O. Dale Williams' participation in this investigation was supported by N.I.H. Training Grant No. T 01 GM 00038 from the National Institute of General Medical Sciences and James E. Grizzle's participation was supported by Special Fellowship No. 1 F03 HS00003-01 from National Center for Health Services Research and Development and by Contract No. V I 005P-341-A with the Veterans Administration.

ANALYSIS OF CATEGORICAL DATA WITH MORE THAN ONE RESPONSE VARIABLE BY LINEAR MODELS

by

O. Dale Williams
and
James E. Grizzle

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 715

January 1970

OREN DALE WILLIAMS. Analysis of Categorical Data With More Than One Response Variable by Linear Models. (Under the direction of JAMES E. GRIZZLE.)

Using the Grizzle, Starmer, Koch linear models approach to the analysis of categorical data as a basis, the author develops

1) relationships between linear contrasts of the logarithm of cell probabilities and tests for marginal independence,
2) methods utilizing relative risk as a dependent variable in a linear model,
3) two methods applicable to data for which the response variable is ordered, one of the methods using mean scores and the other expressing the response variable in terms of a category effect, and
4) extension of the methods for relative risk and for ordered response variables to the multivariate case.

The methods of analysis reduce to a type of weighted least squares regression and the models discussed are similar to those often considered in the analysis of continuous data.

ACKNOWLEDGMENTS

The author appreciates the assistance of his advisor, Dr. James Grizzle, and of the members of his advisory committee, Drs. R. Kuebler, S. Zyzanski, G. Koch, and H. Smith, Jr. He wishes especially to thank Dr. James Grizzle for originally posing the problem and for thoughtfully discussing it at any time during the research. Special thanks also go to Dr. Roy Kuebler for his careful reading of the paper.

The author gratefully acknowledges the assistance of the Division of Research in Epidemiology and Communication Science of the World Health Organization, where most of this research was done.

He wishes to thank his wife, Judy, and others who have assisted him in various stages of work on this project. In particular, the author expresses appreciation to Mrs. Gay Goss for typing the paper in an expeditious manner.

TABLE OF CONTENTS

LIST OF TABLES

Table | Page

CHAPTER I

REVIEW OF THE LITERATURE AND GENERAL FRAMEWORK

1.1. Introduction

The analysis of multidimensional contingency tables has long had undesirable aspects because of the lack of a general, easily implemented approach. Pearson published his chi-square statistic for contingency tables in 1900, long before the first works on analysis of variance appeared. When the analysis of variance technique appeared, it was developed rapidly to become applicable to many of the problems encountered in the analysis of continuous data. Techniques for contingency tables developed much more slowly and only recently have flexible, widely applicable methods begun to appear.

Our review of the literature on the analysis of complex contingency tables provides a background for the general approach presented by Grizzle, Starmer, and Koch [1969], which we outline in the latter part of this chapter. The remaining chapters extend this method and present applications.

1.2. Notation

For this work we consider contingency tables to represent s independent multinomial populations, each with r response categories. We denote the parameters of the s multinomials by P_{ij} and for convenience call them the expected cell probabilities. The observed cell frequencies are denoted by n_{ij} and a typical table of observed cell frequencies appears in Table 1.1. The related expected cell probabilities appear in Table 1.2. We denote the observed relative frequencies by p_{ij} and call these estimates of the P_{ij} observed cell probabilities. Later, we discuss estimates of the P_{ij} obtained under restrictions related to the structure of the table. We call these estimates restricted estimates and denote them by $\hat{p}_{ij}$. For convenience, we present Table 1.1 and Table 1.2 as two-way tables; however, tables of any dimension can be represented similarly.

TABLE 1.1

OBSERVED CELL FREQUENCIES IN A TYPICAL CONTINGENCY TABLE

Multinomial Populations	Categories of Response 1	2	...	r	Total
1	n_{11}	n_{12}	...	n_{1r}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2r}	$n_{2\cdot}$
⋮	⋮	⋮	⋮⋮⋮	⋮	⋮
s	n_{s1}	n_{s2}	...	n_{sr}	$n_{s\cdot}$

TABLE 1.2

EXPECTED CELL PROBABILITIES FOR TABLE 1.1

Multinomial Populations	Categories of Response 1	2	...	r	Total
1	P_{11}	P_{12}	...	P_{1r}	1
2	P_{21}	P_{22}	...	P_{2r}	1
⋮	⋮	⋮	⋮⋮⋮	⋮	⋮
s	P_{s1}	P_{s2}	...	P_{sr}	1

The likelihood function for Table 1.1 is

$$\prod_{i=1}^{s} \frac{n_{i\cdot}!}{\prod_{j=1}^{r} n_{ij}!} \prod_{j=1}^{r} P_{ij}^{n_{ij}} , \qquad (1.2.1)$$

where

$n_{i\cdot}$ = number in i-th population, not a random variable,

n_{ij} = number in j-th response category of the i-th population, a random variable,

P_{ij} = probability of a response in the j-th response category of the i-th population,

and

$$\sum_{j=1}^{r} P_{ij} = 1 \text{ for all } i,$$

$$\sum_{j=1}^{r} n_{ij} = n_{i\cdot} \text{ for all } i, \text{ and}$$

$$\sum_{i=1}^{s} n_{i\cdot} = N \; .$$

We arrange the rs expected cell probabilities in the vector

$$\underset{1\times rs}{\underset{\sim}{P}'} = [\underset{\sim}{P}_1', \underset{\sim}{P}_2', \ldots, \underset{\sim}{P}_s'], \qquad (1.2.2)$$

where the probabilities for the i-th population are represented as

$$\underset{1\times r}{\underset{\sim}{P}_i} = [P_{i1}, P_{i2}, \ldots, P_{ir}].$$

We denote the observed cell probabilities as

$$p_{ij} = n_{ij}/n_{i\cdot}, \qquad (1.2.3)$$

and express them in the vector

$$\underset{1\times rs}{\underset{\sim}{p}'} = [\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_s'], \tag{1.2.4}$$

where

$$\underset{\sim}{p}_i' = [p_{i1}, p_{i2}, \ldots, p_{ir}].$$

The vector $\underset{\sim}{p}_i$ for the i-th population has the covariance matrix $\underset{\sim}{V}(\underset{\sim}{P}_i)$ which, since it is symmetric, we express as

$$\underset{r\times r}{\underset{\sim}{V}(\underset{\sim}{P}_i)} = \frac{1}{n_{i\cdot}} \begin{bmatrix} P_{i1}(1-P_{i1}) & & & \\ -P_{i1}P_{i2} & P_{i2}(1-P_{i2}) & & \\ \vdots & \vdots & \ddots & \\ -P_{i1}P_{ir} & -P_{i2}P_{ir} & \cdots & P_{ir}(1-P_{ir}) \end{bmatrix}. \tag{1.2.5}$$

We denote the sample estimate of $\underset{\sim}{V}(\underset{\sim}{P}_i)$ by $\underset{\sim}{V}(\underset{\sim}{p}_i)$, which consists of $\underset{\sim}{V}(\underset{\sim}{P}_i)$ with the P_{ij} replaced by p_{ij}. Since the s multinomials are assumed to be independent, the covariance matrix for $\underset{\sim}{p}$ can be represented as the symmetric block diagonal matrix

$$\underset{rs\times rs}{\underset{\sim}{V}(\underset{\sim}{P})} = \begin{bmatrix} \underset{\sim}{V}(\underset{\sim}{P}_1) & & & \\ 0 & \underset{\sim}{V}(\underset{\sim}{P}_2) & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \underset{\sim}{V}(\underset{\sim}{P}_s) \end{bmatrix}. \tag{1.2.6}$$

We denote the sample estimate of $\underset{\sim}{V}(\underset{\sim}{P})$ by $\underset{\sim}{V}(\underset{\sim}{p})$.

Bhapkar and Koch [1968a] relate this notation to data analysis by considering the s populations to be factors such as treatments or blocks which have fixed marginal totals and the response categories as responses that have random marginal totals. We shall alter this nomenclature by using the label population to refer to treatments or blocks with fixed marginals. The label factor can then be used to denote a collection of response categories. This slight change allows one to use standard factorial analysis terminology to discuss multifactor interactions both when factors represent subsets of response categories within a single multinomial and when levels of a factor correspond to populations.

1.3. Hypotheses

Bartlett [1935] proposed a hypothesis for the 2×2×2 table that is extended readily to tables of higher dimension and thus allows Pearson's statistic to be used for tables with more than two dimensions. Bartlett formulated his hypothesis by considering one multinomial population with eight categories of response. He supposed that this population can be represented in a 2×2×2 table where 2×2×2 indicates the presence of three variables, each at two levels. The expected cell probabilities appear below in two 2×2 tables where we represent the levels of the three variables by A and $\bar{A}$, B and $\bar{B}$, C and $\bar{C}$, with $\sum_{j=1}^{8} P_j = 1$.

C		
	B	$\bar{B}$
A	P_1	P_2
$\bar{A}$	P_3	P_4

$\bar{C}$		
	B	$\bar{B}$
A	P_5	P_6
$\bar{A}$	P_7	P_8

(1.3.1)

Bartlett's hypothesis is essentially a definition of interaction and much of the recent literature on the analysis of multidimensional contingency tables is based on it. Bartlett proposed a measure of three-way interaction based on differences among the two-way interactions. In more general terms, he considered a measure of association between two factors, namely two-factor interaction, and examined its variation over levels of the third factor. With regard to (1.3.1) the conditions for no two-factor interaction between A and B are

$$\text{for } C, \qquad P_1/P_2 = P_3/P_4;$$

$$\text{for } \bar{C}, \qquad P_5/P_6 = P_7/P_8;$$

so that the ratios P_1P_4/P_2P_3 and P_5P_8/P_6P_7 are measures of two-factor interaction. It follows that a condition for no three-factor interaction is

$$P_1P_4/P_2P_3 = P_5P_8/P_6P_7,$$

or

$$P_1P_4P_6P_7/P_2P_3P_5P_8 = 1. \qquad (1.3.2)$$

Simpson [1951] suggested that definitions of high-order interactions should be independent of the labelling of the variables. For this example his proposal implies that the ratio of the two-factor interactions between A and B for the levels of C should be the same as the ratio of the two-factor interactions between A and C for the levels of B, and so forth. Simpson shows that the conditions specified by Bartlett are symmetric and further states that functions of this ratio, such as its logarithm, are also symmetric. The logarithm of this ratio has appeal because it provides a linear contrast in terms of the $\ell n\ P_j$ analogous to the contrast commonly encountered in factorial experiments. Ratios similar to (1.3.2) or their logarithms have been considered by Norton [1945], Roy and Kastenbaum [1956], Kastenbaum and Lamphier [1959], Bhapkar [1966], Bhapkar and Koch [1968a, 1968b] and others. These ratios provide a technique for expressing interaction which can be generalized easily to higher-dimension tables.

Other measures of association for contingency tables have been presented, for example, by Goodman and Kruskal [1954, 1959] and Mosteller [1968]. However, they are not linearized easily and do not fit into a general approach so readily as the function discussed above.

In general, hypotheses for contingency tables have been made either directly or indirectly in terms of functions of $\underset{\sim}{P}$ in the form

$$\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{0}, \tag{1.3.3}$$

or

$$\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{X}\,\underset{\sim}{\xi}, \tag{1.3.4}$$

where

$$[\underset{\sim}{F}(\underset{\sim}{P})]' = [f_1(\underset{\sim}{P}),\ f_2(\underset{\sim}{P}),\ \ldots,\ f_t(\underset{\sim}{P})]$$

and where the $f_m(\underset{\sim}{P})$, $m = 1, 2, \ldots, t$, are functions of the P_{ij}, $\underset{\sim}{X}$ is a matrix of constants, perhaps a design matrix related to the structure of the table, and $\underset{\sim}{\xi}$ is a vector of unknown model parameters. For example, we can express the hypothesis given by (1.3.2) according to (1.3.3) by letting

$$\underset{\sim}{F}(\underset{\sim}{P}) = \ell nP_1 - \ell nP_2 - \ell nP_3 + \ell nP_4$$
$$- \ell nP_5 + \ell nP_6 + \ell nP_7 - \ell nP_8.$$

We can also examine this hypothesis in terms of (1.3.4) by testing the fit of a model in terms of ℓnP_j such as

$$\ell nP_j = \mu + \alpha + \beta + \gamma + \alpha\beta + \alpha\gamma + \beta\gamma + \alpha\beta\gamma, \qquad (1.3.5)$$

where

μ = overall effect,

α = effect of the A variable on ℓnP_j,

β = effect of the B variable on ℓnP_j,

γ = effect of the C variable on ℓnP_j,

$\alpha\beta$ = effect of AB interaction on ℓnP_j,

and the remaining terms represent the appropriate interaction effects. We obtain this model by letting $\underset{\sim}{F}(\underset{\sim}{P})$ from (1.3.4) construct the ℓnP_j and by using $\underset{\sim}{X}$ and $\underset{\sim}{\xi}$ as matrices typical of experimental design.

Various special cases of (1.3.3) and (1.3.4) have been proposed. Berkson [1944, 1946, 1955, 1968] and others considered a model for the parameters of the likelihood function in (1.2.1) for the special case with s binomials. They considered each binomial to have its parameters P_i and Q_i, $i = 1, 2, \ldots, s$, expressed in terms of the logistic function, that is,

$$P_i = \frac{1}{1 + e^{-(\alpha+\beta x_i)}},$$
$$Q_i = 1 - P_i = \frac{e^{-(\alpha+\beta x_i)}}{1 + e^{-(\alpha+\beta x_i)}} . \qquad (1.3.6)$$

Since $L_i = \ell nP_i - \ell nQ_i = \alpha + \beta\, x_i$ we can write this model in the form of (1.3.4) and investigate hypotheses in terms of α and β.

Plackett [1962], Birch [1963], Goodman [1963b], Mantel [1966], Darroch [1962], and others considered models similar to (1.3.5) which explicitly expressed the logarithm of cell probabilities as linear combinations of "effects."

So far in this section we have dealt with reasonably explicit models and hypotheses and have ignored an approach that does not include the explicit statement of a model. It should be mentioned, however, that some of the literature on the analysis of complex contingency tables is based on the partitioning of the chi-square statistic. The work of investigators such as Irwin [1949] provided the theoretical groundwork for partitioning the statistic. Lancaster [1951], Goodman [1968], and others have used these methods to test hypotheses implicitly the same as those given by (1.3.3) and (1.3.4).

1.4. Estimates

In general, hypotheses such as (1.3.3) or (1.3.4) have been tested by obtaining estimates of the P_{ij} under the constraint that the null hypothesis is true and by then comparing these estimates with the observed cell probabilities. We call estimates of the P_{ij} obtained under the constraint that the null hypothesis is true "restricted estimates" and we express them in the vector

$$\hat{\underset{\sim}{p}}' = [\hat{\underset{\sim}{p}}_1', \hat{\underset{\sim}{p}}_2', \ldots, \hat{\underset{\sim}{p}}_s'], \tag{1.4.1}$$

where the vector for each multinomial population is

$$\hat{\underset{\sim}{p}}_i' = [\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{ir}].$$

Three basic types of estimates have been considered:

1) maximum likelihood estimates (MLEs),

2) minimum χ^2 estimates, where

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{i.}(p_{ij} - \hat{p}_{ij})^2}{\hat{p}_{ij}} ,$$

3) minimum χ_1^2 estimates, where

$$\chi_1^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{i.}(p_{ij} - \hat{p}_{ij})^2}{p_{ij}} .$$

Neyman [1949] has showed that if $\underset{\sim}{F}(\underset{\sim}{P})$ has continuous partial derivatives up to the second order with respect to $\underset{\sim}{P}$ estimates obtained by any of the three methods have the following properties:

1) they are functions of the p_{ij} and do not depend directly on N,

2) they are consistent,

3) as $N \to \infty$ their distribution tends to be normal,

4) the variances of other estimates satisfying 2) and 3) are greater than or equal to their variances, and

5) considered as functions of the P_{ij}, they possess continuous partial derivatives with respect to the P_{ij}.

Such estimates are best asymptotically normal (BAN); hence the three estimates are asymptotically equivalent. This is an important property because we can obtain minimum χ_1^2 estimates by direct solution of linear equations, whereas MLE's and minimum χ^2 estimates generally require iteration. Neyman also showed that, using minimum χ_1^2, we can obtain BAN estimates of $\underset{\sim}{P}$ for a nonlinear function by linearizing the function in a Taylor's series expansion.

Several techniques provide restricted estimates of the cell probabilities. Bartlett [1935] presented a third-degree equation in one unknown which can be solved by iterative techniques to obtain MLE's. Norton [1945] extended Bartlett's concept and provided estimates for the $2\times2\times t$ table. Norton's technique involved solving $t-1$ simultaneous third degree equations in $t-1$ unknowns. Roy and Kastenbaum [1956] considered Bartlett's hypothesis in the general $r\times s\times t$ contingency table and provided a system of $(r-1)(s-1)(t-1)$ simultaneous fourth degree equations in $(r-1)(s-1)(t-1)$ unknowns to obtain MLE's. Kastenbaum and Lamphiear [1959] simplified the problem by applying Newton's method of functional iteration to the equations given by Norton and Roy and Kastenbaum. Birch [1963] used essentially the same iterative scheme that Roy and Kastenbaum presented to get MLE's for the three-way table.

Contingency tables that have missing entries constitute special cases, and Watson [1955], Goodman [1968], Bishop [1969], and Bishop and Fienberg [1969] have presented estimation techniques for obtaining MLE's for different patterns of missing cells.

Other authors considered obtaining estimates for the logistic model described by (1.3.6). Grizzle [1961] provided a technique for MLE's while Berkson [1944, 1946, 1955, 1968] and Hitchcock [1962], among others, discussed minimum χ_1^2 estimates for this model.

Minimum χ^2 estimates rarely have been used, perhaps because, like MLE's, they require iterative techniques that can become laborious. Neyman [1949] and Berkson [1955] presented techniques for obtaining minimum χ^2 estimates, but these papers emphasized minimum χ_1^2 techniques. Berkson indicated that minimum χ_1^2 estimates for (1.3.6) could be obtained by general least squares techniques and others have extended this approach. Grizzle, Starmer, and Koch [1969] consider a method based on the χ_1^2 statistic and least squares techniques.

1.5. Test Statistics

As indicated in Section 1.3, we can formulate various hypotheses in terms of functions of the elements of $\underset{\sim}{P}$. Neyman [1949] has showed that if these functions have continuous partial derivatives up to the second order with respect to $\underset{\sim}{P}$ and the matrix

$$\underset{t\times rs}{\underset{\sim}{H}} = \left[\frac{\partial \underset{\sim}{F}(\underset{\sim}{P})}{\partial \underset{\sim}{P}} \Bigg|_{\underset{\sim}{P} = \underset{\sim}{p}} \right], \quad t \leq s(r-1), \tag{1.5.1}$$

is of full rank, then null hypotheses can be tested by

1) the Pearson χ^2 statistic

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{i\cdot}(p_{ij} - \hat{p}_{ij})^2}{\hat{p}_{ij}},$$

2) the minimum χ_1^2 statistic

$$\chi_1^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{i\cdot}(p_{ij} - \hat{p}_{ij})^2}{p_{ij}}, \text{ or}$$

3) the likelihood ratio χ^2 statistic,

$$\chi_\ell^2 = -2 \sum_{i=1}^{s} \sum_{j=1}^{r} n_{ij} \, \ell n(n_{ij}/n_{i\cdot}\hat{p}_{ij}),$$

provided the $\hat{p}_{ij}$ are BAN estimates of the P_{ij}. Neyman further demonstrated that each of the above statistics has a limiting chi-square distribution with t degrees of freedom as $N \to \infty$ provided:

1) the ratios $n_{i\cdot}/N$ remain constant as $N \to \infty$,

2) there exists at least one solution to (1.3.3) such that $P_{ij} > 0$ for all i, j, and

3) the null hypothesis is true.

Neyman also showed that the three statistics are asymptotically equivalent since, regardless of whether the null hypothesis is true, the probability that any two of them contradicts the other tends to zero as $N \to \infty$.

Some authors have used other chi-square statistics to test hypotheses that can be expressed in the form of (1.3.3). For example, Wald [1943] proposed a general statistic which Bhapkar [1966] applied to categorical data. Wald considers only r-1 of the r p_{ij} given for the $\underset{\sim}{P}_i$ of (1.2.2). If we let

$$\underset{\sim}{p}_i^{*\prime} = [p_{i1}, p_{i2}, \ldots, p_{i(r-1)}],$$

then for the hypothesis in (1.3.3), Wald's statistic is

$$\chi_w^2 = \underset{(1\times t)}{\underset{\sim}{F}'(\underset{\sim}{p}^*)} \; \underset{(t\times t)}{[\underset{\sim}{H}^* \underset{\sim}{C} \underset{\sim}{H}^{*\prime}]^{-1}} \; \underset{(t\times 1)}{\underset{\sim}{F}(\underset{\sim}{p}^*)}$$

where $\underset{\sim}{F}(\underset{\sim}{p}^*)$ is $\underset{\sim}{F}(\underset{\sim}{P}^*)$ evaluated at $\underset{\sim}{P}^* = \underset{\sim}{p}^*$, $\underset{t\times s(r-1)}{\underset{\sim}{H}^*}$ is similar to (1.5.1),

$$\underset{(r-1)s\times(r-1)s}{\underset{\sim}{C}} = \begin{bmatrix} \frac{1}{n_{1\cdot}}(\underset{\sim}{Q}_1 - \underset{\sim}{p}_1^*\underset{\sim}{p}_1^{*\prime}) & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \frac{1}{n_{2\cdot}}(\underset{\sim}{Q}_2 - \underset{\sim}{p}_2^*\underset{\sim}{p}_2^{*\prime}) & \cdots & \underset{\sim}{0} \\ \vdots & \vdots & \cdots & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \frac{1}{n_{s\cdot}}(\underset{\sim}{Q}_s - \underset{\sim}{p}_s^*\underset{\sim}{p}_s^{*\prime}) \end{bmatrix}$$

and $\underset{\sim}{Q}_i$ = diagonal $[p_{i1}, p_{i2}, \ldots, p_{i(r-1)}]$. This statistic has a limiting chi-square distribution as $N \to \infty$ provided the ratios $n_{i\cdot}/N$ remain constant and the null hypothesis is true. Bhapkar [1966] demonstrated that the Wald statistic is algebraically identical to the χ_1^2 statistic when $\underset{\sim}{F}(\underset{\sim}{P})$ is linear or when it is nonlinear and Neyman's [1949] linearization is utilized. This statistic provides a consistent estimate of the variance of $\underset{\sim}{F}(\underset{\sim}{p})$ (Goodman [1963a]) regardless of whether the null hypothesis is true.

