Configural Frequency Analysis -
A Program for 32 Bit Windows Operating Systems

Manual for Program Version 2000

Alexander von Eye¹
Michigan State University

List of Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. New to Version 2000</td>
<td>2</td>
</tr>
<tr>
<td>1. Configural Frequency Analysis</td>
<td>2</td>
</tr>
<tr>
<td>2. The CFA Program</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Classical, First order CFA; Keyboard Input</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Two-Sample CFA with Two Predictors; Keyboard Input</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Second Order CFA; Input via File</td>
<td>18</td>
</tr>
<tr>
<td>2.4 CFA with Covariates; Input via File (Data) and Keyboard (Covariate)</td>
<td>21</td>
</tr>
</tbody>
</table>

¹Address correspondence concerning this manual or the CFA program to Alexander von Eye, Michigan State University, Department of Psychology, 119 Snyder Hall, East Lansing, Michigan 48824-1117; e-mail: VONEYE@MSU.EDU. The author is indebted to Keith E. Niedermeier for helpful comments on earlier versions of this manual, to Matthias Reitzle for identifying a bug in Version 1997, and to Deb Sadowski, Eun-Young Mun, and Adalberto Gonzáles Debén for pointing to problems with an earlier release of program Version 2000.
0. New to Version 2000

*Version 2000* of this Configural Frequency Analysis (CFA) program contains a number of new options and improvements over the previous version, 1997. These options and improvements include:

1. five new definitions of types in two-sample CFA;
2. four new significance tests for two-sample CFA;
3. two new descriptive measures for global CFA;
4. variables can now have up to 999 categories;
5. simplified data input via file; and
6. optimized print output.

In addition, several checks have been included into the program that prevent users from selecting CFA models of an order that is inappropriate for the number of variables under study, and from pursuing analyses with too many variables or covariates. The program now keeps track of the number of covariates that can be included. The following manual explains the new features in the context of the program description.

1. Configural Frequency Analysis

Configural Frequency Analysis (CFA; Lienert, 1969; von Eye, 1990; von Eye, Spiel, & Wood, 1996) is a multivariate method for typological research that involves categorical variables. CFA allows researchers to identify those cells in a cross-classification that contain more (or fewer) cases than expected based on some chance model. When a cell contains more cases than expected it is said to constitute a CFA *type*. When there are fewer cases than expected, a cell is said to constitute a CFA *antitype* (see below). CFA can be applied in both exploratory and confirmatory research. The program described here focuses on exploratory applications.

CFA proceeds in four steps. The first step involves specifying a *CFA Base Model*. The Base Model, which can often be expressed in terms of a log-linear model, involves all variable relationships that are *not* of interest for the hypotheses under study. Specifically, von Eye and
Schuster (1998) proposed that the selection of a CFA base model use the following three criteria. (1) There must be only one way to deviate from the base model. This criterion is needed for clear-cut interpretation of emerging types and antitypes. (2) The sampling scheme under which data were collected must be considered. (3) Base models must be parsimonious.

Examples of base models include global models where all variables have the same status, and regional models where variables are grouped (von Eye, 1990). The classical CFA model of total independence of variables is a global model. It considers the main effects of all variables. Cells that contain more (or fewer) cases than expected from this model indicate local interactions among the variables. These interactions can be of zero or higher order. For \( k \geq 2 \) variables, the classical CFA model can be expressed in terms of the log-frequency model

\[
\log M = X\lambda ,
\]

where \( M \) is the array of expected frequencies in the cross-tabulation, \( X \) is the indicator matrix that contains all vectors needed for the intercept and all main effects, and \( \lambda \) is the parameter vector. \( \hat{m}_i \) is the estimated expected frequency for cell \( i \), where \( i \) goes over all cells. Let \( m_i \) be the \( i \)th element of \( M \). Then, a general null hypothesis for CFA is

\[
H_0: E[n_i] = m_i ,
\]

Exploratory CFA asks, under this null hypothesis, for each cell, whether \( E[n_i] > m_i \). If this is the case, Cell \( i \) is said to constitute a CFA type. If, in contrast, \( E[n_i] < m_i \), Cell \( i \) is said to constitute a CFA antitype. If, statistically, \( E[n_i] = m_i \), Cell \( i \) constitutes neither a type nor an antitype.

An example of a regional CFA model is Prediction CFA (P-CFA), where one group of variables is formed by the predictors and another group is formed by the criteria. The log-frequency model for P-CFA is, in short hand notation \([P][C]\), where P indicates that the model is saturated in the predictors and C indicates that the model is saturated in the criteria.

The second step of CFA involves estimation of expected cell frequencies. For the log-linear CFA base models, this is typically done using maximum likelihood methods (ML). Weighted least squares methods can also be appropriate. However, virtually all CFA applications use ML. The Version 2000 CFA program also uses ML.

The third step of CFA involves performing statistical significance tests. This step first requires protection of the test-wise \( \alpha \). Most typically, researchers protect \( \alpha \) by performing the Bonferroni-adjustment which leads, for t tests, to the adjusted \( \alpha^* = \alpha / t \). Alternatively, the method
proposed by Holm (1979) is often used. Holm’s method considers both the total number of tests, $t$, and the number of tests performed before the $i$th test, for $i = 1, \ldots, t$. Beginning with the second test, Holm’s procedure suggests less conservative decisions concerning the existence of types or antitypes. In two-dimensional or three-dimensional tables, the even less conservative methods proposed by Perli, Hommel, and Lehmacher are recommended (for details see von Eye, 1990; Olejnik, Li, Supattathum, & Huberty, 1998). This CFA program performs a Bonferroni adjustment.

After protecting the test-wise $\alpha$, significance tests are performed. A large number of tests has been proposed for CFA. The following eight tests for global CFA models are part of the CFA program$^2$: (1) binomial test; (2) approximation of the binomial test using Stirling’s formula for approximative calculation of factorials; (3) Pearson $X^2$-component test; (4) z-test; (5) normal approximation of binomial test; (6) Lehmacher’s asymptotic hypergeometric test; (7) Lehmacher’s asymptotic hypergeometric test with Küchenhoff’s continuity correction; and (8) Anscombe’s test. The binomial tests and Lehmacher’s tests require fixed marginal totals. Lehmacher’s tests can be applied only for the classical CFA model of total variable independence with no covariates.

