Set Correlation and Contingency Tables

Jacob Cohen
New York University

Set correlation is a realization of the general multivariate linear model, can be viewed as a multivariate generalization of multiple correlation analysis, and may be employed in the analysis of multivariate data in any form. Set correlation supplements the four methods for analyzing two-way contingency tables described by Zwick and Cramer (1986), and its application to their example is illustrated. It gives the same results for the overall association, and in addition, by the use of nominal scale coding and partialling, it assesses specific hypotheses about the details of the association. Set correlation includes measures of strength of association (including correlations and proportions of variance), significance tests and estimation, power analysis, and computer programs to implement the calculations. *Index terms: canonical analysis, contingency table analysis, correspondence analysis, general multivariate linear model, multivariate analysis of variance, Pearson chi-square, set correlation.*

Zwick and Cramer (1986) showed how the analysis of two-way contingency tables may be accomplished by four methods stemming from different traditions: conventional Pearson chi-square, multivariate analysis of variance (MANOVA), canonical analysis, and correspondence analysis. They explicated the relationship among the methods and showed that they produce equivalent results. The purpose of the present article is to offer set correlation as yet another option for such analyses and to show that its generality results in greater flexibility and information yield, so that it may often be preferred to other methods.

Set Correlation

Set correlation (SC) is a realization of the general multivariate linear model and thus a natural generalization of simple and multiple correlation (Cohen, 1982; reprinted as Appendix 4 of Cohen & Cohen, 1983). In its fixed-model form, it generalizes univariate simple and multiple regression to their multivariate analogues. SC is a multivariate generalization of multiple regression/correlation (MRC) and a general data-analytic method (Cohen & Cohen, 1983; Pedhazur, 1982). It is therefore a general scheme for studying the relationship between two *sets* of variables, X and Y, containing any number of variables (k_X,k_Y).

MRC applications have shown that any information can be represented by a suitable choice of a set of variables; the generality of SC, and its applicability to contingency tables, thus becomes clear. SC also offers various measures of association between sets as well as significance tests and power analysis of

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 4, December 1988, pp. 425–434

0146-6216/88/040425-10$1.75

hypotheses (Cohen, 1988, chap. 10). Unbiased estimates of its major measures of association have been provided (Cohen & Nee, 1984), and computer programs are available for both mainframe computers (Cohen & Nee, 1983) and IBM and compatible microcomputers (Eber & Cohen, 1987).

As is the case for MRC, partialling (residualization) plays an important role in SC. When a set A is partialled from a set B, the k_B variables in the resulting set $B{\cdot}A$ have zero correlations with all the k_A variables in set A. This device is employed in MRC to achieve statistical control (as in the analysis of covariance and the more general analysis of partial variance), to represent conditional relationships (interactions) and curvilinear components, and, with appropriate coding of nominal (categorical) scales, to implement desired contrast functions among groups. It is the last of these that will be exploited in the analysis of contingency tables.

Major Features of SC

Among the many available measures of multivariate association (Cramer & Nicewander, 1979), multivariate $R^2_{Y,X}$ is demonstrably a natural generalization of multiple $R^2_{y\cdot X}$ (van den Burg & Lewis, 1988), and thus may be interpreted as a proportion of (generalized) variance. Using determinants of correlation matrices,

$$R^2_{Y,X} = 1 - \frac{|\mathbb{R}_{YX}|}{|\mathbb{R}_Y|\,|\mathbb{R}_X|} \quad , \tag{1}$$

where $\mathbb{R}_{YX}$ is the full correlation matrix of the Y and X variables,
$\mathbb{R}_Y$ is the matrix of correlations among the variables of set Y, and
$\mathbb{R}_X$ is the matrix of correlations among the variables of set X.
This equation also holds for matrices scaled in terms of variance/covariance or sums of squares/products.

