

Correspondence Analysis applied to psychological research

Laura Doey and Jessica Kurta

University of Ottawa

Correspondence analysis is an exploratory data technique used to analyze categorical data (Benzecri, 1992). It is used in many areas such as marketing and ecology. Correspondence analysis has been used less often in psychological research, although it can be suitably applied. This article discusses the benefits of using correspondence analysis in psychological research and provides a tutorial on how to perform correspondence analysis using the Statistical Package for the Social Sciences (SPSS).

Correspondence analysis (CA) has become most popular in fields such as ecology, where data is collected on the abundance of various animal species in specific sampling units/areas (ter Braak, 1985). The amount of data involved in these samples makes it difficult to get a clear sense of the data at a glance (Palmer, 1993). With the use of CA, an exploratory data technique for categorical data, ecologists have been able to transform these complicated tables into straightforward graphical displays (Hoffman & Franke, 1986). Ecological data is multidimensional making visualization of more than two dimensions difficult. Correspondence Analysis is an ideal technique to analyze this form of data because of its ability to extract the most important dimensions, allowing simplification of the data matrix (Palmer). However, this statistical technique merits further attention within the field of psychological research. In fact, the lack of awareness of this useful statistical method puts psychological researchers at a disadvantage. To demonstrate that CA can be suitably applied to psychological research, this paper will use a research question from community psychology. Additionally, due to the gap in the literature on how to perform CA using statistical software programs, the current paper will describe how to perform CA using SPSS software.

History of CA

CA originated approximately 50 years ago and has been referred to by a variety of names such as dual scaling, method of reciprocal averages and categorical discriminant

analysis (Hoffman, & Franke, 1986). These names are thought to stem from the fact that CA has been used to analyze many different questions and has been given a different name each time it answers a different question. These differing names may also be attributed to the different versions of CA that were being developed in several countries simultaneously (Abdi, & Williams, 2010; StatSoft, Inc., 2010).

Jean-Paul Benzecri, French linguist and data analyst, was an important figure in the initiation of the modern application of CA in the 1960s and 1970s, making CA popular in France (Greenacre, 2007; StatSoft, Inc., 2010). Before 1970, CA was relatively unknown in English speaking countries with only one English publication on CA in existence, written by Benzecri (Clausen, 1988). The application of CA eventually spread to countries such as Japan and England (Clausen, 1988). However, compared to other statistical techniques little was published in English on the topic of CA regardless of its inclusion in American statistical packages such as SPSS in the 1980s (Clausen, 1988). Today CA is popular in some areas of the social sciences, such as marketing and ecology (Hoffman, & Franke, 1986; ter Braak, 1985).

What is CA?

Unlike the many statistical techniques that test hypotheses that have been formed a priori, CA is an exploratory data technique that explores categorical data for which no specific hypotheses have been formed (Storti,

2010). More specifically, CA analyzes two-way or multi-way tables with each row and column becoming a point on a multidimensional graphical map, also called a biplot (Storti, 2010). This biplot typically consists of two or three dimensions (StatSoft, Inc., 2010). Rows with comparable patterns of counts will have points that are close together on the biplot and columns with comparable patterns of counts will also have points that are close together on the biplot (SAS Institute Inc., 2010). The row and column points are shown on the same graphical display allowing for easier visualization of the associations among variables (Storti).

CA uses the chi-square statistic—a weighted Euclidean distance—to measure the distance between points on the biplot (see Clausen, 1988, pp. 12 for equation). In other words, the chi-square distance measures the association between variables. It is important to note that the chi-square distance can be used to examine the associations between categories of the same variable but not between variables of different categories. For example, if the types of mental health services (i.e., community outpatient centers, hospitals etc.; row data) available in various provinces (column data), with abundance as the entries, was examined, it would be possible to obtain the chi-square differences between types of mental health services and between provinces but not between types of mental health services and provinces (Clausen). Due to the fact that CA is a non-parametric statistic, there is no theoretical distribution to which the observed distances can be compared. Therefore, contrary to the classical utilization of the chi-square test, when applied to CA the chi-square test does not reveal whether the association between variables is statistically significant. CA does not support significance testing and is instead used post-hoc as an exploratory method (StatSoft, Inc., 2010).

Dimensions

The goal of CA is to explain the most inertia, or variance, in the model in the least number of dimensions. One way to understand dimensions is that they are comparable to a principal component in factor analysis, the association between the categorical variables (Statsoft Inc., 2010). Some researchers state that the maximum number of dimensions needed to exactly represent the table is the number of rows minus one, or the number of columns minus one (Greenacre, 1984, as cited by Moser, 1989). Other researchers use slightly different rules of thumb when deciding how many dimensions to retain (see Hair et al., 2007). The researcher typically chooses enough dimensions to meet the research objectives (usually two or three). Given the goal of this paper as an introduction to the topic of CA, it will not explore the mathematics involved in calculating dimensions (see Clausen, 1988, p. 17 for additional details).