Woolf [1955] provided yet another statistic. The version of his statistic for a 2×2 table similar to the one for C in (1.3.1) and for the hypothesis

$$\ell n\ P_1 - \ell n\ P_2 - \ell n\ P_3 + \ell n\ P_4 = 0,$$

is

$$X^2 = y_1^2/V_1,$$

where

$$y_1 = \ell n\ n_1 - \ell n\ n_2 - \ell n\ n_3 + \ell n\ n_4, \tag{1.5.2}$$

$$V_1 = 1/n_1 + 1/n_2 + 1/n_3 + 1/n_4,$$

and the n_j are the observed cell frequencies corresponding to the P_j. This statistic has a limiting chi-square distribution with one degree of freedom provided the null hypothesis is true.

Plackett [1962] extended Woolf's log-linear derivation to the hypothesis given by Norton [1945] for the 2×2×t table and to the hypotheses given by Roy and Kastenbaum [1956] and by Kastenbaum and Lamphiear [1959] for the r×s×t table. For the 2×2×t arrangement, Plackett considered a log-linear contrast as given by (1.5.2) for each of the t 2×2 tables; the contrasts are

$$y_k = \log n_{1k} - \log n_{2k} - \log n_{3k} + \log n_{4k}, \quad k = 1,2,\ldots,t,$$

and the variance of the k-th contrast is

$$V_k = 1/n_{1k} + 1/n_{2k} + 1/n_{3k} + 1/n_{4k}.$$

The statistic Plackett gives is

$$X^2 = \sum_{k=1}^{t} (y_k^2/V_k) - \left(\sum_{k=1}^{t} y_k/V_k\right)^2 \Bigg/ \sum_{k=1}^{t} V_k, \quad t \geq 2,$$

which has a limiting chi-square distribution with t-1 degrees of freedom. Plackett also gives a test statistic that utilizes log-linear contrasts for the r×s×t table.

Goodman [1963b] has shown that Plackett's and Woolf's statistics are equivalent to the Wald statistic. Since Bhapkar [1966] showed that Wald's statistic and χ_1^2 were algebraically identical, it follows that all the above are χ_1^2 statistics.

For the model described by (1.3.6) Berkson [1955] proposed a version of the χ_1^2 statistic that is

$$X^2 = \sum_{i=1}^{s} n_i p_i q_i \, (\ell_i - \hat{\ell}_i)^2,$$

where

$$\ell_i = \log(p_i/q_i),$$

$$\hat{\ell}_i = a + bx_i,$$

$\hat{\ell}_i$ representing the estimated value of $L_i = \log\,(P_i/Q_i) = \alpha + \beta\, x_i$, with a and b being estimates of α and β.

Grizzle, Starmer, and Koch [1969] have presented an approach similar to Berkson's, but considerably more general. Since their method provides the nucleus for the present work, we outline it briefly.

1.6. The Grizzle, Starmer, and Koch Method

The Grizzle, Starmer, and Koch [1969] method makes use of the minimum χ_1^2 statistic introduced by Neyman [1949] and is an expansion of material presented by Bhapkar [1966] and Bhapkar and Koch [1968a, 1968b]. Grizzle, Starmer, and Koch consider the hypotheses $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{0}$ and $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{X}\,\underset{\sim}{\xi}$ and they discuss two forms of $\underset{\sim}{F}(\underset{\sim}{P})$:

$$\underset{(t\times 1)}{\underset{\sim}{F}(\underset{\sim}{P})} = \underset{(t\times rs)}{\underset{\sim}{A}}\ \underset{(rs\times 1)}{\underset{\sim}{P}}, \quad t \leq s(r-1), \text{ and} \tag{1.6.1}$$

$$\underset{(t\times 1)}{\underset{\sim}{F}(\underset{\sim}{P})} = \underset{(t\times u)}{\underset{\sim}{K}}\ \ln[\underset{(u\times rs)}{\underset{\sim}{A}}\ \underset{(rs\times 1)}{\underset{\sim}{P}}], \quad u \leq rs, \tag{1.6.2}$$

where $\underset{\sim}{A}$ and $\underset{\sim}{K}$ are matrices of predetermined constants and $\ln \underset{\sim}{A}\,\underset{\sim}{P}$ denotes a $u\times 1$ vector whose elements are the natural logarithms of the elements of the vector $\underset{\sim}{A}\,\underset{\sim}{P}$. They denote the vector $\underset{\sim}{F}(\underset{\sim}{P})$ by

$$[\underset{\sim}{F}(\underset{\sim}{P})]' = [f_1(\underset{\sim}{P}),\ f_2(\underset{\sim}{P}),\ \ldots,\ f_t(\underset{\sim}{P})], \quad t \leq s(r-1), \tag{1.6.3}$$

where the partial derivatives of $f_m(\underset{\sim}{P})$, with respect to $\underset{\sim}{P}$, up to the second order are assumed to exist. Other forms of $\underset{\sim}{F}(\underset{\sim}{P})$ are certainly possible; however, those given in (1.6.1) and (1.6.2) cover a wide variety of problems.

Before proceeding with analyses, we need the estimated variance of $\underset{\sim}{F}(\underset{\sim}{p})$, where

$$[\underset{\sim}{F}(\underset{\sim}{p})]' = [f_1(\underset{\sim}{p}),\ f_2(\underset{\sim}{p}),\ \ldots,\ f_t(\underset{\sim}{p})] \tag{1.6.4}$$

and $f_{\sim m}(\underset{\sim}{p})$ is $f_m(\underset{\sim}{P})$ evaluated at $\underset{\sim}{P} = \underset{\sim}{p}$. We express the estimated variance of $\underset{\sim}{F}(\underset{\sim}{p})$ as

$$\hat{\text{Var}}\ \underset{\sim}{F}(\underset{\sim}{p}) = \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}' = \underset{\sim}{S}, \tag{1.6.5}$$

where $\underset{\sim}{V}(\underset{\sim}{p})$ is the estimated variance of $\underset{\sim}{p}$ (1.2.6) and

$$\underset{\sim}{H} = \left[\left.\frac{\partial \underset{\sim}{F}(\underset{\sim}{p})}{\partial \underset{\sim}{P}}\right|_{\underset{\sim}{P}=\underset{\sim}{p}}\right] .$$

If $\underset{\sim}{F}(\underset{\sim}{p}) = \underset{\sim}{A}\ \underset{\sim}{p}$ then $\underset{\sim}{H} = \underset{\sim}{A}$ and

$$\underset{\sim}{S} = \underset{\sim}{A}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{A}'.$$

When $\underset{\sim}{F}(\underset{\sim}{p})$ is the nonlinear function $\underset{\sim}{K}\ \ell n\ \underset{\sim}{A}\ \underset{\sim}{p}$, we can obtain the estimated large-sample variance by expanding $f_m(\underset{\sim}{P})$ about the point $P = \underset{\sim}{p}$ in a Taylor's series. Neyman [1949] outlined this procedure and showed that, for large samples $\underset{\sim}{F}(\underset{\sim}{P})$ can be written as

$$\underset{\sim}{F}(\underset{\sim}{P}) \approx \underset{\sim}{F}(\underset{\sim}{p}) + \underset{\sim}{H}(\underset{\sim}{P}-\underset{\sim}{p})$$

or

$$\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{F}(\underset{\sim}{P}) \approx \underset{\sim}{H}(\underset{\sim}{p}-\underset{\sim}{P}). \tag{1.6.6}$$

Then the large-sample variance of $\underset{\sim}{F}(\underset{\sim}{P})$ is

$$\begin{aligned}
\underset{\sim}{\text{Var}}[\underset{\sim}{F}(\underset{\sim}{p})] &\approx E\{[\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{F}(\underset{\sim}{p})]\ [\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{F}(\underset{\sim}{P})]'\} \\
&= E[\underset{\sim}{H}(\underset{\sim}{p}-\underset{\sim}{P})\ (\underset{\sim}{p}-\underset{\sim}{P})'\ \underset{\sim}{H}'] \\
&= \underset{\sim}{H}\ E[(\underset{\sim}{p}-\underset{\sim}{P})\ (\underset{\sim}{p}-\underset{\sim}{P})']\ \underset{\sim}{H}' = \underset{\sim}{H}\ \underset{\sim}{\text{Var}}(\underset{\sim}{p})\ \underset{\sim}{H}',
\end{aligned}$$

which is estimated by

$$\underset{\sim}{H}\ \hat{\underset{\sim}{\text{Var}}}(\underset{\sim}{p})\ \underset{\sim}{H}' = \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}' = \underset{\sim}{S}.$$

We will not consider the matrix $\underset{\sim}{S}$ in detail at this time because it can be produced readily by a general computing algorithm and its specific form need not be known by the analyst. We will examine $\underset{\sim}{S}$ later when relationships among different forms of $\underset{\sim}{F}(\underset{\sim}{P})$ are of interest.

The method of analysis is now straightforward weighted least squares regression and can be accomplished by one general algorithm, provided $\underset{\sim}{S}$ is calculated properly. For the present we consider the general regression model

$$\underset{(t\times 1)}{\underset{\sim}{F}(\underset{\sim}{P})} = \underset{(t\times v)}{\underset{\sim}{X}} \; \underset{(v\times 1)}{\underset{\sim}{\xi}} , \qquad v \leq t \leq s(r-1),$$

where the $f_m(\underset{\sim}{p})$ assume the role of dependent variables. We need estimates of the elements of $\underset{\sim}{\xi}$ as well as a test of the fit of the model and tests for hypotheses expressed as $\underset{\sim}{C}\,\underset{\sim}{\xi} = \underset{\sim}{0}$.

Using conventional weighted least squares methods, we can obtain a BAN estimate of $\underset{\sim}{\xi}$ by minimizing the quadratic form

$$\underset{(1\times t)}{(\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{X}\,\underset{\sim}{\xi})'} \quad \underset{(t\times t)}{\underset{\sim}{S}^{-1}} \quad \underset{(t\times 1)}{(\underset{\sim}{F}(\underset{\sim}{p}) - \underset{\sim}{X}\,\underset{\sim}{\xi})} \tag{1.6.7}$$

with respect to $\underset{\sim}{\xi}$. This yields the estimate

$$\hat{\underset{\sim}{\xi}} = \underset{\sim}{b} = (\underset{\sim}{X}'\,\underset{\sim}{S}^{-1}\,\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'\,\underset{\sim}{S}^{-1}\,\underset{\sim}{F}(\underset{\sim}{p}) \tag{1.6.8}$$

which has the estimated variance matrix

$$\hat{\underset{\sim}{\mathrm{Var}}}(\underset{\sim}{b}) = [\underset{\sim}{X}'\,\underset{\sim}{S}^{-1}\,\underset{\sim}{X}]^{-1}.$$

We obtain the minimum value of the quadratic form in (1.6.7) by substituting $\underset{\sim}{b}$ from (1.6.8) for $\underset{\sim}{\xi}$, which gives the usual sum of squares for the residual,

$$[\underset{\sim}{F}(\underset{\sim}{p})]' \underset{\sim}{S}^{-1} [\underset{\sim}{F}(\underset{\sim}{p})] - \underset{\sim}{b}' (\underset{\sim}{X}' \underset{\sim}{S}^{-1} \underset{\sim}{X}) \underset{\sim}{b}.$$

If the model fits, this sum of squares has a limiting chi-square distribution with t-v degrees of freedom.

Given that the model fits, hypotheses concerning the elements of $\underset{\sim}{\xi}$ are of interest. We may express these hypotheses as

$$\underset{(c\times v)}{\underset{\sim}{C}} \quad \underset{(v\times 1)}{\underset{\sim}{\xi}} = \underset{\sim}{0}$$

where $\underset{\sim}{C}$ is a matrix of properly-selected constants. We can test such hypotheses by using the quadratic form

$$\underset{\sim}{b}' \underset{\sim}{C}' [\underset{\sim}{C} (\underset{\sim}{X}' \underset{\sim}{S}^{-1} \underset{\sim}{X})^{-1} \underset{\sim}{C}'] \underset{\sim}{C} \underset{\sim}{b},$$

which, if the null hypothesis is true, has a limiting χ^2 distribution with c degrees of freedom.

The above discussion includes the hypothesis $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{0}$ as a special case with $\underset{\sim}{X} = \underset{\sim}{I}$ and $\underset{\sim}{\xi} = \underset{\sim}{0}$. The only test needed for this situation is the test of the fit of the model provided by the sum of squares

$$[\underset{\sim}{F}(\underset{\sim}{p})] \underset{\sim}{S}^{-1} [\underset{\sim}{F}(\underset{\sim}{p})],$$

which has a limiting chi-square distribution with t degrees of freedom if the model fits.

We note that the calculations for tests of hypotheses and for estimates of $\underset{\sim}{\xi}$ can be performed without the explicit determination of restricted estimates of $\underset{\sim}{P}$. However, we can use the restricted estimates as estimates of the cell probabilities if the hypothesis is true. Many authors, including Mosteller [1968],have discussed restricted estimates. Forthofer, Starmer, and Grizzle [1969] published the expression

$$\hat{\underset{\sim}{p}} = \underset{\sim}{p} - \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{F}(\underset{\sim}{p}) \ . \tag{1.6.9}$$

We now consider the variance of $\hat{\underset{\sim}{p}}$. Clearly

$$\underset{\sim}{\mathrm{Var}}(\hat{\underset{\sim}{p}}) = \underset{\sim}{\mathrm{Var}}(\underset{\sim}{p}) + \underset{\sim}{\mathrm{Var}}[\underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{F}(\underset{\sim}{p})]$$

$$- 2\ \underset{\sim}{\mathrm{Cov}}[\underset{\sim}{p},\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{F}(\underset{\sim}{p})].$$

If we restrict ourselves to the asymptotic variance of $\hat{\underset{\sim}{p}}$ so that we can consider $\underset{\sim}{V}(\underset{\sim}{p})$ and $\underset{\sim}{H}$ as constant, we get the asymptotic relationship

$$\underset{\sim}{\mathrm{Var}}(\hat{\underset{\sim}{p}}) = \underset{\sim}{\mathrm{Var}}(\underset{\sim}{p}) + \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{\mathrm{Var}}[\underset{\sim}{F}(\underset{\sim}{p})]\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})$$

$$- 2\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{\mathrm{Cov}}[\underset{\sim}{p},\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{F}(\underset{\sim}{p})].$$

Substituting for $\underset{\sim}{F}(\underset{\sim}{P})$ from (1.6.6)

$$\underset{\sim}{\mathrm{Cov}}[\underset{\sim}{p},\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{F}(\underset{\sim}{p})] \approx \underset{\sim}{\mathrm{Cov}}\{\underset{\sim}{p},\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ [\underset{\sim}{F}(\underset{\sim}{p}) + \underset{\sim}{H}(\underset{\sim}{p}-\underset{\sim}{P})]\}$$

$$= \underset{\sim}{\mathrm{Cov}}[\underset{\sim}{p},\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}(\underset{\sim}{p}-\underset{\sim}{P})]$$

$$= \underset{\sim}{\mathrm{Cov}}[\underset{\sim}{p},\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{p}]$$

$$= \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{\mathrm{Cov}}[\underset{\sim}{p},\underset{\sim}{p}]$$

$$= \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{\mathrm{Var}}(\underset{\sim}{p}).$$

Since $\underset{\sim}{\mathrm{Var}}[\underset{\sim}{F}(\underset{\sim}{p})] = \underset{\sim}{H}\ \underset{\sim}{\mathrm{Var}}(\underset{\sim}{p})\ \underset{\sim}{H}'$,

$$\underset{\sim}{\mathrm{Var}}(\hat{\underset{\sim}{p}}) \approx \underset{\sim}{\mathrm{Var}}(\underset{\sim}{p}) + \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{\mathrm{Var}}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p}) \tag{1.6.10}$$

$$- 2\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{\mathrm{Var}}(\underset{\sim}{p}).$$

Recalling that $\hat{\underset{\sim}{\text{Var}}}(\underset{\sim}{p}) = \underset{\sim}{V}(\underset{\sim}{p})$ and substituting for $\hat{\underset{\sim}{\text{Var}}}(\underset{\sim}{p})$ in (1.6.10), we get an estimate of the large-sample variance of $\hat{\underset{\sim}{p}}$,

$$\hat{\underset{\sim}{\text{Var}}}(\hat{\underset{\sim}{p}}) = \underset{\sim}{V}(\underset{\sim}{p}) + \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})$$

$$- 2\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p}).$$

Since $\underset{\sim}{S} = \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'$, the above becomes

$$\hat{\underset{\sim}{\text{Var}}}(\hat{\underset{\sim}{p}}) = \underset{\sim}{V}(\underset{\sim}{p}) + \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})$$

$$- 2\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p}),$$

or

$$\hat{\underset{\sim}{\text{Var}}}(\hat{\underset{\sim}{p}}) = \underset{\sim}{V}(\underset{\sim}{p}) - \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p}). \tag{1.6.11}$$

We note that $\underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'\ \underset{\sim}{S}^{-1}\ \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})$ is positive semidefinite and thus the estimates of variance of the elements of $\hat{\underset{\sim}{p}}$ are less than or equal to estimates of variance of the observed cell probabilities.

1.7. Summary

This chapter should provide the background for a methodology allowing the analysis of a wide range of contingency tables. In the following chapters we shall attempt to broaden the range of applicability and also to indicate relationships between this methodology and other techniques.

CHAPTER II

RELATIONSHIPS BETWEEN CERTAIN LINEAR CONTRASTS OF $\ell n\ P_{ijk}$ AND TESTS OF MARGINAL INDEPENDENCE

2.1. Introduction

In this chapter we discuss relationships between restrictions on the expected cell probability parameters, the P_{ijk}, and restrictions on the elements of the vector $\underset{\sim}{\xi}$ in the model $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim\sim}{X\xi}$ for both the linear and logarithmic forms of $\underset{\sim}{F}(\underset{\sim}{P})$.

Primarily, we investigate relationships between certain linear contrasts of the $\ell n\ P_{ijk}$ and tests of marginal independence. We consider these relationships both for complete tables and for the so-called incomplete tables. These relationships provide a means of testing for marginal independence by properly selecting $\underset{\sim}{A}$ and $\underset{\sim}{K}$ and testing $\underset{\sim}{K}\ \ell n\ \underset{\sim\sim}{AP} = \underset{\sim}{0}$. We also present a measure of association suitable for use in a linear model.

2.2. Models and Reparameterization

By definition, probability parameters and their estimates for a multinomial distribution should sum to unity. We now consider some effects of this restriction on the relationships among the elements of $\underset{\sim}{\xi}$ in the model $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim\sim}{X\xi}$ both for $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim\sim}{AP}$ and for $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{K}\ \ell n\ \underset{\sim\sim}{AP}$.

For convenience, we base our discussion on an r×s×t table with expected cell probabilities P_{ijk}, i = 1, 2, ..., r, j = 1, 2, ..., s,

$k = 1, 2, \ldots, t$, with no $P_{ijk} = 0$ because of *a priori* restrictions. This table represents a single multinomial distribution with rst response categories so that $\Sigma_{i,j,k} P_{ijk} = 1$. The dimensions r, s, and t correspond to three factors, R, S, and T with r, s, and t levels respectively.

Some authors have considered models directly in terms of the expected cell probabilities, such as

$$P_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k, \tag{2.2.1}$$

where

μ = overall mean,

α_i = effect associated with i-th level of R,

β_j = effect associated with j-th level of S,

γ_k = effect associated with k-th level of T.

If we sum (2.2.1) over i, j, and k we obtain

$$\sum_{i,j,k} P_{ijk} = (rst)\,\mu + st \sum_i \alpha_i + rt \sum_j \beta_j + rs \sum_k \gamma_k = 1, \tag{2.2.2}$$

so that μ is a function of the other parameters. The design matrix for (2.2.1) is singular, and if we reparameterize by incorporating the usual restrictions

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0, \tag{2.2.3}$$

then it follows that $\mu = 1/(rst)$. Under these conditions the value of μ is known and should not be estimated. We note that if we assume μ to be zero and use the reparameterization (2.2.3) together with (2.2.2), the inconsistency $\Sigma P_{ijk} = 0$ results.

We see that model (2.2.1) must be constructed so that (2.2.2) holds. In addition, since restricted estimates are commonly obtained under the restriction $\Sigma\ \hat{p}_{ijk} = 1$ but not under the additional restriction $0 \leq \hat{p}_{ijk} \leq 1$, models such as (2.2.1) can lead to estimates outside the range $0 \leq \hat{p}_{ijk} \leq 1$.

In contrast to (2.2.1), we can consider

$$\ell n\ P_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k, \qquad (2.2.4)$$

or equivalently

$$P_{ijk} = e^{\mu}\ e^{\alpha_i}\ e^{\beta_j}\ e^{\gamma_k},$$

where α_i, β_j, and γ_k are as described for (2.2.1). For this model

$$\sum_{i,j,k} P_{ijk} = e^{\mu} \sum_i e^{\alpha_i} \sum_j e^{\beta_j} \sum_k e^{\gamma_k} = 1,$$

so that again μ is a function of the other parameters, as Darroch [1962] and others have indicated for this model. Further investigation of (2.2.4) shows that the reparameterization (2.2.3) does not conflict with $\Sigma\ P_{ijk} = 1$ so long as a scaling constant such as μ is present. If μ is deleted, it is necessary that

$$\sum_{i,j,k} P_{ijk} = \sum_i e^{\alpha_i} \sum_j e^{\beta_j} \sum_k e^{\gamma_k} = 1.$$

One reparameterization applicable in this situation is

$$\sum_i e^{\alpha_i} = \sum_j e^{\beta_j} = \sum_k e^{\gamma_k} = 1\ .$$

However, since these restrictions are nonlinear, implementing them in a general procedure is bothersome. We also note that, regardless of whether μ is included in (2.2.4), the properties of the natural logarithm are such that the restriction $\Sigma \hat{p}_{ijk} = 1$ implies the individual $\hat{p}_{ijk}$ are in the range $0 \leq \hat{p}_{ijk} \leq 1$.

