SPRINGER BRIEFS IN STATISTICS

Mark Stemmler

Person-Centered Methods

Configural Frequency Analysis (CFA) and Other Methods for the Analysis of Contingency Tables

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Mark Stemmler

Person-Centered Methods

Configural Frequency Analysis (CFA) and Other Methods for the Analysis of Contingency Tables

Springer

Mark Stemmler
Institute of Psychology
Friedrich-Alexander University
of Erlangen-Nuremberg (FAU)
Erlangen, Germany

ISSN 2191-544X ISSN 2191-5458 (electronic)
ISBN 978-3-319-05535-0 ISBN 978-3-319-05536-7 (eBook)
DOI 10.1007/978-3-319-05536-7
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014937895

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The idea for this book came to me while I was teaching courses during the summer at The Methodology Center at Penn State (Director: Linda Collins). Teaching classes on person-centered methods which do not belong to the standard curriculum at German or American universities was very inspiring to me. The interaction with the students helped me to understand how to explain the content of this class so that it is easy to understand and showed to me how much the students liked this different look at statistics.

This book will take an easy-to-understand look at the statistical approach called the *person-centered method*. Instead of analyzing means, variances and covariances of scale scores as in the common variable-centered approach, the person-centered approach analyzes persons or objects grouped according to their characteristic patterns or configurations in contingency tables. The main focus of the book will be on ***Configural Frequency Analysis*** (CFA; Lienert and Krauth 1975) which is a statistical method that looks for over- and under-frequented cells or patterns. Over-frequented means that the observations in this cell or configuration are observed more often than expected, under-frequented means that this cell or configuration is observed less often than expected. In CFA a pattern or configuration that contains more observed cases than expected is called a *type*; similarly, a pattern or configuration that is less observed than expected are called an *antitype*. CFA is similar to log-linear modeling. In log-linear modeling the goal is to come up with a fitting model including all important variables. Instead of fitting a model, CFA looks at the significant residuals of a log-linear model.

CFA was invented by Gustav A. Lienert, an Austrian physician and professor of psychology, who died in 2001. I was lucky to have met Gustav A. Lienert, who was a very inspiring and enthusiastic person. I am thankful for his cheerfulness and his support. I was introduced to ‘Herrn Lienert’ by Alexander von Eye (Psychology Professor at Michigan State and University of Vienna). I am very thankful to Alex who has introduced me to the field of categorical data analysis.

A number of ideas presented here (especially those in Chap. 6) were proposed by Erwin Lautsch. They were all published in a series of Special Issues on CFA (guest editor together with Alexander von Eye) in the German Journal called Psychology Science (formerly known as the Psychologische Beiträge). Thank you Erwin for sharing your ideas!

One important asset to this book was the development of the R-package ***confreq*** (derived from **con**figural **freq**uency analysis). The open source software R is available at no cost and is developing in a fast and progressive manner. An R-package was also important because there was no readily available software for configural frequency analysis (with exception of a somewhat outdated DOS software written in FORTRAN). ***Confreq*** was written by Jörg-Hendrik Heine (LMU Munich). I met Jörg at our annual statistical meetings in Rothenberge (Northern Germany) organized by Christian Tarnai and Jost Reinecke. Jörg worked diligently on this package for more than 2 years including several setbacks. Many thanks to you Jörg! I am also thankful to Rainer Alexandrowicz (who I also met in Rothenberge) who worked on Stirlings's formula for using the binomial test as part of ***confreq***.

My thanks go out to Amanda Applegate and Heather Foran for proof reading my English. In addition, Heather also addressed to me all the relevant sections which were difficult to understand and not well explained. Her methodological perspective was extremely essential for my writing! Thanks also to Hannah Bracken at Springer for her support in leading my book endeavor.

Finally, I offer my deepest thanks to my wife Susanne and my son Quincy. Thanks for giving me so much comfort and for energizing my life.

Erlangen, Germany Mark Stemmler
Spring 2014

Reference

Lienert, G. A., & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. *Educational Psychology and Measurement, 35*, 231–238.

Contents

Chapter 1
Introducing Person-Centered Methods

Abstract This chapter explains the term *person-centered methods* and how ***Configural Frequency Analysis*** **(CFA)** works. Instead of analyzing means, variances and covariances of scale scores as in the common variable-centered approach, the person-centered approach analyzes persons or objects grouped according to their characteristic configurations in contingency tables. CFA is a statistical method that looks for over- and under-frequented cells or patterns. Over-frequented means, that the observations in this cell or configuration are observed more often than expected, under-frequented means that this configurations is observed less often than expected. In CFA a pattern or configuration that contains more observed cases than expected is called a *type*; similarly, configurations that are less observed than expected are called an *antitype*. In addition, Meehl's paradox (Meehl, J Consult Psychol 14:165–171, 1950) is explained. Meehl's paradox postulates that it is possible to have a bivariate relationship with a zero association or correlation but also a higher order association or correlation. Meehl argued for investigating higher order interactions (beyond bivariate interactions), which can be detected with CFA.

1.1 What Is Configural Frequency Analysis (CFA) Good for?

This chapter takes an easy-to-understand look at the statistical approach called the *person-centered method*. Instead of analyzing means, variances and covariances of scale scores as in the common variable-centered approach, the person-centered approach analyzes persons or objects grouped according to their characteristic patterns or configurations in contingency tables (see Bergman & Magnusson, 1997; Bergman, von Eye, & Magnusson, 2006; Reinecke & Tarnai, 2008; Stemmler & von Eye, 2012). The observed patterns are arranged in tables, ordered by their indices. A certain position in such a table, denoted by a pattern or configuration, is called a cell (Victor, 1989). Such tables are called contingency tables. The main focus of the book will be on ***Configural Frequency Analysis*** (CFA; Lautsch & von Weber,

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_1,

1995; von Eye, 2002; von Eye & Gutiérrez-Penã, 2004) which is a statistical method that looks for over- and under-frequented cells or patterns. Over-frequented means, that the observations in this cell or configuration are observed more often than expected, under-frequented means that this configurations is observed less often than expected. In CFA a pattern or configuration that contains more observed cases than expected is called a *type*; similarly, configurations that are less observed than expected are called an *antitype*. CFA was invented by Gustav A. Lienert, an Austrian physician and professor of psychology, who died in 2001 (Lienert & Krauth, 1975; Stemmler, Lautsch, & Martinke, 2008). CFA is similar to log-linear modeling. In log-linear modeling the goal is to come up with a fitting model including all important variables. Instead of fitting a model, CFA looks at the significant residuals of a log-linear model.

What is a typical research question that can be answered by CFA? Take an example from hydrobiology (Melcher, Lautsch, & Schmutzler, 2012). Let's say one is interested in fish habitats or specially spawning habitats of fish because a sufficient fish stock is important for the ecological system of a river. In logistic regressions you compare places with many fish with places with little fish. Based on logistic regression or loglinear modeling researchers know different important features of the river such as flow velocity, type of structure and substrate of the river bed, and the vegetation of the riverbanks, but they don't know the optimal combination of the features resulting in a typical (i.e., over-frequented) fish habitat.[1] With CFA one can identify significant cells, patterns or configurations. CFA gives answers at the level of individual cells (configurations) instead of the level of variables.

Take, for example, another research question from the field of pediatrics. In a small sample of premature newborns with additional neurological or other health problems (e.g., seizures, need to intubate) one is interested in the healthy (i.e., typical) cognitive development of the children by the age of 5. CFA is able to look at those newborns by searching for characteristic patterns or configurations that allow a normal cognitive development.[2]

There are three commonly used sampling models in the analyses of cross-classified data: (1) poisson, (2) multinomial, and (3) product-multinomial. Here, the different sampling models will not be elaborated, for further information the reader may consult Fienberg (1987). The good news is, that the three sampling schemes lead to the same estimated expected cell values and the same goodness-of-fit-statistics. Usually CFA assumes a multinomial sampling instead of a normal sampling distribution. That is, we usually deal with a fixed obtained sample size N and cross-tabulate each member of the sample according to its values for the underlying variables. This multinomial sampling model can be applied when most of the statistical assumptions for the use of multiple regression or analysis of variance are violated. These assumptions encompass frequent issues, such

[1] By the way, many European fish like a shaded habitat with a fine and coarse substrate depending on high flow velocity.

[2] Girls with intubation but no seizures have the best chances for normal cognitive development.

as small sample size, heteroscedasticity, extreme observations or non-normality. Such violations of important statistical assumptions threaten the statistical validity (Shadish, Cook, & Campbell, 2002), opening up the possibility that the associations found do not hold in reality.

However, multinomial statistics do not belong to the standard curriculum of graduate studies in the humanities or social sciences, leaving the student or researcher with few resources. This book will help to fill this gap by providing an easy to read, hands-on textbook which can be used in any non-introductory undergraduate or graduate statistics course. The textbook requires only knowledge of hypothesis testing or inference statistics but no advanced knowledge of multivariate statistics.

CFA is a very useful statistical tool for the analysis of multiway contingency tables, and CFA can be applied for anything that goes beyond the analysis of more than two categorical variables. The analysis of multidimensional cross-tables has many implications, such as

- Instead of using scale scores one looks at cell frequencies: i.e., one looks for persons or units with characteristic patterns or configurations;
- CFA has few requirements with regard to sample size;
- The underlying sampling distribution is the multinomial distribution, instead of the normal distribution;
- Instead of a linear combination

  $$y = b_0 + b_1X_1 + b_2X_2 + e \tag{1.1}$$

  we are dealing with a multiplicative relationship, which can be transformed into an additive relationship through the logarithm of the equation:

  $$ln\, e_{ij} = \lambda_0 + \lambda_i A_i + \lambda_j B_j. \tag{1.2}$$

- CFA belongs to the non-parametric methods.

1.2 Basics of CFA

The basic procedure of CFA is comparable to analyzing a cross-table with the chi-square statistic. The cross-table consists of *r* rows and *c* columns. Observed values are compared with expected values. The **global chi-square** value for a contingency table with two variables is calculated as follows:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \text{ with df} = (r-1)(c-1) \tag{1.3}$$

where *r* represents the *rows* and *c* represents the *columns*. Each observed frequency o_{ij} has two subscripts, *i* for the row frequencies $i = (1, 2, \ldots, I)$ and *j* for the column

Table 1.1 Data example for a two-by-two cross-table

	Pro	Con	Sum
Males	100	50	150
Females	60	190	250
	160	240	400

frequencies $j = (1, 2, \ldots, J)$. For the denotation of a total row or column, the **index-point notation** is applied. $o_{1.}$ therefore denotes the observed frequencies of row 1 ($o_{2.}$ denotes the observed frequencies of row 2). $o_{.1}$ denotes the observed frequencies of column 1 ($o_{.2}$ denotes the observed frequencies of column 2). The total N is $o_{..} = \sum_{i=1}^{I} o_{i.} = \sum_{j=1}^{J} o_{.j}$. The global chi-square tests the following null and alternative hypotheses: H_0: There is no significant association between the variables involved or the two variables are independent of each other and H_1: There is a significant association between the variables involved or the two variables are not independent of each other.

The expected values or frequencies (e_{ij}) are defined according to the null hypothesis or a **base model** (e.g., the assumption of independence). The assumption of independence is the null hypothesis of a first order CFA. Example (see Table 1.1):

$$e_{ij} = n\,p_{ij} \tag{1.4}$$

$$p_{ij} = p_{i.}p_{.j} \qquad e.g., \; p_{11} = p_{1.}p_{.1} \tag{1.5}$$

e_{ij} = expected frequencies (first subscript = row; second subscript = column)
p_{ij} = cell proportion; $p_{i.}$ = row proportion; $p_{.j}$ = column proportion

The corresponding formal (statistical) hypotheses for two variables can be stated as follows:

$$H_0 : \pi_{ij} = \pi_{i.}\,\pi_{.j}$$

$$H_1 : \pi_{ij} \neq \pi_{i.}\,\pi_{.j}$$

$$p_{i.} = \frac{o_{i.}}{n} \tag{1.6}$$

$$p_{.j} = \frac{o_{.j}}{n} \tag{1.7}$$

$$\begin{aligned} e_{ij} &= \frac{n \times o_{i.} \times o_{.j}}{nn} = \\ &= \frac{o_{i.}o_{.j}}{n} \end{aligned} \tag{1.8}$$

o_{ij} = observed frequencies (first subscript = row; second subscript = column)
n = sample size

$$e_{ij} = \frac{row\,frequencies \times column\,frequencies}{n} \tag{1.9}$$

$$e_{11} = \frac{150 \times 160}{400} = 60$$

$$e_{12} = \frac{150 \times 240}{400} = 90$$

$$e_{21} = \frac{250 \times 160}{400} = 100$$

$$e_{22} = \frac{250 \times 240}{400} = 150$$

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(100-60)^2}{60} + \frac{(50-90)^2}{90} + \frac{(60-100)^2}{100} + \frac{(190-150)^2}{150} \right]$$

$$\chi^2_{emp} = 71.11; \quad df = 1$$

$$\chi^2_{crit} = 6.635, \quad p < 0.01$$

Expected frequencies and the null hypothesis for a three dimensional cross-table are:

$$e_{ijk} = \frac{o_{i..}\,o_{.j.}\,o_{..k}}{n^2} \tag{1.10}$$

$$H_0 : \pi_{ijk} = \pi_{i..}\,\pi_{.j.}\,\pi_{..k}$$

Expected frequencies and the null hypothesis for a four dimensional cross-table are:

$$e_{ijkl} = \frac{o_{i...}\,o_{.j..}\,o_{..k.}\,o_{...l}}{n^3}$$

$$H_0 : \pi_{ijkl} = \pi_{i...}\,\pi_{.j..}\,\pi_{..k.}\,\pi_{...l}$$

We differentiate between the overall or global chi-square value and the **local chi-square value**. A significant global chi-square value is the prerequisite for a significant local chi-square. The local chi-square has one degree of freedom is calculated by:

$$\chi^2_{\mathrm{ij}} = \frac{(\mathrm{o_{ij}} - \mathrm{e_{ij}})^2}{\mathrm{e_{ij}}} \text{ with df} = 1 \tag{1.11}$$

Table 1.2 CFA example for a two-by-two cross-table

		$f_{(o)}$	$f_{(e)}$	χ^2	p-Value	Types/antitypes
Male	Pro	100	60	26.67	0.000	Type
Male	Con	50	90	17.77	0.000	Antitype
Female	Pro	60	100	16.00	0.000	Antitype
Female	Con	190	150	10.67	0.000	Type

If the global chi-square is significant, one may look for significant local chi-square values. Significant local chi-square values represent *types* or *antitypes*. Wickens (1989) preferred the term *outlandish cells*. *Types* represent significantly over-frequented cells ($f_{(o)} > f_{(e)}$) and *antitypes* represent significantly under-frequented cells ($f_{(o)} < f_{(e)}$). For the calculations of local or cell-wise chi-squares and the analysis of types and antitypes see Table 1.2.

One can use a pocket calculator to obtain the local chi-square for each cell (Note: each local chi-square is indicated by two subscripts: first subscript = row; second subscript = column). The corresponding local chi-square for configuration o_{12} in Table 1.1 would be

$$\chi^2_{12} = \frac{(50-90)^2}{90} = 17.77 \quad \text{with df} = 1.$$

By inserting the numbers into Eq. 1.3 you can see that the global chi-square value is 71.111, which is highly significant with $df = 1$. Computer software for CFA will be introduced later (see Chap. 2). There is the danger of an alpha inflation, because we are performing multiple tests. Therefore, we need a two-tailed Bonferroni alpha adjustment while we are looking for over- and under-frequented cells (cf. von Eye, 1990), that is $\alpha^* = 0.025/4 = 0.00625$. Even with $p = 0.00625$ two *types* and two *antitypes* could be identified. One *type* indicates that there are **more** males answering ‘pro’ to the asked question than expected under the null hypothesis, and the other *type* indicates that there are **more** women answering ‘con’ to the asked question than expected under the null hypothesis. To put it differently: Men typically say ‘pro’ to the asked question and women typically say ‘con’ to the asked question. The *antitypes* are explained in the same vein. One *antitype* suggests that there are **less** men than expected under the null hypothesis who answer ‘con’ to the question, and the other *antitype* suggests that there are **less** women than expected under the null hypothesis who answer ‘pro’ to the question. It is uncommon for men to answer ‘con’ to this question, and uncommon for women to answer ‘pro’. Gustav A. Lienert originally wanted to use his CFA only for exploratory purposes, but with the Bonferroni adjustment, hypothesis testing is allowed (Lehmacher, 2000).

Let us look at another (real) data example. It is from Gustav A. Lienert, the Austrian inventor of CFA, his nickname was *Gustl*. The data presented below are from the famous LSD (i.e., acid) studies done with psychology students of him (Krauth & Lienert, 1973), when he was a professor at the University of Marburg in Germany. Throughout his life Lienert was interested in the psychological effects

Table 1.3 Gustav A. Lienert's famous acid data

		C = Narrowed consciousness				
		1 = yes		2 = no		
		T = Thought disturbances				
		Yes	No	Yes	No	
A = Affective disturbances	Yes	20	4	3	15	
	No	1	12	10	0	
		21	16	13	15	$N = 65$

Table 1.4 Technical representation of a cross-table with three-variables

		Item A					
		1		2			
		Item B			Item B		
		1	2		1	2	
Item C	1	o_{111}	o_{112}	$o_{11.}$	o_{121}	o_{122}	$o_{12.}$
	2	o_{211}	o_{212}	$o_{21.}$	o_{221}	o_{222}	$o_{22.}$
		$o_{11.}$	$o_{12.}$		$o_{21.}$	$o_{22.}$	$N = o_{...}$

of pharmaceutical drugs on the human brain. Gustav A. Lienert was a character and he somehow managed to get a sample of acid from a pharmaceutical company. In the 1950s acid was still legal in Germany, and Lienert used it for experiments in his lab. In those times, LSD was an attractive substance for psychologists or medical doctors because it was hypothesized that LSD was mimicking pathological phenomena like psychosis. Today, this hypothesis is widely rejected. Lienert's LSD data expand the above two by two cross-table to three variables. The variables are as follows: C = Narrowed Consciousness; T = Thought Disturbances, and A = Affective Disturbances. Each symptom is rated as 1 = yes or 2 =no, resulting in eight cells.

Let's have a look at the data (see Table 1.3):

A technical representation of a table with three variable is as follows (see Table 1.4): What are the ***degrees of freedom*** for any multiway contingency table? The formula is

$$df = T - \sum_{i=1}^{d} (v_d - 1) - 1 \tag{1.12}$$

with T representing the number of cells or configurations, with $d = 1 \ldots D$ representing the number of variables (dimensions), and v_d the number of categories of a variable. Here, we have $T = 8$ cells, $d = 3$ variables and $v_d = 2$ categories, that is, $df = 8 - (2-1) - (2-1) - (2-1) - 1 = 4$. The corresponding **global chi-square** for a three dimensional table calculated as

Table 1.5 Expected and observed frequencies and the corresponding local chi-square for Lienert's LSD data

Configuration	o_{ijk}	e_{ijk}	Local chi-square	p
1 1 1	20	12.506	4.491	0.034
1 1 2	1	6.848	4.994	0.025
1 2 1	4	11.402	4.805	0.028
1 2 2	12	6.244	5.306	0.021
2 1 1	3	9.464	4.415	0.035
2 1 2	10	5.182	4.478	0.034
2 2 1	15	8.629	4.705	0.030
2 2 2	0	4.725	4.725	0.029

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{s} \frac{(o_{ijk} - e_{ijk})^2}{e_{ijk}} \tag{1.13}$$

where r represents the *rows*, c represents the *columns* and s represents the third dimension, i.e., *stratum*. Each observed frequency o_{ijk} has three subscripts. The index-point notation $o_{1..}$ denotes the observed frequencies of row 1 and so on. The total N is $o_{...} = \sum_{i=1}^{I} = \sum_{j=1}^{J} o_{.j.} = \sum_{k=1}^{K} o_{..k}$. The corresponding expected frequencies are obtained with Eq. 1.10. The assessment for the **local chi-square** has not changed apart from one additional index (see Eq. 1.11).

Let's have a look at the observed and expected frequencies and as well at the local chi-square with its related p-values of the LSD data (see Table 1.5).

The global chi-square $chi^2 = 37.92$ is highly significant. However, due to the Bonferroni adjustment $\alpha^* = 0.025/8 = 0.003125$ no *types* or *antitypes* could be detected. Thus, the results can only be interpreted in an exploratory way. The hypothesized *typical LSD-syndrome* with narrowed consciousness, affective disturbances and thought disturbances was detected 20 times with the corresponding expected values of $e_{111} = 12.506$, which is more often than expected. Overall, the data show that there is great inter-individual variation with regard to the drug response (not to mention its detrimental health effects on regular users).

The above section explained the essence of CFA, which is the search for over- and under-frequented cells (i.e., *types* or *antitypes*). First, a global chi-square is calculated because significant global chi-squares are the prerequisite for the detection of types and antitypes. Then, the local chi-square is obtained on each cell level. For significance testing the Bonferroni adjustment is recommended, otherwise the results may only be interpreted in an exploratory fashion. The calculation of a CFA for a multiway contingency table is tedious, therefore easy-to-use software packages are introduced in Chap. 2. In addition, other test statistics next to the chi-square are presented. The next section presents the need to look for higher order associations.

1.3 Meehl's Paradox

Meehl's paradox (1950) postulates that it is possible to have a bivariate relationship with a zero association or correlation but nevertheless a higher order association or correlation. Let's have a look at the following example. We ask two questions related to alcohol abuse in a sample of heavy drinkers (e.g., young male college students); each item may be answered with either 1 = *yes* or 2 = *no*. Item 1: "Have you ever experienced a black out?" and Item 2: "Have you ever developed a higher tolerance for alcohol?" Based on findings from the alcohol abuse literature, one may conclude that a person who says *yes* to both questions is at serious risk for becoming an alcoholic. This person belongs to the group of alcoholics (A). A person who says *no* to both questions is probably a dissimulating alcoholic, and still belongs to the group of alcoholics (A). And a person who says *yes* to one question has developed some risk for becoming an alcoholic but is still a non-alcoholic. Let's think of a sample of $N = 200$ who would give you the following frequencies; see Table 1.6:

Let's transform this table into a CFA Table; see Table 1.7:

Each two-by-two table results in a zero chi-square of no association (see Tables 1.8–1.10).

Table 1.6 Frequencies for an example of Meehl's paradox

Item 1	Item 2	Alcoholics (A)	Non-alcoholics (NA)
1 = yes	1 = yes	50	0
1 = yes	2 = no	0	50
2 = no	1 = yes	0	50
2 = no	2 = no	50	0

Table 1.7 CFA Table for an example of Meehl's paradox

Configurations		Subject status	$f_{(o)}$	$f_{(e)}$
1 = yes	1 = yes	A	50	25
1 = yes	1 = yes	NA	0	25
1 = yes	2 = no	A	0	25
1 = yes	2 = no	NA	50	25
2 = no	1 = yes	A	0	25
2 = no	1 = yes	NA	50	25
2 = no	2 = no	A	50	25
2 = no	2 = no	NA	0	25

Table 1.8 Crosstabs for Item 1 by Item 2

		Item 2		
		Yes	No	
Item 1	Yes	50	50	100
	No	50	50	100
		100	100	$N = 200$

Table 1.9 Crosstabs for Item 1 by Subject Status

		Subject status		
		A	NA	
Item 1	Yes	50	50	100
	No	50	50	100
		100	100	$N = 200$

Table 1.10 Crosstabs for Item 2 by Subject Status

		Subject status		
		A	NA	
Item 2	Yes	50	50	100
	No	50	50	100
		100	100	$N = 200$

Table 1.11 CFA Table for an example of Meehl's paradox

Configurations		Subject status	$f_{(o)}$	$f_{(e)}$	χ^2	p-Value	Types/antitypes
1 = yes	1 = yes	A	50	25	25	0.000	Type
1 = yes	1 = yes	NA	0	25	25	0.000	Antitype
1 = yes	2 = no	A	0	25	25	0.000	Antitype
1 = yes	2 = no	NA	50	25	25	0.000	Type
2 = no	1 = yes	A	0	25	25	0.000	Antitype
2 = no	1 = yes	NA	50	25	25	0.000	Type
2 = no	2 = no	A	50	25	25	0.000	Type
2 = no	2 = no	NA	0	25	25	0.000	Antitype

The corresponding Phi coefficient is zero.

$$\begin{aligned}\Phi &= \frac{(ad-bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \\ &= \frac{(50\times 50-50\times 50)}{\sqrt{(50+50)(50+50)(50+50)(50+50)}} \\ &= \frac{0}{100} \\ &= 0\end{aligned} \tag{1.14}$$

If one had looked only at the bivariate correlation or association, one may conclude that the relationship between the variables is zero. However, the data has a clear structure. Persons who affirm or negate both items belong to the group of Alcoholics. Persons who either affirm or negate one item belong to the group of Non-Alcoholics. The multivariate association can be only identified within a multinomial approach which is applied by CFA. Let's have a look at the results of the CFA to see whether CFA goes beyond bivariate associations (see Table 1.11):

CFA reveals local associations between the two items and group membership (cf. von Eye, 1990). Therefore, group membership can be inferred based on the response for the two items. In addition, CFA shows that each configuration is reflected either by a *type* or an *antitype* which represents a unique configuration or pattern of states of the three variables. All cases are represented by *types*, and the frequencies for *antitypes* are zero throughout. That is, CFA can differentiate perfectly between the group of alcoholics and non-alcoholics.

