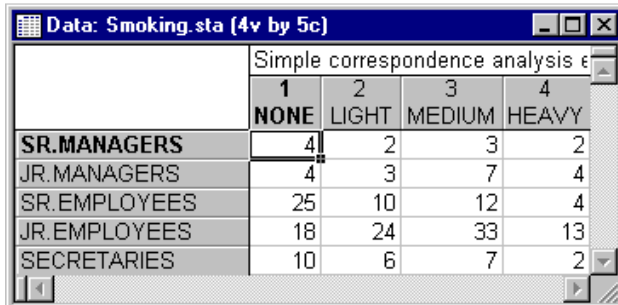


Example 1: Correspondence Analysis and Supplementary Points

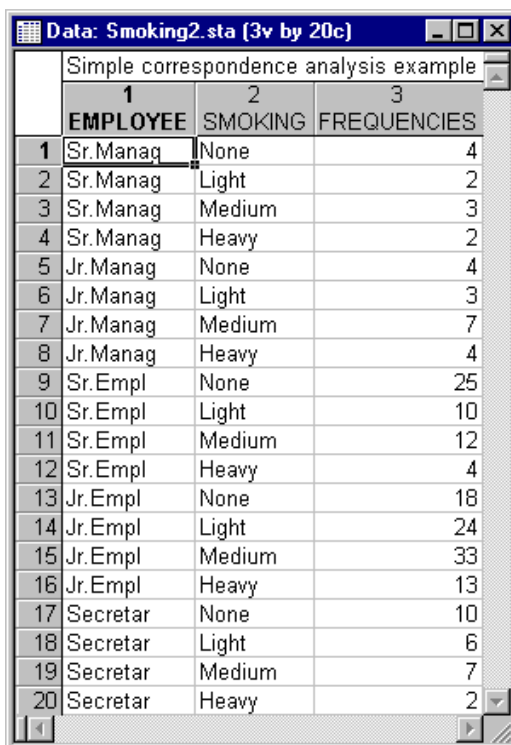
This example is based on a fictitious data set presented in Greenacre (1984, p. 55) to illustrate how to interpret the results of a correspondence analysis. This data set is also discussed in the [Introductory Overview](#). In this example, the different formats of data files accepted by the [Correspondence Analysis](#) module will be illustrated, and the typical results of correspondence analysis will be explained (see also [Computational Details](#)). Also, the use of [supplementary points](#) for aiding in the interpretation of results will be demonstrated.

Open the *Smoking.sta* data file via the [File - Open](#) menu; it is in the /Examples/Datasets directory of *STATISTICA*. This file contains the frequency table, as presented in Greenacre (1984, p. 55).



	1	2	3	4
	NONE	LIGHT	MEDIUM	HEAVY
SR.MANAGERS	4	2	3	2
JR.MANAGERS	4	3	7	4
SR.EMPLOYEES	25	10	12	4
JR.EMPLOYEES	18	24	33	13
SECRETARIES	10	6	7	2

Formats of data files. The [Correspondence Analysis](#) module provides great flexibility with regard to the permissible formats of input data. For example, in addition to the raw frequency table as contained in the file *Smoking*, you could also specify the two-way table by including in the datafile two grouping variables (one for the *Staff* group, another for the *Smoking* category). This format for the table is illustrated in the example data file *Smoking2.sta*.



	1	2	3
	EMPLOYEE	SMOKING	FREQUENCIES
1	Sr.Manag	None	4
2	Sr.Manag	Light	2
3	Sr.Manag	Medium	3
4	Sr.Manag	Heavy	2
5	Jr.Manag	None	4
6	Jr.Manag	Light	3
7	Jr.Manag	Medium	7
8	Jr.Manag	Heavy	4
9	Sr.Empl	None	25
10	Sr.Empl	Light	10
11	Sr.Empl	Medium	12
12	Sr.Empl	Heavy	4
13	Jr.Empl	None	18
14	Jr.Empl	Light	24
15	Jr.Empl	Medium	33
16	Jr.Empl	Heavy	13
17	Secretar	None	10
18	Secretar	Light	6
19	Secretar	Medium	7
20	Secretar	Heavy	2

Finally, you can analyze raw data that are not pretabulated. The data in the example file *Smoking3.sta* are organized in this manner, that is, it only contains two variables (*StaffGrp* and *Smoking*) with codes to indicate to which group each case belongs; there are a total of 193 cases in that file.

