This research was supported by National Institutes of Health, Institute of General Medical Sciences Grants GM-70004-01 and GM-12868-08.

AN ANALYSIS FOR COMPOUNDED LOGARITHMIC-EXPONENTIAL-LINEAR FUNCTIONS OF CATEGORICAL DATA

by

Ronald N. Forthofer and Gary G. Koch

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 801

February 1972

AN ANALYSIS FOR COMPOUNDED LOGARITHMIC - EXPONENTIAL - LINEAR FUNCTIONS OF CATEGORICAL DATA

by

Ronald N. Forthofer
School of Public Health
University of Texas
Houston, Texas 77025, U.S.A.

Gary G. Koch
Department of Biostatistics
University of North Carolina
Chapel Hill, N.C., 27514, U.S.A.

1. INTRODUCTION

One area of application which has become increasingly important to statisticians and other researchers is the analysis of categorical data. Often the principal objective in such investigations is either the testing of appropriate hypotheses or the fitting of simplified models to the multi-dimensional contingency tables which arise when frequency counts are obtained for the respective cross-classifications of specific qualitative variables. Grizzle, Starmer, and Koch [1969] (subsequently abbreviated GSK) have described how linear regression models and weighted least squares can be used for this purpose. The resulting test statistics belong to the class of minimum modified chi-square due to Neyman [1949] which is equivalent to the general quadratic form criteria of Wald [1943]. As such, they have central χ^2-distributions when the corresponding null hypotheses are true. Two alternative approaches to this methodology are that based on maximum likelihood as formulated by Bishop [1969,1971] and Goodman [1970, 1971a, 1971b] and that based on minimum discrimination information as formulated by Ku, Varner, and Kullback [1971].

In each of the previously mentioned papers, primary emphasis was given to the formulation of models and the problems of analysis under various conditions of "no interaction" (see Roy and Kastenbaum [1956] or Bhapkar and Koch [1968a, 1968b]) either in the overall contingency table or the marginals thereof. As a result, attention was essentially restricted to either linear or log-linear functions of cell probabilities. However, there exist a number of common practical situations in which more complex functions must be handled. Examples here include:

1. The analysis of patterns of association in square contingency tables as related to functions of diagonal totals
2. The analysis of rank correlation coefficients as discussed by Goodman and Kruskul [1954, 1959, 1963] and Davis and Quade [1968]
3. The analysis of "ridits" as discussed by Bross [1958]
4. The analysis of partial association as discussed by Mantel and Haenszel [1959] and Mantel [1963].

Each of these problems can be formulated in terms of compounded estimators involving logarithmic, exponential, and linear functions. Thus, such analyses can be undertaken by using an appropriate extension of the GSK approach to this broader class of functions. The remainder of this paper will be concerned with describing and illustrating the use of this more general methodology.

2. STATISTICAL THEORY

We shall assume that there are s populations of elements from which independent random samples of fixed sizes $n_1, n_2, \ldots, n_s$ respectively are selected. The responses of the n_i elements from the i-th population are classified into r categories with n_{ij}, where $j = 1,2,\ldots,r$ denoting the number of elements classified in the j-th response category for the i-th population.

The vector $\underset{\sim}{n}_i$, where $\underset{\sim}{n}_i' = (n_{i1}, n_{i2},\ldots,n_{ir})$, will be assumed to follow the multinomial distribution with parameters n_i and $\underset{\sim}{\pi}_i' = (\pi_{i1}, \pi_{i2},\ldots,\pi_{ir})$. Thus, the relevant product multinomial model is

$$\phi = \prod_{i=1}^{s} \left\{ \frac{n_i!}{\prod_{j=1}^{r} n_{ij}!} \right\} \prod_{j=1}^{r} \pi_{ij}^{n_{ij}} \qquad (2.1)$$

Let $\underset{\sim}{p}_i = (\underset{\sim}{n}_i/n_i)$ and let $\underset{\sim}{p}$ be the compound vector defined by $\underset{\sim}{p}' = (\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_s')$. A consistent estimate for the covariance matrix of $\underset{\sim}{p}$ is given by the block diagonal matrix $\underset{\sim}{V}(\underset{\sim}{p})$ with the matrices $\underset{\sim}{V}_i(\underset{\sim}{p}_i) = (\underset{\sim}{D}_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i')/n_i$ for $i = 1,2,\ldots,s$ representing the main diagonal; here $\underset{\sim}{D}_{\underset{\sim}{p}_i}$ is a diagonal matrix with elements in the vector $\underset{\sim}{p}_i$ on the main diagonal.

Let $F_1(\underset{\sim}{p}), F_2(\underset{\sim}{p}), \ldots, F_u(\underset{\sim}{p})$ be a set of u functions of $\underset{\sim}{p}$, each with partial derivatives up to order two with respect to the elements of $\underset{\sim}{p}$ existing throughout a region containing $\underset{\sim}{\pi} = \underset{\sim}{\varepsilon}(\underset{\sim}{p})$ where $\underset{\sim}{\pi}' = (\underset{\sim}{\pi}_1', \underset{\sim}{\pi}_2', \ldots, \underset{\sim}{\pi}_s')$. If $\underset{\sim}{F} \equiv \underset{\sim}{F}(p)$ is defined by $\underset{\sim}{F}' = \underset{\sim}{F}'(\underset{\sim}{p}) = (F_1(\underset{\sim}{p}), F_2(\underset{\sim}{p}), \ldots, F_u(\underset{\sim}{p}))$, then the covariance matrix of $\underset{\sim}{F}$ can be consistently estimated by $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{H}[\underset{\sim}{V}(\underset{\sim}{p})]\underset{\sim}{H}'$ where $\underset{\sim}{H} = [d\underset{\sim}{F}(\underset{\sim}{x})/d\underset{\sim}{x} \,|\, \underset{\sim}{x} = \underset{\sim}{p}]$. In all applications, the functions comprising $\underset{\sim}{F}$ are chosen so that $\underset{\sim}{V}_{\underset{\sim}{F}}$ is asymptotically non-singular.

The function vector $\underset{\sim}{F}$ is a consistent estimator of $\underset{\sim}{F}(\underset{\sim}{\pi})$. Hence, consideration can then be directed at fitting a linear model $\underset{\sim}{\varepsilon}\{\underset{\sim}{F}(\underset{\sim}{p})\} \doteq \underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}$, where $\underset{\sim}{X}$ is a known (u x t) coefficient matrix of full rank $t \leq u$ and $\underset{\sim}{\beta}$ is an unknown (t x 1) parameter vector. Weighted least squares is applied to determine a BAN estimator $\underset{\sim}{b}$ for $\underset{\sim}{\beta}$ as obtained from the expression

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F} \qquad (2.2)$$

A consistent estimate for the covariance matrix of $\underset{\sim}{b}$ is given by

$$\underset{\sim}{V}_{\underset{\sim}{b}} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1} \qquad (2.3)$$