$R^2_{Y,X}$ may also be written as a function of the q squared canonical correlations (CR^2) where $q = \min(k_Y,k_X)$, the number of variables in the smaller of the two sets:

$$R^2_{Y,X} = 1-(1-CR^2_1)(1-CR^2_2),\ \ldots,\ (1-CR^2_q) \quad . \tag{2}$$

In simple applications, the product of the complements of the CR^2s is the familiar Wilks' (1932) Λ, so $R^2_{Y,X} = 1 - \Lambda$. More generally, Λ is the ratio of the determinant of the error matrix to the determinant of the sum of the hypothesis and error matrices, however scaled (as in Equation 1).

Sets Y and X are generic, that is, set X may be an unpartialled set of independent variables B or a partialled set $B{\cdot}A$; similarly, set Y may be an unpartialled set of dependent variables D or a partialled set $D{\cdot}C$. Depending on whether and by what a set is partialled, the type of X,Y relationship may be "whole," "Y-semipartial," "X-semipartial," or "bipartial," and when $D{\cdot}C$ is related to $B{\cdot}C$, the type of relationship is "partial" (as in bivariate correlation). Matrix formulas for computing $R^2_{Y,X}$ and Λ for these five types of association are given in Cohen (1982, Tables 1 and 2).

When k_Y (or k_X) = 1, multivariate $R^2_{Y,X}$ specializes to multiple $R^2_{y\cdot X}$ (or $R^2_{x\cdot Y}$).

Rao (1975) provided an F test for Λ:

$$F = (L^{-1/s}-1)\,\frac{v}{u} \quad , \tag{3}$$

where

$$u = \text{numerator } df = k_Y k_X \quad , \tag{4}$$

$$v = \text{denominator } df = s\left[N-\max(k_C,k_A)-\frac{k_Y+k_X+3}{2}\right]+1-\frac{u}{2} \quad , \tag{5}$$

and

$$s = \left(\frac{k_Y^2 k_X^2 - 4}{k_Y^2 + k_X^2 - 5}\right)^{1/2}, \quad (6)$$

except that when $k_Y^2 k_X^2 = 4$, $s = 1$. When set C or set A does not exist, its $k = 0$.

The Rao F test specializes to the standard null-hypothesis F test for the multiple R^2 (i.e., $q = 1$). For this case, and when $q = 2$, the test is exact; otherwise, it provides a good approximation. Its Type II error (power) validity has also been demonstrated (Cohen & Nee, 1987).

Application to Contingency Tables

An $I \times J$ contingency table represents the relationship between two categorical (nominal scale) variables according to their joint frequencies. As MRC has made familiar, such group-membership variables may be coded in a variety of ways and subjected to correlational analysis (Cohen & Cohen, 1983). For a scale of C levels, each of these coding methods results in $C - 1$ "score" vectors, which, *when used as a set*, are equivalent and fully describe group membership. For example, for any given dataset, they will produce the same multiple R^2 when used as a set of independent variables. The utility of these different coding methods lies in the fact that each provides, through partialling, a set of contrast functions of group membership, as will be illustrated.

The contingency table used illustratively by Zwick and Cramer (1986) came from a fictitious survey presented by Marascuilo and Levin (1983) as the responses of 500 men to the question "Does a woman have the right to decide whether an unwanted birth can be terminated in the first three months of pregnancy?" Table 1 gives the 3×4 (response alternative $\times$ religion) contingency table.

To approach this problem by means of SC, it is necessary to express the categorical variables in a form suitable for correlation. There is literally an infinitude of different codings of the C levels of a categorical scale into $C - 1$ score vectors (Cohen & Cohen, 1983, chap. 5), all of which, when treated as sets, fully represent the group-membership information. Several of these $C - 1$ vectors have the following property: When the remaining $C - 2$ vectors are partialled from each of these $C - 1$ vectors, the resulting variable implements a specific comparison (contrast) among the C groups. Three such useful coding methods, illustrated in Table 2 for Religion and Abortion Response, are dummy variable coding, effects coding, and contrast coding.

