Psychology Science, Volume 45, 2003 (2), p. 421-429

Log-linear modeling and two-sample CFA in the search of discrimination types

Mark Stemmler[1], C. Raymond Bingham[2]

Summary

Two-sample configural frequency (CFA) is suggested as a useful statistical tool to compare data from pretest-posttest-designs. The investigated data may be difference or improvement scores. The above procedure is recommended because improvement scores from two dependent samples, although metrically scored, are usually non-normally distributed and therefore not suitable for parametric comparisons. The two-sample CFA is compared to log-linear modeling (LLM); the similarities and dissimilarites between the two statistical methods are presented. LLM takes a model fitting approach, that is LLM tests the goodness-of-fit of a null model, which assumes no interactions between the sample or grouping variable and the outcome variables. Instead of a global approach as used by LLM, CFA takes a local or cell level approach, searching for differences between the hypothesized (null) model and the empirical data. The Fisher-Yates test is introduced as a statistic to test for cell patterns or configurations which discriminate between the two samples under investigation. Real data from educational psychological research is used to demonstrate univariate and bivariate two-sample comparisons.

Key words: Two-sample comparison, Configural Frequency Analysis (CFA), two-sample CFA, log-linear modeling (LLM), nonparametric testing, contingency table analysis

[1] PD Dr. Mark Stemmler, Ph.D., University Erlangen-Nuernberg, Department of Psychology, Bismarckstr. 1, D-91054 Erlangen, E-mail: mkstemml@phil.uni-erlangen.de

[2] C. Raymond Bingham, Ph.D., Social and Behavioral Analysis Division, University of Michigan, Transportation Research Institute, 2901 Baxter Road, Ann Arbor, MI 48109-2150, USA, E-mail: rbingham@umich.edu

1. Introduction

In a Lead Article of the journal Applied Psychology (von Eye, Spiel and Wood, 1996) the advantages of configural frequency analysis (CFA) in applied psychological research were presented. This paper can be understood as a sequelae or a supplement, introducing CFA for the analysis of dependent samples, comparable to a dependent t-test in parametric data analysis, an application not outlined in the mentioned Lead Article.

In experimental intervention research improvement scores $Y = X_{\text{Time 2}} - X_{\text{Time 1}}$ are derived from treatment $N_{\text{Treatment}}$ and control samples N_{Control} of N_{Total} individuals (e.g., children, adolescents) as part of pretest-posttest treatment designs. In such designs both samples are observed before and after an intervention, for example, before and after an enhancement program for children. When the improvement scores (Y) are scaled ordinally rather than metrically (or continuously) it is more appropriate to test for group differences using a nonparametric statistical approach such as the Mann-Whitney U test or a median test. However, ratings are often bimodal, one mode resulting from small improvements in the control group and the other mode resulting from large improvements in the treatment group. In such instances the U test (cf. Siegel and Castellan, 1988) should not be applied. Here, the application of other nonparametric tests, for instance, log-linear modeling (LLM) or the two-sample Configural Frequency Analysis (CFA), is recommended. Although CFA is a statistical technique with some tradition (Lienert and Krauth, 1975; Netter, 1996), its merits for evaluation research have been widely neglected in the authors' opinions. In this paper, CFA is presented as a nonparametric statistical tool to test for differences between two dependent samples. The similarities and dissimilarities between CFA and LLM are illustrated using real data from educational psychology.

