 0.00714$ since the $\{n_i\}$ are approximately equal. For the purpose of comparisons, the $\hat{\alpha}_i$ can be treated as approximately independent. Since the scale estimates $\{\hat{\beta}_i\}$ are, for the most part, greater than 1 then the above estimate of variance is conservative. A conservative estimate of $\text{Std}(\hat{\alpha}_i - \hat{\alpha}_j)$ is 0.12. Thus it is clear that the treatment parameters are all significantly different from each other.

To summarise, treatment 5 is best ($\hat{\alpha}_5 = -.78$) and there is a higher consensus of opinion about this ($\hat{\tau}_5 = 2.4$) than about the other treatments. Treatment 3 is worst ($\hat{\alpha}_3 = .88$) and there is a fairly average consensus about this ($\hat{\tau}_3 = 1.4$). There is little consensus about treatment 4 ($\hat{\tau}_4 = .76$) but it rates worse than average ($\hat{\alpha}_4 = .40$).

Chapter 4

LOCALLY MOST POWERFUL RANK TESTS

4.1 Introduction

Ordered categorical data are only partially ordered. For example, in the two sample problem where the observations are laid out in a singly ordered 2×k contingency table of counts $\{n_{ij}\}$, observations in the same column are 'tied' but observations in different columns are strictly ordered. As an alternative to the modelling and estimation procedures of Chapters 2 and 3, we now consider methods whose main purpose is testing of hypotheses. For example in the two sample problem we might wish to test the two hypotheses (i) equality of location and (ii) equality of scale.

We assume a general underlying continuous density f(y). In the example of §4.4,f is assumed to be the logistic density but the theory is quite general. The locally most powerful tests are shown to be rank tests. A special case of the test statistic is the case where there are no ties. In this particular limit the tests are equivalent to the rank tests of Hájek (1962) and Hájek and Sidák (1967), who do not consider the optimum treatment of tied observations.

It should be pointed out that, while the analysis in this chapter concentrates on the two sample problem and associated tests, the methods can be generalised to the many sample problem and regression problems. In these cases the rank test statistics have asymptotically X^2 distributions instead of the normal distributions encountered in the two sample problem.

4.2 Derivation of Locally Most Powerful Tests

We consider tests for hypotheses concerning only a single parameter. For the two sample problem we are particularly interested in hypotheses concerning location and scale, and these two cases are considered separately. The relevant null hypothesis is that the two densities are equal while the alternatives are general one-sided $\mu_1 > \mu_2$ for location and $\sigma_1 > \sigma_2$ for scale.

For the two sample location problem the log likelihood is

$$\ell_1(\underset{\sim}{\theta},\Delta) = \sum_{j=1}^{k} n_{1j} \ln\{F(\theta_j + \tfrac{1}{2}\Delta) - F(\theta_{j-1} + \tfrac{1}{2}\Delta)\}$$

$$+ \sum_{j=1}^{k} n_{2j} \ln\{F(\theta_j - \tfrac{1}{2}\Delta) - F(\theta_{j-1} - \tfrac{1}{2}\Delta)\},$$

where $\{\theta_j\}$ are the unknown category boundaries and $F'(x) = f(x)$ is an arbitrary but known density. The null hypothesis H_O is $H_O : \Delta = 0$. The one-sided alternatives are $H_1 : \Delta > 0$ or $H_1' : \Delta < 0$. The locally most powerful test statistic of H_O against H_1 or H_1' is based on the score function $\frac{\partial \ell_1}{\partial \Delta}(\hat{\underset{\sim}{\theta}}_O, 0)$ (Cox & Hinkley, 1974, p.113) evaluated under the null hypothesis. It is easy to verify that the m.l.e. $\hat{\underset{\sim}{\theta}}_O$, of $\underset{\sim}{\theta}$ under H_O is given by

$$F(\hat{\theta}_{Oj}) = c_j = \frac{1}{n..} \sum_{\ell=1}^{j} n._\ell \qquad (j = 1,\ldots,k-1)$$

where c_j is the cumulative frequency up to, and including category j in the combined sample. Hence

$$\hat{\theta}_{Oj} = F^{-1}(c_j) \qquad (j = 1,\ldots,k-1)$$

where $\hat{\underset{\sim}{\theta}}_O = (\hat{\theta}_{O1} \ldots \hat{\theta}_{O\,k-1})$. The score statistic $\frac{\partial \ell_1}{\partial \Delta}(\hat{\theta}_O, O)$ is given by

$$\frac{\partial \ell_1}{\partial \Delta}(\hat{\theta}_O, O) = \frac{1}{2}\sum_{j=1}^{k}(n_{1j} - n_{2j})\left[\frac{f\{F^{-1}(c_j)\} - f\{F^{-1}(c_{j-1})\}}{c_j - c_{j-1}}\right] \qquad (4.2.1)$$

It is easy to see that (4.4.1) is a rank statistic since the cumulative frequency c_j depends on the maximum rank of the tied observations in category j.

For the scale problem, the log likelihood $\ell_2(\underset{\sim}{\theta}, \tau)$ is

$$\ell_2(\underset{\sim}{\theta}, \tau) = \sum_{j=1}^{k} n_{1j}\ln\{F(\theta_j) - F(\theta_{j-1})\} + n_{2j}\ln\{F(\tau\theta_j) - F(\tau\theta_{j-1})\}.$$

The null hypothesis of equality is $H_O : \tau = 1$ and the locally most powerful test statistic $\frac{\partial \ell_2}{\partial \beta}(\hat{\underset{\sim}{\theta}}_O, 1)$ is

$$\frac{\partial \ell_2}{\partial \beta}(\hat{\underset{\sim}{\theta}}_O, 1) = \sum_{j=1}^{k} n_{2j}\left[\frac{F^{-1}(c_j)f(F^{-1}(c_j)) - F^{-1}(c_{j-1})f(F^{-1}(c_{j-1}))}{c_j - c_{j-1}}\right]. \qquad (4.2.2)$$

We denote the location and scale test statistics by W_1 and W_2 respectively where

$$W_1(f) = \sum_{j=1}^{k} n_{1j}\{\Psi_1(c_j, f) - \Psi_1(c_{j-1}, f)\}/(c_j - c_{j-1}), \qquad (4.2.3)$$

$$W_2(f) = \sum_{j=1}^{k} n_{1j}\{\Psi_2(c_j, f) - \Psi_2(c_{j-1}, f)\}/(c_j - c_{j-1}) \qquad (4.2.4)$$

with $\Psi_1(c_j, f) = f\{F^{-1}(c_j)\}$ and $\Psi_2(c_j) = F^{-1}(c_j)f\{F^{-1}(c_j)\}$.

Note that the statistics $W_1(f)$ and $W_2(f)$ are equivalent to (4.2.1) and (4.2.2) respectively. These test statistics are for location and scale

differences respectively. If we wish to test for other types of departure summarised in a parameter ϕ measuring, say, skewness, the function $\Psi(u,f)$ has the form

$$\Psi(u,f) = \frac{\partial}{\partial\phi} F(\theta;\phi)\Big|_{\theta=F^{-1}(u;\phi_0)} \qquad 0 < u < 1 \qquad (4.2.5)$$

where ϕ_0 is the m.l.e. of ϕ under H_0.

We now investigate some particularly important special cases of the statistics (4.2.3) and (4.2.4). The most common special case occurs when there are no tied observations. In principle, this can be incorporated into the above model by allowing the number of categories to become arbitrarily large. The limiting values of (4.2.3) and (4.2.4) are

$$\sum_{\text{group 1}} \psi_1(u_j)$$

and

$$\sum_{\text{group 1}} \psi_2(u_j)$$

where $\frac{R_j - 1}{n..} < u_j < \frac{R_j}{n..}$, R_j is the rank of the j^{th} observation from sample 1 and the summations are taken over observations in sample 1. It is usual to take $u_j = R_j/(n.. + 1)$ (Hájek, 1962) so that the limiting values are

$$W_1 = \sum_{\text{group 1}} \psi_1\{R_j/(n.. + 1)\} \qquad (4.2.6)$$

and

$$W_2 = \sum_{\text{group 1}} \psi_2\{R_j/(n.. + 1)\} \qquad (4.2.7)$$

with $\psi_1(u) = \frac{\partial}{\partial u}\Psi_1(u;f)$ and $\psi_2(u) = \frac{\partial}{\partial u}\Psi_2(u,f)$. The statistics (4.2.6) and (4.2.7) are precisely the asymptotically most powerful rank tests, for the location and scale problem suggested by Hájek (1962) and Hájek and Sidák (1967), who have proved that these statistics are asymptotically normal under H_0 and under a general class of alternatives. It follows from general likelihood theory that W_1 and W_2 are asymptotically jointly normally distributed.

The functions $\psi_1(u)$ and $\psi_2(u)$ given by

$$\psi_1(u) = f'\{F^{-1}(u)\}/f\{F^{-1}(u)\} \qquad (0 < u < 1)$$

and $$\psi_2(u) = 1 + F^{-1}(u)f'\{F^{-1}(u)\}/f\{F^{-1}(u)\} \qquad (0 < u < 1)$$

are usually known as the score functions for the location and scale problem respectively. Note that if f is symmetric ψ_1 is odd about 1/2 and ψ_2 is even about 1/2. The asymptotic null variance of the location test statistic (4.2.6) is

$$\frac{n_1 n_2}{n} \int_0^1 \psi_1^2(u)\,du \qquad (4.2.8)$$

where $n_1 = n_{1\cdot}$, $n_2 = n_{2\cdot}$ and $n = n_1 + n_2$.

When f is the logistic density, the location test statistic (4.2.6) is equivalent to the ordinary Wilcoxon test statistic while (4.2.3) is the averaged ranks Wilcoxon statistic. The two sided exponential distribution gives the sign test, the normal distribution leads to the expected normal scores test etc.

We now revert to the problem of primary interest, namely optimum rank tests with tied observations. Hájek and Sidák (1967) pp.118-124 suggest various methods for handling tied observations. These methods

include randomisation, averaged scores and mid-ranks. None of these is claimed to be optimal. Behnen (1976) showed that the averaged scores rank statistic is asymptotically superior to randomisation and mid-rank statistics. Note that for the Wilcoxon test, $\psi_1(u)$ is linear and hence the mid-rank and averaged scores statistic are equal. In general though they are not equal. We now show that the locally most powerful statistic (4.2.3) is asymptotically equal to the averaged scores statistic and hence this provides an alternative proof of Behnen's (1976) result.

The ranks of the observations in the j^{th} category are $n\,c_{j-1} + 1$ up to $n\,c_j$. The average score for the j^{th} category is therefore

$$\frac{1}{n_{\cdot j}} \sum_{i=nc_{j-1}+1}^{i=nc_j} \psi\left(\frac{i}{n+1}\right) .$$

For large n, this is approximately equal to the integral

$$\frac{1}{c_j - c_{j-1}} \int_{c_{j-1}}^{c_j} \psi(u)\, du$$

$$= \{\Psi(c_j) - \Psi(c_{j-1})\}/(c_j - c_{j-1}) .$$

Hence the locally most powerful test statistics (4.2.3) and (4.2.4) are asymptotically equal to the averaged scores rank statistic. This concludes the proof that the averaged scores test is asymptotically most powerful.

The asymptotic null variance of the locally most powerful tests W_1 and W_2, and hence the asymptotic variance of the averaged scores rank test for the two sample problem are

$$\frac{n_1 n_2}{n} \sum_{j=1}^{k} \{\Psi_1(c_j) - \Psi_1(c_{j-1})\}^2/(c_j - c_{j-1}) \qquad (4.2.9)$$

and $$\frac{n_1 n_2}{n} \sum_{j=1}^{k} \{\Psi_2(c_j) - \Psi_2(c_{j-1})\}^2/(c_j - c_{j-1}) .$$

For the logistic density the expression (4.2.9) reduces to

$$\frac{n_1 n_2}{n} \sum_{j=1}^{k} (c_j - c_{j-1})(1 - c_j - c_{j-1})^2 . \qquad (4.2.10)$$

The exact variance of the Wilcoxon tied rank statistic is known to be

$$\frac{1}{12} n_1 n_2 (n+1) \{1 - \frac{1}{n(n^2-1)} \sum_{j=1}^{k} (t_j^3 - t_j)\}$$

(Gibbons, 1976, p.165) where t_j is the number of ties in the j^{th} group or category. Since W_1 is related to the Wilcoxon statistic W through

$$W_1 = \frac{2}{n+1} W$$

it follows that the exact variance of W_1 is

$$\frac{n_1 n_2}{3(n+1)} \{1 - \frac{1}{n(n^2-1)} \sum_{j=1}^{k} (t_j^3 - t_j)\}.$$

It follows from the identities (2.4.6) that the exact variance is asymptotically equal to (4.2.10) as expected.

It should be pointed out that since the Wilcoxon test statistic W (or W_1) and the maximum likelihood estimator $\hat{\Delta}$ of Chapter 2 are based on the same likelihood function, tests of $H_0 : \Delta = 0$ based on $\hat{\Delta}$ or W are asymptotically equivalent (Cox & Hinkley, 1974, pp.314-315). However, the Wilcoxon statistic does not provide an estimate of the parameter Δ, but it is a valid test of H_0 against a wide non-parametric class of alternatives, e.g. $H_1 : F_1(x) > F_2(x)$.

Finally, table 4.1 gives the functions $\psi_1(u)$ and $\psi_2(u)$ for some common densities.

Table 4.1

Score functions $\psi_1(u)$ and $\psi_2(u)$ for some common densities

Density	$\psi_1(u)$	$1 + \psi_2(u)$
Normal	$\Phi^{-1}(u)$	$\{\Phi^{-1}(u)\}^2$
Double exp.	sign (2u - 1)	$-\ln\{1 - \lvert 2u-1 \rvert\}$
Logistic	2u - 1	$(2u-1)\ln\{u/(1-u)\}$
Cauchy	$-\sin 2\pi u$	$\cos^2 \pi u$

4.3 Joint Distribution of Several Rank Test Statistics

The methods of §4.2 provide rank statistics for testing specific hypotheses about a single parameter. All other parameters are assumed to remain constant. For most purposes this is an unreasonable assumption. It is often necessary to use a single set of data to make inference about several parameters. The interpretation and analysis of such data sets is greatly facilitated if the parameter estimates or the test statistics are independent or approximately independent.

A simple notation is developed which is useful for calculating joint moments of general rank statistics under various hypotheses. For simplicity we consider only the location scale model and the associated hypotheses for the two sample problem.

$$H_0 : F_1(x) = F_2(x)$$
$$H_1 : F_1(x) = F_2(x + \Delta)$$
$$H_2 : F_1(x) = F_2(\tau x)$$
$$H_3 : F_1(x) = F_2(\tau x + \Delta)$$

It is convenient to consider only densities f(x) which are symmetric since this implies that the score functions $\psi_1(u)$ and $\psi_2(u)$ are odd and even respectively about $u = \frac{1}{2}$. Hence they are orthogonal over (0,1). This relationship is expressed in inner product notation as $\langle\psi_1,\psi_2\rangle = 0$. Note in addition that $\langle\psi_1,1\rangle = \langle\psi_2,1\rangle = 0$.

We will consider the two-sample problem although the analysis could be extended to regression problems. Let π_i be the probability that the i^{th} ordered observation in the combined sample of size n came from the first sample. The joint probability π_{ij} is defined as the probability that both the i^{th} and j^{th} ordered observations are from the first sample.

We define the related functions $\xi_{1n}(u)$, $\xi_{1n}\xi_{1n}^T(u,v)$ by the differential relations

$$d\xi_{1n}\left(\frac{i}{n+1}\right) = \begin{cases} 1/n & \text{if } i^{th} \text{ ranked observation from sample 1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$d^2\xi_{1n}\xi_{1n}^T\left(\frac{i}{n+1}, \frac{j}{n+1}\right) = \begin{cases} \frac{1}{n(n-1)} & \text{if } i^{th} \text{ and } j^{th} \text{ ranked obs. from sample 1} \\ 0 & \text{otherwise} \end{cases}$$

Thus $E\{d\xi_{1n}(u)\}$ is a function which takes the values $\{\pi_i\}$ at the points $i/(n+1)$ and zero elsewhere. Similarly $E\{d^2\xi_{1n}\xi_{1n}^T(u,v)\}$ takes values π_{ij} at the points $(i/(n+1), j/(n+1))$ and zero elsewhere.

