4th ICRIEMS Proceedings
Published by The Faculty Of Mathematics And Natural Sciences
Yogyakarta State University, ISBN 978-602-74529-2-3

Prediction Configural Frequency Analysis (P-CFA) for Indicating Interaction between The Level of Education of Children and Parents

Resa Septiani Pontoh and Defi Yusti Faidah

Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Padjadjaran

[a)]Corresponding author:resa.septiani@unpad.ac.id
[b)] yusti88@gmail.com

Abstract. Configural frequency analysis is an analysis that can see the existence of discrepancy on the model marked by the existence of a significant difference between the frequency of observation and frequency of expectation. This analysis focuses on whether or not the interaction among categories from different variables, and not the interaction among variables. In its development, a person can use this analysis to predict certain events. The basic difference between CFA and P-CFA is that there is no difference between the status of variables on CFA, while p-cfa makes the difference between the predictors and criterias. In addition, Log linier model does not differentiate between independent variables and dependent variables. It is also able to detect the relationship among some variables, including the possibility of a causal relationship. Therefore, this model can be used as the base models in the CFA. In general, CFA and P-CFA usually used to see the characteristics of the causes of an event. Hence, this research will present the concept of the p-cfa and how it works for the real data. The data on this paper is on the level of education of children and parents living in the district of Bandung. From the results, it is known that the parents, both the father and mother of the children who are graduated from high school tend to be followed by their children to be graduated from high school especially if the children studied in urban areas.

INTRODUCTION

Configural Frequency Analysis is an analysis that can see the existence of discrepancies on the model marked by the existence of a significant difference between the frequency of observation and expectation. This analysis focuses on whether or not the interaction among categories from different variables is exist and not the interaction among the variables. The goal of CFA is to detect configuration patterns in the data whether experiencing deviations from the model [1]. The deviation is then called discrepancies. In the CFA, deviation on the configurations will occur when an event more often or more rarely appears from what has been previously expected. An event occured more often than what has been previously expected is called type, while an event rarely occured from what has been previously expected is called antitype [2]. Furthermore, The discrepancies are existing when at least one type and/ or antitype appears. In addition, CFA is a method to identify configurations from the categories of variables in the multidimensional cross-table [3].

CFA is an analysis used for the data in the form of categories. This is based on the understanding of configurations of categories that explain the cell from a cross table. In its development, a person is able to use this analysis to predict certain events. The basic difference is that CFA does not focus on the differences among the status of variables, while P-CFA makes the difference between predictors and criterias. Thus, this analysis is able to collaborate, either multiple predictors or multiple criterias [4]. In general, CFA and P-CFA are usually used to see the characteristics of the causes of an event. Hence, This paper presents the concept of P-CFA and how it works for the real data. This research uses secondary data on the level of education of children and their parents in the district of Bandung.

METHODOLOGY

Similar to CFA, there are five steps on performing Prediction CFA, as follows:

1. selecting a base model,
2. selecting a concept of deviation from independence,
3. selecting a significance test
4. estimating expected frequencies
5. interpreting the type and antitypes

Commonly, there are four groups of base models on CFA which are usually used [5]. The very common model used by CFA is Log linear model. This model is involved to estimate the frequencies of cell on both observation and expectation.

The second and the third base model are priori probabilities for population parameters and substantive model. And, the fourth is concerning the multivarioate distribution. This paper includes log liner as base model to estimate the observed and expected cells' frequencies.

Log Linier Model is used to analyze the pattern of the relationship among groups of categorical variables to see association of two, three or more variables, both simultaneously and partially. The patterns of the relationship among variables can be seen from the interaction among variables itself.

Log Linier Model does not differentiate independent variables and dependent variables [6]. It is also able to detect the relationship among some variables, including the possibility of a causal relationship. Therefore, this method can be used as the base models in the CFA. CFA base model must meet the following three criterias [7]:

1. There is at least one reason for discrepancies (a mismatch between the frequency of observation and frequency expectations marked by the emergence of type or antitype.
2. Parsimoni: basic models should be as simple as possible and at the lowest possible orders.
3. Sampling scheme consideration: sampling scheme from all the variables must be considered.

Different to log linear model, CFA and P-CFA are seldom ignore the goodness of fit since it is not the main interest of CFA. The most important thing in CFA is gathering information if there is discrepancy between model and the fact.