References

Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology, 9*, 291–319.

Bergman, L. R., von Eye, A., & Magnusson, D. (2006). Person-oriented research strategies in developmental psychopathology. In D. Cicchetti & D. J. Cohen (Eds.), *Developmental psychopathology* (2nd ed., pp. 850–888). London: Wiley.

Fienberg, S. E. (1987). The analysis of cross-classified categorical data. 5th printing.

Krauth, J., & Lienert, G. A. (1973). *Die Konfigurationsfrequenzanalyse und ihre Anwendung in Psychologie und Medizin [Configural frequency analysis and its application in psychology and medicine].* Freiburg, Germany: Alber.

Lautsch, E., & von Weber, S. (1995). *Methoden und Anwendung der Konfigurationsfrequenzanalyse (KFA) [Methods and application of configural frequency analysis (CFA)].* Weinheim, Germany: Beltz, Psychologie-Verlags-Union.

Lehmacher, W. (2000). Die Konfigurationsfrequenzanalyse als Komplement des log-linearen Modells [Configural frequency analysis as a complimentary tool to log-linear modeling]. *Psychology Science, 42*(3), 418–427.

Lienert, G. A., & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. *Educational Psychology and Measurement, 35*, 231–238.

Meehl, P. E. (1950). Configural scoring. *Journal of Consulting Psychology, 14*, 165–171.

Melcher, A., Lautsch E. & Schmutzler, S. (2012). Non-parametric methods – Tree and P-CFA – For the ecological evaluation and assessment of suitable aquatic habitats: A contribution to fish psychology. *Psychological Tests and Assessment Modeling, 54*(3), 293–306.

Reinecke, J. & Tarnai, C. (Eds.). (2008). *Klassifikationsanalysen in Theorie und Praxis [Analysis of classifications in theory and practice].* Münster, Germany: Waxmann Verlag.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Stemmler, M., Lautsch, E., & Martinke, D. (Eds.). (2008). *Configural frequency analysis and other non-parametrical methods: A Gustav A. Lienert memorial issue*. Lengerich, Germany: Pabst Publishing.

Stemmler, M., & von Eye, A. (Eds.) (2012). Configural frequency analysis (CFA) and other non-parametrical statistical methods (special issue) – Part I and II. *Psychological Tests and Assessment Modeling, 54*(2 and 3).

Victor, N. (1989). An alternative approach to configural frequency analysis. *Methodika, 3*, 61–73.

von Eye, A. (1990). Introduction to configural frequency analysis: The search for types and 102 antitypes in cross-classifications. Cambridge, UK: Cambridge University Press.

von Eye, A. (2002). *Configural frequency analysis: Methods, models and applications*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A., & Gutiérrez-Penã, E. (2004). Configural frequency analysis: The search for extreme cells. *Journal of Applied Statistics, 31*, 981–997.

Wickens, T. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Chapter 2
CFA Software

Abstract This chapter describes the CFA software that is freely available for use. One is a freeware written by Alexander von Eye (Michigan State University). The other is a R-package called ***confreq*** written by Jörg-Hendrik Heine (LMU Munich). The use of both software packages is described and demonstrated with data examples. Throughout the book both packages built the bases for demonstrating the use of Configural Frequency Analysis (CFA).

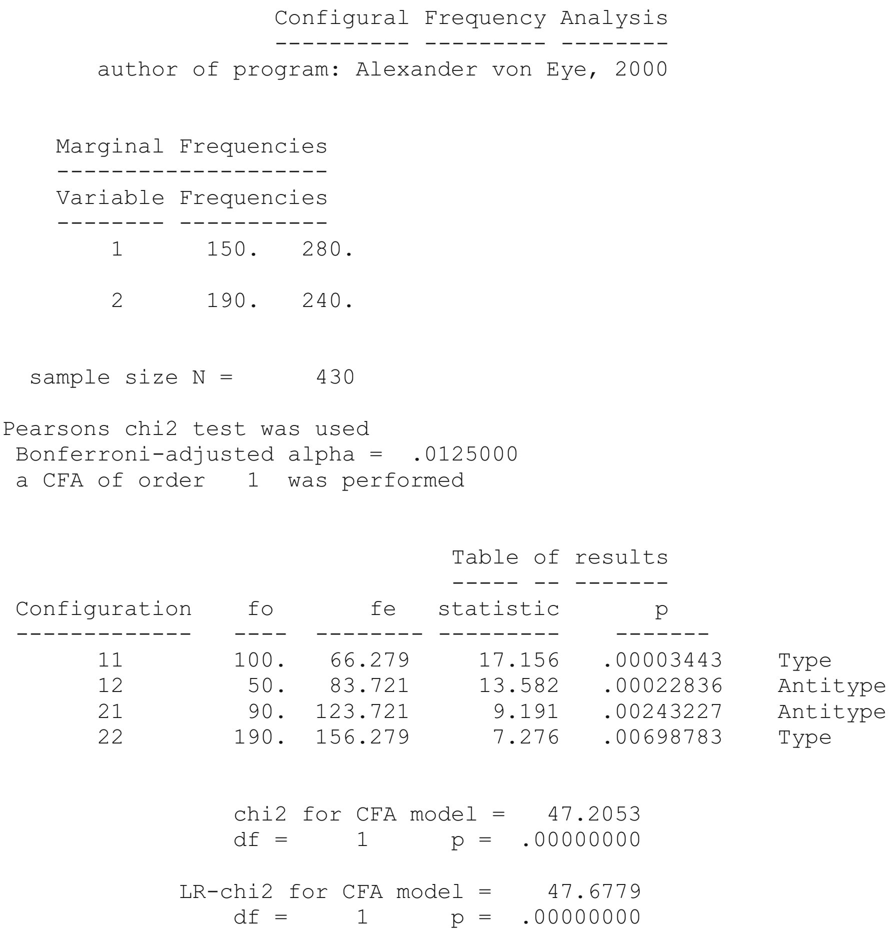
2.1 The Freeware by Alexander von Eye

Alexander von Eye (Michigan State University) has written a CFA program which is available as a freeware and which can be downloaded from his website (http://www.msu.edu/~voneye) (von Eye, 2002). This program, which from now on I will be calling *von Eye program* runs on the DOS level and is therefore is only suitable for Windows PCs. It was so far tested 'under the 32-bit Windows operating system XP Professional, Vista Business and Windows 7 Professional' (p. 248; von Eye, Mair, & Mun, 2010).

The program starts by double-clicking on the file *cfa.exe*. The *von Eye program* is controlled by typing numbers into the program. After it starts, the program asks whether the data will be *entered via a file* $<= 1 >$ *or interactive* $<= 2 >$. We will type the observed frequencies interactively (i.e., "2") from Table 1.1. Next, the *von Eye program* will ask the number of variables $< max = 10 >$. We type "2". Now the program wants the number of categories per variable. The program then calculates the number of cells "4".

Next the program asks for the observed frequency for configuration '11'. What does that mean? The typing of configurations into the computer is guided by simple rules, which are important but easy to learn. Each configuration has two or more indexes (e.g., o_{ijk}). The configurations are numbered starting with 1. Let's say we have three variables. The first variable has two categories, the second has three and the third has two categories. The resulting table will be a 2 by 3 by 2 contingency

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_2,



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


        Marginal Frequencies
        --------------------
        Variable Frequencies
        -------- -----------
            1      150.   280.

            2      190.   240.


      sample size N =      430

  Pearsons chi2 test was used
   Bonferroni-adjusted alpha =  .0125000
   a CFA of order   1  was performed


                                 Table of results
                                 ----- -- -------
   Configuration    fo       fe   statistic       p
   -------------   ----  --------- ---------    -------
         11        100.   66.279     17.156   .00003443    Type
         12         50.   83.721     13.582   .00022836    Antitype
         21         90.  123.721      9.191   .00243227    Antitype
         22        190.  156.279      7.276   .00698783    Type


                    chi2 for CFA model =   47.2053
                    df =     1      p =  .00000000

                LR-chi2 for CFA model =   47.6779
                    df =     1      p =  .00000000
```

Fig. 2.1 Printout of CFA program for data in Table 1 by Alexander von Eye

table with 12 cells. The first configuration will be '111'. The last variable indexed with k is the first to switch. That is, '112'. After all categories of the third variable have been altered, the second variable is altered next: '121', then '122', '131', '132'. Finally, after six configurations, the first variable is altered; that is, '211', '212', '221', '222', '231', and '232'.

We feed in the observed frequencies for each configuration or cell. The total sample size is N = 430. Typing $< yes = 1 >$ will save the data. Then we run a first order CFA $<= 1 >$ with the significance test based on the Pearson's Chi-Square $<= 4 >$ at the 5 % level. Finally we save the file, and for now, we don't want to print the design matrix. Let's look at the print out of the von Eye program (see Fig. 2.1).

By comparing the results of the printout of the von Eye program in Fig. 2.1, one can see that we obtained the same results as with our pocket calculator.

2.2 CFA R-Package

R is an open source software which is suitable for Linux, MacOS X, and Windows. R (R Development Core Team, 2004, 2011) is a program for data analysis, data manipulation and graphical display. The philosophy of R is quite different from other mainstream statistical packages such as SPSS or SAS. For statistical analysis, R requires a syntax which works step by step, while storing intermediate results into objects. In R, the objects can be modified for personal use and plotted easily. R is becoming more and more popular in the field of methodology, partially because R is also easy to connect with LaTeX (see Lamport, 1994). The software can be downloaded for free from the website of **The Comprehensive R Archive Network (CRAN)**
http://cran.r-project.org
Further information or help on R is available at the following website
http://www.r-project.org
at the link *Documentation*. Downloading an extra editor is recommended. *WinEdt* is a suitable editor for Windows (http://www.winedt.com). Another frequently used editor is *Tinn-R* (http://sourceforge.net/projects/tinn-r/). My personal recommendation is to use *R-Studio* (http://www.rstudio.com/). R-Studio is easy to use and very suitable for beginners. It works with four windows. In one windows you may load packages directly from the CRAN server. By simply checking the box of the respective package, they will be installed immediately. There is an extra window for your *R script*, your *workspace*, and the output is listed in the window *console*. If you work only with R, one needs to specify a CRAN server in order to download the software and further packages through the menu item *Documentation, Packages*. Choose a server in your area. In order to run CFA, please download the package ***confreq*** from CRAN or via R-Studio. The name is derived from **con**figural **freq**uency analysis (CFA). This R-package was written by Jörg-Hendrik Heine (University of Munich (LMU)). The reference manual for ***confreq*** can be downloaded from CRAN website (http://cran.r-project.org/web/packages/cfa/index.html).[1]

[1] By the way, I do not recommend the use of the other available R-package called *CFA*. This package includes mistakes, and which will not be further developed by their authors (personal communication with Stefan Funke). For instance, the corresponding z-values are not correctly analyzed (see the book by von Eye, Mair, and Mun (2010) on p. 268 and compare the obtained values with Table 10.6 on p. 185).

Let’s have a look at the corresponding R code of the R-package ***confreq***.

```
> # line starting with the hash key are
> # considered comments no commands!
> # it is recommended to save your R-script
> # and to document it with many comments;
> # this helps when re-using the R-script
> rm(list=ls())
> # clears the workspace in R Studio
> # enter the patterned frequency of
> # Table 1 as a matrix
> # first two columns are the patterns,
> # the third column lists the frequencies
> x1<-c(1,1,100)
> x2<-c(1,2,50)
> x3<-c(2,1,90)
> x4<-c(2,2,190)
> # the four vectors are combined to a matrix
> table1<-rbind(x1,x2,x3,x4)
> table1
> library(confreq)
> # library(confreq) loads the R-package ‘confreq’
> # brings the data matrix into a
> # pattern matrix which can be analyzed
> table1_new <- dat2fre(fre2dat(table1))
> table1_new
> #finally the CFA command for the main effects model
> CFA(table1_new,form="~ A + B")
```

The R syntax works line by line (cf. Alexandrowicz, 2013). Command lines start with “>” and comments are introduced with the hash key. As mentioned before, R works with objects. The configurations are determined by the matrix called “table1” based on the data of Table 1.1. In a 2 by 2 table we are having two main effects, one for each variable (i.e., main effect A and B). Next to typing in configurations and frequencies as a row which will be composed to a matrix, ***confreq*** also reads in EXCEL files, which makes the program even more flexible. The indexed configuration can be easily typed into EXCEL, but it is important to use the file type ‘.CSV’. In Europe, one has to use the file type ‘.CVS2’ which indicates that the separator between digits is a semicolon, instead of a comma, which is used in North America. To give a short example

Table 2.1 R-Syntax for the data of Table 1

Pattern	Observed	Expected	loc.chi.square	loc.df	loc.chi.square.p
1 1	100	66.28	17.16	1.00	0.00
1 2	50	83.72	13.58	1.00	0.00
2 1	90	123.72	9.19	1.00	0.00
2 2	190	156.28	7.28	1.00	0.01

```
> # reading in data from EXCEL into confreq
> contingency_table<-read.csv(file=paste(pfad,"/" +
 ,"table.csv",sep=""),header = FALSE)
> # in Europe one have to use 'read.csv2'
> # but the rest is identical
```

The command for running a Configural Frequency Analysis is “CFA”. The sub-command “form” specifies the main effects and/or interactions. Here, only the main effects A and B were specified. Table 2.1 is a slightly condensed version of the actual table in order to make it comparable to the values obtained through the von Eye program.

CFA uses a number of statistics to test the significance of configurations. So far, only the Pearson chi-square was introduced. However, the *von Eye program* and the CFA R package provide the information on many different test statistics, which will be presented in the next Chap. 3.

Summary: There are two software packages available for CFA: the freeware provided by Alexander von Eye and the R-package ***confreq***. The freeware by von Eye was written in FORTRAN 90 and is only suitable for Windows. The R package also runs on Linux and MAC OS X. The R package is an open source software which is improved and extended constantly. R is also becoming more and more popular in university methods classes.

References

Alexandrowicz, R. (2013). *R in 10 Schritten: Einführung in die statistische Programmierumgebung [R in 10 steps: Introduction to the statistical computing environment]*. Wien: facultas.WUV.

Lamport, L. (1994). *LaTeX: A document preparation system: User’s guide and reference manual*. Reading, MA: Addison-Wesley.

R Development Core Team. (2004). *Manual of the R foundation for statistical computing. R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

R Development Core Team. (2011). *A language and environment for statistical computing [Computer software manual]*. Vienna: R Foundation for Statistical Computing. Available from http://www.R-project.org.

von Eye, A. (2002). *Configural frequency analysis: Methods, models and applications*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A., Mair, P., & Mun, E.-Y. (2010). *Advances in configural frequency analysis*. New York: The Guilford Press.

Chapter 3
Significance Testing in CFA

Abstract This chapter explains six different significance tests which are available in the von Eye program and/or in the R-package ***confreq*** in the search of *types* and *antitypes*. Among them are the binomial and the chi-square test including their normal approximations. The formulas for each test are provided. Advantages and disadvantage of each test are explained.

3.1 The Binomial Test

In von Eye's CFA program this test can be activated under the program section of *the following significance tests are available*. Its pertinent number is "*1*." In R-package ***confreq*** the binomial test is default. In R, each statistic is listed twofold in the printout. First, the value of the statistic is provided and then the p-value is listed. The binomial test can have one of two possible outcomes: (1) a certain configuration will be observed or (2) a certain configuration will not be observed. The underlying idea is that the occurrence of either outcome is based on a Bernoulli process of independent trials (think of flipping a coin). The binomial test estimates a point probability of observing a configuration *o* under a certain expected frequency *e*:

$$B(o) = \binom{n}{o} p^o q^{(n-0)} \tag{3.1}$$

with n = sample size and o = observed frequencies.

$$p = \frac{e}{n} \quad and \quad q = 1 - p.$$

with e = expected frequencies.

Given Meehl's example:

$$p = \frac{25}{200} = 0.125 \quad and \quad q = 1 - 0.125 = 0.875$$

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_3,

the point probability for a certain configuration is

$$B(o_{yesyesA}) = \binom{200}{50} 0.125^{50}\, 0.875^{150} =$$
$$= 0.0085$$

To test for *types* $H_1 : o > e$ the joint probability for o together with o' representing all more extreme observed frequencies for $o' > o$ is compared with a certain significance level (e.g., let's say $\alpha = 5\,\%$). If $o > e$ and o' is more extreme than o, then the corresponding binomial formula for $B_1(o)$ is

$$B_1(o) = \sum_{i=a}^{l} \binom{n}{i} p^i q^{(n-i)} \tag{3.2}$$

where a = o and l = n.

To test for *antitypes* $H_1 : o < e$ and if $o' < o$ and o' is more extreme than o, then the corresponding binomial formula for $B_1(o)$ is

$$B_1(o) = \sum_{i=a}^{l} \binom{n}{i} p^i q^{(n-i)}$$

where a = 0 and l = o.

3.2 Approximation of the Binomial Test Using Stirling's Formula

In von Eye's CFA-program this test can be activated with the number "2". The binomial test according to Stirling is basically a reformulation of the binomial formula:

$$B_1(o) = \binom{n}{o} p^o q^{(n-o)} =$$
$$= \frac{n!}{(n-o)!o!}\, p^o q^{(n-o)}.$$

With n = sample size and o = observed frequencies. The expected frequencies can be calculated through $e = n \times p$ and the proportion p for obtaining a certain configuration is $p = \frac{e}{n}$. By using the formulas for e and p one obtains Stirling's formula:

$$\hat{B}'(o) = \left(\frac{n}{2\pi o(n-o)}\right)^{\frac{1}{2}} \left(\frac{np}{o}\right)^{o} \left(\frac{nq}{n-o}\right)^{(n-o)}. \tag{3.3}$$

In comparison to the traditional binomial testing, Stirling's approximation

- Has generally slightly less power than the exact binomial test;
- Is closest to the exact binomial test, if the difference between the observed and expected frequencies is small.

The binomial test according to Stirling is also available in the R-package ***confreq***, however, the formula is not appropriate for zero observed frequencies, because it is not possible to divide by zero. The output in ***confreq*** in this case is "INF" for infinity!

3.3 Chi-Square Test

In von Eye's CFA-program this test can be activated with number "*3*." In R, this test is the default test, and doesn't need to be activated. The **local chi-square statistic** is calculated as follows

$$\chi^2 = \frac{(\mathrm{o}-\mathrm{e})^2}{\mathrm{e}} \quad \text{with df} = 1 \tag{3.4}$$

where $\mathrm{o} = $ observed frequencies and $\mathrm{e} = $ expected frequencies. The chi-square statistic is based on the assumption that $e > 5$.

3.4 Chi-Square Approximation to the z-Test

In von Eye's CFA-program this test can be activated under the program section of *the following significance tests are available* and its pertinent number is "*4*.". In the CFA R package this test is also a default test. In the output it is listed under "z.Chi" The chi-square approximation to the z-statistic is calculated as follows:

$$z^2_{(\frac{\alpha}{2})} = \chi^2_{(\alpha)} \tag{3.5}$$

$$z^2_{(\frac{\alpha}{2})} = \frac{(o-e)^2}{e} =$$

$$z^2_{(\frac{\alpha}{2})} = \frac{(o-np)^2}{np}$$

$$z = \frac{(o-e)}{\sqrt{np}}.$$

Note that based on Monte Carlo studies (cf. von Eye, 1990), the use of this test statistic is recommended by Alexander von Eye.

3.5 Binomial Approximation to the z-Test

In the von Eye program this test can be activated with the number "*5*".

If $e = n \times p \geq 10$, the binomial approximation is sufficiently accurate. In this case the z-statistic is calculated as follows:

$$z = \frac{(o - np)}{\sqrt{npq}}. \tag{3.6}$$

If $5 \leq n \times p \leq 10$, the binomial approximation needs to apply a so-called continuity correction. The continuity-corrected z-statistic is calculated as follows:

$$z = \frac{(o - np - 0.5)}{\sqrt{npq}}. \tag{3.7}$$

In the R-package *confreq* the continuity correction is activated through the command "ccor=TRUE" (ccor=FALSE is default), here is an example:

```
>CFA(patternfreq,form="~ A + B", ccor=TRUE)
```

3.6 Lehmacher's Asymptotic Test

This test is not available in *confreq*. In the von Eye program this test can be activated with the number "*6*". The Lehmacher's approximation to the z-statistic is calculated as follows:

$$\begin{aligned} z_L &= \frac{(o - np)}{\sigma} \\ z_L &= \frac{(o - e)}{\sigma}, \end{aligned} \tag{3.8}$$

where σ^2 is the exact variance of a hypergeometrical distribution under the assumption that the marginals are fixed. The exact variance is calculated as

$$\sigma^2 = np(1 - p - (n - p)(p - \tilde{p})). \tag{3.9}$$

p is estimated from the marginal under the notion of independence of all variables (here $d = 4$ variables (dimensions)):

$$p = n_{i...} \times n_{.j..} \times \ldots \times n_{...k} / n^d. \tag{3.10}$$

The probability $\tilde{p}$ is estimated by

$$\tilde{p} = (n_{i...} - 1)(n_{.j..} - 1)\dots(n_{...k} - 1)/n^d. \tag{3.11}$$

Lehmacher's test statistic always leads to the largest values (i.e., it is a progressive test). It is true that $|\chi^2| < |z| < |z_L|$. This test statistic is not available in the R-package.

3.7 Küchenhoff's Continuity Correction of Lehmacher's Asymptotic Test

This test is not available in ***confreq***. In the von Eye program this test can be used with number "*7*." In order to avoid overly-liberal test results while using Lehmacher's asymptotic test, Küchenhoff suggested using Lehmacher's test together with the Yates' continuity correction . Here

$$o' = \begin{cases} o - 0.5 & \text{if} \quad o > e \\ o + 0.5 & \text{if} \quad o \le e \end{cases} \tag{3.12}$$

with each local or cell-wise significant test we are conducting multiple test. Therefore, we are facing the danger of alpha inflation. This test statistic is also not available in the R-package.

In order to protect our results from false inferences, both software package have implemented, as a default, a Bonferroni correction (see Chap. 1).

Summary: Although the z-statistic can be recommended in most cases, there is no test that is preferable over all tests. The tests differ in terms of statistical power and the accuracy of the approximation of the sampling distribution. The only exact test is the binomial test. The χ^2-test should not be used because it detects more *types* and fewer *antitypes* (cf. von Eye, 2002).

References

von Eye, A. (1990). *Introduction to configural frequency analysis: The search for types and antitypes in cross-classifications*. Cambridge, UK: Cambridge University Press.

von Eye, A. (2002). *Configural frequency analysis: Methods, models and applications*. Mahwah, NJ: Lawrence Erlbaum.

Chapter 4
CFA and Log-Linear Modeling

Abstract This chapter describes the relationship between log-linear modeling and CFA. Log-linear modeling and CFA may be used as complimentary statistical tools. Log-linear modeling looks for models with an appropriate goodness-of-fit; they can be used to investigate the patterns of association or the structure of dependency among the variables. CFA needs a non-fitting model in order to detect *types* and/or *antitypes*. In CFA and log-linear models, the expected frequencies are calculated according to the underlying null model which is specified in the design matrix using the General Linear Model approach (GLM). Following log-linear modeling hierarchical log-linear modeling is presented. Hi-log models are the best way to determine the structure of dependency among the variables or to find out which interactions are significant. Hi-log modeling is a special form of log-linear modeling. The main effects and interactions are structured hierarchically such that if there are significant higher order interactions in the model, all lower order interactions and main effects must be included. In addition to describing the traditional first-order CFA, a zero-order CFA called Configural Cluster Analysis (CCA) is explained. Finally, the statistic Q describing the pregnancy or precision of a cell is introduced. Small data examples are presented and analyzed with the *von Eye program* as well as with the R-package ***confreq***.

4.1 Log-Linear Modeling: Looking at the Underlying Dependencies

This section describes the relationship between log-linear modeling and CFA. Log-linear modeling is a common statistical tool for the analysis of contingency tables (cf. Langeheine, 1980; von Eye, 1990, 2002). Log-linear models can be used to investigate the patterns of association or the structure of dependency among the variables. They parametrize cell frequencies, or to put it differently, the logarithms of cell frequencies, in terms of main effects and interactions. In CFA and log-linear

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_4,

Table 4.1 Observed frequencies for children's intelligence status by seizure status

		Intelligence		
		Above average	Below average	
Seizures	Present	6	8	14
	Absent	37	5	44
		43	13	$N = 56$

modeling, the expected frequencies are calculated by using the General Linear Model (GLM Kutner, Neter, Nachtsheim, & Li, 2004) approach. The GLM is

$$Y = X\beta + e. \tag{4.1}$$

The expected frequencies are estimated as

$$\hat{Y} = X\beta = Y - e. \tag{4.2}$$

$\hat{Y}$ is a one column vector including the expected frequencies. The number of cells or configurations determine the number of rows of $\hat{Y}$. X is the **design matrix** containing the effect-coded main effect and interaction terms plus the constant. The design matrix X has as many rows as there are cells or configurations, and $m+1$ columns. m is the number of weights; the first weight is always the *constant*, coded with ones. β comprises the weights of the independent variables and is a one-row vector with as many columns as cells. Let's look at the following example. The data were taken from a study at the Erlangen-Nuremberg University Hospital for Children. Newborns with (1) or without seizures (2) were tested with an intelligence test while they attended kindergarten. Children's intelligence was divided into (1) "average or above" and (2) "below average". The following cross-table was developed (see Table 4.1). The aim is to describe the underlying structure (e.g., main effects, interactions) in order to reproduce the observed data. Let's start with a model which assumes independence for the underlying variables (i.e., between variables A and B). A model without any interactions is called a **main effects model**. For independence the following statement must hold:

$$\frac{o_{11}}{o_{.1}} = \frac{o_{12}}{o_{.2}} \tag{4.3}$$

In other words, the proportion of frequencies of variable A_1 in variable B_1 is the same as the proportion of frequencies of variable A_1 in variable B_2. Variable A_1 is distributed equally across variable B. The same holds for variable B across variable A. In the same vein one can state that:

$$\frac{o_{11}}{o_{.1}} = \frac{o_{1.}}{o_{..}} \tag{4.4}$$

Table 4.2 Expected frequencies for children's intelligence status by seizure status

		Intelligence		
		Above average	Below average	
Seizures	Present	10.8	3.3	14
	Absent	32.3	9.8	44
		43	13	$N = 56$

By reformulating equation 4.4 one obtains

$$o_{11} = \frac{o_{1.} \times o_{.1}}{o_{..}}. \tag{4.5}$$

Equation 4.5 describes how the observed frequencies should be distributed under the assumption of independence. Basically it is an estimation and a statement for expected frequencies.

$$e_{ij} = \frac{o_{1.} \times o_{.1}}{o_{..}} \tag{4.6}$$

Therefore the expected frequencies for Table 4.2 are as follows: The expected frequencies have two characteristics

- The marginal values for the expected frequencies of variable A and B are equal to the marginal values of the observed frequencies, and
- Therefore the total sample size is reproduced and it holds $\sum_{i=1}^{I}\sum_{j=1}^{J} e_{ij} = o_{..} = N$

Note: If one compares the above observed with the expected values, one may detect large differences between the observed and expected frequencies. Therefore, our underlying main effects model may not hold. What does the corresponding design matrix for the above data look like? The main effects and interaction terms are **effect coded**; that is, we use coefficients c_i for each category of a variable, which have to sum to zero.

$$\sum c_i = 0 \tag{4.7}$$

Let's assume we have two categories or levels for variable A. Then we have to chose different coefficients for each category, (e.g., 1 and -1). Note that the sum of the coefficients need to be zero. The next variable B also has two levels, but needs to be differently coded than variable A (e.g., -1 and 1). The **design matrix** X for Table 4.1 is as follows:

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \tag{4.8}$$

The assumption that the observed frequencies are reproduced from the marginal values means that variables A and B have a certain effect on e_{ij}. The most interesting thing to note is how big these effects are. This can be expressed through a multiplicative model

$$e_{ij} = \gamma_0 \times \gamma_i A_i \times \gamma_j B_j \times \gamma_{ij} A_i B_j \tag{4.9}$$

which can be transformed into an additive relationship via the natural logarithm

$$ln\, e_{ij} = \lambda_0 + \lambda_i A_i + \lambda_j B_j + \lambda_{ij} A_i B_j. \tag{4.10}$$

ln = the natural logarithm for base e (i.e., e = Euler's number = 2.71828...). Equation 4.10 represents the *log-linear model*, which estimates the e_{ij} as a linear combination consisting of a constant, two main effects and an interaction effect. The λs are parameters, which explain the effects of the variables or variable interactions on the expected frequencies. They cannot be observed from the data but may be estimated. First, the data needs to be read in SPSS. For this purpose we use a SPSS Syntax which reads in frequency data (see Box with Crosstabs Syntax).

```
Data List free
 /seizures intelligence freq.
  weight by freq.
begin data.
1 1  6
1 2  8
2 1 37
2 2  5
end data.

Value Lables
  seizures    1 'present' 2 'absent'
/intelligence 1 'above average'
              2 'below average'.
```

Because the logarithm of zero is not defined, the default SPSS (SPSS IBM Inc., 2010) Delta Option adds 0.5 to each cell, in case there are cells with no frequencies (here the Delta Option is set to 0.01). The following SPSS Syntax was used (see Box with Loglinear Syntax).

```
LOGLINEAR seizures(1,2) intelligence(1,2)
          /CRITERIA = DELTA(0.01)
          /PRINT = DEFAULT
          /DESIGN = seizures  intelligence.
```

In the following the above main effects model and their respective λ parameters can be obtained with SPSS (see Fig. 4.1). Next to each variable name the minimum and maximum values are listed in parentheses. The parameters are estimated such that their sum equals zero; that is $\lambda_2 A_2 = -\lambda_1 A_1$ and $\lambda_2 B_2 = -\lambda_1 B_1$. The parameters for variables A and B are significant (see their respective z-values). A significant main effect reveals that the marginal values are not equally distributed, which is obviously the case for both variables (see Table 4.1). The parameter for the interaction term A by B is not listed because it was set to zero. By having a look at the model fit, one can determine, if the observed data or their underlying relationships can be reproduced only through the main effects. There are two statistics available: (1) the Pearson Chi-square and (2) the Likelihood Ratio Chi-square (LR).

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{4.11}$$

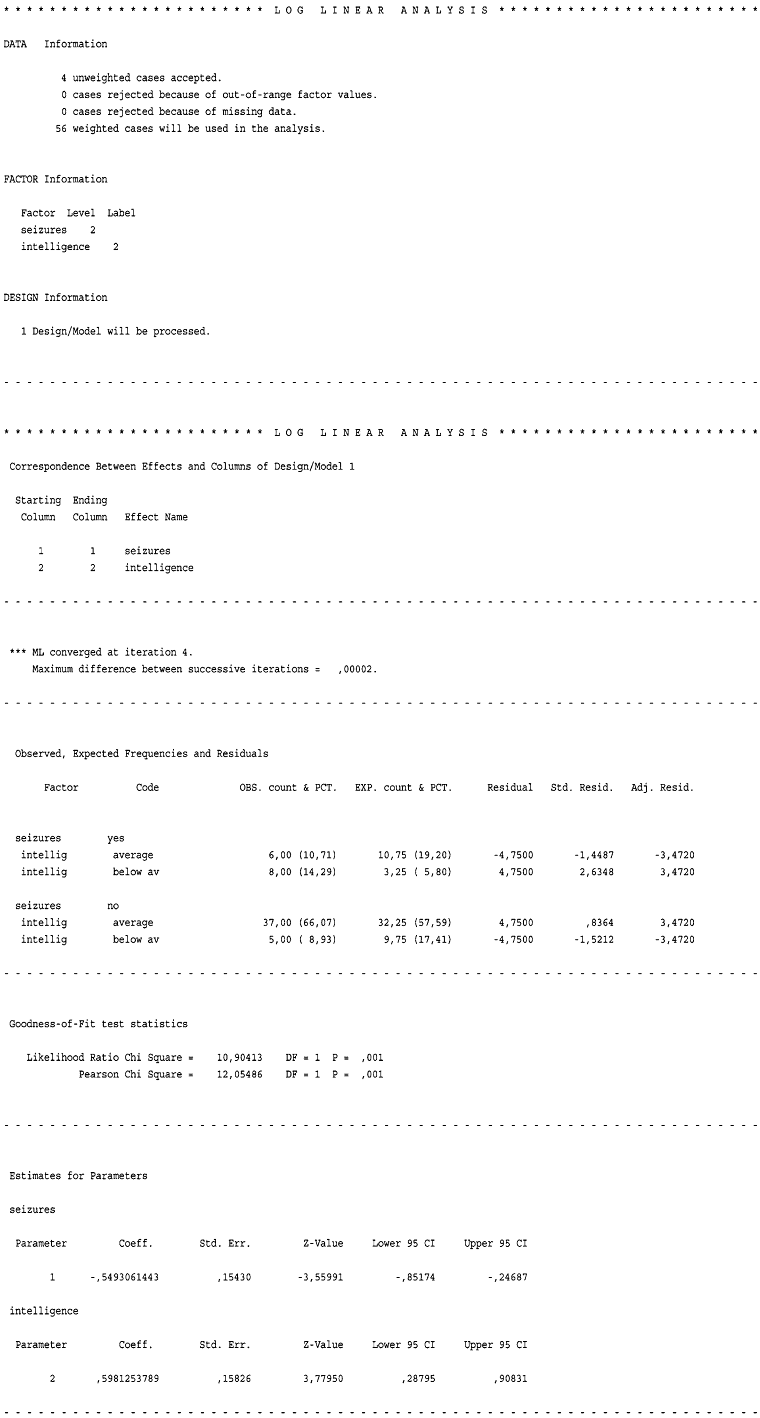
and

$$\mathrm{LR} = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} o_{ij}\, ln \frac{o_{ij}}{e_{ij}} \tag{4.12}$$

A model fits if the respective value of χ^2 or LR is larger than $\alpha = 5\,\%$. Here, both fit statistics indicate a non-fit (LR = 10.90413, df = 1, $p < 0.05$; $\chi^2 = 12.05486$, df = 1, $p < 0.05$; see Fig. 4.1).

Let's have a look at the corresponding R code of the R-package ***confreq***.

```
rm(list=ls())
# clears the workspace in R Studio
# enter the patterned frequency as a matrix
# first two columns are the patterns,
# the third column lists the frequencies
x1<-c(1,1,6)
x2<-c(1,2,8)
x3<-c(2,1,37)
x4<-c(2,2,5)
# the four vectors are combined to a matrix
mat1<-rbind(x1,x2,x3,x4)
mat1
library(confreq)
# loads the R-package 'confreq'
# brings the data matrix into a
# pattern matrix which can be analyzed
mat1_new <- dat2fre(fre2dat(mat1))
mat1_new
#finally the CFA command for the main effects model
CFA(mat1_new,form="~ A + B")
```



```
* * * * * * * * * * * * * * * * * * * * * * L O G   L I N E A R   A N A L Y S I S * * * * * * * * * * * * * * * * * * * * * *

DATA   Information

         4 unweighted cases accepted.
         0 cases rejected because of out-of-range factor values.
         0 cases rejected because of missing data.
        56 weighted cases will be used in the analysis.


FACTOR Information

   Factor  Level  Label
   seizures    2
   intelligence    2


DESIGN Information

   1 Design/Model will be processed.


- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


* * * * * * * * * * * * * * * * * * * * * * L O G   L I N E A R   A N A L Y S I S * * * * * * * * * * * * * * * * * * * * * *

 Correspondence Between Effects and Columns of Design/Model 1

  Starting  Ending
   Column   Column   Effect Name

      1        1     seizures
      2        2     intelligence

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


 *** ML converged at iteration 4.
     Maximum difference between successive iterations =   ,00002.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


  Observed, Expected Frequencies and Residuals

       Factor         Code            OBS. count & PCT.   EXP. count & PCT.      Residual   Std. Resid.   Adj. Resid.


  seizures        yes
   intellig        average                 6,00 (10,71)       10,75 (19,20)       -4,7500       -1,4487       -3,4720
   intellig        below av                8,00 (14,29)        3,25 ( 5,80)        4,7500        2,6348        3,4720

  seizures        no
   intellig        average                37,00 (66,07)       32,25 (57,59)        4,7500         ,8364        3,4720
   intellig        below av                5,00 ( 8,93)        9,75 (17,41)       -4,7500       -1,5212       -3,4720

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


 Goodness-of-Fit test statistics

    Likelihood Ratio Chi Square =    10,90413    DF = 1  P =  ,001
             Pearson Chi Square =    12,05486    DF = 1  P =  ,001


- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


 Estimates for Parameters

 seizures

  Parameter        Coeff.        Std. Err.        Z-Value     Lower 95 CI     Upper 95 CI

        1      -,5493061443          ,15430       -3,55991         -,85174         -,24687

 intelligence

  Parameter        Coeff.        Std. Err.        Z-Value     Lower 95 CI     Upper 95 CI

        2       ,5981253789          ,15826        3,77950          ,28795          ,90831

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Fig. 4.1 Estimated parameters for a main effect model

The (somewhat simplified) printout of the R script looks as follows:

```
$local.test
  pattern observed expected
1      1 1        6    10.75
2      1 2        8     3.25
3      2 1       37    32.25
4      2 2        5     9.75

       z.Chi     p.z.Chi
1 -1.4487364 0.073705603
2  2.6348259 0.004209022
3  0.8364284 0.201456982
4 -1.5212175 0.064102638

$bonferroni.alpha
[1] 0.0125

$global.test
$global.test$chi.square
[1] 12.05486

$global.test$df
df
 1

$global.test$chi.square.p
[1] 0.0005165753

$global.test$alpha
[1] 0.05
```

The R script results in the same global chi-square value: $\chi^2 = 12.05486$, df $= 1$, $p < 0.0005165753$. The output gives also the Bonferroni adjusted alpha of $p = 0.0125$. The alpha level may be adjusted, for instance to a two-sided test, through

```
>CFA(patternfreq,alpha=0.025,form="~ A + B")
```

For all programs, the two fit statistics bear the same result. There is one *type* with the configuration (1,2) which stands for *seizures*=yes and *intelligence*=below average. One can conclude that in this study it was *typical* for newborns with seizures to be below average in intelligence at kindergarten age. The **standardized residuals** in the printout of the SPSS log-linear model (abbreviated as Std. Resid.; see Fig. 4.1) are identical to the option "4" in the CFA program by von Eye called *chi-square approximation to the z-statistic* (see Fig. 4.2) and to the z.Chi-statistic of the R-package ***confreq***. In addition, the *von Eye program* prints off the design matrix

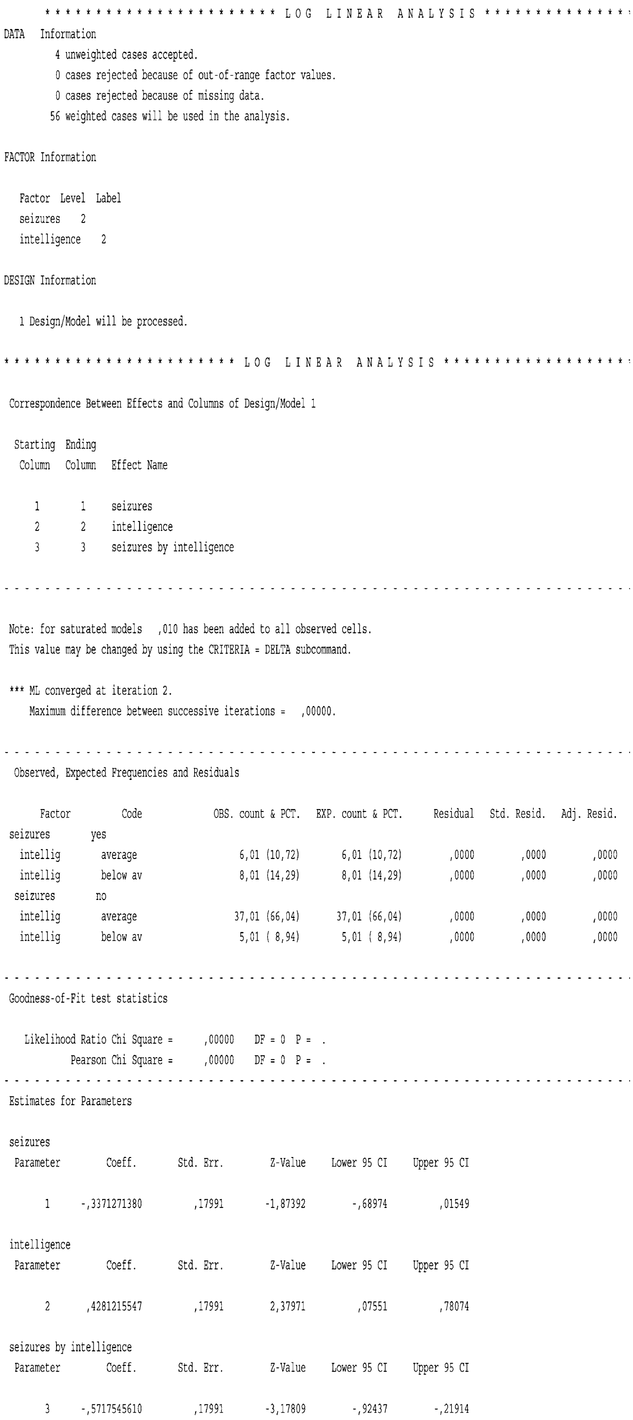
```
                   Configural Frequency Analysis
                   ---------- --------- --------
      author of program: Alexander von Eye, 2000


   Marginal Frequencies
   --------------------
   Variable Frequencies
   -------- -----------
       1       14.    42.

       2       43.    13.


 sample size N =       56

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
-------------   ----  -------- ---------    -------
      11          6.   10.750      -1.449   .07370565    Type
      12          8.    3.250       2.635   .00420906    Type
      21         37.   32.250        .836   .20145692
      22          5.    9.750      -1.521   .06410267


                chi2 for CFA model =   12.0549
                df =     1      p =  .00051658

            LR-chi2 for CFA model =    10.9041
                df =     1      p =  .00095950
```

Fig. 4.2 CFA results of the main effect log-linear model

used which is equal to the one listed above (see Eq. 4.23). However, the constant is not printed because it is redundant. The *von Eye program* ends with an invigorating CARPE DIEM which is Latin and means SEIZE THE DAY (von Eye, Mair, and Mun, 2010).

In order to examine the importance of the interaction term, a **saturated model** is calculated. A saturated model reproduces the observed values perfectly. Although it is somewhat tautological, it is of great value. All parameters are estimated. Let's look at the design matrix of the saturated model:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \tag{4.13}$$



```
      * * * * * * * * * * * * * * * * * * * * * * L O G   L I N E A R   A N A L Y S I S * * * * * * * * * * * * * *
DATA   Information
        4 unweighted cases accepted.
        0 cases rejected because of out-of-range factor values.
        0 cases rejected because of missing data.
       56 weighted cases will be used in the analysis.

FACTOR Information

   Factor  Level  Label
   seizures    2
   intelligence    2

DESIGN Information

   1 Design/Model will be processed.

* * * * * * * * * * * * * * * * * * * * * * L O G   L I N E A R   A N A L Y S I S * * * * * * * * * * * * * * * * * *

 Correspondence Between Effects and Columns of Design/Model 1

  Starting  Ending
   Column   Column   Effect Name

      1       1     seizures
      2       2     intelligence
      3       3     seizures by intelligence

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

 Note: for saturated models   ,010 has been added to all observed cells.
 This value may be changed by using the CRITERIA = DELTA subcommand.

 *** ML converged at iteration 2.
     Maximum difference between successive iterations =   ,00000.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Observed, Expected Frequencies and Residuals

      Factor         Code            OBS. count & PCT.   EXP. count & PCT.     Residual  Std. Resid.  Adj. Resid.
 seizures        yes
   intellig        average                6,01 (10,72)        6,01 (10,72)        ,0000        ,0000        ,0000
   intellig        below av               8,01 (14,29)        8,01 (14,29)        ,0000        ,0000        ,0000
  seizures        no
   intellig        average               37,01 (66,04)       37,01 (66,04)        ,0000        ,0000        ,0000
   intellig        below av               5,01 ( 8,94)        5,01 ( 8,94)        ,0000        ,0000        ,0000

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Goodness-of-Fit test statistics

    Likelihood Ratio Chi Square =      ,00000    DF = 0  P =  .
             Pearson Chi Square =      ,00000    DF = 0  P =  .
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Estimates for Parameters

 seizures
  Parameter        Coeff.       Std. Err.        Z-Value    Lower 95 CI    Upper 95 CI

       1     -,3371271380         ,17991       -1,87392        -,68974         ,01549

 intelligence
  Parameter        Coeff.       Std. Err.        Z-Value    Lower 95 CI    Upper 95 CI

       2      ,4281215547         ,17991        2,37971         ,07551         ,78074

 seizures by intelligence
  Parameter        Coeff.       Std. Err.        Z-Value    Lower 95 CI    Upper 95 CI

       3     -,5717545610         ,17991       -3,17809        -,92437        -,21914
```

Fig. 4.3 Estimated parameters for a log-linear saturated model

Table 4.3 Results of the saturated model using the R-package confreq

Pattern	Observed	Expected	z.Chi	p.z.Chi
1 1	6	6.00	0.00	0.50
1 2	8	8.00	−0.00	0.50
2 1	37	37.00	−0.00	0.50
2 2	5	5.00	0.00	0.50

The interaction term is the result of the multiplication of the two main effects (see the number of columns is equal to the Lambdas in Eq. 4.10). The printout gives parameter estimates for each main effect and the interaction term (see Box with Loglinear Syntax).

```
LOGLINEAR seizures(1,2) intelligence(1,2)
          /CRITERIA = DELTA(0.01)
          /PRINT = DEFAULT
          /DESIGN = seizures   intelligence
          intelligence   BY   seizures.
```

One has to bear in mind that these are multiple test procedures; therefore a Bonferroni alpha adjustment is necessary (cf. von Eye, 1990), that is $\alpha^* = 0.025/3 = 0.00833$ and the corresponding z-value is $|2.40|$. By looking at the z-standardized lambda parameters, one can see that both main effects are not significant (for *seizures* ($\lambda_A = -1.87392$) and for *intelligence* ($\lambda_B = 2.37973$)). But the interaction term *seizures by intelligence* is highly significant ($\lambda_{AxB} = -3.17809$; see Fig. 4.3). In our case, a model that reproduces the observed frequencies satisfactorily needs to include the interaction term; no better or more parsimonious model is available. In addition one can see that the observed values are perfectly reproduced which is constitutive for the saturated model. Both fit statistics indicated a perfect fit with $p = 1.0$. The respective degree of freedom of the model is zero.

The null model may also be obtained through the R-package ***confreq*** by indicating the interaction in addition to the main effects in the CFA command:

```
CFA(patternfreq_neu,form="~ A + B + A:B")
```

Let us have a look at the obtained results (Table 4.3):

CFA can now be used as a complimentary statistical method. The prerequisite for CFA is a **non-fitting** log-linear model, because *types* and *antitypes* represent significant deviations from expected frequencies. The main effects model and the saturated model can be calculated using SPSS as well as the new R-package ***confreq***.

4.2 Hierarchical Log-Linear Modeling

Let us look at examples with three or more variables. The data is taken from Krauth and Lienert (1973). There are two groups of depressed subjects (‘−’ = recovered from minor depression due to treatment; ‘+’ = still suffering from depression while under treatment). Both groups were tested with a temperament scale that consists of three subscales: I = Introversion versus Extraversion (high scores represent Introversion; low scores Extraversion), R = Rigidity (high scores represent rigidity; low scores represent non-rigidity), and V = Vitality (high scores represent high vitality; low scores represent low vitality). ‘+’ represents high scores (above the median), and ‘−’ represents low scores (below the median).