Comparison of CA to Other Statistical Techniques

CA is similar to a number of other statistical techniques. For example, CA and factor analysis are both exploratory methods that attempt to explain the variance in a model and decompose this variance into a low-dimensional representation. In other words, both these techniques attempt to reduce the variability of a model by calculating the minimum number of factors that can explain the most variability in the model (Clausen, 1988; Statsoft Inc., 2010). However, factor analysis determines which variables go together to explain the most covariance between descriptors, whereas CA determines which category variables are associated with one another.

CA is also similar to Principal Component Analysis (PCA). In fact, CA has been described as a “generalized” or “weighted” PCA of a contingency table (Abdi, 2010; SAS Institute Inc., 2010). CA and PCA both present data in a low-dimensional plane that accounts for the model’s main variance. The distances between the points in this low-dimensional space closely resemble the original distances from the high-dimensional space (Fellenberg, Hauser, Brors, Neutzner, Hoheisel, Vingron, 2001). However, PCA extracts which variables explain the largest amount of variance in the data set, whereas the focus of CA is to examine the associations among variables (Fellenberg et al., 2001).

In addition, both CA and cluster analysis are exploratory methods which sort variables based on their degree of correspondence to facilitate the analysis of data, but are not appropriate methods to be used for significance testing (StatSoft, Inc., 2010). Cluster analysis discovers whether different variables are related to one another, whereas CA goes a step further to explain how variables are related (StatSoft, Inc.) Lastly, multidimensional scaling (MDS) is similar to CA in that both methods examine the association between categories of rows and columns, produce a map of these associations and determine the dimensions that best fit the model (SAS Institute Inc., 2010; StatSoft, Inc., 2010). Additionally, both multidimensional scaling and CA have few assumptions that must be met in order for the solution to be accurate (SAS Institute Inc.).

Benefits of CA

One of the benefits of CA is that, as previously mentioned, it can simplify complex data from a potentially large table into a simpler display of categorical variables while preserving all of the valuable information in the data set. This is especially valuable when it would be inappropriate to use a table to display the data because the associations between variables would not be apparent due to the size of the table. It is also important to note that most

other exploratory statistical techniques do not provide a plot of associations among variables.

When other statistical techniques cannot be used to analyze data because certain assumptions are not met, CA becomes useful due to its flexible data requirements (Hoffman & Franke). For example, when a Likert scale is used to collect data, the spaces between descriptors (i.e., “almost never”, “sometimes” and “often”) are not necessarily equivalent. For example, the distance between “almost never” and “sometimes” is not necessarily equivalent to the distance between “sometimes” and “often”. In this type of scenario, CA is a useful technique because it focuses mainly on how variables correspond to one another and not whether there is a significant difference between these variables.

If one wishes to analyze continuous data with CA, the data can be categorized and subsequently analyzed as discrete data. In addition, CA demonstrates how variables are associated by the approximate distance of points to one another on the biplot, and not simply that they are associated. Another benefit of CA is that it can reveal relationships that would not be identified using other non-multivariate statistical techniques, such as performing pairwise comparisons. Moreover, CA presents data using two dual displays—one display for the row data and one display for the column data. This makes analysis of the data easier compared to the many statistical techniques that do not provide dual displays (Hoffman, & Franke, 1986). CA makes it easy to add supplementary data points that may aid in the interpretation of the model into the analysis post-hoc. In other words, CA allows for the addition of row or column points that carry zero inertia to the biplot after it has been constructed. Lastly, CA is also good way to examine data validity and facilitates the treatment of outliers (Fellenberg et al., 2001; Hoffman, & Franke).

Assumptions of CA

Violation of the following assumptions may make the conclusions drawn about the association among variables inaccurate and the biplot a less valuable guide for analyzing the data (Garson, 2008). Firstly, homogeneity of variance across row and column variables must be met (Garson). CA assumes that the statistical properties are similar across rows and columns. For example, there must not be any empty variables (i.e., variables for which all entries consist of zeros). Secondly, CA assumes that the data being analyzed is discrete; however, originally continuous variables can be categorized into discrete variables. Third, the data should be made up of several categories (typically more than three); if CA is used to analyze only two or three categories this analysis is unlikely to be more informative

than the original table itself (Garson). Fourth, all values in the frequency table must be non-negative so that the distances between the points on the biplot are always positive (Garson). CA does not make any distributional assumptions (i.e., assumptions of normality; Garson).

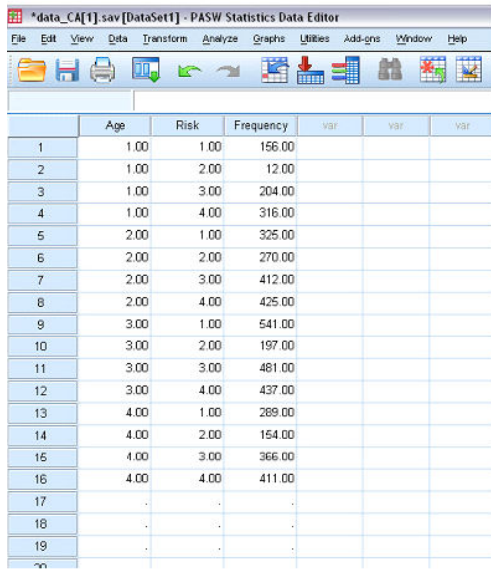
CA in Other Research Areas

CA is a valuable statistical technique in part because it can be used by all researchers and professionals who analyze categorical variables. Therefore, CA is used to analyze research questions across many domains (Greenacre, 1993).