Sampling scheme can affect the base model that is selected. The data taken is assumed has been taken from a populationbased on categories that is previously given and mentiones by a frequency distribution from a cross table. Sampling scheme is used to determine the estimates of frequency of expectations of a cell (von Eye, 2002). The very basic thing to perform a good result on estimating the expected cell frequencies is the sampling scheme.

There are two kinds of sampling schemes which are familiar in CFA and P-CFA. They are multinomial sampling and product multinomial sampling [8]. This paper includes multinomial sampling since the frequencies of the cells are derived genuinely from the respondents.

For the next step, the concept of independence is not too important in P-CFA for one sample. Thus, this paper does not mention this step.

For applying CFA in real data, firstly, to see the interaction among predictors, it can be done by testing the main effect between predictors. If there is a type or antitype means there is interaction between predictors. Log Linier model to see whether or not the interaction among predictors [9] is as follows:

$$LogE(Y_{ijklm}) = \mu + A_i + B_j + C_k + D_l \quad (1)$$

$LogE(Y_{ijklm})$ mentions on expected frequency cell and μ is the intercept. In addition, A_i, B_j, C_k, D_l are main effect on Level of education for children in the category of i, main effect on Gender in the category of j, main effect on education of mothers in the category of k, main effect on education of fathers in the category of l, respectively. If we are statistically sure that there is at least one type/antitype emerged, then we are ready to model interaction between predictors and criterias. The model is explained by equation 2.

$$LogE(Y_{ijklm}) = \mu + A_i + B_j + C_k + D_l + E_m + BC_{jk} + BD_{jl} + ... + BCD_{jkl} + ... + BCDE_{jklm} \quad (2)$$

The model in equation 2 is modified with interactions among predictors and interaction among criterias. However, as mentioned earlier, it is forbidden to make interaction among predictors and criterias.

The next step of this research is to determine the significance test. $H_0 : E[N_i] = E_i$. The Null hypothesis is type and antitype do not emerge. When $H_0 : E[N_i] \neq E_i$, type and antitype will emerge.It means that we reject the null hypothesis.

In other words, if the number of observed frequencies is more than what has been expected, type will emerge, and, if the number of observed frequencies is less than what has been expected, antitype will emerge [11].

$$\alpha^* = \frac{\alpha}{\sum t} \quad (3)$$

It is need to know from (3) that α^* is adjustment of α since the value of α for each different configuration is dfferent with others, then used Bonferroni method by dividing the α = 0.05 by number of configurations happened. $\sum t$ is the number of configurations.

After determining the value, the next step is searching for the test statistic $z = \frac{N_{i-}E_i}{\sqrt{E_i}}$. We then reject Ho if $p - value < \alpha^*$ indicating that type and/or antitype emerge [12].

RESULTS AND DISCUSSION

According to the UNDP Human Development Index 2015 Indonesia was at 110th from 183 countries in the world. As we know, education is one of many factors determining the Human Development Index. Although education is the right for all citizens which is protected by the Law, in fact it could not reach by all citizens. This can be seen from APS in each province in Indonesia. Through APS, it can be known the percentage of population who do not enjoy education..

West Java Province is a province which has a highest number of people living. Most of them are men and 26 percent of the people are living in rural area. According to BPS data (2014), there are 0.7 percent children at the age of 7- 12 did not finish elementary school. In addition, there are 7.16% children at the age of 13-15 did not entry the junior school and 65.4 children at the age of 16 - 18 did not attend to high school.

Therefore, this paper, at the first stage, tries to capture is there any relationship between the level of education of children and their parent. This paper is also included location of the school of the children, not to mention their gender. Thus, Variables examined in this research are parents' education (graduated or not from high school, the location of chldren's schools (urban or rural) and gender (men or women). They are then presented in the form of contingency table to analyze the configurations of the variables which are significantly considered as types or antitypes.

The data of this paper are derived from secondary data of education level in the West Java Province. This paper tried to find the characteristics people in bandung district who tended to graduate from high school based on gender, education of the mother, education of the father and the location of the schoolof the people.

As mentioned above, the base model used in this analysis is log linear model. This model then estimates the expected cell frequencies of all configurations. The first step in P-CFA is identification of association among predictors whether they are independent or not [13]. Figure 1 provides information on the results of the P-CFA among predictors.