Dummy variable coding. The coding for Religion provides a good example. Each vector, unpartialled, is a binary (dichotomous) variable that distinguishes one of the religious groups from the other three, thus *C-O* distinguishes Catholics from the pooled remaining groups. However, when partialled by the other two vectors, it implements the distinction between the group in question and the group coded 0, 0, 0, the Other group in this example. Thus, *C-O* partialling *P-O* and *J-O* (in the conventional notation,

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 Marascuilo and Levin (1983, p. 452).

Table 2
Coding Diagrams for Religion and Abortion Response, and Combined Contrast Coding

a. Religion

	Dummy			Effects			Contrast		
	C-O	*P-O*	*J-O*	*C*	*P*	*J*	*M-M*	*C-P*	*J-O*
Cath	1	0	0	1	0	0	1	1	0
Prot	0	1	0	0	1	0	1	−1	0
Jew	0	0	1	0	0	1	−1	0	1
Oth	0	0	0	−1	−1	−1	−1	0	−1

b. Abortion Response

	Dummy		Effects		Contrast	
	Y-D	*N-D*	*Y*	*N*	*Y-N*	*K-D*
Yes	1	0	1	0	1	1
No	0	1	0	1	−1	1
No op	0	0	−1	−1	0	−2

c. Combined Coding: Religion (Contrast) and Abortion Response (Contrast)

Cell	*N*	*M-M*	*C-P*	*J-O*	*Y-N*	*K-D*
Cath-Yes	76	1	1	0	1	1
Cath-No	64	1	1	0	−1	1
Cath-No op	11	1	1	0	0	−2
Prot-Yes	115	1	−1	0	1	1
Prot-No	82	1	−1	0	−1	1
Prot-No op	6	1	−1	0	0	−2
Jew-Yes	41	−1	0	1	1	1
Jew-No	8	−1	0	1	−1	1
Jew-No op	2	−1	0	1	0	−2
Oth-Yes	77	−1	0	−1	1	1
Oth-No	12	−1	0	−1	−1	1
Oth-No op	6	−1	0	−1	0	−2

C-O·*P-O*,*J-O*) represents a two-group comparison between Catholics and Others, and analogously for the other two partialled variables, *P-O*·*C-O*,*J-O* and *J-O*·*C-O*,*P-O*. Dummy coding of the Abortion Response categories produces *Y-D*·*N-D*, a comparison of Yes respondents and No Opinion (Don't know, or *D*) respondents, and *N-D*·*Y-D*, a comparison of No respondents and No Opinion respondents. Dummy variable coding is optimally used when one of the groups is a reference or control group with which the other $C - 1$ groups are to be compared. The "indicator" variables employed by Zwick and Cramer (1986) to represent Religion and Abortion Response in their MANOVA and canonical analyses are dummy variables, but these methods do not exploit the interpretive meaning of partialled variables.

Effects coding. As can be seen in Tables 2a and 2b, the pattern of effects coding is the same as for dummy coding except that the group coded with a string of 0s is now coded with a string of -1s. If the Protestant "effect" in Table 2a were to be used in a regression analysis (where partialling produces *P*·*C*,*J*), it would yield as its regression coefficient the *Y* mean of the Protestant sample minus the equally-weighted ("unweighted") mean of the means of the four groups, the quantity which is defined as an effect in ANOVA. In SC and correlation analysis generally, when partialled by the other effects-coded variables of the set, it produces a comparison of the group coded 1 with an equally-weighted combination of all the other groups.

Thus, *P·C,J* produces a comparison of the Protestants with equally-weighted Catholics, Jews, and Others. Analogously, the partialled effects-coded Abortion Response *Y·N* in Table 2b compares Yes respondents with an equally-weighted combination of the No and No Opinion groups. Effects coding is optimally used when $C - 1$ groups are to be treated on an equal basis, each to be compared with all the others. One of the groups is inevitably left out; if its effect must be explicitly assessed, the analysis may be repeated with the coding changed to include it.