It has been mentioned that CA has become more popular in ecological research. CA is also popular in marketing research because researchers in this area frequently collect categorical data due to the simplicity of this collection method. (Hoffman, & Franke, 1986). Along with ecology and marketing, CA has also been applied in other areas such as sociology, archaeology, geology and medicine (Greenacre, 2007). Due to the wide application of CA across other domains, the benefits of CA and the simplicity of collecting categorical data in psychological research, it is surprising that CA is not more commonly used in psychology.

In fact, CA could be suitably applied to many different domains within psychological research. CA would be especially relevant for the many burgeoning areas in psychology (i.e., positive psychology) in which it is necessary to determine which questions could be asked or which hypotheses could be formed, due to its exploratory nature (Fellenberg et al., 2001). In addition, categorical data is often easier and less time consuming to collect in psychological research. For example, it is less complex to ask individuals whether or not they have been mildly depressed, moderately depressed or severely depressed over the last six months compared to asking them to describe the severity of their depressive symptoms. CA is often used when a researcher wants to get a general idea of a population before conducting a more complex study. In social psychology, CA may be used to look at the relation between the prevalence of the various sexual orientations in each geographical region in a city. From these results, particular support programs could be put in place to target at-risk individuals who experience difficulty expressing their sexual orientation. In developmental psychology, CA could be used to look at the associations between attachment styles and the types of play children engage in.

As seen above, CA merits more attention within the field of psychological research. To illustrate how CA can be applied within psychology, this paper will use a research question from community psychology to explain how CA is performed using SPSS. This research question will look at



	Age	Risk	Frequency	V3F	V3F	V3F
1	1.00	1.00	156.00			
2	1.00	2.00	12.00			
3	1.00	3.00	204.00			
4	1.00	4.00	316.00			
5	2.00	1.00	325.00			
6	2.00	2.00	270.00			
7	2.00	3.00	412.00			
8	2.00	4.00	425.00			
9	3.00	1.00	541.00			
10	3.00	2.00	197.00			
11	3.00	3.00	481.00			
12	3.00	4.00	437.00			
13	4.00	1.00	269.00			
14	4.00	2.00	154.00			
15	4.00	3.00	366.00			
16	4.00	4.00	411.00			
17	.	.	.			
18	.	.	.			
19	.	.	.			
20	.	.	.			
21	.	.	.			

Figure 1. Entering data into SPSS.

the behaviours that young individuals are most likely to be at risk for developing (i.e., substance abuse, dropping out, violence, mental health issues depending on their age group (i.e., 10-12 years old, 13-15 years old, 16-18 years old, and 19-21 years old).

SPSS tutorial

Using our community psychology example with fabricated data, we can input the data into SPSS following these steps:

Step 1: Entering the Data

First, three variables will be created; variable 1 will be Age, variable 2 will be Risk, and variable 3 will be Frequency. It is necessary to label each variable depending on the number of categories within the variable. For Age, there are four different age groups; 10-12, 13-15, 16-18 and 19-21. In the "Values" tab in SPSS in Variable View, we will give four different values to our Age variable. Value 1 will be age group 10-12, value 2 will be age group 13-15, and so on. The Risk variable also has four categories, and we will label these in the same way. Value 1 will be the risk of Substance Abuse, value 2 will be Drop Out, value 3 will be Violence and value 4 will be Mental Health.

Data can be entered in this way based on the assumption that the frequency for each variable has previously been calculated. If the frequencies of each variable are not calculated and only raw data is available, it is possible to run the same analysis without using Step 2, described below. However, analyzing raw data in this way is uncommon because of the substantial volume of the file. In our example, inputting raw data would produce a file of 4996 lines.

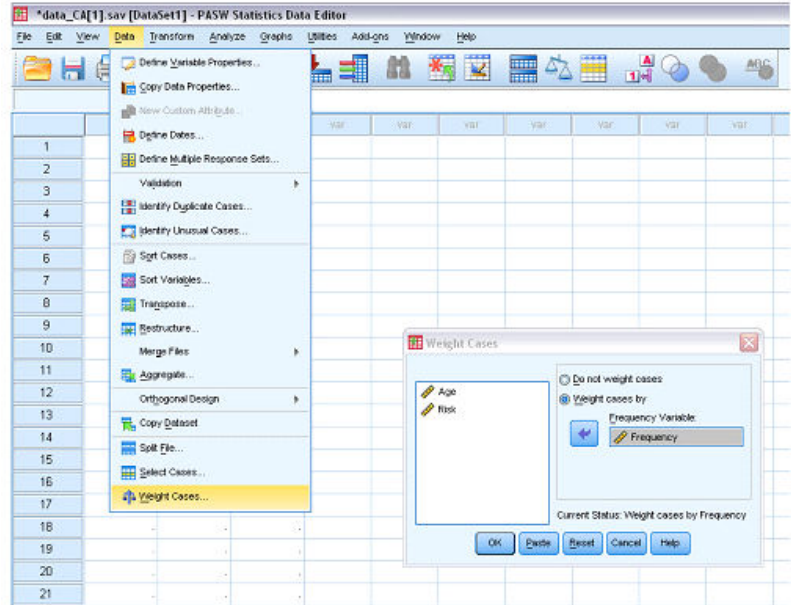


Figure 2. Weighting cases by frequency.