2.3. Tests of Marginal Independence and Measures of Association

2.3.1. General definition of independence

Bhapkar [1961, 1966], Bhapkar and Koch [1968a, 1968b] and Goodman [1968] have given definitions of marginal independence for complete contingency tables. We express these definitions so as to be applicable for incomplete tables. For convenience, we give the definitions for a three-way table, but they can be generalized for tables of other dimensions. We consider three classes of independence and restrict ourselves to tables described by a single multinomial distribution of no more than rst categories with cell probabilities P_{ijk}.

Complete marginal independence is defined by

$$H_1: \quad P_{ijk} = a_i \, b_j \, c_k \tag{2.3.1.1}$$

for all combinations of i, j, k for which $P_{ijk} \neq 0$. For tables with no $P_{ijk} = 0$, we have

$$P_{i\cdot\cdot} = a_i \sum_{j=1}^{s} b_j \sum_{k=1}^{t} c_k,$$
$$P_{\cdot j\cdot} = b_j \sum_{i=1}^{r} a_i \sum_{k=1}^{t} c_k,$$
$$P_{\cdot\cdot k} = c_k \sum_{i=1}^{r} a_i \sum_{j=1}^{s} b_j.$$

Thus if the a_i, b_j, c_k are scaled so that $\Sigma a_i = \Sigma b_j = \Sigma c_k = 1$, the above definition implies the more common expression

$$P_{ijk} = P_{i\cdot\cdot} \, P_{\cdot j\cdot} \, P_{\cdot\cdot k} \,.$$

Three hypotheses define whether the joint contribution of two factors is independent of the third:

$$H_{21}:\ P_{ijk} = a_{ij}\ b_k,$$

$$H_{22}:\ P_{ijk} = a_{ik}\ b_j, \qquad (2.3.1.2)$$

$$H_{23}:\ P_{ijk} = a_{jk}\ b_i,$$

for combinations of i, j, k for which $P_{ijk} \neq 0$. For tables with no $P_{ijk} = 0$, we have

$$P_{ij\cdot} = a_{ij} \sum_{k=1}^{t} b_k,$$

$$P_{\cdot\cdot k} = b_k \sum_{i=1}^{r} \sum_{j=1}^{t} a_{ij},$$

so that if b_k and a_{ij} are scaled so that $\sum_{k=1}^{t} b_k = \sum_{i=1}^{r} \sum_{j=1}^{s} a_{ij} = 1$

then H_{21} implies

$$P_{ijk} = P_{ij\cdot}\ P_{\cdot\cdot k},$$

which is a more common expression. Under these conditions H_{22} and H_{23} imply, respectively,

$$P_{ijk} = P_{i\cdot k}\ P_{\cdot j\cdot}\ ,$$

$$P_{ijk} = P_{\cdot jk}\ P_{i\cdot\cdot}\ .$$

Similarly, three hypotheses define whether each pair of factors is independent within the third:

$$H_{31}:\quad P_{ijk} = a_{ik}\, b_{jk},$$

$$H_{32}:\quad P_{ijk} = a_{ij}\, b_{ik}, \qquad (2.3.1.3)$$

$$H_{33}:\quad P_{ijk} = a_{jk}\, b_{ij},$$

for combinations of i, j, k for which $P_{ijk} \neq 0$. For tables with no $P_{ijk} = 0$, we have, for H_{31},

$$P_{i\cdot k} = a_{ik} \sum_{j=1}^{s} b_{jk},$$

$$P_{\cdot jk} = b_{jk} \sum_{i=1}^{r} a_{ik},$$

$$P_{\cdot\cdot k} = \sum_{i=1}^{r} a_{ik} \sum_{j=1}^{s} b_{jk}.$$

Thus $P_{i\cdot k}\, P_{\cdot jk} /\, P_{\cdot\cdot k} = a_{ik}\, b_{ik}$ and H_{31} implies

$$P_{ijk} = P_{i\cdot k}\, P_{\cdot jk}/P_{\cdot\cdot k},$$

and H_{32} and H_{33} imply, respectively

$$P_{ijk} = P_{ij\cdot}\, P_{i\cdot k}/P_{i\cdot\cdot} \quad ,$$

$$P_{ijk} = P_{ij\cdot}\, P_{\cdot jk}/P_{\cdot j\cdot} \quad .$$

2.3.2. <u>Relationships between definitions of marginal independence and linear contrasts in $\ell n\ P_{ijk}$</u>

The hypotheses expressed in (2.3.1.1), (2.3.1.2), and (2.3.1.3) also can be expressed in terms of linear functions of $\ell n\ P_{ijk}$. These functions are a form of linear contrast and can be generated routinely because of their similarity to contrasts commonly used in the analysis of factorial experiments.

Table 2.1 shows the linear contrasts for a 3×3×3 factorial experiment. Table 2.2 lists the expected cell probabilities under the hypothesis H_{23}. We now derive the relationship between the contrasts in Table 2.1, expressed in the natural logarithm of P_{ijk}, and the hypothesis H_{23} of marginal independence.

We consider the hypothesis H_{23} from (2.3.1.2) and show that testing this hypothesis is equivalent to testing whether the set of linear contrasts in $\ell n\ P_{ijk}$ for RS, RT, and RST interaction is zero. Considering first Table 2.2, we show that relationships among the P_{ijk} under the assumption $P_{ijk} = a_{jk}\ b_i$ imply that the linear contrasts in the $\ell n\ P_{ijk}$ related to the RS, RT, and RST interaction terms in Table 2.1 are zero. We then show that, if the RS, RT, and RST contrasts are zero, $P_{ijk} = a_{jk}\ b_i$ for all i, j, and k.

Consider the first of the four contrasts for RS interaction in Table 2.1. From Table 2.2,

$$\frac{P_{111}\ P_{112}\ P_{113}}{P_{131}\ P_{132}\ P_{133}} = \frac{a_{11}\ a_{12}\ a_{13}}{a_{31}\ a_{32}\ a_{33}}$$

and

$$\frac{P_{311}\ P_{312}\ P_{313}}{P_{331}\ P_{332}\ P_{333}} = \frac{a_{11}\ a_{12}\ a_{13}}{a_{31}\ a_{32}\ a_{33}} \quad ,$$

so that

$$\frac{P_{111}\ P_{112}\ P_{113}\ P_{331}\ P_{332}\ P_{333}}{P_{131}\ P_{132}\ P_{133}\ P_{311}\ P_{312}\ P_{313}} = 1 \ ,$$

or

TABLE 2.1

CONTRASTS FOR A 3×3×3 FACTORIAL EXPERIMENT

Effect	R_1									R_2									R_3								
	S_1			S_2			S_3			S_1			S_2			S_3			S_1			S_2			S_3		
	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3
	P_{111}	P_{112}	P_{113}	P_{121}	P_{122}	P_{123}	P_{131}	P_{132}	P_{133}	P_{211}	P_{212}	P_{213}	P_{221}	P_{222}	P_{223}	P_{231}	P_{232}	P_{233}	P_{311}	P_{312}	P_{313}	P_{321}	P_{322}	P_{323}	P_{331}	P_{332}	P_{333}
R	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	−	−	−	−	−	−	−	−	−
	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−
S	+	+	+	0	0	0	−	−	−	+	+	+	0	0	0	−	−	−	+	+	+	0	0	0	−	−	−
	0	0	0	+	+	+	−	−	−	0	0	0	+	+	+	−	−	−	0	0	0	+	+	+	−	−	−
T	+	0	−	+	0	−	+	0	−	+	0	−	+	0	−	+	0	−	+	0	−	+	0	−	+	0	−
	0	+	−	0	+	−	0	+	−	0	+	−	0	+	−	0	+	−	0	+	−	0	+	−	0	+	−
RS	+	+	+	0	0	0	−	−	−	0	0	0	0	0	0	0	0	0	−	−	−	0	0	0	+	+	+
	0	0	0	+	+	+	−	−	−	0	0	0	0	0	0	0	0	0	0	0	0	−	−	−	+	+	+
	0	0	0	0	0	0	0	0	0	+	+	+	0	0	0	−	−	−	−	−	−	0	0	0	+	+	+
	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	−	−	−	0	0	0	−	−	−	+	+	+
RT	+	0	−	+	0	−	+	0	−	0	0	0	0	0	0	0	0	0	−	0	+	−	0	+	−	0	+
	0	+	−	0	+	−	0	+	−	0	0	0	0	0	0	0	0	0	0	−	+	0	−	+	0	−	+
	0	0	0	0	0	0	0	0	0	+	0	−	+	0	−	+	0	−	−	0	+	−	0	+	−	0	+
	0	0	0	0	0	0	0	0	0	0	+	−	0	+	−	0	+	−	0	−	+	0	−	+	0	−	+
ST	+	0	−	0	0	0	−	0	+	+	0	−	0	0	0	−	0	+	+	0	−	0	0	0	−	0	+
	0	+	−	0	0	0	0	−	+	0	+	−	0	0	0	0	−	+	0	+	−	0	0	0	0	−	+
	0	0	0	+	0	−	−	0	+	0	0	0	+	0	−	−	0	+	0	0	0	+	0	−	−	0	+
	0	0	0	0	+	−	0	−	+	0	0	0	0	+	−	0	−	+	0	0	0	0	+	−	0	−	+
RST	+	0	−	0	0	0	−	0	+	0	0	0	0	0	0	0	0	0	−	0	+	0	0	0	+	0	−
	0	+	−	0	0	0	0	−	+	0	0	0	0	0	0	0	0	0	0	−	+	0	0	0	0	+	−
	0	0	0	+	0	−	−	0	+	0	0	0	0	0	0	0	0	0	0	0	0	−	0	+	+	0	−
	0	0	0	0	+	−	0	−	+	0	0	0	0	0	0	0	0	0	0	0	0	0	−	+	0	+	−
	0	0	0	0	0	0	0	0	0	+	0	−	0	0	0	−	0	+	−	0	+	0	0	0	+	0	−
	0	0	0	0	0	0	0	0	0	0	+	−	0	0	0	0	−	+	0	−	+	0	0	0	0	+	−
	0	0	0	0	0	0	0	0	0	0	0	0	+	0	−	−	0	+	0	0	0	−	0	+	+	0	−
	0	0	0	0	0	0	0	0	0	0	0	0	0	+	−	0	−	+	0	0	0	0	−	+	0	+	−

TABLE 2.2

EXPECTED CELL PROBABILITIES FOR THE HYPOTHESIS H_{23} CLASSIFIED BY FACTOR

		R_1	R_2	R_3
	T_1	$P_{111} = a_{11}\ b_1$	$P_{211} = a_{11}\ b_2$	$P_{311} = a_{11}\ b_3$
S_1	T_2	$P_{112} = a_{12}\ b_1$	$P_{212} = a_{12}\ b_2$	$P_{312} = a_{12}\ b_3$
	T_3	$P_{113} = a_{13}\ b_1$	$P_{213} = a_{13}\ b_2$	$P_{313} = a_{13}\ b_3$
	T_1	$P_{121} = a_{21}\ b_1$	$P_{221} = a_{21}\ b_2$	$P_{321} = a_{21}\ b_3$
S_2	T_2	$P_{122} = a_{22}\ b_1$	$P_{222} = a_{22}\ b_2$	$P_{322} = a_{22}\ b_3$
	T_3	$P_{123} = a_{23}\ b_1$	$P_{223} = a_{23}\ b_2$	$P_{323} = a_{23}\ b_3$
	T_1	$P_{131} = a_{31}\ b_1$	$P_{231} = a_{31}\ b_2$	$P_{331} = a_{31}\ b_3$
S_3	T_2	$P_{132} = a_{32}\ b_1$	$P_{232} = a_{32}\ b_2$	$P_{332} = a_{32}\ b_3$
	T_3	$P_{133} = a_{33}\ b_1$	$P_{233} = a_{33}\ b_2$	$P_{333} = a_{33}\ b_3$

$$\ell n\ P_{111} + \ell n\ P_{112} + \ell n\ P_{113} - \ell n\ P_{131} - \ell n\ P_{132} - \ell n\ P_{133} - \ell n\ P_{311}$$

$$- \ell n\ P_{312} - \ell n\ P_{313} + \ell n\ P_{331} + \ell n\ P_{332} + \ell n\ P_{333} = 0,$$

which is the desired contrast. Similarly, we obtain the other three contrasts for RS interaction. For the first contrast for RT interaction, we get, from Table 2.2,

$$\frac{P_{111}\ P_{121}\ P_{131}}{P_{113}\ P_{123}\ P_{133}} = \frac{a_{11}\ a_{21}\ a_{31}}{a_{13}\ a_{23}\ a_{33}}$$

and

$$\frac{P_{311}\ P_{321}\ P_{331}}{P_{313}\ P_{323}\ P_{333}} = \frac{a_{11}\ a_{21}\ a_{31}}{a_{13}\ a_{23}\ a_{33}} \quad ,$$

so that

$$\frac{P_{111}\, P_{121}\, P_{131}\, P_{313}\, P_{323}\, P_{333}}{P_{113}\, P_{123}\, P_{133}\, P_{311}\, P_{321}\, P_{331}} = 1,$$

or

$$\ell n\, P_{111} - \ell n\, P_{113} + \ell n\, P_{121} - \ell n\, P_{123} + \ell n\, P_{131} - \ell n\, P_{133} - \ell n\, P_{311}$$
$$+ \ell n\, P_{313} - \ell n\, P_{321} + \ell n\, P_{323} - \ell n\, P_{331} + \ell n\, P_{333} = 0,$$

as desired. Again, we obtain similarly the remaining three contrasts for RT interaction.

For the first of the eight contrasts for RST interaction, we see from Table 2.2 that

$$\frac{P_{111}\, P_{331}}{P_{113}\, P_{333}} = \frac{a_{11}\, a_{31}}{a_{13}\, a_{33}}$$

and

$$\frac{P_{131}\, P_{311}}{P_{133}\, P_{313}} = \frac{a_{11}\, a_{31}}{a_{13}\, a_{33}} \quad ,$$

so that

$$\frac{P_{111}\, P_{331}\, P_{133}\, P_{313}}{P_{113}\, P_{333}\, P_{131}\, P_{311}} = 1 \ ,$$

or

$$\ell n\, P_{111} - \ell n\, P_{113} - \ell n\, P_{131} + \ell n\, P_{133} - \ell n\, P_{311} + \ell n\, P_{313}$$
$$+ \ell n\, P_{331} - \ell n\, P_{333} = 0,$$

which correspond to the first of the eight contrasts for RST interaction given in Table 2.1. We can obtain the seven remaining contrasts in a similar fashion.

Given that the RS, RT, and RST contrasts are zero, then, as in the standard factorial experiment, the following model holds:

$$\ell n\ P_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk}$$

or

$$P_{ijk} = e^{\mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk}}$$

where μ is the overall mean, α_i, β_j, γ_k are the effects associated with the i-th, j-th, and k-th levels of R, S, and T, respectively, and $(\beta\gamma)_{jk}$ is the interaction effect associated with the jk-th combinations of S and T. If

$$a_{jk} = e^{\mu + \beta_j + \gamma_k + (\beta\gamma)_{jk}},$$

$$b_i = e^{\alpha_i},$$

then $P_{ijk} = a_{jk}\, b_i$ for all i, j, k and hypothesis H_{23} is satisfied.

Other relationships between hypotheses expressed in terms of marginal probabilities and those expressed in terms of linear contrasts of the $\ell n\ P_{ijk}$ appear in Table 2.3. The relationship between complete marginal independence, hypothesis H_1, and the model

$$P_{ijk} = e^{\mu + \alpha_i + \beta_j + \gamma_k}$$

has been discussed by Goodman [1968] and others.

2.3.3. Relationships between definitions of marginal independence and linear contrasts in $\ell n\ P_{ijk}$ for incomplete tables

We present relationships of the type discussed in Section 2.3.2 for an incomplete table. Consider the 3×3 table with expected cell probabilities

$$
\begin{array}{c|ccc|c}
 & W_1 & W_2 & W_3 & \\
\hline
U_1 & P_{11} & P_{12} & P_{13} & \\
U_2 & P_{21} & P_{22} & P_{23} & \\
U_3 & P_{31} & 0 & P_{33} & \\
\hline
 & & & & 1
\end{array}
\qquad (2.3.3.1)
$$

where $P_{32} = 0$ results from a priori restrictions. The definition of independence equivalent to H_1, (2.3.1.1), for this table is

$$P_{ij} = a_i \, b_j \qquad (2.3.3.2)$$

for all combinations of i, j for which $P_{ij} \neq 0$. Hence independence of V and W implies

$$
\begin{aligned}
P_{11} &= a_1 \, b_1, \\
P_{12} &= a_1 \, b_2, \\
P_{13} &= a_1 \, b_3, \\
P_{21} &= a_2 \, b_1, \\
P_{22} &= a_2 \, b_2, \\
P_{23} &= a_2 \, b_3, \\
P_{31} &= a_3 \, b_1, \\
P_{33} &= a_3 \, b_3.
\end{aligned}
$$

Examining relationships among the P_{ij}, we see that

$$\frac{P_{11}}{P_{13}} = \frac{P_{21}}{P_{23}} = \frac{b_1}{b_3} \text{ or } \frac{P_{11} \, P_{23}}{P_{13} \, P_{21}} = 1,$$

TABLE 2.3

RELATIONSHIPS BETWEEN LINEAR CONTRASTS OF $\ell n\ P_{ijk}$ AND TESTS FOR MARGINAL INDEPENDENCE, FOR A 3×3×3 CONTINGENCY TABLE WITH ONLY THE TOTAL FIXED

Effect	Number of contrasts	Hypotheses P_{ijk} =						
		$a_i\ b_j\ c_k$	$a_{ij}\ b_{ik}$	$a_{ik}\ b_{jk}$	$a_{jk}\ b_{ij}$	$a_{ij}\ b_k$	$a_{ik}\ b_j$	$a_{jk}\ b_i$
R	2	N	N	N	N	N	N	N
S	2	N	N	N	N	N	N	N
T	2	N	N	N	N	N	N	N
RS	4	Y	N	Y	N	N	Y	Y
RT	4	Y	N	N	Y	Y	N	Y
ST	4	Y	Y	N	N	Y	Y	N
RST	8	Y	Y	Y	Y	Y	Y	Y

N indicates that the hypothesis specified by the column heading does not require the associated linear contrasts in $\ell n\ P_{ijk}$ to be zero,

Y indicates that the hypothesis specified by the column heading requires the associated linear contrasts in $\ell n\ P_{ijk}$ to be simultaneously zero.

$$\frac{P_{12}}{P_{13}} = \frac{P_{22}}{P_{23}} = \frac{b_2}{b_3} \text{ or } \frac{P_{12}\,P_{23}}{P_{13}\,P_{22}} = 1,$$

$$\frac{P_{21}}{P_{23}} = \frac{P_{31}}{P_{33}} = \frac{b_1}{b_3} \text{ or } \frac{P_{21}\,P_{33}}{P_{23}\,P_{31}} = 1 .$$

Taking natural logarithms of the above equations, we obtain the constraints

$$\begin{aligned} \ell n\, P_{11} - \ell n\, P_{13} - \ell n\, P_{21} + \ell n\, P_{23} &= 0, \\ \ell n\, P_{12} - \ell n\, P_{13} - \ell n\, P_{22} + \ell n\, P_{23} &= 0, \\ \ell n\, P_{21} - \ell n\, P_{23} - \ell n\, P_{31} + \ell n\, P_{33} &= 0. \end{aligned} \tag{2.3.3.3}$$

If (2.3.3.1) represents a 3×3 factorial experiment with no data for one cell, the estimable contrasts for V×W interaction follow.

W_1			W_2			W_3		
V_1	V_2	V_3	V_1	V_2	V_3	V_1	V_2	V_3
P_{11}	P_{12}	P_{13}	P_{21}	P_{22}	P_{23}	P_{31}	0	P_{33}
+	0	-	-	0	+	0		0
0	+	-	0	-	+	0		0
0	0	0	+	0	-	-		+

The contrasts shown above correspond to the constraints on the $\ell n\, P_{ij}$ given in (2.3.3.3). Thus the hypothesis of marginal independence of V and W requires the linear contrasts of the $\ell n\, P_{ij}$ corresponding to the estimable contrasts for V×W interaction to be zero. If these contrasts are zero, we can express $\ell n\, P_{ij}$ in the model

$$\ln P_{ij} = \mu + \alpha_i + \beta_j, \quad i,j = 1,2,3,(i,j) \neq (3,2),$$

or

$$P_{ij} = e^{\mu + \alpha_i + \beta_j},$$

where
$$\mu = \text{overall mean},$$
$$\alpha_i = \text{effect of i-th level of V},$$
$$\beta_j = \text{effect of j-th level of W}.$$

If we let
$$a_i = e^{\mu + \alpha_i},$$
$$b_j = e^{\beta_j},$$

with the same conditions on i and j as above, then $P_{ij} = a_i b_j$ for all i, j for which $P_{ij} \neq 0$. Thus, if the contrasts in $\ln P_{ij}$ corresponding to the estimable contrasts for V×W interaction are zero, the conditions for complete marginal independence as shown in (2.3.3.2) are satisfied. We can extend this observation easily to higher-dimension tables with any configuration of cells having $P_{ij} = 0$, provided that the appropriate interaction contrasts are estimable.

2.4. A Measure of Association Useful as a Dependent Variable in a Linear Model

In this chapter we have provided the basis for a measure of association useful as a dependent variable in a linear model. To indicate the measure's usefulness, we consider the table shown by (2.3.3.1) and suppose one such table is a cell of a larger r×s×t table. The dimensions of the r×s×t table may correspond to any combination of blocks and factors. However, for this discussion we assume that they correspond to factors R, S, and T with levels r, s, and t respectively.

We are interested in the association between the variables V and W and how this association varies among the combinations of R, S, and T.