If more than three variables are investigated in a log-linear model (e.g., using the data from Table 4.4), the best way to determine the structure of dependency among the variables or to find out which interactions are significant is by using the hierarchical log-linear approach (hi-log models). Hi-log modeling is a form of log-linear modeling. The main effects and interactions are structured hierarchically such that if there are significant higher order interactions in the model, all lower order interactions and main effects must be included. Non-hierarchical log-linear modeling or nonstandard log-linear modeling is also possible but is not discussed here (see Rindskopf (1990) for an explanation). The hi-log SPSS-syntax can be found in the next syntax box.

```
HILOGLINEAR   D(1,2) I(1,2) R(1,2) V(1,2)
              /METHOD = BACKWARD
              /CRITERIA = DELTA(0.01)
              /PRINT = DEFAULT, ESTIM
              /DESIGN.
```

Below, excerpts from the SPSS printout (*Effect k order and higher*) for the hi-log models are presented. In SPSS, k order effects can be explained as follows: *First order effects (k = 1)* are main effects, *second order effects (k = 2)* are

Table 4.4 Data taken from Krauth and Lienert (1973, p. 73)

D+				D+				D−				D−			
	I+				I−				I+				I−		
		V				V				V				V	
		+	−			+	−			+	−4			+	−4
R	+	15	30	R	+	23	22	R	+	25	22	R	+	14	8
	−	9	32		−	14	16		−	46	27		−	47	12

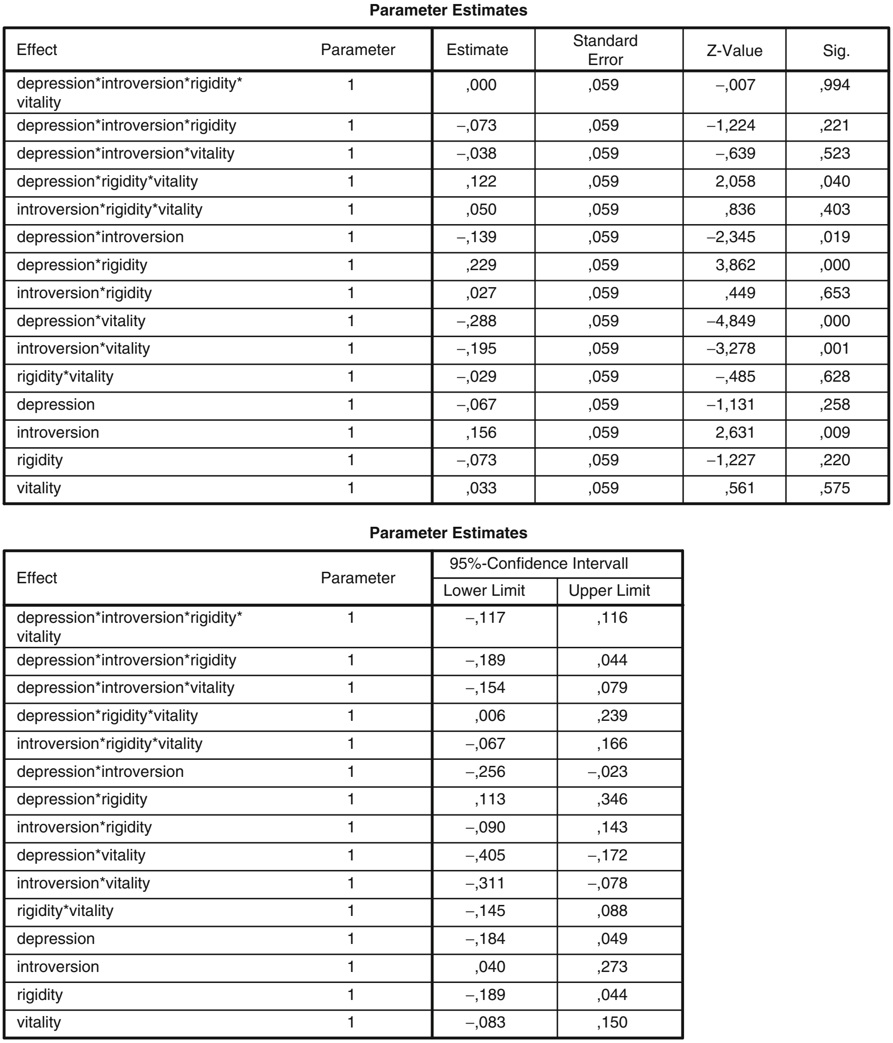
normal interactions between two variables, and *third and higher order effects* are interactions between three and more variables. The SPSS printout is very informative, because it looks at the kth order effects either separately or in sets. The **goodness-of-fit** of the hi-log-linear models for all the effects are listed. A significant chi-square or LR statistic indicates a bad fit, saying that the underlying model does not hold. The upper table of the printout displays the goodness-of-fit when the first order effects and higher are zero (i.e., k = 1 order effects and higher; all the main effects and interactions involved are zero). Of course this model does not fit (LR = 87.201 p = 0.000, and $\chi^2 = 91.569$ p = 0.000). How does the goodness-of-fit looks like if the second order effects and higher (i.e., k = 2 order effects and higher) are zero. Again, this result also leads to a bad fit (LR = 68.895 p = 0.000, and $\chi^2 = 67.272$ p = 0.000).

However, if third order and higher effects are zero the model fits (LR = 8.477 p = 0.132, and $\chi^2 = 8.390$ p = 0.136), indicating that the higher order interactions are not significant. The model adequately representing the dependencies in the presented data involves only significant main effects and first order interactions. Figure 4.4 lists all effects and their level of significance. A model with all main effects and first order interactions seems to suit the data well (see Fig. 4.5). However, one k = 3 interaction is significant: *depression by rigidity by vitality* (p = 0.04), but it is the only significant 3-way interaction. Among the k = 2 interactions a total of four reach the level of significance. They are *depression by introversion* (p = 0.02), *depression by vitality* (p = 0.00), *depression by rigidity* (p = 0.00), and *introversion by vitality* (p = 0.00).

Closely related to the concept of a hierarchical order of the main effects and all interactions in hi-log models is the so called **Lancaster decomposition** (Lancaster, 1951). Lancaster found out, that the **global chi-square** of a first order CFA is composed additively of all possible interactions. Let's look at our example with four the variables D, I, R, and V. The related global chi-square of a first order CFA is denoted as $\chi^2(DIRV)$, in contrast, let's denote the four-way interaction as $\chi^2_{(DIRV)}$. The Lancaster decomposition or partitioning of the global chi-square looks as follows

$$\begin{aligned}\chi^2(DIRV) = {} & \chi^2_{(DIRV)} + \chi^2_{(DIR)} + \chi^2_{(DIV)} + \chi^2_{(IRV)} + \chi^2_{(DRV)} \\ & + \chi^2_{(DI)} + \chi^2_{(DR)} + \chi^2_{(DV)} + \chi^2_{(IR)} + \chi^2_{(IV)} + \chi^2_{(RV)} \end{aligned} \tag{4.14}$$

As can be seen from the upper part of Fig. 4.5 (see index a) a global chi-square means that all k = 2 and higher effects are zero, this result in a $\chi^2 = 68.895$ with $df = 11$. Each component takes one degree of freedom. To add the two-way interactions, one needs to identify the values of all k = 2 order effects are zero (see index b in the upper part of the figure). This results in a $\chi^2 = 60.418$ with $df = 6$, and in order to assess the value of all three-way interaction, one looks at k = 3 order effects are zero (see index b in the upper part of the figure), this results in a $\chi^2 = 8.477$ with $df = 4$. The remaining difference is the value of the chi-square of the four-way interaction which in our example is almost zero and therefore



Parameter Estimates

Effect	Parameter	Estimate	Standard Error	Z-Value	Sig.
depression*introversion*rigidity*vitality	1	,000	,059	–,007	,994
depression*introversion*rigidity	1	–,073	,059	–1,224	,221
depression*introversion*vitality	1	–,038	,059	–,639	,523
depression*rigidity*vitality	1	,122	,059	2,058	,040
introversion*rigidity*vitality	1	,050	,059	,836	,403
depression*introversion	1	–,139	,059	–2,345	,019
depression*rigidity	1	,229	,059	3,862	,000
introversion*rigidity	1	,027	,059	,449	,653
depression*vitality	1	–,288	,059	–4,849	,000
introversion*vitality	1	–,195	,059	–3,278	,001
rigidity*vitality	1	–,029	,059	–,485	,628
depression	1	–,067	,059	–1,131	,258
introversion	1	,156	,059	2,631	,009
rigidity	1	–,073	,059	–1,227	,220
vitality	1	,033	,059	,561	,575

Parameter Estimates

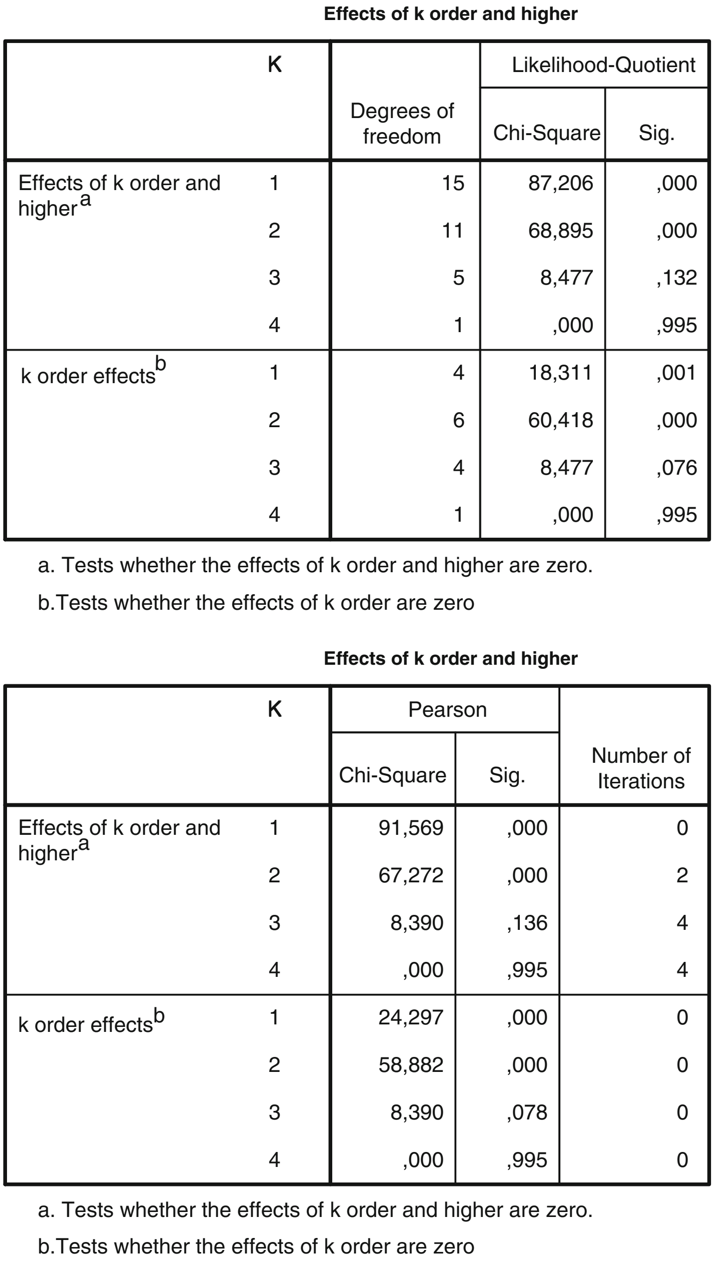
Effect	Parameter	95%-Confidence Intervall	
		Lower Limit	Upper Limit
depression*introversion*rigidity*vitality	1	–,117	,116
depression*introversion*rigidity	1	–,189	,044
depression*introversion*vitality	1	–,154	,079
depression*rigidity*vitality	1	,006	,239
introversion*rigidity*vitality	1	–,067	,166
depression*introversion	1	–,256	–,023
depression*rigidity	1	,113	,346
introversion*rigidity	1	–,090	,143
depression*vitality	1	–,405	–,172
introversion*vitality	1	–,311	–,078
rigidity*vitality	1	–,145	,088
depression	1	–,184	,049
introversion	1	,040	,273
rigidity	1	–,189	,044
vitality	1	–,083	,150

Fig. 4.4 Parameter estimates in a hierarchical log-linear model

non-significant (this interaction has also one degree of freedom). That means, that value of the four-way interaction is obtained by substracting all interaction effects from the global chi-square:

$$\begin{aligned} \chi^2_{(DIRV)} = \chi^2(DIRV) - \chi^2_{(DIR)} - \chi^2_{(DIV)} - \chi^2_{(IRV)} - \chi^2_{(DRV)} \\ - \chi^2_{(DI)} - \chi^2_{(DR)} - \chi^2_{(DV)} - \chi^2_{(IR)} - \chi^2_{(IV)} - \chi^2_{(RV)} \end{aligned} \tag{4.15}$$

and $\chi^2_{(DIRV)} = 68.895 - 60.418 - 8.477 = 0$.



Effects of k order and higher

	K	Degrees of freedom	Likelihood-Quotient	
			Chi-Square	Sig.
Effects of k order and higher[a]	1	15	87,206	,000
	2	11	68,895	,000
	3	5	8,477	,132
	4	1	,000	,995
k order effects[b]	1	4	18,311	,001
	2	6	60,418	,000
	3	4	8,477	,076
	4	1	,000	,995

a. Tests whether the effects of k order and higher are zero.

b.Tests whether the effects of k order are zero

Effects of k order and higher

	K	Pearson		Number of Iterations
		Chi-Square	Sig.	
Effects of k order and higher[a]	1	91,569	,000	0
	2	67,272	,000	2
	3	8,390	,136	4
	4	,000	,995	4
k order effects[b]	1	24,297	,000	0
	2	58,882	,000	0
	3	8,390	,078	0
	4	,000	,995	0

a. Tests whether the effects of k order and higher are zero.

b.Tests whether the effects of k order are zero

Fig. 4.5 K order effects and their goodness-of-fit in a hierarchical log-linear model

4.3 Zero-Order CFA or Configural Cluster Analysis (CCA)

One could also postulate the null hypothesis, that the cells are equally distributed. This is the assumption or underlying hypothesis of the ***zero order CFA*** **or Configural Cluster Analysis (CCA)**, where $e_{ij} = \frac{N}{T}$ and T represents the number of cells. The underlying model includes no main effects or interactions; that is, each cell has the same expected frequency. A CCA might also be run using the SPSS program.

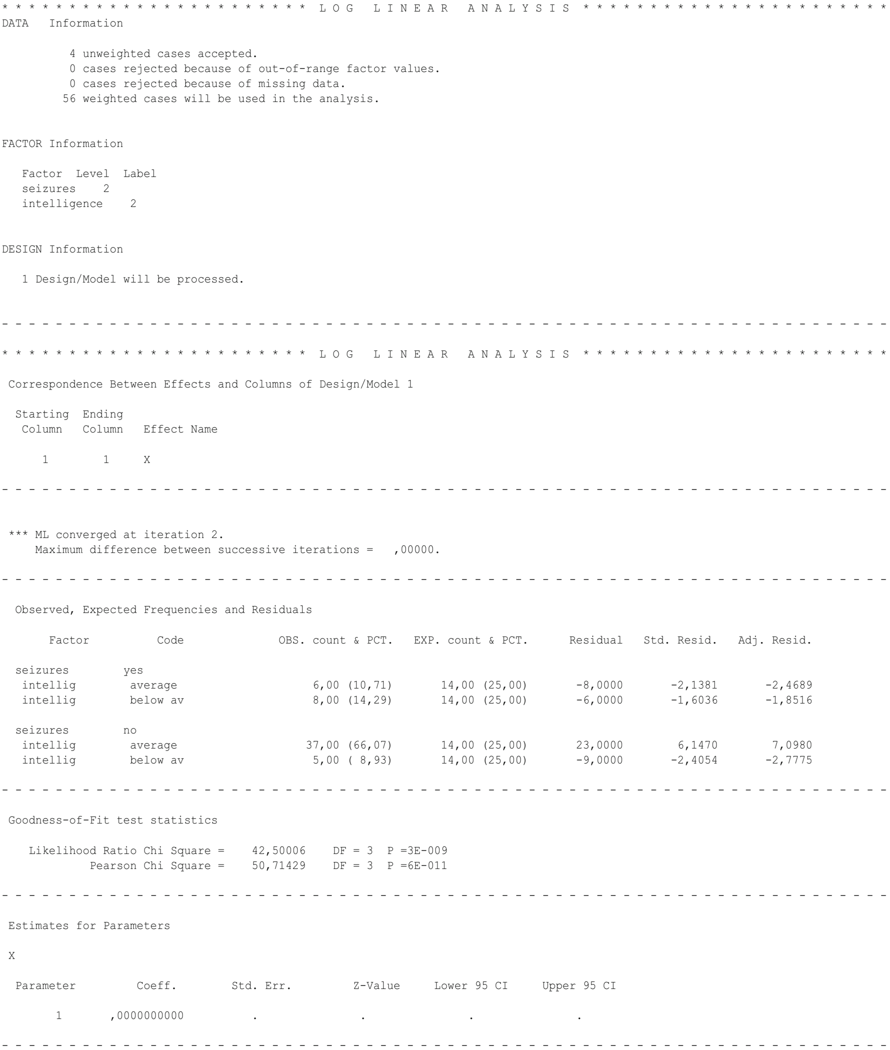
$$\begin{gathered}
\mathit{COMPUTE} \quad X = 1. \\
\mathit{LOGLINEAR}\,\mathit{seizures}(1,2)\,\mathit{intelligence}(1,2)\,\mathit{with}\ X \\
/\mathit{CRITERIA} = \mathit{DELTA}(0.01) \\
/\mathit{PRINT} = \mathit{DEFAULT} \\
/\mathit{DESIGN} = X.
\end{gathered}$$

The trick is to create a variable X which consists only of one value (i.e., $X = 1$), and which will be used as a covariate (see SPSS-Syntax: “with X”). The SPSS printout for a log-linear model based on the expected frequencies according to a zero-order CFA looks as follows (see Fig. 4.6). One can clearly see that the expected frequencies are the same for each configuration. Pearson’s chi-square and the LR statistics are highly significant, indicating that the model does not fit.

The *von Eye program* also offers the possibility of running a zero-order CFA. After typing the cell frequencies into the program, the program produces the output given in Fig. 4.7.

$$\begin{aligned}
\mathit{Here\ are\ the\ current\ options\ for\ CFA\ models}: \\
\mathit{zero\ Order\ CFA} = 0 \\
\mathit{First\ Order\ CFA} = 1 \\
\mathit{For\ any\ higher\ order\ model}:\ \mathit{indicate\ order} \\
\mathit{Two-sample\ CFA} = 20
\end{aligned} \tag{4.16}$$

The printout of the *von Eye program* while using the chi-square approximation to the z-statistic suggests two significant cells under the assumption of equal distribution across all cells: a *type* for “above average” *intelligence* together with “absent” *seizures*, and an *antitype* for “absent” *seizures*, while having “below average”



```
* * * * * * * * * * * * * * * * * * * * * * *  L O G   L I N E A R   A N A L Y S I S  * * * * * * * * * * * * * * * * * * * * * * *
DATA   Information

          4 unweighted cases accepted.
          0 cases rejected because of out-of-range factor values.
          0 cases rejected because of missing data.
         56 weighted cases will be used in the analysis.


FACTOR Information

   Factor  Level  Label
   seizures    2
   intelligence    2


DESIGN Information

   1 Design/Model will be processed.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
* * * * * * * * * * * * * * * * * * * * * * *  L O G   L I N E A R   A N A L Y S I S  * * * * * * * * * * * * * * * * * * * * * * *

 Correspondence Between Effects and Columns of Design/Model 1

  Starting  Ending
   Column   Column   Effect Name

      1        1     X

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -


 *** ML converged at iteration 2.
     Maximum difference between successive iterations =   ,00000.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

  Observed, Expected Frequencies and Residuals

       Factor          Code             OBS. count & PCT.   EXP. count & PCT.      Residual   Std. Resid.   Adj. Resid.

  seizures        yes
   intellig        average                   6,00 (10,71)       14,00 (25,00)       -8,0000       -2,1381       -2,4689
   intellig        below av                  8,00 (14,29)       14,00 (25,00)       -6,0000       -1,6036       -1,8516

  seizures        no
   intellig        average                  37,00 (66,07)       14,00 (25,00)       23,0000        6,1470        7,0980
   intellig        below av                  5,00 ( 8,93)       14,00 (25,00)       -9,0000       -2,4054       -2,7775

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

 Goodness-of-Fit test statistics

    Likelihood Ratio Chi Square =    42,50006    DF = 3  P =3E-009
             Pearson Chi Square =    50,71429    DF = 3  P =6E-011

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

 Estimates for Parameters

 X

  Parameter         Coeff.       Std. Err.         Z-Value     Lower 95 CI     Upper 95 CI

        1      ,0000000000          .               .               .               .

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Fig. 4.6 A Log-linear model based on the expected frequencies of a zero-order CFA

intelligence. The values for the Pearson chi-square and the Likelihood Ratio test (LR) are identical for the SPSS package and the von Eye program.

It is also possible to run a zero-order CFA with R-package ***confreq***. The design matrix simply needs to be adjusted. Let's look at the CCA R-syntax:

```
CFA(patternfreq_neu,form="null")
```

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


       Marginal Frequencies
       --------------------
       Variable Frequencies
       -------- -----------
           1       14.    42.

           2       43.    13.


     sample size N =       56

    the normal z-test was used
    Bonferroni-adjusted alpha =  .0125000
    a CFA of order   0  was performed


                                Table of results
                                ----- -- -------
    Configuration    fo       fe   statistic       p
    -------------   ----  -------- ---------    -------
          11          6.   14.000     -2.138   .01625466
          12          8.   14.000     -1.604   .05440471
          21         37.   14.000      6.147   .00000000    Type
          22          5.   14.000     -2.405   .00807846    Antitype


                    chi2 for CFA model =   50.7143
                    df =     3      p =  .00000000

                LR-chi2 for CFA model =    42.5001
                    df =     3      p =  .00000000
```

Fig. 4.7 Zero-order CFA or configural cluster analysis (CCA) using the von Eye program

The results for ***confreq*** are identical to the *von Eye program* in terms of *types* and *antitypes* while comparing the statistics of the normal approximation of the chi-square values:

```
$local.test
pattern observed expected loc.chi.sq loc.df
 1 1        6       14       4.571429   1
 1 2        8       14       2.571429   1
 2 1       37       14      37.785714   1
 2 2        5       14       5.785714   1
loc.chi.square.p      z.Chi      p.z.Chi
 3.250944e-02      -2.138090 1.625472e-02
 1.088094e-01      -1.603567 5.440472e-02
 7.895786e-10       6.147009 3.947893e-10
 1.615693e-02     -2.405351 8.078466e-03

$bonferroni.alpha
[1] 0.0125

$global.test
$global.test$chi.square
[1] 50.71429

$global.test$df
df
 3

$global.test$chi.square.p
[1] 5.628098e-11

$global.test$alpha
[1] 0.05
```

Both programs reveal the same results in terms of test statistics, global chi-square, and degrees of freedom. In the ***confreq*** printout very small numbers are listed with “e” plus a number e.g., “3.250944e−02” which stands for $3.250944^{1/100}$ which is equal to the number of decimal points: 0.03250944.

Coming back to our main effects model or first order CFA, the *type* with the configuration (1,2), (i.e., *seizures*=yes and *intelligence*=below average) should be investigated in more detail. Gustav A. Lienert, the inventor of CFA, was always looking for a coefficient which can be interpreted similar to the determination coefficient R^2 in multiple regression. This is the statistic Q, which is a **coefficient of precision** or a coefficient of *the pregnancy of a type* (Betzin & Bollmann-Sdorra, 2003; Krauth, 2008). There are two coefficients of precision available, (1) one for $2 \times e_{i,j} - o_{i,j} \geq 0$ and, (2) one for $o_{i,j} > 2 \times e_{i,j}$.

Table 4.5 Adjustment for Q_{max} taken from Betzin and Bollmann-Sdorra (2003)

T	Q_{adj}	T	Q_{adj}	T	Q_{adj}
1	–	6	0.782	15	0.910
2	0.414	7	0.811	20	0.932
3	0.587	8	0.834	50	0.972
4	0.682	9	0.852	75	0.982
5	0.741	10	0.866	100	0.986

In case (1) the formula looks as follows

$$Q_1 = \frac{|o_{ij} - e_{ij}|}{e_{ij}}. \tag{4.17}$$

Q varies between 0 and 1. Zero corresponds to the absence of precision and one to perfect precision. In case (2) the formula looks as follows

$$Q_2 = \frac{|o_{ij} - e_{ij}|}{Max\left\{e_{ij}, N - e_{ij}\right\}} \tag{4.18}$$

where N is the sample size. The second coefficient does not vary between 0 and 1. It needs an adjustment depending on the number of variables (T) involved $Q_{max} = \frac{Q_2}{Q_{adj}}$. This adjustment was calculated by Betzin and Bollmann-Sdorra (2003) (Table 4.5).