In the Data View spreadsheet of our SPSS file, we will have 16 entries, such that each Age category is matched with each Risk category. Our Frequency variable will give the frequency of occurrence in our sample data of each age group with each risk factor (see Figure 1). For example, we can see that Age group 10-12 combined with the Risk factor of Substance Abuse has a frequency of 156 out of our sample of 4996 subjects.

Step 2: Weight the data

Once all of the data has been inputted into SPSS, the next step is to weight the cases by frequency. To do this, click on Data → Weight Cases → Weight Cases by: Frequency Variable: Frequency → OK (see Figure 2). This is done to inform SPSS that the frequencies need not be compiled.

Step 3: Running the Analysis

Once the data is weighted by Frequency, it is now possible to run the analysis in SPSS. To do this, click on Analyze → Dimension Reduction → Correspondence Analysis (see Figure 3). Next, insert each of the Age and Risk variables into the Row and Column profiles respectively. It does not matter which variable is on which axis when running the analysis.

In order to run the full analysis, the range of the rows and columns must be defined for each variable. Under Row, click Define Range. In our Age variable, we only have four categories. We will include all four categories here. Beside Minimum value, enter the value 1; beside Maximum value, enter the value 4; click Update and Continue.

Using CA, it is possible to run a preliminary analysis in which only part of the data is analyzed. With large data sets,

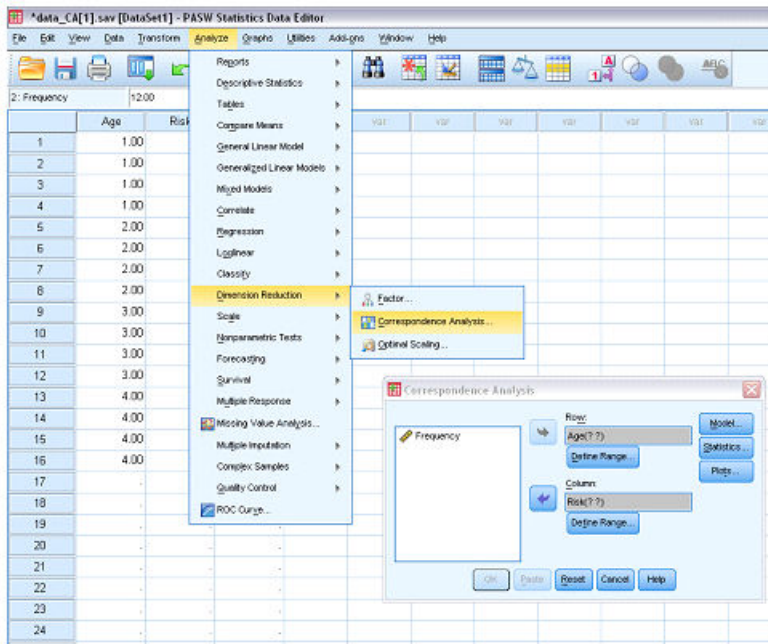


Figure 3. Running correspondence analysis in SPSS.

researchers may choose to include only part of their data in the analysis initially, and later include all variables. This may be useful if there are numerous categories within each variable and the researcher wishes to reduce the noise and focus on a particular association within the dataset before looking at the global picture. In our example, all of the data will be used for analysis.

It is also necessary to define the range for Risk, including all four categories. Beside Minimum value, enter the value 1; beside Maximum value, enter the value 4; click Update and Continue. SPSS will now run a CA using all of our data. If there were more than four categories for the row or column variables, the maximum number would be entered when defining the ranges in order to analyze the entire dataset.

By clicking on Model, the researcher can specify how they would like SPSS to produce the results of the analysis. The first option is choosing the number of dimensions to include in the solution. In this example, the number of dimensions was set at two. This is the default setting in SPSS when running CA; however the dimensions can be increased at the discretion of the researcher depending on the type of research being done. The Distance Measure should be set to Chi square, the Standardization Method should be set to 'Row and column means are removed', and the Standardization Method, depending on how you would like to interpret your results, can be chosen (see Figure 4). In this example, Symmetrical was chosen in order to be able to compare rows to columns (other standardization methods are described below); click Continue.

The next option, Statistics, allows the researcher to choose which output tables to include in the output (see

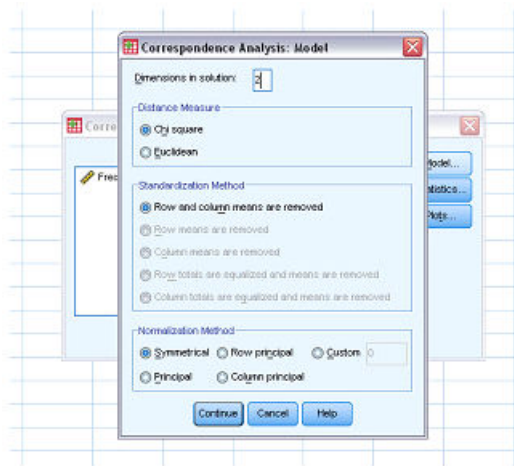


Figure 4. Setting the parameters of the model.