Contrast coding. This form of coding provides for various contrast functions among the groups other than those provided by dummy and effects coding. For example, for four groups, three contrasts may be coded to represent two binary variables and their interaction, as in a 2×2 factorial design. The particular contrasts chosen for religion in Table 1a, when partialled, are those of a simple nested design: *M-M·C-P,J-O* contrasts the equally-weighted combination of the Protestants and Catholics (the Majority religions) with that of the Jews and Others (the Minority religions). *C-P·M-M,J-O* compares the Catholic and Protestant groups (ignoring the other groups), and *J-O·M-M,C-P* compares the Jewish and Other groups (again ignoring the other groups). The contrast coding for the Abortion Response in Table 2b yields *Y-N·K-D*, the comparison of the Yes and No respondents ignoring the No Opinion group, and *K-D·Y-N*, a comparison of the equally-weighted combination of the Yes and No groups with the No Opinion group.

Note that because SC is correlational, its results will be invariant over linear transformations of the coding values in Table 2. Thus, if in Table 2a, *P* were to be coded 3, 6, 3, 0, or in Table 2b, *K-D* 3, 3, 0, *none* of the SC results using these sets would be affected.

The choice of coding method for each variable is dictated by the substantive issues or hypotheses of interest. Assume that in the present problem, the religious comparisons of the nested design are of interest, and they are to be implemented for the Yes-No contrast and for the contrast of having a Yes or No response versus not having an opinion. Thus, the best choice for both variables would be the contrast coding given in Tables 2a and 2b.

Contrast Coding of Religion

Each of the 500 respondents falls in one of the 12 cells of Table 1, and each cell has its distinctive set of "scores" on the three Religion variables and the two Abortion Response variables. Table 2c combines the selected coding to show the "scores" on the five variables for each cell. Thus, for example, the Protestants responding "no" are represented by the values 1, −1, 0, −1, 1. A full score matrix would include 82 such sets of values to represent the 82 cases in that cell. A complete data matrix would contain a row for each of the 500 respondents, each with five scores representing Religion and Abortion Response. With computer packages such as SYSTAT (Wilkinson, 1986), it is unnecessary to physically represent all 500 rows; the 12 rows plus the cell *N*s employed as weights are sufficient.

The SC analysis proceeds with the correlation matrix among the five variables for the 500 cases, given in Table 3. Note that these are simple product-moment correlation coefficients—no partialling has taken place. Note also that although the contrast *coefficients* for Religion given in Table 2 are mutually orthogonal, because the *N*s for the four religions are not equal, the correlations among *M-M*, *C-P*, and *J-O* for the 500 cases are not 0. The same holds for *Y-N* and *K-D*. Because the correlations are for unpartialled variables, their interpretation is at best unclear; in any case, they do not carry the contrasts intended by the hypotheses.

The analytic strategy recommended for SC is to first examine the relationship between the sets made up of multiple variables, and then to pursue those involving their constituents, that is, to determine the source guided by specific a priori hypotheses. To minimize the experimentwise Type I error, when the

relationship involving a set is not significant, its constituent variables are not further analyzed. This is an adaptation of Fisher's method of "protected" t tests, and is the same strategy recommended for MRC (Cohen & Cohen, 1983, pp. 172–176). When applied to contingency tables, the constituent variables are partialled variables that carry the contrasts defined by the hypotheses.

