INFORMATION ANALYSIS OF LINEAR INTERACTIONS IN CONTINGENCY TABLES

BY

S. KULLBACK and D. V. GOKHALE

TECHNICAL REPORT NO. 9
AUGUST 15, 1977

PREPARED UNDER GRANT
DAAG29-77-G-0031
FOR THE U.S. ARMY RESEARCH OFFICE

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Information Analysis of Linear Interactions In Contingency Tables

By

S. Kullback and D.V. Gokhale

TECHNICAL REPORT NO. 9

August 15, 1977

Prepared under Grant DAAG29-77-G-0031

For the U.S. Army Research Office

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Partially supported under Office of Naval Research Contract N00014-76-C-0475 (NR-042-267) and issued as Technical Report No. 249.

THE FINDINGS IN THIS REPORT ARE NOT TO BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, UNLESS SO DESIGNATED BY OTHER AUTHORIZED DOCUMENTS.

INFORMATION ANALYSIS OF LINEAR INTERACTIONS IN CONTINGENCY TABLES

S. KULLBACK
The George Washington University

D.V. GOKHALE
University of California, Riverside

1. INTRODUCTION

The purpose of this article is to illustrate the use of the minimum discrimination information (MDI) approach in studying null hypotheses of no linear interactions in contingency tables of "one response many factors" type. In such contingency tables, the data can be looked upon as a collection of as many multinomial experiments as there are factor-level combinations and each experiment has a number of cells equal to the levels of a response variable. One formulation of a "no linear interaction" hypothesis

is that the cell probabilities of the response variable can be expressed as linear functions of parameters which are structurally less complex. For accounts of different formulations of no-interaction hypotheses and related references, the reader is referred to Bhapkar and Koch [1968] , Darroch [1974].

The "no linear interaction" hypotheses can be formulated as linear constraints on the underlying probabilities, written in

matrix notation as $\underline{Bp} = \underline{0}$. It is possible to apply MDI analysis to obtain estimates of cell frequencies and test various hypotheses and sub-hypotheses. If the hypotheses are "nested" the MDI statistic for the stronger hypothesis (which imposes more constraints) can be analyzed into two components, one measuring disparity between the observed distribution and the weaker hypothesis and the other measuring disparity between the estimated distributions under the two hypotheses. This feature of the MDI statistics is not enjoyed by the chi-square-type or Wald-type statistics used by many authors.

For the sake of clarity of presentation, we will restrict ourselves to the hypotheses of no linear second order interaction in a 2x2x2 table and in a 4x2x2 table. This enables us to compare results with Bhapkar and Koch [1968], who have viewed two sets of data as of the "one response many factors" type. The analysis of a 2x2x2 table shows how the use of an approximation in the MDI statistic leads to a statistic used by Bhapkar and Koch [1968]. The 4x2x2 table is analysed under two hypotheses of no linear interaction of second order, illustrating the analysis of information mentioned in the preceding paragraph.

2. GENERALITIES

For a three-way rxsxt table in which the first variable is a response and the other two variables are factors, one

formulation of no linear second order interaction is given by

$$(2.1)\quad H_o : p(ijk) = \mu(i..) + \mu(ij.) + \mu(i.k),\ i=1,\ldots,r,\ j=1,\ldots,s,\ k=1,\ldots,t,$$

where the p(ijk) are subject to the constraints

$$(2.2)\quad \Sigma_{i=1}^{r} p(ijk) = 1, \text{ for each fixed pair } (jk),$$

and the parameters μ depend only on the indicated indices. The hypothesis H_o is equivalent to the following (r-1)x(s-1)x(t-1) constraints in addition to those in (2.2):

$$(2.3)\quad p(ijk) - p(ijt) - p(isk) + p(ist) = 0,\quad i=1,\ldots,(r-1),\ j=1,\ldots,(s-1),\ k=1,\ldots,(t-1).$$

Writing

$$(2.4)\quad \underline{p} = (p(111), p(211), \ldots, p(r11), p(112), \ldots, p(rst))',$$

where the (jk) indices are in lexicographic order, the constraints (2.2) and (2.3) can be expressed as

$$(2.5)\quad \underline{Bp} = \underline{0}$$

where the vector $\underline{0}$ consists of the first st elements equal to unity and the remaining elements equal to zero. This is illustrated in the examples below.

Let x(ijk) denote the observed frequency in the (ijk)-th cell and $\underline{x}$ denote a vector similar to $\underline{p}$ of (2.4). Also let x(.jk) denote the total number of observations for the (jk)-th

factor combination and let $N = \Sigma_{j=1}^{s}\Sigma_{k=1}^{t}x(.jk)$. Basic to the information analysis is the discrimination information function

$$(2.6) \quad I(p:\pi) = \Sigma_{j=1}^{s}\Sigma_{k=1}^{t}w(jk)\Sigma_{i=1}^{r}p(ijk)\ln[p(ijk)/\pi(ijk)]$$

where $w(jk) = x(.jk)/N$. The vector $\underline{\pi}$ is similar to $\underline{p}$; it is an arbitrary collection of st probability distributions, each on r cells. It is assumed that $x(ijk)$, $p(ijk)$ and $\pi(ijk)$ are positive for all (ijk). The choice of $\underline{\pi}$ depends on the analysis at hand. When it is desired to assess the departure of the data from an external hypothesis (as is the present case), $\underline{\pi}$ is taken to be the vector of observed proportions in each of the st factor combinations.

