Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v40i1.34953

STATISTICS

A new approach to simple correspondence analysis with emphasis on the violation of the independence assumption of the levels of categorical variables

André Luís Alves Costa[1], Carla Regina Guimarães Brighenti[2] and Marcelo Angelo Cirillo[1*]

*[1]Departamento de Ciências Exatas, Universidade Federal de Lavras, Avenida Doutor Sylvio Menicucci, 1001, 37200-000, Lavras, Minas Gerais, Brazil. [2]Departamento de Zootecnia, Universidade Federal de São João del Rei, São João del Rei, Minas Gerais, Brazil. *Author for correspondence. E-mail: macufla@gmail.com*

ABSTRACT. The main hypothesis of correspondence analysis is given by the independence between the levels of categorical variables. Due to violation of this hypothesis, this study aims to improve the technique of correspondence analysis, providing a new approach for the calculation of the coordinates through the residual incorporation by tables in which categories have different levels of correlation. For this purpose, the simulation was made using Monte Carlo to generate frequencies from the correlated binomial distribution BC (n, π, ρ). It was concluded that in all evaluated scenarios the approach is promising in the sense that the objects were better discriminated against conventional approach. Moreover, the proposed procedure for obtaining the coordinates is likely to be used on real data as shown in the application example.

Keywords: binomial; residuals; cophenetic; inertia.

Uma nova abordagem da análise de correspondência simples com ênfase na violação da hipótese de independência dos níveis das variáveis categóricas

RESUMO. A principal hipótese da análise de correspondência é dada pela independência entre os níveis das variáveis categóricas. Decorrente a violação dessa hipótese, esse trabalho tem por objetivo aprimorar a técnica da análise de correspondência, fornecendo uma nova abordagem para o cálculo das coordenadas através da incorporação de resíduos, mediante tabelas em que as categorias apresentam diferentes níveis de correlação. Com esse propósito, utilizou-se a simulação Monte Carlo na geração de frequências provenientes da distribuição binomial correlacionada BC (n, π, ρ). Concluiu-se que em todos os cenários avaliados a abordagem é promissora, no sentido que os objetos foram melhor discriminados em relação a abordagem convencional. Ainda, o procedimento proposto para obtenção das coordenadas é plausível de ser utilizado em dados reais conforme ilustra o exemplo de aplicação.

Palavras-chave: binomial; resíduos; cofenética; inércia.

Introduction

The categorical data analysis involves a number of statistical methods applied to discrete data, represented by quantitative variables. In this context, sampling designs associated with the probabilistic models are proposed with the aim of evidencing relevant information parameters for preserving the measurement scales.

In terms of data organization, it is summarized in the formation of contingency tables, in which the main characteristics of the statistical analysis are characterized by assessing the independence between variables, studying the conditional distributions, understanding the association between categories, as well as a comprehensive graphical representation and interpreting of the results. The above scenario indicates the correspondence analysis technique as an important alternative to contemplate the features mentioned.

According to Guedes, Ivanqui, Martins, and Cochia (1999) is understood as correspondence analysis the exploratory technique of multivariate analysis which allows obtaining a multidimensional graphical representation of the dependence between the rows and/or columns of a contingency table of two inputs, where the rows and columns represent categories or arrangements for categorical variables. The graphical representation is obtained through the distribution of the scores of the categories of rows and columns and marking these categories as points where the scores are used as the coordinates of these points.

The literature includes several papers describing the theoretical aspects and applications on the analysis theory of correspondence, such as: Greenacre (1992; 2007), Blasius, Greenacre, Groenen, and Velden (2009), Beh (2004; 2012).

The feasibility in applying correspondence analysis in a contingency table starts with the application of a chi-square test to evaluate the independence between the levels of categorical variables.

It is worth noting that statistics of this test in correspondence analysis is extremely important to validate the application, such that the dispersion of the points can be represented in the best possible way, in order to provide symmetrical perceptual maps. However, it is known that the chi-square statistic is sensitive to outliers that in the correspondence analysis case, can be identified by dots in which the variable mass of one is far superior to the other.

Regarding the applications of correspondence analysis in real data, which considers the assumption of independence, between the responses of the levels of a categorical variable, innumerable experiments can exemplified in the sensorial analysis, in which, the tasters assign scores on a scale of 0 -10, in relation to some sensorial characteristic (eg flavor).

Therefore, creating classes for these notes and for groups of tasters, defined by age, the data is organized in a frequency table, so that the classes referring to the notes are represented by the columns and the classes of the age group described by lines. In this way, the assumption of independence is contemplated if we assume that the tasters do not present any previous training to perform the tastings, or that they have heterogeneous sensorial perceptions.