In our case $o_{1,2} = 8$ and $e_{1,2} = 3.75$, therefore $8 > 2 \times 3.75 = 7.50$ and Q_2 needs to be calculated, which is

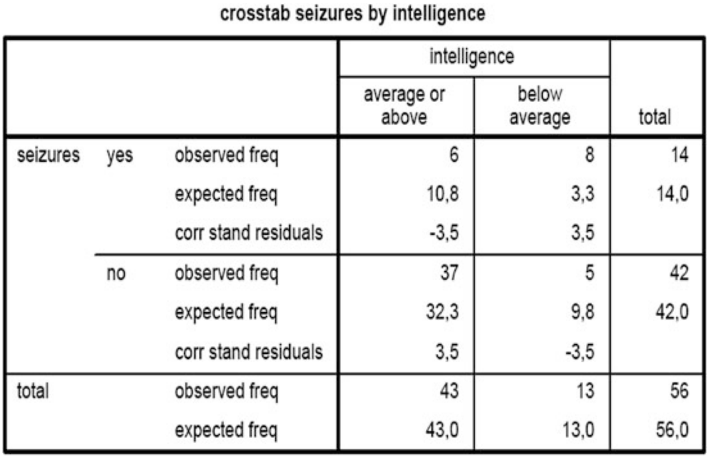
$$Q_2 = \frac{|8 - 3.75|}{Max\{3.75, 56 - 3.75\}} = 0.09.$$

and

$$Q_{max} = \frac{Q_2}{Q_{adj}} = \frac{0.09}{0.414} = 0.22.$$

That is, we have a medium-size coefficient of precision.

Let’s look at the crosstabs procedure in SPSS. The above data are imputed into the normal crosstabs procedure, using the following SPSS Syntax (see Box with Crosstabs Syntax).



crosstab seizures by intelligence

			intelligence		
			average or above	below average	total
seizures	yes	observed freq	6	8	14
		expected freq	10,8	3,3	14,0
		corr stand residuals	-3,5	3,5	
	no	observed freq	37	5	42
		expected freq	32,3	9,8	42,0
		corr stand residuals	3,5	-3,5	
total		observed freq	43	13	56
		expected freq	43,0	13,0	56,0

Fig. 4.8 Standardized residuals using SPSS crosstabs

```
Crosstabs
        /TABLES = seizures BY intelligence
        /FORMAT = AVALUE TABLES
        /STATISTICS = CHISQ
        /CELLS = COUNT EXPECTED ASRESID
        /COUNTASIS.
```

The printout (see Fig. 4.8) displays **corrected standardized residuals (csr)**, which can also be used in the search of *types* or *antitypes*. They are calculated according to the Fuchs-Kenett Test (cf. Fuchs & Kenett, 1980; Lautsch & Thöle, 2003; Haberman, 1977; Stemmler, 1994). The underlying equation is

$$csr_{ij} = \frac{(o_{ij} - e_{ij})}{\sqrt{e_{ij} \times \left(1 - \frac{o_{i.}}{n}\right) \times \left(1 - \frac{o_{.j}}{n}\right)}} \tag{4.19}$$

where n = sample size; e_{ij} are the expected frequencies; o_{ij} are the observed frequencies; $o_{i.}$ are the sum of frequencies for row i (row marginals); $o_{.j}$ are the column marginals of column j. The csr_{ij} needs to be compared to a Bonferroni adjusted alpha (i.e., $\alpha^* = 0.025/4 = 0.00625$) and the corresponding z-value is $|2.50|$. The crosstabs procedure in SPSS also provides residuals which are similar to the *Adjusted Residuals* (abbreviated as Adj. Resid. in the SPSS printout) calculated by the loglinear procedure in SPSS (see Fig. 4.1). Having a non-fitting log-linear model results in significant *types* or *antitypes* which might be detected with a CFA program or in SPSS using a log-linear model or a crosstab.

What is the relationship between CFA and log-linear modeling?

- Log-linear modeling is a statistical tool which is engaged in model fitting; CFA is engaged in residual analysis.
- Both statistical tools use the same procedures for estimating expected frequencies.
- CFA is interested in the interpretation of individual cells.
- CFA attempts to reject the local null hypothesis that is, the fit between observed and expected frequencies on the cell level.
- CFA and log-linear models should be applied in a complimentary way.
- CFA can validate the found global structure of dependence on the local level.

4.4 The Limits of CFA or Different Base Models, Different Types

The calculation of the expected frequencies depends on the chosen **base model**. There are in fact at least four different possible models for such a base model or the respective null hypothesis. The first model is the model of independence or a main effects model, which we used above for our first-order CFA. The expected frequencies can be calculated with the help of a hierarchical log-linear model. This model makes an important assumption; it assumes, that each case in the table was drawn from the same population. The second approach, the Victor-approach to CFA, was introduced by Victor (1989) and Kieser and Victor (1991, 1999). Their underlying null hypothesis is based on the assumption, that the CFA *types* or *antitypes* were drawn from a different population. The third approach is the functional approach to CFA by von Eye and Mair (2007). The fourth approach was introduced by Gutiérrez-Penã and von Eye (2000; 2012), von Eye and Gutiérrez-Penã (2004) it calculates the expected frequencies according to the Bayes Theorem (this approach will not be explained in more details; the interested reader may refer to the listed citations).

Victor and his colleague Kieser (Kieser & Victor 1991, 1999; Victor, 1989) were the first statisticians who draw our attention to the limits of CFA. They used a very simple data example (see Table 4.6; Victor, p. 72) to introduce the idea of the so-called **Victor-cells**. Eyeballing Table 4.6 suggests that configuration $< 1, 1 >$ is an extreme cell, and therefore a *type*. However, running a main effects model or first-order CFA results in none of that(!), neither types nor antitypes (see Table 4.7).

Obviously, the base model of independence was not sensitive enough to detect the *type* in configuration '11'. The problem is that the assumption of independence is assumed for the whole contingency table. In order to actually detect the obvious *type*

Table 4.6 Three by three contingency table with one Victor-cell

				Σ
	10	1	1	12
	1	1	1	3
	1	1	1	3
Σ	12	3	3	n = 18

Table 4.7 Results of a first-order CFA based on the data of the contingency table with one Victor-cell

Pattern	Observed	Expected	chi.square	chi.square.p	z.Chi	p.z.Chi
1 1	10	8.00	0.50	0.48	0.71	0.24
1 2	1	2.00	0.50	0.48	−0.71	0.24
1 3	1	2.00	0.50	0.48	−0.71	0.24
2 1	1	2.00	0.50	0.48	−0.71	0.24
2 2	1	0.50	0.50	0.48	0.71	0.24
2 3	1	0.50	0.50	0.48	0.71	0.24
3 1	1	2.00	0.50	0.48	−0.71	0.24
3 2	1	0.50	0.50	0.48	0.71	0.24
3 3	1	0.50	0.50	0.48	0.71	0.24

Table 4.8 Contingency table with new expected frequencies according to Victor

				Σ
	1	5.5	5.5	12
	5.5	0.3	0.3	6.1
	5.5	0.3	0.3	6.1
Σ	12	6.1	6.1	n = 18

statistically, Victor (1989) suggested one treat the extreme cell with pattern ‘11’ as a **structural zero**. Structural zeros are usually cells which cannot be observed (e.g., a pattern of heavy rain together with a beautiful blue sky). The idea is to calculate new expected frequencies using the **Deming-Stephan-Algorithm** (Haberman, 1977). This is an iterative procedure using the following steps:

- Step 1: Set all extreme cells to zero $n_{ij} = 0$;
- Step 2: Recalculate the new total sample size $N_{..}$ and the new marginal sums of the contingency table;
- Step 3: Recalculate for all suspicious extreme cells the new expected frequencies;
- Step 4: Repeat steps 2 and 3 until the new total sample size $N_{..}$ does not change anymore.

The new expected frequencies according to Victor (1989) are also presented formulas to calculate the new expected frequencies $\hat{x}_{ij}$ directly. Here the formulas for a two-dimensional table are displayed. However, Victor presented formulas for three- and four-dimensional tables (Victor, 1983):

$$\hat{x}_{11} = \tilde{n}_{1.}\tilde{n}_{.1}/(\tilde{n} - \tilde{n}_{1.} - \tilde{n}_{.1})\, with \qquad (4.20)$$

Table 4.9 Three by three contingency table with two extreme outliers

				Σ
	1	10	10	22
	10	10	10	30
	10	10	370	390
Σ	21	30	390	n = 441

$$\tilde{n}_{1.} = n_{1.} - n_{11},$$
$$\tilde{n}_{.1} = n_{.1} - n_{11}\, and,$$
$$\tilde{n} = n - n_{11}$$

$$\hat{x}_{11} = \frac{(12-10)(12-10)}{(18-10)-(12-10)-(12-10)} = 1$$

As a next step the new expected frequencies will be used to detect a Victor-type:

$$\chi^2 = \frac{(\mathrm{n}_{11} - \hat{\mathrm{x}}_{11})^2}{\hat{\mathrm{x}}_{11}} \text{ with df} = 1 \tag{4.21}$$

Using the data from Tables 4.6 and 4.8 result in a

$$\chi^2 = \frac{(10-1)^2}{1} = 81$$

which is highly significant! Dunkl and von Eye (1990) suggest that one should use the variance of the new expected frequencies for the denominator:

$$\chi^2 = \frac{(\mathrm{n}_{ij} - \hat{\mathrm{x}}_{ij})^2}{\mathrm{Var}\,(\hat{\mathrm{x}}_{ij})} \text{ with df} = 1 \text{ and} \tag{4.22}$$

$$\mathrm{Var}\,(\hat{\mathrm{x}}_{ij}) = \frac{\hat{\mathrm{x}}_{ij} + 0.5}{\hat{\mathrm{x}}_{ij} - 0.5}\,\hat{\mathrm{x}}_{ij}$$

After confirming the existence of a Victor-type, one has to test whether the rest of the table is independent. Inserting the data from Tables 4.6 and 4.8 result in a $\chi^2 = 3.68$ which is not significant. If the rest of the table is independent, this is called **quasi-independence** in the presence of one *type*. The procedure is as follows: One first searches for a Victor-type, then the remaining cells are tested for the hypothesis of quasi-independence. If this leads to the rejection of quasi-independence, it can be assumed that another Victor-type exists. Again, after identifying the second Victor-type, the remaining table is tested. If the remaining table is quasi-independent, the procedure ends (cf. von Eye & Stemmler, 1992).

Another interesting data example was introduced by Kieser (Kieser and Victor, 1999, p. 969; see Table 4.9): Again, from eyeballing Table 4.9 one would assume two existing *types*, one in pattern ‘11’ and the other in ‘33’. However, running a first-order CFA results in a surprising result: Cell ‘11’ is the only cell that fulfills the condition of independence, all other cells are declared to represent *types*. In this case the use of a different base model, for instance, the Victor-approach, is recommended.

The other useful approach is the functional CFA (von Eye & Mair, 2007). This approach extends the design matrix in order to blank out extreme cells, while setting them to a structural zero. Let’s take the data from Table 4.6 and blank out the cell ‘11’ with the following design matrix

$$X = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & -1 & 0 \end{pmatrix} \tag{4.23}$$

The first four columns represent the ordinary main-effects model with the constant and the main-effects effect-coded. The fifth column blanks out cell ‘11’. The remaining log-linear model should now result in a fit with no significant residuals being left. The elimination of cells can be based simply on searching for the largest residual. The overall goal is to extend the design matrix until the model fits. While first-order CFA is a one-step procedure, the Victor-approach and the functional CFA are multiple-steps procedures which of course need an Alpha adjustment. Unfortunately, the R-package **confreq** is not yet able to build the above design matrix. However, the R-package **CFA** does (see von Eye, Mair and Mun, 2010).

Summary: Log-linear modeling and CFA may be used as complimentary statistical tools. Log-linear modeling looks for models with an appropriate goodness-of-fit; that is, the corresponding chi-square and LR-values result in p-values lesser than a previously chosen level of significance (α is usually 0.05). An inspection of the residuals reveals deviations from an assumed base model. the residuals indicate whether they were caused by chance, artefacts or other errors in the data. The residuals also indicated whether the model itself needs be modified. CFA needs a non-fitting model ($p > 0.05$) in order to detect *types* and/or *antitypes*. The expected frequencies are calculated according to the underlying null model which is specified in the design matrix using the General Linear Model approach (GLM). CFA and log-linear modeling are based on the same algorithms, however, they pursuit different goals, CFA is “cell-oriented” and log-linear modeling is “dependency-structure oriented” (Victor, 1989).

One may specify a main effects model which corresponds to the first-order CFA or a base model which corresponds to a zero-order CFA or configural cluster analysis (CCA). A saturated model may be used to investigate all available k order effects. Here also a hierarchical log-linear model can be run in SPSS, were all k order effects and all single effects are tested for significance. Once an appropriate null model has detected significant cells or configurations, the pregnancy of the significant cells can be investigated with the test statistic Q. The Fuchs-Kenett-test is another possible significance test in CFA.

Sometimes, one may find a contingency table with extreme cells. In this situation, the first-order CFA model might not be appropriate. As an alternative, the outlier cells can be treated as structural zeros, using the Victor-approach or the functional CFA.

References

Dunkl, E., & von Eye, A. (1990). Kleingruppentests gegen Victor-Typen und -Syndrome [Small group tests against Victor-types and -syndromes]. *Zeitschrift für Klinische Psychologie, Psychopathologie und Psychotherapie [Journal of Clinical Psychology, Psychopathology and Psychotherapy], 1*, 46–51.

Betzin, J., & Bollmann-Sdorra, P. (2003). On the determination coefficient of CFA. *Psychology Science, 45*(2), 400–420.

Fuchs, C., & Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association, 75*, 395–398.

Gutiérrez-Penã, E., & von Eye, A. (2000). A Bayesian approach to configural frequency analysis. *Journal of Mathematical Sociology, 24*, 151–174.

Gutiérrez-Penã, E. (2012). Bayesian predictive configural frequency analysis. *Psychological Tests and Assessment Modeling, 54*(3), 285–292.

Haberman, S. J. (1977). *The analysis of frequency data.* Chicago: University of Chicago Press.

Kieser, M., & Victor, N. (1991). A test procedure for an alternative approach to configural frequency analysis. *Methodika, 5*, 87–97.

Kieser, M., & Victor, N. (1999). Configural frequency analysis (CFA) revisited – A new look at an old apporach. *Biometrical Journal, 41*, 967–983.

Krauth, J. (2008). The pregancy of types and antitypes in CFA. In M. Stemmler, E. Lautsch, & D. Martinke (Eds.), *Configural frequency analysis (CFA) and other non-parametrical methods: Gustav A. Lienert memorial issue* (pp.19–28). Lengerich, Germany: Pabst Science Publishers.

Krauth, J., & Lienert, G. A. (1973). *Die Konfigurationsfrequenzanalyse und ihre Anwendung in Psychologie und Medizin [Configural frequency analysis and its application in psychology and medicine].* Freiburg, Germany: Alber.

Kutner, M. H., Neter, J., Nachtsheim, C. J., & Li, W. (2004). *Applied linear statitical models* (4th ed.). Boston: McGraw-Hill.

Lancaster, H. O. (1951). Complex contingency tables treated by the partition of χ^2. *Journal of the Royal Statistical Society, 13*, 242–249.

Langeheine, R. (1980). *Log-lineare Modelle zur multivariaten Analyse qualitativer Daten [Log-linear models and the multivariate analysis of qualitative data].* München, Germany: R. Oldenbourg Verlag.

Lautsch, E., & Thöle, U. (2003). Classification and explanation of life conceptions using the case of the 14th Shell Youth Study 2002. *Psychology Science, 45*(2), 263–279.

Rindskopf, D. (1990). Nonstandard log-linear models. *Psychological Bulletin, 108*(1), 150–162.

Stemmler, M. (1994). A nonparametrical evaluation of ANOVA and MANOVA designs usign interaction structure analysis. *Biometrical Journal, 36*(8), 911–925.

Victor, N. (1983). A note on contingency tables with one structural zero. *Biometrical Journal, 25*, 283–289.

Victor, N. (1989). An alternative approach to configural frequency analysis. *Methodika, 3*, 61–73.

von Eye, A. (1990). *Introduction to configural frequency analysis: The search for types and antitypes in cross-classifications*. Cambridge, UK: Cambridge University Press.

von Eye, A. (2002). *Configural frequency analysis: Methods, models and applications*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A., & Stemmler, M. (1992). Die Konfigurationsfrequenzanalyse: Kieser's Test gegen Victor's Typen und Syndrome [Configural frequency analyses: Kieser's test versus Victor's types and syndromes]. *Zeitschrift für Klinische Psychologie, Psychopathologie und Psychotherapie [Journal of Clinical Psychology, Psychopathology and Psychotherapy], 2*, 174–178.

von Eye, A., & Gutiérrez-Penã, E. (2004). Configural frequency analysis: The search for extreme cells. *Journal of Applied Statistics, 31*, 981–997.

von Eye, A., & Mair, P. (2007). *A functional approach to configural frequency analysis. Research report series. Department of Statistics and Mathematics* (Vol. 48). Vienna: WU Vienna University of Economics and Business.

von Eye, A., Mair, P., & Mun, E.-Y. (2010). Advances in configural frequency analysis. NewYork: The Guilford Press.

Chapter 5
Longitudinal CFA

Abstract This chapter explains how to use CFA with longitudinal data. Different ways of rearranging the information with the longitudinal data are introduced. First, the analysis of first differences is demonstrated by simply looking at increases or decreases between two time points. Secondly, CFA and visual shape patterning are explained. Here the shape of the curve are used as categories or patterns. Furthermore, a test of marginal homogeneity is provided which tests the null hypothesis of the homogeneity of marginals in a square contingency table. Moreover, a special type, the discrimination type is described. This type differentiates significantly between two independent samples.

5.1 CFA of First Differences

This section explains how to use CFA in longitudinal data. One approach is the use of first differences (von Eye, 2002). For the method of differences consider a series of measures $y_0, y_1, \ldots, y_n$. The first difference between two measures is termed $\Delta y_0 = y_1 - y_0$. In order to use first differences properly two conditions must be fulfilled (cf. von Eye, 2002). (1) The data points used for creating differences must be *equidistant*. (2) The scores that are subtracted from each other are at the interval level.

Let's examine the following data taken from Lienert (1978). $N = 72$ female students were asked to rate their $M = Mood$, their $C = Power\ of\ Concentration$, and $S = Staying\ Power$ on a rating scale ranging from 0 to 100, before (y_0) and after (y_1) their menstruation. We create first differences such that increments are labeled '$+$' and decrements are labeled '$-$' (in cases where the difference is zero, those differences are equally or randomly distributed to either '$+$' or '$-$'). Of course, the choice of coding three options would also be possible. With two options the number of cells in a multidimensional table is smaller. We are now able to ask two kinds of questions. First, we might want to know whether all female students change in their own manner with regard to *Mood*, *Power of Concentration*, and *Staying Power* due

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_5,

Table 5.1 CFA of first differences as a zero-order CFA and first order CFA

				Zero-order CFA		First order CFA	
M	C	S	$f_{(o)}$	$f_{(e)}$	z-statistic	$f_{(e)}$	z-statistic
+	+	+	1	9	−2.667	2.199	−0.809
+	+	−	4	9	−1.667	3.079	0.525
+	−	+	3	9	−2.000	6.134	−1.265
+	−	−	12	9	1.00	8.588	1.164
−	+	+	8	9	−0.333	5.718	0.955
−	+	−	6	9	−1.000	8.005	−0.709
−	−	+	18	9	3.000	15.949	0.514
−	−	−	20	9	3.667	22.329	−0.493
	Σ		72	72	$\chi^{2[0]} = 38.44$	72	$\chi^{2[1]} = 5.81$

to their menstruation, or to put it differently, if there is no systematic change. This is a question of the *zero-order CFA* or Configural Cluster Analysis (CCA). Second, we might want to know whether there is a systematic or *typical female* change in all three investigated characteristics. In the following table (see Table 5.1) we used both CFA versions, *zero-order CFA* and *first order CFA*.

For CCA the degrees of freedom are $df = T - 1$, with T representing the number of cells or configurations, and for first order CFA $df = T - \sum_{i=1}^{d}(v_d - 1) - 1$ with d representing the number of variables, and v the number of categories of a variable. Here we have v = 3 categories with d = 3 variables, that is, $df = 27 - (3-1) - (3-1) - (3-1) - 1 = 20$. Let's look at the results of the CCA first. The Bonferroni adjusted alpha is $\alpha^* = \alpha/8 = 0.003125$ which results in a critical z-value of −2.73. Therefore two *types*, that is '$---$' and '$--+$' evolved, indicating that not all female students changed in the same manner. The overall $\chi^2 - value = 38.44$ is significant for $df = 7$. The question, whether there is a typical stereotypical female change (e.g., down-swing in mood, loss of concentration and staying power) needs to be denied; the *global* $\chi^2 - value = 5.81$ was not significant for $df = 20$. However, menarcheal changes still seem to be a sort of a female confinement, because the majority of 56 women experienced decrements in at least two categories, in comparison to only 16 women, who reported increments in two aspects.

5.2 CFA and Visual Shape Patterns

Krauth (1973) suggested a categorical approach for the analysis of longitudinal data which do not meet the assumptions for parametrical testing. This involves the analysis of the classification of response curves, where the response curves are categories representing change patterns based on first differences. Using the patterns '+', '−', and '=' various representations of the data are possible. Stemmler (1998) suggested visual shape patterns as categories to classify the data. A visual

Table 5.2 Visual shape patterns for four measurement points

			Pattern: T+			
+++	+=+	==+	++=	+==	=++	=+=
			Pattern: T−			
−−−	=−=	=−−	−−=	==−	−=−	−==
			Pattern: U+			
	−−+	−+=	−=+	−++	=−+	
			Pattern: U−			
	+−=	+−−	+=−	=+−	+−−	
			Pattern: *S*±			
			+−+	−+−		
			Pattern: H			
			===			

shape pattern assigns upper case letters and a sign to represent the complete shape of a curve. With M measurement points one can differentiate $M - 1$ curves, such as linear T-shaped, quadratic U-shaped, and cubic S-shaped curves plus H which represents a horizontal line. Table 5.2 lists all possible visual shape patterns for M = 4 measurement points, including ties (i.e., ‘=’).

For instance, linear-shaped curves, denoted T+, represent monotonical increase in scores based on the pattern ‘+ + +’ encompassing slight increase or even ties (e.g., ‘= +’); the opposite would be T− ‘− − −’ including, for example, ‘== −’(see Fig. 5.1). U-shaped curves denoted as U+ (shaped like a regular U) encompass five patterns: ‘= −+’, ‘− + +’, ‘− = +’, ‘− + +’, and ‘− − +’. The same applies to U− (inverted U-shaped curve or N-shaped curve): ‘= +−’, ‘+ − −’, ‘+− =’, ‘+ = −’, and ‘+ − +’, and S− representing the patterns ‘− + −’. H stands for ‘===’ which is a horizontal line.

The data in Stemmler (1998) was taken from a study in which $N = 54$ goldfish were trained to avoid light electroshocks in an aquarium by crossing a light beam or barrier. The dependent variable was the number of avoiding reactions performed by the fish, with higher numbers representing better training or learning. At the end of the training sessions the goldfish received injections of puromycine hydrochloride, a substance which interferes with the ability to learn and to memorize. Four days after the injection the actual testing phase began. For the training and for the testing condition a visual shape patterning was performed. The question was whether there is a stability of the *Visual Shape types* over time or whether the intervention resulted in a change of the frequency of the detected visual shapes or *types*. That is, is there a stability of the detected *types* over time? Table 5.3 shows the frequencies of the visual shapes before and after the medical intervention. The numbers in the main diagonal are listed in parenthesis, because they are not included in the analyses which are looking for change, and the main diagonal represents stability.