For the two sample problem the asymptotically most powerful rank test statistics for location and scale respectively, are of the form

$$W_{1n} = \frac{1}{n}\Sigma\psi_1\{R_i/(n+1)\} \tag{4.3.1}$$

and

$$W_{2n} = \frac{1}{n}\Sigma\psi_2\{R_i/(n+1)\},$$

where the summation is taken over the ranks of the observations in the first sample. The statistics (4.3.1) are conveniently expressed in inner product notation as

$$W_{1n} = \langle\psi_1, \xi_{1n}\rangle \tag{4.3.2}$$

and $$W_{2n} = \langle\psi_2, \xi_{1n}\rangle.$$

The statistics W_{1n} and W_{2n} are based on the derivatives of the log likelihood with nuisance parameters evaluated under H_0. Hence, from general likelihood theory, W_{1n} and W_{2n} are asymptotically jointly normally distributed with zero mean under H_0 and variance-covariance matrix given by Fisher's information matrix. We are interested in their joint distribution under H_0, H_1, H_2, H_3.

The first order moments can be evaluated by taking expectations of (4.3.2). This gives

$$E\{W_{1n}\} = \langle \psi_1, E\{\xi_{1n}\} \rangle$$

$$E\{W_{2n}\} = \langle \psi_2, E\{\xi_{1n}\} \rangle. \qquad (4.3.3)$$

Under H_0, $E\{\xi_{1n}\}$ takes the value n_1/n at each of the points $\frac{i}{n+1}$, $i = 1,\ldots,n$ and hence the expectations in (4.3.3) are both zero. Under H_1, $E\{\xi_{1n}\}$ can be expressed as the sum of a constant and an odd function. Hence $E\{W_{2n}\} = 0$ under H_1. Under H_2, $E\{\xi_{1n}\}$ is even. Hence $E\{W_{1n}\} = 0$ under H_2. Under H_3, the first moments are, in general, non-zero.

The above results depend on ψ_1 being odd and ψ_2 being even. We have already noted that when the data are grouped, averaged scores are asymptotically most powerful. However, averaging the scores usually destroys the symmetry of ψ_1 and ψ_2. Hence the results that $E\{W_{1n}\} = 0$ under H_2 and $E\{W_{2n}\} = 0$ under H_1 do not apply, in general, when the data are grouped.

The second order moments of W_{1n} and W_{2n} are obtained by taking expectations of the equations

$$W_{1n}^2 = \int_0^1 \int_0^1 \psi_1^2(u)\psi_1^2(v)d^2\xi_{1n}\xi_{1n}^T(u,v),$$

$$W_{1n}W_{2n} = \int_0^1 \int_0^1 \psi_1^2(u)\psi_2^2(v)d^2\xi_{1n}\xi_{1n}^T(u,v) \qquad (4.3.4)$$

and $$W_{2n}^2 = \int_0^1 \int_0^1 \psi_2^2(u)\psi_2^2(v)d^2\xi_{1n}\xi_{1n}^T(u,v).$$

In a generalised inner product notation these equations become

$$W_{1n}^2 = \langle\psi_1,\xi_{1n}\xi_{1n}^T,\psi_1\rangle,$$

$$W_{1n}W_{2n} = \langle\psi_1,\xi_{1n}\xi_{1n}^T,\psi_2\rangle \qquad (4.3.5)$$

and $$W_{2n}^2 = \langle\psi_2,\xi_{1n}\xi_{1n}^T,\psi_2\rangle.$$

The analogy with matrix multiplication and quadratic forms is clear. Taking expectations of the equations (4.3.5) gives generalised inner products with kernel $E\{\xi_{1n}\xi_{1n}^T(u,v)\}$. It is easy to verify that under H_0 and H_2 this kernel is symmetric about $u = v$ and about $u = 1-v$. Since ψ_1 is odd and ψ_2 is even it follows that under H_0 and H_2, the statistics W_{1n} and W_{2n} are uncorrelated. This result does not hold under H_1 or H_3, nor does it hold when the data are grouped, except in the special case when the grouping is symmetric, i.e. the number of ties in the i^{th} group is equal to the number of ties in the $(k+1-i)^{th}$ category where k is the number of categories.

For grouped or categorised data we write

$$W_{1n} = \frac{1}{n}\sum_{j=1}^{k} \psi_{1j}\, n_{1j} \qquad (4.3.6)$$

and

$$W_{2n} = \frac{1}{n}\sum_{j=1}^{k} \psi_{2j}\, n_{1j}$$

where ψ_{1j}, ψ_{2j} are the integrated scores for the j^{th} category. Explicitly

$$\psi_{1j} = \frac{1}{c_j - c_{j-1}} \{\Psi_1(c_j) - \Psi_1(c_{j-1})\}, \tag{4.3.7}$$

and similarly for ψ_{2j}.

It follows from (4.3.6) and (4.3.7) that

$$W_{1n} = -\frac{1}{n} \sum_{j=1}^{k} \psi_{1j} n_{2j}$$

and similarly for W_{2n}, so that

$$\sum_{j=1}^{k} \psi_{1j} n_{\cdot j} = \sum_{j=1}^{k} \psi_{2j} n_{\cdot j} = 0 . \tag{4.3.8}$$

This is a useful check for the calculated scores ψ_{1j}, ψ_{2j}.

From (4.2.7) we can calculate the second moments of W_{1n} and W_{2n} under H_0, which are:

$$\mathrm{Var}(W_{1n}) = \frac{n_1 n_2}{n^3} \sum_{j=1}^{k} n_{\cdot j} \psi_{1j}^2$$

$$\mathrm{Var}(W_{2n}) = \frac{n_1 n_2}{n^3} \sum_{j=1}^{k} n_{\cdot j} \psi_{2j}^2 \tag{4.3.9}$$

and

$$\mathrm{Cov}(W_{1n}, W_{2n}) = \frac{n_1 n_2}{n^3} \sum_{j=1}^{k} n_{\cdot j} \psi_{1j} \psi_{2j}$$

Since W_{1n} and W_{2n} are asymptotically joint normal with zero mean under H_0, these second moments specify the asymptotic distribution completely.

4.4 Example

We use the data of §2.9 to illustrate the simultaneous application of two rank tests. The test statistics for location and scale are denoted by W_1 and W_2 respectively. We use a logistic density and the score functions $\psi_1(u)$, $\psi_2(u)$ derived from the logistic density.

$$\psi_1(u) = 2u - 1$$

$$(0 < u < 1) \qquad (4.4.1)$$

$$\psi_2(u) = (2u-1)\ln\{u/(1-u)\} - 1$$

For grouped data it is more useful to have the functions $\Psi_1(u)$ and $\Psi_2(u)$ which are, apart from changes of sign

$$\Psi_1(u) = u(1 - u)$$

$$(0 < u < 1) \qquad (4.4.2)$$

and

$$\Psi_2(u) = u(1 - u)\ln\{u/(1 - u)\} .$$

Table 4.2 demonstrates the steps involved in calculating the location and scale rank statistics W_1 and W_2 as given in (4.3.6).

The original data is in the columns headed groups. Columns 1, 2, 3 are category totals, probabilities and cumulative probabilities respectively. The scores ψ_{1j} in column 4 can be calculated directly from the formula $\psi_{1j} = 1 - c_j - c_{j-1}$. This is purely a consequence of the logistic model since

$$\frac{1}{c_j - c_{j-1}} \{\Psi_1(c_j) - \Psi_1(c_{j-1})\} = 1 - c_j - c_{j-1} .$$

Table 4.2

Calculation of location and scale rank test statistics for logistic models

	Groups		1	2	3	4	5	6
Category	1	2	Total	p_j	c_j	ψ_{1j}	$\Psi_2(c_j)$	ψ_{2j}
0	-	-	-	0	0	-	0	-
1	19	497	516	.3691	.3691	.6309	-.1248	-.3382
2	29	560	589	.4213	.7904	-.1515	.2199	.8182
3	24	269	293	.2096	1.0	-.7904	0	-1.0492
Total	72	1326	1398	1.0				

Column 5 gives the values $\Psi_2(c_j)$ and column 6 is obtained from 5 and 2 from the formula

$$\psi_{2j} = \frac{1}{c_j - c_{j-1}} \{\Psi_2(c_j) - \Psi_2(c_{j-1})\}.$$

Note that the inner products of column 1 with columns 4 and 6 are zero. The statistics nW_1 and nW_2 are simply the inner products of columns 4 and 6 with the numbers in group 1. Their values are -11.6087 and -7.8785 respectively. The second moments of nW_1 and nW_2 are obtained from (4.3.9). They are

$$\text{Var}(nW_1) = 19.7073$$

$$\text{Var}(nW_2) = 37.9034$$

and $$\text{Cov}(nW_1, nW_2) = 2.7360.$$

The standardised statistics $T_1 = W_1/\text{std}(W_1)$ and $T_2 = W_2/\text{std}(W_2)$ have values -2.6150 and -1.2797 with covariance matrix

$$\underset{\sim}{V} = \begin{bmatrix} 1 & .1001 \\ .1001 & 1 \end{bmatrix} .$$

The statistic T_1 should be compared with the ratio $\tilde{\Delta}/\text{std}(\tilde{\Delta})$ in §2.9. For testing H_1 against H_0 we use T_1 which is approximately standard normal, and for H_2 against H_0 we use T_2 which is again standard normal. However, since the tests are not independent we should use

$$(T_1^2 + T_2^2 - 2\rho T_1 T_2)/(1 - \rho^2) \qquad (4.4.3).$$

where $\rho = .1001$ for testing H_3 against H_0. The statistic (4.4.3) is asymptotically X_2^2 and has a value of 7.8848. From the numerical calculations in this example, it appears that when there are only three categories, the combined location-scale statistic (4.4.3) is exactly equal to Pearson's X^2 goodness of fit statistic. Thus T_1 and T_2 are non-orthogonal components of the X^2 statistic. For data with more than three categories, the two statistics are not equal.

In this particular example it is fortunate that the correlation between T_1 and T_2 is small. If the correlation is large, this suggests that either unequal scales or unequal locations adequately explains the differences between the two groups. It may be preferable, from prior information etc., to use the location alternative H_1 rather than the scale alternative H_2, but because of the high correlation the data does not distinguish between the two alternatives. Of course, if the differences are sufficiently great, it may be necessary to use the more general alternative H_3.

Thus if we are prepared, on the grounds of prior information, to accept H_1 as a likely explanation of the differences between the two groups we may wish to test the adequacy of H_1 with H_3 as alternative. It is thus appropriate to do a conditional test based on T_2 given the value of T_1. Thus

$$E(T_2|T_1) = \rho T_1 ,$$

$$Var(T_2|T_1) = Var(T_2)(1 - \rho^2) = 1 - \rho^2$$

since T_1 and T_2 are asymptotically joint normal. For the data of table 4.2 the conditional test statistic

$$\frac{T_2 - \rho T_1}{(1 - \rho^2)^{1/2}}$$

has a value of -1.0230. However in this example there is no ambiguity and the location parameter alone is sufficient to explain the differences between the two groups.

Chapter 5

MATCHED SAMPLES

5.1 Introduction

The first section of this chapter comprises a review of the literature for matched categorised data, matched binary data etc., together with some general remarks on latent structure or latent trait models. The discussion of further parametric models for square contingency tables is postponed to §6.1.

Section 5.2 includes a description of the paired logistic model for binary data and this model is extended in §5.3 to ordered multicategory data. In the extended model, difficulties arise concerning hypothesis testing and parameter estimation, which do not arise in the binary model. To tackle these problems, some results concerning mixtures of binomial distributions and multivariate binomial distributions are outlined in §5.4. These results are used in §5.5 to combine the information from various marginal distributions and hence to obtain a reasonable estimator of the parameter of interest, called Δ. Two such estimators are described in §5.5 and simulation results of §5.6 indicate that the two estimators of Δ are asymptotically unbiased and the result of §5.8 shows that the estimators achieve full asymptotic efficiency at least in a special case. An empirical Bayes procedure is described in §5.7 and an example concerning some radiological data is provided in §5.9.

A matched design is a blocked design with a fixed and equal number of observations per block. In particular, a matched pairs, or paired comparisons design, has two observations per block. Usually, one member of each pair is a 'control' and the second is a 'treated' observation. The principal objective of such an experiment is usually to make inference

about the treatment effect independently of the block effects. Questions about the interaction between blocks and treatments may also be important in many cases, but in this chapter we make the simplifying assumption that the treatment effect remains constant on the logistic scale. In §5.10 a simple test is suggested for checking the validity of the assumption of no interaction.

The purpose of matching or blocking is to reduce the effect of uncontrolled variations and thus to increase the precision of the experiment (Cox, 1958b, p.23; Davies, 1954, p.17). Much of the work on matched designs refers to continuous, usually normally distributed, random variables. As in Chapters 2 and 3, we derive a model for the underlying continuous variable Y and examine the properties of the associated discrete model. For matched pairs the following model is considered in greater detail throughout this chapter.

$$\begin{aligned} E\{Y_{i1}\} &= \lambda_i - \tfrac{1}{2}\Delta, \\ E\{Y_{i2}\} &= \lambda_i + \tfrac{1}{2}\Delta \end{aligned} \qquad (5.1.1)$$

where, conditional on λ_i, Y_{i1} and Y_{i2} are independent random variables with a logistic distribution. The parameter of interest, Δ, is the common difference between the means of each pair $\{Y_{i1}, Y_{i2}\}$, and the block parameters $\{\lambda_i\}$ are nuisance parameters.

More generally, for matched designs with t observations per block and arbitrary but known density, $f(y|\lambda,\tau)$, depending on the block parameter λ and the treatment parameter τ, the joint conditional density of the t observations in the i^{th} block is

$$\prod_{j=1}^{t} f_{Y_j|\lambda_i,\tau_j}(y_j|\lambda_i,\tau_j). \qquad (5.1.2)$$

The formulation (5.1.2) is called the conditional model since the distribution is conditional on the parameters $\{\lambda_i\}$, and in such generality is not considered further. Model (5.1.1) is a particular case of the more general conditional model (5.1.2) and is considered in greater detail throughout this chapter.

Some other specialisations of (5.1.2) have been examined in the statistical literature. In particular, the normal theory model is a special case of (5.1.2). However, the simplicity of the normal theory model for continuous data does not carry over to categorised data. For this reason the logistic version of (5.1.2) is chosen for the analysis of categorised data.

So far no assumptions have been made about the block parameters $\{\lambda_i\}$. In some cases it is reasonable to assume that the $\{\lambda_i\}$ are i.i.d. random variables from some parametric family $G(\lambda)$. Then the joint marginal density of the t observations in a given block is

$$\int \{\prod_{j=1}^{t} f_{Y_j|\lambda,\tau_j}(y_j|\lambda,\tau_j)\}dG(\lambda) \quad . \tag{5.1.3}$$

This type of model is known as a latent structure model (Andersen, 1973; Anderson, 1959; Lazarsfeld, 1950, 1955) and the parameter λ is the latent trait variable.

For matched pairs (5.1.3) defines a joint density for (Y_1,Y_2). Since (5.1.3) is a mixture of conditionally independent random variables, the unconditional variables Y_1 and Y_2 are positively correlated. A particular case of some interest occurs when f and G are both normal distributions. In this case Y_1 and Y_2 are bivariate normal. However, in general the integral in (5.1.3) is rather intractable and this imposes some limitations on the applicability of the model. One particular case which is of special interest for hypothesis testing, is $\Delta = 0$ or more generally in (5.1.3)

$\tau_i = 0$, $i = 1,\ldots,t$. This implies symmetry for the joint distribution of $Y_1\ldots,Y_t$.

We now examine some of the models in the literature for matched categorised data. These are special cases of models for square contingency tables. Two properties which all the models have in common are that they have symmetry as a special case and they allow a complete range of association from independence to complete dependence.

The models for square contingency tables can be roughly divided into two classes. The first type (Bishop et al. 1975, Ch.8) are the log-linear models, such as quasi-symmetry and are discussed more fully in §6.1. The second type is characterised by the explicit fitting of a bivariate distribution to the data. These make use of the ordering among the categories. The distributions used in this context include the bivariate normal and the contingency or C-type distributions (Mardia, 1970, pp.55-73; Plackett, 1965). Both distributions allow a complete range of association, but only the bivariate normal distribution has a simple explanation in terms of a latent trait variable. It should be pointed out that the two classes are not distinct since the bivariate normal model implies quasi symmetry but the converse is not true.

An interesting question arises when an unconditional model such as the bivariate normal model or the bivariate logistic model (5.7.8) are found not to fit the data. It is not clear whether the lack of fit is due to a wrong conditional formulation or whether it is due to a wrong mixing distribution for the nuisance parameters. It is the purpose of the model described in this chapter to make inference about the parameter of interest independently of the nuisance parameters.

