The Philippine Statistician, 1998
Vol. 47, Nos, 1-4, pp 67-75

Correspondence Analysis As Applied To A 6x5 Contingency Data[1]

Liza F. Neri, Emeterio S. Solivas and Zita VJ. Albacea[2]

ABSTRACT

The study focuses on the application of correspondence analysis to a particular 6x5 contingency table. The population of the study is the crosstabulation of graduates from the College of Arts and Sciences, University of the Philippines Los Baños during the ten years period 1987-1996, classified into degree program and year graduated.

Keywords: singular value decomposition, pre- post-decomposition, biplot, row and column coordinates

1. Introduction

Correspondence analysis finds a low-dimensional graphical representation of the association between rows and columns of a contingency table, where categories of rows and columns are depicted as points in a Euclidean space. These points are determined from cell frequencies so that squared distances between certain sets of points in the derived space bear simple relationships to the original tabular entries. One is actually looking for a low-dimensional space, usually a plane, which reflects as accurately as possible the chi-square distances between the points (Greenacre 1994, p. 15).

Correspondence analysis consists of three parts: a pre-decomposition method where the original data or contingency table is transformed by certain transformation procedures; second, the transformed data is subjected to singular value decomposition (SVD) where a set or row and column vectors together with its associated singular values are summarized; and third, a post-decomposition method is applied wherein the row and column vectors are used to come up with the row and column coordinates or scores. Different post-decomposition methods result to different sets of coordinates for the row and column variables.

The paper deals with the description of correspondence analysis and its application to a 6x5 contingency table. Section 2 of the paper deals with matrices and its components as given by the singular value decomposition (SVD), the pre-decomposition method prior to correspondence analysis and the post-decomposition method which gives out the row and column coordinates.

The third section presents the population data in a 6x5 contingency table of number of graduates by degree course and year graduated.

The results of correspondence analysis are presented in the fourth section, including the resulting biplot.

---

[1] Major part of the paper was done as first author's Master of Science in Statistics thesis.

[2] Graduate student, Associate Professor and Assistant Professor, respectively, Institute of Statistics, College of Arts and Sciences, University of the Philippines Los Baños.

The last section summarizes the paper and describes correspondence analysis as an exploratory technique which facilitates interpretation with the use of graphical displays.

2. Theoretical Framework

2.1 Matrix and Singular Value Decomposition

In an I x J matrix A, represented by

$$A = [a_{ij}] \quad \text{where } i=1, 2, ..., I \text{ and } j=1, 2, ..., J,$$

the I rows represent row variable categories and the J columns, column variable categories. The a_{ij} denotes the observed value of the i^{th} row and the j^{th} column in contingency tables which is the number of individuals falling in category i of the row variable and category j of the column variable. The singular value decomposition (SVD) is the method used to decompose a matrix such as this, into row and column structures together with the associated singular values. It decomposes an I x J matrix A as the product of three matrices, U, Γ, and V^T. In symbols,

$$A = U \Gamma V^T \tag{1}$$

where U summarizes the information in the rows of A; the rows in U correspond to the rows in A. Similarly, V summarizes information in the columns of A; the rows in V correspond to columns in A. The columns in the U and V matrices represent the basic components in the structure of the data. The Γ is a diagonal matrix of positive non-zero numbers in descending order: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_K$, called the singular values of A, where k is the rank of A. The first entry λ_1, corresponds to the first column of U and to the first column of V while the second entry λ_2, corresponds to the second column of U and the second column of V, and so on. The values in Γ are the "weights" indicating the relative importance of each dimension in U and V. The columns of U and V, and the elements of Γ are ordered from most important to least important in the overall structure of A. The U and V are called left and right singular matrices, respectively, and must satisfy $U^TU=V^TV=I$. In correspondence analysis, SVD is used to reduce the dimensionality of the rows and columns by using the pre-standardized matrix of the contingency table.

2.2 Pre-decomposition

Correspondence analysis starts off with a contingency table denoted by N with I categories for the row variable and J categories for the column variable, denoted by

$$N = [n_{ij}] \quad \text{where } i=1, 2, ..., I \text{ and } j=1, 2, ..., J$$

and n_{ij} denotes the frequency in the i^{th} row and the j^{th} column. Define P as the correspondence matrix

$$P = [p_{ij}] = (1/n)\, N = [n_{ij}/n]$$

which rescales the original data matrix such that the sum of the elements equals one.