Specifying the analysis. For this example, the example data file *Smoking.sta* will be used. Select [Correspondence Analysis](#) from the [Statistics - Multivariate Exploratory Techniques](#) menu to display the [Correspondence Analysis \(CA\): Table Specifications](#) Startup Panel. In this example, the data file contains frequencies without grouping variables; therefore, select the *Frequencies w/out grouping vars* option button under *Input* on the [Correspondence Analysis \(CA\) tab](#) (if you want to use the file *Smoking2.sta*, you would select the *Frequencies with grouping variables* option button; to use the file *Smoking3.sta*, select the *Raw data (requires tabulation)* option button).

Next select the variables. Click the *Variables with frequencies* button to display the standard [variable selection](#) dialog. Here, select all variables and then click the *OK* button. Note that when you use this data file format (i.e., the input is a tabulated frequency table), *STATISTICA* will interpret the selected variables as the columns of the table to be analyzed, and the cases as the rows of the table. Since the data in file *Smoking.sta* are arranged in that manner, click the *OK* button on the Startup Panel to perform the correspondence analysis. After a few moments the [Correspondence Analysis Results](#) dialog is displayed.

Reviewing the results.

Eigenvalues. If you are not familiar with the correspondence analysis technique, and the most important statistics that are

customarily computed, you may want to review the [Introductory Overview](#) at this point. To reiterate, if you considered the relative row frequencies as coordinates in a space consisting of as many dimensions as there are columns, and the relative column frequencies as coordinates in a space consisting of as many dimensions as there are rows, then the main goal of the analysis is to reconstruct the distances between the row points, and to reconstruct the distances between the column points, in a space defined by as few dimensions as possible. First, click the *Eigenvalues* button on the [Advanced tab](#) to produce the spreadsheet that contains information about the number of dimensions that are necessary to reconstruct the information in the table.

Data: Eigenvalues and Inertia for all Dimensions (Smoking.sta)					
Eigenvalues and Inertia for all Dimensions (Smoking.sta)					
Input Table (Rows x Columns): 5 x 4					
Total Inertia=.08519 Chi²=16.442 df=12 p=.17190					
Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulative Percent	Chi Squares
1	0.273421	0.074759	87.75587	87.7559	14.42851
2	0.100086	0.010017	11.75865	99.5145	1.93332
3	0.020337	0.000414	0.48547	100.0000	0.07982

The first column shows the *Number of dimensions*; a maximum of three dimensions can be extracted, in which case the (relative) frequency table can be reconstructed exactly. The *Singular Values* are computed by the so-called generalized singular value decomposition of the table of relative frequencies (see [Computational Details](#)). The *Eigenvalues* are the squared *Singular Values*, and they will sum to the *Total Inertia*, which is listed in the header of the spreadsheet as .08519. The total inertia is defined as the *Chi-square* value (16.442) divided by the total number of cases (193). Thus, as discussed in the [Introductory Overview](#), the correspondence analysis can also be considered to be a decomposition of the total *Chi-square* value, in much the same way that principal components analysis (see [Factor Analysis](#)) decomposes the total variance/covariance matrix of continuous variables.

As you can see, the dimensions are computed so that the first dimension extracts the most information (i.e., has the highest eigenvalue), the next dimension extracts the second most information, and so on (see also [Computational details](#)). The first dimension in this case extracts 87.76% of the total inertia. The inclusion of the second dimension increases the "explained" inertia to 99.51%.

Note that on the [Quick tab](#) and [Options tab](#) are options under *Number of dimensions* for selecting the number of dimensions to retain in the analysis. You can either directly request a certain *Number of dimensions*, or allow *STATISTICA* to determine the number of dimensions based on the respective user-defined value for the *Cumulative contribution to inertia*. As described in the [Introductory Overview](#), correspondence analysis is mostly a descriptive method, rather than a method for hypothesis testing. Therefore, there are no fixed guidelines as to how to decide on the number of dimensions to interpret. In this case, it is clear that the first two dimensions will explain practically the total inertia for the table. Thus, accept the default 2 dimensions, and click on the *Row and column coordinates* button on the [Advanced tab](#).

