The Statistician 32 (1983) 395–403

On an Exploratory Analysis of Contingency Tables

RON S. KENETT

Abstract: An inherent advantage to simple informative graphical displays is that the experience and intuition of the experimenter who collects the data can contribute to the statistician's data analysis. This is in contrast to numerical techniques where the experimenter gets back from the statistician estimates of parameters derived under more or less valid assumptions. We present here some simple data analytic techniques for investigating two-way and three-way contingency tables. These techniques are demonstrated with data collected in an experiment at Bell Laboratories.

1 Introduction

In recent years, a considerable amount of literature was devoted to contingency tables analysis. The list of references in Bishop *et al.* (1975) is indeed impressively long. However, most such studies take a parametric approach without using any graphical displays which have become, in general, an essential ingredient of any statistical data analysis (see Tukey, 1977).

The use of half-normal plots for identifying outlying cells in contingency tables, as a preliminary step in the smoothing and estimation of cell probabilities, is suggested in the logit case by Cox and Lauh (1967) and by Fienberg (1969) for the loglinear model. Other work on the use of graphical displays in the analysis of contingency tables is presented by Cohen (1980) which applies clustering techniques to the rows and columns of a two-way contingency table. Hartigan and Kleiner (1981) suggest various schemes for representing multi-way tables in two-dimensions. For two-way tables they also exhibit residuals graphically, thus identifying possible outlying cells.

In this work we present some graphical and non-parametric data analytical techniques that can be useful in a preliminary analysis of contingency tables. Throughout this paper we use data collected at Bell Laboratories in an experiment on information retrieval from a database. A general description of the data presented in Tables 1 and 2 follows. In a first experiment two methods (method 1 and method 2) are tested on 11 different items (i, e, n, c, d, r, a, b, t, g, h) and the number of errors during these tests are recorded. Initially only the occurrence or non-occurrence of an error is recorded, so that for each of the 11 items we have a 2×2 contingency table with categories method (method 1, method 2) and number-of-errors (0, 1+) as displayed in Table 1. In Table 2 we have data from a second experiment where only six items were tested on four methods but this time a record was kept of the number-of-errors committed in the test so that we have a $6 \times 4 \times 5$ contingency table with categories item (r, a, b, t, g, h), method (method A, method B, method C, method D) and number-of-errors (0, 1, 2, 3, 4+).

† Ron Kenett is manager of mathematical services at TADIRAN, Telecommunications Division, Petah Tiqva, Israel, and adjunct associate professor in the Engineering School of Tel Aviv University. This work was carried out at Bell Laboratories, Piscataway, NJ 08954.

Section II presents a geometric analysis of interactions in two-way tables demonstrated with the data of Table 1. In section III we describe a methodology for analysing non-parametrically and graphically three-way tables. The methodology is applied to the data of Table 2.

2 Analysis of interactions

One purpose of experiment 1 was to investigate qualitatively and quantitatively interactions between method and number-of-errors across the 11 different items. We proceed by presenting various graphical displays of the data and conclude by giving numerical estimates of interactions which have a nice "geometric" interpretation.

Table 1. The data from experiment 1

Method	*Item*										
	i	e	n	c	d	r	a	b	t	g	h
Method 1											
0	27	57	20	7	7	44	57	42	11	30	6
1+	25	40	10	3	1	48	54	41	33	45	34
Method 2											
0	65	109	37	15	9	88	111	84	33	80	20
1+	21	48	15	5	1	63	54	53	36	59	46
Total	118	254	82	30	18	243	276	220	113	214	106

Table 2. The data from experiment 2

Method	*Item*					
	r	a	b	t	g	h
Method A						
0	18	25	14	5	9	3
1	16	12	9	2	6	4
2	5	4	3	7	6	9
3	2	4	2	2	2	2
4+	4	5	6	10	9	6
Method B						
0	26	32	28	6	21	3
1	13	7	3	4	5	5
2	2	6	4	2	7	3
3	2	4	4	1	3	1
4+	4	12	10	5	7	4
Method C						
0	55	67	51	20	42	11
1	20	9	15	8	20	13
2	7	7	2	4	13	6
3	4	6	4	2	3	5
4+	5	9	7	8	5	3
Method D						
0	33	44	33	13	38	9
1	16	9	16	6	5	14
2	4	4	4	4	4	2
3	1	2	2	2	2	2
4+	6	8	3	2	7	1