A goodness of fit statistic for assessing the extent to which the model characterizes the data is the residual sum of squares

$$X^2 = SS(\underset{\sim}{\varepsilon}\{\underset{\sim}{F}\} = \underset{\sim}{X}\underset{\sim}{\beta}) = \underset{\sim}{F}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F} - \underset{\sim}{b}'(\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})\underset{\sim}{b} \qquad (2.4)$$

which has approximately a chi-square distribution with D.F. = (u - t) in large samples under the hypothesis that the model fits. If the model does adequately

describe the data, tests of linear hypotheses with respect to the parameters comprising $\underset{\sim}{\beta}$ can be undertaken. In particular, for a general hypothesis of the form H_o: $\underset{\sim\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0}$ where $\underset{\sim}{C}$ is a known (d×t) matrix of full rank $d \leq t$, a suitable test statistic is

$$X^2 = SS(\underset{\sim\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0}) = \underset{\sim}{b}'\underset{\sim}{C}'[\underset{\sim}{C}(\underset{\sim}{X}'\underset{\sim\sim}{V}_F^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{C}']^{-1}\underset{\sim\sim}{C}\underset{\sim}{b} \qquad (2.5)$$

which has approximately a chi-square distribution with D.F. = d in large samples under H_0.

Although this formulation is quite general, GSK restricted attention to two types of functions. These were linear functions of the type

$$\underset{\sim}{F}(\underset{\sim}{p}) = \underset{\sim\sim}{A}\underset{\sim}{p} \equiv \underset{\sim}{a} \qquad (2.6)$$

where $\underset{\sim}{A}$ is a known (u × rs) matrix and log-linear functions of the type

$$\underset{\sim}{F}(\underset{\sim}{p}) = \underset{\sim}{K}[\underset{\sim\sim\sim}{\log_e}(\underset{\sim\sim}{A}\underset{\sim}{p})] = \underset{\sim}{K}[\underset{\sim\sim\sim}{\log_e}(\underset{\sim}{a})] = \underset{\sim}{f} \qquad (2.7)$$

where $\underset{\sim}{K}$ is a known (k × u) matrix, $\underset{\sim}{A}$ is as defined in (2.6), and $\underset{\sim\sim\sim}{\log_e}$ transforms a vector to the corresponding vector of natural logarithms. For the linear functions in (2.6), the corresponding estimated covariance matrix is

$$\underset{\sim\sim}{V}_{\underset{\sim}{a}} = \underset{\sim}{A}[\underset{\sim\sim}{V}(\underset{\sim}{p})]\underset{\sim}{A}' \qquad (2.8)$$

while for the log-linear functions in (2.7), the corresponding estimated covariance matrix is

$$\underset{\sim\sim}{V}_{\underset{\sim}{f}} = \underset{\sim\sim}{K}\underset{\sim\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{A}[\underset{\sim\sim}{V}(\underset{\sim}{p})]\underset{\sim}{A}'\underset{\sim\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{K}' \qquad (2.9)$$

where $\underset{\sim\sim}{D}_{\underset{\sim}{a}}$ is a diagonal matrix with elements of the vector $\underset{\sim}{a}$ on the main diagonal.

In this paper, two other general classes of functions will be considered. The first of these are exponential functions of the type

$$\underset{\sim}{F}(\underset{\sim}{p}) = \underset{\sim}{Q}(\underset{\sim\sim\sim}{\exp}\{\underset{\sim}{K}[\underset{\sim\sim\sim}{\log_e}(\underset{\sim\sim}{A}\underset{\sim}{p})]\}) = \underset{\sim}{Q}(\underset{\sim\sim\sim}{\exp}\{\underset{\sim}{f}\}) = \underset{\sim}{g} \qquad (2.10)$$

where $\underset{\sim}{Q}$ is a known (q × k) matrix, $\underset{\sim}{A}$ is as defined in (2.6), $\underset{\sim}{K}$ is as defined in (2.7), and $\underset{\sim\sim\sim}{\exp}$ transforms a vector to the corresponding vector of exponential functions

(i.e., of anti-logrithms). A consistent estimate for the covariance matrix of $\underset{\sim}{g}$ is given in (2.11) where $\underset{\sim}{y} = \underset{\sim\sim}{\exp}(\underset{\sim}{f})$.

$$\underset{\sim}{V}_{\underset{\sim}{g}} = \underset{\sim\sim}{Q}\underset{\sim}{D}_{\underset{\sim}{y}}\underset{\sim\sim}{K}\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p})]\underset{\sim}{A}'\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{K}'\underset{\sim}{D}_{\underset{\sim}{y}}\underset{\sim}{Q}' \tag{2.11}$$

The other class of functions are compounded logarithmic functions of the type

$$\underset{\sim}{F}(\underset{\sim}{p}) = \underset{\sim}{L}\{\underset{\sim\sim\sim}{\log}_e[\underset{\sim}{Q}(\underset{\sim\sim\sim}{\exp}\{\underset{\sim}{K}[\underset{\sim\sim\sim}{\log}_e(\underset{\sim}{A}\underset{\sim}{p})]\})]\} = \underset{\sim}{L}\{\underset{\sim\sim\sim}{\log}_e(\underset{\sim}{g})\} = \underset{\sim}{h} \tag{2.12}$$

where $\underset{\sim}{L}$ is a known $(\ell \times q)$ matrix, $\underset{\sim}{A}$ is defined in (2.6), $\underset{\sim}{K}$ is defined in (2.7), and $\underset{\sim}{Q}$ is defined in (2.10). A consistent estimate for the covariance matrix of $\underset{\sim}{h}$ is

$$\underset{\sim}{V}_{\underset{\sim}{h}} = \underset{\sim\sim}{L}\underset{\sim}{D}_{\underset{\sim}{g}}^{-1}\underset{\sim\sim}{Q}\underset{\sim}{D}_{\underset{\sim}{y}}\underset{\sim\sim}{K}\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p})]\underset{\sim}{A}'\underset{\sim}{D}_{\underset{\sim}{a}}^{-1}\underset{\sim}{K}'\underset{\sim}{D}_{\underset{\sim}{y}}\underset{\sim}{Q}'\underset{\sim}{D}_{\underset{\sim}{g}}^{-1}\underset{\sim}{L}' \tag{2.13}$$

From the results in (2.10) and (2.11) together with those in (2.12) and (2.13), it is apparent that logarithmic, exponential, and linear functions can be compounded to a further extent if necessary. Thus, a considerably broader class of functions can be readily analyzed by the use of linear models than was originally demonstrated by GSK.

Finally, this extended methodology can be undertaken by using the same computer program cited in GSK. The only refinements required are suitable options for additional matrix operations. Also, in some cases, if certain n_{ij} are zero, it may be necessary to replace them by $(1/rn_i)$ in order to prevent $\underset{\sim}{V}_{\underset{\sim}{F}}$ from being singular (see GSK and Berkson [1955] for more details in this respect).

3. EXAMPLES

3.1. An analysis of patterns of association related to diagonals

Let us reconsider the first example discussed by GSK which pertained to the unaided distance vision of women aged 30-39. However, here we shall

analyze the contingency table shown in Table 1 in which the frequencies have been reduced to approximately 10% of their original magnitude.