5.2 Binary Paired Comparisons

The logistic model for binary paired comparisons has been discussed widely in the literature; see, for example, Cox (1958a), (1970), and Altham (1971) who gives a Bayesian analysis. Let X_{i1}, X_{i2} be binary random variables for the i^{th} pair of observations. The logistic model can be written as

$$\begin{aligned} \text{pr}(X_{i1} = x_{i1}|\lambda_i) &= \exp\{(\lambda_i - \tfrac{1}{2}\Delta)x_{i1}\}/\{1 + \exp(\lambda_i - \tfrac{1}{2}\Delta)\}, \\ \text{pr}(X_{i2} = x_{i2}|\lambda_i) &= \exp\{(\lambda_i + \tfrac{1}{2}\Delta)x_{i2}\}/\{1 + \exp(\lambda_i + \tfrac{1}{2}\Delta)\}, \end{aligned} \tag{5.2.1}$$

$$x_{i1}, x_{i2} = 0,1; \quad -\infty < \lambda_i, \Delta < \infty.$$

Thus $\{\lambda_i\}$ are the block parameters or the latent trait variables and Δ is the treatment effect. Then for the i^{th} pair the logistic transforms of the conditional probabilities (5.2.1) are $\lambda_i - \frac{1}{2}\Delta$ and $\lambda_i + \frac{1}{2}\Delta$ respectively, where λ_i is a nuisance parameter characteristic of the i^{th} pair and Δ is a treatment effect assumed constant on the logistic scale. As many authors have noted, the conditional probability

$$\text{prob}(X_{i1} = 0, X_{i2} = 1|X_{i1} + X_{i2} = 1) = e^{\Delta}/(1 + e^{\Delta}) \tag{5.2.2}$$

is independent of the nuisance parameter λ_i. Thus the conditional distribution of the number of (0,1) pairs n_{01}, conditional on the total number of 'mixed' pairs $n_{01} + n_{10}$ is exactly binomial with parameter $e^{\Delta}/(1 - e^{\Delta})$ (Cox, 1970, 55-58). Then the conditional maximum likelihood estimator of Δ is

$$\hat{\Delta}_c = \ln(n_{01}/n_{10}) .$$

Note that this conditional analysis ignores the numbers of 'unmixed' pairs n_{00} and n_{11}. Andersen (1973) has investigated the consistency of the unconditional and conditional maximum likelihood estimator of . The unconditional estimator is inconsistent. Under mild conditions on the sequence of incidental or latent parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, the conditional maximum likelihood estimator, $\hat{\Delta}_c$, converges almost surely to the true value (Andersen, 1973, pp.19-22, 45), and hence is consistent. The restrictions on the sequence of incidental parameters are necessary to ensure that $n_{01} + n_{10} \to \infty$ with probability 1. A sufficient condition on the sequence of incidental parameters, to ensure that $\hat{\Delta}_c$ is consistent, is that the λ_i should be i.i.d. random variables. An alternative condition is that the λ_i should all belong to some compact subsets of the real line. The purpose of these conditions is to ensure that the probability of success or the probability of failure does not become dominant as the sequence of observations gets longer. If the probability of success were to increase too rapidly the number of 'mixed' pairs would not increase beyond a certain point and thus the variance of $\hat{\Delta}_c$ would not tend to zero. For such sequences it seems clear that there is no consistent estimator of Δ and that the information on Δ is bounded as the number of observations increases.

5.3 The Logistic Model for Matched Ordered Categorised Data

One possible, though not a necessary, physical explanation for the model (5.2.1), is in terms of an underlying continuous variable Y. Suppose Y_{i1} and Y_{i2} are two logistic random variables with means $\lambda_i - \frac{1}{2}\Delta$ and $\lambda_i + \frac{1}{2}\Delta$ respectively, and whose variances are conventionally fixed at $\pi^2/3$ which is the variance of the standard logistic distribution. A success (X = 1) is observed when Y is positive and a failure (X = 0) is observed when Y is negative. The derived model for the pair of binary random variables is (5.2.1). We exploit this physical interpretation of the logistic model for binary data, to extend the model to handle ordered categorised variables.

As in §2.1 we introduce k-1 unknown parameters $\theta_1,\ldots,\theta_{k-1}$ to represent the category boundaries. Without loss of generality one of these boundaries can be set equal to zero provided the block parameters $\{\lambda_i\}$ are either arbitrary constants or have a distribution with arbitrary location parameter.

Explicitly, the model for the observed X's is

$$X = j \quad \text{if } \theta_{j-1} \leq Y < \theta_j \qquad (j = 1,\ldots,k) \tag{5.3.1}$$

where $-\infty = \theta_0 < \theta_1 < \ldots < \theta_k = \infty$. In addition

$$\begin{aligned} \text{prob}(Y_{i1} \leq y \mid \lambda_i,\Delta) &= F(y - \lambda_i + \tfrac{1}{2}\Delta) \\ \text{prob}(Y_{i2} \leq y \mid \lambda_i,\Delta) &= F(y - \lambda_i - \tfrac{1}{2}\Delta) \end{aligned} \tag{5.3.2}$$

where $F(X) = e^X/(1 + e^X)$ and, conditional on λ_i, Y_{i1} and Y_{i2} are independent.

As it stands, the model has n incidental or nuisance parameters $\{\lambda_i\}$ and k structural parameters $\theta_1,\ldots,\theta_{k-1}, \Delta$. The terminology is somewhat misleading here since we might consider $\theta_1,\ldots, \theta_{k-1}$ as nuisance parameters. It is easy to see that not all of the parameters are estimable. For example, we can add an arbitrary constant to each of the λ_i's and the same constant to the θ_j's and the model is unchanged. To eliminate this indeterminateness we can arbitrarily fix one of the θ_j or constrain the λ_i to be centred around some fixed point, say, zero. If the $\{\lambda_i\}$ are considered random it is convenient to eliminate the confounding by choosing a distribution for the $\{\lambda_i\}$ which is centred on zero.

From (5.3.1) and (5.3.2) we note that

$$\text{pr}(X_{i1} \leq j \mid \lambda_i, \Delta, \underset{\sim}{\theta}) = F(\theta_j - \lambda_i + \tfrac{1}{2}\Delta) \qquad (j = 1,\ldots,k-1)$$

$$\text{pr}(X_{i2} \leq j \mid \lambda_i, \Delta, \theta) = F(\theta_j - \lambda_i - \tfrac{1}{2}\Delta) \qquad (j = 1,\ldots,k-1)$$

Hence, for $j = 1,\ldots,k-1$

$$\frac{\text{pr}(X_{i1} \leq j,\ X_{i2} > j \mid \lambda_i, \Delta, \underset{\sim}{\theta})}{\text{pr}(X_{i1} > j,\ X_{i2} \leq j \mid \lambda_i, \Delta, \underset{\sim}{\theta})} = e^{\Delta} \qquad (5.3.3)$$

so that the conditional probabilities

$$\text{pr}(X_{i1} \leq j,\ X_{i2} > j \mid X_{i1} \leq j,\ X_{i2} > j \quad \text{or} \quad X_{i1} > j\ X_{i2} \leq j)$$

have the common value of $e^{\Delta}/(1 + e^{\Delta})$ for $j = 1,\ldots,k-1$, and this conditional probability is independent of the nuisance parameters.

We note in passing that the model (5.3.1) and (5.3.2) does not completely specify the k^2 cell probabilities when the $\{\lambda_i\}$ are treated

as random variables with an unknown distribution. On the other hand when the $\{\lambda_i\}$ are treated as fixed parameters the standard methods based on the likelihood function do not, in general, yield consistent estimates even of the structural parameters. However, the conditional model does imply the restriction (5.3.3) on the cell probabilities. It would be of interest to know whether or not the conditional model implies any further restrictions on the cell probabilities. This can be expressed formally by the question "Do there exist functions g and h such that

$$g\{pr(X_1 = i, X_2 = j \mid \lambda, \underline{\theta}, \Delta) \quad i,j = 1,\ldots,k\} = h(\theta, \Delta),$$

where g is a function of the k^2 conditional cell probabilities and h is independent of λ?". Of course, from the implicit function theorem, such functions must exist. In fact, since there are $k^2 - 1$ functionally independent cell probabilities defined by $k + 1$ parameters, there must be $k^2 - 2$ such relationships. We have chosen for simplicity to use only k-1 of these defined in (5.3.3) to estimate Δ. Consequently, our estimate of Δ cannot be expected to be fully efficient unless the statistics associated with (5.3.3) are, in some sense, sufficient for Δ.

Suppose that the observed cell frequencies are $\{m_{ij}\}$ $1 \leq i,j \leq k$ where m_{ij} is the number of pairs of X's for which the first X is i and the second, j. Consider the 2×2 table obtained from the original square k×k table by combining categories 1,2,...,j as, say, failure and categories j+1,...,k as success. There are k-1 such tables and for each the conditional model (5.3.1), (5.3.2) reduces to the model for binary paired comparisons (5.2.1). Let the off-diagonal elements in these tables be r_j and $n_j - r_j$ where r_j is the number of pairs (X_{i1}, X_{i2}) where $X_{i1} \leq j$ and $X_{i2} > j$. Similarly $n_j - r_j$ is the number of pairs where $X_{i1} > j$ and $X_{i2} \leq j$. From

(5.3.3), the conditional distribution of r_j given n_j is binomial with index n_j and parameter $e^{\Delta}/(1 + e^{\Delta})$. We propose to estimate Δ by examining the joint distribution of $\underset{\sim}{r} = (r_1,\ldots,r_{k-1})^T$ conditional on $\underset{\sim}{n} = (n_1,\ldots,n_{k-1})^T$. Unfortunately, the distribution of r_j is not independent of n_i when $i \neq j$ so that, conditional on $\underset{\sim}{n}$ the marginals are no longer binomial and the joint distribution of $\underset{\sim}{r}$ given $\underset{\sim}{n}$ depends on the nuisance parameters. Intuitively, though, we would expect the marginal dependence of r_j on the nuisance parameters to be small.

Since the indices $\{n_j\}$ are interdependent it could be argued that it is not very meaningful to condition on all of them jointly. A second approach might be to examine the joint distribution of $\{r_j\}$ and $\{n_j - r_j\}$ conditional only on the total number of off-diagonal observations. In §5.4 we develop some theory concerning mixture properties of binomial and multinomial distributions. These results are used in §5.5 to derive an estimator for Δ.

5.4 Some Multivariate Binomial Mixtures

In this section we try to develop some multivariate binomial distributions which are useful for describing the joint distribution of the vector $\underset{\sim}{r} = (r_1,\ldots,r_{k-1})^T$ where the elements of $\underset{\sim}{r}$ are defined in §5.3. The covariance structure of $\underset{\sim}{r}$ arises through random components which the elements of $\underset{\sim}{r}$ have in common (Lancaster, 1974; Patil & Joshi, 1968), and through mixing over the indices of the constituent independent random variables. Another equivalent way of generating correlated binomial random variables is to extend the definition of a binomial random variable as the joint distribution of two Poisson random variables whose sum is fixed. We first derive a few simple results for univariate binomial and multinomial distributions.

Let $X_i, i = 1,\ldots,s$ be s independent binomial random variables with indices $\{n_i\}$ and parameters $\{p_i\}$. Suppose, in addition that the $\{n_i\}$ are themselves random variables from a multinomial distribution with index $n = \sum_{j=1}^{s} n_j$ and parameter vector $\underline{\phi} = (\phi_1,\ldots,\phi_s)$ where $\sum_{j=1}^{s} \phi_j = 1$. We are interested in the distribution of ΣX_j and in the joint distribution of $X_1,\ldots,X_s$ conditional only on n. By using probability generating functions it is easy to show that ΣX_j is a binomial random variable with index n and parameter $\bar{p} = \sum_{j=1}^{s} \phi_j p_j$ which is a weighted average of the parameters $\{p_j\}$ of the original independent binomials.

The joint probability generating function for $X_1, X_2, \ldots, X_s$ is

$$[1 - \bar{p} + \sum_{j=1}^{s} t_j \phi_j p_j]^n = \sum_{r_1,\ldots r_s} \cdots \sum t_1^{r_1} t_2^{r_2} \ldots t_s^{r_s} P(X_1 = r_1,\ldots,X_s = r_s | n).$$

That is to say that $X_0, X_1, \ldots, X_s$, where $X_0 = n - \Sigma X_j$, have a joint multinomial distribution with index n and parameter vector $(1 - \bar{p}, \phi_1 p_1, \phi_2 p_2, \ldots, \phi_s p_s)$. In addition the joint distribution of $X_1, X_2, \ldots, X_s$ conditional on ΣX_j is multinomial with index ΣX_j and parameter vector $(\phi_1 p_1/\bar{p}, \phi_2 p_2/\bar{p}, \ldots, \phi_s p_s/\bar{p})$.

These examples demonstrate some of the mixture properties of univariate binomial distributions. Since binomial and multinomial random variables can be generated from independent Poisson random variables with fixed sum, the above reproducibility properties of the binomial and multinomial distributions have simple analogues for the Poisson distribution.

Note that the distribution of $\{X_j\}$ conditional on $n_1,\ldots,n_s$ depends on 2s parameters $n_1,\ldots,n_s, p_1,\ldots,p_s$, whereas the distribution conditional on n depends only on two parameters $n, \bar{p}$.

We now try to extend these ideas to multivariate binomial distributions. For simplicity we consider only the bivariate case. As before we let X_1, X_{12}, X_2 be independent binomial random variables with indices n_1, n_{12}, n_2

and parameters p_1, p_{12}, p_2. We generate correlated variables Y_1, Y_2 by setting

$$Y_1 = X_1 + X_{12} \text{ and } Y_2 = X_2 + X_{12}.$$

The joint distribution of Y_1 and Y_2 depends on the six parameters n_1, n_{12}, n_2, p_1, p_{12}, p_2 and has probability generating function

$$(q_1 + p_1t_1)^{n_1}(q_{12} + p_{12}t_1t_2)^{n_{12}}(q_2 + p_2t_2)^{n_2} . \qquad (5.4.1)$$

When $p_1 = p_{12} = p_2$ this distribution is called bivariate binomial with overlapping trials (Patil & Joshi, 1968, p.60) and both marginals are binomial. In general, however, the marginals are not binomial since the Y's are sums of binomials with different parameters. If, however, the indices are random it is possible that the marginals could be binomial as in the univariate case. For example, if n_2 is fixed and n_1 and n_{12} are independent Poisson random variables whose sum is fixed ($n_1 + n_{12} = N_1$, say) then Y_1 is binomial but Y_2 is the sum of two binomials and has p.g.f.

$$(q_2 + p_2t_2)^{n_2}(1 - \phi_2p_{12} + \phi_2p_{12}t_2)^{N_1} .$$

A natural extension is to let n_1, n_{12}, n_2 have independent Poisson distributions with parameters λ_1, λ_{12}, λ_2 and to fix $n_1 + n_{12} = N_1$ and $n_2 + n_{12} = N_2$. This effectively leaves one degree of freedom among the variables n_1, n_{12}, n_2 and the distribution of n_{12} is

$$p(n_{12}|N_1,N_2,\mu) = C\frac{\mu^{n_{12}}}{(N_1-n_{12})!\, n_{12}!\,(N_2-n_{12})!} \qquad (5.4.2)$$

where $0 \leq n_{12} \leq \min(N_1,N_2)$, $\mu = \lambda_{12}/\lambda_1\lambda_2$ and c is a constant to make the distribution sum to unity. Let $\zeta(t)$ be the probability generating function for n_{12} in (5.4.2). Then

$$\zeta(t) = c \sum_{r=0}^{\min(N_1,N_2)} \frac{\mu^r t^r}{(N_1-r)!\,r!\,(N_2-r)!} \qquad (5.4.3)$$

so that ζ depends on N_1, N_2, μ as well as t. Then from (5.4.1) and (5.4.3) the joint p.g.f. of Y_1 and Y_2 conditional on N_1, N_2 is given by

$$(q_1 + p_1t_1)^{N_1}(q_2 + p_2t_2)^{N_2}\,\zeta\{\frac{q_{12} + p_{12}t_1t_2}{(q_1+p_1t_1)(q_2+p_2t_2)}\}, \qquad (5.4.4)$$

and neither marginal is binomial as mentioned towards the end of §5.3. The first few moments are

$$E(Y_1) = N_1p_1 + \zeta'(1)(p_{12} - p_1), \qquad (5.4.5)$$

$$\mathrm{Var}(Y_1) = N_1p_1q_1 - p_1\zeta'(1)(p_{12}-p_1)+(p_{12}-p_1)^2[\zeta''(1) - \{\zeta'(1)\}^2] \qquad (5.4.6)$$

and

$$\mathrm{Cov}(Y_1,Y_2) = (p_{12}-p_1)(p_{12}-p_2)[\zeta''(1)-\{\zeta'(1)\}^2] + \zeta'(1)\{p_{12}(1-p_1-p_2) + p_1p_2\}. \qquad (5.4.7)$$

It does not appear possible to express $\zeta(t)$ or $\zeta'(1)$ in a simpler form. Of course $\zeta'(1)$ and $\zeta''(1)$ are equal to $E(n_{12})$ and $E\{n_{12}(n_{12}-1)\}$ respectively.

