ANALYSIS OF INTERACTIONS BETWEEN CATEGORICAL VARIABLES

S. Kullback, et al

George Washington University

Prepared for:

Office of Naval Research
Air Force Office of Scientific Research

2 August 1974

UNCLASSIFIED
Security Classification

i

AD- 783686

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)
THE GEORGE WASHINGTON UNIVERSITY
DEPARTMENT OF STATISTICS
WASHINGTON, D.C. 20006

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3 REPORT TITLE
ANALYSIS OF INTERACTIONS BETWEEN CATEGORICAL VARIABLES

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)
TECHNICAL REPORT

5 AUTHOR(S) (First name, middle initial, last name)
S. KULLBACK AND P. N. REEVES

6 REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August 2, 1974	21	6

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0214-0015	
b. PROJECT NO: NR-042-267	22
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10 DISTRIBUTION STATEMENT
Unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
	Office of Naval Research Statistics & Probability Program Arlington, Virginia 22217

13 ABSTRACT

The principle of minimum discrimination information estimation and associated procedures are applied to data from a survey of hospitals to determine the relationship of innovativeness on certain hospital characteristics.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
MULTIDIMENSIONAL CONTINGENCY TABLE ANALYSIS						
CATEGORICAL VARIABLES						
INTERACTION ANALYSIS						

ANALYSIS OF INTERACTIONS BETWEEN CATEGORICAL VARIABLES

by

S. Kullback and P. N. Reeves

TECHNICAL REPORT NO. 22

August 2, 1974

PREPARED UNDER CONTRACT N00014-67-A-0214-0015

(Nr-042-267)

OFFICE OF NAVAL RESEARCH

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, D.C. 20006

Analysis of Interactions Between Categorical Variables

by

S. Kullback and P. N. Reeves
The George Washington University

The principle of minimum discrimination information estimation and associated procedures are applied to data from a survey of hospitals to determine the relationship of innovativeness on certain hospital characteristics.

Introduction

The literature in the health field contains many examples of cases where a researcher has been unable to deal with interaction between variables even though it is generally believed that such interaction does exist. The inability to deal with these interactions stems from the fact that often, data in the health field do not meet the assumptions for use of classical parametric procedures. Until recently none of the non-parametric techniques available could cope with this kind of situation.

A recent article, Johnson et. al. [2] contains a brief list of various techniques now available and gives a detailed description of one particular method for dealing with certain

types of this sort of data. Among the approaches listed but not described in detail [2] refers to the minimum discrimination information methodology reported in [3]. Subsequent research has extended the minimum discrimination information methodology [4], [6], and enhanced its usefulness for health research.

The purpose of this article is two-fold. First, to point out that the reservations noted in [2] do not restrict the usefulness of this approach. Second, to illustrate the technique using health data so that researchers in the health field can become aware of this valuable tool and add it to their armamentarium.

"Multivariate data analysis needs a large and flexible class of hypothetical distributions of free variables indexed by the values of fixed variables. From this class, appropriate subfamilies would be chosen for fitting to specific data sets" [1]. The principle of minimum discrimination information estimation and its basis the minimum discrimination information theorem, which is quite general in its formulation, lead to exponential families of distributions [4], [5], [6]. The exponential families have very useful and desirable statistical properties and contain many subfamilies in common use [1]. "The data analytic attitude to models is empirical rather than theoretical. ...When detailed theoretical understanding is unavailable, a more empirical attitude is natural, so that es-

timation of parameters in models should be seen less as attempts to discover underlying truth and more as data calibrating devices which make it easier to conceive of noisy data in terms of smooth distributions and relations. Exponential families are viewed here as intended for use in the empirical mode. With a given data set, a variety of models may be tried on, and one selected on the grounds of looks and fit" [1]. When the minimum discrimination information estimates provide a satisfactory fit to a set of data a complete analysis, including significance tests and estimates describing the pattern of observations is provided.

We propose to present an example of the use of the principle of minimum discrimination information estimation, its related statistics, exponential family and analysis of information, in relation to a matter of concern to health administrators. The data used are from the field of hospital administration and relate to the matter of innovation in hospitals. We begin with the assumption that the use of electronic data processing (EDP) in hospitals in the late 1960's was innovative. This assumption is substantiated by a variety of surveys of the use of EDP in hospitals. (See Jacobs, Reeves, and Hammond article to be published in Hospitals.) On this basis the data in a survey of hospitals using EDP conducted by Herner and Co. were combined with data from the Guide Issue of Hospitals for the same period

so that a file of records reflecting characteristics of hospitals and levels at which EDP was used by these hospitals was created. The hospitals in this survey were selected by stratified sampling. The stratification (fixed variable) was on the basis of hospital size. All hospitals in the large-size category (200 or more beds) were included in the survey and a ten percent sample was taken of those in the small size category. The data from these files were tabulated and arranged in multiway contingency tables. The analysis of the tables for the large and small hospitals will be described here and interrelated to illustrate the use of the minimum discrimination information estimation technique. Computer programs have been prepared and are available to provide the necessary output for the analysis.