Reviewing the quality and inertias of row and column points. Two spreadsheets will be displayed; one for the row coordinates and one for the column coordinates.

Data: Row Coordinates and Contributions to Inertia (Smoking.sta)*										
Row Coordinates and Contributions to Inertia (Smoking.sta)										
Input Table (Rows x Columns): 5 x 4										
Standardization: Row and column profiles										
Row Name	Row Number	Coord. Dim.1	Coord. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine² Dim.1	Inertia Dim.2	Cosine² Dim.2
SR.MANAGERS	1	-0.065768	0.193737	0.056995	0.892568	0.031376	0.003298	0.092232	0.213558	0.800336
JR.MANAGERS	2	0.258958	0.243305	0.093264	0.991082	0.139467	0.083659	0.526400	0.551151	0.464682
SR.EMPLOYEES	3	-0.380595	0.010660	0.264249	0.999817	0.449750	0.512006	0.999033	0.002998	0.000784
JR.EMPLOYEES	4	0.232952	-0.057744	0.455959	0.999810	0.308354	0.330974	0.941934	0.151772	0.057876
SECRETARIES	5	-0.201089	-0.078911	0.129534	0.998603	0.071053	0.070064	0.865346	0.080522	0.133257

Data: Column Coordinates and Contributions to Inertia (Smoking.sta)*										
Column Coordinates and Contributions to Inertia (Smoking.sta)										
Input Table (Rows x Columns): 5 x 4										
Standardization: Row and column profiles										
Column Name	Column Number	Coord. Dim.1	Coord. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine² Dim.1	Inertia Dim.2	Cosine² Dim.2
NONE	1	-0.393308	0.030492	0.316062	0.999995	0.577372	0.653996	0.994020	0.029336	0.005975
LIGHT	2	0.099456	-0.141064	0.233161	0.984016	0.082860	0.030850	0.326726	0.463174	0.657290
MEDIUM	3	0.196321	-0.007359	0.321244	0.983228	0.148025	0.165617	0.981848	0.001737	0.001380
HEAVY	4	0.293776	0.197766	0.129534	0.994552	0.191743	0.149538	0.684398	0.505754	0.310154

The statistics reported in these spreadsheets are discussed in the [Introductory Overview](#). First look at the *Quality* of the points. The *Quality* of a point is defined as the ratio of the squared distance of the point from the origin in the chosen number of dimensions, over the squared distance from the origin in the space defined by the maximum number of dimensions (remember that the metric here is *Chi-square*, as described in the [Introductory Overview](#)). By analogy to [Factor Analysis](#), the quality of a point is similar in its interpretation to the communality for a variable in factor analysis. As you can see, both the row and column points are represented quite well in the two-dimensional solution; the quality for all points is .89 or higher.

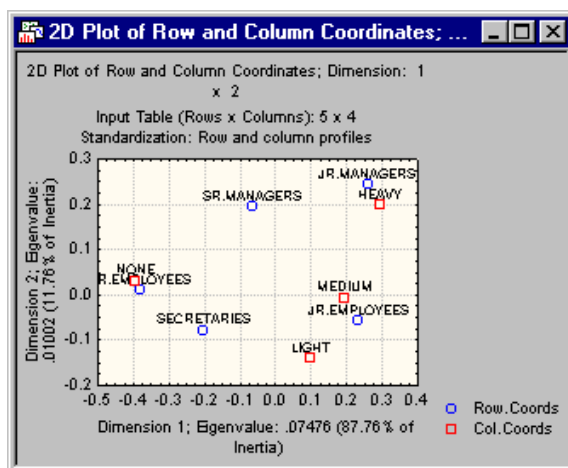
The *Relative inertia* values pertain to the proportion of the total inertia "accounted for" by the respective point. Note that a point may be well represented in a particular solution, but not contribute much to the total inertia. From the spreadsheets shown above one can see that the row that contributes most to the overall inertia is that representing the *Senior Employees*, and the column that contributes most is that representing the *None* smokers.