Table 1

Unaided Distance Vision; 747 Women 30-39

Right Eye	Left Eye: Highest Grade (1)	Second Grade (2)	Third Grade (3)	Lowest Grade (4)	Total
Highest Grade (1)	152	27	12	7	198
Second Grade (2)	23	151	43	8	225
Third Grade (3)	12	36	177	20	245
Lowest Grade (4)	4	8	18	49	79
Total	191	222	250	84	747

One question of interest with respect to these data was whether there was a difference between right eye vision and left eye vision which corresponds to a hypothesis of marginal homogeneity. Various test statistics for this condition have been discussed by several authors. On the other hand, there is also some interest in the association between right eye vision and left eye vision. In particular, if there is a strong relationship between vision in the two eyes, then the frequencies along the main diagonal (i.e., cells 11, 22, 33, 44) should be substantially inflated over what would be expected if right eye vision and left eye vision were independent. Moreover, this inflation might also smoothly decrease in a symmetric manner to depressions with respect to other such diagonals which were successively farther above and/or below the main diagonal. Although this formulation of a pattern of association appears complex, the corresponding analysis can be undertaken in a straightforward manner by using the

methodology outlined in Section 2.

Since the data come from a single multinomial population, the vector $\underset{\sim}{p}$ is defined by expression (3.1);

$$\underset{\sim}{p}' = [152\ 27\ 12\ 7\ 23\ 151\ 43\ 8\ 12\ 36\ 177\ 20\ 4\ 8\ 18\ 49]/747; \tag{3.1}$$

the corresponding estimated covariance matrix is given by $\underset{\sim}{V}(\underset{\sim}{p}) = (\underset{\sim}{D}_{\underset{\sim}{p}} - \underset{\sim}{p}\underset{\sim}{p}')/747$. The matrix $\underset{\sim}{A}$ in (3.2) is applied to $\underset{\sim}{p}$ to form diagonal proportions together with suitable marginal proportions which are used to form estimators for the expected values of the diagonals under the hypothesis of independence. Appropriate pairs

$$\underset{13\times16}{A} = \begin{bmatrix} \underset{\sim}{A}_1 \\ 5\times16 \\ \underset{\sim}{A}_2 \\ 8\times16 \end{bmatrix} \text{ where } \underset{\sim}{A}_1 = \begin{bmatrix} 1\,0\,0\,0 & 0\,1\,0\,0 & 0\,0\,1\,0 & 0\,0\,0\,1 \\ 0\,1\,0\,0 & 0\,0\,1\,0 & 0\,0\,0\,1 & 0\,0\,0\,0 \\ 0\,0\,1\,0 & 0\,0\,0\,1 & 0\,0\,0\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 1\,0\,0\,0 & 0\,1\,0\,0 & 0\,0\,1\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 1\,0\,0\,0 & 0\,1\,0\,0 \end{bmatrix} \tag{3.2}$$

$$\underset{\sim}{A}_2 = \begin{bmatrix} 1\,1\,1\,1 & 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 1\,1\,1\,1 & 0\,0\,0\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 1\,1\,1\,1 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,0\,0\,0 & 1\,1\,1\,1 \\ 1\,0\,0\,0 & 1\,0\,0\,0 & 1\,0\,0\,0 & 1\,0\,0\,0 \\ 0\,1\,0\,0 & 0\,1\,0\,0 & 0\,1\,0\,0 & 0\,1\,0\,0 \\ 0\,0\,1\,0 & 0\,0\,1\,0 & 0\,0\,1\,0 & 0\,0\,1\,0 \\ 0\,0\,0\,1 & 0\,0\,0\,1 & 0\,0\,0\,1 & 0\,0\,0\,1 \end{bmatrix}$$

of these marginal proportions are in effect multiplied together on the log-scale by the $\underset{\sim}{K}$-matrix in (3.3) where $\underset{\sim}{J}_4$ is a (4×1) vector of 1's

$$\underset{21\times13}{\underset{\sim}{K}} = \begin{bmatrix} \underset{\sim}{I}_5 & \underset{\sim}{0}_5 & \underset{\sim}{0}_5 & \underset{\sim}{0}_5 & \underset{\sim}{0}_5 & \underset{\sim}{0}_{5,4} \\ \underset{\sim}{0}_{4,5} & \underset{\sim}{J}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & I_4 \\ \underset{\sim}{0}_{4,5} & \underset{\sim}{0}_4 & \underset{\sim}{J}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & I_4 \\ \underset{\sim}{0}_{4,5} & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{J}_4 & \underset{\sim}{0}_4 & I_4 \\ \underset{\sim}{0}_{4,5} & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{J}_4 & I_4 \end{bmatrix} \tag{3.3}$$

$\underset{\sim}{0}_5$ is a (5×1) vector of 0's, $\underset{\sim}{0}_{5,4}$ is a (5×4) matrix of 0's, $\underset{\sim}{I}_5$ is the (5×5) identity matrix, etc. Corresponding values for observed and expected proportions for the diagonals are obtained with the $\underset{\sim}{Q}$ matrix in (3.4) with $\underset{\sim}{A}_1$ as in (3.2). Finally,

$$\underset{10 \times 21}{\underset{\sim}{Q}} = \begin{bmatrix} \underset{\sim}{I}_5 & \underset{\sim}{0}_{5,16} \\ \underset{\sim}{0}_{5,5} & \underset{\sim}{A}_1 \end{bmatrix} \tag{3.4}$$

the appropriate ratios of these estimates on the log-scale result on using

$$\underset{5 \times 10}{\underset{\sim}{L}} = [\underset{\sim}{I}_5 \quad -\underset{\sim}{I}_5] \tag{3.5}$$

For the data in Table 1, the function vector $\underset{\sim}{F}$ obtained from expression (2.12) with $\underset{\sim}{A}$, $\underset{\sim}{K}$, $\underset{\sim}{Q}$, and $\underset{\sim}{L}$ defined by (3.2), (3.3), (3.4), and (3.5) respectively is

$$\underset{\sim}{F} = \begin{bmatrix} \log_e \{(p_{11}+p_{22}+p_{33}+p_{44})/(p_{10}p_{01}+p_{20}p_{02}+p_{30}p_{03}+p_{40}p_{04})\} \\ \log_e \{(p_{12}+p_{23}+p_{34})/(p_{10}p_{02}+p_{20}p_{03}+(p_{30}p_{04})\} \\ \log_e \{(p_{13}+p_{24})/(p_{10}p_{03}+p_{20}p_{04})\} \\ \log_e \{(p_{21}+p_{32}+p_{43})/(p_{20}p_{01}+p_{30}p_{02}+p_{40}p_{03})\} \\ \log_e \{(p_{31}+p_{42})/(p_{30}p_{01}+p_{40}p_{02})\} \end{bmatrix} = \begin{bmatrix} 0.931 \\ -0.576 \\ -1.523 \\ -0.712 \\ -1.462 \end{bmatrix} \tag{3.6}$$

given in (3.6) where $p_{jj'}$ is the element in $\underset{\sim}{p}$ corresponding to the j-th category of right eye vision and the j'-th category of left eye vision, $p_{j.} = \sum_{j'=1}^{4} p_{jj'}$, and $p_{.j'} = \sum_{j=1}^{4} p_{jj'}$. On applying (2.13), an estimated covariance matrix for $\underset{\sim}{F}$ is the matrix $\underset{\sim}{V}_{\underset{\sim}{F}}$ in (3.7).