Figure 5). By default in SPSS, the first three options are chosen. In our example, we chose to also include Row and Column Profiles, as well as Confidence Statistics for Row Points and Column Points. Again, this is at the discretion of the researcher whether or not to include certain tables; click Continue. For more information on the options provided in this step, click on Help in the bottom right corner of the Correspondence Analysis: Statistics box.

The final option, Plots, allows the researcher to choose how the analysis should be displayed graphically. This is the most important part of the output (see Figure 6). Under Scatterplots, the researcher can choose to display the biplot graphically, as well as only row points and only column points in a separate graph. Displaying Row points or Column points in a scatterplot is useful when comparing row points or column points in order to simplify the data that would be produced in a biplot. Line plots can be used to display row or column categories after standardization and normalization has been performed. In this example, we did not ask for line plots. Finally, Plot Dimensions allows the researcher to choose whether or not to include all of the dimensions that SPSS was asked to produce for the analysis, or restrict the dimensions included in the graphical representation of the data. In this example, SPSS was asked to 'Display all dimensions in the solution'; click Continue.

At this point, the parameters for CA have been set and SPSS can now run the final analysis; click OK (see Figure 7).

Step 4: Interpreting the Output

SPSS will produce a tabulation table, shown in Table 1, called a Correspondence Table. The data given here is based

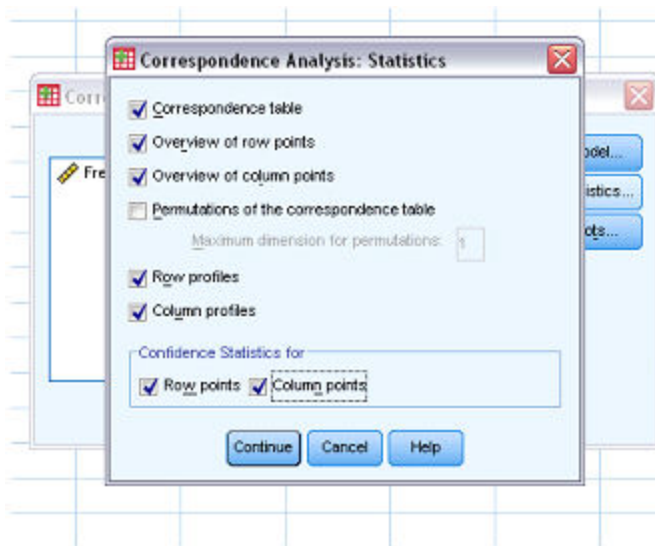


Figure 5. Choosing the statistics to be shown in the SPSS output.

on the data entered for the age and risk variables in SPSS. It will include the frequencies for each of the row and column categories that were given and produce a summation for each of the row categories and column categories called the 'Active Margin'. For example, it can be seen below that all of the frequencies for the age group 10-12 across each of the Risk categories sum to 688. Similarly, all of the frequencies for the risk factor Violence across each of the Age categories sums to 1463.

Next in the SPSS output is a Row Profiles table as shown in Table 2. This table gives the weighted frequency of each of the row points, such that the total for the whole row will sum to 1.

The row profiles are calculated by taking each row point and dividing it by its respective Active Margin for that row. For example, for the age group 10-12 and the risk factor Substance Abuse, the frequency (as given by the Correspondence Table) is 156. The Active Margin for that row is 688. Therefore, $156/688 = .227$. This is done for each value in the table.

Similarly, SPSS produces a table called Column Profiles (Table 3), and these are calculated in the same way as the Row Profiles table.

From the Correspondence Table, we see that age group 16-18 with the risk factor Mental Health has a frequency of 437. The Active Margin for the column of Mental Health is 1589. Therefore, $437/1589 = .275$.

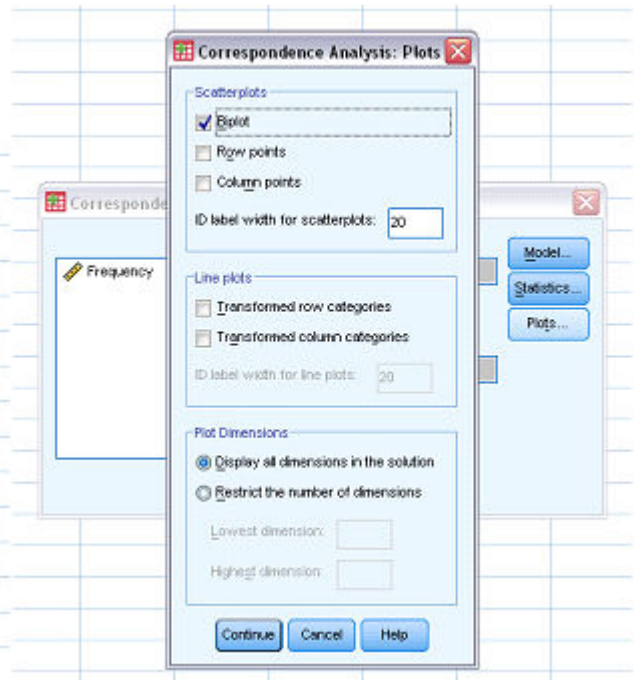


Figure 6. Determining the graphic display of the biplot.