The bivariate distribution of Y_1 and Y_2 can be extended to a multivariate distribution for $Y_1,Y_2,\ldots,Y_s$, but an elaborate notation is needed to describe the general analogues of (5.4.2)-(5.4.4). No essentially new points arise in the general case.

We note that the distribution whose p.g.f. is given by (5.4.4) has the property that if we ignore N_2 or let N_2 be very large then the distribution (5.4.2) becomes binomial with index N_1 and parameter $\lambda_{12}/(\lambda_1 + \lambda_{12})$. Hence, by an earlier result Y_1 is binomial with index N_1 and parameter $(p_1\lambda_1 + p_{12}\lambda_{12})/(\lambda_1 + \lambda_{12})$. The distribution of $r_1,\ldots,r_{k-1}$ conditional on $n_1,\ldots,n_{k-1}$ is the multivariate analogue of the joint distribution of Y_1 and Y_2 conditional on N_1 and N_2.

Another extension of the univariate result is to assume n_1, n_{12}, n_2 to have independent Poisson distributions with parameters λ_1, λ_{12}, λ_2 and to fix the total $N = n_1 + n_2 + n_{12}$. Thus n_1, n_{12}, n_2 are trinomially distributed with index N and parameter vector $\underset{\sim}{\phi} = (\lambda_1/\lambda_\cdot\ ,\ \lambda_{12}/\lambda_\cdot\ ,\ \lambda_2/\lambda_\cdot\ )$ where $\lambda_\cdot = \lambda_1 + \lambda_2 + \lambda_{12}$. Then the joint distribution of Y_1 and Y_2 has p.g.f.

$$(\phi_1 q_1 + \phi_{12} q_{12} + \phi_2 q_2 + \phi_1 p_1 t_1 + \phi_2 p_2 t_2 + \phi_{12} p_{12} t_1 t_2)^N , \qquad (5.4.8)$$

so that the marginal distributions are both binomial. The p.g.f. for the general multivariate analogue of this distribution is given in Patil & Joshi (1968) p.81. The marginal distributions have parameters $\phi_1 p_1 + \phi_{12} p_{12}$ and $\phi_2 p_2 + \phi_{12} p_{12}$ respectively, with common index N. In addition, the conditional distributions of Y_1 given $N_1 = n_1 + n_{12}$ and of Y_2 given $N_2 = n_2 + n_{12}$ are both binomial with indices N_1 and N_2 respectively and parameters $(\phi_1 p_1 + \phi_{12} p_{12})/(\phi_1 + \phi_{12})$ and $(\phi_2 p_2 + \phi_{12} p_{12})/(\phi_2 + \phi_{12})$ respectively.

We note that, as in the univariate case, there is a reduction in the number of parameters from six in (5.4.1) to four in (5.4.8). In the general p-variate analogue the reduction is from $2^{p+1} - 2$ to 2^p. However the version we require has the restriction that all parameters whose indices are not in strict sequence, e.g. p_{13}, p_{124} etc. are all zero.

In this special case the number of parameters is reduced from $p(p+1)$ to $\frac{1}{2}p(p+1) + 1$.

For our present purposes it is useful to write the p.g.f. (5.4.8) in an alternative form which generates the probabilities for $n_1 + n_{12} - r_1 - r_{12}$, $n_2 + n_{12} - r_2 - r_{12}$ as well as for $r_1 + r_{12}$, $r_2 + r_{12}$. The alternative version of (5.4.8) is

$$[\phi_1 q_1 s_1 + \phi_{12} q_{12} s_1 s_2 + \phi_2 q_2 s_2 + \phi_1 p_1 t_1 + \phi_{12} p_{12} t_1 t_2 + \phi_2 p_2 t_2]^N. \quad (5.4.9)$$

The p.g.f. (5.4.9) applies only to the bivariate case. In the application to matched contingency tables, the parameters ϕ, p, q have the following properties:

$$\phi_1 q_1 + \phi_{12} q_{12} = \underset{\Lambda}{E}\, pr(X_1 > 1, X_2 \leq 1|\lambda),$$

$$\phi_1 p_1 + \phi_{12} p_{12} = \underset{\Lambda}{E}\, pr(X_1 \leq 1, X_2 > 1|\lambda),$$

$$\phi_2 q_2 + \phi_{12} q_{12} = \underset{\Lambda}{E}\, pr(X_1 > 2, X_2 \leq 2|\lambda),$$

and

$$\phi_2 p_2 + \phi_{12} p_{12} = \underset{\Lambda}{E}\, pr(X_1 \leq 2, X_2 > 2|\lambda).$$

Hence

$$\phi_1 p_1 + \phi_{12} p_{12} = e^{\Delta}(\phi_1 q_1 + \phi_{12} q_{12})$$

and

$$\phi_2 p_2 + \phi_{12} p_{12} = e^{\Delta}(\phi_2 q_2 + \phi_{12} q_{12}).$$

Note that the marginal p.g.f. obtained from (5.4.9) by putting $t_2 = 1$ is

$$[(\phi_1 q_1 + \phi_{12} q_{12}) s_1 + (\phi_1 p_1 + \phi_{12} p_{12}) t_1 + \phi_2]^N \qquad (5.4.10)$$

and hence the marginal likelihood estimator of Δ is $\hat{\Delta}_1 = \ln\{r_1/(n_1 - r_1)\}$ or $\hat{\Delta}_2 = \ln\{r_2/(n_2 - r_2)\}$ from the second marginal distribution. To obtain an efficient estimator of Δ from the two marginal likelihood estimators we need an estimate of the covariance matrix of r_1, r_2, $n_1 - r_1$, $n_2 - r_2$. It follows from (5.4.10) that

$$E(r_1) = e^{\Delta} E(n_1 - r_1)$$

and

$$E(r_2) = e^{\Delta} E(n_2 - r_2).$$

The second moments are obtained from (5.4.9):

$$\operatorname{cov}(r_1, r_2) = N\phi_{12}p_{12} - (\phi_1 p_1 + \phi_{12} p_{12})(\phi_2 p_2 + \phi_{12} p_{12})$$

$$\operatorname{cov}(n_1 - r_1, n_2 - r_2) = N\phi_{12}q_{12} - (\phi_1 q_1 + \phi_{12} q_{12})(\phi_2 q_2 + \phi_{12} q_{12})$$

$$\operatorname{cov}(r_1, n_1 - r_1) = -(\phi_1 q_1 + \phi_{12} q_{12})(\phi_1 p_1 + \phi_{12} p_{12})$$

$$\operatorname{cov}(r_2, n_1 - r_1) = -(\phi_2 p_2 + \phi_{12} p_{12})(\phi_1 q_1 + \phi_{12} q_{12})$$

etc.

When N is large, the only important covariances are $\operatorname{cov}(r_1, r_2)$ and $\operatorname{cov}(n_1 - r_1, n_2 - r_2)$, which are approximately $N\phi_{12}p_{12}$ and $N\phi_{12}q_{12}$ respectively. It is easy to show that $E(r_{12}|N) = N\phi_{12}p_{12}$ and $E(n_{12} - r_{12}|N) = N\phi_{12}q_{12}$. For small Δ, $p_{12} = 1/2$, so that $\operatorname{cov}(r_1, r_2) = \operatorname{cov}(n_1 - r_1, n_2 - r_2) = \frac{1}{2} E(n_{12}|N)$. This approximate result could have been obtained by considering the n's

fixed and examining the variation of $\underset{\sim}{r}$ relative to $\underset{\sim}{n}$. In §5.5 we show how to combine the information from the marginal likelihoods, in the general case where there are k categories.

5.5 Estimation of Δ

We now consider the problem of estimating Δ in the conditional model (5.3.1)-(5.3.2). Where necessary, we will assume that the $\{\lambda_i\}$ are i.i.d. random variables. In particular this means that the cell counts are multinomial or, equivalently, they are independent Poisson random variables whose sum is fixed. This is the motivation behind the multivariate distributions in §5.4.

In §5.3 it was shown that the distribution of r_j conditional on n_j is binomial with index n_j and parameter $e^{\Delta}/(1 + e^{\Delta})$. It follows that $\ln\{r_j/(n_j-r_j)\}$ tends almost surely to Δ, (Andersen, 1974), or equivalently, $r_j/(n_j - r_j)$ tends a.s. to e^{Δ} for $j = 1,\ldots,k-1$. Of course the indices $\{n_j\}$ are themselves random and interdependent. The analysis of §5.4 shows that the distribution of r_j conditional on $n_1,\ldots,n_{k-1}$ is not binomial and the marginal expectation of r_j depends in general on all the indices $n_1,\ldots,n_{k-1}$. However, the joint distribution of the indices $\underset{\sim}{n}$ is multivariate binomial with index $n = \sum_{i\neq j}\sum m_{ij}$ and parameters p_i, p_{ij}, p_{ijk} etc. which depend on the nuisance parameters $\underset{\sim}{\lambda}$, $\underset{\sim}{\theta}$. This gives a joint distribution for r conditional on n which is the k-1 variate generalisation of (5.4.8), with the restriction that the parameters of all the conditional marginal distributions are equal to $e^{\Delta}/(1 + e^{\Delta})$. The marginal expectation of $\underset{\sim}{r}$ conditional only on n has the form

$$E(\underset{\sim}{r}|n) = E(\underset{\sim}{n}|n)e^{\Delta}/(1 + e^{\Delta}). \qquad (5.5.1)$$

We consider now two estimators of Δ based on $\underset{\sim}{r}$.

$$\tilde{\Delta} = \sum_{j=1}^{k-1} \tilde{w}_j \ln\{(r_j + \tfrac{1}{2})/(n_j - r_j + \tfrac{1}{2})\} \qquad (5.5.2)$$

and

$$\Delta^* = \ln[\{\tfrac{1}{2} + \underset{\sim}{w}^{*T}\underset{\sim}{r}\}/\{\tfrac{1}{2} + \underset{\sim}{w}^{*T}(\underset{\sim}{n} - \underset{\sim}{r})\}] \qquad (5.5.3)$$

where $\sum_j \tilde{w}_j = 1$. These estimators are analogous to the estimators $\tilde{\Delta}$ and Δ^* in the two-sample problem of Chapter 2.

We assume that the case $\Delta = 0$ plays a special role in that it is the dividing point between two qualitatively different possibilities namely $\Delta < 0$ and $\Delta > 0$. Formally, therefore, we construct a null hypothesis of symmetry $H_0 : \Delta = 0$ with a general alternative $H_1 : \Delta \neq 0$. We choose weights $\underset{\sim}{w}^*$ and $\underset{\sim}{\tilde{w}} = (\tilde{w}_1,\ldots,\tilde{w}_{k-1})$ to minimise the variances of Δ^* and $\tilde{\Delta}$ under H_0.

The addition of 1/2 to both numerator and denominator in (5.5.1) and (5.5.2) ensures that the estimates remain finite. In fact it also ensures mean square convergence in addition to almost sure convergence. Several other types of combination are possible. For example, in (5.5.2) we could take logs after summation.

We consider first the null distribution of $\underset{\sim}{r}$ and later examine some of the non-null properties. First we define the symmetric matrix $\underset{\sim}{N} = \{n_{ij}\}$ by

$$n_{ij} = \sum_{\alpha \leq i} \sum_{\beta > j} (m_{\alpha\beta} + m_{\beta\alpha}) \qquad (i \leq j)$$

where $m_{\alpha,\beta}$ is the number of pairs of observations (X_{i1},X_{i2}) which are equal to (α,β). Thus the diagonal elements of $\underset{\sim}{N}$ are equal to $\underset{\sim}{n}$ $(n_{ii} = n_i)$,

and the element n_{ij} is a measure of the random elements which r_i and r_j have in common. We first obtain the null distribution of r conditional on $\underset{\sim}{N}$ and this enables us to get the unconditional first and second moments. Conditioning on $\underset{\sim}{N}$ is equivalent to conditioning on all $\frac{1}{2}k(k-1)$ quantities $\{m_{ij} + m_{ji}\}$ $i < j$ so that, under H_0, the elements of $\underset{\sim}{r}$ are sums of binomial random variables each with parameter 1/2. Hence the null distribution of $\underset{\sim}{r}$ conditional on $\underset{\sim}{N}$ is multivariate binomial with overlapping trials (Patil & Joshi, 1968, p.60). Its joint p.g.f. is given by

$$\prod_{j=1}^{k-1} (q+pt_j)^{m_{j,j+1}+m_{j+1,j}} \prod_{j=1}^{k-2} (q+pt_jt_{j+1})^{m_{j,j+2}+m_{j+2,j}} \ldots (q+pt_1t_2\ldots t_{k-1})^{m_{1k}+m_{k1}}$$

with $p = q = 1/2$. The first two conditional moments are

$$E_0(\underset{\sim}{r}|\underset{\sim}{N}) = \frac{1}{2}\underset{\sim}{n}$$

and

$$V_0(\underset{\sim}{r}|\underset{\sim}{N}) = \frac{1}{4}\underset{\sim}{N}.$$

The same moments conditional only on $\underset{\sim}{n}$ are

$$E_0(\underset{\sim}{r}|\underset{\sim}{n}) = E_0\{E_0(\underset{\sim}{r}|\underset{\sim}{N})|\underset{\sim}{n}\} = \frac{1}{2}\underset{\sim}{n} \qquad (5.5.4)$$

and

$$V_0(\underset{\sim}{r}|\underset{\sim}{n}) = E_0\{V_0(\underset{\sim}{r}|\underset{\sim}{N})|\underset{\sim}{n}\} + V_0\{E_0(\underset{\sim}{r}|\underset{\sim}{N})|\underset{\sim}{n}\}$$

$$= \frac{1}{4}E_0\{\underset{\sim}{N}|\underset{\sim}{n}\} \qquad (5.5.5)$$

The subscript 0 refers to the null distribution with $\Delta = 0$. Equation (5.5.5) tells us that, regardless of the nuisance parameters, $\frac{1}{4}\underset{\sim}{N}$ is an unbiased estimator of the null variance matrix of $\underset{\sim}{r}$ given $\underset{\sim}{n}$.

Note that conditional only on n, the sum of the off diagonal elements of the square table, the null variance matrix becomes

$$V_0(\underset{\sim}{r}|n) = \frac{1}{4} E_0(\underset{\sim}{N}|n) + \frac{1}{4} V_0(\underset{\sim}{n}|n). \tag{5.5.6}$$

It is easiest to obtain an expression for the asymptotic null variances of $\tilde{\Delta}$ and Δ^* conditional on all the elements of $\underset{\sim}{n}$. We do this by ignoring the factors of $\frac{1}{2}$ in the numerators and denominators because these factors have no asymptotic effect on the estimators. Using $\frac{1}{4}\underset{\sim}{N}$ as an estimate of the null variance of $\underset{\sim}{r}$, the weights $\tilde{w}$ and w^* which minimise the null variances of $\tilde{\Delta}$ and Δ^* are

$$\underset{\sim}{\tilde{w}} = \underset{\sim}{D}\underset{\sim}{N}^{-1}\underset{\sim}{n}/(\underset{\sim}{n}^T\underset{\sim}{N}^{-1}\underset{\sim}{n})$$

and

$$\underset{\sim}{w}^* = \underset{\sim}{N}^{-1}\underset{\sim}{n}, \tag{5.5.8}$$

where

$$\underset{\sim}{D} = \text{diag}\{n_1,\ldots,n_{k-1}\}.$$

A similar result can be obtained by examining the joint distribution of $\underset{\sim}{r}$ and $\underset{\sim}{s} = \underset{\sim}{n} - \underset{\sim}{r}$ conditional on n. Under H_0 these have a symmetric multivariate binomial distribution where p.g.f. has the form

$$[q\{\Sigma_i \phi_i u_i + \Sigma_{i,j} \phi_{ij} u_i u_j + \ldots\} + p\{\Sigma\phi_i t_i + \Sigma\phi_{ij} t_i t_j + \ldots\}]^n, \tag{5.5.9}$$

with $p = q = \frac{1}{2}$. Note that the joint p.g.f. of n has the form

$$[\Sigma\phi_i t_i + \Sigma\phi_{ij} t_i t_j + \Sigma\phi_{ijk} t_i t_j t_k + \ldots]^n, \tag{5.5.10}$$

obtained from (5.5.9) by putting $u = t$. From (5.5.10) it follows that

$E(n_{i\cdot}) = n\phi_{i\cdot}$ and $E(n_{ij}) = n\phi_{ij}$, where $\phi_{i\cdot}$ is the summation of all ϕ's which have i as one of their subscripts, and similarly for ϕ_{ij}.