$$\underset{\sim}{V}_{\underset{\sim}{F}} = \begin{bmatrix} 0.74 & -1.20 & -1.26 & -1.27 & -1.38 \\ & 8.56 & -3.64 & -0.17 & 1.03 \\ & & 43.42 & 1.10 & 3.69 \\ & & & 10.28 & -3.87 \\ & & & & 43.99 \end{bmatrix} \times 10^{-3} \tag{3.7}$$

Finally, it is appropriate to note that diagonal functions corresponding to the upper right hand corner cell (i.e., $\log_e\{p_{14}/p_{10}p_{04}\}$) and the lower left hand corner cell (i.e., $\log_e\{p_{41}/p_{40}p_{01}\}$) have been excluded here to prevent possible singularities in $\underset{\sim}{V}_{\underset{\sim}{F}}$.

If right eye vision and left eye vision were independent, then the observed and expected values for the diagonal proportions would be approximately equal and the elements of $\underset{\sim}{F}$ would each be approximately zero. Hence, a statistical test for the hypothesis of independence can be obtained by fitting a suitable linear model which explains essentially all the variation among the components of $\underset{\sim}{F}$ and then testing whether the parameters of that model all equal zero. For $\underset{\sim}{F}$ in (3.6), one model of interest is

$$\underset{\sim}{\varepsilon}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\beta} \qquad \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \tag{3.8}$$

where β_0 is an overall mean while β_1 and β_2 reflect linear and quadratic components of the variation of the elements in $\underset{\sim}{F}$ with respect to the absolute value of the difference between the grades of vision for the right and left eye where the four grades have been scaled as "1," "2," "3," and "4" as shown in Table 1. For the data in this example, the goodness of fit statistic in (2.4) corresponding to the model (3.8) was $X^2 = 0.95$ with D.F. = 2 which is non-significant ($\alpha = .25$). Thus, this model suitably characterizes the vector $\underset{\sim}{F}$.

The estimate $\underset{\sim}{b}$ of $\underset{\sim}{\beta}$ for the model (3.8) and its estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{b}}$, which have been obtained by using (2.2) and (2.3) are shown in (3.9). Since the hypothesis of independence corresponds to H_0: $\beta_0 = \beta_1 = \beta_2 = 0$, an appropriate

$$\underset{\sim}{b} = \begin{bmatrix} 0.930 \\ -1.924 \\ 0.356 \end{bmatrix} \qquad V_{\underset{\sim}{b}} = \begin{bmatrix} 0.73 & -2.90 & .94 \\ & 34.00 & 19.42 \\ & & 12.62 \end{bmatrix} \times 10^{-3} \qquad (3.9)$$

test statistic is (2.5) with $\underset{\sim}{C} = \underset{\sim}{I}_3$ in which case $X^2 = 1657.22$ with D.F. = 3. Thus, the association between right eye vision and left eye vision is statistically significant ($\alpha = .01$). Moreover, for the hypotheses H_0: $\beta_2 = 0$ and H_0: $\beta_1 = 0$, the results are $X^2 = 10.04$ with D.F. = 1 and $X^2 = 108.83$ with D.F. = 1, both of which are significant ($\alpha = .01$). Hence, the model (3.8) together with its estimated parameters in (3.9) represent a concise and interesting characterization of the pattern of association between right eye vision and left eye vision as reflected by these data.

3.2. An analysis of rank correlation

In Table 2, data are shown for the severity of the 'dumping syndrome,' an undesirable sequela of surgery for duodenal ulcer where the four surgical operations are labelled as

A = Drainage and vagotomy
B = 25% resection (antrectomy) and vagotomy
C = 50% resection (hemigastrectomy) and vagotomy
D = 75% resection.

These data were previously considered by GSK in terms of an analysis of variance of the mean scores corresponding to the respective surgical procedure × hospital combinations with none, slight, and moderate being scaled as 1, 2, and 3. Alternatively, the association between the severity of the dumping syndrome and the surgical procedure can be investigated by using rank correlation coefficients described by Goodman and Kruskal [1954, 1959, 1963] and Davis and Quade [1968] since both of these variables are ordinal. Such analyses can be formulated in

terms of the methodology of Section 2.

Table 2

Hospital	Operation A None	Slight	Mod.	Operation B None	Slight	Mod.	Operation C None	Slight	Mod.	Operation D None	Slight	Mcd.
I	23	7	2	23	10	5	20	13	5	24	10	6
II	18	6	1	18	6	2	13	13	2	9	15	2
III	8	6	3	12	4	4	11	6	2	7	7	4
IV	12	9	1	15	3	2	14	8	3	13	6	4
Total	61	28	7	68	23	13	58	40	12	53	38	16

First of all, the data for each hospital will be assumed to have arisen from separate and independent multinomial populations. Let $\underset{\sim}{p}_i = (\underset{\sim}{n}_i / n_i)$ where $\underset{\sim}{n}_i'$ is the vector of frequencies in the i-th row of Table 2 and n_i is their sum, i = I, II, III, IV. The $\underset{\sim}{A}$ matrix in (3.10) produces all cell probabilities together with sums of appropriate concordant and discordant combinations for each of the separate hospitals. Then, the corresponding probabilities of the various types of concordant

$$\underset{96\times 48}{\tilde{A}} = \begin{bmatrix} \underset{\sim}{I}_{12}\,\underset{\sim}{A}_1 & & & \\ & \underset{\sim}{I}_{12}\,\underset{\sim}{A}_1 & & \\ & & \underset{\sim}{I}_{12}\,\underset{\sim}{A}_1 & \\ & & & \underset{\sim}{I}_{12}\,\underset{\sim}{A}_1 \end{bmatrix} \quad \text{where} \quad \underset{12\times 12}{\underset{\sim}{A}_1} = \begin{bmatrix} 000 & 011 & 011 & 011 \\ 000 & 001 & 001 & 001 \\ 000 & 000 & 011 & 011 \\ 000 & 000 & 001 & 001 \\ 000 & 000 & 000 & 011 \\ 000 & 000 & 000 & 001 \\ 000 & 100 & 100 & 100 \\ 000 & 110 & 110 & 110 \\ 000 & 000 & 100 & 100 \\ 000 & 000 & 110 & 110 \\ 000 & 000 & 000 & 100 \\ 000 & 000 & 000 & 110 \end{bmatrix} \qquad (3.10)$$

and discordant pairs are obtained on the log-scale with the $\underset{\sim}{K}$-matrix in (3.11).