The first seven of these tests are described in von Eye (1990). Anscombe’s test can be described as follows. Let $m_i$ be the observed cell frequency for cell $i$, and $\hat{m}_i$ the estimated expected cell frequency. Then the test statistic for Anscombe’s test is

$$r^* = \frac{3\left[ \frac{2}{n_i^2} - \left( m_i - \frac{1}{6} \right)^2 \right] \sum_i \frac{1}{2 m_i^6}}{\sqrt{\sum_i m_i}}$$

which is supposedly more nearly normally distributed than the square root of the better-known Pearson $X^2$ component, $\sqrt{\chi^2} = \frac{m_i - m_i}{\sqrt{m_i}}$.

The Lehmacher test is the most powerful of these eight. However, it requires very large samples in order to avoid non-conservative decisions. The next most powerful test is Lehmacher’s

$^2$For two-sample CFA two variants of the $X^2$-test, the z-test and an approximation of the binomial test are available. Version 2000 contains four additional tests. See Section 2.2.
test with continuity correction, followed by the z-test and the $X^2$-test. The binomial tests are about as powerful as the $X^2$-tests (for a comparison of these tests see von Eye & Rovine, 1988). The characteristics of Anscombe’s test still need to be explored. One constraint that needs to be considered when using the Anscombe test is that estimated expected cell frequencies must be greater than $\hat{m}_i = 0.16667$ for the test to be numerically viable.

The fourth step of CFA involves interpretation of types and antitypes. This interpretation focuses on the characteristics of the individuals in the cells that constitute types and antitypes. Thus, CFA is a pattern-oriented rather than a variable-oriented statistical method (Bergman & El-Khoury, 1995; Cairns, Bergman, & Kagan, 1998).

The following sections describe the use of the CFA program by providing sample runs and print outs.

2. The CFA Program

The CFA program for 32 bit operating systems such as Windows 95, 98, Windows NT, and Windows 2000 was written in Fortran 90 and compiled using the MS Fortran PowerStation 4.0 (1994). The executable program file occupies 337 KB. The program is interactive and keyboard-oriented. Memory allocation is dynamic, and will use as much of a computer’s RAM as needed until either program limits are reached or the entire RAM is used and the computer stalls. Program limits are ten variables. If all ten variables are dichotomous, the cross-tabulation of ten variables has $2^{10} = 1024$ cells. If the ten variables have five categories each, the resulting table has $5^{10} = 9,765,625$ cells. If each of the ten variable has ten categories, the number of cells is 10 billion. This is certainly enough for most purposes. The maximum number of categories per variable is 999. The program expects cell frequencies as input. It cannot create cross-tabulations from raw data.

Installation of a program shortcut. The executable file of the CFA program can be obtained gratis from the author. Suppose the program file is located on the C-drive. In Windows 95, 98, Windows NT, or 2000, a shortcut can be created in various ways. One of the easiest is as follows:

(7) Open Windows’ Explorer;

(8) Open within Explorer the directory where the CFA program is located;

(9) Grab the icon by the CFA program name using the pointing device and drag it to the open area of the Windows screen, outside of the Explorer window.
The result of these operations is that the Windows screen, that is, the screen that is shown after starting the computer displays the shortcut icon to the CFA program. Double clicking this shortcut icon starts the program. Alternatively, the program can be started by double clicking the program name shown by Windows’ Explorer.

The following paragraphs present sample runs to illustrate the most important features of the program.

2.1 Classical First Order CFA; Keyboard Input

The following first table presents a sample run for classical, first order CFA. The data describe a sample of \( N = 65 \) students who were treated with LSD 50 (Lienert, 1964). The following three variables were observed: \( C \) = narrowed consciousness, \( T \) = thought disturbance, and \( A \) = affective disturbance. Each symptom was scaled as either 1 = present or 2 = absent. The cross-classification of the three symptoms has \( 2 \times 2 \times 2 = 8 \) cells. In the present sample run we enter the data via the keyboard.

For users to replicate this sample run we assume that they have the executable file, CFA, of the program on their computer, that the computer runs under Windows 95 or higher, under Windows NT 4.0 or higher, or under Windows 2000, and that there is a program shortcut to the executable file on the screen. If there is no shortcut, the program can be started by double-clicking the program file name within Windows’ Explorer. The following steps must be performed for First Order CFA:

<table>
<thead>
<tr>
<th>Command</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>double click shortcut to CFA program</td>
<td>starts CFA program; program window appears on screen; the program responds with a header telling the user that the CFA program was started and asks whether data will be input via file (=1) or interactively, via the keyboard (=2). We select interactive data input and type 2 (Enter) The program responds by asking for the number of variables. We type</td>
</tr>
<tr>
<td>3 (Enter)</td>
<td>The program asks for number of categories of the first variable. We type</td>
</tr>
<tr>
<td>2 (Enter)</td>
<td>This is repeated until the number of categories for each variable is given. The program follows up by prompting the cell frequencies. We respond to the prompt for the first cell, that is, Cell 111, by typing</td>
</tr>
<tr>
<td>20 (Enter)</td>
<td>To the prompt for the second cell frequency we respond by typing</td>
</tr>
<tr>
<td>1 (Enter)</td>
<td>For the following cells we type</td>
</tr>
<tr>
<td>4, 12, 3, 10, 15, and 0</td>
<td>each number followed by Enter (no commas). When all cell frequencies are keyed in, the program responds by presenting the sample size - in the present example ( N = 65 ) -, and by asking whether the user wishes to save the data (yes = 1; no = 2). We select to save the data and type</td>
</tr>
<tr>
<td>1 (Enter)</td>
<td>The program then asks for the name of the data file. Up to 80 spaces are read. The name must be given in DOS style, that is, including the path. If no path is given, the file will be saved in the currently active directory. In the present example this is the directory that contains the CFA program file. We type</td>
</tr>
<tr>
<td>leuner.dat (Enter)</td>
<td>The program responds by presenting the current program options concerning CFA models. The current version allows one to perform any of the global CFA models, that is, any of the models where the status of all variables is the same. In addition, the program can perform a 2-sample CFA. One indicates the CFA model by typing the order number of the CFA model. For example, for zero order CFA one types 0, for first order CFA one types 1, and so forth. For the present example we select classical first order CFA and type</td>
</tr>
</tbody>
</table>
The program responds by presenting the unidimensional marginal frequencies on the screen and by asking whether the user wishes to include a covariate (yes = 1; no = 2). In the present example we opt not to include covariates and type

The program then presents the eight statistical tests currently included in the program. We select the Lehmann test with continuity correction and indicate our choice by typing

The program then requests input of a significance level. We go with the standard $\alpha = 0.05$ and type

The program now requests the name of the output file. We type

A total of 80 spaces can be used for the file name. The program responds by performing calculations and writing results to the file leuner.out. Finally, the program asks whether the uses wishes that the design matrix, $X$ be printed. In this example we would like to see the design matrix and type

After concluding the analysis the program window disappears.

