

Tim Futing Liao  
Department of Sociology  
University of Essex  
Colchester C04 3SQ  
United Kingdom  
44 1206 872663 (phone)  
44 1206 873410 (fax)  
tfliao@essex.ac.uk  
Word count: 1,289

**Contingency Table.** A contingency table is a statistical table classifying observed data frequencies according to the categories in two or more variables. A table formed by the cross-classification of two variables is called a two-way contingency table (for an example, see Table 1), a table of three variables is termed a three-way contingency table, and in general a table of three or more variables is known as a multiway contingency table. The analysis of multiway contingency tables is sometimes known as multivariate contingency analysis.

All cells in a table may not have observed counts. Sometimes cells in a contingency table contain zero counts, known as empty cells. A zero count may be generated by two distinctive mechanisms. If sample size causes a zero count, say, Chinese American farmers in Illinois, it is known as a sampling zero. If for a cell it is theoretically impossible to have observations, say, male (contraceptive) diaphragm users, it is called a structural zero. Contingency tables with at least one structural zero are termed incomplete tables. When many cells in a contingency table have lower observed frequencies, be they zero or not, such a table is known as a sparse table.

## HISTORICAL DEVELOPMENT

Early work on CATEGORICAL DATA ANALYSIS in the beginning of the 20th century was primarily concerned with the analysis of contingency tables. The well-known debate between Karl Pearson and G. Udny Yule sparked interest in contingency table analysis. How would we

analyze a  $2 \times 2$  contingency table? Pearson was a firm believer in continuous bivariate distributions underlying the observed counts in the cross-classified table; Yule was of the opinion that variables containing discrete categories such as inoculation versus no inoculation would be best treated as discrete without assuming underlying distributions. Pearson's contingency coefficient was calculated based on the approximated underlying correlation of bivariate normal distributions collapsed into discrete cross-classifications, while Yule's  $Q$  was computed using a function of the ODDS RATIO of the observed discrete data directly.

Contributions throughout the 20th century can often be traced by the names of those inventing the statistics or tests for contingency table analysis. Pearson chi-square test, Yule's  $Q$ , Fisher's exact test, Kendall's tau, Cramér's  $V$ , Goodman and Kruskal's tau, and the Cochran-Mantel-Haenszel test (or the Mantel-Haenszel test) are just some examples. These statistics are commonly included in today's statistical software packages (e.g., proc freq in SAS generates most of the above).

Contingency table analysis can be viewed as the foundation for contemporary categorical data analysis. We may see the series of developments in LOGLINEAR MODELING in the latter half of the 20th century as refinements in analyzing contingency tables. Similarly, the advancement in latent trait and latent class analysis is inseparable from the statistical foundation of contingency table analysis.

#### EXAMPLE

We present a classic data table, the Midtown Manhattan data of mental health and parents' socioeconomic status (SES), analyzed by L. Goodman, C. C. Clogg, and many others (see Table 1).

[TABLE 1 ABOUT HERE]

To analyze contingency tables, we need expected frequencies. Let  $F_{ij}$  indicate the expected value for the observed frequency  $f_{ij}$  in row  $i$  and column  $j$  in the table. The  $F_{ij}$  under the model of independence (no association between parents' SES and mental status) is given by

$$F_{ij} = \frac{f_{i+}f_{+j}}{f_{++}},$$

where  $f_{i+}$  gives the column total for the  $i$ th row,  $f_{+j}$  gives the row total for the  $j$ th column, and  $f_{++}$  gives the grand total of the entire table. For example,  $f_{1+} = 64 + 94 + 58 + 46 = 262$ ,  $f_{+1} = 64 + 57 + 57 + 36 + 21 = 307$ , and  $F_{11} = (262 \times 307)/1,660 = 48.5$ . We compute the expected values for the entire table accordingly, and present them in parentheses in Table 1.

The null hypothesis is that the two variables are unrelated or independent of each other. To test the hypothesis, we compute the PEARSON  $\chi^2$  STATISTIC and the LIKELIHOOD RATIO STATISTIC  $L^2$  (or  $G^2$ ):

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \quad \text{and} \quad L^2 = 2 \sum_i \sum_j f_{ij} \log \left( \frac{f_{ij}}{F_{ij}} \right).$$

Applying these formulae, we obtain a Pearson  $\chi^2$  statistic of 45.99 and an  $L^2$  of 47.42. With degrees of freedom of 15 (= the number of rows minus 1 times the number of columns minus 1), we reject the null hypothesis of independence at any conventional significance level and conclude that mental health status and parents SES are associated.