For the same experiment, if we consider that the tasters are trained to carry out the tastings, or that they have homogeneous sensorial perceptions, naturally the answers will be correlated. In this context, the assumption of independence would be violated, and the use of conventional correspondence analysis would not be adequate, that is, the coordinates obtained should be updated with information regarding the dependence of the categorical levels, and incorporation of the residues would be a viable alternative.

Due to this problem, the necessity to add new information to statistical Chi-Square arises, in which the decomposition of singular value is employed. Veloso and Cirillo (2016) in a simulation study proposed a significance test to show major components that best discriminate outliers. They recommend that the samples must be corrected by the distances chi-square Pearson and Yates.

Another alternative is to incorporate Pearson's residues and its different versions described by Lee and Yick (1999), developing a procedure that involves Pearson's residues using the singular value decomposition in relation to the normalization method.

Beh (2012) proposes a procedure to identify the cells that divert from the independence assumption. However, the assumptions relating to such residue are not always satisfied and therefore such findings may lead to questionable conclusions, since these residues do not have unit variance, it is questionable to use them to identify the cells that are not consistent with the aforementioned hypothesis.

Due to the fact that the main assumption of correspondence analysis is checked on the assumption of independence between the categorical variables, represented in rows and columns, it keeps the focus on the incorporation of the standardized waste proposed by Haberman (1973) and residue set proposed By Barnett and Lewis (1994) in the calculation of the scores of simple correspondence analysis, whereas contingency tables structures with varying degrees of correlation between levels of categorical variables. Thus, the simple correspondence analysis is enhanced with getting new coordinates involving those wastes.

Because of what has been mentioned, this study aimed to evaluate and improve the technique of correspondence analysis, providing a new approach for the calculation of residue incorporation by contingency tables in which categories have different levels of correlation.

Material and methods

The generation of contingency tables represented by the matrix $X_{I\times c}$ $X_{I\times c}$, with the I line fixed (i=1,...,I and I=5 or 10), the observed frequencies γ_{ij} (j=1,...,C and C=5 or 10) were simulated according to the binomial distribution, that is $\gamma_{ij}\sim$BC(n,π,ρ) defined by Equation 1:

$$P(y_j|n_j,\pi_j,\rho) = \binom{n_j}{y_j}\pi_j{}^{y_j}(1-\pi_j)^{n_j-y_j}(1-\rho)I_{A1(y_j)} + \pi^{\frac{y_j}{n_j}}(1-\pi_j)^{n_j-\frac{y_j}{n_j}}\rho I_{A2(y_j)} \quad (1)$$

where: A_1=0,1,...,n_j, A2=0,n_j, γ_j=0,...,n_j and 0≤ρ≤1.

Following the procedure given by Cirillo and Ramos (2014), the vector of random variables $Y=(Y_1,Y_2,\ldots,Y_j)$ which each component represents the number of occurrences in the category j for j=1,2,…,C associated to a $\pi_j=(\pi_1, \pi_2,..., \pi_j)$, vector, that corresponds to the binomial probability of success and the mixing binomial rate (n,π) with (1-ρ) probability is a modified Bernoulli distribution represented by *BeM(π)* admitting 0 or *n* values with probability p. Therefore, it should be emphasized that fixed ρ=0 the model BC(n,π,ρ) is equivalent to the common binomial model B(n,π).

For ρ≠0 the model includes extra-binomial variation and ρ≈1 obtains an excess n_j of simulated frequency. In this case, to prevent the matrix frequencies are null, making the realization of matrix operations in the calculation of scores, the null value generated was replaced by Equation 1. Due to this adjustment, the average remains the same obtained by Tallis (1962), but the variance becomes rough, that is Equation 2 and 3:

$$E(Y_j) = n_j\pi_j \quad (2)$$

$$Var(Y_j) \cong \pi_j(1-\pi_j)\{n_j + \rho n_j(n_j-1)\} \quad (3)$$

Following these specifications, the parameter values used in generation of contingency tables are summarized below (Table 1).

Table 1. Parametric values used to generate scenarios for contingency tables for the application of correspondence analysis.

Dimension	Proportion of the binomial	Degree of correlation (ρ)
5×5 and 10×10	0.2	0.2
		0.5
		0.8
	0.5	0.2
		0.5
		0.8
	0.9	0.2
		0.5
		0.8

