

PROC CORRESP:

Different Perspectives for Nominal Analysis

Richard W. Cole

**Systems Analyst
Computation Center**

The University of Texas at Austin

ABSTRACT

Correspondence Analysis as a fundamental approach to the analysis data having any scale of measurement fails to achieve wide acceptance in the United States. As a descriptive and inferential statistical technique, it should be riding the crest of currently popular categorical modeling procedures for a number of reasons: 1) it allows one to plot, in metric space the relationships among nominal categories of variables, to 2) quantify the size and number of dimensions comprising these relationships, and, 3) generate 'optimal scores' that reflect the interval scaled metric values underlying each variable. This paper compares this technique to more traditional categorical modeling procedures and emphasizes its relative advantages for business and behavioral science applications.

INTRODUCTION

Analysis of variables having a nominal scale of measurement has traditionally been accomplished through the use of various approaches such as Pearson's chi-square test of independence, or a likelihood-ratio chi-square test of independence (using maximum likelihood estimation). While the latter has been expanded upon in recent years to a multivariate modeling approach where one seeks to explain the data using some parsimonious set of variables, few researchers, at least in the United States, have examined nominal variables from more of a structural perspective presently available thru what now consistently being called Correspondence Analysis.

This form of analysis has its foundations in a matrix transformation procedure known as singular value decomposition where a matrix of any dimension can be reduced to its basic structure, with resulting dimensions quantified in terms of their explanatory power. These dimensions may be statistically examined, and the levels of each variable (known as points) plotted on these axis. This fundamental approach to contingency table analysis has been rediscovered across, and within numerous disciplines, though the earliest treatment would seem to have been Eckart & Young (1936). At present, correspondence analysis appears to be increasing in popularity with the recent releases of SAS's PROC CORRESP and the SPSS CATEGORIES product. Additionally, publications by the SAGE Press (Weller & Romney, 1990) and the classic text by Greenacre (1984) provide the reader with a thorough introduction to the topic. In this brief paper

a number of examples are given to introduce users to correspondence analysis and the advantages it may provide over traditional tests of independence and fit.

Traditional Approaches to Nominal Analysis

When faced with examining the relationship between two nominal variables most researchers opt for some type of chi-square test of independence where the hypothesis of interest is no interaction or dependency among the variables of interest. That is, the obtained cell frequencies may be explained adequately by the marginal distributions for each of the variables. Having examined the chi-square statistic associated with this test and having found it significant, one may conduct some form of follow-up analysis either by generating contrasts among specific cells or simply examine some form of standardized residuals, which when summed, equal the overall chi-square value. These values that help the researcher to understand relations among the various levels of each variable.

Correspondence Analysis

While the approach above is useful for understanding relations among specific cells and conducting tests of various hypotheses, little may be understood about the interrelationship or dimensionality among the variables. For variables assumed to have linear relationships, various factor, and principal component analysis approaches are used to generate a reduced dimensional space for the variables that may then be plotted along these axes. Correspondence Analysis is the nominal counterpart to such an analysis. By default through PROC CORRESP, output contains information about the dimensions of the data with variance accounted for by a dimension, and its chi-square significance test. Users may output coordinates for the dimensions to SAS/GRAPH and generate a plot of the various levels for each variable. This graphic is more appropriately called a 'profile' containing a point for each level of a variable.

A COMPARISON OF APPROACHES

To illustrate the differences in these two approaches, sample data from Greenacre (1984) involving smoking habits across various levels of employment will be used. Here the researcher has an interest in the "dangers of smoking" at levels of: 1) none, 2) light, 3) medium, and, 4) heavy, for employees categorized as: 1) senior

management, 2) junior management, 3) senior employees, 4) junior employees, and, 5) secretarial staff.

Independence Testing Approach

Table 1 contains a contingency table of these data and results from both traditional chi-square independence tests using SAS's PROC FREQ. Note that these tests of independence fail to detect the presence of an interaction between smoking and employment status. In such a case an examination of the standardized residuals is not necessarily warranted though these values do give an indication of 'heavy' and 'light' cells contributing to the overall chi-square value. The large cell chi-square values represent a discrepancy between the observed and expected cell frequencies (seen in rows one and two of each cell). Based on these results one might conclude while junior employees tend to not be, 'non smokers', senior employees are often found to be 'non smokers'. Still, no overall 'dependency' exists as indicated by the high probability values associated with either test of independence.