Results

First, the relationship between the set X for Religion (REL) and the set Y for Abortion Response (ABO) was examined. The determinants for the total matrix in Table 3, and the Y and X submatrices thereof, were substituted in Equation 1 to find $R^2_{Y.X} = 1 - [.8494/(.9951)(.9280)] = .080$. For the constituents of the Rao F of Equation 3, $k_Y = 2$ and $k_X = 3$ (and both k_C and k_A are 0). Therefore, Equation 4 gives numerator $df\,(u) = 2(3) = 6$, and because the s of Equation 6 works out to 2, Equation 5 gives denominator $df\,(v) = 2[500 - 0 - (2 + 3 + 3)/2] + 1 - 6/2 = 990$. For $R^2_{Y.X}$ between whole sets, Wilks' $\Lambda = 1 - R^2_{Y.X} = 1 - .080 = .920$, so Equation 3 gives $F = (.920^{-1/2} - 1)(990/6) = 7.03$, which with 6 and 990 df is highly significant. (These and the computations that follow were performed using SETCORAN, a computer program for IBM and compatible microcomputers, which provides more detailed output than is presented here, including individual regression analyses on all the variables; see Eber & Cohen, 1987.)

Table 4 displays these results in line 1 and those of the follow-up tests in logical order, using the protection strategy. The unsquared correlations are also included because they give the direction of the differences when single (albeit partialled) variables are correlated. Line 1 shows that there is a modest but significant degree of association overall between REL and ABO, that is, the religious groups differ in their response to the survey question ($R^2_{Y.X} = .080$). However, lines 2 and 3 indicate that although the religious difference with regard to the Yes-No contrast is both material ($R^2_{Y.X} = .073$) and significant, it is neither for the Opinion-No Opinion distinction ($R^2_{Y.X} = .008$). Note that each of these $R^2_{Y.X}$s is a multiple R^2, the special case where one of the sets has a single (albeit partialled) variable. However, such constituent $R^2_{Y.X}$s are *not*, in general, additive. ($T^2_{Y.X}$, another measure of multivariate association, does have additive properties; see Cohen, 1982; van den Burg & Lewis, 1988.)

Because the religious groups show no significant difference in regard to the Opinion-No Opinion contrast, following the protection principle, its relationships with the individual contrasts in REL were not pursued (they were, in this instance, all tiny and nonsignificant). Line 4 shows that Majority-Minority contrast accounts materially ($R^2_{Y.X} = .066$) and significantly for (generalized) variance in the ABO set. When this is followed up by testing the Majority-Minority contrast for the Yes-No contrast in line 5, it, too, is found to account significantly for variance in the latter, also with $R^2_{Y.X} = .066$. Note that this is a simple bipartial relationship, that is, a relationship between two single variables that are differently

Table 3
Correlations Among Abortion Response and Religion Variables, Both Contrast-Coded

	Abortion Response (ABO)		Religion (REL)		
Variable	*Y-N*	*K-D*	*M-M*	*C-P*	*J-O*
Y-N	1.000				
K-D	.070	1.000			
M-M	−.266	.014	1.000		
C-P	−.016	−.084	−.080	1.000	
J-O	−.078	.031	.257	−.021	1.000

Table 4
SC Analysis of Religion Versus
Abortion Response, Both Contrast-Coded

Line	Y	X	$R^2_{Y,X}$	$R_{Y,X}$	F	u,v	p
1	ABO	REL	.080	.283	7.03	6,990	.000
2	*Y-N·K-D*	REL	.073	.270	12.95	3,495	.000
3	*K-D·Y-N*	REL	.008	.092	1.40	3,495	.240
4	ABO	*M-M·C-P,J-O*	.066	.258	17.66	2,495	.000
5	*Y-N·K-D*	*M-M·C-P,J-O*	.066	−.258	35.40	1,496	.000
6	ABO	*C-P·M-M,J-O*	.008	.089	1.99	2,495	.135
7	ABO	*J-O·M-M,C-P*	.001	.032	.24	2,495	.791

Note. ABO = Contrast-coded Abortion Response set: *Y-N, K-D.*
REL = Contrast-coded Religion set: *M-M, C-P, J-O.*

partialled. Thus, $R_{Y,X} = r_{(Y\text{-}N\cdot K\text{-}D)(M\text{-}M\cdot C\text{-}P,J\text{-}O)} = -.258$; the negative sign is meaningful and indicates that the majority religions gave more No (relative to Yes) responses than the minority religions. Finally, lines 6 and 7 show that the other two nested religious contrasts do not relate significantly to the ABO set.