2.1. The simplex representation

Let $X=(X_1, X_2, X_3, X_4)$ be a four-dimensional frequency vector (i.e. $X_i \geqslant 0$, $i=1, 2, 3, 4$ and $\Sigma_{i=1}^{4} X_i = 1$). The set of end points of all possible such vectors X form (in three dimensions) a pyramid like structure called a simplex. There is a natural one-to-one correspondence between the set of all possible 2×2 standardized contingency tables, with entries that add up to one, and the points of a simplex. This correspondence is established by identifying the entries of such a contingency table (say columnwise) with the components of X (see Fienberg and Gilbert 1970). We exploit this relationship to prepresent graphically in Figure 1 the 2×2 standardized tables corresponding to the 11 items symbolized by their letter.

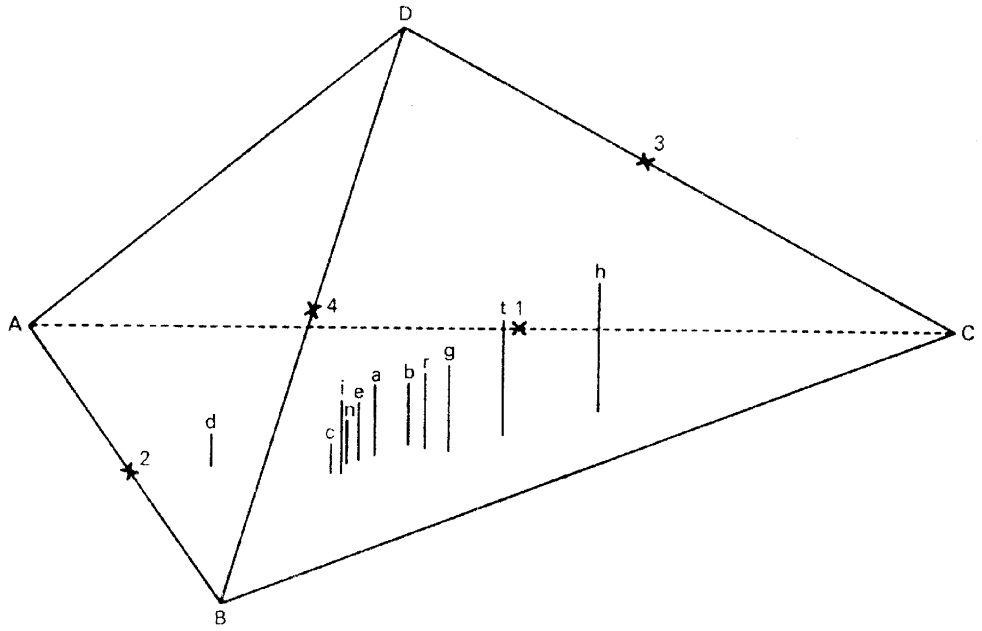

Fig. 1. The simplex representation

On that figure, the entries of any 2×2 table or equivalently the components of any vector X are the lengths of the projections to the four faces of the simplex. We plot on Figure 1 the projections to the lower face corresponding to the values of X_2. To help interpret this figure we add eight points corresponding to the following vectors (tables).

A=(1, 0, 0, 0)	B=(0, 0, 1, 0)
C=(0, 0, 0, 1)	D=(0, 1, 0, 0)
1=(0·5, 0, 0, 0·5)	2=(0·5, 0, 0, 0·5)
3=(0, 0·5, 0, 0·5)	4=(0, 0·5, 0·5, 0)

Point 2 corresponds to a table where every attempt was successful, Point 3 to a table where all attempts result in at least one error so that in both cases there are no interactions between the number of errors made and the method used. On the other hand, Point 1 corresponds to a table where all attempts with method 1 are successful and all method 2 attempts result in at least one error. Point 4 has the same interpretation with method 1 and method 2 interchanged. These two points, therefore, correspond to tables with the strongest interaction since knowledge of the number of errors (0 or 1+) completely characterizes the method used.