The quality for each point, due to each dimension can be found in the columns labeled *Cosine²*. The *Cosine²* values summed across the two dimensions is equal to the total *Quality* value. The relative contribution of each point to the *inertia* for each dimension (remember that the *Eigenvalues* represent the inertias associated with each dimension) is also shown in the spreadsheets above.

Standardization of row and column coordinates. There are several options available on the [Options tab](#) for standardizing the row and column coordinates. Note that the interpretation of the row and column coordinates depends on the method of standardization that is chosen (see also the [Introductory Overview](#)); however, the quality of representation and relative inertia values shown in the spreadsheets above are not affected by the chosen method of standardization.

The coordinates can be computed based either on the matrix of relative row frequencies (*Row profiles* standardization; the analysis is based on the so-called row profile matrix, where the sum of all relative frequencies within each row, across the columns, sums to 1.0), or the relative column frequencies (*Column profiles* standardization; the analysis is based on the so-called column profile matrix, where the sum of all relative frequencies within each column, across the rows, sums to 1.0). In most cases, the *Row & column profiles* standardization is most appropriate (the default). In that case the Euclidean distances between the row points, and the distances between the column points can be interpreted in a meaningful manner (i.e., the distances between the points are *Chi-square* distances; see the [Introductory Overview](#)). However, note that the distances between the row and column points have no meaningful interpretation, regardless of standardization.

Reviewing the row and column coordinates. The best way to quickly review the row and column coordinates is to plot them. On the [Advanced tab](#), click the *Row & col. - 2D* button under *Plots of coordinates*. A [2D scatterplot](#) will be displayed, simultaneously showing the row and column points in the two dimensions (see also Greenacre, 1984, p. 66).



To reiterate, direct comparisons between row and column points are not meaningful. However, you can make meaningful interpretations of the general locations of row and column points, and their relations within each type of point. For example, if you reviewed the 2D graph of the row and column points, you can see that the first (horizontal) dimension, which "accounts for" most of the *inertia* (and is, therefore, the most "important" dimension, explaining most of the differences between the patterns of relative frequencies in the rows of the table, and in the columns of the table), is characterized by *None* smokers on the left, and *Light*, *Medium*, and *Heavy* smokers to the right; the row points that are farthest to the left on this axis are the *Senior Employees* and *Secretaries*. This would suggest that much of the total inertia is due to the difference between non-smokers and smokers, and that there are relatively more non-smokers among *Senior Employees* and *Secretaries*.

Reviewing tables of relative frequencies. You can easily verify this interpretation by reviewing the tables of relative frequencies. On the [Review tab](#), click the *Row percentages* button and then the *Column percentages* button.

Data: Percentages of Row Totals (Smoking)*					
Percentages of Row Totals (Smoking.sta)					
Input Table (Rows x Columns): 5 x 4					
Total Inertia=.08519 Chi²=16.442 df=12 p=.17190					
	NONE	LIGHT	MEDIUM	HEAVY	Total
SR.MANAGERS	36.36364	18.18182	27.27273	18.18182	100.0000
JR.MANAGERS	22.22222	16.66667	38.88889	22.22222	100.0000
SR.EMPLOYEES	49.01961	19.60784	23.52941	7.84314	100.0000
JR.EMPLOYEES	20.45455	27.27273	37.50000	14.77273	100.0000
SECRETARIES	40.00000	24.00000	28.00000	8.00000	100.0000

Data: Percentages of Column Totals (Smoking)*				
Percentages of Column Totals (Smoking.s)				
Input Table (Rows x Columns): 5 x 4				
Total Inertia=.08519 Chi²=16.442 df=12 p=				
	NONE	LIGHT	MEDIUM	HEAVY
SR.MANAGERS	6.5574	4.4444	4.8387	8.0000
JR.MANAGERS	6.5574	6.6667	11.2903	16.0000
SR.EMPLOYEES	40.9836	22.2222	19.3548	16.0000
JR.EMPLOYEES	29.5082	53.3333	53.2258	52.0000
SECRETARIES	16.3934	13.3333	11.2903	8.0000
Total	100.0000	100.0000	100.0000	100.0000

The relative row and column frequencies shown in these spreadsheets support the interpretation of the first dimension: There are a relatively large percentages of *None* smokers among *Senior Employees* and *Secretaries*. This makes the respective row profiles in the table of relative row frequencies, and the respective column profile (*None*) in the table of relative column frequencies different from all the others.