On the basis of these analyses we conclude that there is a distinct relation of innovation on location and length of stay with a common factor for large and small hospitals. The association (measured by the logarithm of the cross-product ratio) between use of EDP and length of stay is the same for the large and small hospitals. The log-odds (logit) of use of EDP in descending order of magnitude within the large hospitals and within the small hospitals are parallel in terms of the combinations of the factors location and length of stay. The usage of EDP is generally greater in the large

hospitals than in the small hospitals except that the best log-odds for the small hospitals is greater than the poorest log-odds for the large hospitals.

Hospital Characteristics Associated With Use of EDP

In a study to identify characteristics which distinguish hospitals which use EDP from those which do not, that is, to identify characteristics which are significantly associated with use of EDP, data on 1176 hospitals, 923 large and 253 small, were collected with respect to use, location, and length of stay. The data appear in the two three-way 2x2x2 contingency tables 1 and 2. In order to determine the relation among the free variables use, location, and length of stay, indexed by size of hospital, and interactions that may exist among these characteristics it seems intuitively clear that an analysis based only on two-way tables would not suffice. We shall analyze the data using the principle of minimum discrimination information estimation and its associated statistics, as presented and discussed in [4], [6].

We shall denote the occurrences in the observed tables 1 and 2 respectively by x(ijk), y(ijk) with

i=1, user; i=2, non-user
j=1, urban; j=2, rural
k=1, short; k=2, long.

The proposed procedure provides estimates for the original data analogous to a regression procedure using sets of observed marginals as explanatory variables and we shall try to find an estimate which does not differ significantly from the observed data. The set of acceptable estimates will indicate the nature of the significant interactions for which we can compute numerical measures.

As a first step in the analysis we shall find "smoothed" estimates of the original data. We shall do this for the large hospitals also even though the data for all large hospitals was collected. We examine the minimum discrimination information estimates obtained by a convergent iterative algorithm starting with a uniform table and successively adjusting for sets of observed marginals. It turns out that the set of two-way marginals are best and the resultant estimates provide a satisfactory fit. The estimated tables have the same two-way and also the same one-way marginals as the original tables [4], [6]. These estimates which we denote by $x_2^*(ijk)$, $y_2^*(ijk)$ respectively for the large and small hospitals are given in tables 3 and 4 and imply no second-order (three-factor) interaction. Note that the estimate for the observed $y(122)=0$ is $y_2^*(122)=0.137$.

The estimates are given analytically by the log-linear representation of an exponential family

$$\ell n\frac{x_2^*(ijk)}{n\pi(ijk)} = L+\tau_1^i T_1^i(ijk)+\tau_1^j T_1^j(ijk)+\tau_1^k T_1^k(ijk)+\tau_{11}^{ij}T_{11}^{ij}(ijk)$$
$$+\tau_{11}^{ik}T_{11}^{ik}(ijk)+\tau_{11}^{jk}T_{11}^{jk}(ijk) \qquad (1)$$

where $n=\sum\sum\sum x(ijk)$, $\pi(ijk)=1/2\times 2\times 2$, L is a normalizing constant, the taus are main-effect and interaction parameters, and the T(ijk) are a set of linearly independent random variables, in this case the indicator functions of the respective marginals. A similar representation holds for $y_2^*(ijk)$. The log-linear representations are shown graphically in Figure 1 [6]. The values in the various columns of Figure 1, zeros or ones, are the values of the respective functions T(ijk). Note that

$$T_{11}^{ij}(ijk)=T_1^i(ijk)T_1^j(ijk), T_{11}^{ik}(ijk)=T_1^i(ijk)T_1^k(ijk),$$

$$T_{11}^{jk}(ijk)=T_1^j(ijk)T_1^k(ijk).$$

To test the goodness-of-fit of the estimates we compute the statistics [4], [6],

$$2I(x:x_2^*)=2\sum\sum\sum x(ijk)\ell n(x(ijk)/x_2^*(ijk))=0.481, \quad 1 \text{ D.F.}$$

$$2I(y:y_2^*)=2\sum\sum\sum y(ijk)\ell n(y(ijk)/y_2^*(ijk))=0.294, \quad 1 \text{ D.F.}$$

Since the statistics are asymptotically distributed as χ^2 we conclude that the "smoothed" values x_2^*, y_2^* are good estimates and we shall use them in our subsequent analysis.