$$\underset{48\times 96}{\underset{\sim}{K}} = \begin{bmatrix} \underset{\sim}{K}_1 \underset{\sim}{I}_{12} & & & \\ & \underset{\sim}{K}_1\underset{\sim}{I}_{12} & & \\ & & \underset{\sim}{K}_1\underset{\sim}{I}_{12} & \\ & & & \underset{\sim}{K}_1\underset{\sim}{I}_{12} \end{bmatrix} \quad \text{where} \quad \underset{12\times 12}{\underset{\sim}{K}_1} = \begin{bmatrix} 100\ 000\ 000\ 000 \\ 010\ 000\ 000\ 000 \\ 000\ 100\ 000\ 000 \\ 000\ 010\ 000\ 000 \\ 000\ 000\ 100\ 000 \\ 000\ 000\ 010\ 000 \\ 010\ 000\ 000\ 000 \\ 001\ 000\ 000\ 000 \\ 000\ 010\ 000\ 000 \\ 000\ 001\ 000\ 000 \\ 000\ 000\ 010\ 000 \\ 000\ 000\ 001\ 000 \end{bmatrix} \tag{3.11}$$

These are combined into overall estimates for the probabilities of concordance and discordance by the use of the $\underset{\sim}{Q}$-matrix in (3.12)

$$\underset{8\times 48}{\underset{\sim}{Q}} = \begin{bmatrix} \underset{\sim}{Q}_1 & & & \\ & \underset{\sim}{Q}_1 & & \\ & & \underset{\sim}{Q}_1 & \\ & & & \underset{\sim}{Q}_1 \end{bmatrix} \quad \text{where} \quad \underset{2\times 12}{\underset{\sim}{Q}_1} = \begin{bmatrix} 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1 \end{bmatrix} \tag{3.12}$$

Finally, the ratio of the probabilities of concordance and discordance for each of the four hospitals are obtained with the $\underset{\sim}{L}$-matrix in (3.13). This final transformation produces indices of association which are analogous to the log cross-product ratios in (2×2) tables; also it will be shown later that these indices can be readily transformed into rank correlation coefficients.

For the data in Table 2, the function vector $\underset{\sim}{F}$ obtained from expression (2.12)

$$\underset{4\times 8}{\underset{\sim}{L}} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \tag{3.13}$$

with $\underset{\sim}{A}$, $\underset{\sim}{K}$, $\underset{\sim}{Q}$ and $\underset{\sim}{L}$ defined by (3.10), (3.11), (3.12), and (3.13) respectively is given in (3.14) together with the corresponding estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{F}}$.

$$\underset{\sim}{F} = \begin{bmatrix} .272 \\ .813 \\ .137 \\ .169 \end{bmatrix} \qquad \underset{\sim}{V}_{\underset{\sim}{F}} = \begin{bmatrix} 5.0464 & & & \\ & 7.9299 & & \\ & & 8.9247 & \\ & & & 8.4030 \end{bmatrix} \times 10^{-2} \qquad (3.14)$$

Once the vector $\underset{\sim}{F}$ has been obtained, a question arises as to the extent to which the rank correlation coefficients comprising $\underset{\sim}{F}$ do not depend on the hospital. This can be investigated by fitting the linear model in (3.15) where β_0 may

$$\underset{\sim}{\varepsilon}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\beta} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 \qquad (3.15)$$

be interpreted as a measure of the partial association between the severity of the dumping syndrome and the extent of the operation. The goodness of fit statistic in (2.4) corresponding to the model (3.15) is $X^2 = 3.72$ with D.F. = 3 which is non-significant ($\alpha = .25$). Hence, the model (3.15) adequately characterizes the vector $\underset{\sim}{F}$ which may be interpreted as meaning that there is no second order interaction between hospitals and surgical operations with respect to their effect on the severity of the dumping syndrome. This result is consistent with that obtained by GSK using another method of analysis. Given this result, the estimate b of β_0 in (3.15) and its estimated variance are

$$b = 0.346 \qquad V_b = 0.018008. \qquad (3.16)$$

For the hypothesis H_0: $\beta_0 = 0$, the test statistic in (2.5) is $X^2 = 6.63$ with D.F. = 1. Thus, this measure of partial association is statistically significant ($\alpha = .01$) which implies that there is a definite relationship between the severity of the dumping syndrome and the surgical operation. This result is also

consistent with that previously obtained by GSK.

Throughout this discussion, the indices $\underset{\sim}{F}$ have been considered. However, using additional operations analogous to (2.10) and (2.12), the actual rank correlation coefficients and their estimated variances can be obtained as

$$\underset{\sim}{G} = (\underset{\sim}{D}_{\underset{\sim}{g}} + \underset{\sim}{I}_4)^{-1} (\underset{\sim}{g} - \underset{\sim}{J}_4) \tag{3.17}$$

$$\underset{\sim}{V}_{\underset{\sim}{G}} = 4(\underset{\sim}{D}_{\underset{\sim}{g}} + \underset{\sim}{I}_4)^{-2} \underset{\sim}{D}_{\underset{\sim}{g}} \underset{\sim}{V}_{\underset{\sim}{F}} \underset{\sim}{D}_{\underset{\sim}{g}} (\underset{\sim}{D}_{\underset{\sim}{g}} + \underset{\sim}{I}_4)^{-2} \tag{3.18}$$

where $\underset{\sim}{g} = \underset{\sim\sim\sim}{\exp}(\underset{\sim}{F})$, $\underset{\sim}{D}_{\underset{\sim}{g}}$ is a diagonal matrix with elements of $\underset{\sim}{g}$ on the diagonal, $\underset{\sim}{J}_4$ is a 4×1 vector of 1's, and $\underset{\sim}{I}_4$ is the 4×4 identity matrix. The resulting rank correlation coefficients $\underset{\sim}{G}$ and $\underset{\sim}{V}_{\underset{\sim}{G}}$ are shown in (3.19). In addition, the partial rank correlation coefficient and its estimated variance can be determined

$$\underset{\sim}{G} = \begin{bmatrix} 0.135 \\ 0.395 \\ 0.068 \\ 0.084 \end{bmatrix} \qquad \underset{\sim}{V}_{\underset{\sim}{G}} = \begin{bmatrix} 1.22 & & & \\ & 1.44 & & \\ & & 2.22 & \\ & & & 2.07 \end{bmatrix} \times 10^{-2} \tag{3.19}$$

by similarly treating b and V_b in (3.16) to obtain

$$\overline{G} = (e^b - 1)/(e^b + 1) = 0.171 \qquad V_{\overline{G}} = \tfrac{1}{4}(1-\overline{G}^2)^2 \, V_b = 0.004242 \tag{3.20}$$

A more direct approach to obtain $\underset{\sim}{G}$ and $\underset{\sim}{V}_{\underset{\sim}{G}}$ is to use a different Q_1 matrix in (3.12). In particular, if

$$\underset{\sim}{Q}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{3.21}$$

and $\underset{\sim}{L}$ is the same as in (3.13), then $\underset{\sim}{G}$ in (3.19) can be obtained from $\underset{\sim}{F}$ based on (3.21) by using $\underset{\sim}{G} = \{2[\underset{\sim\sim\sim}{\exp}(\underset{\sim}{F})] - \underset{\sim}{J}_4\}$. In this event, if $\underset{\sim}{g} = 2[\underset{\sim\sim\sim}{\exp}(\underset{\sim}{F})]$, then $\underset{\sim}{V}_{\underset{\sim}{G}} = \underset{\sim}{D}_{\underset{\sim}{g}} \underset{\sim}{V}_{\underset{\sim}{f}} \underset{\sim}{D}_{\underset{\sim}{g}}$. Finally, both of these methods for the computation of rank correlation

coefficients and their variance yield essentially the same results as would be obtained by using the approach of either Goodman and Kruskal [1954,1959,1963] or Davis and Quade [1968].