Bringing Points 1, 2, 3, and 4 into perspective we can see in Figure 1 that item "d" is the easiest and "h" the most difficult. The remaining items are ordered almost linearly between these two extremes.

Additional insight can be gained by realising that the edge DC consists of all tables with no successes, and edge AB is made up of tables with no errors.

More interpretations on the size of interactions can be derived from Figure 1, but we prefer to make use of the coplanar location of these points and reduce the dimensionality by one. This leads to a second type of graphical displays for 2×2 contingency tables presented next.

2.2 A two-dimensional representation

Let us define

$$D = X_1X_4 - X_2X_3 \tag{1}$$

The quantity D is a measure of interaction that plays an important role in population genetics where it was termed "linkage disequilibrium" by Lewontin and Kojima (1960).

Let

$$\mathrm{f} = X_1 + X_3 \quad \text{and} \quad \mathrm{g} = X_1 + X_2 \tag{2}$$

In the eleven 2×2 tables under investigation f corresponds to the frequency of zero errors and g to the frequency of attempts made with method 1.

It can be easily verified that the components of X are determined by D, f and g through the following set of equations.

$$\begin{aligned} X_1 &= \mathrm{fg} + D \\ X_2 &= (1-\mathrm{f})\mathrm{g} - D \\ X_3 &= \mathrm{f}(1-\mathrm{g}) - D \\ X_4 &= (1-\mathrm{f})(1-\mathrm{g}) + D \end{aligned} \tag{3}$$

or briefly

$$\mathbf{X} = \mathbf{f} \otimes \mathbf{g} + D\mathbf{e} \otimes \mathbf{e} \tag{4}$$

where

$$\begin{aligned} \mathbf{f} &= (\mathrm{f}, 1-\mathrm{f}) \\ \mathbf{g} &= (\mathrm{g}, 1-\mathrm{g}) \\ \mathbf{e} &= (1, -1) \end{aligned}$$

and $\otimes$ is the Kronecker (tensor) product.

Representation (4) is extremely useful because of its intuitive appeal. The first term on the right-hand side of (4) corresponds to a table with no interaction and margins equal to those of the table corresponding to X. The second term is a scalar (D) multiplied by the vector $(1, -1, -1, 1)$, a direction in four space. In other words, the vector X is at a distance D from the independence table $\mathbf{f} \otimes \mathbf{g}$ along the direction $\mathbf{e} \otimes \mathbf{e}$. The direction $\mathbf{e} \otimes \mathbf{e}$ is of special interest since all tables along this direction have the same marginal distributions. The independence table corresponds to the point where $\mathbf{e} \otimes \mathbf{e}$ intersects the surface $D = 0$.

Of particular interest is the case when one of the margins is fixed (say g), since then the inherent three dimensions of X are reduced to two and we can use a two-dimensional plot (say f versus D) instead of the more cumbersome simplex representation. This is applied to the contingency tables under study where g is approximately 0·38 for all 11 items. The results are displayed in Figure 2. When g is not exactly constant, a plot of f versus D corresponds to a projection of the simplex on to two dimensions. The less variable g, the less distortion is induced by that projection.