From the log-linear representation (1) or the graphical presentation in Figure 1, we find that the log-odds or logits of the use of EDP for large hospitals is given by the parametric representation

$$\ell n \frac{x_2^*(111)}{x_2^*(211)} = \tau_1^i + \tau_{11}^{ij} + \tau_{11}^{ik}$$

$$\ell n \frac{x_2^*(112)}{x_2^*(212)} = \tau_1^i + \tau_{11}^{ij}$$

$$\ell n \frac{x_2^*(121)}{x_2^*(221)} = \tau_1^i + \tau_{11}^{ik} \qquad (2)$$

$$\ell n \frac{x_2^*(122)}{x_2^*(222)} = \tau_1^i$$

where the values of the parameters for the estimate $x_2^*(ijk)$ are found to be

$$\tau_1^i = -1.4842,\ \tau_{11}^{ij} = 0.5113,\ \tau_{11}^{ik} = 1.5103.$$

From (2) we also see that for the large hospitals

$$\tau_{11}^{ij} = \ell n \frac{x_2^*(111)x_2^*(221)}{x_2^*(211)x_2^*(121)} = \ell n \frac{x_2^*(112)x_2^*(222)}{x_2^*(212)x_2^*(122)} = 0.5113,$$

that is, the association between usage and location for either short or long stay. Similarly

$$\tau_{11}^{ik} = \ell n \frac{x_2^*(111)x_2^*(212)}{x_2^*(211)x_2^*(112)} = \ell n \frac{x_2^*(121)x_2^*(222)}{x_2^*(221)x_2^*(122)} = 1.5103,$$

that is, the association between usage and stay for either urban or rural location.

For the small hospitals the log-odds or logits are

$$\ell n \frac{y_2^*(111)}{y_2^*(211)} = \tau_1^i + \tau_{11}^{ij} + \tau_{11}^{ik}$$

$$\ell n \frac{y_2^*(112)}{y_2^*(212)} = \tau_1^i + \tau_{11}^{ij}$$

$$\ell n \frac{y_2^*(121)}{y_2^*(221)} = \tau_1^i + \tau_{11}^{ik}$$

$$\ell n \frac{y_2^*(122)}{y_2^*(222)} = \tau_1^i$$

where the values of the parameters for the estimate $y_2^*(ijk)$ are found to be

$$\tau_1^i = -3.3357,\ \tau_{11}^{ij} = 1.3088,\ \tau_{11}^{ik} = 0.9836.$$

For the small hospitals we also have

$$\tau_{11}^{ij} = \ell n \frac{y_2^*(111)y_2^*(221)}{y_2^*(211)y_2^*(121)} = \ell n \frac{y_2^*(112)y_2^*(222)}{y_2^*(212)y_2^*(122)} = 1.3088,$$

that is, the association between usage and location for either short or long stay. Similarly

$$\tau_{11}^{ik} = \ell n \frac{y_2^*(111)y_2^*(212)}{y_2^*(211)y_2^*(112)} = \ell n \frac{y_2^*(121)y_2^*(222)}{y_2^*(221)y_2^*(122)} = 0.9836,$$

that is, the association between usage and stay for either urban or rural locations.

Since the data for the large hospitals reflect observations over all such hospitals, it will be of interest to determine whether there exists a suitable estimate for the

small hospitals, other than $y_2^*(ijk)$, which will have some of its interactions (associations) the same as the corresponding values for the large hospitals. This can be accomplished by using the iterative algorithm fitting various subsets of marginals of $y_2^*(ijk)$ (or the original $y(ijk)$) but starting with a distribution which has the same tau parameters as $x_2^*(ijk)$. The tau parameters of $x_2^*(ijk)$ not affected by the iterative fitting procedure will be "inherited" by the resultant estimate. We shall use the table $v(ijk)=(253/923)x_2^*(ijk)$ which has the same tau parameters as the $x_2^*(ijk)$ table with total adjusted to be the same as the observed total of small hospitals.