2. Configural frequency analysis

Configural Frequency Analysis (CFA) is a nonparametric tool for the analysis of d-dimensional contingency tables (von Eye, 1990; 2002). In CFA the cells of a contingency table, called configurations, are analyzed by comparing expected frequencies to observed frequencies. The binomial test, Pearson's chi-square or asymptotic approximations to the z-statistic are the commonly used test statistics to compare the expected to the observed frequencies (Krauth, 1993; Lautsch and von Weber, 1995). Expected frequencies may be based on any hypothetical model and are usually expressed in terms of a log-linear model, typically a main effect model (Mellenbergh, 1996; von Eye and Nesselroade, 1992). Although CFA was developed independently of log-linear modeling, both methods are similar with regard to residual analysis (Lehmacher, 1984). Residuals are the standardized differences between observed and expected frequencies. CFA and residual analysis in log-linear analysis examine all cells of a contingency table. Whereas log-linear modeling applies a model fitting approach which looks for a match between the model (i.e., the expected frequencies) and the observed frequencies, CFA searches for differences on the cell level. In other words: LLM takes a global and CFA a local approach. Significant differences between observed (f_0) and expected frequencies (f_e) are called "types" if there are more observed than expected frequencies (i.e., $f_0 > f_e$), and "antitypes", if there are fewer observed than expected frequencies (i.e., $f_0 < f_e$). Since its introduction to the scientific community (Lienert, 1969), CFA has developed to a universally useful statistical tool for the analysis of contingency tables (Krauth 1988; Krauth, 1993). The researcher may apply CFA nonparametrically or even parametrically (Spiel and von Eye, 1993) for exploratory purposes and for statistical inference (Krauth, 1993), and for the nonparametric evaluation of MANOVA (Stemmler, 1994) and regression designs (Netter, 1996; Lienert and Netter, 1996).

Conceptually, in intervention research types may identify groups of subjects with certain patterns or variable configurations that are in absolute number significantly greater than expected and which might represent groups for which the intervention was especially effective. Below, the application of LLM and CFA is demonstrated to test for univariate and bivariate differences in a two-sample pretest-posttest design; the demonstration applies real data taken from educational research.

3. Univariate testing for treatment effects in two samples

In this example, 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 data are taken from Lienert, 1978; p. 978). The students were randomly assigned to a treatment ($N_{Treatment} = 18$) and a waiting-list control group ($N_{Control} = 18$). After four weeks, students' reading ability was evaluated by their teacher; the ratings were arranged into three categories based on the kind of changes occuring between pre- and post-test: decreased performance (-), unchanged performance (0), or improved performance (+)(see Table 1). The null hypothesis (H_0) states that the teacher's improvement ratings (i.e., three patterns: -, 0, +) are independent of the grouping or sample variable (i.e., two patterns: treatment and control group). The alternative hypothesis (H_1) postulates that there is an interaction between the teacher's ratings and the grouping variable. The null hypothesis may be expressed in terms of the following log-linear model:

$$(1)\ \log e_{ij} = \lambda_0 + \lambda_{\text{Teacher Rating}} + \lambda_{\text{Group}}$$

The data were analyzed with the program SICFA, a FORTRAN program written by Lautsch and von Weber (1995). The program calculates several test statistics to evaluate the differences between expected and observed frequencies: (1) the z-statistic according to Lienert (1969), which is basically the square root of Pearson's chi-square, (2) the binomial test using Stirlings formula (von Eye, 1990), (3) Lehmacher's asymptotic hypergeometrical test (Lehmacher and Lienert, 1982), and (4) the asymptotic approximation of the z-statistic by Perli, Hommel and Lehmacher (1987). The statistics displayed by the program are listed in Table 1. The authors of SICFA (Lautsch and von Weber, 1995) recommend to use either the Lehmacher's asymptotic hypergeometrical test or the test statistic according to Perli et al. Lehmacher‘s test assumes fixed row and column marginals, which is rarely the case. The test statistic according to Perli et al. seems to be powerful, whereas the z-statistic according to Lienert leads to conservative decisions and should be used primarily for exploratory purposes. The binomial test approximation using Stirlings formula is, according to von Eye (1990), more powerful than the exact binomial test and can best be used when the difference between observed and expected frequencies is large.

In our example, the obtained global chi-square is 12.0 with df = 2 which is significant at the $\alpha = 0.05$ level. This suggests that the null hypothesis of assumed independence of the teacher's ratings and the group variable needs to be rejected. The global chi-square is the same chi-square provided by log-linear modeling ($\chi^2 = 11.94$, df = 2, p = .003), which indicates a mismatch between the expected and the observed frequencies. The log-linear models were calculated using SPSS (SPSS Software Inc.). For reasons of comparisons, Table 1 lists the CFA statistics and also the standardized residuals provided by the log-linear model.