In summary, the SC analysis has not only assessed the overall relationship between religion and abortion response, but has identified its only demonstrable source to be the greater rate of No (compared to Yes) responses of the majority religions compared with the minority religions.

Effects Coding for Religion

What if other specific hypotheses, hence other coding, had been chosen for the nominal scales being related? It would have been quite reasonable to use effects coding for REL (Table 2a), thus contrasting individually three of the religious groups with an equally-weighted combination of the others. On the other hand, no alternative to the contrast coding used for ABO seems sensible. For illustrative purposes, the analysis was repeated with REL effects-coded, leaving the ABO contrast coding unchanged. The resulting correlation matrix is given as Table 5.

The change in coding has resulted in changes in all the correlation coefficients except that between *Y-N* and *K-D*, for which the coding did not change. As was the case in Table 3, because the correlations are for unpartialled variables, their meaning is unclear and, in any case, they do not carry the intended comparisons.

This matrix was subjected to the same analysis as was that of Table 3; the results are summarized in Table 6. It is instructive to compare Table 6 with the summary of the previous analysis summarized in Table 4. Note first (line 1) that the overall relationship between REL, now effects-coded, and ABO is exactly as before. This illustrates the fact that the information carried by a set treated as a whole is invariant over changes in coding. (Nor would line 1 have changed had ABO been differently coded.) For the same reason, lines 2 and 3 are also unchanged from Table 4. As before, there is no evidence that the religions differ with regard to the Opinion-No Opinion contrast (line 3), and that line of follow-up analysis is not pursued further.

Changes begin to occur with line 4, where the analysis follows up the "effects" of REL rather than the contrasts that were its previous constituents. It is found that each of the religions coded has a significant effect on the ABO set (lines 4, 6, and 8), and more particularly, each has an effect with regard to the Yes-No contrast component of the ABO set. The bipartial correlations for *C* (−.192), *P* (−.160), and *J* (.108) show the degree and direction of the effect. Thus, Catholics' tendency to respond Yes rather than

Table 5
Correlations Among Abortion Response (Contrast-Coded) and Religion (Effects-Coded) Variables

	Abortion Response (ABO)		Religion (REL)		
Variable	*Y-N*	*K-D*	*C*	*P*	*J*
Y-N	1.000				
K-D	.070	1.000			
C	−.214	−.029	1.000		
P	−.182	.067	.323	1.000	
J	−.078	.031	.541	.529	1.000

No is significantly less than that of the equally-weighted combination of the remaining groups, and the same is true for Protestants, whereas for Jews it is significantly more; for each group the tendency is to the degree indicated by their bipartial correlations. In summary, from this analysis it would have been learned that each religion departs significantly from the aggregate of the others in regard specifically to the Yes-No response.

Ordinarily, it would not be advisable to run alternative coding methods on the same problem, except for purely data-exploration purposes. Coding methods embody specific hypotheses, and pursuing many possibilities invites a rapid escalation of the risk of "finding" things that are not there (i.e., of the researchwise Type I error). This was done above for illustrative purposes, not to set a bad example. This risk can be controlled by a Bonferroni reduction of the significance criterion, but this is, of course, accompanied by what may be a fatal loss of statistical power (Cohen, 1988, chap. 10).

Discussion

SC generally provides a unified framework within which to systematically study relationships among phenomena, unconstrained by level of measurement. It proceeds systematically to analyze these relationships in terms of specific a priori hypotheses. SC offers correlation coefficients and proportions of variance as measures of association, and a formal basis for hypothesis testing, estimation, and power analysis. These features are brought to bear in the SC analysis of contingency tables, where the use of different coding schemes for nominal scales, together with partialling, provide the desired specificity of analysis, as was illustrated above.