We have examined the incomplete table (2.3.3.1) and have showed that V and W are independent if the three contrasts given by (2.3.3.3) are zero. Consider the vector

$$\begin{bmatrix} u_{1ijk} \\ u_{2ijk} \\ u_{3ijk} \end{bmatrix} = \begin{bmatrix} \ln p_{11ijk} - \ln p_{13ijk} - \ln p_{21ijk} + \ln p_{23ijk} \\ \ln p_{12ijk} - \ln p_{13ijk} - \ln p_{22ijk} + \ln p_{23ijk} \\ \ln p_{21ijk} - \ln p_{23ijk} - \ln p_{31ijk} + \ln p_{33ijk} \end{bmatrix} ,$$

which we denote by $\underset{\sim}{u}_{ijk}$, $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, s$; $k = 1, 2, \ldots, t$. If $\underset{\sim}{u}_{ijk}$ has expected value zero, then V and W are independent within the ijk-th combination of R, S, and T. Thus, considering the vector $\underset{\sim}{u}_{ijk}$ as a measure of association, we can use the model

$$E \begin{bmatrix} u_{1ijk} \\ u_{2ijk} \\ u_{3ijk} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \alpha_{3i} \end{bmatrix} + \begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{bmatrix} + \begin{bmatrix} \gamma_{1k} \\ \gamma_{2k} \\ \gamma_{3k} \end{bmatrix} ,$$

where, for $h = 1, 2, 3$,

μ_h = overall mean for the h-th u,

α_{hi} = effect of the i-th level of R on the h-th u,

β_{hj} = effect of the j-th level of S on the h-th u,

γ_{hk} = effect of the k-th level of T on the h-th u.

We may extend this model easily and in its more general form the vector $\underset{\sim}{u}$ contains the appropriate interaction contrasts in $\ell n\ p_{ijk}$. By using the usual techniques of analysis of experiments, we may construct a linear model, with $\underset{\sim}{u}$ as a dependent variable, which reflects the structure of the overall table. If the subtable is incomplete, as above, then all subtables need to be incomplete identically so that we can construct similar vectors $\underset{\sim}{u}$ within each cell of the overall table.

2.5. An Example of a Multidimensional Complete Table

The data shown in Table 2.4 were collected in an international study of atherosclerosis (Strong, et al., 1968). The overall table is an example with two dependent variables (responses), infarct and myocardial scar, which we denote by i and j respectively, and two independent variables, age and the combination location and race, denoted by k and ℓ. Within each combination of the independent variables we have a 2×2 subtable with observed cell frequencies such as

		Infarct		
		No	Yes	
Myocardial	No	$n_{11k\ell}$	$n_{12k\ell}$	
scar	Yes	$n_{21k\ell}$	$n_{22k\ell}$	
				$n_{\cdot\cdot k\ell}$

(2.5.1)

Neither of the two marginal totals for this subtable is considered fixed and the corresponding expected cell probabilities are $P_{ijk\ell}$ where $\Sigma_{i,j}\ P_{ijk\ell} = 1$ for all combinations of k and ℓ. This subtable corresponds to a 2×2 factorial experiment and we may use the measure of association

$$u_{k\ell} = \ell n\ p_{11k\ell} - \ell n\ p_{12k\ell} - \ell n\ p_{21k\ell} + \ell n\ p_{22k\ell} \qquad (2.5.2)$$

as a dependent variable in a linear model. In Table 2.4 the value of u for each 2×2 subtable is given by the convenient expression

$$e^u = \frac{p_{11}\ p_{22}}{p_{12}\ p_{21}} = \frac{n_{11}\ n_{22}}{n_{12}\ n_{21}} ;$$

thus, for example, the subtable for (Age 55-64, New Orleans Negro) shows

$$e^u = \frac{(10)\quad (4)}{(8)\ (14)} = .357.$$

Fienberg and Gilbert [1970] note that for the 2×2 table

$$u^* = \frac{2}{\pi} \arctan (u)$$

is symmetric around zero, ranges between -1 and +1, and, if the table is arranged properly, is positive for positive association and negative for negative association. The values of u^* are also given in Table 2.4.

We calculate the estimated large-sample variance of $u_{k\ell}$ from $\underset{\sim}{S} = \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'$,

where

$$\underset{\sim}{H}'_{k\ell} = \left[\frac{1}{p_{11k\ell}}\ ,\ \frac{-1}{p_{12k\ell}}\ ,\ \frac{-1}{p_{21k\ell}}\ ,\ \frac{1}{p_{22k\ell}}\right],$$

and

$$\underset{\sim}{V}(\underset{\sim}{p}_{k\ell}) = \begin{bmatrix} p_{11k\ell}(1-p_{11k\ell}) & & & \\ -p_{11k\ell}\ p_{12k\ell} & p_{12k\ell}(1-p_{12k\ell}) & & \\ -p_{11k\ell}\ p_{21k\ell} & -p_{12k\ell}\ p_{21k\ell} & p_{21k\ell}(1-p_{21k\ell}) & \\ -p_{11k\ell}\ p_{22k\ell} & -p_{12k\ell}\ p_{22k\ell} & -p_{21k\ell}\ p_{22k\ell} & p_{22k\ell}(1-p_{22k\ell}) \end{bmatrix}.$$

Performing the indicated multiplications, we get

$$S_{k\ell} = \frac{1}{n_{\bullet\bullet k\ell}} \sum_{i,j} \frac{1}{p_{ijk\ell}} = \sum_{i,j} \frac{1}{n_{ijk\ell}} . \qquad (2.5.3)$$

TABLE 2.4

CASES OF CORONARY HEART DISEASE CLASSIFIED BY TYPE OF LESION, AGE, LOCATION AND RACE

	New Orleans White			Oslo			New Orleans Negro		
	Infarct (age 35-44)								
		No	Yes		No	Yes		No	Yes
Myocardial Scar (age 35-44)	No	9	8	No	7	3	No	4	7
	Yes	6	6	Yes	2	5	Yes	2	3
	$e^u = 1.125$			$e^u = 5.833$			$e^u = .857$		
	$u^* = .075$			$u^* = .672$			$u^* = -.098$		
	Infarct (age 45-54)								
		No	Yes		No	Yes		No	Yes
Myocardial Scar (age 45-54)	No	10	26	No	6	8	No	10	8
	Yes	16	14	Yes	7	11	Yes	14	4
	$e^u = .337$			$e^u = 1.179$			$e^u = .357$		
	$u^* = -.527$			$u^* = .104$			$e^* = -.509$		
	Infarct (age 55-64)								
		No	Yes		No	Yes		No	Yes
Myocardial Scar (age 55-64)	No	18	47	No	10	22	No	4	13
	Yes	28	21	Yes	39	39	Yes	14	2
	$e^u = .287$			$e^u = .455$			$e^u = .044$		
	$u^* = -.570$			$u^* = -.425$			$u^* = -.803$		
	Infarct (age 65-69)								
		No	Yes		No	Yes		No	Yes
Myocardial Scar (age 65-69)	No	3	13	No	5	16	No	0	4
	Yes	11	5	Yes	27	16	Yes	3	2
	$e^u = .105$			$e^u = .185$			$e^u = .000$		
	$u^* = -.734$			$u^* = -.659$			$u^* = -1.00$		

We can examine the effects of the independent variables on $u_{k\ell}$ by considering the model

$$E(u_{k\ell}) = \mu^* + \alpha_k^* + \beta_\ell^* ,$$

where μ^* is the overall mean effect, α_k^*, $k = 1,2,3,4$, is the effect of the k-th age group and β_ℓ^*, $\ell = 1,2,3$, is the effect of the ℓ-th location-race combination. We reparametrize to incorporate the restrictions $\Sigma\, \hat{\alpha}_k^* = 0$ and $\Sigma\, \hat{\beta}_\ell^* = 0$, which is equivalent to calculating the estimates from:

$$E \begin{bmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{31} \\ u_{32} \\ u_{33} \\ u_{41} \\ u_{42} \\ u_{43} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & -1 & -1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & -1 & -1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{bmatrix} ,$$

where the elements of $\underset{\sim}{u}$ are the appropriate $u_{k\ell}$ values taken from Table 2.4. The design matrix is similar to those used in linear models for continuous data. We substituted $1/4n_{\cdot\cdot 43}$ for the zero n_{1143} in order to avoid a singular $\underset{\sim}{S}$ matrix. The remainder of the analysis is identical to weighted multiple regression. We may interpret the residual from fitting the model as the age × location-race interaction with respect to the measure of association u. In addition, we can investigate how this measure of association depends on age and the location-race combination.

The estimated parameters appear in Table 2.5, and the analysis of variance appears in Table 2.6.

TABLE 2.5

ESTIMATED PARAMETERS AND THEIR STANDARD ERRORS FOR THE DATA IN TABLE 2.4

Parameter	Estimate	Standard error
μ	-1.036	.236
α_1	1.496	.443
α_2	.281	.343
α_3	- .413	.296
β_1	- .021	.270
β_2	.776	.293

TABLE 2.6

ANALYSIS OF VARIANCE FOR THE DATA IN TABLE 2.4

Source of Variation	DF		SS	
Age Groups	3		16.61	
Linear trend of age		1		16.28
Remainder		2		.33
Race and Location Combinations	2		7.06	
N.O. white vs N.O. Negro		1		1.59
N.O. white vs Oslo		1		3.55
(N.O. white + N.O. Negro) vs 2 Oslo		1		7.00
Residual	6		2.62	

We compare the residual SS to the tabular value of χ^2 with 6 D.F. and find that the model fits the data adequately. We produce the various sums of squares from the general hypothesis form $\underset{\sim}{C}\,\underset{\sim}{\xi} = \underset{\sim}{0}$, where $\underset{\sim}{C}\,\underset{\sim}{b}$ estimates $\underset{\sim}{C}^*\,\underset{\sim}{\xi}^*$. Thus

$$\underset{\sim}{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

yields the test of homogeneity of age groups. To test for approximate linearity of the age effects (note that the last age group covers only a five-year span), we chose the linear contrast $-3\alpha_1^* - \alpha_2^* + \alpha_3^* + 3\alpha_4^* = 0$. Taking into account the restrictions on the estimates, we find the contrast is estimated by $-6\hat{\alpha}_1 - 4\hat{\alpha}_2 - 2\hat{\alpha}_3$. For testing we might equally well choose $3\hat{\alpha}_1 + 2\hat{\alpha}_2 + \hat{\alpha}_3$, which implies that we could use $\underset{\sim}{C} = (0,3,2,1,0,0)$. We can produce other tests similarly.

We conclude from the analysis that there is no age by location-race interactions and that the measure of association varies linearly with age. The major difference in race and location combination is between Oslo residents and New Orleans residents as shown by the test statistic for the contrast (New Orleans white + New Orleans Negro − 2 Oslo).

2.6. Application to an Incomplete Table and Restricted Estimates

In an earlier discussion, we extended relationships between marginal independence and tests for interaction to include incomplete tables. As an example of an analysis of an incomplete table we have chosen Table 2.7, which Harris and Chi Tu [1929] first presented and which Goodman [1968] recently analyzed. This table is a 4×9 table where the P_{ij}'s associated with half of the cells are restricted to be zero because of a priori conditions.

TABLE 2.7

RELATIONSHIP BETWEEN RADIAL ASYMMETRY AND LOCULAR COMPOSITION IN STAPHYLEA (SERIES A)

Locular composition	Coefficient of radial symmetry .00	.47	.82	.94	1.25	1.41	1.63	1.70	1.89
3 even 0 odd	462 (461)*	-	-	130 (126)	-	-	2 (2)	-	1 (1)
2 even 1 odd	-	614 (612)	138 (135)	-	21 (25)	14 (13)	-	1 (3)	-
1 even 2 odd	-	443 (444)	95 (98)	-	22 (18)	8 (9)	-	5 (3)	-
0 even 3 odd	103 (107)	-	-	35 (31)	-	-	1 (1)	-	0 (0)

*Restricted estimates presented in latter part of Chapter I and further discussed on page 45.

Notice, also, that the lower right-hand cell contains a "true" zero.

We consider these data to represent a sample from a single multinomial distribution having 18 categories corresponding to 18 cells not restricted to be zero a priori, and we arrange the cell probabilities

$$\underset{\sim}{P}' = [P_{11}, P_{14}, P_{17}, P_{19}, P_{22}, P_{23}, P_{25}, P_{26}, P_{28},$$
$$P_{32}, P_{33}, P_{35}, P_{36}, P_{38}, P_{41}, P_{44}, P_{47}, P_{48}].$$

Recalling the discussion relating models and tests of independence, we define a test for interaction based on the vectors in the error space of the model

$$\ln P_{ij} = \mu^* + \alpha_i^* + \beta_j^*.$$

The test consists of determining whether the vectors spanning the error space have expected value zero. Reference to testing interaction in incomplete block designs (Elston and Bush [1964]) and inspection of Table 2.7 show that the estimable interaction contrasts among the $\ell n\ P_{ij}$, which we shall use for testing independence, can be formed by $\underset{\sim}{K}\ \underset{\sim}{\ell n}(\underset{\sim}{P})$, where

$$\underset{\sim}{K} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} .$$

Performing the tests yields $X^2 = 6.24$. Goodman obtained $X^2 = 6.9$ after pooling the last two columns. Hence his test statistic has 6 D.F. while ours has 7. If our definition of independence is accepted, no new methods are required for incomplete tables, and interpretation is compatible with that of complete tables.

We can use the data in Table 2.7 to illustrate another useful device, that of smoothing the data or using restricted estimates. The estimates we give here are for an incomplete table, but the technique is applicable to complete tables also.

If the two ways of classification shown in Table 2.7 are independent, then we can calculate estimates of the cell probabilities under the assumption that the independence restrictions are true. This procedure should smooth out some random irregularities in the data, and Mosteller [1968] has pointed out its desirability. We have shown that hypotheses of independence can be formulated by considering constraints of the form $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{0}$. The restricted estimate, $\hat{\underset{\sim}{p}}$, of $\underset{\sim}{P}$ can be obtained under the assumption that $\underset{\sim}{F}(\underset{\sim}{P}) = \underset{\sim}{0}$. We have given the formula of this estimate in (1.6.9) and we indicated that, at least asymptotically, the variance of $\hat{\underset{\sim}{p}}$ is smaller than the variance of the observed cell probabilities.

An important and seldom-exploited consequence of this result is that, for large samples, the cell frequencies calculated under the assumption of independence are better estimates than the observed cell probabilities when independence prevails. If, however, the preliminary test falsely indicated that the restriction is true, the restricted estimates will be biased.

The cell frequencies calculated under the restriction of independence appear in parentheses in Table 2.7.

CHAPTER III

APPLICATIONS TO RELATIVE RISK

3.1. Introduction

Relative risk, an epidemiological measure often used to describe relationships between physical or social states and disease conditions, has traditionally been applied to single 2×2 tables. Cornfield [1951] and others have described its use. Some authors have considered extensions to more than one 2×2 table. For example, Cornfield [1956] presented an iterative procedure for making multiple comparisons among several relative risks. However, the goal of most work for more than one table has been to obtain an average risk. Woolf [1955] considered the logarithm of relative risk and showed how this measure could be combined for several tables. He included a test for the heterogeneity of the separate measures and a test for non-zero combined relative risk. Other authors, for example Goodman [1963b], considered different techniques of combining risk. Gart [1962] discussed several estimators of the combined risk and indicated that the logarithm of relative risk is both consistent and efficient.

Cornfield [1967] used the logarithm of relative risk in a discriminant function, thus permitting assessment of the contribution of each of several continuous variables to the log risk. However, if all variables are categorical variables, and especially if they do not represent ordered responses, this procedure is undesirable. Grizzle, in unpublished work, used the logarithm of relative risk as a dependent variable in a general linear model and, without violating any assumptions, used categorical variables as independent variables. Grizzle's approach allows one to combine several relative risks when there are no extraneous effects. More important, however, it allows a straightforward evaluation of the separate effects of several categorical variables. In this chapter we expand Grizzle's procedure and consider models for multiple measures of risk.

To fix ideas, we consider two physical or social states, S and $\bar{S}$, which correspond to smoker and non-smoker, respectively, and two disease conditions D and $\bar{D}$, corresponding to with and without disease. Relative risk is the ratio of two conditional probabilities and is defined

$$R = \frac{P(D|S)}{P(D|\bar{S})} .$$

We consider this measure on the logarithmic scale,

$$\ell n\ R = \ell n\ P(D|S) - \ell n\ P(D|\bar{S}) .$$

At least three different sampling schemes generate data for estimates of ℓn R. The three we consider are directly related to the possible types of 2×2 tables. We present estimates of ℓn R for tables with only the total fixed, with one marginal fixed and with the other marginal fixed. We do not consider tables with both marginals fixed.

We also present measures of ℓn R for tables containing information about relative risk for two categories of subjects and for tables containing information about two sources of relative risk for the same subjects.

We indicate the necessary procedures for analyses for each of the measures presented. We include only one example, for a 2×3 subtable with relative risk for two categories of subjects; however, procedures for the other measures are straightforward applications of the same principles.

3.2. Models for Subtables with only the Total Fixed

If a population is surveyed with no *a priori* restrictions placed on the probability of an individual belonging to any of the four cells in a 2×2 subtable, then the subtable represents a four-category multinomial distribution. A typical subtable representing the $k\ell$-th cell in an overall 2×2×t×w table has the observed cell frequencies

	$\bar{S}$	S	
$\bar{D}$	$n_{11k\ell}$	$n_{12k\ell}$	
D	$n_{21k\ell}$	$n_{22k\ell}$	
			$n_{\cdot\cdot k\ell}$

(3.2.1)

where the corresponding $P_{ijk\ell}$ are such that $\Sigma_{i,j}\, P_{ijk\ell} = 1$ for all combinations of k and ℓ. This subtable essentially corresponds to a 2×2 factorial experiment and we can use the measure of association discussed earlier,

$$u_{k\ell} = \ell n\ p_{11k\ell} - \ell n\ p_{12k\ell} - \ell n\ p_{21k\ell} + \ell n\ p_{22k\ell}, \qquad (3.2.2)$$

to measure the relationship between smoking and disease. As discussed earlier, we may consider the structure of the overall table and obtain estimates and tests accordingly.

However, we are interested specifically in relative risk, which we obtain according to the following scheme. The logarithm of the relative risk, given in terms of $P_{ijk\ell}$, for (3.2.1) is

$$\ell n\ R_{k\ell} = \ell n\ \frac{P_{22k\ell}/(P_{12k\ell} + P_{22k\ell})}{P_{21k\ell}/(P_{11k\ell} + P_{21k\ell})}$$

$$= \ell n\ P_{22k\ell} - \ell n\ (P_{12k\ell} + P_{22k\ell})$$

$$- \ell n\ P_{21k\ell} + \ell n\ (P_{11k\ell} + P_{21k\ell}).$$

An estimate of $\ell n\ R_{k\ell}$ is

$$\ell n\ r_{k\ell} = \ell n\ p_{22k\ell} - \ell n\ (p_{12k\ell} + p_{22k\ell}) \qquad (3.2.3)$$

$$- \ell n\ p_{21k\ell} + \ell n\ (p_{11k\ell} + p_{21k\ell}).$$

The estimated large sample variance of $\ell n\ r_{k\ell}$, calculated from $\underset{\sim}{S} = \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'$, is

$$S_{r_{k\ell}} = \frac{n_{11k\ell}}{n_{21k\ell}(n_{11k\ell} + n_{21k\ell})} + \frac{n_{12k\ell}}{n_{22k\ell}(n_{12k\ell} + n_{22k\ell})} .$$

To compare the use of $\ell n\ r_{k\ell}$ versus the use of $u_{k\ell}$ in (3.2.2), we ignore the subscripts k and ℓ and note that the estimated large sample variance of u is

$$S_u = 1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22},$$

which can be written

$$S_u = \frac{n_{11} + n_{21}}{n_{11}\ n_{21}} + \frac{n_{12} + n_{22}}{n_{12}\ n_{22}} ,$$

or

$$S_u = \left(\frac{n_{11} + n_{21}}{n_{11}}\right)^2 \frac{n_{11}}{n_{21}(n_{11} + n_{21})} + \left(\frac{n_{12} + n_{22}}{n_{12}}\right)^2 \frac{n_{12}}{n_{22}(n_{12} + n_{22})} \ .$$

Clearly,

$$\left(\frac{n_{11} + n_{21}}{n_{11}}\right)^2 \geq 1 \text{ and} \left(\frac{n_{12} + n_{22}}{n_{12}}\right)^2 \geq 1,$$

hence $S_u \geq S_r$. This result provides reason to consider carefully the measure utilized. The measure $\ell n\ R_{k\ell}$ is based on conditional distributions, whereas $u_{k\ell}$ is not.