This table is analyzed with the help of Lehmacher’s **test of marginal homogeneity** (Lehmacher, 1980; see also Müller, Netter, & von Eye, 1997). This test searches

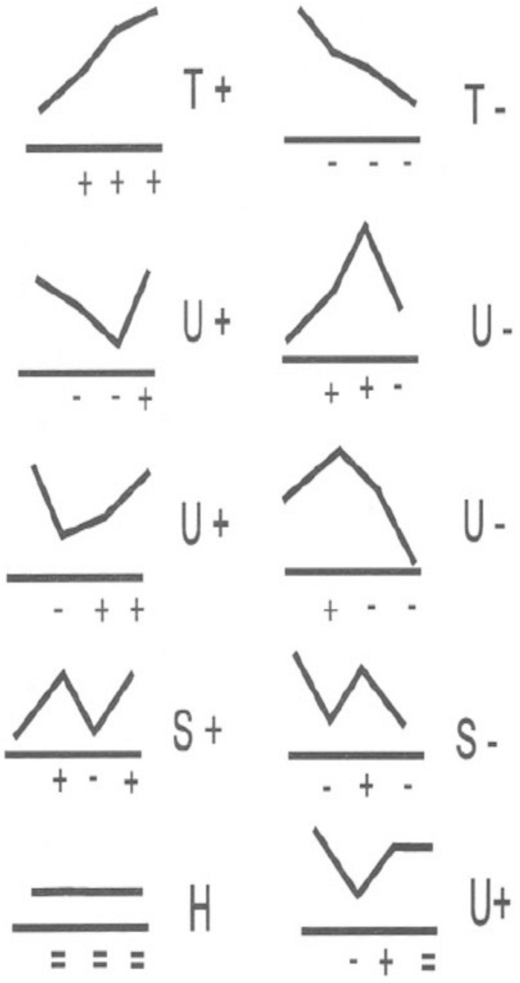


Fig. 5.1 Examples of visual shape patterns based on first differences with four measurement points

for equal probabilities (i.e., $p_{i.} = p_{.i}$) and therefore equal pairs of marginals (i.e., $f_{i.} =$) in a square symmetric contingency table. The null hypothesis is

$$\mathrm{H}_0 : \mathrm{p}_{\mathrm{i.}} = \mathrm{p}_{\mathrm{.i}} \quad \text{for} \quad \text{all} \quad \mathrm{i} = 1(1)\mathrm{r} \tag{5.1}$$

with i representing the columns and rows in a square contingency table. The alternative hypothesis states that at least one pair of marginals is unequal.

$$\mathrm{H}_1 : \mathrm{p}_{\mathrm{i.}} \neq \mathrm{p}_{\mathrm{.i}} \tag{5.2}$$

Table 5.3 Squared contingency table with frequencies of visual shape patterning

Testing condition								
Training condition	T+	T−	U+	U−	S+	S−	H	Σ ...
T+	(0)	7	3	4	2	0	3	19
T−	1	(1)	0	1	0	0	0	2
U+	1	3	(1)	1	0	3	0	8
U−	0	1	2	(0)	0	0	1	4
S+	0	2	1	0	(0)	2	1	6
S−	0	4	0	1	0	(0)	1	6
H	1	4	0	1	0	1	(0)	7
Σ	3	21	6	8	2	6	6	52

Significantly different pairs of marginals may be interpreted as **discrimination types** *types* that differentiate significantly between two (originally) independent samples (cf., Krauth, 1993; Lautsch & von Weber, 1995; von Eye, 1990, 2002). If, for example, 'T−' is a discriminating type and there are more goldfish that follow this pattern in the testing phase than in the training phase, one can conclude, that the injected drug led to a significant monotonical decrease in avoidant reactions (i.e., memory loss) in a number of goldfish. Lehmacher's test is basically a test for asymmetrical change. If we reject the null hypothesis, we know that there was change; if we keep the null hypothesis, there still may be change, but this change is symmetrical in either direction.

Lehmacher's sign test is applied to the data – excluding the data from the main diagonal – with $A_i = f_{i.} - f_{ii}$ for the rows and $B_i = f_{.i} - f_{ii}$ for the columns. Under the null hypothesis the expected frequencies A_i are equal to B_i. For the asymptotic test, the chi-square value is

$$\phi^2 = \frac{(A_i - B_i)^2}{(A_i + B_i)} \quad \text{with 1 df} \tag{5.3}$$

and with i = 1(1)r, if no specific alternative hypothesis was formulated. If, for instance, i = 1 for T+

$$\chi^2 = \frac{(A_1 - B_1)^2}{(A_1 + B_1)} = \frac{(19-3)^2}{(19+3)} = 11.63$$

the resulting chi-square value is $\chi^2 = 11.63$ with df = 1. For each of r Lehmacher's tests a two-tailed hypothesis is tested, because each of the column marginals may be either smaller or larger than the row marginals. Of course the Bonferroni adjustment needs to be applied to each test. Selecting the alpha level as 0.025, with r = 7 simultaneous tests the two-tailed Bonferroni adjusted $\alpha^* = 0.025/7 = 0.0035$ which corresponds to a chi-square value of $\chi^2 = 8.53$ (see Table 5.4). For small expected

Table 5.4 Chi-square values based on Lehmacher's test of marginal homogeneity and the binomial values for each pair of marginals of the visual shape pattern

	i	Training condition	Testing condition	χ^2	Binomial test
1	T+	19	3	11.63*	0.001*
2	T−	2	21	15.70*	<0.001*
3	U+	8	6	0.14	0.395
4	U−	4	8	1.33	0.194
5	S+	6	2	2.00	0.145
6	S−	6	6	0.00	1.000
7	H	7	6	0.08	0.500

$^*p<0.0035$

frequencies it is necessary to obtain an exact test using the binomial test. Because we have only two possible outcomes $p = q = 0.5$, the binomial formula (3.1) for the joint probability looks much simpler

$$B_i = \frac{1}{2}^n \sum_{0}^{x} \binom{n}{x} \tag{5.4}$$

with $n = A_i + B_i$ and x is the minimum of (A_i, B_i). The corresponding values of the exact binomial test are also listed in Table 5.4.

References

Krauth, J. (1973). Nichtparametrische Ansätze zur Auswertung von Verlaufskurven [Nonparametrical approaches to the analysis of curves]. *Biometrical Journal, 15*, 557–566.

Krauth, J. (1993). *Einführung in die Konfigurationsfrequenzanalyse* [Introduction to configural frequency analysis]. Weinheim, Germany: Beltz-Verlag.

Lautsch, E., & von Weber, S. (1995). *Methoden und Anwendungen der Konfigurationsfrequenzanalyse (KFA)* [Methods and applications of configural frequency analysis]. Weinheim, Germany: Beltz/Psychologie Verlags Union.

Lehmacher, W. (1980). Simultaneous sign tests for marginal homogeneity of square contingency tables. *Biometrical Journal, 22*(8), 795–798.

Lienert, G. A. (1978). *Verteilungsfreie Methoden in der Biostatistik (Band II)* [Non-parametrical methods in the field of biometrics (Vol. II)]. Meisenheim am Glan, Germany: Hain.

Müller, M. J., Netter, P., & von Eye, A. (1997). Catecholamine response curves of male hypertensives identified by Lehmacher's two sample configural frequency analysis. *Biometrical Journal, 39*, 29–38.

Stemmler, M. (1998). Nonparametrical analysis of change patterns in dependent samples. *Methods of Psychological Research Online, 3*(29), 23–36.

von Eye, A. (1990). *Introduction to configural frequency analysis: The search for types and antitypes in cross-classifications*. Cambridge, UK: Cambridge University Press.

von Eye, A. (2002). *Configural frequency analysis: Methods, models and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chapter 6
Other Person-Centered Methods Serving as Complimentary Tools to CFA

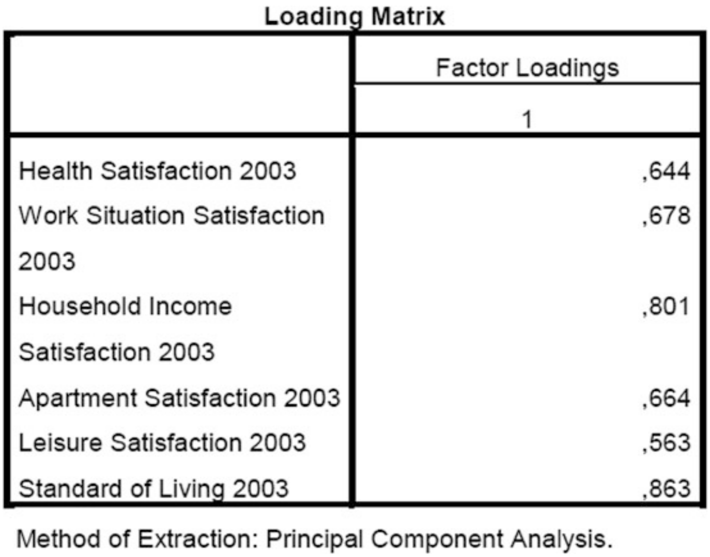
Abstract This chapter explains the use of other person-centered methods as complimentary tools to CFA. Among them are CHAID-model (Chi-Square Automatic Interaction Detection) as a model for contrasting groups, latent class analysis (LCA) which is comparable to a factor analysis using categorical variables and (multiple) correspondence analysis (CA) which is a technique for dimensional reduction and perceptual mapping. Using small data examples, the essence of each statistical method is explained and its close relationship to CFA is demonstrated. CFA may always be used as a complimentary tool offering additional insight into the data.

6.1 Answer Tree and CFA

In this section we will be using CFA to take a more in-depth look at psychometric scaling. For instance, instead of looking at factor loadings to determine which items contribute the most to a latent factor, we look at groups of subjects with typical answers on a scale. Let's say we measure *Life Satisfaction* with several items. We want to go beyond the traditional psychometric analyses and look at what aspects of life satisfaction lead people to rate their life as highly satisfied or dissatisfied. In a second step we might want to use latent class analysis (LCA) to explore the dimensionality of scales. All statistical procedures presented here use CFA and other complimentary tools to get further insight into the scale structure.

For our illustration we use the data from a German Socio-Economic Panel called SOEP. The SOEP is an annual survey on the German infrastructure conducted by the Deutsches Institut für Wirtschaftsforschung (DIW; German Institute for Economic Research) in Berlin. In the selected data set, the overall *Life Satisfaction* of Germans was measured from 2003 through 2006, each year on a scale ranging from 0 to 10, with '10' being the highest level of *Life Satisfaction*. We categorize the data to come up with three categories: we have increases and decreases in *Life Satisfaction*. But it is also possible to come up with a third category, *stability* (i.e., '='). We code changes of 0–2 scores from 1 year to another as '2', representing stability. A '1'

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_6,



Loading Matrix

	Factor Loadings
	1
Health Satisfaction 2003	,644
Work Situation Satisfaction 2003	,678
Household Income Satisfaction 2003	,801
Apartment Satisfaction 2003	,664
Leisure Satisfaction 2003	,563
Standard of Living 2003	,863

Method of Extraction: Principal Component Analysis.

Fig. 6.1 Factor loadings of the items measuring life satisfaction

refers to a decrease of at least three scores (i.e., -3); a ‘3’ represents an increase of at least three scores (i.e., $+3$). The underlying question is whether there are typical or significant patterns of change or stability in terms of *Life Satisfaction*.

First, we run a factor analysis using the original SOEP raw data by including another six items or indicators of life satisfaction which were measured in 2003 (the variable names are listed in parentheses): (1) *Satisfaction with Workplace (Work03)*, (2) *Satisfaction with Health Status (Health03)*, (3) *Satisfaction with Household Income (Income03)*, (4) *Satisfaction with Housing Conditions (Apart03)*, (5) *Satisfaction with Leisure Activities (Leisure03)*, and (6) *Satisfaction with Standard of Living (Standard03)*. All items are loading on one factor, which explains 50.3 % of the variance of all items (see SPSS printout in Fig. 6.1).

In order to run a first-order CFA, we select four items with the highest loadings: Work03, Income03, Apart03, and Standard03, and we conduct a median-split such that a ‘1’ represents *low satisfaction* and a ‘2’ *high satisfaction*. By looking exclusively at *types* we detect a total of five. The configurations ‘1111’ and ‘2222’ represent, two complimentary types and those subjects who are unsatisfied with all aspects of their life and those who, overall, are satisfied with all aspects. The complimentary types ‘1222’ and ‘2111’ represent persons who are satisfied with all aspects of their life except with *Work* or who are satisfied only with *Work* but with nothing else. This result indicates that a person’s working condition is an essential aspect of our judgement of *Life Satisfaction*. The fifth type represents people who are generally satisfied with their life but unhappy with their housing condition (configuration ‘2221’) (see Fig. 6.2 and Table 6.1).

The von Eye program and **confreq** reveal identical results. In the next step we will be using the program SPSS to create a summary index of the four items

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


           Marginal Frequencies
           --------------------
           Variable Frequencies
           -------- -----------
               1     3354.  3721.

               2     3688.  3387.

               3     4226.  2849.

               4     4625.  2450.


         sample size N =     7075

        the normal z-test was used
        Bonferroni-adjusted alpha =  .0031250
        a CFA of order   1  was performed


                                  Table of results
                                  ----- -- -------
        Configuration    fo       fe   statistic       p
        -------------   ----  -------- ---------    -------
            1111       1406.  682.678     27.684   .00000000    Type
            1112        307.  361.635     -2.873   .00203309    Antitype
            1121        167.  460.234    -13.669   .00000000    Antitype
            1122        124.  243.800     -7.673   .00000000    Antitype
            1211        299.  626.960    -13.098   .00000000    Antitype
            1212        127.  332.120    -11.255   .00000000    Antitype
            1221        356.  422.672     -3.243   .00059157    Antitype
            1222        568.  223.902     22.996   .00000000    Type
            2111       1200.  757.378     16.083   .00000000    Type
            2112        230.  401.205     -8.547   .00000000    Antitype
            2121        140.  510.594    -16.401   .00000000    Antitype
            2122        114.  270.477     -9.514   .00000000    Antitype
            2211        483.  695.563     -8.060   .00000000    Antitype
            2212        174.  368.461    -10.131   .00000000    Antitype
            2221        574.  468.921      4.853   .00000061    Type
            2222        806.  248.401     35.379   .00000000    Type


                        chi2 for CFA model = 3991.9562
                        df =    11      p =  .00000000

                    LR-chi2 for CFA model =  3478.5211
                        df =    11      p =  .00000000
```

Fig. 6.2 First order CFA with life satisfaction items using the von Eye program

Table 6.1 First order CFA with life satisfaction items using the R-package confreq

Pattern	Observed	Expected	loc.chi.square	loc.df	loc.chi.square.p	z.Chi	p.z.Chi
1 1 1 1	1,406	682.68	766.39	1.00	0.00	27.68	0.00
1 1 1 2	307	361.63	8.25	1.00	0.00	−2.87	0.00
1 1 2 1	167	460.23	186.83	1.00	0.00	−13.67	0.00
1 1 2 2	124	243.80	58.87	1.00	0.00	−7.67	0.00
1 2 1 1	299	626.96	171.55	1.00	0.00	−13.10	0.00
1 2 1 2	127	332.12	126.68	1.00	0.00	−11.26	0.00
1 2 2 1	356	422.67	10.52	1.00	0.00	−3.24	0.00
1 2 2 2	568	223.90	528.82	1.00	0.00	23.00	0.00
2 1 1 1	1,200	757.38	258.67	1.00	0.00	16.08	0.00
2 1 1 2	230	401.21	73.06	1.00	0.00	−8.55	0.00
2 1 2 1	140	510.59	268.98	1.00	0.00	−16.40	0.00
2 1 2 2	114	270.48	90.53	1.00	0.00	−9.51	0.00
2 2 1 1	483	695.56	64.96	1.00	0.00	−8.06	0.00
2 2 1 2	174	368.46	102.63	1.00	0.00	−10.13	0.00
2 2 2 1	574	468.92	23.55	1.00	0.00	4.85	0.00
2 2 2 2	806	248.40	1,251.67	1.00	0.00	35.38	0.00

measuring different aspects of life satisfaction. Subsequently, we split this total score into three equivalent categories using the 33rd and 66th percentiles. The resulting categories are 1 = low life satisfaction, 2 = medium life satisfaction, and 3 = high life satisfaction.

In case that we have many categorized variables and we define one variable as a dependent variable, a technique known as CHAID may be applied by using the SPSS module *Answer Tree* and the option *Exhaustive CHAID*. "…CHAID partitions the data into mutually exclusive, exhaustive, subsets that best describe the dependent variable" (Kass, 1980, p. 119). That means that we apply the CHAID-model (CHAID stands for **Ch**i-Square **A**utomatic **I**nteraction **D**etection) as a model for contrasting groups (cf. Lautsch & Plichta, 2003; Lautsch & Thöle, 2005). The use of this program results in a graphical illustration called a **tree diagram**. CHAID is a stepwise procedure; first the program searches for the best predictor of the dependent variable by partitioning the data. The chi-square statistic is used to pick the best predictor, similar to how the F-value is used in stepwise regression to decide which variable should be included or excluded (Kass, 1980). In our example the dependent variable is the categorized total score of Life Satisfaction based on the four items with the highest factor loadings (i.e., low, medium, and high). The independent variables are the dichotomized items measuring different aspects of Life Satisfaction. One advantage of this procedure is the ability to investigate the effects of several independent variables on a dependent variable without any restriction on the level of measurement. Basically, the procedure follows a step-by-step hierarchical bivariate analysis. First, the best predictor is identified, which reduces the largest amount of variance in the dependent variable. Then the next-best variable is introduced.

Instead of using a categorical independent variable it is also possible to use interval-level variables (cf. Haughton & Oulabi, 1997). *Answer Tree* also offers the option to optimize the level of measurement with each step as the variance of the dependent variable is reduced (e.g., by categorizing interval level variables or by reducing the number of categories of nominal or ordinal variables).[1] See Fig. 6.3 for the printout. The branches of each tree divide the dependent variable into separate groups and thus can be considered nodes or configurations. These configurations will be used to 'predict' the three categories of the total score. For these analyses we use SPSS again and the following syntax (see Box with Compute and IF Syntax).

```
COMPUTE Node = 0.
        if((incomecat = 1)and(workcat = 1)and(housecat = 1))Node = 1.
        if((incomecat = 1)and(workcat = 1)and(housecat = 2))Node = 2.
        if((incomecat = 1)and(workcat = 2)and(housecat = 1))Node = 3.
        if((incomecat = 1)and(workcat = 2)and(housecat = 2))Node = 4.
        if((incomecat = 2)and(workcat = 1)and(housecat = 1))Node = 5.
        if((incomecat = 2)and(workcat = 1)and(housecat = 2))Node = 6.
        if((incomecat = 2)and(workcat = 2)and(standardcat = 1))Node = 7.
        if((incomecat = 2)and(workcat = 2)and(standardcat = 2))Node = 8.