Zwick and Cramer (1986) showed that Pearson chi-square analysis, MANOVA, canonical analysis, and a form of correspondence analysis all yield equivalent results when applied to an $I \times J$ contingency table, using Table 1 for illustration. To implement their analyses, I and J were expressed respectively as $I - 1$ and $J - 1$ "indicator" (specifically, dummy) variables, and the correlation matrix among these variables was determined. They showed that the matrix equation whose solution yields the squared canonical correlations between the two sets is readily transformed into a form whose trace is the Pillai-Bartlett statistic used in MANOVA and other multivariate methods. Another version of the canonical analysis yields the same function computed for the conventional Pearson chi-square. Finally, they showed how the solution of this version of the canonical analysis may be converted to the solution of a first-order correspondence analysis.

The results of the four methods are equivalent in the sense that they produce the same significance test results for the overall association between I and J. The Pillai-Bartlett statistic is the sum of the squared canonical correlations, which, when multiplied by N, is approximately chi-square distributed with

$df = k_Y k_X$; given the demonstrated equivalence, all the methods produce the same chi-square value for the same df. (For the example, $\chi^2 = 40.17$, $df = 6$, and $p = .000$.) Equation 2 shows that $R^2_{Y.X}$, and therefore Wilks' Λ, is also a function of the squared canonical correlations. The Rao F test applied to $R^2_{Y.X}$ is another of several multivariate test statistics that are somewhat more accurate than the Pillai-Bartlett chi-square procedure, although for large samples the tests are virtually interchangeable. (The Rao test is preferred primarily because, as noted, it is a direct generalization of the conventional F test on multiple R^2, and because evidence exists of its robustness in power estimation; see below.)

It was therefore inevitable that the SC analysis would show exactly the same highly significant overall association between Religion and Abortion Response (line 1 in Tables 4 and 6) as was found by the other methods. Following this first step, the SC analyst proceeds systematically to analyze the association in terms of specific a priori hypotheses, each with measures of the strength of association and a significance test.

SC does not do everything. Correspondence analysis is a form of metric multidimensional scaling ("dual scaling"; Tenenhaus & Young, 1985) primarily employed to generate optimal scale values for the categories of nominal scales. The scale values are optimal in the sense that a variable so scaled will yield a maximum F ratio as the dependent variable in an ANOVA with the other categorical variable's groups. The utility of the method lies in its ability to represent the rows and columns as $I + J$ points in a q-dimensional space, of which the largest two dimensions are often sufficient for representation. By portraying such characteristics as religions and abortion responses as points on a graph, affinities and disparities can be observed both within and between religions and responses. It is easy to see the virtues of correspondence analysis as an exploratory tool. Note, however, that the scale values for I and J, having been generated to be optimal relative to each other, have no necessary meaning when either is related to some third variable.

Zwick and Cramer showed, moreover, that the solution of the matrix equation of the canonical analysis also produces the optimal weights; it follows that the MANOVA analysis does as well. Thus, except for the Pearson chi-square test, all of the methods they described may be employed to obtain this useful portrait of the two-way contingency table. Correspondence analysis also generalizes to multi-way frequency tables.

SC may also be used in the analysis of contingency tables of higher order. Assume that in the above example, data were available for the region in which the respondent lives. These data could be coded (say, using effects coding) and partialled from REL and ABO to assess their relationship "controlling"

Table 6
SC Analysis of Effects-Coded Religion
Versus Contrast-Coded Abortion Response

