# THE STUDY OF CELLS IN THE ANALYSIS OF CONTINGENCY TABLES FROM THE VIEWPOINT OF RUDAS, CLOGG AND LINDSAY MIXTURE INDEX OF FIT.

Adalberto González Debén y Jesús E. Sánchez García
Instituto de Cibernética, Matemática y Física, CITMA

**RESUMEN**
En este trabajo se presenta el estado del arte del índice mixto de falta de ajuste, propuesto por Rudas Clogg y Lindsay en 1994. Se discute el análisis de la clase latente no estructurada, considerado como un análisis de residuos de nuevo tipo. Se concluye que el mismo resulta idóneo siempre que la falta de ajuste se deba a la existencia de individuos "sobrantes" , pero que no sucede lo mismo en el caso de que en una o varias celdas haya individuos "faltantes". Por último, se define la influencia de una celda en la falta de ajuste de un modelo que puede ser usada como una medida descriptiva complementaria en el analisis de residuos usual y en el analisis de las frecuencias de las configuraciones.

**ABSTRACT**
In this paper the state of the art of the mixture index of fit, proposed by Rudas, Clogg &Lindsay in 1994 is presented. The analysis of the unstructured latent class considered as a residual analysis of a new kind is discussed. It is concluded that it is most adequate whenever the lack of fit is due to the existence of "spare" individuals, but it is not the same if there are "lacking" individuals in several cells. At last, the influence of a cell on the lack of fit of a model is defined, which can be used as a complementary descriptive measure in the usual residual analysis and in the configural frequency analysis.

**KEY WORDS:** Latent Class Analysis, Rudas-Clogg-Lindsay Mixture Index of Fit, Residual Analysis

**MSC** 62H17

## 1. INTRODUCTION

Rudas, Clogg & Lindsay (1994) propose an index of fit that constitutes a rather new use of the latent class model (see also Rudas, 1998 and Rudas 2002). This index is general, easy to interpret and does not have the limitations of usual procedures related with sample size.

Let H be a model proposed for a contingency table $P=\{P_y\}$, where *y* runs over all possible response patterns. Let $\pi=P(X=2)$ and $1-\pi=P(X=1)$ be a latent class distribution for a 2-class model. The following family of saturated models $H_\pi$, with two latent classes, is defined,

$$P_y=(1-\pi)Q_y+\pi R_y$$

where $P_y=P(Y=y), Q_y=P(Y=y/X=1),$ and $R_y=P(Y=y/X=2)$ have the following characteristics:

- Model H is valid in the first latent class *Q*
- There is no assumption concerning the second latent class *R*.

The formulation without restriction for the second component can be considered as a representation of the un-modelled heterogeneity, the variation not described by model *H*.

The Family $H_\pi$ is very general, it contains model *H* (for $\pi=0$) up to the saturated one (when $\pi=1$.) The classical latent class model is a particular case where the independence model *H* is valid in both latent classes. Goodman's model (1975) and the extended models of Dayton & Macready (1980) are included. In these cases, *H* is a linear scaling model.

The family of models $H_\pi$ has the monotony property, i. e.:

$$\text{If } \pi < \pi' \text{ then } H_\pi \subset H_{\pi'}.$$

Rudas, Clogg & Lindsay's mixture index of fit (RCL) is defined, on the basis of this property, as the minimal value of $\pi$, $\pi = \pi^*$, for which $H_\pi$ is saturated.

The $\pi^*$ value is interpreted as the proportion of the population intrinsically not described by *H*. The lack of fit of *H* is concentrated on R. Complementarily, $1-\pi^*$ is also an index of fit: it is the proportion of the population intrinsically described by *H*.

