Paper 227-2008

PROC CORRESP for Categorical Data: Correspondence Analysis (CA) for Discovery, Display, and Decision-Making

Wendy Dickinson, Ph.D., Ringling College of Art and Design, Sarasota, FL
Bruce W. Hall, Ed.D., University of South Florida, Tampa, FL

ABSTRACT

Over the past few decades, correspondence analysis has gained "an international reputation as a powerful statistical tool for the graphical analysis of contingency tables" (Beh, 2004, p.257). Correspondence analysis is an exploratory data technique (Hoffman, Franke, 1986) related to both the biplot and principal components analysis that yields a multidimensional representation of the association between the row and column categories of a two-way contingency table (Friendly, 1991).

Correspondence analysis is designed to show how data deviate from expectation (observed values vs. expected values) when the row and column variables are independent (Friendly, 1991). Correspondence analysis thus produces a two-dimensional graphical plot of the observed data variation (Wheater et al, 2003), which can then be examined for behavioral overlap between variables.

The contemporary application of correspondence analysis to categorical data can provide a powerful tool in the arsenal of statistical methodology – a tool to help interpret and describe relationships between categorical variables. For example, correspondence analysis could help us interpret measures of association, agree/disagree responses for dichotomous survey items, and pass/fail categories for educational assessment instruments.

A simple scenario is provided to illustrate the application of correspondence analysis to a categorical dataset, utilizing SAS® 9.1 code to demonstrate the PROC CORRESP procedure for categorical data. This scenario example provides an introduction to the mathematical theory behind correspondence analysis, and utilizes PROC CORRESP (SAS® 9.1) to analyze a dataset, produce quantitative output, and provide a graphical display for both informed decision-making and effective dissemination of research results.

INTRODUCTION

THEORETICAL OVERVIEW

Over the past few decades correspondence analysis has gained "an international reputation as a powerful statistical tool for the graphical analysis of contingency tables" (Beh, 2004, p.257). Correspondence analysis is an exploratory data technique (Hoffman, Franke,1986) related to both the biplot and principal components analysis that yields a multidimensional representation of the association between the row and column categories of a two-way contingency table (Friendly, 1991).

Correspondence analysis, an ordination technique (Palmer, 1993), examines the proportional goodness-of-fit between recorded variables (Gabriel, 2002), and shows data deviation from expectation within the row and column categories (Friendly, 1991). Correspondence analysis produces a two-dimensional graphical plot of the observed data variation (Wheater et al, 2003),

which can then be examined for behavioral overlap between variables. A variant of this method, subset correspondence analysis (Greenacre & Pardo, 2006) uses only selected portions of the data to enter into the analysis and subsequent algorithmic activity. This paper focuses on an illustrative scenario example utilizing the complete dataset.

MATHEMATICS OF CORRESPONDENCE ANALYSIS

Each cell within the contingency table contributes to the χ^2 statistic value. Per Friendly (1991, p.514), the matrix of deviations from independence, **D**, is expressed in terms of the proportions, **pij**:

$$\mathbf{D} = \{dij\} = \frac{(pij - pi + p + j)}{\sqrt{pi + p + j}} = \frac{(fij - eij)}{\sqrt{eij}\sqrt{n}}$$

Therefore, correspondence analysis is designed to show how data deviate from expectation (observed values vs. expected values) when the row and column variables are independent (Friendly, 1991).

PURPOSE

The purpose of this paper is to provide an overview of correspondence analysis, and present an actual application (simple scenario example) of PROC CORRESP (SAS 9.1) to a categorical data set. The PROC CORRESP code and methodology, quantitative output, and graphical variable displays of the correspondence analysis are presented.

METHODS

First, a scenario is provided for the reader to illustrate the application of correspondence analysis to a categorical dataset. Second, SAS 9.1 code was developed to demonstrate the PROC CORRESP procedure for categorical data. Next, the scenario dataset output and results are presented. Lastly, practical implications for correspondence analysis are provided. The scenario is provided below.

Correspondence Analysis: Scenario

You are the external evaluator for a new museum display consisting of several free-standing, room-like educational exhibits. The museum management would like you to determine how visitors to the museum are interacting with the new exhibits, based on age. You decide to record each visitor's age, and then observe how each visitor interacts with the exhibit.

The two categorical variables are thus age group, and the observed exhibit behavior. Observed behaviors include looking only at the exhibit from the outside, peeking inside the door to the exhibit space, and entering into the exhibit space. Age grouping was delineated as: young (19 years and under), midage (20-39 years), and older (40 and over). Over one hour, you observe the visitors(N = 300) in the museum exhibit area. The raw data is recorded in Table 1, Data for Museum Behaviors.