Comparisons can be made between row or column categories by taking their profiles. Profiles are a set of percentages, calculated for row or column frequencies. These are used instead of the original entries because the latter do not yield a meaningful interpretation of distances between row or column points. Row totals are denoted by $n_{i.}$ and column totals by $n_{.j}$. Hence, the vector of row profiles is $\boldsymbol{R} = \boldsymbol{N}\boldsymbol{D}_{\boldsymbol{R}}^{-1} = [n_{ij}/n_{i.}]$, where $\boldsymbol{D}_{\boldsymbol{R}}$ is the diagonal matrix of row totals. Similarly, the vector of column profiles is $\boldsymbol{C} = \boldsymbol{N}\boldsymbol{D}_{\boldsymbol{C}}^{-1} = [n_{ij}/n_{.j}]$, where $\boldsymbol{D}_{\boldsymbol{C}}$ is the diagonal matrix of column totals. Masses are associated with profiles and are defined as $r_i = n_{i.}/n$ and $c_j = n_{.j}/n$ for rows and columns, respectively. These are thought of as weights of profiles such that in a cloud of profiles in space, the average profile lies in an average or central position, but tends to lie more towards the profiles which have higher mass.

In terms of $\boldsymbol{P}$, the row profiles can be written as $\boldsymbol{D}_{\boldsymbol{r}}^{-1}\boldsymbol{P}$ where $\boldsymbol{D}_{\boldsymbol{r}}$ is the diagonal matrix of the row masses. Similarly, the column profiles can be written as $\boldsymbol{P}\boldsymbol{D}_{\boldsymbol{c}}^{-1}$ where $\boldsymbol{D}_{\boldsymbol{c}}$ is the diagonal matrix of column masses.

Prior to SVD, the data matrix $\boldsymbol{N}$ is transformed to remove differences in row and column totals while leaving the pattern of association within the table unaffected. The transformation will express each cell as a proportion and is given as follows:

$$n_{ij}^{*} = n_{ij} \Big/ \left(\sum_{i} n_{ij} \sum_{j} n_{ij} \right)^{1/2} = n_{ij} \Big/ \left(n_{i.} n_{.j} \right)^{1/2}.$$

Subjecting matrix $\boldsymbol{N}$ to SVD will give the left and right singular matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, and the matrix of singular values, Γ. Different pre-decomposition methods will give different values of $\boldsymbol{U}$, Γ, and $\boldsymbol{V}$. Using these matrices, the post-decomposition method follows.

2.3 Post-decomposition

A major objective of correspondence analysis is to find two sets of coordinates, one for the rows of the contingency table and another for the columns, in as few dimensions as possible, usually in two dimensions. These coordinates would make squared distances between row or column points correspond directly to squared distances between rows or columns of the contingency table.

Four possible row coordinates, denoted by DA, DAD, DAD1/2, and DAID1/2, are defined as follows:

$$DA = D_r^{-1/2} U = \left[\sqrt{\frac{n}{n_i}} \cdot u_{ij} \right]; \tag{2}$$

$$DAD = D_r^{-1/2} U \Gamma = \left[\sqrt{\frac{n}{n_{i.}}} \cdot u_{ij} \cdot \lambda_k \right]; \tag{3}$$

$$DAD1/2 = D_r^{-1/2} U \Gamma^{1/2} = \left[\sqrt{\frac{n}{n_{i.}}} \cdot u_{ij} \cdot \sqrt{\lambda_k} \right]; \text{ and} \tag{4}$$

$$DAID1/2 = D_r^{-1/2}U(\Gamma + I)^{1/2} = \left[\sqrt{\frac{n}{n_i}} \cdot u_{ij} \cdot \sqrt{\lambda_k + 1}\right] \quad (5)$$