Supplementary points. An important aspect of correspondence analysis is to represent row and/or points that were not part of the original analysis in the same coordinate system as the regular points (see also the [Introductory Overview](#)). Greenacre (1984, Table 3.5) provides an example of this procedure, in the context of this data set. Specifically, suppose you had available information about the national averages concerning the different categories of smoking, and information about the number of employees in each staff group that did or did not consume alcohol.

	Smoking Category			
	<i>None</i>	<i>Light</i>	<i>Medium</i>	<i>Heavy</i>
National Average	42%	29%	20%	9%

Staff Group	Alcohol	
	<i>Yes</i>	<i>No</i>
Senior Managers	0	11
Junior Managers	1	17
Senior Employees	5	46
Junior Employees	10	78
Secretaries	7	18

Specifying a supplementary row. From the [Supplementary points tab](#), first click the *Add row points* button. The [Supplementary Row Points](#) dialog is displayed where you can specify the supplementary row points. Remember that in row profile standardization, the analysis is performed on the relative row frequencies, which will sum to 1.0; thus, it does not matter whether you entered here 42 or .42, i.e., percentages or proportions, the results would be the same either way.

To enter a supplementary row, first type a name or label for the row into the first column of the spreadsheet (e.g., type *Average*). Next type in the values 42, 29, 20, and 9 under the respective column headers *None*, *Light*, *Medium*, and *Heavy*.

Point	Name of supp.pnt	NONE	LIGHT	MEDIUM	HEAVY
1	Average	42	29	20	9
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					

OK Cancel

To accept these values, exit the dialog by clicking the *OK* button; if you exit the dialog by closing it or clicking the *Cancel* button, your entries will be discarded.

Specifying supplementary columns. Next click the *Add column point* button, and enter the supplementary column frequencies shown above. Enter the frequencies row-by-row, and then click the *OK* button again.

Supplementary Column Points (Smoking.sta)

Supplementary Column Points (Smoking.sta)
Enter the values (frequencies) for the new (supplementary) points; then click OK.

Point	Name of supp.pnt	SR.MANAGERS	JR.MANAGERS	SR.EMPLOYEES	JR.EMPLOYEES	SECRETARIES
1	Alcohol Yes	0	1	5	10	7
2	Alcohol No	11	17	46	78	18
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						

OK Cancel

Reviewing statistics for supplementary points. After specifying the supplementary rows, whenever you select any of the plots of the coordinates, or when you select the spreadsheets of row and column coordinates, the resulting displays will incorporate the results for the supplementary rows and columns. For example, shown below are the coordinate values and related statistics that are displayed, along with the statistics for the standard row and column points reviewed earlier, after you click the *Row and column coordinates* button on the [Advanced tab](#).

Data: Row Coordinates and Contributions to Inertia (Smoking)*

Row Coordinates and Contributions to Inertia (Smoking.sta)
Input Table (Rows x Columns): 5 x 4
Standardization: Row and column profiles

Row Name	Row Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine ² Dim.1	Inertia Dim.2	Cosine ² Dim.2
SR.MANAGERS	1	-0.065768	0.193737	0.056995	0.892568	0.031376	0.003298	0.092232	0.213558	0.800336
JR.MANAGERS	2	0.258958	0.243305	0.093264	0.991082	0.139467	0.083659	0.526400	0.551151	0.464682
SR.EMPLOYEES	3	-0.380595	0.010660	0.264249	0.999817	0.449750	0.512006	0.999033	0.002998	0.000784
JR.EMPLOYEES	4	0.232952	-0.057744	0.455959	0.999810	0.308354	0.330974	0.941934	0.151772	0.057876
SECRETARIES	5	-0.201089	-0.078911	0.129534	0.998603	0.071053	0.070064	0.865346	0.080522	0.133257
Average		-0.258368	-0.117648		0.761324			0.630578		0.130746