FORMATS Node(F1.0).
VARIABLE LABELS Node   'innerstructure of life satisfaction'.
VALUE LABELS Node
                  1”not at all satisfied(Node = 1)”
                  2”satisfied with appartment(Node = 2)”
                  3”statisfied with work place(Node = 3)”
                  4”satisfied with work and appartment(Node = 4)”
                  5”satisfied with income(Node5)”
                  6”satisfied with income and appartment(Node = 6)”
                  7”satisfied with income and work(Node = 7)”
                  8”satisfied with income, work and standard of living(Node = 8).”
FREQUENCIES VARIABLES = Node.
```

[1] Useful options in SPSS: allow only 30 subjects in parent nodes and 10 in child nodes. Uncheck the Bonferroni adjustment button.

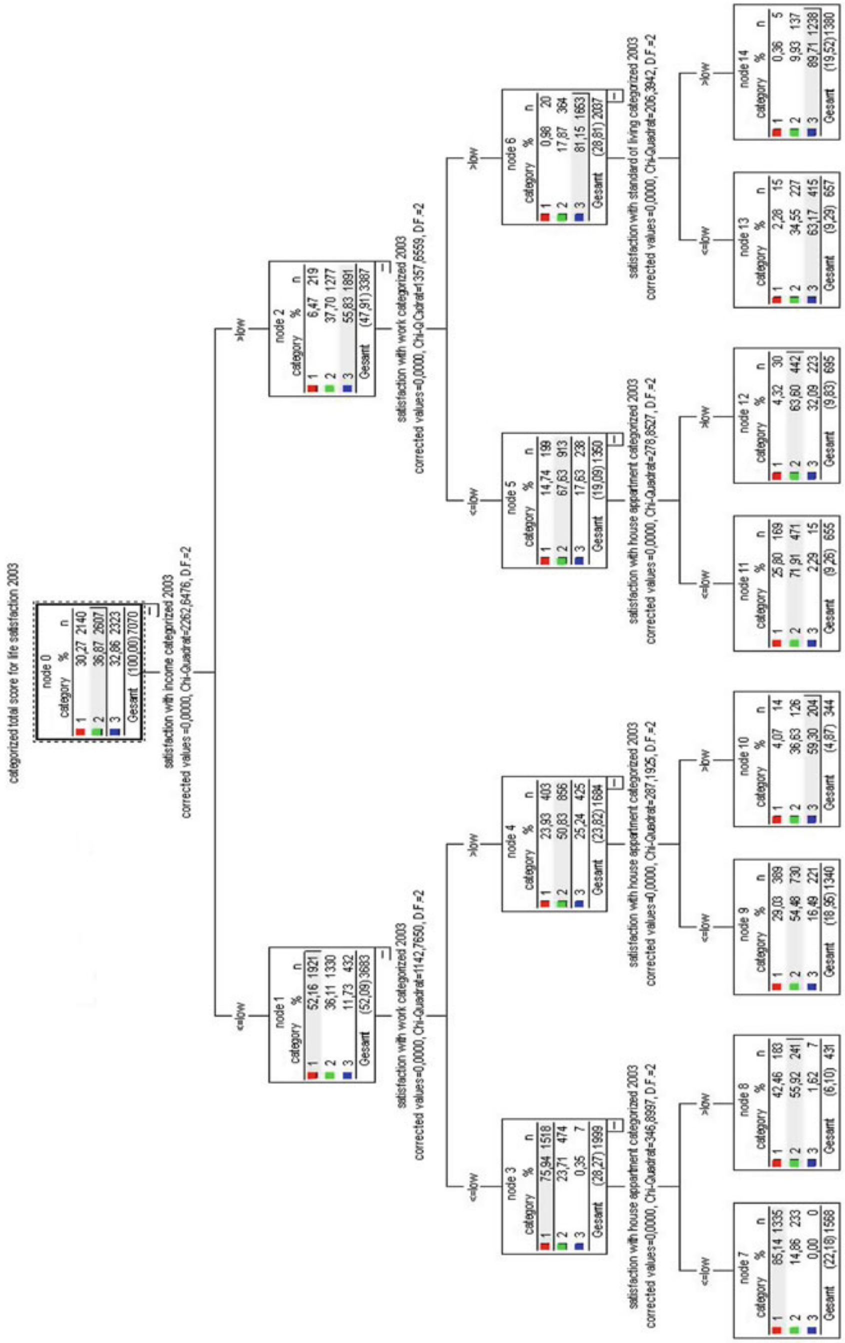
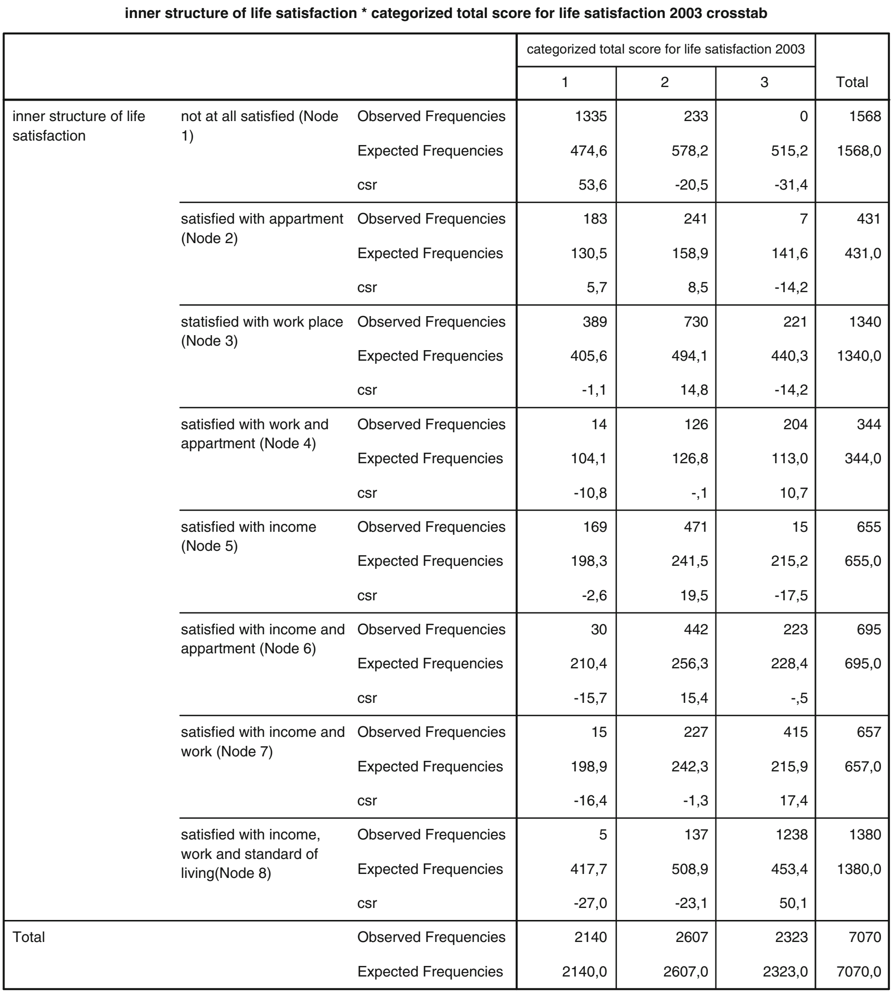


Fig. 6.3 Graphical illustration of the contrasting groups in AnswerTree. (Note. Here is a translation of the following German words: 'Gesamt' = total; 'Chi-Quadrat' = chi-square)



inner structure of life satisfaction * categorized total score for life satisfaction 2003 crosstab

			categorized total score for life satisfaction 2003			Total
			1	2	3	
inner structure of life satisfaction	not at all satisfied (Node 1)	Observed Frequencies	1335	233	0	1568
		Expected Frequencies	474,6	578,2	515,2	1568,0
		csr	53,6	-20,5	-31,4	
	satisfied with appartment (Node 2)	Observed Frequencies	183	241	7	431
		Expected Frequencies	130,5	158,9	141,6	431,0
		csr	5,7	8,5	-14,2	
	statisfied with work place (Node 3)	Observed Frequencies	389	730	221	1340
		Expected Frequencies	405,6	494,1	440,3	1340,0
		csr	-1,1	14,8	-14,2	
	satisfied with work and appartment (Node 4)	Observed Frequencies	14	126	204	344
		Expected Frequencies	104,1	126,8	113,0	344,0
		csr	-10,8	-,1	10,7	
	satisfied with income (Node 5)	Observed Frequencies	169	471	15	655
		Expected Frequencies	198,3	241,5	215,2	655,0
		csr	-2,6	19,5	-17,5	
	satisfied with income and appartment (Node 6)	Observed Frequencies	30	442	223	695
		Expected Frequencies	210,4	256,3	228,4	695,0
		csr	-15,7	15,4	-,5	
	satisfied with income and work (Node 7)	Observed Frequencies	15	227	415	657
		Expected Frequencies	198,9	242,3	215,9	657,0
		csr	-16,4	-1,3	17,4	
	satisfied with income, work and standard of living(Node 8)	Observed Frequencies	5	137	1238	1380
		Expected Frequencies	417,7	508,9	453,4	1380,0
		csr	-27,0	-23,1	50,1	
Total		Observed Frequencies	2140	2607	2323	7070
		Expected Frequencies	2140,0	2607,0	2323,0	7070,0

Fig. 6.4 Detecting the inner structure of life satisfaction using the corrected standardized residuals (csr)

By simply doing a Crosstab analysis in SPSS with the nodes as the row variable and the three categories of the *Life Satisfaction* as the three columns, we can identify a number of *types* by using the Fuchs-Kenett-Test or the corrected standardized residuals (see Eq. 4.19; see Fig. 6.4). We may interpret the findings as follows. First, we start with a trivial result: Those who rated each life satisfaction item in the lower median are considered overall low in their life satisfaction (Node 1). Being satisfied with your housing condition may still lead to a low overall life satisfaction or at the most to a medium life satisfaction (Node 2). Being content with your workplace results in at least medium life satisfaction (Node 3). Having two areas of high satisfaction, like work and housing situation results in high life satisfaction (Node 4). Being satisfied with income alone or income and one's housing situation lead to medium life satisfaction (Node 5 and Node 6). High life satisfaction occurs if one is satisfied with income and work (Node 7). Thus, one's satisfaction with their work place is of higher value than one's satisfaction with their housing condition. Of course, if one is highly satisfied with the work place, income, and standard of living, those subjects belong to the highest category for Life Satisfaction.

Summary: In case one has categorical data and one defines one variable as the dependent variable, variables predicting this dependent variable can be detected by using CHAID. The use of this program results in a graphical illustration called a **tree diagram**. CHAID is a stepwise procedure; first the program searches for the best predictor by partitioning the data. The chi-square statistic is used to pick the best predictor. The branches of each tree can now considered as nodes or as configurations which divide the dependent variable into separate groups. These configurations will be used to 'predict' the categories of the dependent variable. With the help of the Fuchs-Kenett statistic (i.e., CSR) *types* may be detected. Therefore, CHAID and CFA can be considered useful complimentary statistical tools.

6.2 Latent Class Analysis and CFA

Latent class analysis (LCA) is most often seen as an equivalent to factor analysis (FA). While FA extracts latent continuous factors from a pool of continuous variables, LCA extracts latent categorical factors or classes from a pool of categorical variables. Both statistical tools aim at data reduction. Although CFA and LCA are very similar in their approach, they are rarely performed in combination (cf. Lautsch & Plichta, 2003, 2005). From an applied point of view, they can in fact be considered as complementary statistical tools for type exploration and confirmation. CFA can be seen as a LCA on the manifest level, where all *types* are considered latent classes. The term latent class is preferred over the term cluster in this situation. For the analysis of typologies, LCA can be used as a probabilistic cluster analysis. Therefore, LCA is introduced and described below. The benefit of comparing the results of a LCA with the results of a CFA will be presented.

The most important aspect of LCA is *local (stochastic) independence*. This means that the subjects of a sample are divided into groups or classes, such that within a class the characteristics or variables of the subjects are independent (i.e., the chi-square value of the respective contingency table is zero). The LCA provides the following parameter estimates: (1) the *latent class probabilities*, that is the probability of belonging to a class, from which we can infer the class sizes, (2) the conditional probability of being a member of a class given a response set on variables called *class membership probabilities*, (3) the conditional probability of

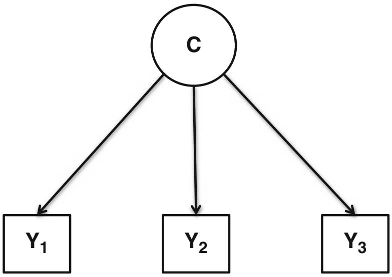


Fig. 6.5 Figural representation of a latent class

giving a response set on variables depending on the class membership (this is called *conditional response probabilities*), and (4) the probabilities of obtaining a *response pattern* i.e., responses on variables. The following picture shows a latent class solution with the latent class C and three (manifest) indicator variables (Fig. 6.5). The arrows from C to the variables can be seen as regressions coefficients, but they are response probabilities depending on the particular class. The indicators variables are binary or ordered categorical variables.

The following elaborations are taken from Vermunt and Magidson (2003). Let's call the latent class X and Y_l one of the L observed variables, where $1 \leq l \leq L$. In addition, let C be the number of latent classes and D_l the number of levels of Y_l. LCAs are indexed by $x, x = 1, 2, \ldots C$, and a particular response of Y_l by $y_l, y_l = 1, 2, \ldots, D_l$. Y represents a vector and y is used to refer to a complete response pattern.

The conditional response probabilities are used to interpret the structure of types defined by the latent class. The above-mentioned probabilities are used to calculate expected frequencies. A model-fit tests how well observed and expected frequencies match each other. Therefore, LCA is a probabilistic model. The observed frequency $f(o)_{ijk}$ of each cell are reproduced by the following formula:

$$f(o)_{y_1,y_2,y_3} = N \sum_{x=1}^{C} \delta_x \rho_{y_1/x} \rho_{y_2/x} \rho_{y_3/x} \tag{6.1}$$

where N is the sample size, δ is the latent class probability and $\rho_{y_1 y_2 y_3/x}$ represents the conditional probability of being a member of class X_i given a response 1 or 2 on variable Y_1,Y_2, and Y_3.

Let's look at the following data example by Lazarsfeld and Henry (1968) where $N = 1{,}000$ subjects need to solve questions or problems (i.e., A, B, and C). They either '1' = solved or '2' = did not solve the problems (Table 6.2).

The basic idea of LCA is that the probability of obtaining the observed configurations or the observed response patterns y, $P(Y = y)$, is a weighted average of the C conditional response probabilities P(Y = y—X = x); that is

Table 6.2 Data example by Lazarsfeld and Henry (1968)

Y_1	Y_2	Y_3	f_o	$P(X=1\|\mathbf{Y}=y)$	$P(X=2\|\mathbf{Y}=y)$
1	1	1	220	0.982	0.018
1	1	2	160	0.900	0.100
1	2	1	60	0.400	0.600
1	2	2	160	0.100	0.900
2	1	1	60	0.900	0.100
2	1	2	60	0.400	0.600
2	2	1	60	0.100	0.900
2	2	2	220	0.018	0.982

Table 6.3 Results of LCA based on the Lazarsfeld and Henry data

	X = 1 (Master)	X = 2 (Non-Master)
$P(X=x)$	0.50	0.50
$P(Y_1=1\|X=x)$	0.80	0.40
$P(Y_2=1\|X=x)$	0.90	0.10
$P(Y_3=1\|X=x)$	0.60	0.20

$$P(Y=y)=\sum_{x=1}^{C}P(X=x)P(Y=y|X=x). \tag{6.2}$$

$P(X=x)$ denotes the proportions of persons belonging to class x. The idea of local stochastic independence is manifested in the following formula:

$$P(Y=y|X=x)=\prod_{l=1}^{L}P(Y_l=y_l|X=x). \tag{6.3}$$

The conditional response probabilities can be used to name the classes, similar to how we use factor loadings to interpret the factors. Combining the two equations 6.2 and 6.3 we obtain the following model for $P(Y=y)$:

$$P(Y=y)=\sum_{x=1}^{C}P(X=x)\prod_{l=1}^{L}P(Y_l=y_l|X=x). \tag{6.4}$$

MPLUS reveals the following results (Table 6.3). In MPLUS the conditional response probabilities are listed in the data set called *2CLASS.DAT*. The results indicate that a two-class model fits the data perfectly (chi-square value is zero). The classes are equal in size (i.e., 0.50). The conditional probability for answering each of the three questions correctly (i.e., ‘1’) while belonging to Class 1 are $\rho_{y_1}=0.80$, $\rho_{y_2}=0.90$, and $\rho_{y_3}=0.60$. The complementary probability (adding up to one) indicate the probability of answering the three questions correctly, while belonging to Class 2. Obviously, Class 1 consists of knowledgable subjects (i.e., Masters),

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


   Marginal Frequencies
   --------------------
   Variable Frequencies
   -------- -----------
       1      600.   400.

       2      500.   500.

       3      400.   600.


 sample size N =     1000

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        220.  120.000       9.129   .00000000    Type
     112        160.  180.000      -1.491   .06801859
     121         60.  120.000      -5.477   .00000002    Antitype
     122        160.  180.000      -1.491   .06801859
     211         60.   80.000      -2.236   .01267362
     212         60.  120.000      -5.477   .00000002    Antitype
     221         60.   80.000      -2.236   .01267362
     222        220.  120.000       9.129   .00000000    Type


                chi2 for CFA model =  241.1111
                df =     4      p =  .00000000

            LR-chi2 for CFA model =   222.6193
                df =     4      p =  .00000000


                                      CARPE DIEM
```

Fig. 6.6 Results of CFA based on the Lazarsfeld and Henry data using the von Eye program

while Class 2 represents the opposite (i.e., Non-Masters). Let us enter the data into a CFA (see Fig. 6.6 and Table 6.4). By looking exclusively at *types* one can easily detect that the two class latent class structure is very well reproduced by the CFA. Each *type* represents one category of a latent class.

Table 6.4 Results of CFA based on the Lazarsfeld and Henry data using the R-package confreq

Pattern	Observed	Expected	loc.chi.square	loc.df	loc.chi.square.p	z.Chi	p.z.Chi
1 1 1	220	120.00	83.33	1.00	0.00	9.13	0.00
1 1 2	160	180.00	2.22	1.00	0.14	−1.49	0.07
1 2 1	60	120.00	30.00	1.00	0.00	−5.48	0.00
1 2 2	160	180.00	2.22	1.00	0.14	−1.49	0.07
2 1 1	60	80.00	5.00	1.00	0.03	−2.24	0.01
2 1 2	60	120.00	30.00	1.00	0.00	−5.48	0.00
2 2 1	60	80.00	5.00	1.00	0.03	−2.24	0.01
2 2 2	220	120.00	83.33	1.00	0.00	9.13	0.00

Table 6.5 Latent class proportions and corresponding configuration probabilities

Question	Response	$\rho_{y_1,y_2,y_3/1}$	$\rho_{y_1,y_2,y_3/2}$
1	1	**0.80**	0.40
	2	0.20	**0.60**
2	1	**0.90**	0.10
	2	0.10	**0.90**
3	1	**0.60**	0.20
	2	0.40	**0.80**

Listing the latent class proportions and the corresponding configuration probabilities stress the use of LCA in exploring and confirming *types* (Table 6.5). By taking the highest (bold) conditional probabilities (i.e. the response probability while belonging to a certain class) the two *types*, ‘111’ and ‘222’ can be perfectly reproduced. High conditional probabilities go along with many observed frequencies. Therefore, we look specifically at types, while comparing LCA with CFA. Here, the two detected types represent the two ends of one latent continuum. Usually CFA is more sensitive in finding *types* or latent classes than LCA. As in CFA the latent classes found in LCA include higher order interactions.

6.3 Correspondence Analysis and CFA

Another statistical tool that investigates the relationship between persons or objects in contingency tables is correspondence analysis (CA) (Borg & Groenen, 2005; Hair, Black, Babin, Anderson, & Tatham, 2006; Hair, Black, Babin, & Anderson, 2010). CA examines the relationships between categories of nominal data in a cross-tabulation based on their associations and CA presents the results in a graphical description called a perceptual map. In a perceptual map, persons or objects are plotted such that their proximity represent closeness or strong relationships (cf. Lautsch & Thöle, 2003). Sometimes CA is also referred to as homogeneity analysis (the respective program in SPSS is called HOMALS (HOMogeneity Analysis of

Table 6.6 Cross-tabulation of the Lazarsfeld and Henry data (1968)

		Y1			
		Solved		Not solved	
		Y2		Y2	
		Solved	Not solved	Solved	Not solved
Y3	Solved				
	o_{ijk}	220	60	60	60
	e_{ijk}	120	120	80	80
	d	−100	60	20	20
	signed χ^2	83.33	−30.00	−5.0	−5.0
Y3	Not solved				
	o_{ijk}	160	160	60	220
	e_{ijk}	180	180	120	120
	d	20	20	60	−100
	signed χ^2	−2.20	−2.20	−30.00	83.33
					$N = 1{,}000$

d difference between observed and expected frequency

Alternating Least Squares)). While CA and homogeneity analysis refer to the bivariate analysis of categorical variables, multiple correspondence analysis (MCA) stands for the multivariate analysis of nominal variables. Similar to factor analysis (FA), CA also aims at reducing the dimensionality. The extracted dimensions can be seen as latent factors or dimensions. The reported discrimination measures can be seen as equivalent to factor loadings in a factor analysis (Greenacre, 1989; Greenacre & Blasius, 1994). However, in contrast to CFA and LCA, MCA uses only bivariate associations, no higher-order information.

CA uses the chi-square statistic (see Eq. 1.11) based on the cell frequencies as a measure of similarity or association. The expected frequencies are calculated assuming independence between the variables (see Eq. 1.9). We use the data example from Lazarsfeld and Henry (1968) and we list the observed and expected frequencies as well as the Chi-Square statistics and the difference between the expected and observed frequencies in the table (see Table 6.6):

In CA, the chi-square values are the measures of proximity or association. The absolute value of the chi-square denotes the degree of association, but all chi-square values are positive and therefore, the direction of the similarity is removed. To restore the directionality, we add a sign to this statistic, but the reversed sign(!) of the difference scores between the expected and the observed frequencies. This has been done already in Table 6.6. Now, the positive values stand for greater association and the negative values for less association. In terms of CFA, one would say, that we are exclusively searching for *types*. The *types* in CFA will eventually come out as persons who are closely together in the perceptual map. The 'signed chi-square values' are used to create a latent space based on orthogonal dimensions upon which the categories of the variables involved can be placed to represent the strength of association through the nearness of the persons. In Fig. 6.7 the two *types* of Masters

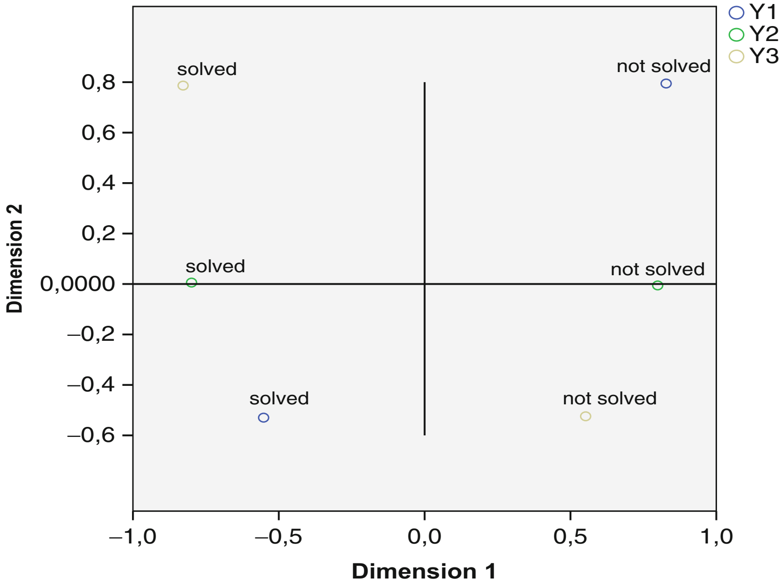


Fig. 6.7 Perceptual map from correspondence analysis based on the Lazarsfeld and Henry data

and Non-Masters are well represented on the left and the right side of the perceptual map. The respective SPSS-syntax for using more than two variables can be found in the next syntax box.

```
HOMALS
        /VARIABLES = Y1(2) Y2(2) Y3(2)
        /ANALYSIS = Y1 Y2 Y3
        /DIMENSION = 2
        /PRINT FREQ EIGEN DISCRIM QUANT
        /PLOT QUANT OBJECT NDIM(ALL, MAX)
        /MAXITER = 100
        /CONVERGENCE = .00001.
```

One of the disadvantages of CA is the indeterminacy of the number of dimensions. In the case of two categorical variables involved, the number of maximum possible dimensions is given through the smaller of the number of rows or columns minus one. For example, with five columns and four rows, the maximum number of dimensions would be three, which is four minus one. In the multivariate case, the maximum number of dimensions is

$$D_{max} = p - s \tag{6.5}$$

with p = number of categories and s = number of variables. However, usually a two dimensional perceptual map is pursued. As with similar methods, like in factor analysis, where we also have the problem of the indeterminacy of the number of factors, the 'researcher should balance interpretability versus parsimony of the data representation' (Hair et al., 2006, p. 673). A statistical measure that helps to decide whether an additional dimension is helpful is **inertia** (I). The total inertia for two categorical variables is calculated through

$$I_{tot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{\chi^2}{n}. \tag{6.6}$$

In the multivariate case, the maximum inertia is

$$I_{tot} = \frac{p}{s} - 1 \tag{6.7}$$

In real life, the maximum inertia is rarely reached; a measure for the obtained inertia is the sum of the eigenvalues. In our data example, the maximum inertia would be $8/3 - 1 = 1.667$. The printout in SPSS offers up to three dimensions with the following eigenvalues:

$$I_{dim} : I_1 = 0.518; I_2 = 0.278; I_3 = 0.205$$

The obtained inertia is the sum of eigenvalues:

$$I_{obtained} = \sum_{i=1}^{dim} E_i = 0.518 + 0.278 + 0.205 = 1.00$$

The eigenvalues divided by the obtained inertia times 100 % indicate the amount of variance explained by one dimension. Therefore, the three maximum possible dimensions in our case would explain $d_1 = 51.8\,\%$, $d_2 = 27.8\,\%$ and $d_3 = 20.5\,\%$ of the total variance.

Summary: CFA uses the information of categorial variables in multiway contingency tables. There are related multivariate statistical tools like Answer Tree (CHAID), Latent Class Analysis (LCA) and Correspondence Analysis (CA) that are also based on categorical variables. All mentioned statistical tools are similar because they apply the chi-square statistic to calculate associations or similarities. CFA may be employed in combination with these statistical tools in terms of *type* exploration and confirmation. CHAID investigates the underlying structure of the identified types in terms of independent variables. In LCA, the obtained latent factors or classes may be interpreted as *types* and in CA the *types* can be presented in a graphical manor.

References

Borg, I., & Groenen, P. J. K. (2005). *Modern multidimensional scaling – Theory and applications* (2nd ed.). New York: Springer.

Greenacre, M. J. (1989). *Theory and applications of correspondence analysis* (3rd printing). London: Academic.

Greenacre, M. J., & Blasius, J. (1994). *Correspondence analysis in the social sciences: Recent developments and applications*. London: Academic.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Haughton, D., & Oulabi, S. (1997). Direct marketing modeling with CART and CHAID. *Journal of Direct Marketing, 11*(4), 42–52.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Lautsch, E., & Plichta, M. (2003). Configural frequency analysis (CFA), multiple correspondence analysis (MCA) and latent class analysis (LCA): An empirical comparison. *Psychology Science, 45*(2), 298–323.

Lautsch, E., & Plichta, M. M. (2005). Configural frequency analysis (CFA) and latent class analysis (LCA): Are the outcomes complementary? *Psychology Science, 45*(3/4), 424–430.

Lautsch, E., & Thöle, U. (2003). Classification and explanation of life conceptions using the case of the 14th shell youth study 2002. *Psychology Science, 45*(2), 263–279.

Lautsch, E., & Thöle, U. (2005). Identification and analysis of types of personality in a brief measure of the Big-Five personality domains. *Psychology Science, 47*(3/4), 479–500.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghthon Mifflin.

Vermunt, J. K., & Magidson, J. (2003) *Latent GOLD choice 3.0. User's guide*. Belmont, MA: Statistical Innovations.

Chapter 7
CFA and Its Derivatives

Abstract This chapter introduces several derivatives of CFA that can be used for different purposes. First, there is **Prediction CFA (P-CFA)**. This version of CFA is comparable to multiple regression. One variable is defines as the dependent variable or criterion, which is usually measured with a certain time lag with regard to the other independent variables or predictors. Second, there is **Interaction Structure Analysis (ISA)**. ISA uses an extended definition of interactions, which cannot be analyzed with log-linear modeling. Instead of searching for singular *types* or *antitypes*, one can search for *biprediction* types by looking for regional instead of local contingency. Finally, **two-sample CFA** is introduced as a very useful statistical tool similar to t-tests for independent samples. This derivative of CFA searches for *types* which differentiate the two samples under investigation, so called *discrimination* types.