The Summary table (Table 4) is the most important table provided in the SPSS output for CA.

CA uses the chi-square statistic to test for total variance explained, along with the associated probability. The chi-square statistic is high when there is a high correspondence between the rows and columns of a table (Fellenberg, Hauser, Brors, Neutzner, Hoheisel, Vingron, 2001). The first thing to look at in the summary table is whether or not the model is significant. In this example, our model is highly significant at the .000 level, with an alpha of .05 and a chi-square value of 210.373. We also see that SPSS has generated three dimensions to explain our model. In CA, SPSS only produces dimensions that can be interpreted, rather than including all dimensions that explain something about the

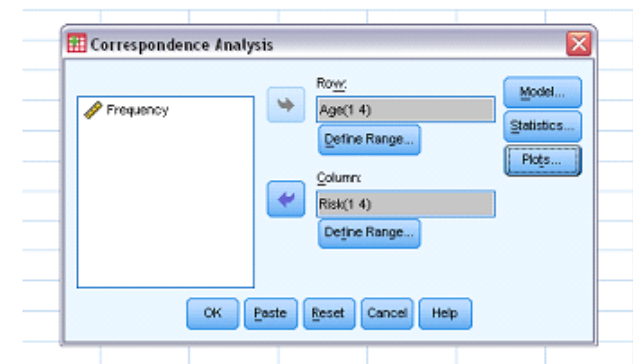


Figure 7. A view of the final window before running the analysis.

Age	Risk				
	Substance Abuse	Drop Out	Violence	Mental Health	Active Margin
10-12	156	12	204	316	688
13-15	325	270	412	425	1432
16-18	541	197	481	437	1656
19-21	289	154	366	411	1220
Active Margin	1311	633	1463	1589	4996

Table 1. Correspondence Table.

Age	Risk				
	Substance Abuse	Drop Out	Violence	Mental Health	Active Margin
10-12	.227	.017	.297	.459	1.000
13-15	.227	.189	.288	.297	1.000
16-18	.327	.119	.290	.264	1.000
19-21	.237	.126	.300	.337	1.000
Mass	.262	.127	.293	.318	

Table 2. Row Profile Table.

model. For this reason, inertia does not always add up to 100%. The Inertia column gives the total variance explained by each dimension in the model. In our model, the total inertia (total variance explained) is 4.2%. This indicates that for our model, knowing something about Age explains around 4% of something about Risk and vice versa. This association is weak, but still highly significant as indicated by our chi square statistic.

Each dimension is listed according to the amount of variance explained in the model. Dimension 1 will always explain the most variance in the model, followed by Dimension 2 and so on. In this example, Dimension 1 explains 3% of the total 4.2% of variance accounted for. Furthermore, Dimension 2 explains 1.2% of the total 4.2% of variance accounted for. Dimension 3 explains 0% of the total variance accounted for, and would therefore be dropped from further analysis. The Singular Value column gives the square roots of the eigenvalues, which describes “the maximum canonical correlation between the categories of the variables in analysis for any given dimension” (Garson, 2008). In CA, eigenvalues and inertia are synonymous in that, “each axis has an eigenvalue whose sum equals the inertia of the cloud (mass of points; Benzecri, 1992).

The values in the Proportion of Inertia column give the percent of variance that each dimension explains of the total variance explained by the model. In this example, Dimension 1 explains approximately 71% of the total 4.2% of variance explained in the model. Furthermore, Dimension 2 explains approximately 29% of the 4.2% of variance explained in the model. Dimension 3 explains too little of the total variance explained to be kept for further analysis.

There is no “rule of thumb” or criteria for keeping or rejecting dimensions for analysis based on proportion of inertia; it depends on the research question and the researcher decides what is clinically significant versus statistically significant for any given case. In essence, this example dictates that there are two dimensions that can explain the most variance between risk factors and age group. Some research questions may reveal that three dimensions are necessary to explain most of the variance.

The Overview Row Points (Table 5) gives information on how each of the row points is plotted in the final biplot. The ‘Mass’ column in this table indicates the proportion of each age group with respect to all age groups in the analysis. The column ‘Score in Dimension’ indicates the coordinates in each dimension (1 and 2) where each row category will be situated on the biplot. Inertia again reflects variance. The ‘Contribution’ column reflects how well each of the points load onto each of the dimensions, as well as how well the extraction of dimensions explains each of the points. In this example, we see that the 10-12 age group loads heavily on Dimension 1 (74%) and not heavily on Dimension 2 (~ 2%). It can also be seen that the extraction of Dimension 1 explains 99% of the variance in the 10-12 age group across risk factor, whereas the extraction of Dimension 2 only explains around 1% of the variance in the 10-12 age group across risk factor. As seen in Table 6, the Overview Column Points gives the same information for the plotting of column points on the biplot.