The model of independence can also be understood from the ODDS RATIOS computed from the expected frequencies. One may calculate the cross-product ratio for every adjacent  $2 \times 2$  table contained in the contingency table to understand the independence assumption. That is,  $(48.5 \times 88.8)/(95.0 \times 45.3) \approx 1.0$ ,  $(95.0 \times 53.4)/(57.1 \times 88.8) \approx 1.0$ , . . . ,  $(57.8 \times 50.9)/62.1 \times$

47.3)  $\approx 1.0$ . One quickly discovers that all such odds ratios are approximately one! An odds ratio of one indicates that there is no relationship between the two variables forming the table.

Both variables are ordinal in nature, but that information is not considered in the current analysis. Further fine-tuning via LOGLINEAR MODELING will be necessary to analyze the type of association between the two variables.

## COLLAPSIBILITY OF CONTINGENCY TABLES

A prominent issue in contingency table analysis is that of collapsibility. Simply put, a contingency table is collapsible if the fit of a particular statistical model is not significantly affected by either combining its dimensions (i.e., removing variables) or combining some categories in a variable (i.e., reducing the number of response categories) or both. In the current example, someone may suspect that there exists little difference between the mild and the moderate symptoms, noticing the differences between the observed and the expected counts in the two middle columns in Table 1 are uniformly below 1. Table 2 presents the observed data (and the expected counts) with those two categories collapsed.

[TABLE 2 ABOUT HERE]

From Table 2 we arrive at a new set of test statistics, a Pearson  $\chi^2$  of 43.51 and an  $L^2$  of 44.94, with 10 degrees of freedom based on 6 rows and 3 columns (a still significant result at the .05 level). A LIKELIHOOD RATIO TEST based on either the Pearson  $\chi^2$  statistics or the  $L^2$ s by taking the difference of the new and the old values gives a test statistic of 2.47 with 5 (= 15 – 10) degrees of freedom. This result, which follows a  $\chi^2$  distribution, is highly insignificant, suggesting that the two middle symptom categories can be collapsed without losing much

information. The reader is encouraged to find out whether the parents' SES categories can be collapsed.

TIM FUTING LIAO

*References*

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.

Fienberg, S. E. (1981). *The Analysis of Cross-Classified Categorical Data* (2nd ed.). Cambridge, MA: MIT Press.

Rudas, T. (1997). *Odds Ratios in the Analysis of Contingency Tables*. Thousand Oaks, CA: Sage.

Table 1: A Contingency Table of Parents' Socioeconomic Status (SES) and Mental Health Status

(expected values in parentheses)

Mental Health Status				
Parents' SES	Well	Mild Symptom	Moderate Symptom	Impaired
A (high)	64 (48.5)	94 (95.0)	58 (57.1)	46 (61.4)
B	57 (45.3)	94 (88.8)	54 (53.4)	40 (57.4)
C	57 (53.1)	105 (104.1)	65 (62.6)	60 (67.3)
D	72 (71.0)	141 (139.3)	77 (83.7)	94 (90.0)
E	36 (49.0)	97 (96.1)	54 (57.8)	78 (62.1)
F (low)	21 (40.1)	71 (78.7)	54 (47.3)	71 (50.9)

Table 2: A Contingency Table of Parents' Socioeconomic Status (SES) and Mental Health Status with Collapsed Middle Symptom Categories (expected values in parentheses)

Mental Health Status			
Parents' SES	Well	Mild or Moderate Symptom	Impaired
A (high)	64 (48.5)	152 (95.0)	46 (61.4)
B	57 (45.3)	148 (88.8)	40 (57.4)
C	57 (53.1)	170 (104.1)	60 (67.3)
D	72 (71.0)	218 (139.3)	94 (90.0)
E	36 (49.0)	151 (96.1)	78 (62.1)
F (low)	21 (40.1)	125 (78.7)	71 (50.9)