Similarly for the column coordinates, DB, DBD, DBD1/2, and DBID1/2 are defined as follows:

$$DB = D_c^{-1/2}V = \left[\sqrt{\frac{n}{n_j}} \cdot v_{ij}\right]; \quad (6)$$

$$DBD = D_c^{-1/2}V\Gamma = \left[\sqrt{\frac{n}{n_j}} \cdot v_{ij} \cdot \lambda_k\right]; \quad (7)$$

$$DBD1/2 = D_c^{-1/2}V\Gamma^{1/2} = \left[\sqrt{\frac{n}{n_j}} \cdot v_{ij} \cdot \sqrt{\lambda_k}\right]; \text{ and} \quad (8)$$

$$DBID1/2 = D_c^{-1/2}V(\Gamma + I)^{1/2} = \left[\sqrt{\frac{n}{n_j}} \cdot v_{ij} \cdot \sqrt{\lambda_k + 1}\right]. \quad (9)$$

These coordinates are the options used by correspondence analysis. Since the DAD-DBB pair gives the most efficient biplot, this paper will apply only the said post-decomposition method.

After the post-decomposition, a biplot of the two-dimensional coordinate scores is constructed. The first coordinate corresponds to the vertical (Y) axis and the second coordinate to the horizontal (X) axis.

3. Methodology

The study considered the DAD-DBD post-decomposition normalization as the plotting option. This plotting option was provided by SAS through the CORRESP procedure.

The data set used as a population in this study consists of the graduates from the College of Arts and Sciences, UP Los Baños during the period 1987-1996. The resulting data set was a two-way contingency table by degree program and year of graduation. The row variable which is the degree program was classified into six categories (degree programs grouped according to areas of specialization), namely: (1) Communication Arts; (2) Sociology; (3) Statistics, Mathematics and Applied Mathematics; (4) Biology, Zoology, and Botany; (5) Chemistry and Agricultural Chemistry; and (6) Computer Science. The column variable was broken down into five categories, namely; (1) 1987-88; (2) 1989-90; (3) 1991-92; (4) 1993-94; and (5) 1995-96. Table 1 gives the data used in this study.

Table 1. Contingency table of CAS graduates by degree programs and year of graduation.

DEGREE PROGRAMS	YEAR OF GRADUATION				
	1987-88	1989-90	1991-92	1993-94	1995-96
Communication Arts	47	44	54	71	98
Sociology	25	35	24	48	79
Statistics, Mathematics & Applied Mathematics	146	121	137	120	162
Biology, Zoology & Botany	195	226	293	369	401
Chemistry & Agricultural Chemistry	31	40	57	51	60
Computer Science	49	80	106	83	113

The SPSS software was used to facilitate sampling and matrix computations.

4. Results And Discussion

The contingency table presented in Table 1 showed the dominance of Biology, Zoology and Botany graduates and followed by the graduates of Statistics, Mathematics, and Applied Mathematics degree programs. The lowest number of graduates were from Sociology, gaining through the years, together with Chemistry and Agricultural Chemistry degree programs. The inertia and chi-square decomposition of the data set are shown in Table 2. About 77% of the total chi-square value was explained by the first dimension alone. Together with the second dimension, 88% of the total inertia was explained.

Table 2. Inertia and chi-square decomposition of the population data.

SINGULAR VALUES	PRINCIPAL INERTIAS	CHI-SQUARES	χ^2 PERCENTAGES
0.113	0.0137	42.75	57.25
0.082	0.0068	22.83	30.58
0.050	0.0025	8.28	11.08
0.016	0.0002	0.81	1.08
	0.0222	74.67	

Table 3 contains the summary statistics for the row points. The quality column shows which category is represented well in the two-dimensional map. Statistics, Mathematics and Applied Mathematics together with Chemistry and Agricultural Chemistry were well-represented, 99.95% and 97.75% of their inertia, respectively, while the group of Biology, Zoology and Botany courses had the lowest with 68.67%. Not surprisingly, the Biology, Zoology and Botany category had the highest mass owing to its having the largest number of graduates for the past years. Conversely, Sociology has the lowest mass. A point with a high value of inertia means a point far from the average profile.

Table 3. Summary statistics of the row points

	QUALITY	MASS	INERTIA
Communication Arts	0.8525	0.0933	0.0552
Sociology	0.8524	0.0627	0.2307
Statistics, Mathematics & Applied Mathematics	0.9995	0.2039	0.4156
Biology, Zoology & Botany	0.6867	0.4410	0.1150
Chemistry & Agricultural Chemistry	0.9775	0.0710	0.0368
Computer Science	0.7109	0.1281	0.1467

The partial contribution to inertia for the row points is presented in table 4.