It follows from (5.5.9) that the null variance of $(\underset{\sim}{w}^{*T}\underset{\sim}{r})/(\underset{\sim}{w}^{*T}\underset{\sim}{s})$ is asymptotically given by

$$V_0\{\underset{\sim}{w}^{*T}\underset{\sim}{r}/\underset{\sim}{w}^{*T}\underset{\sim}{s}\} \approx \underset{\sim}{w}^{*T}E(\underset{\sim}{N})\underset{\sim}{w}^{*}/[\tfrac{1}{2}\,\underset{\sim}{w}^{*T}E(\underset{\sim}{n})]^2,$$

and hence

$$\underset{\sim}{w}^{*} = [E(\underset{\sim}{N})]^{-1}E(\underset{\sim}{n}),$$

so the obvious estimator of $\underset{\sim}{w}^{*}$ is $\underset{\sim}{w}^{*} = \underset{\sim}{N}^{-1}\underset{\sim}{n}$.

The asymptotic null variance of Δ^* is therefore given by

$$V_0(\Delta^*) = 4/\underset{\sim}{n}^{T}\underset{\sim}{N}^{-1}\underset{\sim}{n}\,. \tag{5.5.11}$$

The asymptotic null variance of $\tilde{\Delta}$ is also given by (5.5.11). The non-null variances of $\tilde{\Delta}$ and Δ^* can be approximated by (5.5.11) for small Δ, but for slightly larger values of Δ the following variance estimator is suggested.

$$V(\Delta^*) = V(\tilde{\Delta}) = 4(1 + \tfrac{1}{4}\Delta^2)/\underset{\sim}{n}^{T}\underset{\sim}{N}^{-1}\underset{\sim}{n} \tag{5.5.12}$$

This approximation is suggested by the relation

$$V(r_j|n_j) = n_j\, e^{\Delta}/(1 + e^{\Delta})^2 = \tfrac{1}{4}\,n_j(1 - \tfrac{1}{4}\Delta^2 \ldots)$$

The covariances are not deflated by the same factor. Nevertheless, (5.5.12) seems to be a reasonable approximation for medium values of Δ.

If the observed matrix $\underset{\sim}{N}$ is singular, as can happen when the original data are sparse, it is sufficient to find $\tilde{\underset{\sim}{w}}$, $\underset{\sim}{w}^*$ which satisfy

$$\underset{\sim}{N}\underset{\sim}{D}^{-1}\tilde{\underset{\sim}{w}} = \underset{\sim}{n} \qquad \text{with } \underset{\sim}{1}^T\tilde{\underset{\sim}{w}} = 1$$

and

$$\underset{\sim}{N}\underset{\sim}{w}^* = \underset{\sim}{n}$$

where $\underset{\sim}{1}$ is the unit vector. In this case the asymptotic variance of both estimators is $(4 + \Delta^2)/(\underset{\sim}{n}^T\underset{\sim}{w}^*)$.

An interesting mathematical problem arises concerning the consistency of Δ^* when the complete ranking of all 2n observations is available. We note that for a fixed number of categories, both Δ^* and $\tilde{\Delta}$ are consistent. However, when the complete ranking is used the effective number of categories increases with n and it can be shown that Δ^* reduces to

$$\exp(\Delta^*) = \frac{(\text{No of pairs where } X_2 > X_1) + \frac{1}{2}}{(\text{No of pairs where } X_1 > X_2) + \frac{1}{2}}$$

and that, for small Δ, $\operatorname{plim}(\Delta^*) = 2/3\Delta + O(\Delta^3)$. Hence Δ^* is inconsistent for completely ranked observations. It seems likely that $\tilde{\Delta}$ is also inconsistent in the same limit but it does not have a simple limiting form.

It seems reasonable, therefore, to expect that when the number of categories is large, the estimator Δ^* will be biased towards the origin.

5.6 Simulation Results for the Estimators $\tilde{\Delta}$, Δ^*

To test the adequacy of $\tilde{\Delta}$ and Δ^* in both medium (n = 100) and large (n = 1000) samples, four categories were taken so that the square tables each had 16 cells. This gives an average cell count of 6 and 60 for the medium and large samples respectively. However, many of the observations lay on the diagonal cells and did not enter into the analysis. Thus, the effective sample size or the number of off diagonal elements was considerably less than n.

Pairs of continuous random variables Y_{i1}, Y_{i2} were generated as the sum of a uniform and a logistic random variable

$$Y_{i1} = U_i + Z_{i1} ,$$

$$Y_{i2} = U_i + Z_{i2} ,$$

where $\{Z_{i1}, Z_{i2}\}$ are independent logistic random variables with mean $(-\frac{1}{2}\delta, \frac{1}{2}\delta)$ and variance 1. The uniform random variables $\{U_i\}$ had a range of (0,8). Hence the correlation of Y_{i1} and Y_{i2} was 16/19 = 0.84. This high correlation reduces the efficiency of the simulation since many observations fall in the diagonal cells of the table and do not enter in the analysis. However, correlations of about .8 are encountered in real data in practice, and in this sense the generated data mimic real data.

The category boundary points $\underset{\sim}{\theta}$ were chosen to be $\underset{\sim}{\theta} = (-\infty, 2, 4, 6, \infty)$ so that reasonable numbers of observations fall in each category. The range of true values of δ was 0,(.1),.9. This is equivalent to values of Δ in the range (0,1.6) since $\Delta = \delta\pi/\sqrt{3}$.

The data in table 5.1 show the true value of δ together with the average of 100 large-sample and the average of 100 small-sample estimates

$\tilde{\Delta}$ and Δ^*. The statistics $\tilde{\Delta}$ and Δ^* or equivalently, $\tilde{\delta}$ and δ^* were calculated on the same data sets and hence are highly correlated. Unlike the corresponding unpaired estimators, neither $\tilde{\Delta}$ nor Δ^* show any bias towards the origin even in medium sized samples.

Table 5.1

Simulation results for paired sample estimators $\tilde{\Delta}$, Δ^*.

	'Small' sample n = 100		'Large' sample n = 1000	
δ	$\tilde{\delta}$	δ^*	$\tilde{\delta}$	δ^*
0	.019	.016	.002	.002
.1	.117	.119	.103	.103
.2	.199	.203	.198	.197
.3	.288	.296	.288	.289
.4	.402	.414	.403	.404
.5	.496	.515	.502	.504
.6	.594	.608	.595	.597
.7	.676	.696	.701	.704
.8	.808	.840	.807	.812
.9	.887	.916	.899	.902

All entries are the means of 100 repetitions at each value of $\delta(=\Delta\sqrt{3}/\pi)$.

The estimated standard deviation from (5.5.12) was adequate for values of δ in the range studied here. The standard deviations of $\tilde{\delta}$ and δ^* for the small samples ranged from .018 at $\delta = 0$ to .022 at $\delta = 1$, while for the large samples the corresponding range was .005 to .007.

FIG. 5.1

Graph of simulation results for paired estimators $\tilde{\Delta}, \Delta^*$ in small samples. (see table 5.1)

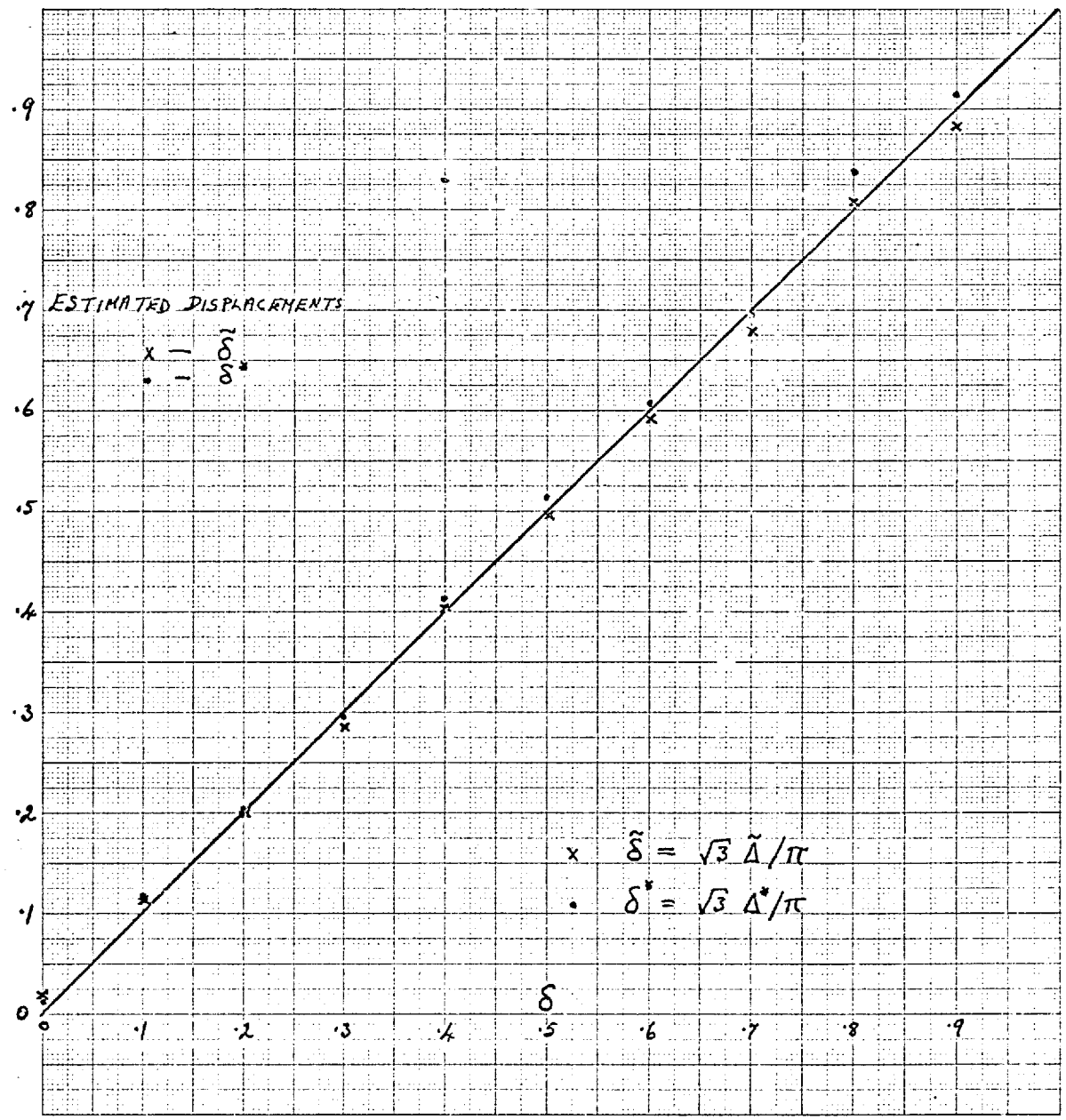
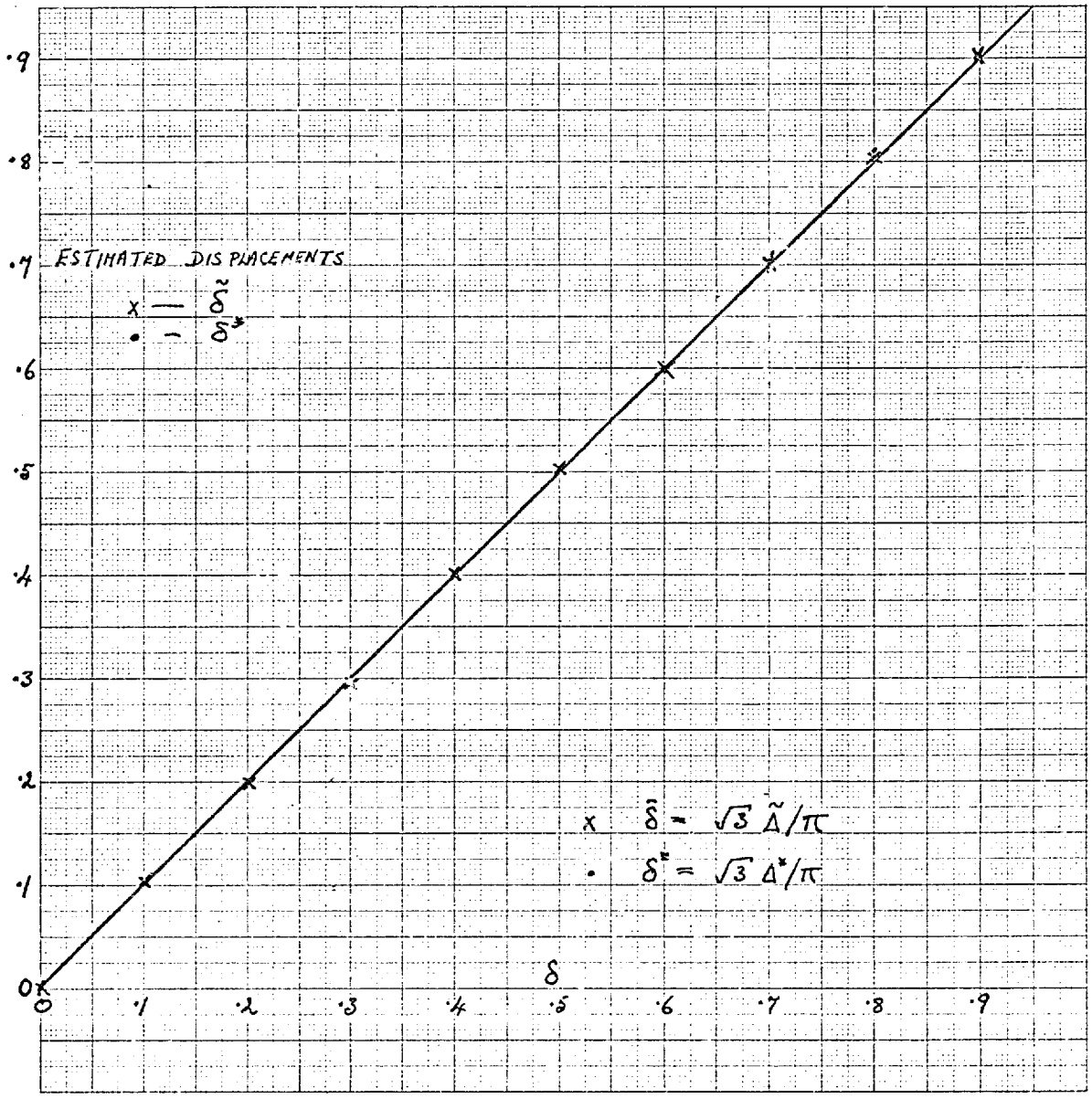


FIG. 5.2

Graph of simulation results for paired estimators $\tilde{\Delta}, \Delta^*$ in large samples. (see table 5.1)



5.7 A Random Effects Model

Suppose the nuisance parameters $\{\lambda_i\}$ are i.i.d. random variables with known distribution function $G(\lambda;\underset{\sim}{\alpha})$ which depends on the unknown parameters $\underset{\sim}{\alpha}$. The parameter of interest, Δ, is assumed fixed. A similar problem concerning the linear model is considered by Cox (1975). As in §5.1 let Y_1 and Y_2 be the unobservable continuous random variables. From (5.1.3) the joint marginal density of pairs of variables (Y_1,Y_2) in the same block is

$$\int\{\prod_{j=1}^{2} f_{Y_j|\lambda}(y_j|\lambda,\Delta)\}dG(\lambda;\underset{\sim}{\alpha}). \tag{5.7.1}$$

If the parametric form of both f and G is known then (5.7.1) defines the joint distribution of pairs of observations in the same block. All pairs have independent identical distributions which depend on the parameters Δ, $\underset{\sim}{\alpha}$. Since the number of parameters is now fixed, the method of maximum likelihood based on the marginal distribution of (Y_1,Y_2) in (5.7.1) yields consistent estimates of the parameters Δ, $\underset{\sim}{\alpha}$.

In many cases of interest, the distribution of the nuisance parameters $\{\lambda_i\}$ is unknown. If the family of distributions $G(\lambda;\underset{\sim}{\alpha})$ is sufficiently flexible we could expect at least one member of the family to be a close approximation to the true distribution of the $\{\lambda_i\}$. It is not clear how much this empirical Bayes procedure is affected by an inflexible choice of prior distribution for the nuisance parameters, but in any case it is possible to test for the adequacy of the chosen model (5.7.1). What this test does not tell us is whether the inadequacy lies in the conditional formulation (5.7.2) of the model or in a poor approximation to the distribution of the nuisance parameters.

The joint density of the continuous unobservable variables Y_1 and Y_2 conditional on λ is

$$f_{Y_1Y_2|\lambda}(y_1,y_2|\lambda,\Delta) = \frac{\exp(y_1 + y_2 - 2\lambda)}{\{1 + \exp(y_1-\lambda+\frac{1}{2}\Delta)\}^2\{1 + \exp(y_2-\lambda-\frac{1}{2}\Delta)\}^2} . \qquad (5.7.2)$$

Since it involves only a linear transformation we consider, without loss of generality, the symmetric case with $\Delta = 0$.