7.1 Prediction-CFA

CFA searches for *types* and *antitypes* which represent significant deviations from the null hypothesis. Normally, the null hypothesis is the assumption of independence between the variables involved. That is, in terms of a log-linear model, the null hypothesis represents the main effects model. However, if we change the null hypothesis, we can test a number of different models. For example, in prediction CFA (P-CFA; cf. Lösel & Stemmler, 2012; Stemmler & Lösel, 2012; Stemmler, Lösel, Beelmann, & Jaursch, 2008; von Eye, 2002) predictors are assumed to be independent from the criterion. Let's say the variables A and B are the predictors and variable C is the criterion. The respective log-linear model, which represents the underlying null hypothesis, is a model which is saturated within each set of predictors and criterions. It looks like the following:

$$ln\, e_{ijk} = \lambda_0 + \lambda_i A_i + \lambda_j B_j + \lambda_k C_k + \lambda_{ij} AB_{ij}. \tag{7.1}$$

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_7,

Table 7.1 Prediction CFA for the predictors Gender, Externalizing and Internalizing Problems in kindergarten and intensive behavior problems in the classroom as the criterion

Cell index				Prediction CFA		
Ge	Ex	In	Cb	$f(o_{ijk})$	$f(e_{ijk})$	z_{ijkl}
1	−	−	−	98	99.83	−0.183
1	−	−	+	21	19.17	0.418
1	−	+	−	29	31.04	−0.366
1	−	+	+	8	5.96	0.835
1	+	−	−	31	37.75	−1.098
1	+	−	+	14	7.25	2.507
1	+	+	−	12	16.78	−1.166
1	+	+	+	8	3.22	2.662
2	−	−	−	138	124.16	1.242
2	−	−	+	10	23.84	−2.834
2	−	+	−	39	35.23	0.634
2	−	+	+	3	6.77	−1.447
2	+	−	−	18	20.13	−0.475
2	+	−	+	6	3.87	1.085
2	+	+	−	10	10.07	−0.921
2	+	+	+	2	1.93	0.048

$ln\, e_{ijk}$ is the natural logarithm of the expected frequencies, λ_0 is the intercept, λ_i is the parameter of variable A, λ_j is the parameter of variable B, and $\lambda_{ij}AB_{ij}$ represents the interaction between the predictors. The lambda parameter can be interpreted similarly to beta weights in a regression equation. If the base model does not fit, there must be an interaction between the predictor and the criterion. Let's have a look at the following data from the Erlangen-Nuremberg Development and Prevention Study (cf. Stemmler, Lösel, Beelmann, Jaursch, & Zenkert, 2005). The predictors are Ge = gender (1 = boys, 2 = girls), Ex = externalizing behavior, and In=internalizing behavior rated by kindergarten teachers ('+' is behavior problems above the 75th percentile, '−' behavior problems below the 75th percentile). The criterion was Cb = classroom behavior ('+' = three and more behavior problems mentioned in the school report cards, '−' less than three behavior problems mentioned). The following data evolved (Table 7.1) The prediction CFA may also be obtained through the R-package ***Confreq*** by indicating the interaction in addition to the main effects in the CFA command. Note that the set of predictors need to saturated in the model (Table 7.2):

```
CFA(patternfreq_neu,form="~ A + B + C + D +
 A:B + A:C + B:C + A:B:C")
```

Let us have a look at the obtained results: The base model for the prediction CFA revealed no satisfactory fit (LR = 29.77, df = 7, $p < 0.001$). This indicates that there were differences between the observed and estimated frequencies. Because we are

Table 7.2 Results of a prediction CFA using the R-package confreq

Pattern	Observed	Expected	chi.square	chi.square.p	z.Chi	p.z.Chi
1 1 1 1	98	99.83	0.03	0.85	−0.18	0.43
1 1 1 2	21	19.17	0.18	0.68	0.42	0.34
1 1 2 1	29	31.04	0.13	0.71	−0.37	0.36
1 1 2 2	8	5.96	0.70	0.40	0.84	0.20 0
1 2 1 1	31	37.75	1.21	0.27	−1.10	0.14
1 2 1 2	14	7.25	6.29	0.01	2.51	0.01
1 2 2 1	12	16.78	1.36	0.24	−1.17	0.12
1 2 2 2	8	3.22	7.09	0.01	2.66	0.00
2 1 1 1	138	124.16	1.54	0.21	1.24	0.11
2 1 1 2	10	23.84	8.03	0.00	−2.83	0.00
2 1 2 1	39	35.23	0.40	0.53	0.63	0.26
2 1 2 2	3	6.77	2.10	0.15	−1.45	0.07
2 2 1 1	18	20.13	0.23	0.63	−0.48	0.32
2 2 1 2	6	3.87	1.18	0.28	1.09	0.14
2 2 2 1	10	10.07	0.00	0.98	−0.02	0.49
2 2 2 2	2	1.93	0.00	0.96	0.05	0.48

only interested in *types* we use the one-tailed Bonferroni adjustment $\alpha^* = 0.05/8 = 0.0065$ which corresponds to a z-value of $|2.48|$. We identify two *types*. The configuration '1 + − +' indicates that there were more boys than expected under the null model who had serious classroom behavior problems. This configuration made up 6.2 % of the boys' sample. The second significant *type* '1 + + +' shows that there were more boys than expected who had serious behavior problems in the first grade and who were high in both Externalizing and Internalizing Problems in kindergarten. This *type* contained 3.6 % of the male group.

7.2 Interaction Structure Analysis (ISA)

Interaction Structure Analysis (ISA) (i.e., grouping variables into two groups) uses an extended definition of interactions (cf. Stemmler, 2000). Traditionally, with three variables, let's say A, B, and C, only one interaction of the second order is possible, that is A by B by C (i.e., ABC). The definition for interactions in ISA goes back to Lancaster (1969) who argues that with T variables there is more than just one T-order interaction possible. With three variables, there would also be three second order interactions, (i.e. A.BC, B.AC, and C.AB). The decimal point divides the variables into two groups: A.BC means that there is an interaction between variable A and B and C lumped together. For further rules of interactions in ISA we postulate (see von Eye, 1990; p. 83)

- "If there is no second order interaction and no first order interaction (defined as associations), the three variables are totally independent.

Table 7.3 Survival time of rats ('−' short, '=' medium, and '+' long) were investigated while receiving two kinds of conditions

	Condition two							
Condition one	A		B		C		D	
	31	=	82	+	43	=	45	=
I	45	=	110	+	45	=	71	=
	46	=	88	+	63	=	66	=
	43	=	72	+	76	+	62	=
	36	=	92	+	44	=	56	=
II	29	=	61	=	35	=	102	+
	40	=	49	=	31	=	71	=
	23	−	124	+	40	=	38	=
	22	−	30	=	23	−	30	=
III	21	−	37	=	25	=	36	=
	18	−	38	=	24	−	31	=
	23	−	29	=	22	−	33	=

- If there are no first order interactions, there may be second order interactions nevertheless.
- The existence of second order interactions follows from the existence of first order interactions." For instance, if with three variables A.B holds true, also does A.BC.

With T variables $r_T = \frac{1}{2}(3^T + 1) - 2^T$ interactions evolve; that is, with T = 3 variables $r_3 = 6$, with T = 4 variables $r_4 = 25$ interactions, and with T = 5 variables $r_5 = 90$. Therefore, with four and more variables ISA becomes time consuming. One alternative is to group the variables by a rationale. For example, in Stemmler (1994) $N = 48$ laboratory rats were given two treatments in a search for appropriate antidotes. The first condition consisted of three different poisons (i.e., I, II, and III), and the second condition consisted of four different antidotes (i.e., A, B. C, and D). The dependent variable was survival time (see Table 7.3). Table 7.3 can be transformed into Table 7.4 by grouping the data according to the survival times (i.e., −, =, and +). Now the tripartite survival time functions as the response variable in an ISA with c = 3 response classes. As Table 7.4 shows, poison I in combination with antidote B leads to prolonged survival times (i.e., $0_{21} = 4$). This is a treatment-response *type* according to a H_1 which is a local alternative to the H_0 of no treatment effects. We use the Fuchs-Kenett-Test to calculated the corrected standardized residuals for $e_{21} = \frac{(4 \times 8)}{48} = 0.66$

$$csr_{21} = \frac{(4 - 0.66)}{\sqrt{0.66 \times \left(1 - \frac{4}{48}\right) \times \left(1 - \frac{8}{48}\right)}} = 4.704$$

Table 7.4 Survival time of rats ('−' short, '=' medium, and '+' long) were investigated while receiving two kinds of treatment

		Response classes			Row marginals
		f+	f=	f−	
	A	0	4	0	4
I	B	$4^*(=a)$	0	0(=b)	4(=A)
	C	1	3	0	4
	D	0	4	0	4
	A	0	3	1	4
II	B	2	2	0	4
	C	0	4	0	4
	D	1	3	4	4
	A	0(=c)	0	$4^*(=d)$	4(=B)
III	B	0	4	0	4
	C	0	1	3	4
	D	0	4	0	4
Column marginals		8(=C)	32	8(=D)	N = 48

Table 7.5 Collapsed fourfold table for Fisher's exact test

	f+	Others	Row marginals
I/B	$a = 4^*$	b = 0	A = 4
Others	c = 4	d = 40	B = 44
Columns marginals	C = 8	D = 40	N = 48

The calculated z-statistic far exceeds the two-tailed Bonferroni limit $0.025/36 = 3.279$. Thus the treatment combination I/B is most effective with regard to the survival prolongation. The asymptotic test by Fuchs and Kenett (1980) may not be valid because of small expected frequencies ($e_{ij} < 5$), and the **Fisher's exact test** (cf. Siegel & Castellan, 1988; von Eye, 2002) may be more appropriate. However, the Table 7.4 needs to be transformed into a fourfold table as shown in Table 7.5. The one-sided tail probability p for the treatment-response *type* I/B can be calculated as follows

$$p(I/B) = \frac{A!B!C!D!}{N!a!b!c!d!} = \frac{4!44!8!40!}{48!4!0!4!40!} = 0.0003597 \tag{7.2}$$

The two-sided probability $p = 2 \times (0.00036) = 0.00072$ is smaller than the Bonferroni alpha adjustment $0.025/36 = 0.000694$. Thus the I/B response type has also been verified by Fisher's exact test. Fisher's test is exact; no assumptions need to be made concerning an approximation of a test statistic to a sampling distribution. It is cumbersome to calculate. Therefore, it is rarely an option in CFA programs.

7.3 Biprediction-Type

In most cases, a singular *type* does not occur, but rather there are most often either no *types* at all or two *types* combined with two *antitypes*. In such a case a favorable treatment (T+) is usually compared with an unfavorable one (T−), if the response variable is dichotomized at the sample median (to become either X+ or X−). In such a case, the fourfold table is usually overfrequented in cells *a* and *d* (i.e., *types*) and underfrequented in cells *b* and *c* (i.e., *antitypes*). The *types* of *a* and *d* and the *antitypes* of *b* and *c* may be considered to define an *interaction type* in terms of an ISA. The most common term, however, is **biprediction type** (Lienert & Netter, 1987; Stemmler, 1994).

Under the assumption that a treatment response table of *r* treatment combinations and *c* response configurations implies a treatment-response fourfold table with 2 *types* and 2 *antitypes*, the global chi-square equation (see Eq. 1.3) may be decomposed into an orthogonal fourfold component with $df = 1$ and a residual component with $(r-1)(c-1)-1$ degrees of freedom. An easy way to decompose this formula was suggested by Kimball (1954); this decomposition was called **regional contingency** by Havránek and Lienert (1984). In terms of a z-statistic, Kimball's interaction type between two treatment modalities and two response classes can be calculated by

$$z = \frac{A(Cd-Dc)-B(Cb-Da)}{\sqrt{\frac{ABCD(A+B)(C+D)}{N}}} \tag{7.3}$$

For Eq. 7.3 any $r\,x\,c$ table needs to be collapsed to a $3x3$ table Table 7.6 The treatment-response *type* I/B f+ with the frequency a^* and the III/A f− *type* with the frequency d^* have been called biprediction types because a^* and d^* are predicted simultaneously. In the context of ISA the term *interaction type* is preferred. Let's apply Kimball's decomposition test to Table 7.4

$$z = \frac{4(8\times 4-8\times 0)-4(8\times 0-8\times 4)}{\sqrt{\frac{4\times 4\times 8\times 8(4+4)(8+8)}{48}}} = 4.92$$

For the $r = (3\times 4) = 12$ treatment combinations and $c = 2$ extreme response (i.e., f+ and f−) there are

Table 7.6 Application of Kimball's chi-square decomposition

	f+	f−	Others	Row marginals
I/B	a^*	b	x	A
III/A	c	d^*	x	B
Other	x	x	x	x
Columns marginals	C	D	x	N

$$R = \binom{12}{2}\binom{2}{1} = 132$$

fourfold associations to be tested simultaneously. The 5 % Bonferroni limit is therefore $z(0.05/132) = z(0.000379) = 3.37$. Since $z = 4.92$ exceeds this limit we can rely on the existence of the explored types.

7.4 Two-Sample CFA

In the following, the two-sample CFA (Stemmler & Bingham, 2004; Stemmler & Hammond, 1997; von Eye, 2002) is illustrated. The two-sample CFA is comparable to the t-test in parametric statistical analysis. However, methods exist for the comparison of three or more groups. The underlying assumption representing the null hypothesis is that the two samples were drawn from the same population. The same expected frequencies for each group configuration applies, and deviations from the frequency distribution should only be random. In other words, let's say A and B are variables characterizing the grouping variable C, then the contingency tables of A and B together need to be homogenous across C. The underlying null hypothesis is $H_0 : \pi_{ABC} = \pi_{AB}\pi_C$ and $H_1 : \pi_{ABC} \neq \pi_{AB}\pi_C$.

The following data example is taken from Lienert (1978, p. 978). A pretest-posttest treatment design is used to assess improvement in school performance (i.e., reading ability) in a sample of students ($N = 36$) suffering from dyslexia. The students were randomized according to either a treatment group ($N_T = 19$) or a waiting-list control group ($N_C = 18$). In addition, the teachers and the students rated the students' performance on reading ability occurring between the pre- and post-test. Teachers rated whether the students' ability has '−' *decreased*, '0' *not changed*, or '+' *improved*. Students rated whether they felt that their reading had '+' *improved* or had '=' *not improved*. Combining these self-assigned ratings with the teacher-assigned ratings resulted in the following three by two by two table (see Table 7.7). Let's type the data into von Eye's CFA program and use the *Two-sample CFA = 20*-option as one of the mentioned CFA models. Instead of using zeros it is better to type in a non-zero integer like 0.01, in order to get the correct degrees of freedom. The program rounds the data mathematically, such that if you type in 0.5, the actual integer listed will be 1.0. Two-sample CFA does not differentiate between *types* and *antitypes*; instead **discrimination types** (see page 55) comparable to the longitudinal CFA are listed. For the detection of discrimination types in two-sample CFA, the original tables needs to be rewritten, because each configuration is compared across the two groups (cf. von Eye, 2002; see Table 7.8). Testing the patterns '+ =' for teachers' and students' ratings, the following table results (see Table 7.9). For the detection of *types* several statistical tests are available. First, the exact Fisher's test calculates the probability of a cell frequency as follows (see Eq. 7.2)

Table 7.7 Two-sample CFA: listing of self-ratings and teacher ratings of dyslexic children

Teacher	Student	Group	f(o)	f(e)
+	+	T	6	5.00
+	+	C	4	5.00
+	=	T	3	8.50
+	=	C	14	8.50
0	+	T	4	2.00
0	+	C	0	2.00
0	=	T	2	1.00
0	=	C	0	1.00
−	+	T	0	0.00
−	+	C	0	0.00
−	=	T	3	1.50
−	=	C	0	1.50

Table 7.8 Two-by-two cross-classification for two-sample CFA testing

Configuration	Groups		Row
P_1P_2	A	B	Totals
ij	$a = N_{ijA}$	$b = N_{ijB}$	$A = N_{ij}$
All others combined	$c = N_{..A} - N_{ijA}$	$d = N_{..B} - N_{ijB}$	$B = N - N_{ij}$
Columns total	$C = N_{..A}$	$D = N_{..B}$	N

Table 7.9 Testing the configuration '+ =' against all other patterns

Configuration	Groups		Row
	Treatment group	Control group	Totals
+ =	a = 3	b = 14	A = 17
All others combined	c = 15	d = 4	B = 19
Columns total	C = 18	D = 18	N = 36

$$p(a) = \frac{A!B!C!D!}{N!a!b!c!d!} =$$

$$p(a) = \frac{17!19!18!18!}{36!3!14!15!4!} = 0.000305$$

Using the data from Table 7.9 this test is significant even after using the Bonferroni alpha adjustment $0.05/6 = 0.0083$. Next comes the traditional χ^2-test with $df = 1$:

$$\chi^2 = \frac{N(a \times d - b \times c)^2}{ABCD} = \tag{7.4}$$

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


    Marginal Frequencies
    --------------------
    Variable Frequencies
    -------- -----------
        1      27.    6.    3.

        2      14.   22.

        3      18.   18.


   sample size N =   37.
           The chi2 test will be performed with continuity correction


 Bonferroni-adjusted alpha =  .0083333

                                 Table of results
                                 ----------------
Configuration   f     statistic        p       pi*   Type?
------------- -----   ---------     -------  ------  -----


      111        6.
      112        4.       .161      .687945    .174
   ---------------------------------------------------------------------
      121        3.
      122       14.     10.706      .001068    .388  Discrimination Type
   ---------------------------------------------------------------------
      211        4.
      212        0.      2.365      .124087    .492
   ---------------------------------------------------------------------
      221        2.
      222        0.       .425      .514385    .479
   ---------------------------------------------------------------------
      311        0.
      312        0.      5.011      .025181    .008
   ---------------------------------------------------------------------
      321        3.
      322        0.      1.307      .252930    .488
   ---------------------------------------------------------------------
```

Fig. 7.1 Printout of CFA program for the twosample-CFA and the data in Table 7.7

$$\chi^2 = \frac{36(3\times 4 - 14\times 15)^2}{17\times 19\times 18\times 18} = 13.486$$

The resulting chi-square value is also highly significant, because the corresponding Bonferroni adjusted χ^2-value is $|5.76|$. The traditional chi-square test works best if the sample size is large. When the sample size is not large continuity correction is recommended:

$$\chi^2 = \frac{N|a \times d - b \times c| - 0.5 \times N^2}{ABCD} = \tag{7.5}$$

$$\chi^2 = \frac{36|3 \times 4 - 14 \times 15| - 0.5 \times 36)^2}{17 \times 19 \times 18 \times 18} = 11.1455$$

Let's have a look at the print-out of the CFA-program (see Fig. 7.1). The corresponding chi-square value with the continuity correction is slightly different (11.145 versus 10.706), because von Eye's program used $N = 37$, based on the inserted non-zero integers which we used for all zero cells, instead of the original $N = 36$. Otherwise the values would match perfectly.

References

Fuchs, C., & Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association, 75*, 395–398.

Havránek, T., & Lienert, G. A. (1984). Local and regional versus global contingency testing. *Biometrical Journal, 26*, 483–494.

Kimball, A. W. (1954). Short-cut formulas for the exact partitioning of chi-square in contingency tables. *Biometrics, 10*, 452–458.

Lancaster. H. O. (1969), The Chi-Square Distribution. New York: John Wiley & Sons. Inc.

Lienert, G. A. (1978). *Verteilungsfreie Methoden in der Biostatistik (Band II)* [Non-parametrical methods in the field of biometrics (Vol. II)]. Meisenheim am Glan, Germany: Hain.

Lienert, G. A., & Netter, P. (1987). Nonparametric analysis of treatment-reponse tables by bipredictive configural frequency analysis. *Methods of Information in Medicine, 26*, 89–92.

Lösel, F., & Stemmler, M. (2012). Continuity and patterns of externalizing and internalizing behavior problems in girls: A variable- and person-oriented study from preschool to youth age. *Psychological Tests and Assessment Modeling, 54*(3), 308–319.

Siegel, S., & Castellan, J. R. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Stemmler, M. (1994). A nonparametrical evaluation of ANOVA and MANOVA designs using interaction structure analysis. *Biometrical Journal, 36*(8), 911–925.

Stemmler, M. (2000). A foreward method for interaction struture analysis (ISA): Application in ANOVA design. *Psychology Science, 42*(3), 494–503.

Stemmler, M., & Bingham, C. R. (2004). Nonparametric testing of improvement scores in a two sample prepost design. *Psychology Science, 45*, 208–216.

Stemmler, M., & Hammond, S. (1997). Configural frequency analysis of dependent samples for intra-patient treatment comparisons. *Studia Psychologica, 39*, 167–175.

Stemmler, M., & Lösel, F. (2012). The stability of externalizing behavior in boys from preschool age to adolescence: A person-oriented analysis. *Psychological Tests and Assessment Modeling, 54*(2), 195–207.

Stemmler, M., Lösel, F., Beelmann, A., & Jaursch, S. (2008). A configural perspective on the stability of externalizing problem behavior in children: Results from the Erlangen-Nuremberg development and prevention study. In M. Stemmler, E. Lautsch, & D. Martinke (Eds.), *Configural frequency analysis and other non-parametrical methods: A Gustav A. Lienert memorial issue* (pp. 70–83). Lengerich, Germany: Pabst Publishing Company.

Stemmler, M., Lösel, F., Beelmann, A., Jaursch, S., & Zenkert, B. (2005). Child problem behavior in kindergarten and in primary school: A comparison between prediction configural frequency analysis and multiple regression. *Psychology Science, 47*(3/4), 467–478.

von Eye, A. (1990). *Introduction to configural frequency analysis: The search for types and antitypes in cross-classifications*. Cambridge, UK: Cambridge University Press.

von Eye, A. (2002). *Configural frequency analysis: Methods, models and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Errata

Person-Centered Methods

Mark Stemmler

Institute of Psychology, Friedrich-Alexander University of Erlangen-Nuremberg (FAU), Erlangen, Germany

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7, © Springer International Publishing Switzerland 2014

DOI 978-3-319-05536-7_8

The paperback and online versions of the book contain some errors, and the corrections to these versions are given on the following pages.

The online version of the original book can be found at
http://dx.doi.org/10.1007/978-3-319-05536-7

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7_8,
© Springer International Publishing Switzerland 2014

Frontmatter

Page IV, Sixth line from the top: The copyright holder of this book should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

1
Introducing Person-Centered Methods

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

2
CFA Software

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

3
Significance Testing in CFA

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

4
CFA and Log-Linear Modeling

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

5
Longitudinal CFA

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

6
Other Person-Centered Methods Serving as Complimentary Tools to CFA

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

7
CFA and Its Derivatives

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

Glossary

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

Index

The copyright holder of this chapter should be changed from Springer International Publishing Switzerland 2014 to The Author(s) 2014.

Glossary

Antitype Represents or indicates an under-frequented cell ($f_{(o)} < f_{(e)}$).

Base Model Is the underlying model to calculate the expected frequencies; it usually the independence model. Different base models result in different types or *antitypes*.

Bonferroni Alpha Adjustment In case of multiple test procedures a Bonferroni alpha adjustment is necessary. It divides the alpha level by the number of tests, e.g., with a two-sided test of $\alpha = 0.05$ and three tests, the new alpha level would be $\alpha^* = 0.025/3 = 0.00833$.

Chi-Square Automatic Interaction Detection (CHAID) In case that we have many categorized variables and we define one variable as a dependent. CHAID partitions the data into mutually exclusive, exhaustive, subsets that best describes the dependent variable. One applies a CHAID-model for contrasting groups. The use of this program results in a graphical illustration called a tree diagram.

Coefficient of Precision Is a coefficient which can be interpreted similar to the determination coefficient R^2 in multiple regression. This is the statistic Q, which is a coefficient of precision or a coefficient of *the pregnancy of a type*.

Configural Cluster Analysis (CCA) Is a zero-order CFA where the underlying model includes no main effects or interactions; that is, each cell has the same expected frequency.

Configural Frequency Analysis (CFA) Is a statistical method that looks for over- and under-frequented cells or patterns in a contingency table. Over-frequented means, that the observations in this cell or configuration are observed more often than expected, under-frequented means that this configurations is observed less often than expected.

Confreq An R-package using Configural Frequency Analysis and the log-linear modeling approach for analyzing contingency tables.

Correspondence Analysis (CA) Is another statistical tool that investigates the relationship between persons or objects in contingency tables. CA examines the relationships between categories of nominal data in a cross-tabulation based on their associations and CA presents the results in a graphical description called

DOI 10.1007/978-3-319-05536-7, © Springer International Publishing Switzerland 2014

a perceptual map. In a perceptual map, persons or objects are plotted such that their proximity represent closeness or strong relationships. Sometimes CA is also referred to as homogeneity analysis.

Design Matrix X a is design matrix containing the effect-coded main effect and interaction terms plus the constant. The design matrix X has as many rows as there are cells or configurations, and $m+1$ columns. m is the number of weights; the first weight is always the constant, coded with ones.

Discrimination Type Is a type that differentiates significantly between two (originally) independent samples.

Global Chi-square Is a statistic referring to the whole contingency table. It is the sum of all cell-wise deviations between observed and expected frequencies using the chi-square statistic.

Local Chi-square Is a statistic referring to one cell or configuration, where it represents the calculated deviation between the observed and expected frequencies using the chi-square statistic.

Lancaster decomposition Lancaster (1969) found out, that the global chi-square of a first order CFA is composed additively of all possible interactions.

Log-linear modeling Is a statistical tool to investigate the underlying structure of dependency in a contingency table. The logarithm of the expected frequencies can be expressed in a linear function of parameters. The parameters indicate the impact of main effects and interactions on the data in the contingency table.

Latent Class Analysis; LCA Is most often seen as an equivalent to factor analysis (FA). While FA extracts latent continuous factors from a pool of continuous variables, LCA extracts latent categorical factors or classes from a pool of categorical variables. Both statistical tools aim at data reduction.

Longitudinal CFA This version of CFA tests the stability or instability of configurations over time. Configurations define observations of patterns of one sample over time.

Meehl's paradox A data example constructed by Meehl where there are no bivariate associations or correlations but higher order associations which allow the exact prediction of a group membership in a 2 by 2 by 2 contingency table.

Quasi-Independence Means, that after blanking out a certain cell, the remaining contingency table has to be independent and therefore, the respective chi-square must not be significant.

R Software R is an open source software which is suitable for Linux, MacOS X, and Windows. R (R Development Core Team, 2011) is a program for data analysis, data manipulation and graphical display.

Saturated Model Is a model that reproduces the observed values perfectly. It includes all main and interaction effects.

Structural Zeros Are usually cells which cannot be observed (e.g., a pattern of heavy rain together with a beautiful blue sky).

Test of Marginal Homogeneity This table searches for equal probabilities (i.e., $p_{i.} = p_{.i}$) and therefore equal pairs of marginals (i.e., $f_{i.} =$) in a square symmetric contingency table.

Type Represents or indicates an over-frequented cell ($f_{(o)} > f_{(e)}$).

Index

M. Stemmler, *Person-Centered Methods*, SpringerBriefs in Statistics,
DOI 10.1007/978-3-319-05536-7, © Springer International Publishing Switzerland 2014