We summarize the procedure: starting the iterative fitting algorithm with $v(ijk)$ (recall that $y(ijk)$ and $y_2^*(ijk)$ have the same two-way and one-way marginals)

	Marginals fitted	Estimate	Tau Parameters "inherited" from v(ijk)
a)	$y(i.k), y(.jk)$	$u_a^*(ijk)$	τ_{11}^{ij}
b)	$y(ij.), y(.jk)$	$u_b^*(ijk)$	τ_{11}^{ik}
c)	$y(ij.), y(i.k)$	$u_c^*(ijk)$	τ_{11}^{jk}
d)	$y(.jk), y(i..)$	$u_d^*(ijk)$	$\tau_{11}^{ij}, \tau_{11}^{ik}$
e)	$y(i.k), y(.j.)$	$u_e^*(ijk)$	$\tau_{11}^{ij}, \tau_{11}^{jk}$
f)	$y(ij.), y(..k)$	$u_f^*(ijk)$	$\tau_{11}^{ik}, \tau_{11}^{jk}$
g)	$y(i..), y(.j.), y(..k)$	$u_g^*(ijk)$	$\tau_{11}^{ij}, \tau_{11}^{ik}, \tau_{11}^{jk}$

In order to test whether the u* estimates differ significantly from the y_2^* estimates, that is, whether the interaction parameters in y_2^* differ significantly from the interaction parameters in u* "inherited" from x_2^* or v, we compute the statistic

$$2I(y_2^*:u_m^*)=2\sum\sum\sum y_2^*(ijk)\ell n(y_2^*(ijk)/u_m^*(ijk))$$

which is asymptotically distributed as χ^2 with 1 D.F. for m=a,b,c, 2 D.F. for m=d,e,f, 3 D.F. for m=g.

The only case which yielded a non-significant value was $u_b^*(ijk)$ for which

$$2I(y_2^*:u_b^*) = 0.408, \quad 1 \text{ D.F.}$$

The values of $u_b^*(ijk)$ are given in Table 5.

The log-linear representation for $u_b^*(ijk)$ in terms of v(ijk) is

$$\ell n\frac{u_b^*(ijk)}{v(ijk)}=L+\tau_1^i T_1^i(ijk)+\tau_1^j T_1^j(ijk)+\tau_1^k T_1^k(ijk) \qquad (3)$$
$$+\tau_{11}^{ij}T_{11}^{ij}(ijk)+\tau_{11}^{jk}T_{11}^{jk}(ijk)$$

Note that τ_{11}^{ik} does not appear explicitly in (3). By using the log-linear representation for v(ijk) itself we also get the reparametrization or log-linear representation for $u_b^*(ijk)$ in terms of the uniform distribution

$$\ell n \frac{u_b^*(ijk)}{n\pi(ijk)} = L + \tau_1^i T_1^i(ijk) + \tau_1^j T_1^j(ijk) + \tau_1^k T_1^k(ijk)$$
$$+ \tau_{11}^{ij} T_{11}^{ij}(ijk) + \tau_{11}^{ik} T_{11}^{ik}(ijk) + \tau_{11}^{jk} T_{11}^{jk}(ijk) \qquad (4)$$

We remark that the numerical values of the taus in (3) and (4) are not the same.

The log-odds or logits of the use of EDP for small hospitals may now be given by the parametric representation

$$\begin{aligned}
\ell n \frac{u_b^*(111)}{u_b^*(211)} &= \tau_1^i + \tau_{11}^{ij} + \tau_{11}^{ik} \\
\ell n \frac{u_b^*(112)}{u_b^*(212)} &= \tau_1^i + \tau_{11}^{ij} \\
\ell n \frac{u_b^*(121)}{u_b^*(221)} &= \tau_1^i + \tau_{11}^{ik} \\
\ell n \frac{u_b^*(122)}{u_b^*(222)} &= \tau_1^i
\end{aligned} \qquad (5)$$

where the values of the parameters in (5) are

$$\tau_1^i = -3.8569, \quad \tau_{11}^{ij} = 1.3354, \quad \tau_{11}^{ik} = 1.5103.$$

For the small hospitals we now have the associations

$$\tau_{11}^{ij} = \ell n \frac{u_b^*(111)u_b^*(221)}{u_b^*(221)u_b^*(121)} = \ell n \frac{u_b^*(112)u_b^*(222)}{u_b^*(212)u_b^*(122)} = 1.3354$$

and

$$\tau_{11}^{ik} = \ell n \frac{u_b^*(111)u_b^*(212)}{u_b^*(211)u_b^*(112)} = \ell n \frac{u_b^*(121)u_b^*(222)}{u_b^*(221)u_b^*(122)} = 1.5103.$$

Note that τ_{11}^{ij}, the association between usage and location for the small hospitals is still different from that for the large hospitals, but that the association between usage and stay, τ_{11}^{ik}, is now the same for both large and small hospitals.