The above sample run resulted in the following output file (to save space the following protocol was slightly edited):

```
Configural Frequency Analysis
---------- --------- --------
author of program: Alexander von Eye, 2000
Marginal Frequencies
Variable Frequencies
--------- ---------
1 37. 28.
2 34. 31.
3 42. 23.
```
Lehmachers test with continuity correction was used.
Bonferroni-adjusted alpha = .0062500.
a CFA of order 1 was performed.

Table of results

<table>
<thead>
<tr>
<th>Configuration</th>
<th>fo</th>
<th>fe</th>
<th>statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>20.</td>
<td>12.506</td>
<td>3.183</td>
<td>.00072795</td>
</tr>
<tr>
<td>112</td>
<td>1.</td>
<td>6.848</td>
<td>-2.800</td>
<td>.00255136</td>
</tr>
<tr>
<td>121</td>
<td>4.</td>
<td>11.402</td>
<td>-3.198</td>
<td>.00069093</td>
</tr>
<tr>
<td>122</td>
<td>12.</td>
<td>6.244</td>
<td>2.819</td>
<td>.00240908</td>
</tr>
<tr>
<td>211</td>
<td>3.</td>
<td>9.464</td>
<td>-2.874</td>
<td>.00207264</td>
</tr>
<tr>
<td>212</td>
<td>10.</td>
<td>5.182</td>
<td>2.442</td>
<td>.00730272</td>
</tr>
<tr>
<td>221</td>
<td>15.</td>
<td>8.629</td>
<td>2.887</td>
<td>.00194350</td>
</tr>
<tr>
<td>222</td>
<td>0.</td>
<td>4.725</td>
<td>-2.458</td>
<td>.00698741</td>
</tr>
</tbody>
</table>

Chi2 for CFA model = 37.9198
df = 4   p = .0000012

Descriptive indicators of types and antitypes

<table>
<thead>
<tr>
<th>cell</th>
<th>Rel. Risk</th>
<th>Rank</th>
<th>logP</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>1.599</td>
<td>4</td>
<td>1.549</td>
<td>2</td>
</tr>
<tr>
<td>112</td>
<td>.146</td>
<td>7</td>
<td>.730</td>
<td>7</td>
</tr>
<tr>
<td>121</td>
<td>.351</td>
<td>5</td>
<td>.938</td>
<td>5</td>
</tr>
<tr>
<td>122</td>
<td>1.922</td>
<td>2</td>
<td>1.595</td>
<td>1</td>
</tr>
<tr>
<td>211</td>
<td>.317</td>
<td>6</td>
<td>.787</td>
<td>6</td>
</tr>
<tr>
<td>212</td>
<td>1.930</td>
<td>1</td>
<td>1.421</td>
<td>4</td>
</tr>
<tr>
<td>221</td>
<td>1.738</td>
<td>3</td>
<td>1.536</td>
<td>3</td>
</tr>
<tr>
<td>222</td>
<td>.000</td>
<td>8</td>
<td>.515</td>
<td>8</td>
</tr>
</tbody>
</table>

Design Matrix

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>-1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>-1.0</td>
<td>1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>-1.0</td>
<td>-1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
</tr>
</tbody>
</table>

CARPE DIEM

Read from the top to the bottom this print out can be interpreted as follows. After the program title and a authorship statement the program presents the marginal frequencies. The order of the variables is the same as the order in which the variables were input. The sample size is given next. In the next line, the program confirms the choice of significance test. In the present example, this was Lehmacher’s test with Küchenhoff’s continuity correction. Bonferroni adjustment of the test-wise \( \alpha \) resulted in the adjusted \( \alpha^* = 0.05/8 = 0.00625 \). A statement confirming that a first order CFA is performed is followed by the table of results. The columns of this table contain (1) the indices of the cells of the cross-tabulation; (2) the observed cell frequencies, labeled fo; (3) the estimated expected
cell frequencies, labeled $f_e$; (4) the values of the selected test statistic$^3$; (5) the one-sided tail probabilities of the tests statistic; and (6) if applicable, the designation of a configuration as constituting a type or an antitype.

The present analysis suggests that there exist three types and three antitypes. For purposes of illustration we interpret the first type and the first antitype. The first type has cell-index pattern 111. It suggests that LSD 50 causes more participants than expected from chance to experience all three symptoms, that is, narrowed consciousness, thought disturbances, and affective disturbances. The chance model had been specified in the CFA base model in which we had postulated that the three symptoms are not associated (this is the log-linear main effect model or model of variable independence). The first antitype has cell-index pattern 112. It suggests that presence of the first two symptoms and absence of the third symptom co-occur less often than expected from the chance model. For a substantive interpretation of the complete results see Lienert (1964).

Under the frequency table there is information on the goodness-of-fit of the CFA base model. This is given in units of a Pearson $X^2$ that is followed by its degrees of freedom and tail probability.

The table below the significance test results is new to Version 2000 of this CFA program. This table displays two coefficients that are also used for data mining in large sparse contingency tables and in Bayesian analysis of cross-classifications (DuMouchel, 1999; von Eye, & Gutiérrez-Peña, in preparation). These two coefficients are the relative risk ratio, RR, and $LogP$. RR is defined as

$$RR_i = \frac{n_i}{m_i},$$

where $i$ indexes the cells in a cross-tabulation. RR is easy to interpret. It indicates the relative frequency of the occurrence of a configuration, given the expectation from the base model. For example, a score of RR = 10 suggests that ten times as many cases were observed

$^3$When the binomial tests are selected, this column is omitted.
as expected from the base model. RR can assume very large values when the expected cell frequencies are small. LogP is defined as

$$\log P_i = -\log_{10}(Pr[X \geq n_i]),$$

where $X \sim \text{Poisson}(m_i)$. LogP can be interpreted as the Poisson probability that the observed cell frequency is smaller than the expected cell frequency. For example, for Cell 122 in the above example, we calculated LogP = 1.595. This indicates a probability of $p = 10^{-1.595} = 0.0254$.

It is important to note that the status of RR and LogP in this context is that of descriptive measures rather than significance tests. Therefore, rather than printing a probability for each measure, the scores are ranked and the ranks are printed. Thus, two goals can be accomplished. First, when the sample size-to-table size ratio is large enough, the usual CFA significance tests can be employed and interpreted. Second, when the table is sparse and the tests can not be taken seriously any longer, the descriptive measures can be used as indicators of the degree to which the discrepancy between n and m is extreme. Type and antitype decisions can then be based on selecting the $\alpha\%$ most extreme discrepancies.