**2. $2 \times 2$ CONTINGENCY TABLE**

An explicit formula for $\pi^*$ can be obtained for the case of two cross-classified dichotomic variables. Table 1 represents the observed frequencies.

| Rows/Columns | Column 1 | Column 2 | Total |
|---|---|---|---|
| Row 1 | A | B | A+B |
| Row 2 | C | D | C+D |
| Total | A+C | B+D | n |

**Table 1**. Observed frequencies in a 2x2 table.

Let us suppose $AD\,BC > 0$, and $A > D$; the perfect value corresponding to cell (2,2) is:

$$x = \frac{BC}{A},$$

the decomposition into two latent classes has the following form:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ C & \frac{BC}{A} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & D - \frac{BC}{A} \end{bmatrix}.$$

Finally,

$$\pi^* = \frac{AD - BC}{nA}.$$

It can be seen that the value of the index agrees exactly with one of the differences between the observed value and the value of the perfect cell (expressed as proportion of the whole sample).

Note that if $A = D$, although $\pi^*$ remains unchanged, two equivalent decompositions of the original table can be obtained depending on the cell being assigned the perfect value.

**3. TWO-WAY CONTINGENCY TABLE**

Rudas *et al.* (1994) use as example in their paper the case of the two-way contingency table and the model of independence. The results of the analysis of one of the examples used by those authors are presented, as a way of illustrating their ideas.

| Eyes/Hair | Black | Brown | Red | Blond |
|---|---|---|---|---|
| Brown | 68 | 119 | 26 | 7 |
| Blue | 20 | 84 | 17 | 94 |
| Gray | 15 | 54 | 14 | 10 |
| Green | 5 | 29 | 14 | 16 |

**Table 2**. Tabla 1, page 625, Rudas *et al.* (1994)

It handles with cross-tabulation of eye and hair colors for a sample of 592 individuals. The data appear in table 2.

The lack of fit estimation for the independence model for this example is $\hat{\pi}^* = 0.2958$. This means that approximately 30% of the population does not follow the independence model.

Rudas *et al.* (1994) propose an EM algorithm for the estimation of parameter $\pi^*$. Moreover, they gave a lower confidence limit, $\pi_L^*$, based on a likelihood ratio statistic $L^2$, and equal to 2.70 (the 90$^{th}$ percentile of the chi-square distribution with 1 degree of freedom.) The confidence interval is only lower bounded because by definition of $\pi^*$ model $H_\pi$ is saturated for every value $\pi > \pi^*$.

**4. RECENT DEVELOPMENTS**

The aforementioned paper left some questions unanswered and they motivated ulterior investigations. It was still to be answered, to know:

1. A measure of precision for the estimator $\hat{\pi}^*$
2. The study of effect of sample zeros
3. The extension of the method to more complex models
4. The investigation of the possibilities and limitations of the analysis of the unstructured latent class.

With respect to the measure of precision for the estimator $\hat{\pi}^*$, Rudas *et al.* (1994) and Clogg, Rudas & Xi (1995) proposed the difference $\hat{\pi}_L^* - \hat{\pi}^*$. Dayton (1999) proposed the estimation of the standard error of $\hat{\pi}^*$ by the Jackknife procedure.

With respect to the sample zeros the two usual alternatives have been explored: (1) to add a small positive constant to each cell, and (2) to consider the sample zeros as structural zeros.

Rudas & Zwick (1997) used the first variant. They compared the results obtained when adding different values and they conclude that for sparse tables it is recommendable to smooth the observed distribution by adding $\varepsilon = 0.1.$ Rudas (2002) recommends the same analysis for being sure that the results do not depend on the used value.

Formann (2000) used the second variant and he finds it satisfactory, even more, based on this solution he recommends this method as the best for the case of sparse tables.

With respect to the extension of the method to more complex models Clogg (1995) and Clogg, Rudas & Xi (1995) used this approach for the analysis of the structure in mobility tables. González (1998) uses the row-column association model and compares its goodness of fit with that of the independence model. Dayton (2003) applies the RCL index to the latent class and the Rasch models.