From Figure 1, it is known that type and antitype emerge meaning that the indepency among predictors are achieved, four types and seven antitypes. Thus, it indicates that the next analysis can be done.

Interaction of Predictors and Criterias

The analysis between the predictors and criterias is run to see the discrepancies occured as described in Figure 2. Based on this, it can be seen the emergence of type indicating that there are deviations from the base model formed. This deviation is the result of configuration variables. The third field shows the level of education of children. The two digits of the first number show the configuration variables predictors. Based on the Figure 2, it can be seen the emergence of type indicating that there are deviations from the base model formed.

Configuration	fo	fe	statistic	p	
1111	218.	230.280	-.809	.20919072	
1112	211.	144.594	5.522	.00000002	Type
1121	47.	71.963	-2.943	.00162727	Antitype
1122	15.	45.186	-4.491	.00000355	Antitype
1211	3.	33.687	-5.287	.00000006	Antitype
1212	0.	21.152	-4.599	.00000212	Antitype
1221	68.	10.527	17.713	.00000000	Type
1222	2.	6.610	-1.793	.03647689	
2111	203.	198.433	.324	.37288589	
2112	163.	124.597	3.440	.00029050	Type
2121	49.	62.010	-1.652	.04925011	
2122	10.	38.937	-4.637	.00000177	Antitype
2211	2.	29.028	-5.017	.00000026	Antitype
2212	0.	18.227	-4.269	.00000981	Antitype
2221	55.	9.071	15.249	.00000000	Type
2222	4.	5.696	-.711	.23866067	

chi2 for CFA model = 738.6781
df = 11 p = .00000000

Figure 1. Result of interaction among predictors

Configuration	fo	fe	statistic	p	
11	128.	133.914	-.511	.30464751	
12	90.	84.086	.645	.25947246	
21	177.	129.614	4.162	.00001577	Type
22	34.	81.386	-5.253	.00000008	Antitype
31	14.	28.871	-2.768	.00282275	
32	33.	18.129	3.493	.00023906	Type
41	7.	9.214	-.729	.23285944	
42	8.	5.786	.921	.17863840	
51	2.	1.843	.116	.45392236	
52	1.	1.157	-.146	.44192774	
61	0.	.000	.000	.49999081	
62	0.	.000	.000	.49999272	
71	13.	41.771	-4.452	.00000426	Antitype
72	55.	26.229	5.618	.00000001	Type
81	1.	1.229	-.206	.41831123	
82	1.	.771	.260	.39733939	
91	118.	124.700	-.600	.27425753	
92	85.	78.300	.757	.22447367	
101	150.	100.129	4.984	.00000031	Type
102	13.	62.871	-6.290	.00000000	Antitype
111	18.	30.100	-2.205	.01371035	
112	31.	18.900	2.783	.00269081	
121	5.	6.143	-.461	.32235903	
122	5.	3.857	.582	.28031215	
131	0.	1.229	-1.108	.13384256	
132	2.	.771	1.399	.08093802	
141	0.	.000	.000	.49999081	
142	0.	.000	.000	.49999272	
151	11.	33.786	-3.920	.00004428	Antitype
152	44.	21.214	4.947	.00000038	Type
161	1.	2.457	-.930	.17629419	
162	3.	1.543	1.173	.12037563	

chi2 for CFA model = 242.1134
df = 15 p = .00000000

Figure 2. Result on Interaction among Predictors and Criterias

Followings are some examples on describing configurations mentioned by figured 1:

- 1111: A man whom mother and father were graduated from high school and he studied in urban area
- 1112: A man whom mother and father were graduated from high school and he studied in urban rural area
- 1121: A man whom mother was graduated and father was not graduated from high school and he studied in urban area
- 2111: A woman whom mother and father were graduated from high school and she studied in urban area

- 2222: A woman whom mother and father were not graduated from high school and she studied in rural area

CFA focus on the configuration of the results of the analysis marked by the emergence of type or antitype. The results show that the configuration model is significant. While from the configuration, types or antitypes explaining that the configuration fits to the base model.