In the above example, the configurations identified as types and antitypes are among the most extreme ones in the rank order of LogP values. Note, however, that the most extreme RR (Configuration 212) constitutes neither a type nor an antitype. For details how these measures relate to each other see DuMouchel (1999) or von Eye and Gutiérrez-Peña (in preparation).

The last part of the print out - optional in Version 2000 of this CFA program - is the design matrix that was used to estimate the expected cell frequencies. The design matrix contains all vectors needed for the main effects and interactions in the model. The effects are expressed in terms of effect coding. The constant vector is implied. Covariates are part of this protocol if they are part of the CFA base model.

CARPE DIEM means SEIZE THE DAY.

2.2 Two-Sample CFA with Two Predictors; Keyboard Input

Two-sample CFA allows researchers to compare two independent groups of individuals. This variant of CFA can only find discrimination types (no discrimination antitypes). The reason is that if there are more cases than expected from the base model in one group, there must be fewer cases than expected in the other group. This is by necessity because CFA typically estimates cell frequencies
such that the marginal frequencies are reproduced. The two exceptions to this strategy are von Eye’s (1985) CFA of directed relationships and CFA of differences (von Eye, 1990, Chap. 6), which is not log-linear. The log-linear model for two-sample CFA is \([P][G]\), where \(P\) comprises all variables used to discriminate between the two groups, and \(G\) is the grouping variable.

The following sample run re-analyzes Lienert’s suicide data (see von Eye, 1990, Table 5.18). The data describe suicide patterns in pre- and post-WW II Germany for males (=1) and females (=2). In the years 1952 (=1) and 1944 (=2) the numbers of incidences were counted in which suicide was committed by gassing (=1), hanging (=2), use of drug overdose (=3), drowning (=4), cutting veins (=5), shooting (=6), and jumping (=7). The base model for the following gender comparison is \([\text{Year; Means of Suicide}][\text{Gender}]\). This model is saturated in the predictors, that is, it takes into account the main effects for Year and Means of Suicide and the interaction between Year and Means of Suicide. In addition, the model assumes independence between the two predictors and Gender. Therefore, if there is an interaction between the two predictors and Gender, there must be a difference between the gender groups and Means of Suicide, for a given year.

The following example pursues two objectives. The first is to illustrate how the CFA program can be used to perform two-sample CFA. The second objective is to show how one can perform CFA for regional models, that is for models where variables differ in status. In the present example there are two predictors and one grouping variable. The interaction between the two predictors is part of the two-sample CFA base model, but the interactions between the two predictors and the criterion are not part of the base model. The following paragraphs illustrate how to estimate expected cell frequencies for a model with two interacting predictors and one independent criterion.

Consider the above model, \([P][G]\). The cross-tabulation of the two predictors is \(P_1 \times P_2\). This cross-tabulation has \(I \times J\) cells. It contains all the information available. The saturated model also exhausts all available information. Therefore, one can use this table and declare the cells of this table the categories of a composite predictor. Suppose, for example, that \(I = J = 2\). Then, the indices of the cells of the cross-tabulation of these two variables are 11, 12, 21, and 22. Now, we declare these four cells to be the four categories of a composite predictor and obtain for the indices 1 = 11, 2 = 12, 3 = 21, and 4 = 22. This applies accordingly for three or more variable categories, there are three or more predictor variables, and two or more criterion variables in Prediction CFA.
When using the CFA program for two-sample CFA we indicate to the program that we have a dichotomous variable for the grouping. This must be the last in the list of variables, that is, the fastest changing variable. The first variable is either a composite predictor that results from crossing all predictors, or a series of one or more predictor variables. Results do not depend on the definition of the predictors. The following Table summarizes data and command input via keyboard.

<table>
<thead>
<tr>
<th>Command</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>double click shortcut to CFA program</td>
<td>starts CFA program; program window appears on screen. The program asks whether data will be entered via file (=1) or keyboard (=2). We enter data via keyboard and type 2</td>
</tr>
<tr>
<td>2</td>
<td>The program now asks for the number of variables. We type 2 (Enter)</td>
</tr>
<tr>
<td>2 (Enter)</td>
<td>The program prompts the number of categories for the predictor. We have a composite predictor that results from crossing a 2-category with a 7-category variable. Thus, we have a 14-category composite predictor and type 14 (Enter)</td>
</tr>
<tr>
<td>14 (Enter)</td>
<td>For the gender variable we type 2 (Enter)</td>
</tr>
<tr>
<td>The program then prompts the cell frequencies. We type 52, 47, 31, 14, 44, 97, 20, 10, 22, 5, 3, 0, 2, 2, 16, 61, 76, 35, 7, 9, 19, 54, 15, 4, 35, 11, 9, 2</td>
<td>Each of these numbers is followed by Enter (no commas need to be entered). The first number in this pattern is the frequency with which males committed suicide by a given means and in a given year. The second number is the frequency for this pattern for females. After completion of data input the program asks whether the user wishes to save the data. We type 1 (Enter)</td>
</tr>
<tr>
<td>suicide.dat (Enter)</td>
<td>to indicate that yes. After the prompt we give suicide.dat (Enter) for the data file name. Up to 80 spaces can be used for the file name.</td>
</tr>
</tbody>
</table>
The program then asks what model the user wishes to run. We type 20 (Enter) to indicate that we want a two-sample CFA. The program then presents the marginal frequencies and requests the significance level. We type .05 (Enter) The program then prompts the name for the output file. We type suicide.out (Enter) The program responds by presenting the current options for significance tests. Our samples are relatively large. Therefore we can select one of the z-tests. We select the z-approximation of the binomial test and type 3 (Enter) The program then asks whether the user wishes to perform a first order CFA using the same data. We indicate that no by typing 2 (Enter) The program closes and the program window disappears.