3.3 An analysis using "ridits"

This example is based on Schotz's [1966] study of lone drivers in injury producing accidents. The relevant data which have also been analyzed by Bhapkar and Koch [1968a] are shown in Table 3.

Table 3

Driver Group	Accident Type	Accident Severity				Total
		Minor	Moderate	Severe	Extreme	
Lone Driver	Rollover	21	567	1356	644	2588
Lone Driver	Non-Rollover	996	5454	2773	1256	10479
Injured Driver with passengers	Rollover	18	553	1734	869	3174
Injured Driver with passengers	Non-Rollover	679	4561	2516	1092	8848
Total		1714	11135	8379	3861	25089

One way of investigating the variable accident severity is to assign scores to the categories "minor," "moderate," "severe," and "extreme" and then to analyze the average values corresponding to the driver group × accident type categories. A useful method for such scaling, which has been suggested by Bross [1958], is based on "ridits." Here these scores will be derived from the cumulative distribution of the ordinally valued variable accident severity in the combined total population. The methodology of section 2 applies to "ridits" because it enables both their direct calculation as well as their analysis in terms of driver group and accident type effects. Other aspects of their use are discussed by Bross [1958].

Let us now view the data as coming from a single multinomial population. Let $\underset{\sim}{p}$ denote the vector defined in (3.22). The $\underset{\sim}{A}$-matrix in (3.23) produces all

$$\underset{\sim}{p}' = [21,567,1356,644,996,5454,2773,1256,18,553,1734,869,679,4561,2516,1092]/25089 \quad (3.22)$$

cell probabilities, all marginal probabilities for driver group × accident type combinations, and the "ridit" scores. Each cell probability is then multiplied by the appropriate "ridit" and divided by the appropriate marginal probability on the log-scale with the $\underset{\sim}{K}$-matrix in (3.24). Finally, the average "ridit"

$$\underset{24\times16}{\underset{\sim}{A}} = \begin{bmatrix} \underset{\sim}{I}_{16} \\ \\ \underset{\sim}{A}_1 \\ {\scriptstyle 8\times16} \end{bmatrix} \text{ where } \underset{\sim}{A}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 1 & \frac{1}{2} & 0 & 0 & 1 & \frac{1}{2} & 0 & 0 & 1 & \frac{1}{2} & 0 & 0 & 1 & \frac{1}{2} & 0 & 0 \\ 1 & 1 & \frac{1}{2} & 0 & 1 & 1 & \frac{1}{2} & 0 & 1 & 1 & \frac{1}{2} & 0 & 1 & 1 & \frac{1}{2} & 0 \\ 1 & 1 & 1 & \frac{1}{2} & 1 & 1 & 1 & \frac{1}{2} & 1 & 1 & 1 & \frac{1}{2} & 1 & 1 & 1 & \frac{1}{2} \end{bmatrix}$$

scores are derived by performing the additions corresponding to the $\underset{\sim}{Q}$-matrix in (3.25).

$$\underset{16\times24}{\underset{\sim}{K}} = \begin{bmatrix} \underset{\sim}{I}_{16} & \underset{\sim}{K}_1 \end{bmatrix} \text{ where } \underset{16\times8}{\underset{\sim}{K}_1} = \begin{bmatrix} -\underset{\sim}{J}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{I}_4 \\ \underset{\sim}{0}_4 & -\underset{\sim}{J}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{I}_4 \\ \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & -\underset{\sim}{J}_4 & \underset{\sim}{0}_4 & \underset{\sim}{I}_4 \\ \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & \underset{\sim}{0}_4 & -\underset{\sim}{J}_4 & \underset{\sim}{I}_4 \end{bmatrix} \tag{3.24}$$

$$\underset{\sim}{Q} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{3.25}$$

The resulting $\underset{\sim}{F}$ vector obtained from (2.10) with $\underset{\sim}{A}$, $\underset{\sim}{K}$, and $\underset{\sim}{Q}$ as defined in (3.23), (3.24) and (3.25) is given in (3.26) together with the corresponding estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{F}}$. The effects of driver group and accident type on the "ridit" score can be evaluated by fitting the linear model in (3.27). Where β_0 is an overall

$$\underset{\sim}{F} = \begin{bmatrix} 0.649 \\ 0.445 \\ 0.674 \\ 0.459 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{V}_{\underset{\sim}{F}} = \begin{bmatrix} 18.10 & -1.82 & -1.21 & -1.79 \\ & 3.96 & -1.60 & -2.68 \\ & & 13.12 & -1.56 \\ & & & 5.18 \end{bmatrix} \times 10^{-6} \tag{3.26}$$

$$\underset{\sim}{\varepsilon}\{F\} = \underset{\sim\sim}{X\beta} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}. \tag{3.27}$$

mean, β_1 is a driver group effect, and β_2 is an accident type effect. The goodness of fit statistic in (2.4) corresponding to the model (3.27) is $X^2 = 2.31$ with D.F. = 1 which is non-significant ($\alpha = 0.10$). This result implies that there is no interaction between driver group and accident type with respect to the ridit score. The estimate $\underset{\sim}{b}$ of $\underset{\sim}{\beta}$ in (3.27) and its estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{b}}$ are shown in (3.28). Hence, it can be verified that the test

$$\underset{\sim}{b} = \begin{bmatrix} 0.557 \\ -0.009 \\ 0.105 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{V}_{\underset{\sim}{b}} = \begin{bmatrix} 1.15 & 0.06 & 1.53 \\ & 2.53 & 0.30 \\ & & 2.86 \end{bmatrix} \times 10^{-6} \tag{3.28}$$

statistics for the hypotheses H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$ are $X^2 = 31.18$ with D.F.=1 and $X^2 = 3874.29$ with D.F.=1 respectively, both of which are significant ($\alpha = .01$).