Table 4. Partial contributions to inertia for the row points \

	DIM 1	DIM 2
Communication Arts	0.0260	0.1053
Sociology	0.1775	0.3107
Statistics, Mathematics & Applied Mathematics	0.6657	0.1119
Biology, Zoology & Botany	0.1279	0.0189
Chemistry & Agricultural Chemistry	0.0016	0.1146
Computer Science	0.0014	0.3386

The partial contributions to inertia measures to what extent the geometric orientation of an axis is determined by the single variable categories. Hence, for the first dimension, Statistics, Mathematics and Applied Mathematics with Sociology are relatively important. These two courses define the first dimension. Computer Science and Sociology are the courses which define the second dimension. These courses define the opposite poles of each dimension.

Table 5 shows the row coordinates of the row variable (degree programs).

DEGREE PROGRAMS	DIM 1	DIM 2
Communication Arts	-0.06	-0.09
Sociology	-0.19	0.18
Statistics, Mathematics & Applied Mathematics	0.20	-0.06
Biology, Zoology & Botany	-0.06	0.02
Chemistry & Agricultural Chemistry	0.02	0.10
Computer Science	0.01	0.13

The first dimension separated Communication Arts, Sociology, Biology, Zoology and Botany from Statistics, Mathematics, and Applied Mathematics; Chemistry and Agricultural Chemistry and Computer Science. Dimension 2, on the other hand, grouped together Communication Arts and Statistics, Mathematics, and Applied Mathematics as one, Sociology, Biology, Zoology, and Botany; Chemistry and Agricultural Chemistry; and Computer Science as another.

The results of the column categories are given in tables 6 and 7. All columns are represented by a capital letter, that is, 1987-88 is represented by the letter A, 1989-90 by B, 1991-92 by C, 1993-4 by numbers 1 to 5, that is, 1987-88 by 1, 1989-90 by 2 and so on.

From table 6, the quality column shows that the years 1991-92 and 1987-88 are well-represented in the two-dimensional map, with 99.07% and 97.33% of their inertias, respectively. The category 1989-90 is poorly represented with only a value of 44.6%. The years 1995-96 has the highest mass, meaning the last two years got the highest number of graduates during the period. Looking at the mass column, the values are increasing owing to the increasing number of graduates through the years.

Table 6. Summary statistics of the column points

	QUALITY	MASS	INERTIA
1987-88	0.9733	0.1465	0.3708
1989-90	0.4460	0.1623	0.0563
1991-92	0.9907	0.1994	0.2075
1993-94	0.7068	0.2205	0.1803
1995-96	0.8607	0.2713	0.0851

Table 7 shows the partial contributions to inertia for the column points. The years 1987-88 and 1993-94 define the first dimension while 1991-92 and 1987-88 define the second dimension. These categories were represented by the points on the opposite poles of each dimension.

Table 7. Partial contributions to inertia for the column points

	DIM 1	DIM 2
1987-88	0.4903	0.2623
1989-90	0.0362	0.0143
1991-92	0.0568	0.5658
1993-94	0.2219	0.0014
1995-96	0.1948	0.1561

On the other hand, the coordinates of the column variable (year of graduation) are summarized in Table 8.