The following protocol contains the slightly edited result file, suicide.out.

```
Configural Frequency Analysis
----------- ------------
author of program: Alexander von Eye, 2000

Marginal Frequencies
---------------------
Variable Frequencies
------- -----------
 1  99.  45.  141.  30.  27.  3.  4.  77.  111.  16.  73.  19.  46.  11.
 2  351. 351.

sample size N = 702.
The z-approximation of the binomial test will be performed

Bonferroni-adjusted alpha = .0035714

Table of results
----------- ----- -------------- ---- ---- ----
Configuration f statistic p pi* Type?
----------- ----- -------------- ---- ---- ----
```
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>52.</td>
<td>0.542</td>
<td>0.293839</td>
<td>0.404</td>
</tr>
<tr>
<td>21</td>
<td>31.</td>
<td>2.620</td>
<td>0.004402</td>
<td>0.179</td>
</tr>
<tr>
<td>31</td>
<td>44.</td>
<td>-4.993</td>
<td>0.000000</td>
<td>0.454</td>
</tr>
<tr>
<td>41</td>
<td>20.</td>
<td>1.866</td>
<td>0.031017</td>
<td>0.050</td>
</tr>
<tr>
<td>51</td>
<td>22.</td>
<td>3.336</td>
<td>0.000424</td>
<td>0.222</td>
</tr>
<tr>
<td>61</td>
<td>3.</td>
<td>1.736</td>
<td>0.041303</td>
<td>0.500</td>
</tr>
<tr>
<td>71</td>
<td>2.</td>
<td>0.000</td>
<td>0.500000</td>
<td>0.389</td>
</tr>
<tr>
<td>81</td>
<td>16.</td>
<td>-5.435</td>
<td>0.000000</td>
<td>0.426</td>
</tr>
<tr>
<td>91</td>
<td>19.</td>
<td>-4.328</td>
<td>0.000000</td>
<td>0.417</td>
</tr>
<tr>
<td>101</td>
<td>7.</td>
<td>.506</td>
<td>0.306499</td>
<td>0.847</td>
</tr>
<tr>
<td>111</td>
<td>19.</td>
<td>.516</td>
<td>0.041303</td>
<td>0.331</td>
</tr>
<tr>
<td>121</td>
<td>15.</td>
<td>2.558</td>
<td>0.005257</td>
<td>0.723</td>
</tr>
<tr>
<td>131</td>
<td>35.</td>
<td>3.661</td>
<td>0.000126</td>
<td>0.091</td>
</tr>
<tr>
<td>141</td>
<td>9.</td>
<td>2.127</td>
<td>0.016697</td>
<td>0.389</td>
</tr>
</tbody>
</table>

### Alternative Measures of Deviation from Independence

| f1 f2 lambda lambda t rho delta theta |
|---|---|---|---|---|---|---|---|---|---|
| .297 | .200 | .020 | .020 | .018 | .055 | .038 | .035 | .035 | .217 |
| .537 | .539 | .578 | .578 | .542 | .29559 | .29506 | .28155 | .28155 | .29391 |
| .212 | .104 | .099 | .099 | .847 | .085 | .042 | .055 | .055 | .331 |
| 2.483 | 2.474 | 1.782 | 1.782 | 2.555 | 2.483 | 2.474 | 1.782 | 1.782 | 2.555 |
| .00652 | .00668 | .03736 | .03736 | .00530 | .00652 | .00668 | .03736 | .03736 | .00530 |
| .245 | .196 | -1.888 | 1.888 | -9.800 | .051 | .041 | .030 | .030 | .201 |
| .00000 | .00000 | .00000 | .00000 | .00000 | .00000 | .00000 | .00000 | .00000 |
| .810 | .073 | .070 | .070 | .723 | .103 | .041 | .069 | .069 | .395 |
| 1.755 | 1.764 | 1.024 | 1.024 | 1.831 | 1.755 | 1.764 | 1.024 | 1.024 | 1.831 |
| .03959 | .03833 | .15301 | .15301 | .03358 | .03959 | .03833 | .15301 | .15301 | .03358 |
| .383 | .147 | .126 | .126 | 1.532 | .383 | .147 | .126 | .126 | 1.532 |
Read from the top, this print out can be interpreted in a fashion parallel to the print out in Section 2.1. The table of results, however, is arranged differently. More specifically, the table of results presents the frequency for a predictor pattern for the two groups always in pairs of lines. Here the frequencies for the males appear first and the frequencies for the females appear second. The information whether a discrimination type was found always appears in the second line. For example, consider the third pair of lines, that is, the lines with indices 31 and 32. This is the predictor pattern *Suicide by drug overdose in 1952*. The discrimination type suggests that this pattern is
observed more often in females ($m_{32} = 97$) than in males ($m_{31} = 44$). In contrast, the next discrimination type suggests that, in 1952, males committed suicide by cutting veins more often than females ($m_{51} = 22$ versus $m_{52} = 5$). The remaining three discrimination types can be interpreted accordingly.

New to *Version 2000*, the column between the tail probabilities and the designation of a pair of cells as discrimination type displays the coefficient $\pi^*$. This coefficient is based on a definition of goodness of fit proposed by Rudas, Clogg, and Lindsay (1994), and was introduced for use in CFA by Gonzáles Debén (1998; see also Gonzáles Debén & Méndez Ramírez, 1999). Consider a 2 x 2 cross-classification with cell frequencies A, B, C, and D, and with sample size n. Then, $\pi^*$ is

\[ \pi^* = \frac{AD - BC}{An} \]  

The range of admissible values for $\pi^*$ is $0 \leq \pi^* \leq 1$. If $BC > AD$, $\pi^*$ can become negative and greater than 1. Therefore, the program calculates

\[ \pi^* = \frac{BC - AD}{Bn} \text{ if } BC > AD. \]

Also new to *Version 2000* of this CFA program, the block of significance test results is followed by a block of five other measures of deviation from independence and their significance tests. These measures, $\lambda$, $\tilde{\lambda}$, $\rho$, $\Delta$, and $\theta$ were introduced by Goodman (1991) and proposed for use in CFA by von Eye, Spiel, and Rovine (1995; cf. Gonzáles Debén & Méndez Ramírez, 1999). These measures are defined as follows. $\rho$ is the correlation in a 2 x 2 table. $\theta$ is the log odds ratio. $\Delta = |\rho|$. $\lambda$ is the log-linear parameter $|\log(\theta)|/4$, and

\[ \tilde{\lambda} = |\log(\theta)| \sqrt{(A+B)(C+D)(A+C)(B+D)}. \]

$\Delta$ and $\rho$ are related to Pearson's $X^2$. Both are marginal-dependent. $\lambda$ is marginal-free. It is related to the odds ratio, $\theta$ which is marginal-free too. $\tilde{\lambda}$ is a marginal-dependent quantity related to $\theta$. It can be interpreted as a weighted log-linear interaction with row- and column marginals as weights. For details see Goodman (1991) and von Eye et al. (1995; cf. Gonzáles Debén & Méndez Ramírez, 1999). For each of these measures a standard error, a z-score, and a one-sided tail probability is printed. For $\lambda$, $\tilde{\lambda}$, $\rho$, and $\Delta$ these values are estimated using the jack-knifing procedure described in the appendix of von Eye et al. (1995). The standard error of the log odds ratio is estimated as described, e.g., by Christensen (1997, p. 30). If one of the comparison frequencies is zero, the z-score and the tail probabilities are not estimated, and the standard errors are printed as zero. The present example suggests that these five measure can lead to quite discrepant appraisals of the two samples.
The only exception includes \( \rho \) and \( \Delta \) which differ only in sign (if the correlation is negative).