In this example, the risk factor for Drop Out loads well onto Dimension 1 (65%), and not as well on Dimension 2 (~ 18%). Furthermore, Dimension 1 explains around 90% of the

Age	Risk				
	Substance Abuse	Drop Out	Violence	Mental Health	Mass
10-12	.119	.019	.139	.199	.138
13-15	.248	.427	.282	.267	.287
16-18	.413	.311	.329	.275	.331
19-21	.220	.243	.250	.259	.244
Active Margin	1.000	1.000	1.000	1.000	1.000

Table 3. Column Profile Table.

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.173	.030			.713	.713	.011	.102
2	.110	.012			.285	.998	.014	
3	.008	.000			.002	1.000		
Total		.042	210.373	.000 ^a	1.000	1.000		

a. 9 degrees of freedom

Table 4. Summary Table.

Overview Row Points^a

Age	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		Total
					1	2	1	2	
10-12	.138	-.962	.111	.022	.735	.015	.991	.008	1.000
13-15	.287	.373	.336	.010	.230	.295	.660	.339	.999
16-18	.331	.124	-.456	.008	.029	.630	.104	.896	1.000
19-21	.244	-.063	.163	.001	.006	.059	.181	.767	.947
Active Total	1.000			.042	1.000	1.000			

a. Symmetrical normalization

Table 5. Overview Row Points.

variance in Drop Out across age group, and Dimension 2 explains around 10% of the variance in Drop Out across age group.

Tables 7 and 8, Confidence Row Points and Confidence Column Points, provide the standard deviations of row and column scores in each dimension, which is used to assess the precision of the estimates of points on their axes, much like confidence intervals are used in other statistical analyses.

Finally, SPSS produces a biplot, which provides a visual display of each of the values in the dataset plotted with their axes. This provides a global view of the trends within the data. In this example, because only two dimensions were extracted, SPSS can display the results in 2D in the form of a biplot. In the event that three dimensions would be used, a 3D graph would be produced to represent each dimension. When using a biplot, the chi-square statistic reveals the strength of trends within the data, which is based on the point distances of categories. The distance between any row points or column points gives a measure of their similarity

(or dissimilarity). Points that are mapped close to one another have similar profiles, whereas points mapped far away from one another have very different profiles. Distances between row and column points are interpreted differently. Only general statements can be made about observed trends; precise conclusions cannot be drawn. Because we asked SPSS to standardize our data using symmetrical normalization, we can compare rows to columns in a general fashion. Standardization in CA allows for a more evenly weighted distribution among large differences and small differences in distances between points, so that they can be compared without larger differences skewing the data and overbearing the smaller differences (Storti, 2010). Symmetrical normalization is a technique used to standardize row and column data so as to be able to make general comparisons between the two. Other forms of standardization allow you to compare row variable points or column variable points, or rows or columns, but not rows to columns (see Garson, 2008 for

Risk	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		Total
					1	2	1	2	
Substance Abuse	.262	.087	-.511	.008	.011	.626	.044	.955	.999
Drop Out	.127	.946	.392	.022	.654	.178	.902	.098	1.000
Violence	.293	-.024	.006	.000	.001	.000	.374	.016	.390
Mental Health	.318	-.426	.260	.012	.333	.196	.809	.190	.999
Active Total	1.000			.042	1.000	1.000			

a. Symmetrical normalization

Table 6. Overview of Column Points.

further information on other standardization techniques for CA).

All of our data has been graphically represented by the biplot above. With the caveat that this particular model only explains ~4% of risk factors based on age, some general trends can be seen. For example, we see that the age group of 16-18 is particularly at risk for problems with Substance Abuse. We also see that the age group 10-12 is not particularly at risk for Drop Out, but is more at risk for problems with Mental Health. Age group 19-21 appears to be more prone risk factors such as Violence and Mental Health rather than Substance Abuse and Drop Out. Age group 13-15 appears to have a marginal risk for Violence, Mental Health and Drop Out, but not as much for Substance Abuse.

Conclusion

CA is a statistical technique that is used primarily by social scientists and behavioural researchers to explore the relations among multivariate categorical variables (de Leeuw, 2005; Hoffman & Franke, 1986). CA is used less frequently in psychological research than in other areas, but could be suitably applied to various psychological research questions. In fact, psychological researchers would be at a disadvantage if they were not aware of the many benefits of CA, especially the graphical map this statistical technique provides, which facilitates the visualization of the associations between the rows and columns of a table. The goal of the current paper is to show how CA can be used to examine psychological data using the example of the associations between various age groups and degree of risk for developing problematic behaviors (i.e., substance abuse, dropping out, violence, mental health issues). This is only one of the many research questions that could be explored within the realm of psychology using CA.

Furthermore, with the results that are found using CA, additional research can be done to answer more specific research questions. For example, our data shows that 16-18 year olds are most at risk for substance abuse. With this information, a social psychologist could explore what makes

16-18 year olds more susceptible to substance abuse or what substances are most commonly abused in this age group. This could facilitate prevention and intervention programs in targeting at-risk individuals within this age group, as well as discovering areas of resilience within this age group. In general, CA provides an extremely useful general picture of associations between variables, and follow-up statistics can provide a more in-depth look at a particular research in question.