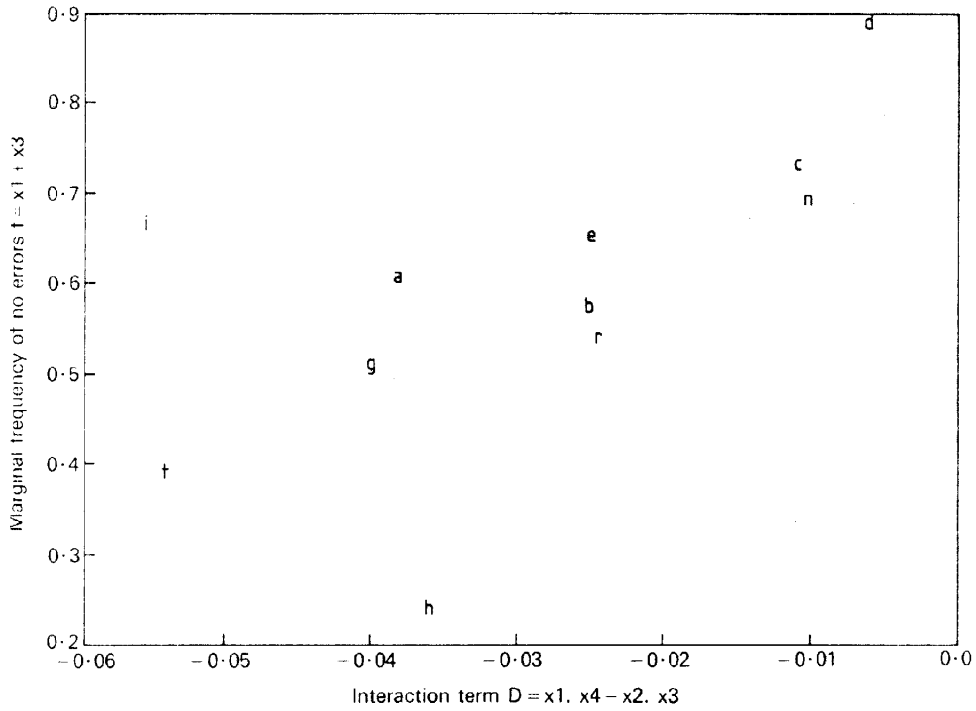

Fig. 2. Tables of method 1 versus method 2

One main observation can be made from Figure 2:

There is a general increasing trend of the items indicating that the more difficult items are associated with higher interactions. In other words, method 1 and method 2 perform similarly on easy items and differently on difficult ones where method 2 is less error prone.

The geometric interpretation of D makes it an appealing measure of interaction. However, it is not standardized with respect to "neighbouring volume" in the simplex. More explicitly, tables corresponding to points near corners will have intrinsically smaller D values than those corresponding to more central points. To correct for this effect we suggest to first compute, D_M, the distance between $\mathbf{f}\otimes\mathbf{g}$, the projection of X on the surface of independence ($D=0$), and the face of the simplex closest to $\mathbf{X}$ along $\mathbf{e}\otimes\mathbf{e}$. We then define the relative disequilibrium $\hat{D}=D/D_M$ and use this as a measure of interaction. The explicit computation of $\hat{D}$ is performed by the following algorithm:

Definition of the relative disequilibrium $\hat{D}$:

If $D>0$

then

if $X_3 < X_2$

then $\hat{D}=\dfrac{D}{D+X_3}$

else $\hat{D}=\dfrac{D}{D+X_2}$

else

if $X_1 < X_4$

then $\hat{D} = \frac{D}{D - X_1}$

else $\hat{D} = \frac{D}{D - X_4}$

In Table 3 we display the $\hat{D}$ values for the 11 tables corresponding to the different items tested. The relative disequilibrium produces a slightly different ordering of the items than D, enhancing the linear increasing trend of Figure 2. The following are properties of $\hat{D}$.

Table 3. Interactions of method and number-of-errors

Item	n	t	i	g	a	r	e	b	d	c	h
D	0·389	0·358	0·267	0·222	0·164	0·120	0·118	0·116	0·100	0·062	0·054
α	0·406	0·364	0·349	0·491	0·513	0·656	0·646	0·646	0·775	0·778	0·811

Properties of $\hat{D}$:

(1) $0 \leqslant \hat{D} \leqslant 1$.
(2) $\hat{D} = 0$ corresponds to tables with independent variables.
(3) Tables with one zero entry or tables with two diagonal zero entries have $\hat{D} = 1$.
(4) Tables with two zero entries on the same row or column or with three zero entries have $\hat{D} = 0$.