The MDI estimates $x^*(ijk) = Np^*(ijk)$ are such that the discrimination information (2.6) is minimized subject to the constraints $\underline{Bp} = \underline{\theta}$. In other words, $p^*(ijk)$ is the distribution which satisfies the hypothesized constraints (2.5) and is "closest" (in the MDI sense) to the observed distribution. There are several convergent iterative computer algorithms for obtaining $x^*(ijk)$. One is described in the Appendix.

The MDI statistic $2I(x^*:x) = 2NI(p^*:\pi)$ has a chi-square distribution in large samples with degrees of freedom equal to rank(B) - st.

3. THE 2x2x2 TABLE

Consider the probabilities of a 2 x 2 x 2 contingency table (Table 1).

Table 1

	B j=1		β j=2	
	C k=1	γ k=2	C k=1	γ k=2
i=1 A	p(111)	p(112)	p(121)	p(122)
i=2 α	p(211)	p(212)	p(221)	p(222)

The experimental procedure selects a fixed number of observations under the four possible combinations of the factors (B,β), (C,γ) and determines the number of occurrences of (A,α) for each case. In effect then the procedure is examining four binomials with

(3.1) $p(1jk) + p(2jk) = 1,\ j=1,2, k = 1,2.$

The corresponding observed values are shown in table 2, It is desired to test whether the observed values are consistent with a null hypothesis of no interaction on a linear scale,

Table 2

	j=1		j=2	
	k=1	k=2	k=1	k=2
i=1	x(111)	x(112)	x(121)	x(122)
i=2	x(211)	x(212)	x(221)	x(222)
	x(.11)	x(.12)	x(.21)	x(.22)

that is

$$H_0: \quad p(111) - p(112) = p(121) - p(122)$$

$$(3.2) \qquad \text{or } p(111) - p(112) - p(121) + p(122) = 0.$$

We shall determine estimates for the cell entries subject to the null hypothesis and compare the estimated and observed values. The estimated table is given in table 3 where the λ's are to be determined.

Table 3

	j=1		j=2	
	k=1	k=2	k=1	k=2
i=1	$x(111)+\lambda_1$	$x(112)+\lambda_2$	$x(121)+\lambda_3$	$x(122)+\lambda_4$
i=2	$x(211)-\lambda_1$	$x(212)-\lambda_2$	$x(221)-\lambda_3$	$x(222)-\lambda_4$
	x(.11)	x(.12)	x(.21)	x(.22)

We shall use the principle of minimum discrimination information estimation and thus determine the λ's which minimize

$$(3.3)\quad \begin{cases} (x(111)+\lambda_1)\,\ell n\dfrac{x(111)+\lambda_1}{x(111)} + (x(211)-\lambda_1)\,\ell n\dfrac{x(211)-\lambda_1}{x(211)} \\ +(x(112)+\lambda_2)\,\ell n\dfrac{x(112)+\lambda_2}{x(112)} + (x(212)-\lambda_2)\,\ell n\dfrac{x(212)-\lambda_2}{x(212)} \\ +(x(121)+\lambda_3)\,\ell n\dfrac{x(121)+\lambda_3}{x(121)} + (x(221)-\lambda_3)\,\ell n\dfrac{x(221)-\lambda_3}{x(221)} \\ +(x(122)+\lambda_4)\,\ell n\dfrac{x(122)+\lambda_4}{x(122)} + (x(222)-\lambda_4)\,\ell n\dfrac{x(222)-\lambda_4}{x(222)} \\ +\tau\left(\dfrac{x(111)+\lambda_1}{x(.11)} - \dfrac{x(112)+\lambda_2}{x(.12)} - \dfrac{x(121)+\lambda_3}{x(.21)} + \dfrac{x(122)+\lambda_4}{x(122)}\right), \end{cases}$$

where τ is a Lagrange undetermined multiplier and (3.2) is reflected by the condition

$$(3.4)\quad \frac{x(111)+\lambda_1}{x(.11)} - \frac{x(112)+\lambda_2}{x(.12)} - \frac{x(121)+\lambda_3}{x(.21)} + \frac{x(122)+\lambda_4}{x(.22)} = 0.$$

Differentiating (3.3) with respect to $\lambda_1,\ldots,\lambda_4$ leads to the "normal" equations

$$(3.5)\quad \begin{cases} \ell n\dfrac{x(111)+\lambda_1}{x(111)} - \ell n\dfrac{x(211)-\lambda_1}{x(211)} + \dfrac{\tau}{x(.11)} = 0, \\ \ell n\dfrac{x(112)+\lambda_2}{x(112)} - \ell n\dfrac{x(212)-\lambda_2}{x(212)} - \dfrac{\tau}{x(.12)} = 0, \\ \ell n\dfrac{x(121)+\lambda_3}{x(121)} - \ell n\dfrac{x(221)-\lambda_3}{x(221)} - \dfrac{\tau}{x(.21)} = 0, \\ \ell n\dfrac{x(122)+\lambda_4}{x(122)} - \ell n\dfrac{x(222)-\lambda_4}{x(222)} + \dfrac{\tau}{x(.22)} = 0. \end{cases}$$

There are a number of different iterative approaches to determine the solution to (3.5) but our interest here is to examine the relation of an approximate solution to other proposed methods.