To look for significant deviations from the hypothesized model on the cell level we apply the two-sample CFA, which is comparable to the dependent sample t-test in parametric statistical analysis. Two types and two antitypes were consistently detected through Lehmacher's test statistic and the approximation of the z-statistic according to Perli et al. Significant deviations on the cell level exist for those students who were rated as improved by the teacher and for those who were rated as unchanged. There were more treatment students than expected (i.e., type) who were rated as improved by the teachers; in the same category there were fewer control students than expected (i.e., antitype). Furthermore, there were fewer treatment students than expected who were rated as unchanged and more control students than expected in the same category. The standardized residuals show the same pattern (see z-scores above ±1.96).

In order to search for so called *discrimination types* (Lautsch and von Weber, 1995), that is types or significant cell patterns that significantly differentiate between the two samples, Table 1 was transformed into three, two by two (2 x 2) tables (see Table 2), one table for each pairwise comparison of improvement ratings derived by comparing one group to the other two groups combined: {'+' versus '-,0'}, {'0' versus '-,+'} and {'-' versus '0,+'}.

Table 1:
Application of LLM and CFA to test the independence between teacher's rating and the grouping variable; data are taken from a study of 36 students suffering from dyslexia

Teacher Rating	Group	f(o)	f(e)	Approx. of the z-statistic (Lienert)	Approx. of the Binomial test (Stirling)	Lehmacher's test	Approx. of the z-statistic (Perli et al.)	Standard. Residuals
+	T	18	13.5	1.22	1.37	3.04 T	3.46 T	4.49*
+	C	9	13.5	1.22	1.39	3.04 A	3.46 A	-4.49*
0	T	0	3.0	1.73	1.71	2.20	2.68 A	-3.00*
0	C	6	3.0	1.73	1.44	2.20	2.68 T	3.00*
-	T	0	1.5	1.22	0.79	1.19	1.81	-1.50
-	C	3	1.5	1.22	0.88	1.19	1.81	1.50

Note. Learning disabled children (N = 36) were divided at random into two equally sized groups of $N_T = N_C = 18$. The treatment group N_T received a new reading enhancement program while the control group N_C was on a waiting list. After 4 weeks of training all N children were rated by their teachers as 'improved'(+), 'unchanged' (0) or 'decreased' (-). The new alpha level based on Bonferroni adjustment is α^*: 0.05 / 6 = .0083; the corresponding z-value is 2.640. T = Type and A = Antitype. Total chi-square = 12.0, df = 2, p<.05; *p<.05.

Table 2:
Transformed data of Table 1

		Treatment Group	Control Group
Teachers' Ratings	+	a = 18*	b = 9
	-,0	c = 0	d = 9
	p = 0.000516		

		Treatment Group	Control Group
Teachers' Ratings	0	a = 0	b = 6*
	-,+	c = 18	d = 12
	p = 0.00953		

		Treatment Group	Control Group
Teachers' Ratings	-	a = 0	b = 3
	0,+	c = 18	d = 15
	p = 0.11428		

Note. For the application of a two-sample CFA, Table 1 is transformed into three two by two (2 x 2) tables. In the first table the teacher's ratings of '0' and '-' are collapsed together, in the second table the ratings of '+' and '-' are collapsed, and in the third the ratings of '+' and '0'. The p-values were derived from the one-tailed Fisher's exact test. The asterisk indicates a significant type. $^{*}p < 0.05$ (Bonferroni adjusted).

The exact one-tailed Fisher-Yates test (Lienert, 1978; von Eye, 1990) for Table 2 yields the following probabilities: $p_{(+ \text{ versus } 0,-)} = 0.000516$; $p_{(0 \text{ versus } +,-)} = 0.00953$; and $p_{(\text{versus } +,0)} = 0.11428$. Using Bonferroni's adjusted $\alpha^* = 0.05/3 = 0.0167$, the two probabilities $p_{(+ \text{ versus } 0,-)}$ and $p_{(0 \text{ versus} + ,-)}$ are still significant.

In terms of a two-sample CFA one might say that there are two significant types discriminating significantly between treatment and control group. These *discrimination types* can be detected in the Table 1; one type implies that the enhancement program is effective (see number of improved scores: $f(T_+) = 18$), since significantly more subjects showed improvement in the treatment than in the control group. The other type states that significantly more students in the control group experienced no change in their reading abilities (see number of unchanged scores: $f(T_0) = 6$); the two samples of students do not differ significantly in the number whose reading improvement ratings decreased.