González & Méndez (2000) propose the use of the index as a new concept of type in the configural frequency analysis for two samples. Von Eye (2001) included this descriptive measure in this program *CONFIGURAL FREQUENCY ANALYSIS, version 2000.*

Xi & Lindsay (1996) propose a new computing method for the RCL index by using sequential quadratic programming (SQP.) The advantages are (1) it has a higher convergence rate than the EM algorithm; (2) it is more general in the sense that it can be applied to any loglinear model, and (3) it does not explicitly require the maximum likelihood estimations of the parameters within each class.

Dayton (1999) includes the RCL index in the chapter "Determination of fit of a model to the data" and he presents the implementation of SQP algorithm in Excel. With this procedure he measures the fit of different models such as: the CLCM with two latent class and some scaling models. Moreover, he gives a practical rule for interpreting the $\pi^*$ value (less than 10% of lack of fit is a reasonable fit.)

At last, Formann (2000) defines the generalized RCL index where the number of components representing model H can be more than one.

This index of fit is rather new, but it can be seen that there is an increase in its diffusion and new application in different contexts.

## 5. ANALYSIS OF THE UNSTRUCTURED LATENT CLASS

With respect to the second latent class, Clogg *et al.* (1995) used this approach for quantifying the structure in mobility tables. According to them, for a basal model (independence, quasi-independence, and uniform association) the RCL index can be interpreted as the amount of structure contained in the data and that is out of the basal model. This structure accumulates in R and its study constitutes a residual analysis of a new kind. Clogg, Rudas & Matthews (1997) propose the use of simple graphs to visualize the second latent class and to study the unmodelled structure.

In González & Méndez (2001) this new approach is compared to the usual residual analysis. A conclusion in their paper is that the analysis of the non-structured latent class is one more tool to study the causes of the lack of fit of a model and to detect possible atypical cells. In this sense, the analysis of the unstructured latent class, could be seen as a residual analysis of a new kind, it could be used as a complement of the standardized residual analysis, the analysis of eliminated residuals and the variation of the likelihood ratio statistic.

The frequency estimations for latent classes 1 and 2, respectively, for the example of hair and eye color cross-classification are shown in Tables 3 and 4. It has already been mentioned that 30% of the population does not follow the independence ($\hat{\pi}^* = 0.2958$.)

| EYES/HAIR | Black | Brown | Red | Blond |
|---|---|---|---|---|
| Brown | 28.35 | 119 | 24.09 | 7 |
| Blue | 20 | 84 | 17 | 4.95 |
| Gray | 12.86 | 54 | 10.94 | 3.17 |
| Green | 5 | 21.01 | 4.26 | 0.36 |

**Table 3**. First latent class (follows the model of independence)

| EYES/HAIR | Black | Brown | Red | Blond |
|---|---|---|---|---|
| Brown | **39.65** | 0 | 1.91 | 0 |
| Blue | 0 | 0 | 0 | **89.05** |
| Gray | 2.14 | 0 | 3.06 | 6.83 |
| Green | 0 | 7.99 | 9.74 | 15.64 |

**Table 4**. Second latent class (non structured)

Note that all cells in the first latent class are perfect and the second class contains the "spare" individuals. The majority of them are grouped in cells (1,1) and (2,4) and both together have 73.5% of individuals in the unstructured latent class. According with this, both variables are dependent because there is an excess of individuals with brown eyes and black hair, as well as blue-eyed with blond hair.

Several presentations could be used to analyze the unstructured latent class. In Table 4 the raw data were presented. It is also possible to use: (1) the percentages with respect to the amount of individuals in the analysis, (2) the percentages with respect to the amount of individuals in the latent class, and (3) the percentages with respect to the amount of individuals in each cell (see Clogg *et al.*, 1997; González and Méndez, 2001.)