Assuming that the ratios of λ's to the observed values are small, we use the approximations

$$\ell n \frac{x(111)+\lambda_1}{x(111)} \simeq \frac{\lambda_1}{x(111)}, \quad \ell n \frac{x(211)-\lambda_1}{x(211)} \simeq -\frac{\lambda_1}{x(211)}, \text{ etc.,}$$

in (3.5) and get

$$(3.6)\quad \begin{cases} \dfrac{\lambda_1}{x(111)} + \dfrac{\lambda_1}{x(211)} + \dfrac{\tau}{x(.11)} = 0 = \lambda_1 \dfrac{x(.11)}{x(111)\,x(211)} + \dfrac{\tau}{x(.11)}, \\ \dfrac{\lambda_2}{x(112)} + \dfrac{\lambda_2}{x(212)} - \dfrac{\tau}{x(.12)} = 0 = \lambda_2 \dfrac{x(.12)}{x(112)\,x(212)} - \dfrac{\tau}{x(.12)}, \\ \dfrac{\lambda_3}{x(121)} + \dfrac{\lambda_3}{x(221)} - \dfrac{\tau}{x(.21)} = 0 = \lambda_3 \dfrac{x(.21)}{x(121\;\; x(221)} - \dfrac{\tau}{x(.21)}, \\ \dfrac{\lambda_4}{x(122)} + \dfrac{\lambda_4}{x(222)} + \dfrac{\tau}{x(.22)} = 0 = \lambda_4 \dfrac{x(.22)}{x(122)\,x(222)} + \dfrac{\tau}{x(.22)}. \end{cases}$$

From (3.6) and (3.4) we have, introducing the notation

$$x(1ij) = x(.ij)\hat{p}(ij),\ x(2ij) = x(.ij)\hat{q}(ij),\ \hat{p}(ij) + \hat{q}(ij) = 1,$$

$$(3.7)\quad \begin{cases} \lambda_1 = -\dfrac{x(111)x(211)}{(x(.11))^2}\tau = -\hat{p}(11)\hat{q}(11)\tau, \\ \lambda_2 = \dfrac{x(112)x(212)}{(x(.12))^2}\tau = \hat{p}(12)\hat{q}(12)\tau, \\ \lambda_3 = \dfrac{x(121)x(221)}{(x(.21))^2}\tau = \hat{p}(21)\hat{q}(21)\tau, \\ \lambda_4 = -\dfrac{x(122)x(222)}{(x(.22))^2}\tau = -\hat{p}(22)\hat{q}(22)\tau, \\ \tau = \dfrac{\hat{p}(11) - \hat{p}(12) - \hat{p}(21) + \hat{p}(22)}{\dfrac{\hat{p}(11)\hat{q}(11)}{x(.11)} + \dfrac{\hat{p}(12)\hat{q}(12)}{x(.12)} + \dfrac{\hat{p}(21)\hat{q}(21)}{x(.21)} + \dfrac{\hat{p}(22)\hat{q}(22)}{x(.22)}}. \end{cases}$$

Let us write

$$x^*(111) = x(111) + \lambda_1,\ x^*(211) = x(211) - \lambda_1\ ,$$
$$x^*(112) = x(112) + \lambda_2,\ x^*(212) = x(212) - \lambda_2\ , \tag{3.8}$$
$$\text{etc.}$$

where the λ's satisfy (3.5).

If we also use the quadratic approximations

$$2\{(x(111)+\lambda_1)\ \ell n\ \frac{x(111)+\lambda_1}{x(111)} + (x(211)-\lambda_1)\ \ell n\ \frac{x(211)-\lambda_1}{x(211)}\} \tag{3.9}$$
$$\doteq \lambda_1^2\ (\frac{1}{x(111)} + \frac{1}{x(211)}) = \lambda_1^2\ \frac{x(.11)}{x(111)x(211)} = \frac{\lambda_1^2}{x(.11)\hat{p}(11)\hat{q}(11)}$$

then we get for the minimum discrimination information statistic

$$2I(x^*:x) = 2\Sigma\ \Sigma\ \Sigma\ \ x^*(ijk)\ \ell n\ \frac{x^*(ijk)}{x(ijk)}$$
$$\doteq \tau^2\ \{\ \frac{\hat{p}(11)\hat{q}(11)}{x(.11)} + \frac{\hat{p}(12)\hat{q}(12)}{x(.12)} + \frac{\hat{p}(21)\hat{q}(21)}{x(.21)} + \frac{\hat{p}(22)\hat{q}(22)}{x(.22)}\}$$
$$= \frac{(\hat{p}(11)-\hat{p}(12)-\hat{p}(21)+\hat{p}(22))^2}{\frac{\hat{p}(11)\hat{q}(11)}{x(.11)} + \frac{\hat{p}(12)\hat{q}(12)}{x(.12)} + \frac{\hat{p}(21)\hat{q}(21)}{x(.21)} + \frac{\hat{p}(22)\hat{q}(22)}{x(.22)}} \tag{3.10}$$
$$= \lambda_1^2(\frac{1}{x(111)} + \frac{1}{x(211)}) + \lambda_2^2(\frac{1}{x(112)} + \frac{1}{x(212)}) + \ldots + \lambda_4^2(\frac{1}{x(122)} + \frac{1}{x(222)})\ .$$

Note that the last value in (3.10) is the modified Neyman χ_1^2

$$\chi_1^2 = \Sigma\ \frac{(obs-exp)^2}{obs}\ . \tag{3.11}$$

and indeed the equations in (3.6) are those to determine the minimum modified χ^2 estimates. The next to last value in (3.10) is the statistic given by Bhapkar and Koch [1968, p. 116] based on a

criterion due to Wald. The square root of this value is the statistic used by Snedecor and Cochran [1967, p. 496].