The analysis given here is different from the more common use of "ridits" as known constant coefficients rather than scores derived from the observed data. For that approach the data for the four rows of Table 3 are usually viewed as coming from separate and independent multinomial populations. In this event, the value of $\underset{\sim}{F}$ is the same as in (3.26); but $\underset{\sim}{V}_{\underset{\sim}{F}}$ is now the diagonal matrix shown in (3.29). Thus, the goodness of fit test for (3.27) is $X^2 = 2.29$ with D.F. = 1. Also, the estimate $\underset{\sim}{b}$ is essentially the same as that shown in

$$\underset{\sim}{V}_{\underset{\sim}{F}} = \begin{bmatrix} 19.49 & & & \\ & 6.72 & & \\ & & 14.17 & \\ & & & 7.79 \end{bmatrix} \times 10^{-6}. \tag{3.29}$$

(3.28), but the test statistics for $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ are slightly smaller with values of $X^2 = 31.17$ and $X^2 = 3707.00$ respectively. For the case in which there are several independent multinomial populations, this approach has some appeal; but since the "ridit" scores used here have still been determined from the observed data, some questions arise as to whether the components of $\underset{\sim}{F}$ can be viewed as independent. This difficulty can, however, be avoided by using the methods of Section 2 in a somewhat different manner than what was originally described for these data. In particular, the appropriate $\underset{\sim}{A}$ is given in (3.30) with $n_1 = 2588$, $n_2 = 10479$, $n_3 = 3174$, $n_4 = 8848$, and $n = 25089$;

$$\underset{20\times16}{\underset{\sim}{A}} = \begin{bmatrix} \underset{\sim}{I}_{16} \\ \underset{4\times16}{\underset{\sim}{A}_1} \end{bmatrix} \text{ where } \underset{\sim}{A}_1 = \frac{1}{n} \begin{bmatrix} \frac{1}{2}n_1 & 0 & 0 & 0 & \frac{1}{2}n_2 & 0 & 0 & 0 & \frac{1}{2}n_3 & 0 & 0 & 0 & \frac{1}{2}n_4 & 0 & 0 & 0 \\ n_1 & \frac{1}{2}n_1 & 0 & 0 & n_2 & \frac{1}{2}n_2 & 0 & 0 & n_3 & \frac{1}{2}n_3 & 0 & 0 & n_4 & \frac{1}{2}n_4 & 0 & 0 \\ n_1 & n_1 & \frac{1}{2}n_1 & 0 & n_2 & n_2 & \frac{1}{2}n_2 & 0 & n_3 & n_3 & \frac{1}{2}n_3 & 0 & n_4 & n_4 & \frac{1}{2}n_4 & 0 \\ n_1 & n_1 & n_1 & \frac{1}{2}n_1 & n_2 & n_2 & n_2 & \frac{1}{2}n_2 & n_3 & n_3 & n_3 & \frac{1}{2}n_3 & n_4 & n_4 & n_4 & \frac{1}{2}n_4 \end{bmatrix} \tag{3.30}$$

Similarly, the appropriate $\underset{\sim}{K}$ is given in (3.31). If the $\underset{\sim}{Q}$ in (3.25) were used,

$$\underset{16\times20}{K} = [\underset{\sim}{I}_{16} \quad \underset{\sim}{K}_1] \text{ where } \underset{16\times4}{\underset{\sim}{K}_1} = \begin{bmatrix} \underset{\sim}{I}_4 \\ \underset{\sim}{I}_4 \\ \underset{\sim}{I}_4 \\ \underset{\sim}{I}_4 \end{bmatrix} \tag{3.31}$$

then the resulting $\underset{\sim}{V}_{\underset{\sim}{F}}$ would be singular since determination of the ridits in this situation removes one of the degrees of freedom associated with the four driver group x accident type combinations. Hence, to obtain results analogous to those previously given, one can use the $\underset{\sim}{Q}$ matrix in (3.32).

$$\underset{\sim}{Q} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{3.32}$$

The resulting function vector $\underset{\sim}{F}$ obtained from (2.10) with $\underset{\sim}{A}$, $\underset{\sim}{K}$, and $\underset{\sim}{Q}$ as defined in (3.30), (3.31), and (3.32) is given in (3.33) together with the corresponding estimated covariance matrix $\underset{\sim}{V}_F$

$$\underset{\sim}{F} = \begin{bmatrix} -0.040 \\ 0.420 \\ -0.011 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{V}_F = \begin{bmatrix} 48.15 & 6.57 & 19.15 \\ & 46.11 & 4.28 \\ & & 48.15 \end{bmatrix} \times 10^{-6} \qquad (3.33)$$

From these results, it follows that test statistics for driver group effects accident type effects, and their interaction are $X^2 = 32.73$, $X^2 = 3825.63$, and $X^2 = 2.29$ respectively, each with D.F. = 1.

Finally, other types of ridit analysis like that of Williams and Grizzle [1972] based on partial scores can be undertaken within this same framework.

3.4. An analysis of partial association

The data in Table 4 are from the Kihlberg, Narragon, and Campbell [1964] study of the relationship between car size and accident injuries for drivers who were alone at the time of an automobile accident. Here, we shall assume

Table 4

Car Weight	Accident Type	Driver Not Ejected		Driver Ejected	
		Not Severe	Severe	Not Severe	Severe
Small	Collision	350	150	26	23
Small	Roll-over	60	112	19	80
Standard	Collision	1878	1022	111	161
Standard	Roll-over	148	404	22	265

that each of the rows of Table 4 have arisen from separate and independent multinomial populations. In this case, one question of interest is the manner in which the partial association between accident severity and driver ejection after adjustment for car weight depends on accident type. Given that the relationship between accident severity and ejection is similar for the two car weights, this

problem can be dealt with by considering for each accident type, an index which is the ratio of the observed number of severe accidents where the driver was ejected to the expected number of such accidents when accident severity and driver ejection are independent. This index, which has been also extensively discussed by Campbell [1970], can be analyzed by the methods described in Section 2.

The $\underset{\sim}{A}$-matrix in (3.30) produces the cell probability of a severe accident with driver ejected together with the corresponding marginals for each car weight × accident type combination. The $\underset{\sim}{K}$-matrix in (3.31) then produces the appropriate

$$\underset{12\times16}{\underset{\sim}{A}} = \begin{bmatrix} \underset{\sim}{A}_1 & & & \\ & \underset{\sim}{A}_1 & & \\ & & \underset{\sim}{A}_1 & \\ & & & \underset{\sim}{A}_1 \end{bmatrix} \qquad \text{where} \qquad \underset{3\times4}{\underset{\sim}{A}_1} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{3.30}$$

observed and expected proportions on the log-scale. These observed and expected

$$\underset{8\times12}{\underset{\sim}{K}} = \begin{bmatrix} \underset{\sim}{K}_1 & & & \\ & \underset{\sim}{K}_1 & & \\ & & \underset{\sim}{K}_1 & \\ & & & \underset{\sim}{K}_1 \end{bmatrix} \qquad \text{where} \qquad \underset{\sim}{K}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \tag{3.31}$$

values are then added across car weight within each accident type by use of the $\underset{\sim}{Q}$-matrix in (3.32). The index representing the ratio of observed over expected

$$\underset{4\times8}{\underset{\sim}{Q}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.32}$$

for each accident type is obtained on the log-scale with the $\underset{\sim}{L}$ matrix in (3.33). For the data in Table 4, the function vector $\underset{\sim}{F}$ obtained from expression (2.12)

$$\underset{\sim}{L} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{3.33}$$

with $\underset{\sim}{A}$, $\underset{\sim}{K}$, $\underset{\sim}{Q}$, and $\underset{\sim}{L}$ defined by (3.30), (3.31), (3.32). and (3.33) is given in (3.34) together with its estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{F}}$.

$$\underset{\sim}{F} = \begin{bmatrix} 0.433 \\ 0.139 \end{bmatrix} \qquad \underset{\sim}{V}_{\underset{\sim}{F}} = \begin{bmatrix} 50.11 & \\ & 5.47 \end{bmatrix} \times 10^{-4} \qquad (3.34)$$

If the linear model in (3.35) is fitted to $\underset{\sim}{F}$ in (3.34) then the goodness of fit

$$\underset{\sim}{\varepsilon}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\beta} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \beta_0 \qquad (3.35)$$

test indicates whether the partial association index between accident severity and driver ejection adjusted for car weight is the same for the two accident types. The appropriate test statistic obtained from (2.4) is $X^2 = 15.49$ with D.F. = 1 which is significant ($\alpha = .01$). Hence, the relationship of severity to driver ejection is different for collision and roll-over accidents.