The overall table under consideration is a 2×2×t×w table. If we let the dimensions denoted by t and w correspond to factors T and W with t and w levels, respectively, we can construct a model in terms of $\ell n\ R_{k\ell}$ for the entire table. We consider, as an example, the model

$$\ell n\ R_{k\ell} = \mu^* + \alpha_k^* + \beta_\ell^* \ , \tag{3.2.4}$$

where μ^* represents an overall mean and α_k^* and β_ℓ^* represent T and W effects. We express this model in terms of $\underset{\sim}{K}\ \ell n\ \underset{\sim}{A}\ \underset{\sim}{P} = \underset{\sim}{X}\ \underset{\sim}{\xi}$ by letting

$$\underset{1\times 4}{\underset{\sim}{P}'_{k\ell}} = [P_{11k\ell}, P_{12k\ell}, P_{21k\ell}, P_{22k\ell}],$$

$$\underset{1\times(4tw)}{\underset{\sim}{P}'} = [\underset{\sim}{P}'_{11}, \underset{\sim}{P}'_{12}, \ldots, \underset{\sim}{P}'_{1t}, \ldots, \underset{\sim}{P}'_{w1}, \underset{\sim}{P}'_{w2}, \ldots, \underset{\sim}{P}'_{wt}], \tag{3.2.5}$$

$$\underset{\sim}{A}_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \text{, and}$$

$$\underset{\sim}{K}_1 = [1 \quad -1 \quad -1 \quad 1].$$

Then we can write $\underset{\sim}{A}$ and $\underset{\sim}{K}$ as the Kronechker products,

$$\underset{4tw\times 4tw}{\underset{\sim}{A}} = \underset{tw\times tw}{\underset{\sim}{I}} \otimes \underset{4\times 4}{\underset{\sim}{A}_1} \text{, and} \tag{3.2.6}$$

$$\underset{tw\times 4tw}{\underset{\sim}{K}} = \underset{tw\times tw}{\underset{\sim}{I}} \otimes \underset{1\times 4}{\underset{\sim}{K}_1} \ .$$

We have $\underset{\sim}{K}_1 \; \ell n \; \underset{\sim}{A}_1 \; \underset{\sim}{P}_{k\ell} = \ell n \; R_{k\ell}$, so that $\underset{\sim}{K} \; \ell n \; \underset{\sim}{A} \; \underset{\sim}{P} = \underset{\sim}{X} \; \underset{\sim}{\xi}$ expresses the tw $\ell n \; R_{k\ell}$'s according to the design given by $\underset{\sim}{X}$. For this example we select $\underset{\sim}{X}$ to incorporate the restrictions $\Sigma \; \hat{\alpha}^* = \Sigma \; \hat{\beta}^* = 0$ and as such it provides the parameter vector

$$\underset{\sim}{\xi}' = [\mu, \alpha_1, \alpha_2, \ldots, \alpha_{t-1}, \beta_1, \beta_2, \ldots, \beta_{w-1}]. \tag{3.2.7}$$

The test of the fit of this model provides a test for no T×W interaction with regard to $\ell n \; R_{k\ell}$. If the model fits, we can compare combinations of elements of $\underset{\sim}{\xi}$ by selecting $\underset{\sim}{C}$ appropriately and testing $\underset{\sim}{C} \; \underset{\sim}{\xi} = \underset{\sim}{0}$. If there is no T×W interaction, no difference among the levels of T, and no difference among the levels of W, we can obtain an estimate of the average of the tw separate risks by considering the model

$$\underset{\sim}{K} \; \ell n \; \underset{tw\times 1}{\underset{\sim}{A} \; \underset{\sim}{P}} = \underset{tw\times 1}{[1, 1, \ldots, 1]'}\mu \ .$$

This model provides an estimate of μ which is analogous to the weighted geometric mean discussed by Gart [1962]. We may test the hypothesis that the average risk is zero in terms of $\underset{\sim}{C} \; \underset{\sim}{\xi} = \underset{\sim}{0}$ with $\underset{\sim}{C}$ being the scalar 1 and $\underset{\sim}{\xi}$ being the scalar μ. This estimate of μ is not adjusted for the presence of any effects of the factors T or W. If T and W effects are present, an adjusted estimate can be obtained by using (3.2.4).

3.3. Subtables with Marginals for S and $\bar{S}$ Fixed

Still concerning ourselves with relationships between smoking and disease, we now consider the marginals for S and $\bar{S}$ fixed. One source for data of this type is the so-called prospective study where smokers and non-smokers are determined and each group is followed to obtain estimates of the proportions that develop disease.

As before, we consider an overal 2×2×t×w table, but one for which a typical 2×2 subtable of observed frequencies is

	$\bar{S}$	S
$\bar{D}$	$n_{11k\ell}$	$n_{12k\ell}$
D	$n_{21k\ell}$	$n_{22k\ell}$
	$n_{\cdot 1k\ell}$	$n_{\cdot 2k\ell}$

(3.3.1)

where the corresponding $P_{ijk\ell}$ are such that $\Sigma_i\, P_{ijk\ell} = 1$ for $j = 1,2$ for all combinations of k and ℓ.

The logarithm of the relative risk for this subtable in terms of the $P_{ijk\ell}$ is

$$\ell n\; R_{k\ell} = \ell n\; [P_{22k\ell}/P_{21k\ell}] = \ell n\; P_{22k\ell} - \ell n\; P_{21k\ell},$$

which we estimate by

$$\ell n\; r_{k\ell} = \ell n\; p_{22k\ell} - \ell n\; p_{21k\ell}\;. \qquad (3.3.2)$$

The estimated large sample variance of $\ell n\; r_{k\ell}$, calculated from $\underset{\sim}{S} = \underset{\sim}{H}\; \underset{\sim}{V}(\underset{\sim}{p})\; \underset{\sim}{H}'$, is

$$S_{r_{k\ell}} = \frac{p_{11k\ell}}{p_{21k\ell}} + \frac{p_{12k\ell}}{p_{22k\ell}} = \frac{n_{11k\ell}}{n_{21k\ell}} + \frac{n_{12k\ell}}{n_{22k\ell}}\;. \qquad (3.3.3)$$

We can use the model in (3.2.4) for the measure given in (3.3.2). If we let

$$\underset{\sim}{P}_{k\ell} = [P_{11k\ell}, P_{21k\ell}, P_{12k\ell}, P_{22k\ell}],$$

$$\underset{\sim}{A}_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\underset{\sim}{K}_1 = [1 \quad -1],$$

then $\underset{\sim}{K}_1 \, \ell n \, \underset{\sim}{A}_1 \, \underset{\sim}{P}_{k\ell} = \ell n \, R_{k\ell}$. We may now consider the $\ell n \, R_{k\ell}$ in a model similar to that indicated by the parameter vector given in (3.2.7). The remainder of the analysis can be done as indicated in Section 3.2.

3.4. Subtables with Marginals for D and $\bar{D}$ Fixed

In many types of research the disease groups, such as subjects with and subjects without lung cancer, are first determined and then the proportions of smokers in each group are estimated. Thus the marginal frequencies associated with D and $\bar{D}$ are fixed.

We retain our concept of an overall 2×2×t×w table, but now the observed frequencies for a typical subtable are given by

	$\bar{S}$	S	
$\bar{D}$	$n_{11k\ell}$	$n_{12k\ell}$	$n_{1\cdot k\ell}$
D	$n_{21k\ell}$	$n_{22k\ell}$	$n_{2\cdot k\ell}$

(3.4.1)

with the corresponding $P_{ijk\ell}$ such that $\Sigma_j \, P_{ijk\ell} = 1$ for i = 1,2 and all combinations of k and ℓ.

Because they arise from retrospective studies, tables similar to (3.4.1) are relatively common, and consequently considerable attention has been given to their analyses. Relative risk is a desirable measure for these tables because of the inferences it allows; however, relative risk is based on $P(D|S)$ and $P(D|\bar{S})$ where as (3.4.1) provides estimates of $P(S|D)$ and $P(\bar{S}|D)$. Using a well-known relationship for conditional probabilities, we can express

$$R = \frac{P(D|S)}{P(D|\bar{S})} = \frac{P(S|D)\ P(\bar{S})}{P(\bar{S}|D)\ P(S)} .$$

Writing R in logarithmic form provides

$$\ell n\ R = \ell n\ P(S|D) - \ell n\ P(\bar{S}|D) - \ell n\ P(S) + \ell n\ P(\bar{S}),$$

which we want to express in terms of the $P_{ijk\ell}$ in (3.4.1). We can obtain readily the conditional probabilities $P(S|D)$ and $P(\bar{S}|D)$, which are $P_{22k\ell}$ and $P_{21k\ell}$ respectively; however, $P(S)$ and $P(\bar{S})$ are more bothersome.

By considering the elements of (3.4.1) somewhat differently, Cornfield [1951] has shown that if $P(D)$ is small relative to $P(S|\bar{D})$ and $P(\bar{S}|\bar{D})$ then $P(S)$ and $P(\bar{S})$ are approximated by $P_{12k\ell}$ and $P_{11k\ell}$ respectively. Under these conditions

$$\ell n\ R_{k\ell} \approx \ell n\ P_{22k\ell} - \ell n\ P_{21k\ell} - \ell n\ P_{12k\ell} + \ell n\ P_{11k\ell},$$

which we estimate by

$$\ell n\ r_{k\ell} = \ell n\ p_{22k\ell} - \ell n\ p_{21k\ell} - \ell n\ p_{12k\ell} + \ell n\ p_{11k\ell}.$$

In terms of the $n_{ijk\ell}$, this estimate is

$$\ell n\ r_{k\ell} = \ell n\ n_{22k\ell} - \ell n\ n_{21k\ell} - \ell n\ n_{12k\ell} + \ell n\ n_{11k\ell}$$

and has estimated asymptotic variance

$$S_{k\ell} = 1/n_{11k\ell} + 1/n_{12k\ell} + 1/n_{21k\ell} + 1/n_{22k\ell}.$$

We note that the estimate of relative risk for (3.4.1) and its variance are identical to the estimate of association (2.5.2) and its variance (2.5.3) given for subtable (2.5.1). Thus investigating association, as we have defined it, in a 2×2 table representing a 4-category multinomial distribution is equivalent to investigating relative risk for (3.4.1), provided that $P(S)$ and $P(\bar{S})$ are approximated by $P_{22k\ell}$ and $P_{21k\ell}$.

We can execute the analysis for the overall 2×2×t×w table in the same manner as shown in Section 3.2.

3.5. Relative Risk for Two Categories of "Diseased" Subjects

We now consider simultaneously the measures of relative risk for two categories of disease, making use of a specific example. Data for examples of this type can be arranged into 2×3 subtables. Our discussion pertains to subtables with only the total fixed; however, we can also extend the methods presented in Sections 3.3 and 3.4 to 2×3 tables.

We analyze the data in Table 3.1, which is a 2×3×2×3 table with subtables similar to (3.5.1)

	Abnormal ECG			
	No	Possible	Definite	
$\overline{\text{CHD}}$	$n_{11k\ell}$	$n_{12k\ell}$	$n_{13k\ell}$	
CHD	$n_{21k\ell}$	$n_{22k\ell}$	$n_{23k\ell}$	
				$n_{\cdot\cdot k\ell}$

(3.5.1)

where the $P_{ijk\ell}$ are such that $\Sigma_{i,j}\, P_{ijk\ell} = 1$ for all combinations of k and ℓ. This subtable provides relative risks for two categories of subjects. The measures are

$$R_1 = \frac{P(\text{CHD}|\text{Possible abnormal ECG})}{P(\text{CHD}|\text{No abnormal ECG})}$$

and

$$R_2 = \frac{P(\text{CHD}|\text{Definite abnormal ECG})}{P(\text{CHD}|\text{No abnormal ECG})} \quad .$$

We are interested in $\ell n\ R_1$ and $\ell n\ R_2$, which we estimate (ignoring the subscripts k and ℓ for the present) by

$$\ell n\ r_1 = \ell n\ p_{22} - \ell n(p_{12} + p_{22}) - \ell n\ p_{21} + \ell n(p_{11} + p_{21})$$

and

$$\ell n\ r_2 = \ell n\ p_{23} - \ell n(p_{13} + p_{23}) - \ell n\ p_{21} + \ell n(p_{11} + p_{21}).$$

Consequently, the vector $[\ell n\ r_1, \ell n\ r_2]'$ contains measures of relative risk for two classifications of subjects with disease.

The estimated large sample variance of $[\ell n\ r_1, \ell n\ r_2]'$, calculated from $\underset{\sim}{S} = \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'$, is

$$\underset{\sim}{S} = \begin{bmatrix} \dfrac{n_{11}}{n_{21}(n_{11}+n_{21})} + \dfrac{n_{12}}{n_{22}(n_{12}+n_{22})} & , & \dfrac{n_{11}}{n_{21}(n_{11}+n_{21})} \\ \\ \dfrac{n_{11}}{n_{21}(n_{11}+n_{21})} & , & \dfrac{n_{11}}{n_{21}(n_{11}+n_{21})} + \dfrac{n_{13}}{n_{23}(n_{13}+n_{23})} \end{bmatrix}$$

We present the remainder of the analysis in terms of Table 3.1, which is extracted from The Framingham Study, Section 9, Table 9-A-14, Exam 1. This table shows the subjects of the study classified by three categories of abnormal ECG and by presence or absence of CHD, for males and females, for three age groups. We assume the six 2×3 subtables are constructed as (3.5.1). Let k = 1,2 and ℓ = 1,2,3 be subscripts for

TABLE 3.1

NUMBER OF SUBJECTS CLASSIFIED BY CORONARY HEART DISEASE (CHD) CATEGORY AND ABNORMAL ECG CATEGORY FOR MALES AND FEMALES FOR THREE AGE GROUPS; DATA EXTRACTED FROM THE FRAMINGHAM STUDY, SECTION 9, TABLE 9-A-14, EXAM 1

Males / **Females**

Age 35-44

Males — Abnormal ECG

	No	Possible	Definite	Sum
$\overline{\text{CHD}}$	764	61	33	858
CHD	2	1	4	7
Sum	766	62	37	865

$\ell n\ r_1 = 1.82$ $\ell n\ r_2 = 3.72$

Females — Abnormal ECG

	No	Possible	Definite	Sum
$\overline{\text{CHD}}$	999	64	28	1091
CHD	2	1	1	4
Sum	1001	65	29	1095

$\ell n\ r_1 = 2.04$ $\ell n\ r_2 = 2.84$

Age 45-54

Males — Abnormal ECG

	No	Possible	Definite	Sum
$\overline{\text{CHD}}$	607	57	41	705
CHD	7	6	13	26
Sum	614	63	54	731

$\ell n\ r_1 = 2.12$ $\ell n\ r_2 = 3.05$

Females — Abnormal ECG

	No	Possible	Definite	Sum
$\overline{\text{CHD}}$	740	79	51	870
CHD	6	6	1	13
Sum	746	85	52	883

$\ell n\ r_1 = 2.17$ $\ell n\ r_2 = 0.87$

Age 55-64

Males — Abnormal ECG

	No	Possible	Definite	Sum
$\overline{\text{CHD}}$	256	38	36	330
CHD	6	4	8	18
Sum	262	42	44	348

$\ell n\ r_1 = 1.42$ $\ell n\ r_2 = 2.07$

Females — Abnormal ECG

	No	Possible	Definite	Sum
$\overline{\text{CHD}}$	344	53	29	426
CHD	4	3	4	11
Sum	348	56	33	437

$\ell n\ r_1 = 1.53$ $\ell n\ r_2 = 2.36$

the sex and age groups, respectively, and let

$$\underset{1\times 6}{\underset{\sim}{P}'_{k\ell}} = [P_{11k\ell}, P_{12k\ell}, P_{13k\ell}, P_{21k\ell}, P_{22k\ell}, P_{23k\ell}],$$

$$\underset{6\times 6}{\underset{\sim}{A}_1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix},$$

$$\underset{2\times 6}{\underset{\sim}{K}_1} = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 & -1 \end{bmatrix}.$$

The expression $\underset{\sim}{K}_1 \; \ell n \; \underset{\sim}{A}_1 \; \underset{\sim}{P}_{k\ell}$ provides

$$\begin{bmatrix} \ell n \; r_{1k\ell} \\ \ell n \; r_{2k\ell} \end{bmatrix} = \underset{\sim}{K}_1 \; \ell n \; \underset{\sim}{A}_1 \; \underset{\sim}{P}_{k\ell}$$

and we have one such vector for each of the six subtables. We relate these vectors to the structure of the overall table according to the model

$$\begin{bmatrix} \ell n \; R_{1k\ell} \\ \ell n \; R_{2k\ell} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \alpha^*_{1k} \\ \alpha^*_{2k} \end{bmatrix} + \begin{bmatrix} \beta^*_{1\ell} \\ \beta^*_{2\ell} \end{bmatrix}, \quad k = 1,2 \quad , \quad \ell = 1,2,3,$$

where

μ_h = overall mean for the h-th ℓn R, h = 1,2,

α_{hk} = sex effect for the h-th ℓn R, and

β_{hk} = age effect for the h-th ℓn R.

We express the model in terms of $\underset{\sim}{K}\ \ell n\ \underset{\sim}{A}\ \underset{\sim}{P} = \underset{\sim}{X}\ \underset{\sim}{\xi}$ by letting

$$\underset{12\times 36}{\underset{\sim}{K}} = \underset{6\times 6}{\underset{\sim}{I}} \otimes \underset{2\times 6}{\underset{\sim}{K}_1}$$

$$\underset{36\times 36}{\underset{\sim}{A}} = \underset{6\times 6}{\underset{\sim}{I}} \otimes \underset{6\times 6}{\underset{\sim}{A}_1}$$

$$\underset{1\times 36}{\underset{\sim}{P}'} = [\underset{\sim}{P}'_{11}\ \underset{\sim}{P}'_{21}\ \underset{\sim}{P}'_{12}\ \underset{\sim}{P}'_{22}\ \underset{\sim}{P}'_{13}\ \underset{\sim}{P}'_{23}],$$

and

$$\underset{12\times 8}{\underset{\sim}{X}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & -1 \\ 1 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & -1 \end{bmatrix} .$$

We have constructed the design matrix $\underset{\sim}{X}$ to incorporate the restriction $\Sigma_k\ \hat{\alpha}^*_{hk} = \Sigma_\ell\ \hat{\beta}^*_{h\ell} = 0$ for $h = 1,2$. We represent the parameter vector as

$$\underset{\sim}{\xi}' = [\mu_1, \mu_2, \alpha_1, \alpha_2, \beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}] ,$$

where

$$\begin{aligned} \alpha^*_{h1} &= \alpha_h, \qquad h = 1,2, \\ \alpha^*_{h2} &= -\alpha_h, \\ \beta^*_{h1} &= \beta_{h1}, \\ \beta^*_{h2} &= \beta_{h2}, \\ \beta^*_{h3} &= -\beta_{h1} - \beta_{h2}. \end{aligned}$$

The analysis provides the parameter estimates and their standard errors as appear in Table 3.2. It is clear from examining these estimates in relation to their standard errors that there is little evidence of either age or sex effect on the vector $[\ell n\ r_1, \ell n\ r_2]'$. We summarize the analysis in Table 3.3.

TABLE 3.2

PARAMETERS AND ESTIMATES FOR DATA IN TABLE 3.1

Parameter	Estimate	Standard Error	Parameter	Estimate	Standard Error
μ_1	1.85	0.354	β_{11}	0.08	0.610
μ_2	2.68	0.328	β_{21}	0.68	0.496
α_1	-0.06	0.287	β_{12}	0.30	0.419
α_2	0.28	0.307	β_{22}	-0.09	0.382

The residual SS provides a test of the fit of the model, and we see the model fits adequately. Hence we can investigate the main effects for

TABLE 3.3

ANALYSIS OF VARIANCE FOR DATA IN TABLE 3.1

Source	DF	SS
Sex	2	1.22
Age	4	3.49
Residual	4	3.47

age and sex. We obtain the SS for sex by testing the hypothesis $\underset{\sim}{C}_1 \underset{\sim}{\xi} = \underset{\sim}{0}$, where

$$C_1 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} .$$

Since the rank of C_1 is two, there are two degrees of freedom for sex. Similarly, for the SS for age we test $\underset{\sim}{C}_2 \, \underset{\sim}{\xi} = \underset{\sim}{0}$, where

$$\underset{\sim}{C}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

Since observed age and sex effects are not significant, we may obtain estimates of the average risk for both $\ell n \, R_1$ and $\ell n \, R_2$. We do this by using the model $\underset{\sim}{K} \, \ell n \, \underset{\sim}{A} \, \underset{\sim}{P} = \underset{\sim}{X} \, \underset{\sim}{\xi}_1$, where $\underset{\sim}{K}$, $\underset{\sim}{A}$, and $\underset{\sim}{P}$ are as before, but

$$\underset{\sim}{X}' = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix},$$

$$\underset{\sim}{\xi}_1' = [\mu_1, \mu_2] .$$

This model provides the estimate

$$\hat{\mu}_1 = 1.90,$$

$$\hat{\mu}_2 = 2.66.$$

We can examine the hypotheses $\mu_1 = 0$ and $\mu_2 = 0$ by testing $\underset{\sim}{C}_3 \, \underset{\sim}{\xi}_1 = \underset{\sim}{0}$ and $\underset{\sim}{C}_4 \, \underset{\sim}{\xi}_1 = \underset{\sim}{0}$ with

$$\underset{\sim}{C}_3 = [1,0],$$

$$\underset{\sim}{C}_4 = [0,1].$$

We can reject these hypotheses since the sum of squares obtained in testing them, 44.42 and 103.46, are values of statistics each of which follows a limiting chi-square distribution with one degree of freedom. We may now compare the two average risks to test whether subjects in the abnormal ECG category "definite" have a relative risk different from that for subjects in the category "possible." We effect this comparison by testing $\underset{\sim}{C}_5 \underset{\sim}{\xi}_1 = \underset{\sim}{0}$, with

$$\underset{\sim}{C}_5 = [1, -1].$$

This test provides the sum of squares 7.42 as value of a statistic which follows a limiting chi-square distribution with one degree of freedom. Thus we can reject the null hypothesis ($p < .05$).

The results of the analysis of Table 3.1 indicate there is no evidence that either $\ell n\ R_1$ or $\ell n\ R_2$ varies among the combinations of age and sex. There is evidence that both subjects in the "possible" and in the "definite" categories of abnormal ECG have greater risk for CHD than subjects in the category "no." Further, we have evidence that the relative risk for subjects in the category "definite" is greater than that for subjects in the category "possible."

The method presented here for 2×3 subtables allows the simultaneous analyses of two categories of relative risk. One category is for subjects with "possible" abnormal ECG and the other is for subjects with "definite" abnormal ECG. The model presented comprises the simultaneous consideration of two univariate linear models, one for each category of relative risk.

3.6. Two Sources of Risk for the Same Subject

We now examine the situation where each individual is subjected to two sources of risk. For convenience, we continue by discussing the relationship between smoking and disease (D), except we now add a new dimension, say hypertension. We examine simultaneously the risk for the disease pertaining to smoking and to hypertension. We can also measure how these two relative risks are related.

We present the measures of relative risk in terms of a 2×4×t×w table where t and w represent the levels of the factors T and W as discussed in Section 3.2. A typical 2×4 subtable has the expected cell probabilities as shown in (3.6.1) where H and $\bar{H}$ denote presence and absence of hypertension, respectively, and the other headings are for presence and absence of smoking and disease.

