A Multivariate Perspective on the Analysis of Categorical Data

Rebecca Zwick
Educational Testing Service

Elliot M. Cramer
University of North Carolina at Chapel Hill

Psychological research often involves analysis of an $I \times J$ contingency table consisting of the responses of J groups of individuals on a criterion variable with I nominal categories. The conventional statistical approach for comparing responses across groups is the Pearson chi-square test. Alternatively, this analysis can be viewed as a multivariate analysis of variance with binary dependent variables, a canonical correlation analysis with two sets of binary variables, or a form of correspondence analysis. Although these analysis approaches stem from different traditions, they produce equivalent results when applied to an $I \times J$ table.

In psychological research, an investigator may often wish to compare several groups of individuals on a categorical response variable, such as a measure of attitude. The conventional statistical test in this case is the Pearson chi-square. In this paper, the relations between the Pearson chi-square and several multivariate statistical techniques are illustrated. Specifically, it is shown that the chi-square test for an $I \times J$ contingency table can be viewed as a one-way multivariate analysis of variance (MANOVA) with J groups and $I - 1$ binary dependent variables. This MANOVA approach is, in turn, computationally identical to the canonical analysis of contingency tables, as presented by Kendall and Stuart (1967, pp. 569–574; see also Isaac & Milligan, 1983). Correspondence analysis, an approach that is viewed primarily as a scaling method, yields equivalent results as well.

The following example from Marascuilo and Levin (1983, p. 452) is used throughout for illustration. Table 1 presents data from a hypothetical nationwide survey in which 500 randomly selected men were asked the question, ''Does a woman have the right to decide whether an unwanted birth can be terminated during the first three months of pregnancy?'' The response choices were ''Yes'', ''No'', and ''No opinion''. There were four groups of respondents: Catholic, Protestant, Jewish, and Other. It is of interest to determine whether the four religious groups differ in terms of the proportions of ''Yes'', ''No'', and ''No opinion'' responses.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 2, June 1986, pp. 141–145
© Copyright 1986 Applied Psychological Measurement Inc.
0146-6216/86/020141-05$1.50

Conventional Pearson Chi-Square Approach

To test the hypothesis that there are no response differences across religious groups, a researcher can compute the ordinary Pearson chi-square statistic. A convenient computational form is

$$X^2 = N \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{f_{ij}^2}{f_{i+} f_{+j}} - 1 \right) \quad , \qquad (1)$$

where f_{ij} is the frequency of the ith response ($i = 1,2, \ldots, I$) in the jth group ($j = 1,2, \ldots, J$),

f_{i+} is the marginal frequency for the ith response category,

Table 1
Responses of 500 Men to Abortion Survey

	Religion				
Response	Catholic	Protestant	Jewish	Other	Total
Yes	76	115	41	77	309
No	64	82	8	12	166
No opinion	11	6	2	6	25
Total	151	203	51	95	500

Note: From *Multivariate Statistics in the Social Sciences* (p. 452) by L. A. Marascuilo and J. R. Levin (1983). Monterey, CA: Brooks/Cole. Reprinted by permission.

f_{+j} is the marginal frequency for the jth group, and

N is the total sample size.

Equation 1 is equivalent to the familiar form,

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} [(o_{ij} - e_{ij})^2/e_{ij}] \quad , \tag{2}$$

where o_{ij} and e_{ij} are the observed and estimated expected values, respectively, for the i,j cell. The obtained X^2 value is compared to the $100(1 - \alpha)$ percentile of the χ^2 distribution with $(I - 1)(J - 1)$ degrees of freedom. In the present example, $X^2 = 40.17$, which exceeds $\chi^2_{6:95} = 12.59$, the critical value for $\alpha = .05$. Therefore, the null hypothesis is rejected: There is reason to believe that the probabilities of ''Yes'', ''No'', and ''No opinion'' responses to the abortion survey question differ across the four religious groups. Post hoc contrasts could be performed to explore the nature of these differences.