4. Bivariate testing for treatment effects in two samples

Table 1 contains the data based on a univariate (i.e., the teacher's rating) two-sample comparison. The design was extended to a bivariate case including students, self-assigned improvement ratings, as well as teacher-assigned improvement ratings (the student data are unpublished data provided by G. A. Lienert). At the end of the training period the dyslexic children were asked whether they felt their reading had improved (↑) or had not improved

(=). Combining these self-assigned ratings with the teacher-assigned ratings resulted in the three by two by two table of 12 improvement ratings as illustrated in Table 3a. Data analysis was done using CFA and LLM. Again, the H_o postulates an independence model, allowing no interactions between the grouping variable and any other variables in the model; interactions between the non-grouping variables are allowed. The H_1 assumes at least one interaction between the grouping variable and any variable in the model. The log-linear model may be stated as follows (TR = teacher's ratings; SR = students' ratings):

$$(2)\ \log e_{ij} = \lambda_0 + \lambda_{TR} + \lambda_{SR} + \lambda_{Group} + \lambda_{TR\ SR}$$

Table 3a:
Bivariate two sample CFA - listing of self-ratings and teacher ratings of the dyslexic children

Teacher	**Student**	**Group**	**f(o)**	**f(e)**	χ^2	**Std. Resid.**	**Residuals**
+	↑	T	6	5.00	0.2	0.45	1.00
+	↑	C	4	5.00	0.2	-0.45	-1.00
Test for a discrimination type: $\chi^2 = 0.40$, df = 1, p = n.s.							
Teacher	Student	Group	f(o)	f(e)	χ^2	Std. Resid.	Residuals
+	=	T	3	8.50	3.56	-1.88	-5.49
+	=	C	14	8.50	3.56	1.88	5.49
Test for a discrimination type: $\chi^2 = 7.11$, df = 1, p < .05							
Teacher	Student	Group	f(o)	f(e)	χ^2	Std. Resid.	Residuals
0	↑	T	4	2.00	2.00	1.41	2.00
0	↑	C	0	2.00	2.00	-1.41	-2.00
Test for a discrimination type: $\chi^2 = 4.00$, df = 1, p = n.s.							
Teacher	Student	Group	f(o)	f(e)	χ^2	Std. Resid.	Residuals
0	=	T	2	1.00	1.00	0.99	0.99
0	=	C	0	1.00	1.00	-0.99	-0.99
Test for a discrimination type: $\chi^2 = 2.00$, df = 1, p = n.s.							
Teacher	Student	Group	f(o)	f(e)	χ^2	Std. Resid.	Residuals
-	↑	T	0	0	-	-	-
-	↑	C	0	0	-	-	-
-							
Teacher	Student	Group	f(o)	f(e)	χ^2	Std. Resid.	Residuals
-	=	T	3	1.50	1.50	1.23	1.50
-	=	C	0	1.50	1.50	-1.23	-1.50

Note. The teacher's ratings were arranged as 'improved' (+), 'unchanged' (0) or 'decreased' (-); students' ratings as 'improved' (↑) or 'unchanged' (=). T = treatment group; C = control group. * = significant discrimination type after Bonferroni adjustment $\alpha^* = \alpha/6 = 0.0083$, the corresponding $\chi^2 = 6.97$. The total $\chi^2 = 16.51$ with df = 5, p < .05.

Now, the CFA program by von Eye (2002) was applied. The expected frequencies calculated by the program were consistent with those provided by SPSS. Again, CFA statistics and the standardized residuals from LLM are listed for reasons of comparisons. The two-sample CFA adds the chi-square components for each specific pattern from the two samples and tests the combined chi-square component for a discrimination type; it is also possible to add the absolute values of the standardized residuals and interprete them as z-statistics. The total chi-square provided by the log-linear model was $\chi^2 = 16.51$, df = 5, p = .006; thus the postulated null model of no interactions between the grouping variable and the other variables in the model needs to be rejected. Any deviation from the null model can be interpreted as a discrimination type, that is, a type that differentiates significantly between the two groups; here, no differentiation between types and antitypes is performed. By applying the Bonferroni adjustment for multiple testing α^*: 0.05 / 6 = 0.0083 (the corresponding chi-square is 6.97), one discrimination type was detected with CFA. There were fewer treatment students than expected by the null model who rated themselves as unchanged but were rated by the teachers as improved. At the same time, there were more control students who rated themselves as unchanged but were rated by the teachers as improved. Again, the Fisher-Yates test may be performed to double-check for a discrimination type with a = 3, b = 14, c = 18-3 = 15, and d = 18-14 = 4 resulting in a 1-tailed p-value of p = 0.000305 (see Table 3b). For R = 6 sign patterns in Table 3a, the Bonferroni adjustment is $\alpha/6 = 0.0083$. Therefore, the bivariate rating type (T+, S=) has a post-hoc significance below the 5% level, suggesting that this type differentiates significantly between the two samples. Thus, the Fisher-Yates test and the two-sample CFA lead to same result.