For each generated contingency table, r_{ij} and $r_{ij}^{\star}$ proposed respectively by Barnett and Lewis (1994) and Haberman (1973) were calculated by Equation 4, 5 and 6:

$$R_{BL} = \begin{bmatrix} r_{11} & \cdots & r_{1c} \\ \vdots & r_{ij} & \vdots \\ r_{l1} & \cdots & r_{lc} \end{bmatrix} \text{ in which } r_{ij} = \frac{y_{ij}-e_{ij}}{\sqrt{e_{ij}}} \quad (4)$$

for, (l=1,…,n ; $_j$=1,…,n), $e_{ij}=n\times p_{ij}p_{ij}$ with $p_{ij} = E\left[\frac{x_{ij}}{n}\right]$ estimated by the equation $\frac{x_{i+}x_{+j}}{n}$; (5)

$$R_H = \begin{bmatrix} r_{11}^{*} & \cdots & r_{1c}^{*} \\ \vdots & r_{ij}^{*} & \vdots \\ r_{l1}^{*} & \cdots & r_{lc}^{*} \end{bmatrix} \text{ given } r_{ij} = \frac{r_{ij}^{*}}{\sqrt{\left(1-\frac{y_{i+}}{n}\right)\left(1-\frac{y_{+j}}{n}\right)}} \quad (6)$$

Following these specifications, the scores of correspondence analysis in the conventional way C_1 and C_2 modified with the incorporation of $C_1^{R_{BL}}$ and $C_2^{R_{BL}}$, $C_1^{R_H}$ and $C_2^{R_H}$ residues, were obtained respectively for the profiles of 'row' and 'column' represented by Equation 7 at 12:

$$C_1 = \begin{bmatrix} \frac{1}{\left(\frac{n_{1+}}{n}\right)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\left(\frac{n_{p+}}{n}\right)} \end{bmatrix} \begin{bmatrix} u_{11}\sqrt{\frac{n_{1+}}{n}} & \cdots & u_{1k}\sqrt{\frac{n_{1+}}{n}} \\ \vdots & \ddots & \vdots \\ u_{p1}\sqrt{\frac{n_{p+}}{n}} & \cdots & u_{pk}\sqrt{\frac{n_{p+}}{n}} \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix} \quad (7)$$

$$C_2 = \begin{bmatrix} \frac{1}{\left(\frac{n_{1+}}{n}\right)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\left(\frac{n_{+q}}{n}\right)} \end{bmatrix} \begin{bmatrix} v_{11}\sqrt{\frac{n_{+1}}{n}} & \cdots & v_{1k}\sqrt{\frac{n_{+1}}{n}} \\ \vdots & \ddots & \vdots \\ v_{q1}\sqrt{\frac{n_{+q}}{n}} & \cdots & v_{qk}\sqrt{\frac{n_{+q}}{n}} \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix} \quad (8)$$

$$C_1^{(R_{BL})} = \begin{bmatrix} r_{11} & \cdots & r_{1c} \\ \vdots & r_{ij} & \vdots \\ r_{l1} & \cdots & r_{lc} \end{bmatrix} \begin{bmatrix} u_{11}\left(\frac{n_{1+}}{n}\right)^{-\frac{1}{2}} & \cdots & u_{1k}\left(\frac{n_{1+}}{n}\right)^{-\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ u_{p1}\left(\frac{n_{p+}}{n}\right)^{-\frac{1}{2}} & \cdots & u_{pk}\left(\frac{n_{p+}}{n}\right)^{-\frac{1}{2}} \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix} \quad (9)$$

$$C_2^{(R_{BL})} = \begin{bmatrix} r_{11} & \cdots & r_{1c} \\ \vdots & r_{ij} & \vdots \\ r_{l1} & \cdots & r_{lc} \end{bmatrix} \begin{bmatrix} v_{11}\left(\frac{n_{1+}}{n}\right)^{-\frac{1}{2}} & \cdots & v_{1k}\left(\frac{n_{1+}}{n}\right)^{-\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ v_{q1}\left(\frac{n_{+q}}{n}\right)^{-\frac{1}{2}} & \cdots & v_{qk}\left(\frac{n_{+q}}{n}\right)^{-\frac{1}{2}} \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix} \quad (10)$$

$$C_1^{(R_H)} = \begin{bmatrix} r_{11}^{*} & \cdots & r_{1c}^{*} \\ \vdots & r_{ij}^{*} & \vdots \\ r_{l1}^{*} & \cdots & r_{lc}^{*} \end{bmatrix} \begin{bmatrix} u_{11}\left(\frac{n_{1+}}{n}\right)^{-\frac{1}{2}} & \cdots & u_{1k}\left(\frac{n_{1+}}{n}\right)^{-\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ u_{p1}\left(\frac{n_{p+}}{n}\right)^{-\frac{1}{2}} & \cdots & u_{pk}\left(\frac{n_{p+}}{n}\right)^{-\frac{1}{2}} \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix} \quad (11)$$