There is no design matrix included in the protocol of two-sample CFA. However, the design matrix used by the program is created using the same method as the design matrices for the main effect models.

2.3 Second Order CFA; Data Input via File

This section illustrates the use of second order CFA and data input via file. The base model of first order CFA does not consider any variable interactions. In contrast, the base model of second order CFA considers all pair-wise interactions. For example, consider the three variables, A, B, and C. In bracket notation, the base model for first order CFA of these variables is, \([A][B][C]\). The base model for second order CFA of these three variables is \([AB][AC][BC]\). This is a hierarchical log-linear model that implies the lower order terms, that is, in the present example, the main effects of all variables.

To illustrate second order CFA we use the LSD data from Section 2.1 again. We now assume that these data are available in the file named leuner.dat. The following print out displays the contents of this file:

```
3 2 2 2
20.
1.
4.
12.
3.
10.
10.
0.
```

This file shows how data files must be structured to be readable for the CFA program. In a first string, the CFA program expects information about the size of the cross-tabulation to be analyzed. Specifically, the program expects to read the number of variables and then, for each variable, the number of categories. For both the number of variables and the number of categories for each variable, three places are used. In the present example, the first line of the data file indicates that we have three variables with two categories each. Next, the program expects to read the observed cell frequencies. It is important to note that the cells must be in the proper order, with the fastest changing variable being the last in the array. Please notice the periods after the frequencies. The format in which the frequencies are read is \((x, f6.0)\), where the \(x\) indicates a blank at the beginning of the row. If the frequencies are presented with the period, they can appear anywhere within the six
columns. If the period is omitted, the last digit must be placed in the sixth column of the format, that is the seventh column of the line.

The following commands must be issued to perform Second Order CFA with the LSD data.

<table>
<thead>
<tr>
<th>Command</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>double click shortcut to CFA program</td>
<td>starts program, opens program window on screen. The program then asks whether the data will be entered via file or keyboard (interactively). We select input via file and type</td>
</tr>
<tr>
<td>1 (Enter)</td>
<td>The program then prompts the name for the input file. We type</td>
</tr>
<tr>
<td>leuner.dat (Enter)</td>
<td>The program confirms on the screen that this file is now open and presents the CFA model options. To calculate a second order CFA we type</td>
</tr>
<tr>
<td>2 (Enter)</td>
<td>The program confirms the selection, presents the marginal frequencies, and asks whether a covariate will be included. We have no covariate and type</td>
</tr>
<tr>
<td>2 (Enter)</td>
<td>The program then presents the available significance tests. Because the sample size is relatively small we can select the binomial test. Thus, we type</td>
</tr>
<tr>
<td>1 (Enter)</td>
<td>The program then prompts the significance level. We type</td>
</tr>
<tr>
<td>.05 (Enter)</td>
<td>The program then asks for the name of the output file. We type</td>
</tr>
<tr>
<td>leuner2.out (Enter)</td>
<td>The program writes the results to the file leuner2.out. We opt to include the design matrix in the output and key</td>
</tr>
</tbody>
</table>
The following protocol presents the results of second order CFA of the LSD data:

Configural Frequency Analysis

author of program: Alexander von Eye, 2000

Marginal Frequencies

Variable Frequencies

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37.</td>
<td>28.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>34.</td>
<td>31.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>42.</td>
<td>23.</td>
<td></td>
</tr>
</tbody>
</table>

sample size N = 65
Bonferroni-adjusted alpha = .0062500
a CFA of order 2 was performed
significance testing used binomial test

Table of results

<table>
<thead>
<tr>
<th>Configuration</th>
<th>fo</th>
<th>fe</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>20.</td>
<td>14.200</td>
<td>.05983935</td>
</tr>
<tr>
<td>112</td>
<td>1.</td>
<td>6.800</td>
<td>.00652806</td>
</tr>
<tr>
<td>121</td>
<td>4.</td>
<td>9.800</td>
<td>.02423844</td>
</tr>
<tr>
<td>122</td>
<td>12.</td>
<td>6.200</td>
<td>.01898418</td>
</tr>
<tr>
<td>211</td>
<td>3.</td>
<td>8.800</td>
<td>.01798567</td>
</tr>
<tr>
<td>212</td>
<td>10.</td>
<td>4.200</td>
<td>.00860371</td>
</tr>
<tr>
<td>221</td>
<td>15.</td>
<td>9.200</td>
<td>.03572556</td>
</tr>
<tr>
<td>222</td>
<td>0.</td>
<td>5.800</td>
<td>.00229865</td>
</tr>
</tbody>
</table>

Antitype

chi2 for CFA model = 37.4653
df = 1   p = .00000000

Descriptive indicators of types and antitypes

<table>
<thead>
<tr>
<th>cell</th>
<th>Rel. Risk</th>
<th>Rank</th>
<th>logP</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>1.408</td>
<td>4</td>
<td>1.130</td>
<td>4</td>
</tr>
<tr>
<td>112</td>
<td>.147</td>
<td>7</td>
<td>.719</td>
<td>6</td>
</tr>
<tr>
<td>121</td>
<td>.408</td>
<td>5</td>
<td>.644</td>
<td>8</td>
</tr>
<tr>
<td>122</td>
<td>1.936</td>
<td>2</td>
<td>1.615</td>
<td>2</td>
</tr>
<tr>
<td>211</td>
<td>.341</td>
<td>6</td>
<td>.661</td>
<td>7</td>
</tr>
<tr>
<td>212</td>
<td>2.381</td>
<td>1</td>
<td>1.959</td>
<td>1</td>
</tr>
<tr>
<td>221</td>
<td>1.630</td>
<td>3</td>
<td>1.344</td>
<td>3</td>
</tr>
<tr>
<td>222</td>
<td>.000</td>
<td>8</td>
<td>.770</td>
<td>5</td>
</tr>
</tbody>
</table>

Design Matrix

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>-1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>-1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

This command concludes the run and closes the program window.
As is obvious from the model specification, considering all pair-wise interactions carries the model closer to the saturated model than the base model of first order CFA. Thus, the second order CFA estimated expected cell frequencies will, on average, be closer to the observed cell frequencies than the first order CFA estimated expected cell frequencies. A comparison of the results from the protocol in Section 2.1 with the present protocol confirms this. The discrepancies between the observed and the expected cell frequencies are smaller, and there is only one antitype left. This is the antitype of those individuals that did not experience any of the LSD effects. Under the second order CFA base model 5.8 cases were expected to show no effects. However, none were observed.