The analysis of the unstructured latent class, in its nature, could be seen as a residual analysis of a new kind. It is most adequate whenever the lack of fit is due to the existence of "spare" individuals as in the analyzed example. Nevertheless, in the case of one or several cells with "lacking" subjects this analysis is not so evident (González, 1999). In what follows, it is presented a new result for solving this difficulty.

## 6. INFLUENCE OF A CELL ON THE LACK OF FIT OF A MODEL

Let H be a proposed model for a contingency table P. Let $\pi^*$ be the index of fit of model *H*, and let $\pi^*_y$ that of model *H* considering cell *y* as a structural zero. The influence of a cell *y* on the lack of fit of a model H is defined as follows:

$$ILF_y = \frac{\pi^* - \pi^*_y}{\pi^*}.$$

It is interpreted as the relative reduction of the lack of fit for model H when cell *y* is ignored.

Table 5 shows the index of fit of the quasi-independence model for each cell in the contingency table under analysis, and Table 6 shows the influence of each cell on the lack of fit (percentages.)

Note that cell (1,4) containing the individuals with brown eyes and blond hair is also noteworthy. This cell has a zero in the unstructured latent class (see Table 4.) To the analysis already accomplished, it could be added that the low quantity of subjects with brown eyes and blond hair also contributes to the lack of fit of the independence model.

| EYES/HAIR | black | brown | Red | Blond |
|---|---|---|---|---|
| Brown | **0.2270** | 0.2899 | 0.2926 | **0.2351** |
| Blue | 0.2906 | 0.289 | 0.2909 | **0.1454** |
| Gray | 0.2925 | 0.2884 | 0.2906 | 0.2843 |
| Green | 0.2788 | 0.2823 | 0.2793 | 0.2709 |

**Table 5.** Lack of fit index for the quasi-independence model

| EYES/HAIR | Black | Brown | Red | Blond |
|---|---|---|---|---|
| Brown | **23.26** | 1.99 | 1.08 | **20.52** |
| Blue | 1.76 | 2.30 | 1.66 | **50.85** |
| Gray | 1.12 | 2.50 | 1.76 | 3.89 |
| Green | 5.75 | 4.56 | 5.58 | 8.42 |

**Table 6**. Influence of each cell on the lack of fit

Victor (1989) and Kieser & Victor (1991) develop an alternative approach to the standard configural frequency analysis. They say it is not appropriate to estimate the expected values of the whole contingency table by using the information of the cells under analysis, because it is supposed that they do not come from the same population. These authors propose to consider the cells constituting possible types or antitypes as structural zeros. Kieser & Victor (1999) formulate this new approach in terms of the loglinear model and they propose methods to carry out the exploratory as well the confirmatory analysis.

The definition of the influence of a cell proposed in this paper is akin to Kieser & Victor's approach. It is natural to consider this measure as complement of the analysis proposed by these authors.

## 7. CONCLUSIONS

The mixture index of fit proposed by Rudas, Clogg & Lindsay in 1994 is a descriptive measure easy to interpret and with the possibility of a general use. It is not intended to be a substitute for the other known measures, but to use it as their complement. In the last 10 years, it has experienced a continuous advance in the computational aspects as well as in the variety of applications.

The analysis of the unstructured latent class, in its nature, could be seen as a residual analysis of a new kind. It is best used whenever the lack of fit is due to the existence of "spare" subjects as is the case in the analyzed example. Nevertheless, in the case of one or several cells with "lacking" subjects this analysis is not so evident.

The definition of influence of a cell on the lack of fit of a model proposed in this paper can be used as a complementary descriptive measure in the usual residual analysis and in the configural frequency analysis.

**Acknowledgments:** The authors thank Prof. Dr. Jeroen Vermunt for the information and the software given to them.

**RECEIVED** FEBRUARY 2009
**REVISED** MAY 2009

## REFERENCES

[1] CLOGG, C.C. (1995): Latent Class Models. In: ARMINGER, C., CLOGG, C.C., and SOBEL, M.E. (Eds.) : **Handbook of statistical modeling for the social and behavioral sciences**. New York. Plenum Press.