There is considerable difficulty in finding a parametric family of distributions for $\{\lambda_i\}$ which is reasonably flexible and for which the integral

$$\int_{-\infty}^{\infty} f_{Y_1Y_2|\lambda}(y_1,y_2;\lambda)\,dG(\lambda;\underset{\sim}{\alpha})$$

can easily be evaluated. The simplest density $g(\cdot)$ which I could find is

$$g_1(\lambda;\underset{\sim}{\alpha}) = \exp\{(\alpha_3-\lambda)\alpha_1\}/[\{1 + \exp(\alpha_3-\lambda)\}^{\alpha_1+\alpha_2} B(\alpha_1,\alpha_2)]$$

where $\underset{\sim}{\alpha} = (\alpha_1,\alpha_2,\alpha_3)$; $\alpha_1,\alpha_2 > 0$ and $B(\alpha_1,\alpha_2)$ is the beta function. The parameter α_3 is simply a location parameter and can be ignored without loss of generality since the category boundary parameters $\underset{\sim}{\theta}$ take care of location information. It is therefore sufficient to consider the restricted family

$$g(\lambda;\underset{\sim}{\alpha}) = \exp(-\lambda\alpha_1)/[\{1 + \exp(-\lambda)\}^{(\alpha_1+\alpha_2)} B(\alpha_1,\alpha_2)] , \qquad (5.7.3)$$

where $\underset{\sim}{\alpha} = (\alpha_1,\alpha_2)$.

We now consider some properties of the family of densities (5.7.3).

(i) If Z is a beta random variable with parameters α_1 and α_2 then $\ln\{Z/(1 - Z)\}$ has the density (5.7.3).

(ii) The m.g.f. of (5.7.3) is

$$\frac{B(\alpha_1 - t\ ,\ \alpha_2 + t)}{B(\alpha_1,\alpha_2)}\ , \tag{5.7.4}$$

so the cumulant generating function is

$$\psi(\alpha_1-t) + \psi(\alpha_2+t) - \psi(\alpha_1) - \psi(\alpha_2)\ , \tag{5.7.5}$$

where ψ is the log-gamma function. (The psi-function given in Abramowitz & Stegan (1970) is the derivative of the log-gamma function.)

(iii) The density g is unimodal, the maximum frequency occurring at $\lambda = \ln(\alpha_2/\alpha_1)$.

From (5.7.5) the cumulants, κ_r, are

$$\kappa_r = \psi^{(r)}(\alpha_2) + (-1)^r\psi^{(r)}(\alpha_1) \tag{5.7.6}$$

where $\psi^{(r)}$ is the r^{th} derivative of ψ. For the relationship between the cumulants, κ_r, and the moments μ_r see Kendall and Stuart (1969, vol.1, p.68). The mean and variance are given by κ_1 and κ_2. These are

$$\mu = \kappa_1 = \psi'(\alpha_2) - \psi'(\alpha_1)$$

and

$$\sigma^2 = \kappa_2 = \psi''(\alpha_2) + \psi''(\alpha_1)\ .$$

When $\alpha_1 = \alpha_2$ the distribution $g(\lambda;\underset{\sim}{\alpha})$ is symmetric about zero, while $\alpha_1 > \alpha_2$ means that the distribution is negatively skewed and vice-versa.

We now consider the joint marginal distribution of pairs (Y_1,Y_2) after the nuisance parameter λ has been removed by integration. The joint cumulative distribution function of Y_1 and Y_2 is

$$F(y_1 y_2;\underline{\alpha}) = \int_{-\infty}^{\infty} \frac{\exp(y_1+y_2-2\lambda)\exp(-\lambda\alpha_1)}{\{1+\exp(y_1-\lambda)\}\{1+\exp(y_2-\lambda)\}\{1+\exp(-\lambda)\}^{\alpha_1+\alpha_2}}\, d\lambda \quad . \tag{5.7.7}$$

$$= \begin{cases} \dfrac{\alpha_1}{(\alpha_1+\alpha_2)(e^{-y_2} - e^{-y_1})} \{I(e^{-y_2}) - I(e^{-y_1})\} & (y_1 \neq y_2) \\ \dfrac{\alpha_1}{(\alpha_1+\alpha_2)} I'(e^{-y}) & (y_1 = y_2 = y) \end{cases} \tag{5.7.8}$$

where

$$I(e^{-y}) = e^{-y(\alpha_1+1)}\, {}_2F_1(\alpha_1+\alpha_2,\ \alpha_1+1;\ \alpha_1+\alpha_2+1;\ 1-e^{-y})$$

and

$$I'(e^{-y}) = \frac{\alpha_1+1}{(\alpha_1+\alpha_2+1)} e^{-\alpha_1 y}\, {}_2F_1(\alpha_1+\alpha_2,\ \alpha_1+2;\ \alpha_1+\alpha_2+2;\ 1-e^{-y}) \ ,$$

(see Gradshteyn & Ryzhik (1965), 3.315, p.305). The function ${}_2F_1$ is the hypergeometric function which can be computed from a power series convergent when the argument is less than 1 in modulus. For other values of the argument there are recurrence relations given in Abramowitz & Stegun (1970) Ch.15.

There are considerable computational problems associated with fitting the bivariate cumulative distribution (5.7.8). The main problem is the existence of numerous poles of both the hypergeometric function and the gamma function. An additional problem is the slow convergence rate of the hypergeometric series expansion for certain values of the argument. Until a fast reliable algorithm is found for evaluating the hypergeometric function it seems unlikely that (5.7.8) will provide a practical alternative to standard bivariate distributions.

Despite these problems, a Fortran program was written to evaluate the distribution (5.7.8). The series expansion was used to calculate the

hypergeometric function when the argument was small and translation formulae were used for large values of the argument. To avoid poles in the gamma function, its reciprocal was calculated instead.

An application is given in §5.9.

5.8 Random Pairing

Suppose pairs are formed at random, for example by deliberate matching on variables which are in fact unrelated to the factor in question. In this rather unusual case both the two-sample model of chapter 3 and the paired model are applicable. Intuitively, though, it seems that we would do better by using the model with fewer parameters i.e. the independent samples model.

Armitage (1975) investigated the problem of random pairing for binary variables and concluded that the unmatched model was generally more efficient than the matched model. He uses as his criterion the difference between the asymptotic variances of the cross-ratio and the equivalent estimator for matched binary data. The two models are asymptotically equally efficient when there is no treatment effect. In this section it is shown that this result can be extended to multi-category models based on the logistic distribution.

We consider first the variance of the estimator of Δ from the paired or matched model described in §5.5. To distinguish the two estimators the subscript p is used for the paired estimators. From (5.5.12) the variances of $\tilde{\Delta}_p$ and Δ^*_p are given by

$$V_p = V(\tilde{\Delta}_p) = 4(1 + \tfrac{1}{4}\Delta^2)/(\underset{\sim}{n}^T \underset{\sim}{N}^{-1} \underset{\sim}{n}). \qquad (5.8.1)$$

Random pairing implies that all the block parameters $\{\lambda_i\}$ are equal and hence independence in the square table of observations. Hence

$$pr(X_1 \leq i,\ X_2 \geq j) \;=\; pr(X_1 \leq i)\ pr(X_2 \geq j). \qquad i < j$$

We consider the null case and define: $\gamma_i = pr(X_1 \leq i) = pr(X_2 \leq i)$. Let the total number of pairs of observations be n. It follows that the elements $\{n_{ij}\}$ of N have expectation

$$E_0(n_{ij}) \;=\; 2n\gamma_i(1 - \gamma_j), \quad i \leq j \tag{5.8.2}$$

and the elements $\{n_i\}$ of $\underset{\sim}{n}$ have expectation

$$E_0(n_i) \;=\; 2n\gamma_i(1 - \gamma_i).$$

From (5.8.2) we see that $E_0(\underset{\sim}{N})$ is the Green's matrix encountered in Chapter 2. The inverse is a symmetric Jacobi matrix, so the asymptotic value of V_p when $\Delta = 0$ is

$$V_p \approx 2/n \left\{ \sum_{j=1}^{k-1} \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}) \right\}^{-1} \tag{5.8.3}$$

where $\pi_j = \gamma_j - \gamma_{j-1}$. It would be of interest to know the behaviour of V_p for non-zero values of Δ but I have not been able to get even a first order approximation to V_p for non-zero Δ.

For the two sample model we get from (2.3.3)

$$V(\Delta^*) \;=\; V(\tilde{\Delta}) \;=\; \left\{ \frac{n}{2} \sum_{j=1}^{k-1} \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}) \right\}^{-1}$$

when $\Delta = 0$. Thus there is no asymptotic loss of efficiency when the true value of Δ is zero, and this extends Armitage's result for binary data. It seems intuitively clear, although there is no proof, that $\tilde{\Delta}_p$ is inefficient compared to $\tilde{\Delta}$ when the true value of Δ is non-zero.

5.9 Example

This example, from Wise & Oldham (1963), concerns the degree of pneumoconiosis in coalface workers as measured radiologically. There are 8 mines denoted by the letters A,...,H with sample sizes ranging from 33 to 148. At every site each individual was radiographed at the beginning and at the end of a 2.5 year period. The degree of pneumoconiosis is classified on a four category scale indicating increasing severity of the disease. A typical table is given below.

Table 5.2

Paired readings for 82 coalface workers at mine 'G' from table 1 of Wise & Oldham (1963).

		Second reading				
	Category	1	2	3	4	Total
First	1	43	8	3	0	54
reading	2	2	2	5	3	12
	3	1	0	7	2	10
	4	0	0	1	5	6
	Total	46	10	16	10	82

It is fairly clear from table 5.2 that there is a strong association between pairs of readings as we would expect and that, on the average, the second readings are higher than the first. Thus there appears to be some evidence of disease progression at mine 'G'. The interesting question to ask is whether or not the progression is the same at each site, and if not to identify those sites which show most progression.

It should be stressed that we are interested in progression only and not in the absolute levels of the disease. Thus it is possible, though in practice probably unlikely, that the sample chosen at a particular site may have a high average level of pneumoconiosis but show little or no change over a 2.5 year period. Conversely, on a new site with a young workforce, the average level may be low but the progression rate may be fast. It is therefore desirable that we should be able to estimate progression independently of the absolute level of the disease. This is precisely the role of the matched pairs model.

To compare values of $\tilde{\Delta}$ obtained from different tables an extra assumption is necessary. The observer or reader variation must be logistic with constant variance throughout. The actual logistic form is unlikely to be crucial, but the assumption of constant observer variance from table to table is very important since the observer variance determines the scale on which Δ is measured. In this particular data set, all readings were by the same panel and hence it is reasonable to expect the variance to remain constant throughout.

To compute $\tilde{\Delta}$ and Δ^* we need the quantities $\underset{\sim}{N}$, $\underset{\sim}{n}$ and $\underset{\sim}{r}$ for each table. For table 6 these are

$$\underset{\sim}{N} = \begin{bmatrix} 14 & 4 & 0 \\ 4 & 12 & 3 \\ 0 & 3 & 6 \end{bmatrix}, \quad \underset{\sim}{n} = \begin{bmatrix} 14 \\ 12 \\ 6 \end{bmatrix}, \quad \underset{\sim}{r} = \begin{bmatrix} 11 \\ 11 \\ 5 \end{bmatrix}.$$

The weights $\tilde{w}$ and w^* are (.523, .283, .194) and (.847, .534, .733) respectively. The two estimates, $\tilde{\Delta}$ and Δ^* are 1.45 and 1.50 respectively with common standard deviation .53.

Table 7 gives the estimates $\tilde{\Delta}$ and Δ^* together with their standard deviations for the other mines A ... H from the data in table 1 of Wise & Oldham (1963). In the matched pairs model the quantity e^{Δ} has a conditional odds interpretation: it is the odds of observing progression conditional on observing a change.

Table 5.3

Progression estimates $\tilde{\Delta}$, Δ^* for 8 mines A ... H.

	MINES							
	A	B	C	D	E	F	G	H
$\tilde{\Delta}$	.84	.37	2.38	2.88	3.16	3.20	1.45	1.90
Δ^*	.92	.51	2.22	3.22	3.70	3.26	1.50	1.62
Std	.49	1.18	.66	.74	1.11	.60	.53	.58
n	90	33	87	83	82	148	82	84

There is strong evidence of positive progression in all mines except for B and possibly A. Among the other mines it is clear that D, E, F have greatest progression. There may be an explanation for this in terms of location, type of coal, work conditions etc. but no information on such factors is available.

It is of interest to compare the present analysis with other methods which use only the information in the marginals of the table. The data have been analysed by Hutchinson (1976) using an exponential model and by Wise & Oldham (1963) using a Normal distribution for the marginals. Table 5.4 shows the results of these analyses for mine 'G' together with the

estimate obtained by using the method of Clayton (1974) outlined in Chapter 2. All three methods show some evidence of progression although none is conclusive.

Table 5.4

Progression estimates for mine 'G' from marginals only

	Model	Estimate	Std	ratio
H	exponential	.27	.17	1.59
W & O	Normal	.29	.18	1.61
C	Logistic	.48	.31	1.55

It should be emphasised that the different estimates in table 5.4 are not directly comparable although we would expect the estimates from the normal and logistic models to be approximately in the ratio $\sqrt{3}/\pi$: 1 (Cox (1970) pp.26-29). In addition, since the paired model has the conditional variances fixed and models based on the marginals have the marginal variances fixed we would expect matched estimates and marginal estimates to be related through the correlation of pairs of observations. We write this relationship as

$$\Delta_p = \Delta_M/(1 - \rho)^{1/2} \qquad (5.9.1)$$

where Δ_p is the paired estimator and Δ_M is the marginal estimator. In fact there is no justification for examining the marginals alone except when $\rho = 0$ and this corresponds to the case of random pairing discussed in §5.8. A model for matched data, which estimates the marginal by odds ratio Δ_M is discussed in §6.3.

The empirical Bayes model of §5.7 was also used to analyse the data in table 5.2. The model was first fitted with $\alpha_1 = \alpha_2$ so that the mixing distribution of the nuisance parameters was symmetric. The maximum likelihood estimates of the parameters $(\hat{\alpha}, \hat{\Delta}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ were (.265, 1.624, 2.184, 4.144, 7.619) and std$(\hat{\Delta})$ = .52. A X^2 goodness of fit statistic was calculated although some of the cell frequences are small and this had a value of 11.05 on 10 d.f.

In fact the parameter α determines the correlation structure in the distribution (5.7.8), since

$$\rho = 2\psi''(\alpha)/\{2\psi''(\alpha) + \pi^2/3\} \tag{5.9.2}$$

or more generally, when $\alpha_1 \neq \alpha_2$

$$\rho = \rho\{\psi''(\alpha_1) + \psi''(\alpha_2)\}/\{\psi''(\alpha_1) + \psi''(\alpha_2) + \pi^2/3\} \tag{5.9.3}$$

where $\psi(\alpha)$ is the log-gamma function at α. Substitution of $\hat{\alpha}$ in (5.9.2) gives $\hat{\rho}$ = .9036 for the data of table 5.1. (Tables of the digamma and trigamma function are given in Abramowitz & Stegun (1970), Ch.6.)

Using this value of ρ substituted in (5.9.1) gives the approximate relationship $\Delta_p = 3.2\ \Delta_M$. This relationship is verified by examining the estimates of Δ for table 5.2.

$\tilde{\Delta}_p$ = 1.45 - 1.50 from table 5.3

$\tilde{\Delta}_M$ = .48 from table 5.4.

The ratio is approximately 3 as expected.

In conclusion we point out that, for such highly correlated data, there is considerable gain from using the paired model. The asymptotic gain in efficiency increases as the correlation increases.

5.10 Discussion

It is nearly always preferable in any statistical application to use a model whose adequacy to explain the data can be tested. The conditional model described in §5.2 and §5.3 where the nuisance parameters $\{\lambda_i\}$ are unrestricted cannot easily be tested for goodness of fit (Cox (1970), p.58). However it is in principle possible to test that the sequence of estimators

$$\hat{\Delta}_j = \ln\{(r_j + \tfrac{1}{2})/(n_j - r_j + \tfrac{1}{2})\}$$

have common mean Δ. A suitable test statistic would be

$$4(\underset{\sim}{\hat{\Delta}} - \tilde{\Delta}\underset{\sim}{1})^T \underset{\sim}{N}^{-1} (\underset{\sim}{\hat{\Delta}} - \tilde{\Delta}\underset{\sim}{1}) \qquad (5.10.1)$$

where $\underset{\sim}{\hat{\Delta}} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_{k-1})^T$; $\tilde{\Delta}$ is given by (5.5.2) or (5.5.3) and $\underset{\sim}{1}$ is the unit vector. The statistic (5.10.1) has a distribution which is asymptotically X^2 on k-2 degrees of freedom when the model is true. Large values of (5.10.1) indicate some interaction between blocks and treatments.