For comparison purposes we also compute the cross product ratio, $\alpha = X_1X_4/X_2X_3$, which is a more traditional measure of interaction (see Bishop, *et al.*, 1975). $\hat{D}$ has an advantage over α in that it does not "blow up" for small X_2 and X_3. This permits a more meaningful comparison of interactions between different contingency tables. For an illustration of the difference between $\hat{D}$ and α consider the following 2×2 contingency tables A and B:

A		B	
0·50	0·45	0·50	0·40
0·005	0·045	0·0101	0·0899

Both tables yield the same $\hat{D}$ value of 0·802 but table A has $\alpha = 10$ and table B has $\alpha = 11{\cdot}126$. From the definition of $\hat{D}$ we know that the points corresponding to tables A and B within the three-dimensional simplex are at the same relative distance from points corresponding to tables exhibiting independence with the same margins as in A and B. However, the α measure indicates more intraction in table B than in table A.

The classical measures of association found in the literature (see Bishop *et al.*, 1975, p. 374) are direct functions of either α or of the correlation coefficient, ρ. The relative disequilibrium is a direct function of neither as can be seen in expressions (5) and (6).

For $D > 0$, $X_3 < X_2$ we have that:

$$\hat{D} = \frac{X_3(\alpha - 1)}{1 + X_3(\alpha - 1)} \tag{5}$$

and also

$$\hat{D}=\frac{f(1-f)g(1-g)\rho}{X_3+f(1-f)g(1-g)\rho} \tag{6}$$

Let x_i and D_0 be the observed values of X_i ($i=1, \ldots, 4$) and D respectively then

$$\hat{D}_0=\frac{D_0}{D_0+x_3} \tag{7}$$

is the maximum likelihood estimate of $\hat{D}$ (when $D>0$ and $X_3<X_2$) under the multinomial sampling model and under the product binomial model where either the observed row or the observed column totals are fixed.

Using results of Goodman and Kruskal (1972) we can show that under both types of sampling models, σ^2, the variance of the asymptotic distribution of $\sqrt{n}\,(\hat{D}_0-\hat{D})$ is given by

$$\sigma^2=\frac{X_3^2}{(D+X_3)^4}\{(X_2-X_3)D+X_1X_4\} \tag{8}$$

which we estimate by replacing X_i by x_i ($i=1, \ldots, 4$) in (8).

In particular, for $D=0$, we have that

$$\sigma^2=\frac{X_2}{X_3} \tag{9}$$

Bishop *et al.* (1975, p. 377) give under the same sampling models, the asymptotic variance of α, σ^2 ($\hat{\alpha}$), which becomes for $D=0$

$$\sigma^2(\hat{\alpha})=\frac{1}{X_2X_3} \tag{10}$$

Comparing (9) and (10) yields that for tables with independence the variance of $\hat{D}_0$ is smaller than the variance of the observed value of α, $\hat{\alpha}$. For a further comparison of $\hat{D}$ and α see Kenett (1980).

3 Analysis of similarities and differences between methods

In experiment 2 the experimenter was interested in whether the four methods had similar number-of-errors distributions and if not which method was different. The data we analyse consists of a $6\times4\times5$ contingency table. The non-parametric methodology we use is a new approach for an exploratory analysis of such tables.

3.1. The methodology

Our analysis goal is to determine how different the four methods are from each other. The methodology we use is looking at distances between number-of-errors distributions and results in partial orderings of the methods on the basis of such distances.

The technique we propose involves two decisions

(a) Choose a *base* or *standard number-of-errors distribution* to which the different methods will be compared.
(b) Choose a measure of distance between distributions. We decided here on the χ^2 statistic.

Following these choices the analysis involves three steps.

(i) For each item and for each method compute the distance from the number-of-errors frequency vector to the corresponding standard frequency vector. For each pair of

methods, calculate the number of items, out of six, for which method 1 is closer to the standard than method 2.

(ii) Since there are six items the most extreme comparison is 6:0. Assuming that the six items are a random sample from a large collection of testable items we invoke the sign test to determine whether a comparison between two methods is significant or not. A comparison of 5:1 or a more extreme one results in a P-value of 13 per cent. Since we are after qualitative conclusions only, we will consider such comparisons significant. We could have also used more powerful rank order tests at the expense of simplicity in the technique.