MANOVA Approach

In order to treat the contingency table analysis as a MANOVA, it is necessary to express the dependent variable in terms of $I - 1$ indicator variables, denoted here as I_1 and I_2. I_1 can be viewed as an indicator for a ''Yes'' response. Respondents who answered ''Yes'' are given a score of 1 on I_1; all other respondents receive a score of 0. Similarly, I_2 is an indicator for a ''No'' response. The result of this coding is that respondents who answered ''Yes'' receive the scores $I_1 = 1$, $I_2 = 0$; those who answered ''No'' receive the scores $I_1 = 0$, $I_2 = 1$; and those who stated they had ''No opinion'' are scored $I_1 = 0$, $I_2 = 0$.

It is now possible to compute hypothesis ($\mathbf{H}$) and total ($\mathbf{T}$) sums of squares and cross-products matrices, as is done in ordinary MANOVA. These matrices are analogous to the between and total sums of squares in univariate analysis of variance (ANOVA). That is, the diagonal elements h_{11} and h_{22} of $\mathbf{H}$ are the between-group sums of squares for the indicator variables I_1 and I_2, respectively, and the off-diagonal element $h_{12} = h_{21}$ is the between-group sum of cross-products of I_1 and I_2. The elements of $\mathbf{T}$ are defined in a corresponding fashion. In the present example,

$$\mathbf{H} = \begin{bmatrix} 7.81 & -7.77 \\ -7.77 & 7.91 \end{bmatrix} \quad , \tag{3}$$

and

$$\mathbf{T} = \begin{bmatrix} 118.04 & -102.59 \\ -102.59 & 110.89 \end{bmatrix} \quad . \tag{4}$$

(All figures are rounded to two decimal places throughout the paper.) As shown by Kshirsagar (1972, p. 383), the Pearson chi-square statistic can now be computed as $X^2 = NV$, where $V =$

$\mathrm{tr}(\mathbf{T}^{-1}\mathbf{H})$ is the Pillai-Bartlett statistic (Bartlett, 1939; Pillai, 1955), a well-known MANOVA criterion. The Pearson chi-square is expressed in a similar form by Koch and Bhapkar (1982, p. 448). Here,

$$V = \mathrm{tr}(\mathbf{T}^{-1}\mathbf{H}) = .03 + .05 = .08 \quad . \tag{5}$$

An equivalent means of obtaining V is through solution of the equation

$$\mathbf{Hu} = \theta\mathbf{Tu} \quad . \tag{6}$$

The eigenvalues θ_i of $\mathbf{T}^{-1}\mathbf{H}$ are .07 and .01, which, of course, sum to the trace;

$$\begin{aligned} X^2 = NV &= N \sum_{i=1}^{S} \theta_i \\ &= 500\,(.08) = 40.17 \quad , \end{aligned} \tag{7}$$

where $S = \min(I - 1, J - 1)$ is the number of non-zero eigenvalues.

Canonical Analysis of Contingency Tables

The application of canonical correlation analysis to contingency tables is treated in detail in Kendall and Stuart (1967, pp. 568–573). One way to conduct a canonical analysis of a contingency table is to create $I - 1$ indicator variables, corresponding to the I row categories, and $J - 1$ indicator variables, corresponding to the J column categories. In the present example, two indicator variables can be created to represent the three possible responses to the survey question, as described above. Similarly, religious affiliation can be represented as three indicator variables. Then, an ordinary canonical correlation analysis is performed, relating the $I - 1$ variables representing the row categories (survey response) to the $J - 1$ variables representing the column categories (religion). As in any canonical correlation problem, the equation to be solved is

$$(\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy})\mathbf{u} = \theta\mathbf{R}_{yy}\mathbf{u} \quad . \tag{8}$$