Based on the results of the data analysis, types appear on:

2 1 : A man whom mother and father were not graduated from high school and he studied in rural area, tended to not graduate from high school,

3 2 : A man whom mother was not graduated and father was graduated from high school, he studied in rural area, tended to not graduate from high school,

7 2 : A man whom mother was not graduated and father was graduated from high school, he studied in urban area, tended to graduate from high school,

10 1: A woman whom mother and father were not graduated from high school, she studied in rural area, tended to not graduate from high school,

15 2: A woman whom mother and father were not graduated from high school, she studied in urban area, tended to graduate from high school,

Based on the results of the data analysis, antitypes appear on:

2 4 : A man whom mother and father were not graduated from high school, he studied in rural area, tended to not graduate from high school,

7 1 : A man whom mother was not graduated and father was graduated from high school, he studied in urban area, tended to graduate from high school,

10 2 : A woman whom mother and father were not graduated from high school, she studied in rural area, tended to not graduate from high school,

15 1 : A woman whom mother and father were graduated, she studied in urban area, tended to not graduate from high school,

From the conclusion, it can be indicated that the education of parents and the location of the schools in urban areas tend to encourage a child to graduate from high school.

CONCLUSION

P-CFA will combine the variables predictors become one of variables with many categories. The basic difference is CFA does not focus on the difference between the status of variables, while P-CFA makes the difference between predictors and criterias. Log Linier Model is used as based model since it does not differentiate independent variables and dependent variables. It also can detect relationship among some variables, including the possibility of a causal relationship. Thus, this analysis is able to collaborate, either multiple predictors or multiple criterias. The variables examined in this research are parents' education, the location of childrens' schools and gender presented in the contingency table to analyze the variables which is significantly considered as types or antitypes. This paper involves secondary data education level of pople in Bandung District to see the characteristics of education of people based on gender, education of their parents and school location. From the data analysis, it can be shown that types and antitypes are emerged. Thus, it can be concluded that the education of the parents and the location of the schools (in urban areas) tended to encourage a child to be graduated from high school. It is needed to do further analysis to see the accuracy and validation of the prediction.

ACKNOWLEDGEMENT

The authors thank to Allah SWT and also Faculty of Mathematics and Science, Department of Statistics, Universitas Padjadjaran of the funding.

REFERENCES

1. von Eye, A., & Gutiérrez-Peña, E. (2004). Configural Frequency Analysis - the search for extreme cells. Journal of Applied Statistics, 31, 981 - 997.

2. von Eye, A. 2002. *Configural Frequency Analysis: Methods, Models, and Aplications*. Lawrence Erlbaum Associates: London.
3. Heilmann, W.-R., Lienert, G. A., & Maly, V. (January 01, 1979). Prediction Models in Configural Frequency Analysis. *Biometrical Journal, 21*, 1, 79-86.
4. Mair,P. K., Eye, A. V. 1994. Conjugate Prediction models for Configural Frequency Analysis. *Australian Journal of Statistics, Volume 37 (2008), Number 2, 161–173*
5. Agresti, A. 2007. *An Introductional to Categorical Analysis*. John Willey & Sons, Inc. New York.
6. von, E. A., Mair, P., & Bogat, G. A. (January 01, 2005). Prediction models for Configural Frequency Analysis. *Psychology Science, 47*, 342-355.
7. von, E. A. (January 01, 2004). Base models for Configural Frequency Analysis. *Psychology Science, 46*, 1, 150-170.
8. von Eye, A., & Schuster, C. (1998). On the specification of models for Configural Frequency Analysis - sampling schemes in Prediction CFA. Methods of Psychological Research - online, 3, 55 – 73
9. von Eye, A., & Bogat, G.A. (2006). Logistic regression and prediction Configural Frequency Analysis - A comparison.
10. von Eye, A., & Brandtstädter, J. (1997). Configural Frequency Analysis as a searching device for possible causal relationships. Methods of Psychological Research - Online, 2, 2, 1 - 23.
11. Gutierrez-Pena, E., and von Eye, A. (2000). A Bayesian approach to configural frequency analysis. Journal of Mathematical Sociology,24, 151-174
12. Mair, P. (2007). A framework to interpret nonstandard log-linear models. Austrian Journal of Statistics,36, 1-15.
13. Mair, P., and von Eye, A. (2007). Application scenarios for nonstandard log-linear models. Psychological Methods, 139-156.