References

Abdi, H., & Williams, L. J. (2010). *Correspondence Analysis. Encyclopedia of Research Design*. Thousand Oaks, CA: Sage

Benzecri, J.-P. (1992). *Correspondence Analysis Handbook*. New York: Marcel Decker.

Böckenholt, U., & Takane, Y. (1994). *Linear constraints in correspondence analysis*. In Greenacre, M., & Blasius, J. (Eds.), *Correspondence Analysis in the Social Sciences: Recent Developments and Applications* (pp. 112-127). London: Academic Press.

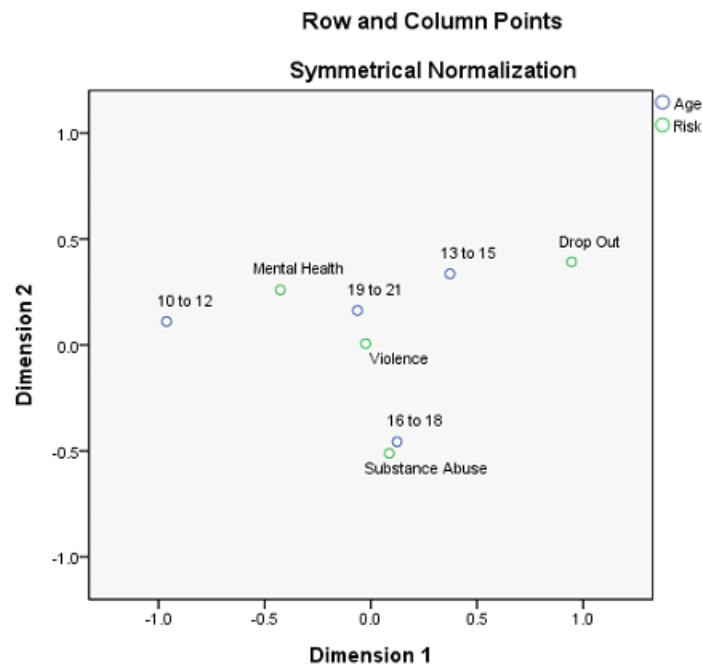
Clausen, S. E. (1988). *Applied Correspondence Analysis: An Introduction*. Thousand Oaks, CA: Sage.

Age	Standard Deviation in Dimension		Correlation
	1	2	1-2
10-12	.046	.132	.540
13-15	.079	.061	-.662
16-18	.094	.036	.473
19-21	.065	.072	.016

Table 7. Confidence Row Points.

Risk	Standard Deviation in Dimension		Correlation
	1	2	1-2
Substance Abuse	.105	.045	.225
Drop Out	.078	.122	-.864
Violence	.054	.067	-.008
Mental Health	.063	.065	.741

Table 8. Confidence Column Points.



Plot 1. A biplot displaying various risk factors among adolescents and how they relate to specific age groups on two dimensions.

- De Leeuw, J. (2005). *Review of Correspondence Analysis and Data Coding with Java and R*. Journal of Statistical Software, 14, 230-232.
- Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., & Vingron, M. (2001). *Correspondence analysis applied to microarray data*. Proceedings of the National Academy of Sciences, 98, 10781-10786.
- Garson, D. (2008). *Correspondence Analysis*, from Statnotes: Topics in Multivariate Analysis. Retrieved 04/01/2010 from <http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm>.
- Greenacre, M. J. (1984). Theory and Applications of Correspondence Analysis. Academic Press: New York.
- Greenacre, M. J. (2007). Correspondence analysis in practice. Boca Raton, Florida: Taylor and Francis Group.
- Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tatham, R. L. (2007). Multivariate Data Analysis. Toronto: Prentice Hall.
- Hoffman, D. L., Franke, G. R. (1986). *Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research*. Journal of Marketing Research, 23, 213-227.
- Moser, B. E. (1989). *Exploring contingency tables with correspondence analysis*. Cabios, 5, 183-189.
- Palmer, M. W. (1993). *Putting things together in even better order: The advantages of canonical correspondence analysis*. Ecology, 74, 2215-2230.
- SAS Institute Inc. (2010). Knowledge Base/Focus Areas: Statistics. Retrieved 19/08/2010 from <http://support.sas.com/rnd/app/da/market/stat.html>.
- Statistics Solutions, Inc. (2009). *Correspondence Analysis*, from Statistics Solutions. Retrieved 04/01/2010 from <http://www.statisticssolutions.com/methods-chapter/statistical-tests/correspondence-analysis/>.
- StatSoft, Inc. (2010). *Correspondence Analysis*, from Electronic Statistics Textbook. Retrieved 04/01/2010 from <http://www.statsoft.com/textbook/>.
- Storti, D. (2010). *Correspondence Analysis*, from Unesco. Retrieved 04/01/2010 from http://www.unesco.org/webworld/idams/advguide/Chapt6_5.htm.
- Ter Braak, C. J. F. (1985). *Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model*. Biometrics, 41, 859-873.

Manuscript received 12 November 2010.

Manuscript accepted 1 February 2011.