In this example, $\mathbf{R}_{yy}$ is the $(I - 1) \times (I - 1)$ correlation matrix of the indicator variables for survey response, $\mathbf{R}_{xx}$ is the $(J - 1) \times (J - 1)$ correlation matrix of the indicator variables for religion, and $\mathbf{R}_{yx} = \mathbf{R}'_{xy}$ is the $(I - 1) \times (J - 1)$ matrix of cross-correlations of the indicator variables representing survey response with those representing religion. The $S = \min(I - 1, J - 1)$ non-zero eigenvalues (θ_i) of Equation 8 are the squared canonical correlations between two sets of variables. The matrix $\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$ is equal to $\mathbf{T}^{-1}\mathbf{H}$; thus, Equation 8 is identical to Equation 6. Therefore, the canonical approach again yields

$$X^2 = N \sum_{i=1}^{S} \theta_i = 40.17 \quad . \tag{9}$$

Kendall and Stuart showed that an equivalent means of conducting this analysis involves the eigen equation

$$\mathbf{FF'v} = \theta\mathbf{v} \quad , \tag{10}$$

where $\mathbf{F}$ is an $I \times J$ matrix with elements $f_{ij}/(f_{i+}\,f_{+j})^{1/2}$. When this method is used, $\theta_i = 1$ is always a solution. This eigenvalue, which is of no interpretive value, is discarded. The remaining S non-zero eigenvalues are equal to those obtained through solution of Equations 6 or 8. Therefore, X^2 can be expressed as $N[\mathrm{tr}(\mathbf{FF'}) - 1]$, which is equivalent to Equation 1.

Correspondence Analysis

A related technique for the analysis of two-way contingency tables is correspondence analysis. A review of the multiple origins of this method and its applications to the case of multidimensional contingency tables is presented by Tenenhaus and Young (1985; see also Greenacre, 1984; Hill, 1974, 1982). A primary goal of correspondence analysis is the derivation of optimal scale values for the row and column categories of a contingency table.

Before discussing the method in detail, it should be pointed out that the MANOVA and canonical correlation approaches described above can also be viewed in a scaling context (e.g., McKeon, 1964). For instance, in an experiment with J groups and a categorical response variable, scale values could be assigned to each response category and a univariate ANOVA performed on these values. The scale values that would maximize the ANOVA F statistic are the elements of the eigenvector corresponding to the largest eigenvalue obtained via the MANOVA approach; that is, the first discriminant function.

This vector has $I - 1$ elements; the Ith category receives a scale value of 0. The optimal scale values for the independent variable, religion, could also be obtained via the MANOVA approach by interchanging the role of the independent and dependent variables, computing new $\mathbb{H}$ and $\mathbb{T}$ matrices, and again obtaining the first discriminant function. A researcher might wish to conduct a MANOVA in a preliminary study to obtain optimal scale values for use in univariate analyses to be performed in later studies.

Although the derivation of optimal scale values is rarely mentioned as a primary goal of MANOVA, it is often one of the stated purposes for performing a canonical analysis of a contingency table (e.g., Kendall & Stuart, 1967). The canonical approach is typically applied in situations in which it is considered desirable to seek optimal scale values for both the row and column categories. From a statistical perspective, however, the canonical problem is identical to the MANOVA approach, because it is necessary to scale only one of the two variables in order to maximize the relationship between them. If Equation 8 is used to perform the canonical analysis, the eigenvector associated with the largest eigenvalue is the vector of optimal scale values for the rows of the contingency table. As in the MANOVA approach, the vector has only $I - 1$ elements; the Ith scale value is equal to 0. Scale values for the columns can be obtained by substituting $\mathbb{R}_{xy}\mathbb{R}_{yy}^{-1}\mathbb{R}_{yx}$ for $\mathbb{R}_{yx}\mathbb{R}_{xx}^{-1}\mathbb{R}_{xy}$ and $\mathbb{R}_{xx}$ for $\mathbb{R}_{yy}$ in Equation 8. A value of 0 is assigned to the Jth column category.

For purposes of obtaining scale values, the Kendall and Stuart algorithm is less convenient than Equation 8. In their method, the eigenvectors ($\mathbb{v}$) corresponding to the row categories are obtained from Equation 10; the eigenvectors ($\mathbb{v}^*$) associated with the column categories are obtained by substituting $\mathbb{F}'\mathbb{F}$ for $\mathbb{F}\mathbb{F}'$ in the equation. The eigenvectors corresponding to the solution $\theta_i = 1$ are discarded. The eigenvectors $\mathbb{v}_1$ and $\mathbb{v}_1^*$ corresponding to the largest non-trivial eigenvalue must then be rescaled to obtain the optimal scale values for row and column categories, respectively (see Kendall & Stuart, p. 571).