In addition, it is of interest to test the significance of each of the indices in $\underset{\sim}{F}$. This can be done by using the linear model in (3.36) and testing the hypotheses H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$. The results based on (2.5) are $X^2 = 37.42$ and $X^2 = 35.32$ respectively, each of which is significant ($\alpha = .01$).

$$\underset{\sim}{\varepsilon}\{\underset{\sim}{F}\} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \qquad (3.36)$$

For the two accident types, each of these test statistics is analogous to what would be obtained using the method of Mantel and Haenszel [1959] or Mantel [1963] except that they are based on the log-scale functions $\underset{\sim}{F}$ with estimated covariance matrix obtained from (2.13). The advantage of the approach suggested here is that it permits the investigation of other models for indices of partial association like that in (3.35).

An alternative approach to this problem, is to use $\underset{\sim}{K}$ in (3.37) to calculate the observed over expected ratio on the log scale for each car weight × accident

$$\underset{\sim}{K} = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{bmatrix} \qquad (3.37)$$

type combination. The resulting function vector $\underset{\sim}{F}$ (there is no $\underset{\sim}{Q}$ or $\underset{\sim}{L}$ matrix here), is then considered in terms of the linear model in (3.38) where β_0 is an overall mean, β_1 is a car weight effect, and β_2 is an accident type effect.

$$\underset{\sim}{\varepsilon}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\beta} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad (3.38)$$

The goodness of fit statistic from (2.4) is $X^2 = 0.09$ with D.F. = 1. Also, test statistics for H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$ from (2.5) are $X^2 = 0.18$ with D.F. = 1 and $X^2 = 39.33$ with D.F. = 1 respectively. Since the goodness of fit statistic and the statistic for H_0: $\beta_1 = 0$ are non-significant ($\alpha = .10$), it follows that this measure of association between accident severity and driver ejection does not depend on car weight. Thus, this assumption which was required by our initial analysis in (3.30) - (3.36) is appropriate for these data. As a result, the test statistic for H_0: $\beta_2 = 0$ and the goodness of fit statistic for (3.35) are both directed at the effect of accident type; and it is useful to note that they have similar values. The advantage of this analysis is that it allows the association between accident severity and ejection to be investigated at the more refined level corresponding to all car weight × accident type combinations. However, in many situations of this kind, the sample sizes are sometimes too small for such an extensive analysis. Hence, the approach which was emphasized at the beginning of this section may be more feasible and appropriate.

Finally, one can note that the results obtained here are somewhat different

from those obtained by Bhapkar and Koch [1968a] in a previous consideration of these data. They used the log - cross-product-ratio measure of association between accident severity and driver ejection and, on applying a model similar to (3.38), found that it was not significantly ($\alpha = .05$) affected by car weight and accident type. Thus, the log-observed-over-expected-ratio for the severe accident with driver ejected category is a more sensitive index to accident type effects than the log-cross-product-ratio in this application. This statement, however, should be interpreted carefully because use of the log-observed-over-expected ratio for a given category is justified only when the corresponding cell in the 2×2 table can be viewed as having the greatest practical importance.

REFERENCES

Berkson, J. [1955]. Maximum likelihood and minimum χ^2 estimates of the logistic function. J. Amer. Statist. Ass. 50, 130-62.

Bhapkar, V.P. and Koch, G.G. [1968a]. Hypotheses of 'no interaction' in multidimensional contingency tables. Technometrics 10, 107-23.

Bhapkar, V.P. and Koch, G.G. [1968b]. On the hypotheses of 'no interaction' in contingency tables. Biometrics 24, 567-94.

Bishop, Y.M.M. [1969]. Full contingency tables, logits, and split contingency tables. Biometrics 25, 383-99.

Bishop, Y.M.M. [1971]. Effects of collapsing multidimensional contingency tables. Biometrics 27, 545-62.

Bross, I.D.J. [1958]. How to use 'ridit' analysis. Biometrics 14, 18-38.

Campbell, B.J. [1970]. Driver injury in automobile accidents involving certain car models. University of North Carolina Highway Safety Research Center Report.

Davis, C.E. and Quade, D. [1968]. "On comparing the correlations within two pairs of variables," Biometrics 24, 987-996.

Goodman, L.A. [1970]. The multivariate analysis of qualitative data: interactions among multiple classifications. J. Amer. Statist. Ass. 65, 226-56.

Goodman, L.A. [1971a]. The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. J. Amer. Statist. Ass. 66, 339-44.

Goodman, L.A. [1971b]. The analysis of multidimensional contingency tables; Stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics 13, 33-61.

Goodman, L.A. and Kruskal, W.H. [1954]. Measures of association for cross classifications. J. Amer. Statist. Ass. 49, 734-64.

Goodman, L.A. and Kruskal, W.H. [1959]. Measures of association for cross classifications. II: Further discussion and references. J. Amer. Statist. Ass. 54, 123-63.

Goodman, L.A. and Kruskal, W.H. [1963]. Measures of association for cross classification. III: Approximate sampling theory. J. Amer. Statist. Ass. 58, 310-64.

Grizzle, J.E., Starmer, C.F., and Koch, C.G. [1969]. Analysis of categorical data by linear models. Biometrics 25, 489-504.

Kihlberg, J.K., Narragon, E.A., and Campbell, B.J. [1964]. Automobile crash injury in relation to car size. Cornell Aeronautical Laboratory Report No. V.J.-1823-R11.

Ku, H.H., Varner, R., and Kullback, S. [1971]. Analysis of multidimensional contingency tables, J. Amer. Statist. Ass. 66, 55-64.

Mantel, N. and Haenszel, W. [1959]. Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst. 22, 719-48.

Mantel, N. [1963]. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. J. Amer. Statist. Ass. 58, 690-700.

Neyman, J. [1949]. Contribution to the theory of the χ^2 test. Pp. 239-73 in Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles.

Roy, S.N. and Kastenbaum, M.A. [1956]. On the hypothesis of no interaction in a multiway contingency table. Ann. Math. Statist. 27, 749-57.

Schotz, W.E. [1966]. The lone driver involved in injury-producing accidents. Cornell Aeronautical Laboratory Report No. VJ-1823-R19.

Wald, A. [1943]. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54, 426-82.

Williams, O. D. and Grizzle, J. E. [1972]. Analysis of contingency tables having ordered response categories. J. Amer. Statist. Ass. 66,