Data: Column Coordinates and Contributions to Inertia (Smoking)*

Column Coordinates and Contributions to Inertia (Smoking.sta)
Input Table (Rows x Columns): 5 x 4
Standardization: Row and column profiles

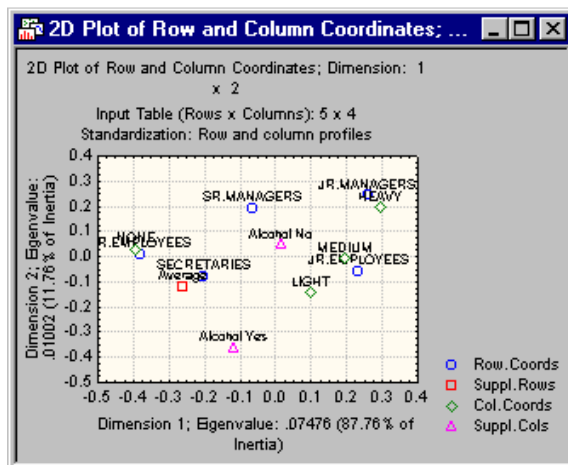
Column Name	Column Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine ² Dim.1	Inertia Dim.2	Cosine ² Dim.2
NONE	1	-0.393308	0.030492	0.316062	0.999995	0.577372	0.653996	0.994020	0.029336	0.005975
LIGHT	2	0.099456	-0.141064	0.233161	0.984016	0.082860	0.030850	0.326726	0.463174	0.657290
MEDIUM	3	0.196321	-0.007359	0.321244	0.983228	0.148025	0.165617	0.981848	0.001737	0.001380
HEAVY	4	0.293776	0.197766	0.129534	0.994552	0.191743	0.149538	0.684398	0.505754	0.310154
Alcohol Yes		-0.114829	-0.361956		0.438575			0.040104		0.398471
Alcohol No		0.015536	0.048971		0.438575			0.040104		0.398471

The interpretation of these statistics is the same as that for the points that were used to perform the analysis (see also the [Introductory Overview](#)). It appears that the two-dimensional solution represents the new row point *Average* (i.e., national average) very well (the *Quality* is .7613). The new column points are not quite as well represented, however, still, over 40% of the total squared (weighted) distance of these points from the origin in the space defined by the maximum number of dimensions is "accounted for" by the two-factor solution (the *Quality* is equal to .4386 for both supplementary column points).

At this point, you may want to try to enter as supplementary row and column points the respective column and row totals for the entire table. You will see that those points will be represented by coordinates that are equal to 0 for all dimensions. This illustrates

that the space defined by the two dimensions is weighted by the respective column and row totals, which define the origin of the coordinate system. Thus, you could interpret the distances of the points from the origin as (*Chi-square*) distances from the respective column and row totals.

Plots with supplementary points. Now produce the combined 2D scatterplot again, for both the row and column points. Click the *Row & col. - 2D* button under *Plots of coordinates* on the [Advanced tab](#).



The supplementary row point for the national *Average* will be plotted on the left side of the origin for the horizontal axis (the coordinate value is $-.2584$; see the first table shown above). Thus, one may infer that there are relatively more *None* smokers on average in the nation than there are in the current sample.

The supplementary column points *Alcohol Yes* and *Alcohol No* approximately line up along the second axis, which also appears to distinguish between different degrees of smoking, i.e., *Light*, *Medium*, and *Heavy* (as mentioned above, the first axis appears to distinguish between *None* smokers and smokers). Thus, there is some indication that *Heavy* smokers are also more likely to consume alcohol (specifically, the pattern of frequencies across the staff groups for *Alcohol* is more similar to the pattern of frequencies for the *Heavy* and *Medium* smokers). However, remember that correspondence analysis is primarily a descriptive and/or [exploratory technique](#) to represent categorical data in graphical displays, and no claims of statistical significance are implied (see the [Introductory Overview](#); see also [Elementary Concepts in Statistics](#)).

See also, [Correspondence Analysis - Index](#).