2.4 CFA with Covariates; Input via File (Frequencies) and Keyboard (Covariate)
This section illustrates the use of covariates in CFA (Glück & von Eye, in press; von Eye, Spiel, & Rovine, in press). Consider, for example, the log-linear CFA base model given in (1). The inclusion of covariates leads to the model,

\[
\log M = X_b \lambda_b + X_c \lambda_c,
\]

where \(X_c\) is a matrix with the covariates in its columns, and \(\lambda_c\) is the

---

4 Notice that the binomial test is also less powerful than Lehmacher’s test. Thus, differences in power can also contribute to this difference in results. The Lehmacher test is not applicable in Second Order CFA. Therefore, a direct comparison between results from the two base models is not possible when Lehmacher’s test is used.
parameter for the covariate. Subscript \( b \) refers to the base mode. Using a covariate implies that more information than in standard base models is used when estimating expected cell frequencies. As a result, the expected cell frequencies typically (but not always) are closer to the observed cell frequencies, and it is less likely that types and antitypes will emerge. Covariates in CFA are particularly useful when there is information that may systematically vary over the cells of a cross-classification.

The following example illustrates the use of covariates by re-analyzing the data presented by Khamis (1996; cf. von Eye, Spiel, & Rovine, in press). The data describe the use of Cigarettes (C), Alcohol (A), and Marijuana (M) in a sample of \( N = 2276 \) high school students. Each drug was scored as either used (= 1) or not used (= 2). These data can be analyzed using, for instance, log-linear modeling (Khamis, 1996) or CFA (see below). Now suppose that, after a first analysis it becomes known in an imaginary re-analysis that all of those students that use both marijuana and alcohol also have police records for traffic violations (\( V = 1 \)), and none of the others are known for traffic violations (\( V = 2 \)). One may now ask whether knowledge of this covariate changes CFA results. The following equation gives the CFA base model with covariate for the present example. The base model is a log-linear main effects model, that is, a model that includes all main effects but no interaction.

\[
\log \left( \hat{m}_{ijk} \right) = \lambda_0 + \lambda_C \hat{m}_{111} + \lambda_A \hat{m}_{112} + \lambda_M \hat{m}_{121} + \lambda_{CM} \hat{m}_{122} + \lambda_{AM} \hat{m}_{211} + \lambda_{AM} \hat{m}_{212} + \lambda_{AM} \hat{m}_{221} + \lambda_{AM} \hat{m}_{222}
\]

The vector on the left hand side of the equation represents the expected cell frequencies, \( \hat{m}_{ijk} \). The matrix right after the equal sign is the indicator matrix. The first column in this matrix, a column of constants, is needed for estimation of the ‘grand mean parameter,’ \( \lambda_0 \). The following three columns contain the indicator variables for the main effects of variables C, M, and A. The second summand in this equation contains the vector for the covariate, multiplied by the one-element vector for the
covariate parameter.

Table 1 (from von Eye et al., in press) summarizes the results of standard, first order CFA of these data without the covariate. The results with covariate appear in the following output protocol. CFA was performed using the normal approximation of the binomial test with Bonferroni adjustment of the testwise $\alpha$. The adjusted $\alpha^*$ was 0.00625. Types are labeled with T; antitypes are labeled with A.

Table 1: CFA of Khamis’ Drug Use Data

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Observed Frequencies</th>
<th>CFA without Covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>f_e</td>
<td>z</td>
</tr>
<tr>
<td>CAM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>279</td>
<td>64.88</td>
</tr>
<tr>
<td>112</td>
<td>2</td>
<td>47.33</td>
</tr>
<tr>
<td>121</td>
<td>456</td>
<td>386.70</td>
</tr>
<tr>
<td>122</td>
<td>44</td>
<td>282.09</td>
</tr>
<tr>
<td>211</td>
<td>43</td>
<td>124.19</td>
</tr>
<tr>
<td>212</td>
<td>3</td>
<td>90.60</td>
</tr>
<tr>
<td>221</td>
<td>538</td>
<td>740.23</td>
</tr>
<tr>
<td>222</td>
<td>911</td>
<td>539.98</td>
</tr>
</tbody>
</table>

The application of first order CFA with no covariate suggests that more high school students than expected from the assumption of variable independence use all three drugs, Marijuana, Alcohol, and Cigarettes (Type 111); fewer students than expected use only Cigarettes and Alcohol (Antitype 112); more students than expected use only Marijuana and Cigarettes (Type 121); fewer students than expected use only Cigarettes (Antitype 122), only Alcohol and Marijuana (Antitype 211), only Alcohol (Antitype 212), or only Marijuana (Antitype 221); and more students than expected do not
use any of the three drugs (Type 222).

Also considering the (hypothetical) citation record creates a different picture (the complete output follows below) (cf. Mellenbergh, 1996). The discrepancies between the observed and the expected cell frequencies are, on average, smaller and the overall $X^2$ is smaller by almost one half (824.16 from 1411.39). In spite of the large sample size, the resulting pattern of types and antitypes is no longer the same. Configuration 112 no longer constitutes an antitype and neither does Configuration 212.

The following table and output illustrate the use of the CFA program for first order CFA with a covariate. We assume that the data are stored in a file named „Khamis2.dat.“ This file only contains the frequencies of the cross-tabulation. The covariate will be entered via the keyboard. The following output displays the data file:

```
3  2  2  2
279.
2.
456.
44.
43.
3.
538.
911.
```

The following commands are needed to perform first order CFA of Khamis’ drug data with a covariate.

<table>
<thead>
<tr>
<th>Command</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>double click shortcut to CFA program</td>
<td>starts program. The program asks how the data will be entered. We type</td>
</tr>
<tr>
<td>1 (Enter)</td>
<td>thus indicating that the data will be read from a file. The program then prompts the name of the data file. We type</td>
</tr>
<tr>
<td>khamis2.dat (Enter)</td>
<td>The program responds by confirming that this file has been opened and presents the CFA model options. We type</td>
</tr>
<tr>
<td>1 (Enter)</td>
<td>to indicate that we wish to calculate a first order CFA. Next the program asks whether we would like to include a covariate. We type</td>
</tr>
</tbody>
</table>
1 (Enter) to indicate that we wish to use a covariate. The program requests the values of the covariate for each cell. We type 1, 2, 2, 1, 2, 2, 1, 2, 2 each value followed by Enter (commas do not need to be entered). The program then asks whether another covariate will be entered. We type 2 (Enter) thus indicating that we have only one covariate. We then type 4 (Enter) to indicate our selection of the z-test, we type .05 (Enter) to indicate the significance level, and we type khamis2.out (Enter) to name the output file. Entering 1 (Enter) includes the design matrix in the output.