One of several variants of correspondence analysis, which Hill (1974) called first-order correspondence analysis, is illustrated here for the abortion survey example. The analysis was conducted using a modification of a SAS macro by Young and Sarle (personal communication, 1983). First-order correspondence analysis requires that an indicator matrix, denoted here as $\mathbb{A}$, be constructed. The row dimension of $\mathbb{A}$ is equal to the number of respondents (here, 500). The column dimension is $I + J$, which, in this example, is equal to 7. A respondent receives a 1 in the columns of $\mathbb{A}$ that correspond to the row and column of the original contingency table in which that respondent falls. All other values for the respondent are 0. For example, if the columns of $\mathbb{A}$ are denoted as I_1, I_2, I_3, J_1, J_2, J_3, and J_4, the 41 respondents in the 1,3 cell of Table 1 each have 1 in column I_1 and J_3 and 0 elsewhere.

The matrix $\mathbb{A}$ is then transformed as follows:

$$\mathbb{A}^* = (\mathbb{R}^*)^{-1/2}\,\mathbb{A}(\mathbb{C}^*)^{-1/2} \quad , \tag{11}$$

where $(\mathbb{R}^*)^{-1/2}$ is a diagonal matrix containing the reciprocal square roots of the row totals of $\mathbb{A}$, and $(\mathbb{C}^*)^{-1/2}$ is a diagonal matrix containing the reciprocal square roots of the column totals of $\mathbb{A}$. Following this, a singular value decomposition of $\mathbb{A}^*$ (see Strang, 1980, p. 142) is performed to obtain the eigenvectors of $\mathbb{A}^{*\prime}\mathbb{A}^*$, the eigenvectors of $\mathbb{A}^*\mathbb{A}^{*\prime}$, and the (common) eigenvalues of $\mathbb{A}^{*\prime}\mathbb{A}^*$ and $\mathbb{A}^*\mathbb{A}^{*\prime}$. Note that $\mathbb{A}^{*\prime}\mathbb{A}^*$ can be represented as

$$\mathbb{A}^{*\prime}\mathbb{A}^* = 1/2 \left[\begin{array}{c|c} \mathbb{I} & \mathbb{F}\mathbb{F}' \\ \hline \mathbb{F}'\mathbb{F} & \mathbb{I} \end{array}\right] \quad , \tag{12}$$

where $\mathbb{F}$ is defined as in Equation 10. It is not surprising, then, that the eigenstructures of $\mathbb{A}^{*\prime}\mathbb{A}^*$ and $\mathbb{F}\mathbb{F}'$ are related in a specifiable manner. The number of non-zero eigenvalues of $\mathbb{A}^{*\prime}\mathbb{A}^*$ (or $\mathbb{A}^*\mathbb{A}^{*\prime}$) is equal to $I + J - 1$ for a two-way table. As in Equation 10, one eigenvalue is equal to 1. The remaining non-zero eigenvalues (λ_i) of $\mathbb{A}^{*\prime}\mathbb{A}^*$ are related to the S remaining non-zero eigenvalues θ_i of $\mathbb{F}\mathbb{F}'$ as follows: For each θ_i, first-order correspondence analysis yields two solutions, $\lambda_i = (1 + \sqrt{\theta_i})/2$ and $\lambda_{i'} = (1 - \sqrt{\theta_i})/2$ (Hill, 1974, p. 347). In addition, for each zero eigenvalue of $\mathbb{F}\mathbb{F}'$, first-order correspondence analysis yields an eigenvalue of $(1 + \sqrt{0})/2 = (1 - \sqrt{0})/2 = .50$. (In general, there will be $|J - I|$ such eigenvalues.)