In accordance with the minimum discrimination information theorem (Kullback [1959]) the log-linear representation for $x^*(ijk)$ is given graphically as in figure 1 where the interpretation is

$$
\begin{aligned}
&\ell n \frac{x^*(111)}{x(111)} = L_1 - \tau/x(.11) \ , \\
&\ell n \frac{x^*(211)}{x(111)} = L_1 \ , \\
&\ell n \frac{x^*(112)}{x(112)} = L_2 + \tau/x(.12) \ , \\
&\ell n \frac{x^*(212)}{x(212)} = L_2 \ , \\
&\cdots \\
&\ell n \frac{x^*(222)}{x(222)} = L_4 \ .
\end{aligned} \tag{3.12}
$$

Recalling (3.8) we see that (3.12) in fact leads to (3.5).

If we write

$$
\begin{aligned}
\theta^* &= \frac{x^*(111)}{x(.11)} - \frac{x^*(112)}{x(.12)} - \frac{x^*(121)}{x(.21)} + \frac{x^*(122)}{x(.22)} = p^*(11) - p^*(12) - p^*(21) + p^*(22) , \\
\hat{\theta} &= \frac{x(111)}{x(.11)} - \frac{x(112)}{x(.12)} - \frac{x(121)}{x(.21)} + \frac{x(122)}{x(.22)} = \hat{p}(11) - \hat{p}(12) - \hat{p}(21) + \hat{p}(22) ,
\end{aligned} \tag{3.13}
$$

then as shown in Kullback [1959, p. 101-106]

$$2I(x^* : x) \simeq (\theta^* - \hat{\theta})^2 / \hat{\sigma}^2 \ , \tag{3.14}$$

where $\hat{\sigma}^2$ is determined as follows. Let $\underline{T}$ denote the 8 x 5 matrix in figure 1, that is,

$$(3.15)\quad \underline{T}= \begin{bmatrix} 1 & 0 & 0 & 0 & -1/x(.11) \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & +1/x(.12) \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & +1/x(.21) \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1/x(.22) \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

and $\underline{D}_x$ the 8 x 8 diagonal matrix with entries x(ijk), that is,

$$(3.16)\quad \underline{D}_x= \begin{bmatrix} x(111) & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & x(211) & & & & & & \vdots \\ \vdots & & x(112) & & & & & \vdots \\ \vdots & & & x(212) & & & & \vdots \\ \vdots & & & & x(121) & & & \vdots \\ \vdots & & & & & x(221) & & \vdots \\ \vdots & & & & & & x(122) & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x(222) \end{bmatrix}.$$

Compute the 5 x 5 matrix $\underline{S} = \underline{T}'\underline{D}_x\underline{T}$ and partition it as follows

$$(3.17)\quad \underline{S} = \begin{bmatrix} \underline{S}_{11} & \underline{S}_{12} \\ \underline{S}_{21} & \underline{S}_{22} \end{bmatrix}, \quad \underline{S}_{11} \text{ is } 4 \times 4,\ \underline{S}_{22} \text{ is } 1 \times 1,\ \underline{S}_{21} = \underline{S}'_{12} \text{ is } 1 \times 4,$$

then $\hat{\sigma}^2$ in (3.14) is given by

$$(3.18)\quad \hat{\sigma}^2 = \underline{S}_{22} - \underline{S}_{21}\,\underline{S}_{11}^{-1}\underline{S}_{12}\ .$$

It may be verified that this results in

$$(3.19)\quad \hat{\sigma}^2 = \frac{x(111)x(211)}{(x(.11))^3} + \frac{x(112)x(212)}{(x(.12))^3} + \frac{x(121)x(221)}{(x(.21))^3} + \frac{x(122)x(222)}{(x(.22))^3}$$

$$= \frac{\hat{p}(11)\hat{q}(11)}{x(.11)} + \frac{\hat{p}(12)\hat{q}(12)}{x(.12)} + \frac{\hat{p}(21)\hat{q}(21)}{x(.21)} + \frac{\hat{p}(22)\hat{q}(22)}{x(.22)}\ .$$

But θ^* in (3.13) is zero and we see that (3.14) is indeed the next-to-last value in (3.10). It is interesting to note that $2I(x^*:x)$ can be approximated without necessarily computing the values of $x^*(ijk)$.

Note now that in order to express the hypothesis H_o of (3.2) in the form $\underline{B}\,\underline{p} = \underline{0}$, we can let

$$\underline{p} = (p(111), p(211), p(112), p(212), p(121), p(221), p(122), p(222))',$$

(3.20)
$$\underline{B} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \end{bmatrix}$$

and

(3.21) $\underline{0} = (1,1,1,1,0)'$.