	$\bar{S}$		S		
	$\bar{H}$	H	$\bar{H}$	H	
$\bar{D}$	$P_{11k\ell}$	$P_{12k\ell}$	$P_{13k\ell}$	$P_{14k\ell}$	
D	$P_{21k\ell}$	$P_{22k\ell}$	$P_{23k\ell}$	$P_{24k\ell}$	
					1

(3.6.1)

where $\Sigma_{i,j} P_{ijk\ell} = 1$ for all combinations of k and ℓ. We let

$$R_S = \frac{P(D|S)}{P(D|\bar{S})}$$

and

$$R_H = \frac{P(D|H)}{P(D|\bar{H})} \quad ,$$

whence for (3.6.1),

$$R_{S_{k\ell}} = \frac{(P_{23k\ell} + P_{24k\ell})/(P_{13k\ell} + P_{14k\ell} + P_{23k\ell} + P_{24k\ell})}{(P_{21k\ell} + P_{22k\ell})/(P_{11k\ell} + P_{12k\ell} + P_{21k\ell} + P_{22k\ell})}$$

and

$$R_{H_{k\ell}} = \frac{(P_{22k\ell} + P_{24k\ell})/(P_{12k\ell} + P_{14k\ell} + P_{22k\ell} + P_{24k\ell})}{(P_{21k\ell} + P_{23k\ell})/(P_{11k\ell} + P_{13k\ell} + P_{21k\ell} + P_{23k\ell})} .$$

We wish to consider the logarithm of these measures of relative risk, so that if we let

$$\underset{1\times 8}{\tilde{P}'_{k\ell}} = [P_{11k\ell}, P_{12k\ell}, P_{13k\ell}, P_{14k\ell}, P_{21k\ell}, P_{22k\ell}, P_{23k\ell}, P_{24k\ell}],$$

$$\underset{8\times 8}{\tilde{A}_1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

and

$$\underset{2\times 8}{\tilde{K}_1} = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{bmatrix} ,$$

then

$$\tilde{K}_1 \ \ell n \ \tilde{A}_1 \ \tilde{P}_{k\ell} = \begin{bmatrix} \ell n \ R_{S_{k\ell}} \\ \ell n \ R_{H_{k\ell}} \end{bmatrix} .$$

An estimate of $\tilde{K}_1 \ \ell n \ \tilde{A}_1 \ \tilde{P}_{k\ell}$ is $\tilde{K}_1 \ \ell n \ \tilde{A}_1 \ \tilde{p}_{k\ell}$. The estimated large sample variance of this estimate can be calculated from $\tilde{S} = \tilde{H} \ \tilde{V}(\tilde{p}) \ \tilde{H}'$ and can be written

$$\underset{2\times 2}{\underset{\sim}{S}} = \begin{bmatrix} \widehat{\mathrm{Var}}[\ell n\ r_S] & \widehat{\mathrm{Cov}}[\ell n\ r_S,\ \ell n\ r_H] \\ \widehat{\mathrm{Cov}}[\ell n\ r_S,\ \ell n\ r_H] & \widehat{\mathrm{Var}}[\ell n\ r_H] \end{bmatrix}.$$

From $\underset{\sim}{S}$ we can obtain readily a measure of the correlation between the two measures of relative risk.

A model for the overall table is

$$\begin{bmatrix} \ell n\ R_{S_{k\ell}} \\ \ell n\ R_{H_{k\ell}} \end{bmatrix} = \begin{bmatrix} \mu_S \\ \mu_H \end{bmatrix} + \begin{bmatrix} \alpha_{S_k} \\ \alpha_{H_k} \end{bmatrix} + \begin{bmatrix} \beta_{S_\ell} \\ \beta_{H_\ell} \end{bmatrix},$$

where

μ_S = overall mean for $\ell n\ R_S$,

μ_H = overall mean for $\ell n\ R_H$,

α_{S_k} = effect of k-th level of T on $\ell n\ R_S$,

α_{H_k} = effect of k-th level of T on $\ell n\ R_H$

β_{S_ℓ} = effect of ℓ-th level of W on $\ell n\ R_S$, and

β_{H_ℓ} = effect of ℓ-th level of W on $\ell n\ R_H$.

This model can be written as $\underset{\sim}{K}\ \ell n\ \underset{\sim}{A}\ \underset{\sim}{P} = \underset{\sim}{X}\ \underset{\sim}{\xi}$ with the matrices constructed in much the same way as they were for the example in Section 3.5; consequently the remainder of the analysis can be completed by following the procedures given in Section 3.5. The additional calculation for the correlation between the log relative risks is done easily.

3.7. Summary

We have discussed three different functions which reflect three different sampling schemes commonly encountered in estimating relative risk. We have shown how these functions can be used as dependent variables in a general linear model for which all independent variables are categorical variables. The proposed functions and their statistical analysis constitute a method which is easy to use and which allows the analysis to reflect properly the sampling scheme. We have also applied this method to data in 2×3 subtables and have examined the relationship between the two categories of relative risk provided by such tables. This method can be extended to be applicable for subtables of other dimensions; however, there may be some tables for which a meaningful measure of relative risk cannot be defined.

In addition we have presented a technique for situations where individuals are subjected to risk from more than one source. This technique is actually a multivariate technique and could as well have been discussed in Chapter V, which contains other multivariate techniques, however we have placed it here in order to have a more complete discussion of relative risk.

CHAPTER IV

METHODS FOR ORDERED RESPONSE CATEGORIES

4.1. Introduction

Many areas of research providing categorical data have observations classified into ordered response categories. In some cases the categories represent intervals of a continuous measurement, but often they indicate such rankings as low, medium, and high, or 1-plus, 2-plus, etc. Such data often have been analyzed by techniques commonly used for unordered categories. However, the analysis of this kind of data is improved generally by analytical techniques which allow one to consider the relationships among the ordered categories.

We present two possible methods of analysis allowing consideration of the relationships among ordered categories. We call one technique the scoring method, because it utilizes scores determined for each of the ordered categories and allows us to calculate a mean score for each multinomial distribution. This method is related to a regression approach discussed by Sen [1967] which was extended by Ghosh [1969] to include c-sample methods with multivariate stochastic scores. We have derived the other technique from a model proposed initially by Thurstone [1959] and discussed by Bock and Jones [1967]. Our second method provides estimates of the distances between categories by considering models for cumulative probabilities.

4.2. The Scoring Method

We present an example of the scoring method for an r×s×t table representing rs independent multinomial distributions each with t ordered response categories. The p_{ijk} for this table are such that $\Sigma_k\, p_{ijk} = 1$ for all combinations of i and j. For each p_{ijk} we determine a score z_{ijk} and we write the scores for each multinomial in the vector

$$\underset{1\times t}{\underset{\sim}{z}'_{ij}} = [z_{ij1}, z_{ij2}, \ldots, z_{ijt}], \quad i = 1,2,\ldots,r, \quad j = 1,2,\ldots,s.$$

We arrange the scores for the entire table in the block diagonal matrix

$$\underset{rs\times rst}{\underset{\sim}{Z}} = \text{Diagonal } [\underset{\sim}{z}'_{11}, \underset{\sim}{z}'_{12}, \ldots, \underset{\sim}{z}'_{1s}, \ldots, \underset{\sim}{z}'_{r1}, \underset{\sim}{z}'_{r2}, \ldots, \underset{\sim}{z}'_{rs}].$$

The scalar $\underset{\sim}{z}'_{ij}\, \underset{\sim}{p}_{ij}$ is a mean score for the ij-th multinomial and $\underset{\sim}{Z}\underset{\sim}{p}$ is the vector of mean scores for the r×s×t table.

Selection of the z_{ijk} determines the relationship of the response categories. We can select the ranks of the response categories to be the z_{ijk} and, in this case, $z_{ijk} = k$ and $\underset{\sim}{z}'_{ij} = [1,2,\ldots,t]$ for all i and j. This selection is appropriate for equally spaced categories. If the categories are not equally-spaced, but distances among them are known, we can alter $\underset{\sim}{z}_{ij}$ to reflect the actual spacing. Often, however, we do not know the distances and we cannot assume them to be equal. For such situations we prefer that the scores be determined from the data and, consequently, we use percentile scores for the z_{ijk}. Bross [1958] describes the calculation of percentile scores from an empirical reference distribution. Garrett [1953] describes some of their properties and gives several examples of their use. We calculate the scores from the marginal distribution for the ordered categories of the data being analyzed.

We demonstrate the scoring method, using percentile scores as the z_{ijk}, by considering the data in Table 4.1, taken from Bahr [1969]. These data are a tabulation of indices for the consumption of alcoholic beverages for residents of three locations, classified by number of years lived at that location. In the analysis we consider the index of drinking to be the ordered response variable, the locations to correspond to blocks as in a blocked design, and "number of years in quarter" to be the design variable of principal interest.

If locations did not represent blocks and if we wished to test for homogeneity of the effects of location and "number of years in quarters" as in a factorial experiment, we would calculate marginal percentile scores from the combined data for the entire table. Under these conditions $\underset{\sim}{z}_{ij} = \underset{\sim}{z}$ for all combinations of i and j, where i = 1,2,3, is the index for location and j = 1,2,3, is the index for number of years in quarters. However, if the experiment contains non-homogeneous blocks, we should calculate the scores within each block. For data such as in Table 4.1, locations are considered blocks and we calculate the scores within each location.

To calculate the percentile scores, we let

$$p_{i\cdot k} = n_{i\cdot k}/n_{i\cdot\cdot} \qquad i = 1,2,3, \qquad k = 1,2,3,$$

where k is the index for extent of drinking, be the observed marginal probabilities for block i. Then $p_{i\cdot 1}$ is the observed probability of a subject in block i having a quantitative index of "light or abstention," $p_{i\cdot 2}$ is the observed probability in block i for the index "moderate,"

TABLE 4.1

NUMBER OF SUBJECTS CLASSIFIED BY EXTENT OF CURRENT DRINKING AND NUMBER OF YEARS LIVED IN GROUP QUARTERS FOR THREE LOCATIONS, DATA EXTRACTED FROM BAHR [1969], TABLE 6, PAGE 374

Quantitative index of extent of drinking	Bowery: Number of years in quarters 0	1-4	5+	Total	Percentile score
Light and abstention	25	21	20	66	.1701
Moderate	21	18	19	58	.4897
Heavy or spree	26	23	21	70	.8196
Total	72	62	60	194	
Mean score	.498	.504	.499		
Standard error (x10)	.322	.347	.346		

	Camp: Number of years in quarters 0	1-4	5+	Total	Percentile score
Light and abstention	29	16	8	53	.1352
Moderate	27	13	11	51	.4005
Heavy or spree	38	24	30	92	.7635
Total	94	53	49	196	
Mean score	.466	.486	.581		
Standard error (x10)	.275	.374	.351		

	Park Slope: Number of years in quarters 0	1-4	5+	Total	Percentile score
Light and abstention	44	18	6	68	.2833
Moderate	19	9	8	36	.7167
Heavy or spree	9	4	3	16	.9333
Total	72	31	17	120	
Mean score	.479	.493	.602		
Standard error (x10)	.298	.258	.601		

and similarly for $p_{i\cdot 3}$. The percentile scores for the i-th block

$$a_{i1} = p_{i\cdot 1}/2$$

$$a_{i2} = p_{i\cdot 1} + p_{i\cdot 2}/2$$

$$a_{i3} = p_{i\cdot 1} + p_{1\cdot 2} + p_{i\cdot 3}.$$

These scores are related to ranks, since subjects in block i in the category corresponding to the marginal probability $p_{i\cdot 1}$ are tied for the lowest rank in that block and their midrank is proportional to $p_{i\cdot 1}/2$. Those in the category corresponding to $p_{i\cdot 2}$ are tied for the second rank and their midrank is proportional to $p_{i\cdot 1} + p_{i\cdot 2}/2$, and so forth.

The percentile scores and the mean scores for each multinomial distribution appear in Table 4.1. We can express the percentile scores for each of the three locations in the vectors

$$\underset{\sim}{z}_1' = [.1701 \quad .4897 \quad .8196],$$

$$\underset{\sim}{z}_2' = [.1352 \quad .4005 \quad .7653],$$

$$\underset{\sim}{z}_3' = [.2833 \quad .7167 \quad .9333].$$

Then the complete matriz $\underset{\sim}{Z}$ is the block diagonal matrix

$$\underset{9\times 27}{\underset{\sim}{Z}} = \text{Diagonal } [\underset{\sim}{z}_1', \underset{\sim}{z}_1', \underset{\sim}{z}_1', \underset{\sim}{z}_2', \underset{\sim}{z}_2', \underset{\sim}{z}_2', \underset{\sim}{z}_3', \underset{\sim}{z}_3', \underset{\sim}{z}_3'].$$

A model for such data is

$$E[\underset{\sim}{z}_{ij}' \, \underset{\sim}{p}_{ij}] = \mu^* + \alpha_i^* + \beta_j^*,$$

where μ^* is the overall mean; α_i^*, i = 1,2,3, is the effect associated with the i-th location or block; and β_j^*, j = 1,2,3, is the effect associated with numbers of years lived at the location. This model can be written

$$\underset{9\times 27}{\underset{\sim}{Z}} \quad \underset{27\times 1}{\underset{\sim}{P}} = \underset{9\times 5}{\underset{\sim}{X}} \quad \underset{5\times 1}{\underset{\sim}{\xi}} \ ,$$

where $\underset{\sim}{X}$ has been constructed to incorporate the restriction $\Sigma\hat{\alpha}_i^* = \Sigma\hat{\beta}_j^* = 0$ and is given by

$$\underset{9\times 5}{\underset{\sim}{X}} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{bmatrix} ,$$

with

$$\underset{\sim}{\xi}' = [\mu, \alpha_1, \alpha_2, \beta_1, \beta_2] \ .$$

The remainder of the analysis is straightforward. The estimated parameters and their standard errors appear in Table 4.2. We show the results for tests of hypotheses in the analysis of variance in Table 4.3. Each of the SS has a limiting central chi-square distribution if its corresponding null hypothesis is true. The residual SS provides a test of the fit of the model, and the SS for "Years lived at location" provides a test for absence of this effect. Thus we see a significant "Years lived at location" effect with most of the difference coming from the 0 vs 5+ comparison.

TABLE 4.2

ESTIMATED PARAMETERS AND THEIR STANDARD ERRORS
FOR THE DATA IN TABLE 4.1

Parameter	Estimate	Standard error
μ	.508	.012
α_1	-.007	.017
α_2	.000	.016
β_1	-.029	.016
β_2	-.012	.018

TABLE 4.3

ANALYSIS OF VARIANCE FOR THE DATA IN TABLE 4.1

Source of Variation	DF	SS
Locations	2	0.20
Years lived at location	2	6.11
0 vs 5+	1	5.99
0 vs 1-4	1	0.36
Residual	4	4.18

Examination of the mean scores in Table 4.1 indicates that the mean rank score is related nonlinearly to the time in residence in Camp and Park Slope while there is very little effect of time in residence for the Bowery. In Table 4.3, we compared 0 with 5+ and 0 with 1-4 for years lived at location. We could, of course, have chosen to test instead for linear or quadratic effects of this variable.

4.3. An Approach Derived from Thurstone's Model

Thurstone [1959] presents a model which permits the analysis of data arranged in ordered response categories. He assumes that the ordered categories are derived from an underlying continuous distribution of a specified form. He utilizes this distribution to obtain a model for placing objects into the ordered categories. Bock and Jones [1967, chapter 8] discuss Thurstone's model with specific reference to tables with s independent multinomial distributions, each with r ordered response categories. We examine their discussion and use it as a basis for a technique permitting an investigation of interaction involving category effects.

First we need to consider the following notation:

Multinomial population	Ordered category 1	2	...	r	Total
1	P_{11}	P_{12}	...	P_{1r}	1
2	P_{21}	P_{22}	...	P_{2r}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	P_{s1}	P_{s2}	...	P_{sr}	1

(4.3.1)

where $\Sigma_k P_{ik} = 1$ for all i. Two simultaneously operating processes are integral to the Bock and Jones discussion and to describe them we denote the objects in the i-th multinomial population by X_{ij}, $j = 1,2,\ldots,n_{i.}$ and the categories by C_k, $k = 1,2,\ldots,r$. Then we assume the object X_{ij} is evaluated and assigned the value

$$v_{ij} = \mu_i + e_{ij}$$

where

$$\mu_i = \text{mean effect for population i, and}$$

$$e_{ij} = \text{random effect associated with object j in population i.}$$

Further, we assume that the categories are evaluated according to the model

$$w_{ik} = \tau_k + d_{ik},$$

where

τ_k = overall mean value associated with the k-th category, and

d_{ik} = random effect associated with the evaluation of the k-th category for the i-th population.

Bock and Jones then construct a model for placing objects into categories using

$$\begin{aligned} v_{ijk} &= v_{ij} - w_{ik} \\ &= \mu_i - \tau_k + (e_{ij} - d_{ik}), \end{aligned} \tag{4.3.2}$$

with $E(e_{ij} - e_{ik}) = 0$ and $Var(e_{ij} - d_{ik}) = \sigma_i^2$. According to this model, if $v_{ijk} \leq 0$, then the object X_{ij} is placed into category C_k or below. Thus we may conveniently consider the cumulative probabilities

$$\begin{aligned} \phi_{i1} &= P_{i1}, \\ \phi_{i2} &= P_{i1} + P_{i2}, \\ &\vdots \\ \phi_{i(r-1)} &= P_{i1} + P_{i2} + \ldots + P_{i(r-1)}, \\ \phi_{ir} &= 1. \end{aligned} \tag{4.3.3}$$

Using this notation, $\phi_{ik} = P(v_{ijk} \leq 0)$; we may now include the form of the underlying continuous distribution. We consider the logistic distribution with the density

$$g(z) = \frac{1}{4}\,\mathrm{sech}^2\left(\frac{z}{2}\right),$$

which has mean zero and variance $\pi^2/3$. We have selected this distribution both because of its similarity to the normal distribution and also because it allows us to use variables of the form $\ell n\ [p/(1-p)]$, which can be expressed in terms of $\underset{\sim}{K}\ \ell n\ \underset{\sim}{A}\ \underset{\sim}{p}$. To relate this distribution to our problem, we transform to mean $\mu_i - \tau_k$ and variance σ_i^2 by letting

$$t = \frac{z\,\sigma_i\,\sqrt{3}}{\pi} + \mu_i - \tau_k \quad .$$

Thus

$$\phi_{ik} = \int_{-\infty}^{0} \frac{1}{4}\,\mathrm{sech}^2\ \frac{t-(\mu_i-\tau_k)}{2\,\sigma_i\,\sqrt{3}/\pi}\ \frac{\pi}{\sigma_i\,\sqrt{3}}\ dt$$

$$= \frac{1}{2}\tanh\ \frac{t-(\mu_i-\tau_k)}{2\,\sigma_i\,\sqrt{3}/\pi}\ \Bigg|_{-\infty}^{0} \quad .$$

But $\tanh(u) = \dfrac{e^{2u}-1}{e^{2u}+1}$, so that

$$\phi_{ik} = \frac{1}{2}\ \frac{e^{\frac{t-(\mu_i-\tau_k)}{\sigma_i\,\sqrt{3}/\pi}} - 1}{e^{\frac{t-(\mu_i-\tau_k)}{\sigma_i\,\sqrt{3}/\pi}} + 1}\ \Bigg|_{-\infty}^{0}$$

$$= \frac{1}{2}\left[\frac{e^{\frac{-(\mu_i - \tau_k)}{\sigma_i \sqrt{3}/\pi}} - 1}{e^{\frac{-(\mu_i - \tau_k)}{\sigma_i \sqrt{3}/\pi}} + 1} + 1\right]$$

$$= e^{\frac{-(\mu_i - \tau_k)}{\sigma_i \sqrt{3}/\pi}} \Bigg/ \left[e^{\frac{-(\mu_i - \tau_k)}{\sigma_i \sqrt{3}/\pi}} + 1\right] ,$$

and

$$1 - \phi_{ik} = 1 \Bigg/ \left[e^{\frac{-(\mu_i - \tau_k)}{\sigma_i \sqrt{3}/\pi}} + 1\right] .$$

Thus we can write a model involving the population effect and the category effect,

$$\ell n[\phi_{ik}/(1 - \phi_{ik})] = - (\mu_i - \tau_k)\pi/\sigma_i\sqrt{3} \quad . \qquad (4.3.4)$$

This model is scaled by the quantity $-\pi/\sigma_i \sqrt{3}$. Consequently, if more than one population is involved and we wish to obtain estimates of τ_k on the same scale across populations, then the σ_i must all either be known or be equal. We note also that this model contains no terms similar to $(\mu\tau)_{ik}$ and thus does not permit consideration of interaction between populations and categories.

Clearly, the above model, as considered by Bock and Jones, requires the assumptions

(1) the true underlying continuous distribution is the logistic distribution,

(2) there is no interaction of the form $(\mu\tau)_{ik}$, and

(3) the variances of the underlying distributions are equal.

We now propose an alternative to the Bock and Jones procedure that avoids the necessity of requirements (1) and (2) above. We note that an interaction term such as $(\mu\tau)_{ik}$ is easily included in the model given in (4.3.4); however, if the s multinomial distributions are cross-classified in a factorial experiment, then $(\mu\tau)_{ik}$ represents a higher order interaction term. We organize the analysis so that we can investigate low-order interaction involving category effects and, since we use the least-squares procedure of Grizzle, Starmer, and Koch, we avoid problems associated with the form of the underlying continuous distribution.

To demonstrate the procedure, we examine the data in Table 4.4, which Bock and Jones [1967, chapter 8] previously analyzed. The Acceptance Branch of the Food and Container Institute for the Armed Forces obtained these data in a survey of U. S. Armed Forces personnel and the data concern a preference for black olives. These data were collected from six populations which were cross-classified into a 2×3 factorial experiment with one factor being "urbanization," at the levels rural and urban, and the other being "location," at the levels NE, MW, and SW. For the data collection, interviewers asked the subjects to indicate their preference for black olives on a nine-category scale, later reduced to six categories by combining categories 2-3, 4-5, and 8-9. Under the conditions of this survey, the subject evaluated the ordered categories. Consequently, we are interested in any information concerning the failure of subjects in different populations to evaluate their preference according to the same scale. In the analysis such information appears in the form of interactions for category effects.