Line	Y	X	$R^2_{Y.X}$	$R_{Y.X}$	F	u,v	p
1	ABO	REL	.080	.283	7.03	6,990	.000
2	Y-N·K-D	REL	.073	.270	12.95	3,495	.000
3	K-D·Y-N	REL	.008	.092	1.40	3,495	.240
4	ABO	C·P,J	.040	.201	10.70	2,495	.000
5	Y-N·K-D	C·P,J	.037	−.192	19.80	1,496	.000
6	ABO	P·C,J	.030	.172	7.84	2,495	.001
7	Y-N·K-D	P·C,J	.026	−.160	13.74	1,496	.000
8	ABO	J·P,C	.012	.110	3.24	2,495	.039
9	Y-N·K-D	J·P,C	.012	.108	6.26	1,496	.012

Note. ABO = Contrast-coded Abortion Response set: Y-N, K-D.
REL = Effects-coded Religion set: C, P, J.

for region (i.e., their pooled within-region associations). The rest of the analysis would be as before, except that region would additionally be partialled in the X and Y sets throughout. Also, as a condition for pooling, or as an issue in its own right, the analyst might wish to assess the interaction of religion by region (i.e., the homogeneity of the REL vs. ABO association over regions). Because SC is a multivariate generalization of MRC, this can be accomplished exactly as in MRC, using as the X set the REL × ABO product set from which REL and ABO have both been partialled, and using the Y set as before.

Indeed, other variables to be included in the analysis need not be coded as nominal scales. Quantitative variables, such as years of education and/or annual income, can be treated as was region, either as single variables or as sets (Cohen & Cohen, 1983, chap. 6, 8).

A problem posed by all of the standard multivariate significance tests is their assumption of multivariate normality for the dependent variable set, which is patently not met by contingency tables. Some monte carlo work nearing completion provides some reassurance of the robustness of the Rao F test. For five null-association 3 × 4 contingency tables with marginals of varying degrees of skewness and for $N = 60$, 120, and 240, actual Type I error rates were determined for the nominal .01 and .05 levels using 2,000 replications for each of these 15 combinations of conditions. The actual rates overall were .012 and .050, respectively. (Results for $N = 30$ were poorer: .023 and .056.) Also, for five 3 × 4 tables of varying degrees of association, power values computed using the noncentral F distribution hardly differed from the monte carlo results (overall mean difference = + .004).

References

Cohen, J. (1982). Set correlation as a general multivariate data-analytic method. *Multivariate Behavioral Research, 17,* 301–341.

Cohen, J. (1988). *Statistical power analysis* (2nd ed.). Hillsdale NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences* (2nd ed.). Hillsdale NJ: Erlbaum.

Cohen, J., & Nee, J. C. N. (1983). CORSET, a FORTRAN IV program for set correlation. *Educational and Psychological Measurement, 43,* 817–820.

Cohen, J., & Nee, J. C. N. (1984). Estimators for two measures of association for set correlation. *Educational and Psychological Measurement, 44,* 907–917.

Cohen, J., & Nee, J. C. N. (1987). A comparison of two noncentral F approximations, with applications to power analysis in set correlation. *Multivariate Behavioral Research, 22,* 483–490.

Cramer, E. M., & Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika, 44,* 43–54.

Eber, H. W., & Cohen, J. (1987). *SETCORAN. A PC program to implement set correlation as a general multivariate data-analytic method* [Computer program]. Atlanta: Psychological Resources.

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

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston.

Rao, C. R. (1975). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

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.

van den Burg, W., & Lewis, C. Some properties of two measures of multivariate association. *Psychometrika, 53,* 109–122.

Wilkinson, L. (1986). *SYSTAT: The system for statistics*. Evanston IL: SYSTAT.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika, 24,* 471–494.

Zwick, R., & Cramer, E. M. (1986). A multivariate perspective on the analysis of categorical data. *Applied Psychological Measurement, 10,* 141–145.

Acknowledgments

Patricia Cohen, as always, provided a helpful critique.

Author's Address

Send requests for reprints or further information to Jacob Cohen, Department of Psychology, New York University, 6 Washington Place, New York NY 10003, U.S.A.