With the empirical Bayes procedure it is possible to test the adequacy of the model as shown in the example of §5.9. Unfortunately it is not easy to interpret a large X^2 value since it oould arise from a false conditional model or from a false distribution for the nuisance parameters.

An alternative hierarchy of parametric models which is closely related to the models discussed in this chapter, is developed in Chapter 6.

Chapter 6

FURTHER MODELS FOR SQUARE CONTINGENCY TABLES

6.1 Introduction

One important criticism of the conditional model described in Chapter 5 is that its adequacy to explain the data cannot be tested directly. The adequacy of the empirical Bayes model can be tested, but if it is found to be inadequate the inadequacy may be in the mixing distribution, $G(\lambda;\alpha)$, and not in the conditional formulation. The models described in this chapter have been developed to retain some of the useful properties of the paired logistic model with the additional property of being easily tested for goodness of fit. Further, the models form a hierarchy, so that if a particular version is found to be inadequate, a more general version can be fitted.

The most common model in the literature for square contingency tables is the model of quasi-symmetry defined by Caussinus (1965). This model can be expressed in log-linear form (Bishop et al. (1975)) but the multiplicative version is

$$\pi_{ij} = c\,\frac{\alpha_i}{\alpha_j}\,\phi_{ij} \qquad (6.1.1)$$

with $\phi_{ij} = \phi_{ji}$, $\Sigma\Sigma\phi_{ij} = 1$, $\alpha_1 = 1$ and c is a constant to make $\Sigma\Sigma\pi_{ij} = 1$. Symmetry is a special case of quasi-symmetry obtained by the restriction $\alpha_i = 1$, $i = 1,\ldots,k$.

In §6.2 a useful invariance criterion is suggested for deciding whether a particular model is suitable for analysing data on a nominal or on an ordinal scale. It is shown in §6.3 that quasi-symmetry is suited to data on a nominal scale whereas an alternative model, p-symmetry,

is suitable only for data on an ordered scale.

A third model is also described. This model has the advantage of being easily extended to higher dimensional matched tables.

6.2 Invariance Properties of Models for Ordinal and Nominal Data

We consider two invariance properties which may enable us to decide which of two models is appropriate for a given situation. This analysis relates only to square contingency tables although slightly different invariance properties could be suggested for rectangular tables. The first transformation considered is the somewhat trivial row to column interchange. In the categorised matched data problem, it is a matter of taste or convention which variable to place in the rows. The corresponding model should be, in some sense, invariant under this transformation. This transformation will not be considered further because all the models considered are so invariant.

The second, and more important transformation considered is the permutation transform. The permutation transformation involves reordering or permuting both rows and columns of the square table. It is understood that the same permutation transformation is applied to both rows and columns. We formally define permutation invariance as follows.

Let $\underset{\sim}{\Pi}(\underset{\sim}{\theta}) = \{\pi_{ij}(\underset{\sim}{\theta})\}$ $1 \leq i, j \leq k$ be a model for the k^2 cell probabilities which depends on a parameter vector $\underset{\sim}{\theta}$. The set of permutation transformations $\mathcal{P}$ form a group whose typical element $\underset{\sim}{T}_\alpha$ is a $k \times k$ unitary matrix which transforms the vector $\underset{\sim}{J} = (1,2,\ldots,k)^T$ to the permutation $\underset{\sim}{\alpha} = (\alpha_1,\ldots,\alpha_k)^T$. Thus $\underset{\sim}{T}_\alpha \underset{\sim}{J} = \underset{\sim}{\alpha}$ and the elements $\{t_{ij}(\underset{\sim}{\alpha})\}$ of $\underset{\sim}{T}_\alpha$ are given by

$$t_{ij}(\underset{\sim}{\alpha}) = \begin{cases} 1 & (j = \alpha_i) \\ 0 & \text{otherwise.} \end{cases}$$

The special subgroup $\mathcal{P}_\nu$ of $\mathcal{P}$ consisting of the identity I_k and the reverse permutation matrix, T_ν, which sends $\underset{\sim}{J}$ to the reverse permutation $(k, k-1,\ldots,2,1)^T$ is of particular importance for ordered data.

The model $\underset{\sim}{\Pi}(\underset{\sim}{\theta})$ is said to be permutation invariant if, for every $\underset{\sim}{\theta}$ in the parameter space, Θ, there exists a $\theta^* \in \Theta$ depending on $\underset{\sim}{\theta}$ and T such that

$$\underset{\sim}{T}\underset{\sim}{\Pi}(\underset{\sim}{\theta})\underset{\sim}{T}^T = \underset{\sim}{\Pi}(\underset{\sim}{\theta}^*) \qquad (6.2.1)$$

for every T in the permutation group $\mathcal{P}$.

Models which satisfy (6.2.1) but only for transformations $\underset{\sim}{T}$ in $\mathcal{P}_\nu$ are said to be palindromic invariant.

Roughly speaking, the definition (6.2.1) means that if the data can be explained by the model $\underset{\sim}{\Pi}(\underset{\sim}{\theta})$, then if the rows and columns are presented in a different order, the new data can be explained by the same model $\underset{\sim}{\Pi}(\underset{\sim}{\theta}^*)$ with just a change in the value of the parameter. Usually $\underset{\sim}{\theta}^*$ can be obtained from $\underset{\sim}{\theta}$ by a permutation transformation, although this restriction is not necessary. Similarly, models which are palindromic invariant can accept the data only in a specified order or its reverse. Clearly, palindromic invariance is a desirable property for ordinal data and permutation invariance is a corresponding property for nominal data.

6.3 Some Parametric Models and their Properties

To contrast the differing properties of permutation invariance and palindromic invariance both types of model are given. Let X_1 and X_2 be the row and column variables respectively where both variables take possible values 1, 2, ..., k. Define

$$\pi_{ij} = pr(X_1 = i,\ X_2 = j) \qquad (1 \leq i, j \leq k)$$

$$P_{ij} = \sum_{\alpha \leq i,} \sum_{\beta > j} \pi_{\alpha\beta} \qquad (1 \leq i < j \leq k)$$

$$Q_{ij} = \sum_{\alpha > j} \sum_{\beta \leq i} \pi_{\alpha\beta} \qquad (1 \leq i < j \leq k)$$

and

$$\Phi_{ij} = \sum_{\alpha \leq i} \sum_{\beta \leq j} \pi_{\alpha\beta} \qquad (1 \leq i,j \leq k.)$$

We note that P_{ij}, Q_{ij} and Φ_{ij} have little meaning except in the context of ordered categories. Three models are considered.

Model I: Quasi-symmetry

$$\pi_{ij} = c \frac{\alpha_i}{\alpha_j} \phi_{ij} \qquad (1 \leq i,j \leq k) \qquad (6.3.1)$$

where $\phi_{ij} = \phi_{ji}$, $\Sigma\Sigma\phi_{ij} = 1$, $\alpha_1 = 1$ and c is a constant to make $\Sigma\Sigma\pi_{ij} = 1$. This model is permutation invariant and is one of the log linear models considered by Bishop et al. (1975) p.286 and others. The log linear version of (6.3.1) in the notation of Bishop et al. is

$$\ln(\pi_{ij}) = u_0 + u_{1(i)} + u_{2(j)} + u_{12(ij)} \qquad (6.3.2)$$

with $u_{12(ij)} = u_{12(ji)}$ and further linear constraints on the parameters which make (6.3.2) equivalent to (6.3.1). The properties of quasi-symmetry are not discussed here since these appear elsewhere in the literature, but we note in passing that marginal homogeneity plus quasi-symmetry implies symmetry.

We now investigate the invariance properties of quasi-symmetry. It is easy to verify from (6.3.2) that if the rows and columns are permuted in the same way the new model is

$$\ln(\pi'_{ij}) = u_0 + u'_{1(i)} + u'_{2(j)} + u'_{12(i,j)}$$

where

$$\pi'_{ij} = \pi_{\alpha_i \alpha_j}, \quad u'_{12(i,j)} = u_{12(\alpha_i,\alpha_j)},$$

$u'_{1(i)} = u_{1(\alpha_i)}$ and $u'_{2(j)} = u_{2(\alpha_j)}$ where $\alpha_1,\ldots,\alpha_k$ is a permutation of the numbers $1,2,\ldots,k$.

Model II: p-symmetry

$$\begin{aligned} P_{ij} &= c\, e^{\frac{1}{2}\Delta_i} \frac{\alpha_i}{\alpha_{j-1}} \psi_{ij} \qquad (1 \le i < j \le k) \\ Q_{ij} &= c\, e^{-\frac{1}{2}\Delta_i} \frac{\alpha_{j-1}}{\alpha_i} \psi_{ij} \qquad (1 \le i < j \le k) \\ \pi_{ii} &= \psi_{ii} \qquad (1 \le i \le k) \end{aligned} \tag{6.3.3}$$

where $\psi_{ij} = \psi_{ji}$, $\alpha_1 = 1$, c is chosen to make $\Sigma\Sigma\pi_{ij} = 1$ and since $\{\psi_{ij}\}$ are probabilities they satisfy the estimability condition

$$\sum_{i=1}^{k} \psi_{ii} + \sum_{|i-j|=1} \psi_{ij} - \sum_{|i-j|=2} \psi_{ij} = 1.$$

This estimability condition is the analogue of the constraint $\Sigma\Sigma\phi_{ij} = 1$ for quasi symmetry as can be seen by examining the multiplicities of inclusion for the k^2 cells in the above expression.

The model for p-symmetry has several interesting properties which are in sharp contrast to quasi-symmetry. Special cases of p-symmetry include marginal homogeneity, conditional symmetry and symmetry. Marginal homogeneity is obtained by putting $\Delta_i = 0$, $i = 1,\ldots,k-1$, but this does not imply symmetry. Conditional symmetry is obtained by putting $\alpha_i = 1$, $i = 1,\ldots,k-1$ and $\Delta_i = \Delta$, $i = 1,\ldots,k-1$, and has the conditional interpretation

$$pr(X_1 = i,\ X_2 = j \mid X_1 < X_2) = pr(X_1 = j,\ X_2 = i \mid X_2 < X_1) \quad (i < j).$$

For a discussion of this model see Bishop et al. (1970), pp.285-286. The model for p-symmetry has a similar conditional interpretation

$$pr(X_1 \leq i,\ X_2 > i \mid X_1 \leq i,\ X_2 > i \text{ or } X_1 > i,\ X_2 \leq i) = e^{\Delta_i}/(1 + e^{\Delta_i}), \tag{6.3.4}$$

since $P_{i,i+1}/Q_{i,i+1} = e^{\Delta_i}$, $i = 1,\ldots,k-1$. This property is related to the paired logistic model described in Chapter 5. The more useful version of p-symmetry has the restriction that all the parameters $\{\Delta_i\}$ are equal. It is easy to show that p-symmetry is not equivalent to the paired logistic model of Chapter 5 in the sense that no distribution for the nuisance parameters can produce p-symmetry. Despite this, they are sufficiently alike and the parameters Δ_i can be interpreted in the same way for both models, i.e. as a measure of the lack of marginal homogeneity.

We now show that p-symmetry is palindromic invariant. After a reverse permutation transformation the relation between the transformed P'_{ij}, Q'_{ij} and the original P_{ij}, Q_{ij} is

$$P'_{ij} = Q_{k-j+1,k-i+1} \quad (i < j)$$

$$Q'_{ij} = P_{k-j+1,k-i+1} \quad (i < j),$$

and

$$\pi'_{ii} = \pi_{k-i+1,k-i+1} \quad (i = j).$$

These can, in turn, be expressed as functions of the original parameters and hence the model is palindromic invariant. It is easy to verify that p-symmetry is not invariant under general permutation transformations.

Model III: The discrete bivariate logistic model (DBL)

This is a model for the cumulative bivariate probability Φ_{ij} which shares some of the properties of the paired logistic model of Chapter 5. It is

$$\begin{aligned}
\Phi_{ij} &= \frac{\alpha_i \beta_j \lambda_{ij}^2}{(1 + \alpha_i \lambda_{ij})(1 + \beta_j \lambda_{ij})} && (1 \leq i,j < k) \\
\Phi_{ik} &= \frac{\alpha_i \lambda_{ik}}{1 + \alpha_i \lambda_{ik}} && (1 \leq i < k) \\
\Phi_{ij} &= \frac{\beta_j \lambda_{kj}}{1 + \beta_j \lambda_{kj}} && (1 \leq j < k) \\
\Phi_{kk} &= 1 \qquad (\lambda_{ij} = \lambda_{ji}) \, .
\end{aligned} \tag{6.3.5}$$

Some estimability conditions are necessary since we can multiply α_i and β_j by arbitrary factors and divide λ_{ij} by the same factors to leave the model unchanged. The simplest constraints are $\lambda_{ik} = 1$, $i = 1,\ldots,k-1$. Since $\lambda_{ik} = \lambda_{ki}$, the cumulative marginal probabilities are now $\alpha_i/(1 + \alpha_i)$ and $\beta_j/(1 + \beta_j)$ so we can easily interpret lack of marginal homogeneity by referring to the logistic distribution. Marginal homogeneity is obtained by putting $\alpha_i = \beta_i$ and, like quasi-symmetry, this also implies symmetry. Sometimes it is useful to summarise the lack of marginal homogeneity in a single parameter. We do this by constraining $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ by $\ln(\alpha_i) - \ln(\beta_i) = d$. Note that d is measured on the marginal scale whereas the paired estimator Δ is measured on a conditional distribution scale. As pointed out in §5.10, d and Δ are related through the correlation ρ.

$$d = \Delta(1 - \rho)^{1/2}$$

The DBL model does not have the property (6.3.4) but it does have the advantage over p-symmetry that it yields estimates of the category boundaries on a logistic scale. Thus the category boundaries are $\frac{1}{2}\ln(\alpha_i\beta_i)$ and the lack of marginal homogeneity is summarised in the parameters $\ln(\alpha_i/\beta_i) = d_i$.

The DBL model is related to the logistic model of Chapter 5 by taking an arbitrary distribution on the nuisance parameter λ so that

$$\Phi_{ij} = \text{pr}(X_1 \leq i,\ X_2 \leq j) = E\left[\frac{\exp(\theta_i + \phi_j - 2\lambda)}{\{1+\exp(\theta_i-\lambda)\}\{1+\exp(\phi_j-\lambda)\}}\right]$$

where expectation is taken over the distribution of λ. Despite the close similarity between the two models, they are not equivalent since the DBL model is not palindromic invariant whereas the paired logistic model is palindromic invariant.

It is sometimes necessary to extend these models to three or more dimensions. Of the models considered here, only quasi symmetry and the DBL model have general multivariate analogues. For quasi-symmetry the general multivariate analogue can be written in log-linear form in the notation of Bishop et al.

$$\ln\pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{j(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(ji)} + u_{123}(ijk)$$

with $u_{12(ij)} = u_{12(ji)}$ etc. $u_{123(ijk)} = u_{123(jik)}$ ... etc.

There are further estimability constraints on the parameters.

The discrete trivariate logistic model is

$$\Phi_{ij\ell} = \frac{\alpha_i\beta_j\gamma_\ell\lambda^3_{ij\ell}}{(1+\alpha_i\lambda_{ij\ell})(1+\beta_j\lambda_{ij\ell})(1+\gamma_\ell\lambda_{ij\ell})} \qquad 1 \leq i,j,\ell < k \qquad (6.3.6)$$

where $\lambda_{ij\ell} = \lambda_{ji\ell} = \lambda_{i\ell j}$ etc., with two and one-dimensional marginals as in (6.3.5). In (6.3.6) homogeneity of the one-dimensional marginals implies symmetry in the two-dimensional marginals and also homogeneity of the three two-dimensional marginals. The multivariate generalisation is straightforward.

In addition, the DBL model provides an estimate of the correlation which is a useful parameter for summarising a further aspect of the data. We note that for positive random variables X, Y with joint density $f(x,y)$ and cumulative distribution $F(x,y)$ the expectation of XY is

$$E(XY) = \int_0^\infty \int xy\, f(x,y)\,dx\,dy$$

$$= \int_0^\infty \int_0^\infty \{1 - F_1(x) - F_2(y) + F(x,y)\}\,dx\,dy,$$

so that

$$\mathrm{cov}(X,Y) = \int_0^\infty \int_0^\infty \{F(x,y) - F_1(x)F_2(y)\}\,dx\,dy \qquad (6.3.7)$$

where $F_1(x) = F(x,\infty)$ and $F_2(y) = F(\infty,y)$.