[2] CLOGG, C.C., RUDAS, T. and XI, L. (1995): A new index of structure for the analysis of models for mobility tables and other cross-classifications. **Sociological Methodology,** 25, 197-223.

[3] CLOGG, C.C., RUDAS, T. and MATTHEWS, S. (1997): Analysis of contingency tables using graphical displays based on the mixture index of fit. In: JORG, B. & GREENACRE, M. (Eds.): **Visualization of Categorical Data,** 425-439, Academic Press, N. York.

[4] DAYTON, C.M. (1999): **Latent class scaling analysis.** Sage, Thousand Oaks.

[5] DAYTON, C.M. (2003): Applications and computational strategies for the two-point mixture index of fit. **British Journal of Mathematical and Statistical Psychology.** 56, 1-13

[6] DAYTON, C.M, and MACREADY, G.B. (1980): A scaling model with response errors and intrinsically unscalable respondents. **Psychometrika.** 45, 343-356

[7] FORMANN, A.K. (2000): Rater agreement and the generalized Rudas-Clogg-Lindsay index of fit. **Statistics in Medicine.** 19, 1881-1888.

[8] GONZÁLEZ, A. (1998): Experiencias con un nuevo índice de falta de ajuste en el análisis de tablas de contingencia. **Thesis for the Master Degree,** University of Havana.

[9] GONZÁLEZ, A. (1999): Experiencias con el uso de un nuevo índice de falta de ajuste en tablas de contingencia. **Paper presented at the Cuba-Mexico Statistics Meeting**., La Habana, Cuba.

[10] GONZÁLEZ, A. y MÉNDEZ, I. (2000): Un nuevo concepto de tipo en el análisis de las frecuencias de las configuraciones en dos muestras. **Multiciência,** 4, 7-17.

[11] GONZÁLEZ, A. y MÉNDEZ, I. (2001): Comparación entre diferentes métodos de análisis de residuos en tablas de contingencia y un nuevo enfoque. **Investigación Operacional,** 22, 170-178

[12] GOODMAN, L.A. (1975): A new model for scaling response patterns: An aplication of the quasi-independence concept. **Journal of the American Statistical Association**, 70, 755-768

[13] KIESER, M. y VICTOR, N. (1991): A test procedure for an alternative approach to configural frequency analysis. **Methodika**, 5. 87-97.

[14] KIESER, M. and VICTOR, N. (1999): Configural frequency analysis (CFA) revisited: a new look at an old approach. **Biometrical Journal**, 41, 15-22

[15] RUDAS, T. (1998): The mixture index of fit. In: ANUSKA, F. (Ed.): **Advances in methodology, Data Analysis and Statistics,** 14, 15-22

[16] RUDAS, T. (2002): A latent class approach to measuring the fit of a statistical model. In: HAGENAARS, J. & MCCUTCHEON A. (Eds.): **Applied latent class analysis, Cambridge**, Cambridge University Press, 345-365

[17] RUDAS, T., CLOGG, C.C. and LINDSAY, B.C. (1994): A new index of fit based on mixture methods for the analysis of contingency tables. **Journal of the Royal Statistical Society, Series B,** 56, 623-639

[18] RUDAS, T. and ZWICK, R. (1997): Estimating the importance of differential item functioning, **Journal of Educational and Behavioral Statistics**, 22, 31-45.

[19] VICTOR, N. (1989): An alternative approach to configural frequency analysis, **Methodika**, 3, 61-73.

[20] VON EYE, A. (2001): Configural frequency analysis 0 version 2000. A program for 32 bit windows operating system. **Methods of psychological research online**, 6, 129-139.

[21] XI, L. and LINDSAY, B. (1996): A note on calculating the pi* index of fit for the analysis of contingency tables. **Sociological methods & research,** 25, 248-259