Figure 1

i	j	k	L_1	L_2	L_3	L_4	τ
1	1	1	1				-1/x(.11)
2	1	1	1				
1	1	2		1			+1/x(.12)
2	1	2		1			
1	2	1			1		+1/x(.21)
2	2	1			1		
1	2	2				1	-1/x(.22)
2	2	2				1	

We shall illustrate the preceding discussion by Bartlett's data on root cuttings used also as an example by Snedecor and Cochran [1967], Bhapkar and Koch [1968], Berkson [1972].

The following Table 4 from Bartlett [1935], who refers to data from Hoblyn and Palmer, is the result of an experiment designed to investigate the propagation of plum root stocks from root cuttings. There were 240 cuttings for each of the four treatments.

Table 4.

	At Once j=1		In Spring j=2	
	Long k=1	Short k=2	Long k=1	Short k=2
Dead i=1	84	133	156	209
Alive i=2	156	107	84	31
	240	240	240	240

By using the $\underline{B}$ and $\underline{0}$ defined in (3.20) and (3.21) and the iterative algorithm described in the Appendix the MDI estimates $x^*(ijk)$ of the cell-frequencies are obtained as

82.883 134.213 157.117 208.448

157.117 105.787 82.883 31.552

They agree within round-off errors with those obtained by Berkson [1972]. The MDI statistic $2I(x^*:x)$ equals 0.0819 with one D.F.

4 . THE 4x2x2 TABLE

Analysis of hypotheses of no linear interaction in a 4x2x2 table is illustrated by Schotz's data Table 5 on drivers in injury producing accidents, taken from Table III of Bhapkar and Koch [1968], who regard accident severity as response and the other two classifications as factors.

Table 5.

	Accident Severity (i)	Minor	Moderate	Moderately Severe	Severe to Extreme	Total
Driver Group (k)	"ridit" Accident (r) Type (j)	.05	.33	.71	.93	
Lone Driver	Rollover	21	567	1356	644	2588
	Non-rollover	996	5454	2773	1256	10479
	Sub-total	1017	6021	4129	1900	13067
Injured Driver with Passengers	Rollover	18	553	1734	869	3174
	Non-rollover	679	4561	2516	1092	8848
	Sub-total	697	5114	4250	1961	12022
Total		1714	11135	8379	3861	25089

Let us ignore the numerical severity "ridit" scores r_i, i=1,...4 and consider the hypothesis of no linear second order interaction formulated in (2.3). The B matrix is

Cell index: 111 211 311 411 112 212 312 412 121 221 321 421 122 222 322 422

(4.1) $\underline{B}=$

1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
1	0	0	0	-1	0	0	0	-1	0	0	0	1	0	0	0
0	1	0	0	0	-1	0	0	0	-1	0	0	0	1	0	0
0	0	1	0	0	0	-1	0	0	0	-1	0	0	0	1	0

and

(4.2) $$\underline{\theta} = (1,1,1,1,0,0,0)'.$$

Using the algorithm described in the Appendix, the MDI estimates of cell frequencies come out to be

x*(111)=	27.32	x*(211)=	531.50	x*(311)=	1359.16	x*(411)=	670.02
x*(121)=	932.32	x*(221)=	5535.91	x*(321)=	2768.49	x*(421)=	1242.28
x*(112)=	14.59	x*(212)=	583.70	x*(312)=	1733.23	x*(412)=	842.48
x*(122)=	734.45	x*(222)=	4884.30	x*(322)=	2522.48	x*(422)=	1106.77

The MDI statistic 2I(x*:x) with 3 d.f. is 19.703, which is significant at 5% level, showing that the data do <u>not</u> support the hypothesis of no linear second order interaction as given by (2.3).

It is interesting to examine here the hypothesis of no linear second order interaction with respect to average "ridits" considered by Bhapkar and Koch [1968]. The hypothesis is

$$H_1 : A_k = \Sigma_{i=1}^{4} r_i [p(i1k)-p(i2k)] = A, \quad k=1,2,$$

where A is a constant. This is equivalent to $A_1 - A_2 = 0$. The 5x16 matrix B_1 corresponding to H_1 has the same first 4 rows as $\underline{B}$ and

the fifth row is

$$(r_1, r_2, r_3, r_4, -r_1, -r_2, -r_3, -r_4, -r_1, -r_2, -r_3, -r_4, r_1, r_2, r_3, r_4).$$

The vector $\underline{\theta}_1$ equals $(1,1,1,1,0)'$.

The MDI estimates $x^*_1(ijk)$ of cell frequencies are given below:

$x^*_1(111) = 19.96$	$x^*_1(211) = 551.09$	$x^*_1(311) = 1359.62$	$x^*_1(411) = 657.3$
$x^*_1(121) = 1004.55$	$x^*_1(221) = 5470.00$	$x^*_1(321) = 2759.92$	$x^*_1(421) = 1244.5$
$x^*_1(112) = 18.80$	$x^*_1(212) = 566.69$	$x^*_1(312) = 1732.79$	$x^*_1(412) = 855.7$
$x^*_1(122) = 671.88$	$x^*_1(222) = 4543.52$	$x^*_1(322) = 2529.15$	$x^*_1(422) = 1103.4$

The MDI statistic $2I(x^*_1 : x)$ is 1.980 with 1 d.f. This should be compared with the value 2.02 obtained by Bhapkar and Koch [1968] for their Wald-type statistic.