The following protocol displays the contents of the output file khamis2.out:

```
Configural Frequency Analysis
---------- --------- --------
author of program: Alexander von Eye, 2000

Marginal Frequencies
--------------------
Variable Frequencies
-------- -----------
1     781. 1495. 
2     327. 1949. 
3    1316.  960. 
sample size N =  2276
the normal z-test was used
Bonferroni-adjusted alpha = .0062500
a CFA of order   1  was performed

Table of results
---------------
Configuration    fo       fe   statistic       p
--------------- ----  -------- --------- -----
111    279.  110.493     16.031   .00000000    Type
112      2.    1.716       .217   .41409490
121    456.  341.087      6.222   .00000000    Type
122      44.  327.704    -15.672   .00000000    Antitype
211      43.  211.507    -11.587   .00000000    Antitype
212      3.    3.284      -.157   .43767758
221    538.  652.913     -4.497   .00003458    Antitype
222    911.  627.296     11.327   .00000000    Type
```
Descriptive indicators of types and antitypes

<table>
<thead>
<tr>
<th>cell</th>
<th>Rel. Risk</th>
<th>Rank</th>
<th>logP</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>2.525</td>
<td>1</td>
<td>40.467</td>
<td>3</td>
</tr>
<tr>
<td>112</td>
<td>1.166</td>
<td>4</td>
<td>.298</td>
<td>7</td>
</tr>
<tr>
<td>121</td>
<td>1.337</td>
<td>3</td>
<td>8.890</td>
<td>5</td>
</tr>
<tr>
<td>122</td>
<td>.134</td>
<td>8</td>
<td>83.438</td>
<td>1</td>
</tr>
<tr>
<td>211</td>
<td>.203</td>
<td>7</td>
<td>42.527</td>
<td>2</td>
</tr>
<tr>
<td>212</td>
<td>.913</td>
<td>5</td>
<td>.231</td>
<td>8</td>
</tr>
<tr>
<td>221</td>
<td>.824</td>
<td>6</td>
<td>5.703</td>
<td>6</td>
</tr>
<tr>
<td>222</td>
<td>1.452</td>
<td>2</td>
<td>25.896</td>
<td>4</td>
</tr>
</tbody>
</table>

Design Matrix

```
1.0 1.0 1.0 1.0
1.0 1.0 -1.0 2.0
1.0 -1.0 1.0 2.0
1.0 -1.0 -1.0 2.0
-1.0 1.0 1.0 1.0
-1.0 1.0 -1.0 2.0
-1.0 -1.0 1.0 2.0
-1.0 -1.0 -1.0 2.0
```

CARPE DIEM

This protocol largely replicates the results presented by von Eye et al. (in press). Differences to the second panel of Table 1, above, reflect differences in the power of the statistical tests. Note that Lehmacher’s tests are not applicable when covariates are used.

3. CFA and Log-linear Models

The relationship between log-linear modeling and CFA has often been disputed (e.g., Krauth, 1980; Lehmacher, 1984; von Eye, 1990). The two poles of the discussion are that CFA is either nothing else than residual analysis in log-linear modeling, or log-linear modeling and CFA are completely unrelated although they use the same methods of estimation of expected cell frequencies. The present program allows users to estimate virtually any log-linear model, including non-hierarchical and non-standard models, even if it does not constitute a model suited for CFA. The overall model fit will be calculated. Parameters will not be estimated. Users need to key in the complete design matrix (excluding the column with the constant scores).

Consider the following example (from von Eye & Niedermeier, 1999, p. 74 - 77 and p. 92ff). The data describe the effect that client-centered psychotherapy has on introverted inpatients. In a control group design a no treatment group was compared to a treatment group. The cross-classification was formed by the three variables Patient Group (G; 1 = treatment; 2 = control), Status
at pretest (B; 1 = critical symptom present; 2 = other symptoms present), and Status at posttest (P; 1 = critical symptom cured, 2 = critical or other symptoms present). The main hypothesis for the following analyses contrasts the number of patients in the treatment group that is cured from the symptoms with the number of patients cured in the control group. The design matrix for this hypothesis was taken from von Eye and Niedermeier (1999; Table 3.12).

To estimate the overall goodness-of-fit for the log-linear model that tests this hypothesis, one proceeds with the CFA program as usual. The question concerning the order of CFA must be answered with $0$, indicating that a zero-order CFA is requested. This is the only CFA model that allows one to key in all vectors of the design matrix (except the constant vector which is implied). The vectors in the design matrix that are necessary for the log-linear model are keyed in as covariates (see Section 2.4 of this manual). The results of the sample run appear in the following, slightly edited print out.

```
Configural Frequency Analysis
---------- --------- --------
author of program: Alexander von Eye, 2000

Marginal Frequencies
-------------------
Variable Frequencies
------- ----------
 1    112. 102.
 2     58.  156.
 3     47.  167.
sample size N = 214
Bonferroni-adjusted alpha = .0062500
a CFA of order 0 was performed
significance testing used binomial test

Table of results
----- ------- -------
Configuration fo  fe      p
----------- ---- ------- -------
 111      11. 11.082 .56929259
 112      21. 21.918 .47431824
 121      22. 16.418 .09941508
 122      58. 61.582 .32392977
 211       3.  3.082 .62882868
 212      23. 21.918 .43624475
 221      11. 16.418 .09837529
```
In this printout, the CFA results must not be interpreted because the non-standard log-linear model that was fitted to the data does not meet the criteria for a CFA model. However, the estimated expected cell frequencies and the overall goodness-of-fit results are the same as reported by von Eye and Niedermeier (1999). The overall model fit is excellent and the hypothesis is confirmed (see von Eye & Niedermeier, 1999, p. 77).

4. Numerical Accuracy

All real-valued calculations in this program are performed with double precision. However, the approximation routines for the significance tests can be trusted only up to 16 decimal places. Therefore, the new descriptive measures for the global models of CFA are to be preferred for the designation of configurations as types or antitypes (1) when the discrepancies between the observed and the expected cell frequencies are that big that 16 decimal places are insufficient to describe the
tail probabilities and (2) when the expected cell frequencies are so small that the approximation characteristics of the asymptotic tests are uncertain. The descriptive measures and the test statistics themselves are accurate up to 308 decimal places.

4. References


Wissenschaftliche Beiträge.