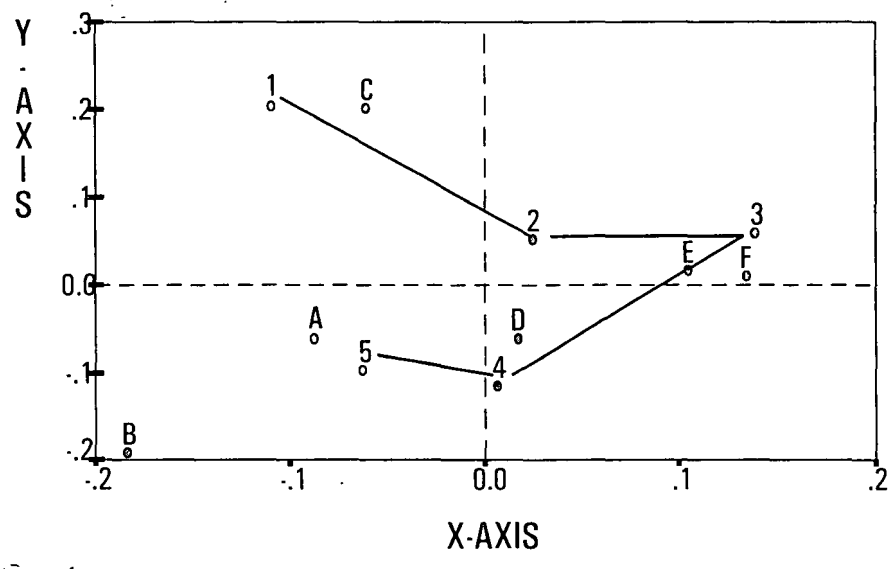
Table 8. The resulting two-dimensional coordinates of years of graduation

YEARS OF GRADUATION	DIM 1	DIM 2
1987-88	0.21	-0.11
1989-90	0.05	0.02
1991-92	0.06	0.14
1993-94	- 0.11	0.01
1995-96	-0.10	-0.06

The first six years, represented by categories 1987-88, 1989-90, and 1991-92 were separated from the last four years, that is, categories 1993-94 and 1995-96, for the first dimension. The second dimension grouped together 1987-88 and 1995-96, the first and last categories, respectively.

The plot of the row and column points gives an overall view of both variables on the plane. The plot of the DAD-DBD coordinates is shown in Figure 1.

Figure 1. Two dimensional plot of the population data using DAD-DBD coordinates.



Legend:

A	Communication Arts,	1	1987-88
B	Sociology	2	1989-90
C	Statistics, Mathematics & Applied Mathematics	3	1991-92
D	Biology, Zoology & Botany	4	1993-94
E	Chemistry & Agricultural Chemistry	5	1995-96
F	Computer Science		

The vertical axis defined the direction of change, 1987-88 at the topmost part and 1995-96 at the lower part. Computer Science, Chemistry and Agricultural Chemistry were

clustered close together away Sociology. This dimension separated mathematically inclined courses from non-mathematical ones. The other axis separated Communication Arts, Sociology, Statistics, Mathematics and Applied Mathematics from others. A row and column points close together imply a high frequency in that cell. A high number of Biology, Zoology and Botany graduates occurred in 1993-1994.

5. Summary and Conclusion

Correspondence analysis is an exploratory technique which describes a table of numerical information in the form of contingency tables. It transforms such tables into graphical displays which facilitates interpretation.

For the population data of CAS graduates for 1987-1996, the highest number of graduates was from Biology, Zoology and Botany courses. On the other hand, the years 1995 and 1996 produced the largest number of graduates.

From the results of correspondence analysis using the population data set, the first two dimensions explained 88% of the total inertia. Only the Statistics, Mathematics and Applied Mathematics category together with the Chemistry and Agricultural Chemistry category were explained pretty well in the map. For the column categories, the years 1991-92 and 1987-88 were the categories which were represented well in the two-dimensional map. The first dimension was defined by the Statistics, Mathematics and Applied Mathematics and the Sociology row categories together with the years 1987-88 and 1993-94. Dimension 2 was defined by Computer Science and Sociology categories combined with the years 1991-92 and 1987-88.

6. References

ANDERSEN, E. B. 1990. *The Statistical Analysis of Categorical Data.* Springer-Verlag, Germany.

GREENACRE, M. and BLASIUS, J. 1994. *Correspondence Analysis in the Social Sciences Recent Developments and Applications.* Academic Press, London.

JACKSON, J. E. 1991. *A User's Guide to Principal Components.* John Wiley and Sons, USA.

JAMBU, M. 1991. *Exploratory and Multivariate Data Analysis.* Academic Press, USA.

KRZANOWSKY, W. J. 1988. *Principles of Multivariate Analysis A User's Perspective.* Oxford University Press, USA.

SAS Institute Inc. (eds). *SAS User's Guide Proc Corresp.* SAS Institute Inc. Cary, NC, USA.