Now observe that H_1 is implied by the stronger hypothesis H_o given by (2.3), since the fifth row of $\underline{B}_1$ can be expressed as a linear combination of rows of $\underline{B}$. To see this let B(h) denote the h-th row of $\underline{B}$ of (4.1), then

$$B_1(5) = r_1 B(5) + r_2 B(6) + r_3 B(7) + r_4 [B(1) - B(2) - B(3) + B(4) - B(5) - B(6) - B(7)].$$

Hence we can analyze the information $2I(x^* : x)$ as follows:

Analysis of Information

Component due to	Information	D.F.	Chi-square (5%)
H_o	$2I(x^* : x) = 19.703$	3	7.815
	$2I(x^* : x^*_1) = 17.723$	2	5.991
H_1	$2I(x^*_1 : x) = 1.980$	1	3.841

We see that the data do not provide statistically significant evidence against the hypothesis H_1 of no second-order interaction with respect to average "ridits". In other words, this hypothesis does explain the departure from the hypothesis (2.3) of no linear second-order interaction.

Further analysis of these data can be done in two ways; in terms of "ridit" values and in terms of the non-quantitative contrasts among p(ijk) given by the last three rows of the matrix B of (4.1).

"Ridits" : Note that the data are not consistent with the hypothesis of no linear second-order interaction ($2I(x^*:x)=19.703$, 3 d.f.), while they can be regarded as consistent with the hypothesis H_1 of equality of means of the "ridit" values (r_1, r_2, r_3, r_4) of the four distributions ($2I(x_1^*:x)=1.980$, 1 d.f.). The remaining two degrees of freedom can be associated respectively with the hypotheses of equality of second and third moments of the "ridit" values. the hypothesis H_2 of equality of means and second moments (which is equivalent to the hypothesis of equality of means and variances) corresponds to a 6x16 matrix, B_2, say, which has the first five rows as in B_1 and the sixth row is

$$(r_1^2, r_2^2, r_3^2, r_4^2, -r_1^2, -r_2^2, -r_3^2, -r_4^2, -r_1^2, -r_2^2, -r_3^2, -r_4^2, r_1^2, r_2^2, r_3^2, r_4^2)$$

and $\theta_2=(1,1,1,1,0,0)'$.

Under H_2 the MDI statistic $2I(x_2^*:x)$ comes out to be 10.036. The difference 10.036-1.980=8.056 is the contribution due to the additional constraint in B_2 as compared to B_1, assignable to equality of variances. Finally the difference 19.703-10.036=9.667 is the contribution due to equality of the third moments in

addition to the equality of the first two moments. Since each of these differences is asymptotically a chi-square with one degree of freedom , we conclude that though there is no significant second- order linear interaction with respect to mean "ridits", there appears to be a significant contribution due to heterogeneity of the second and third moments of the four "ridit" distributions.

Non-quantitative approach. A different line of analysis treats the response variable (accident severity) as a qualitative variable ignoring "ridit" values. In this case, since the overall hypothesis of no linear second-order interaction leads to a significant MDI statistic ($2I(x^{*}:x)=19.703$, 3 d.f.) it may be of interest to examine which of the three constraints (given by the last three rows of the matrix B in (4.1)) contribute significantly to $2I(x^{*}:x)$. For this purpose , we set up several B-matrices omitting one or two rows from the last three rows of (4.1) each time. For example, the B-matrix without the seventh row corresponds to the (weaker) hypothesis

$$H_3: \quad \begin{aligned} &p(111)-p(112)-p(121)+p(122)=0, \\ &p(211)-p(212)-p(221)+p(222)=0. \end{aligned}$$

Implicit in H_3 is the third constraint

$$[p(311)+p(411)]-[p(312)+p(412)]-[p(321)+p(421)]+[p(322)+p(422)]=0.$$

Hence H_3 tests no linear second-order interaction with respect to levels 1 and 2 combining levels 3 and 4 of the response. Note that under these weaker hypotheses the MDI statistics will give a value not larger than 19.703. The analysis is summarized in Table 6.

Omitting rows 5 and 6 of the matrix B of (4.1) corresponds to the hypothesis of no linear second-order interaction in a 2x2x2 table with level 3, pooling all the remaining levels. This is the only hypothesis with which the data are consistent. Thus it appears that levels 1 and 2 of accident severity both jointly and separately account for a major (significant) contribution towards the presence of a linear second-order interaction.

Table 6 also indicates a way of reducing categories in a contingency table with the inherent qualities of the observed data least affected. Thus if the given 4x2x2 table is to be reduced to a 3x2x2 table, this should be done by combining levels 3 and 4. Similarly, if a 2x2x2 table is required as a partial summary of the 4x2x2 table one should examine all the possible ways of pooling the levels of the response variable and select the way in which the maximum contribution to the linear second-order interaction is retained. The possible ways are level

(1) against (2)+(3)+(4), (2) against (1)+(3)+(4),

(3) against (1)+(2)+(4), (1)+(2) against (3)+(4),

(1)+(3) against (2)+(4), and (1)+(4) against (2)+(3).