Correspondence Analysis Approach

With Correspondence Analysis quite a different picture emerges. Figure 1 contains a plot of the first two correspondence analysis dimensions calculated from the data. In this plot, dimension 1 (along the x axis) may be expressed along a no smoking to heavy smoking continuum with medium smoking best representing the smoking end of the continuum (due to its distance from the origin and its low dimension 2 value). A reexamination of the contingency table analysis of Table 1 may support this initial interpretation. Note that high cell chi-square values are only obtained for no smoking (in the junior-emp and senior-emp cells) which may therefore be viewed as contributing primarily to the definition of Dimension 1. As might be expected, the three smoking categories (light, medium, and heavy) are represented on the smoking end of Dimension 1 with nonsmoking independently representing the nonsmoking end of the continuum. As can be seen from the PROC CORRESP output in Table 2, Dimension 1 explains 87 percent of the variance (distances across points) in the data. For the smoking levels we may examine inertia values that are analogous to variance accounted for within a factor, but in this case within a dimension. Of obvious interest is the partial inertia for the no smoking category having a value of .6539. This indicates that the

TABLE 1

Employment By Smoking Habits

Rows-Employment Status

Columns-Smoking Habits

Frequency					Total
Expected Freq					
Cell Chi-Square	hvysmoke	ltsmoke	medsmoke	nosmoke	
junior-emp	13	24	33	18	88
	11.399	20.518	28.269	27.813	
	0.2249	0.5909	0.7916	3.4625	
junior-mgmt	4	3	7	4	18
	2.3316	4.1969	5.7824	5.6891	
	1.1938	0.3413	0.2564	0.5015	
secret	2	6	7	10	25
	3.2383	5.829	8.0311	7.9016	
	0.4735	0.005	0.1324	0.5573	
senior-emp	4	10	12	25	51
	6.6062	11.891	16.383	16.119	
	1.0282	0.3008	1.1728	4.8929	
senior-mgmt	2	2	3	4	11
	1.4249	2.5648	3.5337	3.4767	
	0.2321	0.1244	0.0806	0.0788	
Total	25	45	62	61	193
	12.95	23.32	32.12	31.61	100.00

STATISTICS

Statistic	DF	Value	Prob
Chi-Square	12	16.442	0.172
Likelihood Ratio Chi-Square	12	16.348	0.176

relative contribution of this category to Dimension 1 is around 65 percent. This is a clear indication that no smoking plays a major role in defining Dimension 1. Dimension 2 would seem reflective of some construct ranging from light to heavy smoking, though its interpretation is more difficult. Regarding employment status, notice that senior employees and junior employees fall roughly along Dimension 1 as well. This finding would seem to be consistent with findings from the standardized residuals especially in the case of nonsmoking senior employees who are overly abundant.

TABLE 2

Smoking Habits by Employment Level Analysis

Inertia and Chi-Square Decomposition

Singular Values	Principal Inertias	Chi- Squares	Percents	18	36	54	72	90
				-----+-----+-----+-----+-----				
0.27342	0.07476	14.4285	87.76%	*****				
0.10009	0.01002	1.9333	11.76%	***				
0.02034	0.00041	0.0798	0.49%					
	-----	-----						
	0.08519	16.4416	(Degrees of Freedom = 12)					

Summary Statistics for the Row Points

	Quality	Mass	Inertia
junior-emp	0.999810	0.455959	0.308354
junior-mgmt	0.991082	0.093264	0.139467
secret	0.998603	0.129534	0.071053
senior-emp	0.999817	0.264249	0.449750
senior-mgmt	0.892568	0.056995	0.031376

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
junior-emp	0.330974	0.151772
junior-mgmt	0.083659	0.551151
secret	0.070064	0.080522
senior-emp	0.512006	0.002998
senior-mgmt	0.003298	0.213558

Summary Statistics for the Column Points

	Quality	Mass	Inertia
hvysmoke	0.994552	0.129534	0.191743
ltsmoke	0.984016	0.233161	0.082860
medsmoke	0.983228	0.321244	0.148025
nosmoke	0.999995	0.316062	0.577372

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
hvysmoke	0.149538	0.505754
ltsmoke	0.030850	0.463174
medsmoke	0.165617	0.001737
nosmoke	0.653996	0.029336

Figure 1

Scatter plot of Dimension 1 vs Dimension 2 for the 'smoke' variable. The plot shows four clusters: 'senior-ngat' and 'junior-ngat' in the top half, and 'senior-emp' and 'junior-emp' in the bottom half. 'nosmoke' is on the left and 'ltsmoke' is on the right. The x-axis ranges from -0.5 to 0.5, and the y-axis ranges from -0.3 to 0.3.

41

stressed as a graphical approach to understanding the relationship among variables, this technique can be used in conjunction with more traditional expectancy table approaches to examine information other than hypothesis testing.

REFERENCES

- Eckart, C., & Young, G. (1936). The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, 211-218.
- Greenacre, M. (1984) Theory and applications of correspondence analysis. Academic Press, London.
- Weller, S. C., & Romney, A. K. (1990). *Metric Scaling: Correspondence Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-075. Newbury Park, CA: Sage.

Richard W. Cole
Systems Analyst
Computation Center
University of Texas @ Austin
Austin, TX 78701-1110

INTERNET: rcole@emx.cc.utexas.edu
BITNET: rwc@utxvm