Arranging the log-odds of usage in descending order of magnitude within the large hospitals and within the small hospitals we find

Large hospitals	Factors	Small hospitals
$\ell n \frac{x_2^*(111)}{x_2^*(211)} = 0.5374$	Urban,Short	$\ell n \frac{u_b^*(111)}{u_b^*(211)} = -1.0111$
$\ell n \frac{x_2^*(121)}{x_2^*(221)} = 0.0262$	Rural,Short	$\ell n \frac{u_b^*(121)}{u_b^*(221)} = -2.3466$
$\ell n \frac{x_2^*(112)}{x_2^*(212)} = -0.9729$	Urban,Long	$\ell n \frac{u_b^*(112)}{u_b^*(212)} = -2.5214$
$\ell n \frac{x_2^*(122)}{x_2^*(222)} = -1.4841$	Rural,Long	$\ell n \frac{u_b^*(122)}{u_b^*(222)} = -3.8569$

Conclusion

There are many instances in which a researcher needs some technique that will allow him to take into consideration the interactions of many variables, particularly qualitative, that do not meet the stringent assumptions underlying parametric

statistical testing procedures. Contingency table analysis based upon the minimum discrimination information technique is a tool that is available to fill this need. We have seen that application of this technique to certain types of problems mentioned in [2] is indeed feasible. We have illustrated the use of this technique by showing that innovation in hospitals as indicated by the adoption of EDP is significantly associated with location and length of stay, the latter association being the same for both large and small hospitals. Furthermore, innovativeness is most pronounced for large hospitals with short stay and least for small hospitals with long stay.

Acknowledgement

The work of the first author was partially supported by the Air Force Office of Scientific Research, Office of Aerospace Research, U.S. Air Force under Grant AFOSR-69-1513. Computations were done at the Computer Center of The George Washington University, using programs for the analysis of contingency tables prepared by Professor C. T. Ireland and modifications by Mrs. Marian Fisher.

Table 1

Large Hospitals x(ijk)

	Urban		Rural		
	Short	Long	Short	Long	
User	376	40	52	15	483
Non-user	217	112	54	57	440
	593	152	106	72	923

Table 2

Small Hospitals y(ijk)

	Urban		Rural		
	Short	Long	Short	Long	
User	28	2	11	0	41
Non-user	80	14	114	4	212
	108	16	125	4	253

Table 3

Large Hospitals $x_2^*(ijk)$

	Urban		Rural		
	Short	Long	Short	Long	
User	374.305	41.694	53.695	13.306	483.000
Non-user	218.693	110.308	52.307	58.692	440.000
	592.998	152.002	106.002	71.998	923.000

Table 4

Small Hospitals $y_2^*(ijk)$

	Urban		Rural		
	Short	Long	Short	Long	
User	28.137	1.863	10.863	0.137	41.000
Non-user	79.863	14.137	114.137	3.863	212.000
	108.000	16.000	125.000	4.000	253.000

Table 5

Small Hospitals $u_b^*(ijk)$

	Urban		Rural		
	Short	Long	Short	Long	
User	28.810	1.190	10.917	0.083	41.000
Non-user	79.190	14.810	114.083	3.917	212.000
	108.000	16.000	125.000	4.000	253.000

Figure 1

i	j	k	L	τ_1^i	τ_1^j	τ_1^k	τ_{11}^{ij}	τ_{11}^{ik}	τ_{11}^{jk}
1	1	1	1	1	1	1	1	1	1
1	1	2	1	1	1		1		
1	2	1	1	1		1		1	
1	2	2	1	1					
2	1	1	1		1	1			1
2	1	2	1		1				
2	2	1	1			1			
2	2	2	1						

References

1. Dempster, A. P. An overview of multivariate data analysis. Journal of Multivariate Analysis 1:316, 1971.

2. Johnson, William D. and Koch, Gary. Analysis of qualitative data: linear functions. Health Service Research 5:358, 1970.

3. Ku, H. H. and Kullback, S. Interaction in multidimensional contingency tables: An information theoretic approach. Nat. Bur. Stand. J. Res. 72B:159, 1968.

4. Ku, H. H., Varner, R. N., Kullback, S. On the analysis of multidimensional contingency tables. Journal of the American Statistical Association 66:55, 1971.

5. Kullback, S. Information Theory and Statistics, New York: Wiley, 1959, New York: Dover Publications Inc., 1968.

6. Kullback, S. Minimum discrimination information estimation and application. Invited paper presented to SIXTEENTH CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT AND TESTING, U.S.A. LOGISTICS MANAGEMENT CENTER, FT. LEE, VA. 21 OCTOBER 1970. ARO-D Report 71-3, 1-38. Proceedings of the conference.