TABLE 4.4

FREQUENCIES OF PREFERENCES FOR BLACK OLIVES CLASSIFIED BY RURAL, URBAN, AND LOCATION, DATA TAKEN FROM TABLE 8.11, PAGE 244 OF BOCK AND JONES [1967]

Urbanization	Location	a	b	c	d	e	f	Total
		Ordered Category						
Rural	NE	23	18	20	18	10	15	104
	MW	30	22	21	17	8	12	110
	SW	11	9	26	19	17	24	106
Urban	NE	18	17	18	18	6	25	102
	MW	20	15	12	17	16	28	108
	SW	12	9	23	21	19	30	114

To develop our procedure, we need the following notation. Let the expected cell probabilities be denoted by P_{ijk}, where i = 1,2 corresponds to Rural and Urban, j = 1,2,3, corresponds to NE, MW, and SW, k = 1,2,...,6, corresponds to the ordered response categories, a,b,...,f. We write these probabilities in the vector

$$\underset{\sim}{P}' = [\underset{\sim}{P}'_{11}, \underset{\sim}{P}'_{12}, \underset{\sim}{P}'_{13}, \underset{\sim}{P}'_{21}, \underset{\sim}{P}'_{22}, \underset{\sim}{P}'_{23}],$$

where

$$\underset{\sim}{P}'_{ij} = [P_{ij1}, P_{ij2}, \ldots, P_{ij6}].$$

We now construct the logits for the cumulative probabilities in the same way that Bock and Jones constructed them. For each ij combination we construct five logits similar to

$$\ell_{ijh} = \ell n\, [\phi_{ijh}/(1 - \phi_{ijh})], \quad h = 1,2,\ldots, 5,$$

where the ϕ_{ijh} are the cumulative probabilities defined in (4.3.3). If, for each ij combination, we let

$$\underset{1\times 5}{\underset{\sim}{\ell}'_{ij}} = [\ell_{ij1}, \ell_{ij2}, \ldots, \ell_{ij5}],$$

$$\underset{5\times 10}{\underset{\sim}{K}^{*}} = \begin{bmatrix} 1 & -1 & & & & & & & & \\ 0 & 0 & 1 & -1 & & & & & & \\ 0 & 0 & 0 & 0 & 1 & -1 & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

and

$$\underset{10\times 6}{\underset{\sim}{A}^{*}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

then

$$\underset{\sim}{\ell}_{ij} = \underset{\sim}{K}^{*} \,\ell n\, \underset{\sim}{A}^{*} \,\underset{\sim}{P}_{ij}.$$

We now express each of the five logits in a separate linear model such as

$$\ell_{ijh} = \mu_h + \alpha^{*}_{ih} + \beta^{*}_{jh},$$

and we consider the five models for the five logits simultaneously in the form

$$\begin{bmatrix} \ell_{ij1} \\ \ell_{ij2} \\ \ell_{ij3} \\ \ell_{ij4} \\ \ell_{ij5} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} + \begin{bmatrix} \alpha^*_{i1} \\ \alpha^*_{i2} \\ \alpha^*_{i3} \\ \alpha^*_{i4} \\ \alpha^*_{i5} \end{bmatrix} + \begin{bmatrix} \beta^*_{j1} \\ \beta^*_{j2} \\ \beta^*_{j3} \\ \beta^*_{j4} \\ \beta^*_{j5} \end{bmatrix} , \tag{4.3.5}$$

where

μ_h = overall mean for the h-th logit, h = 1, 2, ..., 5,

α^*_{ih} = effect of i-th level of urbanization on the h-th logit,

β^*_{jh} = effect of j-th level of location on the h-th logit.

We can express this complete model as $\underset{\sim}{K}\, \ell n\, \underset{\sim}{A}\, \underset{\sim}{P} = \underset{\sim}{X}\, \underset{\sim}{\xi}$, where

$$\underset{30\times 60}{\underset{\sim}{K}} = \underset{6\times 6}{\underset{\sim}{I}} \otimes \underset{5\times 10}{\underset{\sim}{K}^*} ,$$

$$\underset{60\times 36}{\underset{\sim}{A}} = \underset{6\times 6}{\underset{\sim}{I}} \otimes \underset{10\times 6}{\underset{\sim}{A}^*} ,$$

and

$$\underset{30\times 20}{\underset{\sim}{X}} = \begin{bmatrix} \underset{\sim}{I}_5 & \underset{\sim}{I}_5 & \underset{\sim}{I}_5 & \underset{\sim}{0}_5 \\ \underset{\sim}{I}_5 & \underset{\sim}{I}_5 & \underset{\sim}{0}_5 & \underset{\sim}{I}_5 \\ \underset{\sim}{I}_5 & \underset{\sim}{I}_5 & -\underset{\sim}{I}_5 & -\underset{\sim}{I}_5 \\ \underset{\sim}{I}_5 & -\underset{\sim}{I}_5 & \underset{\sim}{I}_5 & \underset{\sim}{0}_5 \\ \underset{\sim}{I}_5 & -\underset{\sim}{I}_5 & \underset{\sim}{0}_5 & \underset{\sim}{I}_5 \\ \underset{\sim}{I}_5 & -\underset{\sim}{I}_5 & -\underset{\sim}{I}_5 & -\underset{\sim}{I}_5 \end{bmatrix} , \tag{4.3.6}$$

with $\underset{\sim}{I}_5$ representing a 5×5 identity matrix, $-\underset{\sim}{I}_5$ representing the negative of a 5×5 identity matrix and $\underset{\sim}{0}_5$ representing the 5×5 null matrix. The vector of parameters is

$$\underset{\sim}{\xi}' = [\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \qquad (4.3.7)$$

$$\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25}],$$

with μ_h as defined for (4.3.5) and the urbanization and location effects reparameterized so that $\sum_i \hat{\alpha}_{ih} = \sum_j \hat{\beta}_{jh} = 0$ for all h. Thus

$$\hat{\alpha}^*_{1h} = \hat{\alpha}_h,$$

$$\hat{\alpha}^*_{2h} = \hat{\alpha}_h,$$

$$\hat{\beta}^*_{1h} = \hat{\beta}_{1h},$$

$$\hat{\beta}^*_{2h} = \hat{\beta}_{2h},$$

$$\hat{\beta}^*_{3h} = -\hat{\beta}_{1h} - \hat{\beta}_{2h}, \quad h = 1, 2, \ldots, 5.$$

We note that if we arrange the parameter vector as in (4.3.7), then we can write the design matrix $\underset{\sim}{X}$ with the usual +1's, -1's, and 0's being replaced in the same pattern by $\underset{\sim}{I}_5$'s, $-\underset{\sim}{I}_5$'s, and $\underset{\sim}{0}_5$'s.

This model produces the estimates of parameters and their standard errors given in Table 4.5. The μ_h are scale constants for the categories adjusted for location and urbanization effect. We are interested in relationships among the categories, but first we need to look for interaction between category "effects" and the other effects. We summarize information concerning these interactions and other effects in Table 4.6. The test for the fit of the model (4.3.5) is a test for the combined Urbanization × Location and Category × Urbanization ×

TABLE 4.5

ESTIMATES OF PARAMETERS AND THEIR STANDARD ERRORS FOR THE DATA IN TABLE 4.4

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
μ_1	-1.585	.108	β_{11}	.190	.148
μ_2	- .806	.088	β_{12}	.269	.122
μ_3	.017	.080	β_{13}	.202	.114
μ_4	.753	.087	β_{14}	.245	.125
μ_5	1.358	.101	β_{15}	.052	.141
α_1	.133	.104	β_{21}	.370	.143
α_2	.144	.085	β_{22}	.396	.120
α_3	.226	.080	β_{23}	.176	.113
α_4	.250	.086	β_{24}	.127	.123
α_5	.300	.100	β_{25}	.161	.142

Location interaction effects. We can examine the Category × Location interaction by testing simultaneously whether the five β_{1h} as well as the five β_{2h} are equal. We can do this by testing $\underset{\sim}{C}\,\underset{\sim}{\xi} = \underset{\sim}{0}$, where

$$\underset{8\times 20}{\underset{\sim}{C}} = \begin{bmatrix} \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{C}_1 & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{0} & \underset{\sim}{C}_1 \end{bmatrix},$$

with

$$\underset{4\times 5}{\underset{\sim}{C}_1} = \begin{bmatrix} 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

and $\underset{\sim}{0}$ being the 4×5 null matrix. Likewise we test for no Category × Urbanization interaction with

$$\underset{4\times 20}{\underset{\sim}{C}} = [\underset{\sim}{0} \quad \underset{\sim}{C}_1 \quad \underset{\sim}{0} \quad \underset{\sim}{0}].$$

TABLE 4.6

ANALYSIS OF VARIANCE FOR THE DATA IN TABLE 4.4

Source	DF			SS
Category	4			526.36*
Urbanization	1		8.95*	11.50*
Category × Urbanization	4		2.52	
Location	2			18.86*
NE vs SW		1	12.84*	
MW vs SW		1	15.51*	
Category × Location	8			14.91
Urbanization × Location	2			
Category × Urbanization × Location	8			7.85

*Significant at $P < .05$.

The tests for the remaining effects are produced similarly.

Table 4.6 provides a convenient display of the results of the various tests. Given that their respective null hypotheses are true, each of the SS in Table 4.6 has a limiting central chi-square distribution with the indicated number of degrees of freedom. These results provide some evidence ($P < .06$) of Category × Location interaction and strong evidence of a location effect ($P < .0001$) and an urbanization effect ($P < .003$).

If we ignore the difficulties arising from the possible Category × Location interaction, we can examine closely the relationships among the categories. Our primary interest is to see whether the categories are equally spaced, which we can ascertain by testing simultaneously

$$\mu_2 - \mu_1 = \mu_3 - \mu_2,$$
$$\mu_2 - \mu_1 = \mu_4 - \mu_3,$$
$$\mu_2 - \mu_1 = \mu_5 - \mu_4.$$

We perform this test in terms of $\underset{\sim}{C}\ \underset{\sim}{\xi} = \underset{\sim}{0}$ with

$$\underset{3\times 20}{\underset{\sim}{C}} = [\underset{\sim}{C}_2, \underset{\sim}{0}],$$

where

$$\underset{3\times 5}{\underset{\sim}{C}_2} = \begin{bmatrix} -1 & 2 & -1 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

and $\underset{\sim}{0}$ is the 3×15 null matrix.

The estimated contrasts for this test are

$$\hat{\mu}_2 - \hat{\mu}_1 - (\hat{\mu}_3 - \hat{\mu}_2) = -.0448,$$
$$\hat{\mu}_2 - \hat{\mu}_1 - (\hat{\mu}_4 - \hat{\mu}_5) = .0424,$$
$$\hat{\mu}_2 - \hat{\mu}_1 - (\hat{\mu}_5 - \hat{\mu}_4) = .1734,$$

with standard errors .1111, .1062, and .1068, respectively. The test provides a value of 5.32 for the sum of squares. If the null hypothesis is true, this statistic follows an asymptotic chi-square distribution with three degrees of freedom. Hence we have little evidence ($P < .15$) that the categories are not equally spaced on the logit scale.

We consider this approach to the analysis of data in ordered response categories to be appealing for two reasons. First, it allows the examination of category effects, which is especially useful in medical research involving the rankings 1-plus, 2-plus, etc. We can test hypotheses concerning distances between the categories, so that we can measure the ability of the subject to distinguish between, say, 1-plus and 2-plus. The interaction of category effects with other effects provides information about the stability of the category scale under various conditions or for various groups. Second, this approach requires no assumptions about the form of the underlying continuous distributions.

CHAPTER V

MULTIVARIATE TECHNIQUES

5.1. Introduction

In this chapter we present techniques analogous to some of the methods commonly used in conventional multivariate analysis. In Section 3.6 we discussed a multivariate technique for relative risk; however, we indicate here in greater detail the relationship between the methods we propose and the conventional methods for multivariate analysis. We include two examples for ordered response categories.

In multivariate analysis of variance we obtain for the i-th individual a vector of p observations which is generally denoted by

$$\underset{\sim}{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ip}].$$

The elements of $\underset{\sim}{y}_i$ usually are not independent, because they represent measurements on one subject. We organize n independent vectors in the matrix

$$\underset{n\times p}{\underset{\sim}{Y}} = \begin{bmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{y}_n \end{bmatrix}$$

and we express $\underset{\sim}{Y}$ in the multivariate general linear model

$$\underset{n\times p}{E[\underset{\sim}{Y}]} = \underset{n\times q}{\underset{\sim}{B}} \quad \underset{q\times p}{\underset{\sim}{\xi}} \ .$$

We can "string out" this model by arranging the observations in the vector

$$\underset{np\times 1}{\underset{\sim}{Y}^{*}} = \begin{bmatrix} \underset{\sim}{y}_1' \\ \underset{\sim}{y}_2' \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{y}_n' \end{bmatrix}$$

Consequently we can write the model as

$$\underset{np\times 1}{E[\underset{\sim}{Y}^{*}]} = \underset{np\times q^{*}}{\underset{\sim}{B}^{*}} \quad \underset{q^{*}\times 1}{\underset{\sim}{\xi}^{*}} \ ,$$

where $\underset{\sim}{B}^{*}$ and $\underset{\sim}{\xi}^{*}$ have been obtained by restructuring $\underset{\sim}{B}$ and $\underset{\sim}{\xi}$ so that they correspond to $\underset{\sim}{Y}^{*}$. If we let the vector of observed probabilities, $\underset{\sim}{p}_i$, play the role of $\underset{\sim}{y}_i$, then the appearance of the above model resembles the model we have been considering for categorical data.

With conventional multivariate analysis we are often interested in investigating functions of the elements of $\underset{\sim}{y}_i$, because they provide information about independence of variables and equality of mean values or their generalizations. We encounter analogous situations for data in contingency tables. Instead of a vector $\underset{\sim}{y}_i$ of observations for an individual, we have a vector $\underset{\sim}{p}_i$ of observed cell probabilities for a multinomial distribution. Rather than forming new variables within each $\underset{\sim}{y}_i$, we form functions of the cell probabilities within each $\underset{\sim}{p}_i$. We consider functions such as linear combinations of the p_{ij} or of the

$\ell n\ p_{ij}$ within the i-th multinomial. Linear functions of the p_{ij} allow cells to be pooled and arbitrary linear scores to be formed for each multinomial population. We can use linear functions of the $\ell n\ p_{ij}$ to construct measures of association which we can then examine according to the condition of the experiment.

5.2. Multivariate Scoring Method

The techniques for ordered response categories presented in Chapter IV are extended readily to include tables with more than one ordered response variable. We consider first a multivariate extension of the scoring method and describe its use in terms of the data in Table 5.1.

The data in Table 5.1 show the evaluation of patient's pain and requirements for medication after surgery for duodenal ulcer. A typical subtable from Table 5.1, showing the observed cell frequencies, is given by (5.2.1),

Operation i

Response 2 (medication)	Response 1 (pain) 1	2	3	Sum
1	n_{i11}	n_{i12}	n_{i13}	$n_{i1\cdot}$
2	n_{i21}	n_{i22}	n_{i23}	$n_{i2\cdot}$
3	n_{i31}	n_{i32}	n_{i33}	$n_{i3\cdot}$
4	n_{i41}	n_{i42}	n_{i43}	$n_{i4\cdot}$
Sum	$n_{i\cdot 1}$	$n_{i\cdot 2}$	$n_{i\cdot 3}$	$n_{i\cdot\cdot}$

, (5.2.1)

TABLE 5.1

PATIENTS CLASSIFIED ACCORDING TO ULCER PAIN, REQUIREMENTS FOR MEDICATION, AND TYPE OF OPERATION

	Pain			
Operation VD				
Medication	None	Slight	Moderate	Total
None	170	18	7	195
Rarely	7	5	0	12
Occasional	8	8	4	20
Regularly	0	3	14	17
Total	185	34	25	244
Operation VA				
Medication	None	Slight	Moderate	Total
None	170	22	8	200
Rarely	7	7	1	15
Occasional	5	8	8	21
Regularly	2	1	9	12
Total	184	38	26	248
Operation VH				
Medication	None	Slight	Moderate	Total
None	176	26	14	216
Rarely	8	6	3	17
Occasional	5	5	2	12
Regularly	2	5	9	16
Total	191	24	28	261
Operation RA				
Medication	None	Slight	Moderate	Total
None	181	17	10	208
Rarely	6	12	2	20
Occasional	6	7	3	16
Regularly	2	3	11	16
Total	195	39	26	260

where the corresponding P_{ijk} are such that $\Sigma_{j,k} P_{ijk} = 1$ for all i. This subtable contains two ordered response variables for which we need to obtain scores.

Let z_{1ik}, k = 1,2,3, denote the scores for the three categories of response 1 for operation i and let z_{2ij}, j = 1,2,3,4, denote the scores for the four categories of response 2 for operation i. We now consider the p_{ijk} under two scoring systems simultaneously. To do this we write the expected cell probabilities for operation i in the vector

$$\underset{1\times 12}{\underset{\sim}{P}_i'} = [P_{i11}, P_{i12}, P_{i13}, \ldots, P_{i41}, P_{i42}, P_{i43}], \tag{5.2.2}$$

and we let

$$\underset{1\times 12}{\underset{\sim}{z}_{1i}'} = [z_{1i1}, z_{1i2}, z_{1i3}, \ldots, z_{1i1}, z_{1i2}, z_{1i3}], \tag{5.2.3}$$

$$\underset{1\times 12}{\underset{\sim}{z}_{2i}'} = [z_{2i1}, z_{2i1}, z_{2i1}, \ldots, z_{2i4}, z_{2i4}, z_{2i4}]. \tag{5.2.4}$$

Then $\underset{\sim}{z}'_{1i} \underset{\sim}{P}_i$ is a mean score for operation i. This is equivalent to applying the scores to the marginal probabilities for response 1. Similarly, $\underset{\sim}{z}_{2i}' \underset{\sim}{P}_i$ is a mean score for operation i obtained from the marginals of response 2.

If we let

$$\underset{1\times 48}{\underset{\sim}{P}'} = [\underset{\sim}{P}_1', \underset{\sim}{P}_2', \underset{\sim}{P}_3', \underset{\sim}{P}_4'], \tag{5.2.5}$$

$$\underset{8\times 48}{\underset{\sim}{Z}} = \begin{bmatrix} \underset{\sim}{z}'_{11} & 0 & 0 & 0 \\ \underset{\sim}{z}'_{21} & 0 & 0 & 0 \\ 0 & \underset{\sim}{z}'_{12} & 0 & 0 \\ 0 & \underset{\sim}{z}'_{22} & 0 & 0 \\ 0 & 0 & \underset{\sim}{z}'_{13} & 0 \\ 0 & 0 & \underset{\sim}{z}'_{23} & 0 \\ 0 & 0 & 0 & \underset{\sim}{z}'_{14} \\ 0 & 0 & 0 & \underset{\sim}{z}'_{24} \end{bmatrix}, \qquad (5.2.6)$$

then $\underset{\sim}{Z}\,\underset{\sim}{P}$ represents the mean scores for Table 5.1, with two scores for each operation. We use percentile scores for elements of $\underset{\sim}{z}_{1i}$ and $\underset{\sim}{z}_{2i}$ because of their analogy to ranking, which we discussed in Section 4.3. We calculate one set of percentile scores from the marginal totals, over the entire table, for response 1, and another set for response 2.

The observed marginal distribution for response 1, computed for the complete table is given by

$$p_{\cdot\cdot k} = \sum_i n_{i\cdot k}/n_{\cdot\cdot\cdot}, \qquad k = 1,2,3,$$

where $n_{\cdot\cdot\cdot} = \Sigma_i\, n_{i\cdot\cdot}$ is the observed total for the complete table. Similarly the observed marginal distribution for response 2 is given by

$$p_{\cdot j\cdot} = \sum_i n_{ij\cdot}/n_{\cdot\cdot\cdot}, \qquad j = 1,2,3,4.$$

The percentile scores for response 1 are the same for all four operations and they can be written $z_{1ik} = z_{1k}$ where

$$z_{11} = p_{\cdot\cdot 1}/2,$$
$$z_{12} = p_{\cdot\cdot 1} + p_{\cdot\cdot 2}/2,$$
$$z_{13} = p_{\cdot\cdot 1} + p_{\cdot\cdot 2} + p_{\cdot\cdot 3}/2.$$

Similarly, the scores for response 2 are

$$z_{21} = p_{\cdot 1 \cdot}/2,$$
$$z_{22} = p_{\cdot 1 \cdot} + p_{\cdot 2 \cdot}/2,$$
$$z_{23} = p_{\cdot 1 \cdot} + p_{\cdot 2 \cdot} + p_{\cdot 3 \cdot}/2,$$
$$z_{24} = p_{\cdot 1 \cdot} + p_{\cdot 2 \cdot} + p_{\cdot 3 \cdot} + p_{\cdot 4 \cdot}/2.$$

The values for these scores, rounded to the nearest whole per cent, are

$$z_{11} = 37,$$
$$z_{12} = 82,$$
$$z_{13} = 95,$$

and

$$z_{21} = 40,$$
$$z_{22} = 84,$$
$$z_{23} = 91,$$
$$z_{24} = 97.$$

The vectors $\underset{\sim}{z}_1$ and $\underset{\sim}{z}_2$, corresponding to (5.2.3) and (5.2.4) respectively, are

$$\underset{\sim}{z}_1' = [37, 82, 95, 37, 82, 95, 37, 82, 95, 37, 82, 95]$$

and

$$\underset{\sim}{z}_2' = [40, 40, 40, 84, 84, 84, 91, 91, 91, 97, 97, 97].$$

The multiplication $\underset{\sim}{Z}\ \underset{\sim}{p}$, with $\underset{\sim}{Z}$ as shown in (5.2.6), provides the mean scores given in Table 5.2. We calculate the estimated large sample covariance matrix for the scores from the form $\underset{\sim}{S} = \underset{\sim}{H}\ \underset{\sim}{V}(\underset{\sim}{p})\ \underset{\sim}{H}'$. The variances, covariances, and correlation coefficients for the mean

TABLE 5.2

MEAN SCORES FOR RESPONSE 1 AND RESPONSE 2
FOR THE DATA IN TABLE 5.1

	Operation			
	VD	VA	VH	RA
Response 1	49.25	49.98	50.46	49.55
Response 2	50.40	49.74	48.71	50.03

scores appear in Table 5.3 for each operation. We calculate the correlation coefficient according to the standard formula and give it below the diagonal.