Now $\Phi_{ij} - \Phi_{ik}\Phi_{kj}$ is analogous to the expression in the integrand of (6.3.7) and is given by

$$\Phi_{ij} - \Phi_{ik}\Phi_{kj} = \frac{\alpha_i\beta_j(\lambda_{ij}-1)\{1+\lambda_{ij}(1+\alpha_i+\beta_j)\}}{(1+\alpha_i)(1+\beta_j)(1+\alpha_i\lambda_{ij})(1+\beta_j\lambda_{ij})} \qquad (6.3.8)$$

which is positive when $\lambda_{ij} > 1$. To approximate (6.3.7) using (6.3.8) we need a finite upper limit for the integral so we assume that X and Y have been transformed on to the range (0,1). Then (6.3.7) is approximated by a weighted sum of $\Phi_{ij} - \Phi_{ik}\Phi_{kj}$. An appropriate expression for the

correlation, ρ, is

$$\rho = 12 \sum_{i=1}^{k} \sum_{j=1}^{k} (a_i - a_{i-1})(b_j - b_{j-1})\{\Phi_{ij} - \Phi_{ik}\Phi_{kj}\} \qquad (6.3.9)$$

where $a_i = \alpha_i/(1 + \alpha_i)$, $b_j = \beta_j/(1 + \beta_j)$, $a_0 = b_0 = 0$ and $a_k = b_k = 1$. What we have done here is to transform the marginals so that they are uniform and then estimate the correlation of the transformed variables.

An easier method of estimating the correlation is to use the estimates of the category medians on the logistic scales and then compute the ordinary product moment correlation coefficients using these scores. The category medians are the logit transforms of $\{\frac{1}{2}(a_i + a_{i-1})\}$ for the rows and $\{\frac{1}{2}(b_j + b_{j-1})\}$ for the columns. The two methods should give similar results.

Finally, table 6.1 gives the number of parameters associated with each model and the degrees of freedom left over for testing the adequacy of the model.

Table 6.1

Number of parameters and degrees of freedom under different models

Model	Restrictions	Parameters	d.f.
quasi symmetry	none	$\frac{1}{2}(k-1)(k+4)$	$\frac{1}{2}(k-1)(k-2)$
symmetry	none	$\frac{1}{2}k(k+1)-1$	$\frac{1}{2}k(k-1)$
p-symmetry	none	$\frac{1}{2}k(k+5)-4$	$\frac{1}{2}(k-2)(k-3)$
p-symmetry	$\Delta_i = \Delta$	$\frac{1}{2}(k-1)(k+4)$	$\frac{1}{2}(k-1)(k-2)$
p-symmetry	$\alpha_i = 1$, $\Delta_i = \Delta$	$\frac{1}{2}(k(k+1)$	$\frac{1}{2}k(k-1)-1$
DBL	none	$\frac{1}{2}(k-1)(k+4)$	$\frac{1}{2}(k-1)(k-2)$
DBL	$\alpha_i/\beta_i = e^{\Delta}$	$\frac{1}{2}k(k+1)$	$\frac{1}{2}k(k-1)-1$

6.4 Example

This example from Stuart, 1953, concerns the unaided distance vision of 7477 women aged 30-39 employed in Royal Ordnance factories in Britain during the period 1943-46. The data have been extensively analysed in the statistical literature; see, for example, Stuart (1953, 1955) and Bishop et al. (1975), p.284, for further analyses. The row variable represents the right eye grade and the column variable the left eye grade. The categories are ordered from highest (1) to lowest (4).

Table 6.2 gives the data and the fitted values under restricted p-symmetry ($\Delta_i = \Delta$) are given in table 6.3. The fitted values under quasi symmetry are given in table 6.4.

The restricted p-symmetry model gives a reasonable fit ($X^2 = 6.2$ on 3 d.f.) and is a slight improvement over quasi-symmetry ($X^2 = 7.3$ on 3 d.f.). The maximum likelihood estimate of Δ is 0.167 with approximate standard deviation .046. Thus the lack of marginal homogeneity is adequately summarised in the single parameter. The interpretation of this result is that the left eye is on average worse than the right eye. The model for p-symmetry does not unfortunately give a measure of correlation or association between the two variables.

A second interpretation of the parameter Δ in the model for p-symmetry involves the conditional odds ratio. This interpretation is the same as that given to the parameter Δ in the paired logistic model of Chapter 5. Thus we define the conditional odds of the event $X \leq j$ to be

$$\text{odds}(X_{i1} \leq j | \lambda_i) = \text{pr}(X_{i1} \leq j | \lambda_i)/\text{pr}(X_{i1} > j | \lambda_i).$$

The model for p-symmetry implies

$$\frac{\text{odds}(X_{i1} \leq j|\lambda_i)}{\text{odds}(X_{i2} \leq j|\lambda_i)} = e^{\Delta_j} \qquad (j = 1,\ldots,k-1)$$

whereas restricted p-symmetry and the logistic model of Chapter 5 both imply

$$\frac{\text{odds}(X_{i1} \leq j|\lambda_i)}{\text{odds}(X_{i2} \leq j|\lambda_i)} = e^{\Delta} \qquad (j = 1,\ldots,k-1)$$

This should be contrasted with the marginal odds ratios

$$\frac{\text{odds}(X_1 \leq j)}{\text{odds}(X_2 \leq j)} = \frac{\text{pr}(X_1 \leq j)\text{pr}(X_2 > j)}{\text{pr}(X_2 \leq j)\text{pr}(X_1 > j)} \qquad (j = 1,\ldots,k-1)$$

which are not constant under any of the models so far considered except when the variables X_1 and X_2 are independent. In general, the marginal odds ratio is smaller than the conditional odds ratio, since pairs of observations (X_1X_2) are positively correlated.

Table 6.5 gives the expected frequencies under the discrete bivariate logistic model (6.3.5) with the restriction that the marginal odds ratios are equal ($\ln \alpha_i - \ln \beta_i = \Delta_M$). It was mentioned without proof in §6.3 that the DBL model is not palindromic. Table 6.6 gives the expected frequencies when the order of the categories is reversed. Although the fitted frequencies are different when the categories are reversed, the parameter estimates are very similar. For table 6.5 the estimates of the log marginal odds ratio and the category boundaries are .046 and (-1.048, 0.237, 2.101) while for table 6.6 the estimates are -0.044 and (-2.101, -.238, 1.048) respectively. Ideally we would expect the sign of the log odds ratio to change and the category boundaries to be in the reverse order with sign changed. The estimated standard deviation of the log odds ratio is .016.

Table 6.2

Unaided distance vision of 7477 women aged 30-39 employed in Royal Ordnance Factories from 1943 to 1946

		Left Eye Grade				
		Highest			Lowest	
		(1)	(2)	(3)	(4)	Total
Right eye	High (1)	1520	266	124	66	1976
	(2)	234	1512	432	78	2256
	(3)	117	362	1772	205	2456
	Low (4)	36	82	179	492	789
	Total	1907	2222	2507	841	7477

Source: Stuart (1953)

Table 6.3

Fitted frequencies under restricted p-symmetry ($\Delta_i = \Delta$)

		Left Eye Grade				
		Highest				
		(1)	(2)	(3)	(4)	Total
Right eye	High (1)	1520.000	264.490	133.212	58.946	1976.648
	(2)	235.589	1512.000	423.397	86.778	2257.764
	(3)	107.301	370.905	1772.000	204.176	2454.382
	Low (4)	43.683	72.667	179.856	492.000	788.206
	Total	1906.573	2220.062	2508.465	841.900	7477.000

$X^2_{ps} = 6.2$ on 3 d.f.

Table 6.4

Fitted frequencies under quasi symmetry

		Left Eye Grade (1)	(2)	(3)	(4)	Total
Right eye: High	(1)	1520.000	263.380	133.584	59.036	1976.000
	(2)	236.620	1512.000	418.986	88.394	2256.000
	(3)	107.416	375.014	1772.000	201.570	2456.000
Low	(4)	42.964	71.606	182.430	492.000	789.000
Total		1907.000	2222.000	2507.000	841.000	7477.000

X^2_{Qs} = 7.3 on 3 d.f.

Table 6.5

Fitted frequencies under discrete bivariate logistic model

	Left Eye Grade (1)	(2)	(3)	(4)	
(1)	1519.493	262.291	135.810	56.141	1973.735
(2)	237.248	1511.828	410.419	88.533	2248.028
(3)	107.402	382.072	1772.115	194.853	2456.442
(4)	44.096	73.327	188.998	492.374	798.795
Total	1908.239	2229.518	2507.342	831.901	7477.000

X^2 = 10.73 on 5 d.f.

Table 6.6

Fitted frequencies under discrete bivariate logistic model with category order reversed.

		Left Eye Grade (4)	(3)	(2)	(1)	Total
Right eye	(4)	492.740	185.897	75.398	45.150	799.185
	(3)	197.803	1770.610	382.227	105.562	2456.202
	(2)	85.341	410.252	1513.048	239.757	2248.398
	(1)	55.334	138.698	259.830	1519.352	1973.214
Total		831.218	2505.457	2230.503	1909.821	7476.999

$X^2 = 10.95$ on 5 d.f.

Under the unrestricted discrete bivariate logistic model there is only a slight reduction in the X^2 goodness of fit statistic ($X^2 = 9.46$ on 3 d.f.). This indicates that the difference between the two marginals is adequately summarised in a single parameter.

Finally, we note, without explanation, the residual pattern in tables 6.3, 6.4, 6.5 and 6.6. Under quasi symmetry and under p-symmetry the diagonals are fitted exactly so that the residuals are zero on the diagonal. Under the D.B.L. model the residuals on the diagonal are not zero but they are small. For all the tables 6.3-6.6 the residual pattern is essentially

0	+	-	+
-	0	+	-
+	-	0	+
-	+	-	0

.

It is unlikely that this pattern is random since it occurs in all the models considered. Therefore there is some aspect of the data, connected with the above residual pattern, which all of the models ignore. It looks like this pattern is connected with the difference, left minus right, between the two margins.

The estimated category medians are -1.903, .367, .969 and 2.854 from the D.B.L. model. These give a product moment correlation estimate of .701 which is somewhat larger than the value of .633 obtained by Stuart (1953) using a variation of Kendall's rank correlation coefficient.

The correlation estimator (6.3.9) gives a value of .674. This particular correlation estimator is more difficult to calculate than the usual product moment estimator. It is difficult to say that any one of these correlation estimators is preferable to all the others, but it is important to be consistent when comparing the degree of association in two tables.

REFERENCES

Abramowitz, M. and Stegun, I.A. (1970) Editors; Handbook of Mathematical Functions. New York, Dover.

Aitchison, J. and Silvey, S.D. (1957). The generalisation of probit analysis to the case of multiple responses. Biometrika, 44, 131-140.

Altham, P.M.E. (1971). The analysis of matched proportions. Biometrika, 58, 561-576.

Andersen, E.B. (1973). Conditional Inference and Models for Measuring. Copenhagen, Mentalhygiejnisk Forlag.

Anderson, T.W. (1959). Some scaling methods and estimation procedures in the latent class model. In Probability and Statistics, pp.9-38, Edited by Ulf Grenander. New York, Wiley.

Armitage, P. (1970). Statistical Methods in Medical Research. Oxford, Blackwell.

Armitage, P. (1975). The use of the cross-ratio in aetiological studies. In Perspectives in Probability and Statistics, papers in honour of M.S. Bartlett on the occasion of his sixty-fifth birthday. pp.349-355. Edited by J. Gani, London, Academic Press.

Ashford, J.R. (1959a). A problem of subjective classification in industrial medicine. Applied Statistics, 8, 168-185.

Ashford, J.R. (1959b). Analysis of data from semi-quantal responses. Biometrics 15, 573-581.

Behnen, K. (1976). Asymptotic comparison of rank tests for the regression problem when ties are present. Ann. Statist. 4, 157-174.

Berkson, J. (1951). Why I prefer logits to probits. Biometrics 7, 327-339.

Bishop, Y.M., Fienberg, S.E. and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, M.I.T. Press.

Bradley, R.A., Katti, S.K. and Coons, I.J. (1962). Optimal scaling for ordered categories. Psychometrika 27, 355-374.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. Ann. Fac. Sci. Univ. Toulouse, 29, 77-192.

Chambers, E.A. and Cox, D.R. (1967). Discrimination between alternative binary response models. Biometrika, 54, 573-578.

Clayton, D.G. (1974). Some odds-ratio statistics for the analysis of ordered categorical data. Biometrika, 61, 525-531.

Cox, D.R. (1958a). Planning of Experiments. London, Wiley.

Cox, D.R. (1958b). Two further applications of a model for binary regression. Biometrika, 45, 562-565.

Cox, D.R. (1966). Some procedures connected with the logistic quality response curve. In Research Papers in Statistics; Feitschrift fur J. Neyman. pp.55-71. Edited by F.N. David. London, Wiley.

Cox, D.R. (1970). The Analysis of Binary Data. London, Methuen.

Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. London, Chapman & Hall.

Davies, O.L. (Editor) (1954). The Design and Analysis of Industrial Experiments. Edinburgh, Oliver and Boyd.

Fienberg, S.E. (1975). The observational study - a review. J. Amer. Statist. Assoc., 70, 521-523.

Fisher, R.A. (1963). Statistical Methods for Research Workers. 13th edition. Edinburgh, Oliver and Boyd.

Gart, J.J. and Zweifel, J.R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika, 54, 181-187.

Gibbons, J.D. (1975). Nonparametric methods for quantitative analysis. New York, Holt, Rinehart and Winston.

Gradshteyn, I.S. and Ryzhik, J.M. (1965). Tables of Integrals, Series and Products. 4th edition. London, Academic Press.

Haberman, S.J. (1974). The Analysis of Frequency Data. Chicago, Chicago University Press.

Hájek, J. (1962). Asymptotically most powerful rank order tests. Ann. Math. Statist. 33, 1124-1147.

Hájek, J. and Sidák, Z (1967). Theory of Rank Tests. London, Academic Press.

Holmes, M.C. and Williams, R.E.O. (1954). The distribution of carriers of Streptococcus pyrogenes among 2413 healthy children. J. Hyg. Camb. 52, 165-179.

Hutchinson, T.P. (1976). The usefulness of signal detection theory in the analysis of ordinal data from diverse fields. Draft report - Traffic Studies Group, University College, London

Jacobsen, M. (1975). Quantifying radiological changes in simple pneumoconiosis. Applied Statistics, 24, 229-249.

Karlin, S. (1968). Total Positivity vol.1. Stanford, Stanford University Press.

Kendall, M.G. and Stuart, A. (1969). The Advanced Theory of Statistics, vol.1, 3rd edition. London, Griffin.

Kendall, M.G. and Stuart, A. (1973). The Advanced Theory of Statistics, vol.2, 3rd edition. London, Griffin.

Lancaster, H.O. (1974). Multivariate binomial distributions. In Studies in Probability and Statistics: Papers in honour of E.J.G. Pitman. Edited by E.J. Williams, Jerusalem Academic Press.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In Measurement and Prediction, vol.4 of Studies in Social Psychology in World War 2. Edited by Stouffer et al. Princeton University Press. Reprinted (1966) New York, Wiley.

Lazarsfeld, P.F. (1955). Recent developments in latent structure analysis. Sociometry 18, 391-403.

Lazarsfeld, P.F. and Henry, N.W. (1918). Latent Structure Analysis. Boston, Houghton Mifflin Company.

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.

Nishisato, S. and Arri, P.S. Nonlinear programming approach to optimal scaling of partially ordered categories. Psychometrika, 40, 525-548.

Patil, G.G. and Joshi, S.W. (1968). A Dictionary and Bibliography of Discrete Distributions. Edinburgh, Oliver and Boyd.

Pearson, K. (1901). Mathematical contributions to the theory of evolution. Phil. Trans. Roy. Soc., 195, 79-150.

Plackett, R.L. (1965). On a class of bivariate distributions. J. Amer. Statist. Assoc., 60, 516-522.

Plackett, R.L. (1974). The Analysis of Categorical Data. London, Griffin.

Roy, S.N. and Sarhan, A.E. (1956). On inverting a class of patterned matrices. Biometrika, 43, 227-231.

Simon, G. (1974). Alternative analyses for the singly ordered contingency table. J. Amer. Statist. Assoc., 69, 971-976.

Snell, E.J. (1964). A scaling procedure for ordered categorical data. Biometrics, 20, 592-607.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. Biometrika, 40, 105-110.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. Biometrics, 42, 412-416.

Tukey, J.W. (1970). Exploratory Data Analysis (Limited preliminary edition). Reading, Mass., Addison Wesley.

Wise, M.E. and Oldham, P.D. Estimating progression of coalworkers' simple pneumoconiosis. Brit. J. Industr. Med., 20, 124-144.