SAS code for correspondence analysis (CA)

```
*****************************************************;
*Wendy Dickinson and Bruce Hall
*Museum.sas, SAS version 9.1 (2008)
*Modified from Friendly, SAS version 6.0 (1991)
*****************************************************;
*input categorical data from visitor observations  *;
data museum;
input enters peeksin lookonly agegroup$;
cards;
        10 30 60   young
        30 50 20   midage
        70 20 10   older
        ;

title1 'Museum Behaviors and Age';

*****************************************************;
*invoke the correspondence procedure
*define variables
*agegroup is the row variable
*behavior is the column variable
*****************************************************;
proc corresp data=museum out=behavior;
var enters peeksin lookonly;
id agegroup;

proc print data = behavior;
var _type_  agegroup dim1 dim2 quality;
run;

data label;
      set behavior;
      xsys = '2'; ysys = '2';
      x = dim1;
      y = dim2;
      text= agegroup;
      size = 1.4;
      function = 'LABEL';
            run;
*****************************************************;
*invoke SAS/GRAPH to create visual display
*using coordinate system
*define axis 1 and axis 2
*****************************************************;
      proc gplot data = coord;
            plot dim1 * dim2
            /anno = label frame
            href = 0 vref = 0 lvref= 3 lhref= 3
            vaxis = axis2  haxis = axis1
            vminor = 1 hminor = 1;

            axis1 length = 3 in order= (-1 to 1 by .5)
                  label = (h= 1.3   'Dimension 1');
            axis2 length = 3 in order = (-1 to 1 by .5)
                  label = (h=1.3  a=90 r=0 'Dimension 2');
                  symbol v=none;
      run;
```

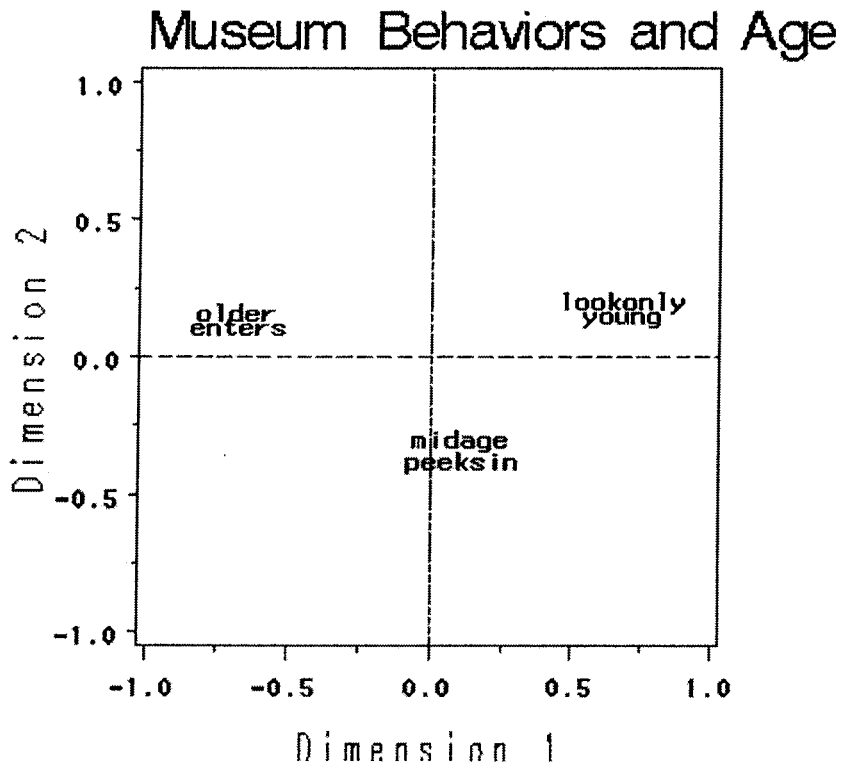
RESULTS

Row and column coordinates

SAS output results from the Correspondence Analysis (scenario example), with row and column coordinates, are documented in Tables 2 and 3.

The correspondence analysis graphical output is shown in Figure 1, Museum Behaviors and Age. In comparison to the traditional chi-square table (numbers only) output, this visual graph output greatly improves the communication of the underlying correspondence between observed and expected variable information. (For an excellent discussion of traditional frequency tables, see Cody and Smith, 2006). As shown below, Dimension 1 depicts the (column coordinates) of observed behaviors (columns 1 and 2). Dimension 2 depicts the associated location differences (row coordinates) between museum visitors based on age. The two-dimensional graphical display combines the row and column coordinates to create a useful visual summary of the association between the variables of age group and behavior.

Figure 1. Museum Behaviors and Age



The Quality Variable

In Table 4, the "quality" variable provides a quantitative measure of the proportion of a point's distance from the origin, accounted for by the two-dimensional solution (DIM1, DIM2) within PROC CORRESP (Friendly,1991). In this example scenario, the quality values indicate that each of the points is being represented quite well in the two-dimensional solution.