TABLE 5.3

VARIANCES, COVARIANCES, AND CORRELATION COEFFICIENTS FOR
THE MEAN SCORES GIVEN IN TABLE 5.2

Operation							
VD		VA		VH		RA	
1.959	1.063	1.944	.966	1.937	.718	1.856	.964
.57	1.762	.55	1.612	.43	1.414	.57	1.570

We now proceed with the analysis. Since there are only four types of operations and no other classification of the design variables, we can test all the hypotheses of interest simply by contrasting the mean scores for the different types of operations. However, we choose to follow a course more in keeping with a general method and we choose a model

$$E \begin{bmatrix} \underset{\sim}{z}_1' \underset{\sim}{p}_i \\ \underset{\sim}{z}_2' \underset{\sim}{p}_i \end{bmatrix} = \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix} ,$$

which provides estimates of only the mean values for each operation for each response. In a sense, such a model is degenerate since the best estimates of μ_{1i} and μ_{2i} are simply $\underset{\sim}{z}_1' \underset{\sim}{p}_i$ and $\underset{\sim}{z}_2' \underset{\sim}{p}_i$. We express this model in matrix form as

$$E \begin{bmatrix} \underset{\sim}{z}_1' \underset{\sim}{p}_1 \\ \underset{\sim}{z}_2' \underset{\sim}{p}_1 \\ \underset{\sim}{z}_1' \underset{\sim}{p}_2 \\ \underset{\sim}{z}_2' \underset{\sim}{p}_2 \\ \underset{\sim}{z}_1' \underset{\sim}{p}_3 \\ \underset{\sim}{z}_2' \underset{\sim}{p}_3 \\ \underset{\sim}{z}_1' \underset{\sim}{p}_4 \\ \underset{\sim}{z}_2' \underset{\sim}{p}_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \\ \mu_{41} \\ \mu_{42} \end{bmatrix},$$

which yields the $\underset{\sim}{z}_1' \underset{\sim}{p}_i$ and $\underset{\sim}{z}_2' \underset{\sim}{p}_i$ as best estimates of the μ_{1i} and μ_{2i}. The residual sum of squares for this model is zero, but we can still test contrasts among the μ_{1i} and μ_{2i}.

If we choose

$$\underset{\sim}{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \end{bmatrix}$$

in the general form $\underset{\sim}{C}\,\underset{\sim}{\xi} = \underset{\sim}{0}$, where

$$\underset{\sim}{\xi}' = [\mu_{11}\ \mu_{12}\ \mu_{21}\ \mu_{22}\ \mu_{31}\ \mu_{32}\ \mu_{41}\ \mu_{42}],$$

We produce

$$\mu_{11} - \mu_{41} = 0,$$
$$\mu_{21} - \mu_{41} = 0,$$
$$\mu_{31} - \mu_{41} = 0,$$

which can be true if and only if

$$\mu_{11} = \mu_{21} = \mu_{31} = \mu_{41}.$$

The resulting test statistic is $X^2 = .43$, referred to a chi-square distribution with 3 D.F. Thus we conclude that the operations are homogeneous in their effect on pain. Choosing

$$\underset{\sim}{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

results in a test of $\mu_{12} = \mu_{22} = \mu_{32} = \mu_{42}$, which yields $X^2 = 1.04$ for a chi-square variable with 3 D.F., which implies homogeneity with respect to requirement for medication. We can test both hypotheses of homogeneity simultaneously by choosing

$$\underset{\sim}{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix},$$

which yields $X^2 = 2.79$ with 6 D.F.

Thus we conclude that the operations do not differ with respect to pain or to requirements for medication, but that pain and requirement for medication are correlated. The estimate of the correlation obtained from the pooled variance-covariance matrix is "$\hat{\rho}$" = .53.

The extension to any number of dimensions and to any design is obvious. Analogous to multivariate analysis, we can form new variables from the original multinomially distributed variables, fit models, test hypotheses, and estimate the correlations among the variables produced.

5.3. Multivariate Cumulative Logits

In Section 4.3 we analyzed data classified into ordered response categories by expressing the cumulative logits for the ordered response variable in terms of a category effect and a treatment effect. We now extend this procedure so as to be applicable to data with two ordered response variables. We consider again the data in Table 5.1 and the subtable given in (5.2.1).

For (5.2.1) we construct a set of cumulative logits for response 1 and a set for response 2. We express each logit of each set in terms of a category effect and an operation effect. The vector of expected cell probabilities for operation i appears in (5.2.2). The cumulative probabilities for response 1, operation i, are

$$\phi_{i11} = P_{i11} + P_{i21} + P_{i31} + P_{i41},$$

$$\phi_{i12} = \phi_{i11} + P_{i12} + P_{i22} + P_{i23} + P_{i24},$$

$$\phi_{i13} = 1.$$

For response 2, operation i, we get

$$\phi_{i21} = P_{i11} + P_{i12} + P_{i13},$$

$$\phi_{i22} = \phi_{i21} + {}_{i21} + P_{i22} + P_{i23},$$

$$\phi_{i23} = \phi_{i22} + P_{i31} + P_{i32} + P_{i33},$$

$$\phi_{i24} = 1.$$

We consider the cumulative logits for response 1,

$$\ell_{i11} = \ell n\, [\phi_{i11}/(1 - \phi_{i11})],$$

and

$$\ell_{i12} = \ell n\, [\phi_{i12}/(1 - \phi_{i12})],$$

and for response 2,

$$\ell_{i21} = \ell n\, [\phi_{i21}/(1 - \phi_{i21})],$$

$$\ell_{i22} = \ell n\, [\phi_{i22}/(1 - \phi_{i22})],$$

and

$$\ell_{i23} = \ell n\, [\phi_{i23}/(1 - \phi_{i23})],$$

in the model

$$\begin{bmatrix} \ell_{i11} \\ \ell_{i12} \\ \ell_{i21} \\ \ell_{i22} \\ \ell_{i23} \end{bmatrix} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} + \begin{bmatrix} \alpha^*_{i11} \\ \alpha^*_{i12} \\ \alpha^*_{i21} \\ \alpha^*_{i22} \\ \alpha^*_{i23} \end{bmatrix} , \quad i = 1,2,3,4 \tag{5.3.1}$$

where

$\mu_{k\ell}$ = location constant for the $k\ell$-th logit,

$\alpha^*_{ik\ell}$ = effect of operation i for $k\ell$-th logit.

We can write the logits in terms of $\underset{\sim}{K}_1\ \ell n\ \underset{\sim}{A}_1\ \underset{\sim}{P}_i$, where

$$\underset{5\times10}{\underset{\sim}{K}_1} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} ,$$

$$\underset{10\times 12}{\underset{\sim}{A}_1} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix},$$

and $\underset{\sim}{P}_i$ is given in (5.2.2). If we let

$$\underset{1\times 5}{\underset{\sim}{\ell}_i'} = [\ell_{i11}, \ell_{i12}, \ell_{i21}, \ell_{i22}, \ell_{i23}],$$

$$\underset{1\times 20}{\underset{\sim}{\ell}'} = [\ell_1', \ell_2', \ell_3', \ell_4'],$$

$$\underset{20\times 40}{\underset{\sim}{K}} = \underset{4\times 4}{\underset{\sim}{I}} \otimes \underset{5\times 10}{\underset{\sim}{K}_1},$$

$$\underset{40\times 48}{\underset{\sim}{A}} = \underset{4\times 4}{\underset{\sim}{I}} \otimes \underset{10\times 12}{\underset{\sim}{A}_1},$$

then $\underset{20\times 1}{\underset{\sim}{\ell}} = \underset{20\times 40}{\underset{\sim}{K}} \; \ell n \; \underset{40\times 48}{\underset{\sim}{A}} \; \underset{48\times 1}{\underset{\sim}{P}}$.

Consequently, we can express (5.3.1) as

$$\underset{20\times 1}{\underset{\sim}{\ell}} = \underset{20\times 20}{\underset{\sim}{X}} \; \underset{20\times 1}{\underset{\sim}{\xi}}, \tag{5.3.2}$$

incorporating the restrictions $\Sigma_i \hat{\alpha}_{ik\ell} = 0$ for all k,ℓ, so that

$$\underset{20\times 20}{\underset{\sim}{X}} = \begin{bmatrix} \underset{\sim}{I}_5 & \underset{\sim}{I}_5 & \underset{\sim}{0}_5 & \underset{\sim}{0}_5 \\ \underset{\sim}{I}_5 & \underset{\sim}{0}_5 & \underset{\sim}{I}_5 & \underset{\sim}{0}_5 \\ \underset{\sim}{I}_5 & \underset{\sim}{0}_5 & \underset{\sim}{0}_5 & \underset{\sim}{I}_5 \\ \underset{\sim}{I}_5 & -\underset{\sim}{I}_5 & -\underset{\sim}{I}_5 & -\underset{\sim}{I}_5 \end{bmatrix},$$

with $\underset{\sim}{I}_5$, $\underset{\sim}{0}_5$, and $-\underset{\sim}{I}_5$ denoting the 5×5 identity matrix, zero matrix, and negative identity matrix, respectively. The parameters are given in

$$\underset{\sim}{\xi}' = [\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \mu_{23}, \alpha_{111}, \alpha_{112}, \alpha_{121}, \alpha_{122}, \alpha_{123},$$
$$\alpha_{211}, \alpha_{212}, \alpha_{221}, \alpha_{222}, \alpha_{223}, \alpha_{311}, \alpha_{312}, \alpha_{321}, \alpha_{322}, \alpha_{323}],$$

where

$$\begin{aligned} \alpha^*_{1k\ell} &= \alpha_{1k\ell}, \\ \alpha^*_{2k\ell} &= \alpha_{2k\ell}, \\ \alpha^*_{3k\ell} &= \alpha_{3k\ell}, \\ \alpha^*_{4k\ell} &= -\alpha_{1k\ell} - \alpha_{2k\ell} - \alpha_{3k\ell}. \end{aligned}$$

We have used model (5.3.2) to obtain the estimates of parameters and their standard errors given in Table 5.4. We can now examine relationships between the response variables and the operations. We examine the hypotheses that there are no differences among the operations with respect to response 1 by testing $\underset{\sim}{C}_1 \underset{\sim}{\xi} = \underset{\sim}{0}$ where

$$\underset{\sim}{C}_1 = [\underset{6\times 5}{\underset{\sim}{0}} \quad \underset{6\times 5}{\underset{\sim}{C}_2} \quad \underset{6\times 5}{\underset{\sim}{C}_3} \quad \underset{6\times 5}{\underset{\sim}{C}_4}],$$

with

$$\underset{6\times 5}{\underset{\sim}{C}_2} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

TABLE 5.4

ESTIMATES OF PARAMETERS AND THEIR STANDARD ERRORS FOR THE MODEL (5.3.2)

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
μ_{11}	1.075	.072	α_{211}	-.019	.126
μ_{12}	2.158	.103	α_{212}	-.013	.179
μ_{21}	1.441	.080	α_{221}	-.014	.139
μ_{22}	1.920	.094	α_{222}	-.045	.162
μ_{23}	2.756	.133	α_{223}	.223	.248
μ_{111}	.067	.128	α_{311}	-.715	.387
μ_{112}	.012	.181	α_{312}	-.039	.175
μ_{121}	-.599	.438	α_{321}	.128	.141
μ_{122}	-.198	.158	α_{322}	.199	.170
μ_{123}	-.165	.222	α_{323}	-.027	.226

$$\underset{6\times 5}{\underset{\sim}{C}_3} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\underset{\sim}{C}_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The value of the SS for testing this hypothesis is .61, which is referred to a chi-square distribution with 6 degrees of freedom. Thus we find no evidence of differences among the four operations with regard to response 1.

We test for no differences among operations with respect to response 2 by using $\underset{\sim}{C}_5 \underset{\sim}{\xi} = \underset{\sim}{0}$, where

$$C_5 = [\underset{9\times5}{\underset{\sim}{0}} \quad \underset{9\times5}{\underset{\sim}{C}_6} \quad \underset{9\times5}{\underset{\sim}{C}_7} \quad \underset{9\times5}{\underset{\sim}{C}_8}],$$

with

$$\underset{9\times5}{\underset{\sim}{C}_6} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\underset{9\times5}{\underset{\sim}{C}_7} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\underset{9\times 5}{\underset{\sim}{C}_8} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

The value of the SS for testing $\underset{\sim}{C}_5 \, \underset{\sim}{\xi} = \underset{\sim}{0}$ is 6.66, which is referred to a chi-square distribution with 9 degrees of freedom. As for response 1, we observe no evidence of differences among the four operations with regard to response 2.

The estimates of μ_{11} and μ_{12}, and of μ_{21}, μ_{22}, and μ_{23} are scale constants for the three categories of response 1 and for the four categories of response 2. Thus we can obtain information about the relationship between these two sets of categories by examining the variance-covariance matrix for the estimates of the five $\mu_{k\ell}$'s. This information appears in Table 5.5, with the variances shown on the diagonal, the covariances above the diagonal, and the correlation coefficients below.

We are also interested in relationships among the ordered categories for response 1 and for response 2. For example, if $\mu_{11} = \mu_{12}$, we might better consider the three categories of response 1 to be only two categories.

TABLE 5.5

ESTIMATES OF THE VARIANCES, COVARIANCES, AND CORRELATION COEFFICIENTS FOR THE ESTIMATES OF THE CATEGORY EFFECTS FROM TABLE 5.1

		μ_{11}	μ_{12}	μ_{21}	μ_{22}	μ_{23}
Response 1	μ_{11}	.0052	.0043	.0029	.0031	.0036
	μ_{12}	.58	.0106	.0031	.0044	.0069
Response 2	μ_{21}	.50	.38	.0064	.0059	.0055
	μ_{22}	.45	.45	.78	.0089	.0083
	μ_{23}	.38	.50	.52	.66	.0177

We test $\mu_{11} = \mu_{12}$ in the form $\underset{\sim}{C}\,\underset{\sim}{\xi} = \underset{\sim}{0}$, with

$$\underset{1\times 20}{\underset{\sim}{C}} = [1, -1, \underset{1\times 18}{\underset{\sim}{0}}\ ] .$$

The value of the statistic for testing this hypothesis is 163.22, which is referred to a chi-square distribution with one degree of freedom. Consequently, we conclude that there are three distinct categories for response 1.

For response 2, if the distances between the successive categories are equal, then

$$\mu_{22} - \mu_{21} = \mu_{23} - \mu_{22}$$

or

$$-\mu_{21} + 2\mu_{22} - \mu_{23} = 0.$$

We then test the hypothesis that the distances between the successive categories are equal by testing $\underset{\sim}{C}\,\underset{\sim}{\xi} = \underset{\sim}{0}$ with

$$\underset{1\times 20}{\underset{\sim}{C}} = [0,\ 0,\ -1,\ 2,\ -1,\ \underset{1\times 15}{\underset{\sim}{0}}\ ].$$

The value of the statistic for testing this hypothesis is 9.15, which is referred to chi-square with one degree of freedom. We conclude, therefore, that distances between the categories, are not equal.

The technique presented in this section can be applied to data with more than two ordered response variables and to tables representing more complex experimental designs merely by properly constructing cumulative logits and by using an appropriate design matrix.

CHAPTER VI

SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

This work has used the Grizzle, Starmer, and Koch [1969] method for the analysis of contingency tables as a basis for methods of analyzing relative risk as well as data with one or more ordered response variables. In addition, relationships between linear contrasts in $\ell n\ P_{ijk}$, tests of marginal independence, and analysis of incomplete tables have been classified.

The methods presented are for minimum χ_1^2 estimates and tests are based on asymptotic theory. Consequently, one of the remaining problems requiring consideration is a documentation of the small sample properties of these procedures. In the process of investigating small sample properties, one might compare the methods of estimating relative risk discussed in Chapter III with the discriminant function approach discussed by Cornfield [1967].

Investigation concerning the distribution of the correlation coefficient[1] calculated in Chapter V would be of value, since knowledge of this distribution would allow tests for zero correlation among estimates of parameters derived from multinomial populations when the

---

[1]M. Hills. 1969. On looking at large correlation matrices, Biometrika 56:249-53, discussed the transformation $z = \ell n[(1+\hat{\rho})/(1-\hat{\rho})]$, where, if $\rho = 0$, z follows approximately a normal distribution with mean zero and standard deviation approximately $1/\sqrt{45}$; if his conjecture that the distribution of z is little affected by the dependence inherent in multinomial data is true, this problem may be solved readily.

response variable has a natural order.

Also, it should be worth investigating whether it is worthwhile to expand the Forthofer, Starmer, and Koch [1969] computer program to calculate maximum likelihood estimates.

BIBLIOGRAPHY

Bahr, H. M. 1969. Institutional life, drinking, and disaffiliation. Social Problems 16: 365-75.

Bartlett, M. S. 1935. Contingency table interactions. Journal of the Royal Statistical Society Supplement 2: 248-52.

Berkson, J. 1944. Application of the logistic function to bio-assay. Journal of the American Statistical Association 39: 357-65.

Berkson, J. 1946. Approximation of chi-square by "probits" and by "logits." Journal of the American Statistical Association 41: 70-4.

Berkson, J. 1955. Maximum likelihood and minimum χ^2 estimates of the logistic function. Journal of the American Statistical Association 50: 130-62.

Berkson, J. 1968. Application of minimum logit χ^2 estimate to a problem of Grizzle with a notation on the problem of no interaction. Biometrics 24: 75-95.

Bhapkar, V. P. 1961. Some tests for categorical data. Annals of Mathematical Statistics 32: 72-83.

Bhapkar, V. P. 1966. A note on the equivalence of two test criteria for hypotheses in categorical data. Journal of the American Statistical Association 61: 228-35.

Bhapkar, V. P., and Koch, G. G. 1968a. On the hypothesis of 'no interaction' in contingency tables. Biometrics 24: 567-94.

Bhapkar, V. P., and Koch, G. G. 1968b. Hypotheses of no interaction in multidimensional contingency tables. Technometrics 10: 107-22.

Birch, M. W. 1963. Maximum likelihood in three-way contingency tables. Journal of the Royal Statistical Society Series B 25: 220-33.

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

Bishop, Y. M., and Fienberg, S. E. 1969. Incomplete two-dimensional contingency tables. Biometrics 25: 119-28.

Bock, R. D., and Jones, L. V. 1967. The Measurement and Prediction of Judgment and Choice. San Francisco: Holden-Day.

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

Cornfield, J. 1951. A method of estimating comparative rates from clinical data, application to cancer of the lung, breast, and cervix. Journal of the National Cancer Institute 11: 1269-1275.

Cornfield, J. 1956. A statistical problem arising from retrospective studies. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4. Berkeley and Los Angeles: University of California Press (pp. 135-48).

Cornfield, J. 1967. Discriminant functions. Review of the International Statistical Institute 35: 142-53.

Darroch, J. N. 1962. Interaction in multi-factor contingency tables. Journal of the Royal Statistical Society Series B 24: 251-63.

Elston, R. C., and Bush, N. 1964. The hypotheses that can be tested when there are interactions in an analysis of variance model. Biometrics 20: 681-98.

Fienberg, S. E., and Gilbert, J. P. 1970. Geometry of a 2×2 contingency table. Journal of the American Statistical Association 65: 694-701.

Forthofer, R. N., Starmer, C. F., and Grizzle, J. E. 1969. A program for the analysis of categorical data by linear models. University of North Carolina Institute of Statistics Mimeo Series No. 604.

Framingham Study Group. 1968. The Framingham Study (Section 9, table 9-A-14, exam 1).

Garrett, H. E. 1953. Statistics in Psychology and Education, Fourth Edition. New York: Longmans, Green and Company.

Gart, J. J. 1962. On the combination of relative risk. Biometrics 18: 601-10.

Ghosh, M. 1969. Asymptotically optimal nonparametric tests for miscellaneous problems of linear regression. University of North Carolina Institute of Statistics Mimeo Series No. 634.

Goodman, L. A. 1963a. On Plackett's test for contingency table interactions. Journal of the Royal Statistical Society Series B 25: 179-88.

Goodman, L. A. 1963b. On methods for comparing contingency tables. Journal of the Royal Statistical Society Series A 126: 94-108.

Goodman, L. A. 1968. The analysis of cross-classified data; independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association 63: 1091-131.

Goodman, L. A., and Kruskal, W. H. 1954. Measures of association for cross classifications. Journal of the American Statistical Association 49: 732-64.

Goodman, L. A., and Kruskal, W. H. 1959. Measures of association for cross classification. II: Further discussion and references. Journal of the American Statistical Association 58: 310-64.

Grizzle, J. E. 1961. A new method of testing hypotheses and estimating parameters for the logistic model. Biometrics 17: 372-85.

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

Harris, J. E., and Tu, C. 1929. A second category of limitations in the applicability of the contingency coefficient. Journal of the American Statistical Association 24: 367-75.

Hitchcock, S. E. 1962. A note on the estimation of the parameters of the logistic function using the minimum logit χ^2 method. Biometrika 49: 250-2.

Irwin, J. O. 1949. A note on the subdivision of χ^2 into components. Biometrika 36: 103-4.

Kastenbaum, M. A., and Lamphier, D. E. 1959. Calculation of chi-square to test the no three-factor interaction hypothesis. Biometrika 15: 107-15.

Lancaster, H. O. 1951. Complex contingency tables treated by the partition of χ^2. Journal of the Royal Statistical Society Series B 18: 242-9.

Mantel, N. 1966. Models for complex contingency tables and polychotomous dosage response curves. Biometrics 22: 83-95.

Mosteller, F. 1968. Association and estimation in contingency tables. Journal of the American Statistical Association 63: 1-28.

Neyman, J. 1949. Contributions to the theory of the χ^2 test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles: University of California Press (pp. 239-72).

Norton, H. W. 1945. Calculation of chi-square for complex contingency tables. Journal of the American Statistical Association 40: 251-8.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophy Magazine Series 5 50: 157-75.

Placket, R. L. 1962. A note on interactions in contingency tables. Journal of the Royal Statistical Society Series B 24: 162-6.

Roy, S. N., and Kastenbaum, M. A. 1956. On the hypothesis of no interaction in a multiway contingency table. The Annals of Mathematical Statistics 27: 749-57.

Sen, P. K. 1967. Asymptotically most powerful rank order tests for grouped data. The Annals of Mathematical Statistics 38: 1229-39.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society Series B 13: 238-41.

Strong, J. P., Solberg, L. A., and Restrepo, C. 1968. Atherosclerosis in persons with coronary heart disease. Laboratory Investigation 18: 527-38.

Thurstone, L. L. 1959. The Measurement of Values. Chicago: University of Chicago Press.

Wald, A. 1943. Test of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54: 426-82.

Watson, G. S. 1956. Missing and "mixed-up" frequencies in contingency tables. Biometrics 12: 47-50.

Woolf, B. 1955. On estimating the relation between blood group and disease. Annals of Human Genetics 19: 251-53.