(iii) From the above comparisons determine a consistent partial ordering of the methods (with respect to distances from the chosen standard). Methods are arranged as follows. For significant comparisons we have them one next to each other (the larger on the right), for non-significant comparisons we put them one on top of the other (cf. Figures 3 and 4).

This technique is applied repeatedly in the remainder of this section for various choices of standards. These different choices will enable us to answer several questions regarding similarities and differences of the methods under investigation.

3.2. Distances from a synthetic method producing the average number of errors

Our first choice for a standard is the average number of errors (over the four methods), per item, weighted by sample sizes. Computing χ^2 distances of the methods from this "average" method produces the numbers in Table 4.

Table 4. Distances from average method

Item	*Method*			
	A	B	C	D
r	2·22	0·48	0·93	0·65
a	2·78	1·82	1·41	0·87
b	2·29	5·21	1·52	2·55
t	5·01	0·49	0·89	2·44
g	5·38	0·94	2·81	3·89
h	4·96	1·08	0·82	4·18

These numbers produce comparisons which result in the partial ordering displayed in Figure 3. It is immediately apparent that method A stands out as most distant from the average indicating that this method performed differently from the rest.

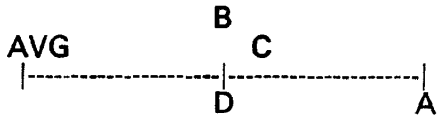

Fig. 3. Distances from average method

3.3. Distances from the separate methods as standards

Another possible choice is to use each separate method as a standard. This produces the partial orderings of Figure 4. Note that two methods on top of each other are not similar; it is only their position relative to the standard that is similar. One two-dimensional arrange-

ment of the four methods is consistent with the partial orderings of Figure 4. This is presented in Figure 5. Again method A stands out as yielding a different performance. The experimenters were able in this case to interpret the horizontal and vertical axes of Figure 5, on the basis of the actual technical differences between the information retrieval methods.

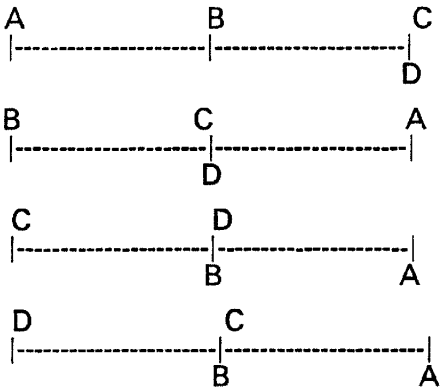

Fig. 4. Distances from separate methods

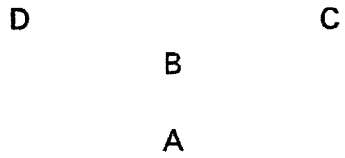

Fig. 5. Two-dimensional representation of methods consistent with partial orderings

References

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, Mass.

Cohen, A. (1980). On the graphical display of the significant components in two-way contingency tables. *Commun. Statist – Theor. Methods* A9 (10) 1025–41.

Cox, D. R. and Lauh, E. (1967). A note on the graphical analysis of multidimensional contingency tables. *Technometrics* **9**, 481–8.

Fienberg, S. E. (1969). Preliminary graphical analysis and quasi-independence for two-way contingency tables. *Appled Statistics* **18**, 153–68.

Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association* **65**, 694–701.

Goodman, L. A. and Kruskal, W. H. (1972). Measures of association for cross classifications. IV: Simplification of asymptotic variances. *Journal of the American Statistical Association* **67**, 415–21.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. *Computer Science and Statistics: 13th Symposium on the Interface*, Pittsburgh, Penn.

Kenett, R. S. (1980). Relative disequilibrium in contingency tables. *Technical Report No. 596*, Dept. of Statistics, University of Wisconsin, Madison.

Lewontin, R. C. and Kojima, K. I. (1960). The evolutionary dynamics of complex polymorphisms, *Evolution* **14**, 458–72.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.