Chi-Square decomposition and singular values

Singular values and the chi-square decomposition are shown in Table 5. As discussed by Friendly, the singular values "are the canonical correlations between the optimally scaled categories" (1991, p.517).If we assign the DIM1 scores for age group and behavior to the total visitor set of 300 behavior observations, the correlation of these variables would be .5546. By examining the DIM2 scores, this review provides a second orthogonal scaling of the two categorical variables, yielding a correlation value of .2537 (Friendly,1991).).

IMPLICATIONS

By looking at the graph of the association between age and behavior, we can determine, based on their age group, which museum visitors are likely to look only(if they are young), peek inside(if they are in the midage group), or enter into the exhibit (if they are older).

This type of visual summary provides the viewer with a clear understanding of the association between age and behavior recorded in the data. This clear understanding might have been masked, however, if we had merely presented the reader with a cross-tabs table of numbers.

While the use of correspondence analysis is relatively new to the educational research field, its "history can be traced back at least 50 years" (Hoffman & Franke, p.213, 1986).**T**he contemporary application of correspondence analysis to categorical data can provide a powerful tool in the arsenal of statistical methodology – a tool to help interpret and describe relationships between categorical variables. In example, correspondence analysis could help us interpret the measure of association between variables of gender, ethnicity, and school type, variables of agree/disagree for dichotomous survey response categories, and categories of passing/failing for educational assessment instruments.

This example scenario provides an introduction to both the mathematical theory behind correspondence analysis, and utilizes PROC CORRESP (SAS 9.1) to analyze a dataset, produce quantitative output, and provide a graphical display for effective dissemination of research results.

REFERENCES

Beh, E.J. (2004). Simple Correspondence Analysis: A bibliographic overview. *International Statistical Review 72(2), 257-284.*

Cody, R.P. & Smith, J.K. (2006). Applied Statistics and the SAS Programming Language, Fifth Edition. Pearson Education, Inc., Upper Saddle River: NJ.

Friendly, M. (1991). SAS System for Statistical Graphics, First Edition. SAS Institute, Inc., Cary: NC.

Gabriel, K.R. (2002). Goodness of fit of biplots and correspondence plots. *Biometrika, 89(2), 423-436.*

Greenacre, M., & Pardo, R. (2006). Subset Correspondence Analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research, 35(2), 193-218.*

Hoffman, D. L. & Franke, G.R. (1986). Correspondence Analysis: Graphical representation of categorical data in marketing research. *Journal of Marketing Research, 23(3), 213-227.*

Palmer, M. (1993). Putting things in even better order: The advantages of canonical correspondence analysis. *Ecology, 74(8), 2215-2230.*

Wheater, C. P., Cook, P.A., Clark, P, Syed, Q. and Bellis, M.A. (2003). Re-emerging Syphilis: a detrended correspondence analysis of the behavior of HIV positive and negative gay men.

Retrieved from http://www.biomedcentral.com/1471-2458/3/34

ACKNOWLEDGMENTS

The Authors would like to thank Lisa Adkins for her invaluable assistance in the preparation of this document.

CONTACT INFORMATION

Wendy B. Dickinson, Ph.D.
Acting Coordinator of Mathematics
Ringling College of Art and Design
2700 North Tamiami Trail
Sarasota, FL 34234
Work Phone: 941.359.7521
Fax: 941/953/5071
E-mail: wdickins@ringling.edu

Table 1

Date for Museum Behaviors

Age grouping	Looks only	Peeks in	Enters
Young	60	30	10
Midage	20	50	30
Older	10	20	70

Table 2

The CORRESP Procedure Output: row coordinates

	Dim 1	Dim 2
Young	0.6827	0.1766
Midage	-0.0071	-0.3588
Older	-0.6756	0.1822

Table 3

The CORRESP Procedure Output: column coordinates

	Dim1	Dim2
enters	-0.6669	0.1346
peeksin	0.1193	-0.3546
lookonly	0.6825	0.2295

Table 4

Correspondence Analysis Code Output

Obs	_TYPE_	State	Dim1	Dim2	Quality
1	INERTIA				
2	OBS	young	0.68270	0.17661	1
3	OBS	midage	-0.00705	-0.35880	1
4	OBS	older	-0.67565	0.18220	1
5	VAR	enters	-0.66686	0.13457	1
6	VAR	peeksin	0.11929	-0.35464	1
7	VAR	lookonly	0.68251	0.22957	1

Table 5

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.55457	0.30754	92.263	82.69	82.69
0.25372	0.06437	19.312	17.31	100.00
Total	0.37192	111.576	100	