The MDI statistics corresponding to the first three combinations are given in table 6 as the three entries 11.803,9.750, and 0.032 respectively. To find the MDI statistics corresponding to the remaining three combinations one can add the last two rows of the B-matrices when rows 7, 6, 5, are omitted one at a time. This gives the MDI statistics as 3.517, 1.538, and 8.078 respectively. The largest of these MDI statistics is 11.803, showing that levels 2, 3, and 4 should be pooled and level 1 be retained in the 2x2x2 table.

The analysis above shows that levels 1 and 2 are the main contributors to the departure from the hypothesis of no linear second-order interaction.

Table 6

MDI Statistics Under Different B-matrices

Operation on rows of (4.1)	MDI statistic	D.F.
Delete (7)	18.385	2
Delete (6)	12.125	2
Delete (5)	13.188	2
Delete (6), (7)	11.803	1
Delete (5), (7)	9.750	1
Delete (5), (6)	0.032	1
Delete (7), add (5) and (6)	3.517	1
Delete (6), add (5) and (7)	1.538	1
Delete (5), add (6) and (7)	8.088	1

5. Acknowledgment

The work of the first author was supported in part by Research Grant HL 15191 from National Heart and Lung Institute, National Institutes of Health, Bethesda Maryland.

APPENDIX

Described below is an algorithm to obtain $x^*(ijk) = Np^*(ijk)$ which minimize the discriminationinformation function (2.6) subject to the constraints $\underline{B}\underline{p} = \underline{\theta}$, where $\underline{p}$ is as in (2.4). (Gokhale [1974]).

With $w(jk) = x(.jk)/N$, multiply the first r elements of $\underline{p}$ by $w(11)$, the second r elements by $w(12)$, and so on. The vector so obtained can be written as $\underline{W}\underline{p}$, where $\underline{W}$ is a diagonal matrix, the entries in the first r diagonal positions being $w(11)$, those in the next r diagonal positions being $w(12)$, etc. In fact, it is easy to see that $\underline{W}\underline{p}$ is a probability distribution over the rst cells. The constraints $\underline{B}\underline{p} = \underline{\theta}$ can be written as

$$\underline{B}\underline{W}^{-1}\underline{W}\underline{p} = \underline{\theta} = \underline{C}(\underline{W}\underline{p}), \text{ say.}$$

The elements of $\underline{W}\underline{p}$ can be indexed by a subscript t, say. It is thus sufficient to consider the problem of minimizing

(A.1) $I(P:\Pi) = \Sigma_t P_t \ln(P_t/\Pi_t)$

with respect to the constraints

(A.2) $\underline{C}\underline{P} = \underline{\theta}$.

Note that $\underline{C} = \underline{B}\underline{W}^{-1}$, $\underline{P} = \underline{W}\underline{p}$ and $\underline{\Pi} = \underline{W}\underline{\pi}$.

Assume now that the rows of $\underline{C}$ are linearly independent. There exists a unique $\underline{P}^*$ which minimizes (A.1) and satisfies

(A.3) $\underline{\ln P^*} = \underline{\ln \Pi} + \underline{C}'\underline{\lambda}$

where $\underline{\ln a}$ denotes $(\ln a_1, \ldots, \ln a_t)'$ and $\underline{\lambda}$ is a vector of Lagrangian multipliers. (See Kullback [1959]). Let

(A.4) $\underline{C}^+ = \underline{C}'(\underline{C}\underline{C}')^{-1}$ and $\underline{R} = \underline{C}^+\underline{C}$.

Then equation (A.3) is equivalent to

(A.5) $(\underline{I}-\underline{R})\ (\ell n\ \underline{P}^* - \ell n\ \underline{\Pi}) = 0.$

The symmetric and idempotent matrix $\underline{R}$ projects vectors of dimension equal to that of $\underline{P}$ onto the space spanned by rows of $\underline{C}$. Let

$$U = \{\underline{z} : \underline{C}^+\underline{\theta} + (\underline{I}-\underline{R})\underline{z} > 0 \},$$

where for a vector $\underline{x}$, $\underline{x} > 0$ denotes that every element of $\underline{x}$ is positive. Then for every $\underline{z} \in U$, $\underline{C}^+\underline{\theta}+(\underline{I}-\underline{R})\underline{z}$ is a solution of (A.2). Conversely, for every probability vector $\underline{P}$ which satisfies (A.2), there exists a $\underline{z} \in U$ such that $\underline{P} = \underline{C}^+\underline{\theta}+(\underline{I}-\underline{R})\underline{z}$. The first assertion is easy to verify and the second follows by setting $\underline{z} = \underline{P}$ and noting that $\underline{C}^+\underline{\theta} = \underline{RP}$. Consider (A.1) as a function of $\underline{z}$ defined over U. Then

$$I(z) > 0,$$

the gradient $G(z)$ of $I(z)$ at z is

(A.6) $G(z) = (\underline{I}-\underline{R})\ (\ell n\ \underline{P}(z) - \ell n\ \underline{\Pi}\ )$

and the Hessian of I at z is

(A.7) $H(z) = (\underline{I}-\underline{R})[\underline{\Delta}(P(z))]^{-1}(\underline{I}-\underline{R}),$

where $\underline{\Delta}$ (b) denotes a diagonal matrix with elements of vector b in the diagonal.