In the present example, the eigenvalues that result from application of Equation 10 are 1, .07, .01, and 0. The first-order correspondence analysis yields the eigenvalues 1, $(1 + \sqrt{.07})/2 = .63$, $(1 - \sqrt{.07})/2 = .37$, $(1 + \sqrt{.01})/2 = .54$, $(1 - \sqrt{.01})/2 = .46$, and $(1 + \sqrt{0})/2 = (1 - \sqrt{0})/2 = .50$. (The eigenvalue of 1 can be eliminated by centering each column of $\mathbf{A}^*$.)

The $(I + J)$–dimensional eigenvector $\mathbf{w}_1$ associated with the largest non-trivial eigenvalue of $\mathbf{A}^{*\prime}\mathbf{A}^*$ is related to the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_1^*$ of Equation 10 as follows:

$$\mathbf{w}_1 = k \left[\begin{array}{c} \mathbf{v}_1 \\ \hline \mathbf{v}_1^* \end{array}\right] , \tag{13}$$

where k is a constant of proportionality. Like $\mathbf{v}_1$ and $\mathbf{v}_1^*$, $\mathbf{w}_1$ must be rescaled to obtain the optimal scale values.

In addition to producing optimal scale values for row and column categories of a contingency table, correspondence analysis is also viewed as a means of obtaining optimal scale values for individuals (Tenenhaus & Young, 1985), which are derived from the N–dimensional eigenvectors of $\mathbf{A}^*\mathbf{A}^{*\prime}$. By making use of the relation between the eigenstructures of $\mathbf{A}^{*\prime}\mathbf{A}^*$ and $\mathbf{F}\mathbf{F}'$ and the principles of singular value decomposition, however, the eigenvectors of $\mathbf{A}^*\mathbf{A}^{*\prime}$ can be derived from the eigenvectors ($\mathbf{v}$) of Equation 10. Therefore, in the case of a two-way contingency table, correspondence analysis does not provide any information that could not be obtained by performing a canonical analysis. An advantage of correspondence analysis is that it can be generalized to the case of multidimensional contingency tables simply by expanding the number of columns of $\mathbf{A}$ (see Skinner & Shew, 1982; Tenenhaus & Young, 1985).

References

Bartlett, M. (1939). A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society, 35,* 180–185.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Hill, M. O. (1974). Correspondence analysis: A neglected multivariate method. *Applied Statistics, 23,* 340–354.

Hill, M. O. (1982). Correspondence analysis. In Kotz, S., & Johnson, N. L. (Eds.), *Encyclopedia of statistical sciences* (Vol. 2, pp. 204–210). New York: Wiley.

Isaac, P. D., & Milligan, G. W. (1983). A comment on the use of canonical correlation in the analysis of contingency tables. *Psychological Bulletin, 93,* 378–381.

Kendall, M. G., & Stuart, A. (1967). *The advanced theory of statistics* (2nd ed., Vol. 2). London: Charles Griffin.

Koch, G. G., & Bhapkar, V. P. (1982). Chi-square tests. In Kotz, S., & Johnson, N. L. (Eds.), *Encyclopedia of statistical sciences* (Vol. 1, pp. 442–457). New York: Wiley.

Kshirsagar, A. M. (1972). *Multivariate analysis.* New York: Marcel Dekker.

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences.* Monterey CA: Brooks/Cole.

McKeon, J. J. (1964). Canonical analysis: Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory. *Psychometric Monograph* (No. 13).

Pillai, K. (1955). Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics, 26,* 117–121.

Skinner, H. A., & Shew, W. (1982). Dimensional analysis of rank-order and categorical data. *Applied Psychological Measurement, 6,* 41–45.

Strang, G. (1980). *Linear algebra and its applications.* New York: Academic Press.

Tenenhaus, M., & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika, 50,* 91–119.

Acknowledgments

The authors thank Forrest W. Young for his help on the section about correspondence analysis, and Al Beaton and Paul Holland for their comments on the manuscript. This investigation was supported in part by a National Research Service Award (No. 5-T32-MH15745) from the National Institute of Mental Health to the University of North Carolina at Chapel Hill, where the first author was a postdoctoral fellow.

Author's Address

Send requests for reprints or further information to Rebecca Zwick, Educational Testing Service, Princeton NJ 08541.