Table 3b:
Listing of data from Table 3a transformed such that the pattern '+ =' is tested against all other patterns

T S	Treatment Group	Control Group
+ =	a = 3	b = 14*
other	c = 15	d = 4
	$N_T = 18$	$N_C = 18$
	p = 0.000305	

Note. T = teacher; S = student. The asterisk indicates a significant type. $^*p < .05$ (Bonferroni adjusted).

Of course the design may be expanded to three or more improvement ratings by including other variables, such as the parents' view of the child's learning progress. In this case, the design needs to be adapted accordingly, but the test rationale stays the same.

5. Discussion

Configural Frequency Analysis is a very useful tool for nonparametrically testing two-sample differences in applied psychological research. This version of CFA may be compared to the parametric dependent sample t-test. CFA compares the expected frequencies based on the specified null hypothesis (usually an independence model) with the observed frequencies. Whereas LLM takes a global or model fitting approach (Langeheine, 1984); CFA takes a local cell level approach to look at differences between the expected and observed frequencies. CFA also identfies which cells account for the significant sample differences. Those cells were called *discrimination types*. The traditional Fisher-Yates test and the two-sample CFA were suggested for the detection of discriminating cell patterns.

For two-sample comparisons LLM and CFA can be treated as complementary statistical tools. LLM can be used to look for the underlying associations in the dataset. On the cell level standardized residuals may indicate deviations from a pre-specified model. A significant global chi-square assumed, CFA is able to identify the types that discriminate between the two samples.

In general, improvement scores derived from pretest-posttest designs only yield conclusive discrimination types when the following two criteria are met: (1) when the control and treatment samples are randomly selected from a given population (e.g., in our case the population of children suffering from dyslexia); and (2) when the improvement ratings are made ‘blindly’ by a teacher who does not know the group assignments of the children being rated. In the present example the improvement ratings were derived from essays written before and after an intervention (i.e., enhancement training in the treatment group and regular training in the waiting-list control group). These essays were compared by a teacher who was blind to group membership of the children, and the post-treatment essays were rated as having improved (improved performance: +), remained the same (unchanged performance: 0), or declined (decreased performance: -) relative to the pre-treatment essays.

However, the experimental approach based on randomized assignments to treatment and control groups is rarely realized in educational research. More often a quasi-experimental approach is utilized in which two parallel, already existing classes of children are exposed to different teaching methods. If such a quasi-experimental design is used, care must be taken to rule out alternative explanations for the resultant improvement ratings. For example, improvement ratings from two parallel classes of students are comparable only if the students in the two classes have the same average IQ. Classes with higher average IQs will show greater improvements; however this will be an artifact of their IQ and not an effect of the training, because students with higher IQs are likely to profit more from whatever training they receive. Therefore, if obtained from two non-randomly assigned groups, improvement ratings are not directly comparable, even if the ratings are obtained from a blind rater.

The pretest-posttest design can be generalized to more than one improvement criterion if a longitudinal, rather than a cross-sectional design, is used. Single-measurement-point designs, such as the one illustrated in Table 1, may be expanded to a longitudinal design with three measurement points. Improvement ratings across the three measurements can be used to yield four, two-period improvement patterns resembling response curves: ($\uparrow\uparrow$), ($\uparrow$=), (=$\uparrow$), (= =) (where (=) corresponds to no improvement and ($\uparrow$) indicates improvement (cf. Lehmacher, 1987). Assuming that the treatment was given between the first and second time of measurement, ($\uparrow$=) would represent a short-term learning effect which does not further improve after the second measurement, while ($\uparrow\uparrow$) would indicate a long-term effect of increasing learning abilities, and (= =) would indicate no improvement of learning abilities. Finally, a pattern of no improvement followed by enhanced performance (=$\uparrow$) would indicate a delayed effect.