Being idempotent, $\underline{I}-\underline{R}$ is positive definite,

so that I is a convex function of z over the convex set U.

Thus for a z_o satisfying $G(z_o) = 0$, $I(z)$ assumes its minimum over U. In fact, $G(z_o) = 0$ implies that the corresponding $\underline{P}(z_o)$ satisfies (A.5) in view of (A.6).

At the s-th iteration the algorithm uses a vector $\underline{z}(s)$ in U and the corresponding $\underline{P}(s) = \underline{C}^{+}\underline{\theta}+(\underline{I}-\underline{R})\underline{z}(s)$. If

(A.8) $|G(s)| = |G[z(s)]| \leq \varepsilon$,

where $\varepsilon > 0$ is chosen according to the required accuracy, the procedure is terminated and $\underline{P}^{*}$ is set equal to $\underline{P}(s)$. If (A.8) does not hold, the direction $\underline{D}(s)$ of maximum rate of decrease in $I(z)$ at $\underline{z}(s)$ is obtained by norming $(-G(s))$. A positive constant $c(s)$, sufficiently small, is then found such that with $\underline{z}(s+1) = \underline{z}(s)+c(s)\underline{D}(s)$ and

$\underline{P}(s+1)=\underline{C}^{+}\underline{\theta}+(\underline{I}-\underline{R})\underline{z}(s+1)$

(A.9) $\underline{P}(s+1) > 0$

and

(A.10) $I(s+1) < I(s)$.

The $(s+1)$-th iteration is started with $\underline{z}(s+1)$ and $\underline{P}(s+1)$. One way of finding $c(s)$ is to first set it equal to unity. It is repeatedly doubled until one of (A.9) or (A.10) is violated. If (A.9) or (A.10) do not hold with $c(s) = 1$, it is repeatedly halved until they do.

Consider now the choice of $\underline{z}(1)$ and $\underline{P}(1)$. If some $\hat{\underline{p}} > 0$ is known to satisfy (A.2), we set $\underline{P}(1) = \underline{z}(1) = \hat{\underline{p}}$. If not, $\hat{\underline{p}}$ can be

found easily, though by trial and error, by several methods. One method is to compute $\underline{q}(\xi) = \underline{\Delta}(\xi)\underline{C}'[\underline{C}\underline{\Delta}(\xi)\underline{C}']^{-1}\ \underline{\theta}$ for a positive probability vector $\underline{\xi}$ and set $\hat{\underline{p}} = \underline{q}(\xi)$ if the latter is positive. Another method is to check whether $\underline{C}^{+}\underline{\theta}+(\underline{I}-\underline{R})\underline{\xi}$ is positive and, if so, set it equal to $\underline{P}(1)$. Usually, putting $\underline{\xi}$ equal to the observed probability vector gives the desired value of $\hat{\underline{p}}$. In fact, then $\underline{q}\ (\xi)$ is the "minimum modified chi-square" estimate of $\underline{P}$ subject to (A.2), which minimizes $\Sigma(P_t - \xi_t)^2/\xi_t$, while $\underline{C}^{+}\underline{\theta}+(\underline{I}-\underline{R})\underline{\xi}$ minimizes the Euclidean distance between $\underline{P}$ and $\underline{\xi}$. As such, these $\hat{\underline{p}}$ serve as good starting points for the iterations.

The numerical computations of sections 3 and 4 were programmed in APL/360.

REFERENCES

BERKSON, J., 1972. Minimum discrimination information, the 'no interaction' problem, and the logistic function. Biometrics, 28, 443-468.

BHAPKAR, V. P. and KOCH, GARY G., 1968. Hypotheses of 'no interaction' in multidimensional contingency tables. Technometrics, 10, 107-123.

DARROCH, J. N., 1974. Multiplicative and additive interaction in contingency tables. Biometrika, 61, 2, 207-214.

GOKHALE, D. V., 1974. A steepest descent algorithm for minimum discrimination information. Unpublished.

KULLBACK, S., 1959, Information Theory and Statistics. J. Wiley amd Sons, Inc., New York, Dover Edition, 1968.

SNEDECOR, G. W. and COCHRAN, W. G., 1967. Two-way classifications with unequal numbers and proportions. Statistical Methods (6th Edn.). Iowa State University Press, Ames, Iowa. 495-496.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 9	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Information Analysis of Linear Interactions In Contingency Tables		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 9 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) S. Kullback and D.V. Gokhale		8. CONTRACT OR GRANT NUMBER(s) DAAG29-77-G-0031
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P-14435-M
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE August 15, 1977 13. NUMBER OF PAGES 25
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents. This report partially supported under Office of Naval Research Contract N00014-76-C-0475 (NR-042-267) and issued as Technical Report No. 249.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Contingency tables, no linear interactions, minimum discrimination information approach

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The use of the minimum discrimination information (MDI) approach in studying null hypotheses of no interactions on a linear scale in contingency tables of "one response many factors" type is illustrated. Quadratic approximations to the M.D.I. statistic are related to Wald-type statistics and Neyman's modified chi-square. Follow up analyses when the null hypothesis is not satisfied are illustrated.