References

1. Krauth, J. (1988). Distribution-free statistics: An application-oriented approach. Amsterdam: Elsevier.
2. Krauth, J. (1993). Einführung in die Konfigurationsfrequenzanalyse (KFA). Ein multivariates nichtparametrisches Verfahren zum Nachweis und zur Interpretation von Typen und Syndromen. Weinheim: Beltz Psychologie Verlags Union.

3. Krauth, J. (1996). Good typal analysis must be based on a precise definition of types. Applied Psychology, 4, 334-337.
4. Langeheine, R. (1984). Explorative Techniken zur Identifikation von Strukturen in großen Kontingenztafeln. Zeitschrift für Sozialpsychologie, 254-268.
5. Lautsch, E., and von Weber, S. (1995). Methoden und Anwendungen der Konfigurationsfrequenzanlyse (KFA). Weinheim: Beltz; Psychologie Verlags Union.
6. Lehmacher, W. (1984). Die Beziehung der Konfigurationsfrequenzanalyse zu anderen Methoden der Analyse qualitativer Daten. In H. Enke and J. Haerting (Eds.), Biometrie und Biostatistik in der Medizin und verwandten Gebieten, (pp. 83-91). Halle: Martin-Luther-Universität Halle-Wittenberg.
7. Lehmacher, W. (1987). Verlaufskurven und cross over. Heidelberg: Springer.
8. Lehmacher, W., und Lienert, G. A. (1982). Die Konfigurationsfrequenzanalyse. XVI. Neue Tests gegen Typen und Syndrome. Zeitschrift für klinische Psychologie und Psychotherapie, 30, 5-11.
9. Lienert, G. A, and Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. Educational Psychological Measurement, 35, 231-238.
10. Lienert, G. A. (1969). Die Konfigurationsfrequenzanalyse als Klassifikationsmittel in der klinischen Psychologie. In M. Irle (Eds.), Bericht über den 26. Kongress der Deutschen Gesellschaft für Psychologie 1968 in Tübingen, (pp. 244-253). Göttingen: Hogrefe.
11. Lienert, G. A. (1978). Verteilungsfreie Methoden in der Biostatistik. Band II. Meisenheim am Glan: Verlag Anton Hain.
12. Lienert, G. A. and Netter, P. (1996). LSD response in Eysenckian trait types identified by polypredictive CFA. Personality and Individual Differences, 21, 845-850.
13. Mellenbergh, G. J. (1996). Other null model, other (anti)type. Applied Psychology, 4, 329-330.
14. Netter, P. (1996). Prediction CFA as a search for types: History and specifications. Applied Psychology, 4, 338-344.
15. Perli, H. G., Hommel, W., and Lehmacher, W. (1987). Test procedure in configural frequency analysis (CFA) controlling the local and multiple level. Biometrical Journal, 27, 885-93.
16. Siegel, S., and Castellan, N. J. (1988). Nonparametric Statistics (2nd ed.). New York: McGraw Hill.
17. Spiel, C., and von Eye, A. (1993). Configural frequency analysis as a parametric method for the search of types and antitypes. Biometrical Journal, 35, 151-164.
18. Stemmler, M. (1994). A nonparametrical evaluation of ANOVA and MANOVA designs using interaction structure analysis. Biometrical Journal, 36, 911-925.
19. von Eye, A. (1990). Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications. Cambridge: Cambridge University Press.
20. von Eye, A. (2002). Configural Frequency Analysis - methods, models, and applications. Mahwah, NJ: Lawrence Erlbaum Publishers.
21. von Eye, A., Spiel, C., and Wood, P. K. (1996). Configural frequency analysis in applied psychological research. Applied Psychology, 4, 301-327.
22. von Eye, A., and Nesselroade, J. R. (1992). Types of change: Application of configural frequency analysis in repeated measurement designs. Experimental Aging Research, 4, 169-183.