$$C_2^{(R_H)} = \begin{bmatrix} r_{11}^{*} & \cdots & r_{1c}^{*} \\ \vdots & r_{ij}^{*} & \vdots \\ r_{l1}^{*} & \cdots & r_{lc}^{*} \end{bmatrix} \begin{bmatrix} v_{11}\left(\frac{n_{+1}}{n}\right)^{-\frac{1}{2}} & \cdots & v_{1k}\left(\frac{n_{+1}}{n}\right)^{-\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ v_{q1}\left(\frac{n_{+q}}{n}\right)^{-\frac{1}{2}} & \cdots & v_{qk}\left(\frac{n_{+q}}{n}\right)^{-\frac{1}{2}} \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{bmatrix} \quad (12)$$

where $(\lambda_1,...,\lambda_k)$ eigenvalues $(u_{1l},...,u_{pk})$ and $(v_{11},\ldots, v_{qk})$ eigenvectors were calculated from the decomposition of singular values. Where $R_{BL}=r_{ij}=R_{BL}\times R^{T}_{BL}$ and $R_H=r_{ij}^{\star}=R_H\times R^{T}_{H}$ for the line profiles and $R_{BL}=r_{ij}=R_{BL}{}^{T}\times R_{BL}$ and $R_H=r_{ij}^{\star}=R_H{}^{T}\times R_H$ for the column profiles.

Given that there are no theoretical restrictions regarding the incorporation of the residues in obtaining the coordinates, considering frequency tables with independence between the levels of the categorical variables, the cophenetic correlation coefficient was used, in order to compare the similarity between the coordinates obtained from the conventional way and with the incorporation of the residues. Further details of the description of this coefficient are described below.

The proximity analysis of modified coordinates with the incorporation of the residues compared to conventional coordinates was performed by the cophenetic coefficient correlation. The phenetic matrix was determined by S, dissimilarity matrix, where each coordinate as given by the Euclidean distance between the data and the cophenetic array of cophenetic distances to a hierarchical grouping generated by the average connection method, this being used in order to avoid threads. Like this, the cophenetic correlation coefficient is given by Equation 13:

$$r_{cof}=\frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})(s_{ij}-\bar{s})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij}-\bar{s})^2}} \quad (13)$$

where: c_{ij} is the value of similarity between individuals i and j, obtained from the cophenetic matrix and s_{ij} similarity values between individuals i and j obtained from the similarity matrix in which Equation 14:

$$\bar{c}=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}c_{ij} \text{ and}$$
$$\bar{s}=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}s_{ij} \quad (14)$$

To obtain the results, we computed the average of cophenetic correlation coefficient, the total inertia and the scores given 2.000 Monte Carlo achievements.

Results and discussion

The results presented in Table 2 show that the incorporation of residues in the calculation of scores provided an improvement to the dimension tables in the sense that the coordinates had higher correlation to data grouped hierarchically considering the average linkage method. A most promising result was observed when the binomial samples were generated considering high degree of correlation ($\rho=0.8$) for tables larger than 10×10.

Table 2. Best performances of the cophenetic correlation coefficient between Euclidian distance matrices and the cophenetic matrix by the average linkage method average.

Dimension	Binomial Proportion	Degree of correlation (ρ)	Conventional	Barnett and Lewis	Haberman
5×5	0.2	0.2	0.2485	0.6607	0.6488
	0.5	0.8	0.1383	0.8215	0.8246
	0.9	0.5	0.5621	0.7916	0.7977
10×10	0.2	0.2	0.6199	0.1587	0.1554
	0.5	0.8	0.5144	0.4372	0.4388
	0.9	0.5	0.4703	0.3516	0.3513

The correlation between the dissimilarity matrix obtained by Euclidean distance applied to the data and the matrix obtained by the average connection method was higher in the conventional approach when considering the generated binomial samples $\pi=0.2$ in all degrees of correlation.

Regarding the perceptual maps, considering tables of 5×5 (Figure 1) and 10x10 (Figure 2) dimension, we have a case in which the cophenetic correlation coefficient (Table 2) presented a higher evaluation by using the residues.

The results shown in Figure 1 indicate that the incorporation of the residues calculation provided a better discrimination of the objects described in the 'row' and 'column' directions. Increasing the size of contingency tables (Figure 2). It was observed that the centroid was not altered, however, the association between the categorical levels of the objects described in the 'rows' and 'columns' were identified. This result suggests that the increase in the degree of correlation provides a better performance of the correspondence analysis in determining the associations when the incorporation of the residues is used.

In both situations, the results corroborate the recommendations made by Beh (2012), which are based on the violation of the assumption of independence and the variability of points is affected, being reflected in the asymmetry of the quadrants.

Regarding the percentage of the returned inertia on the axes, given the cases illustrated in Figure 1 and 2, by the results in Table 3, it is noted that for the first two axes, the results were suitable as they show high ratios. It is possible to observe that when the size of the contingency table is increased, the explanation ratio is reduced.

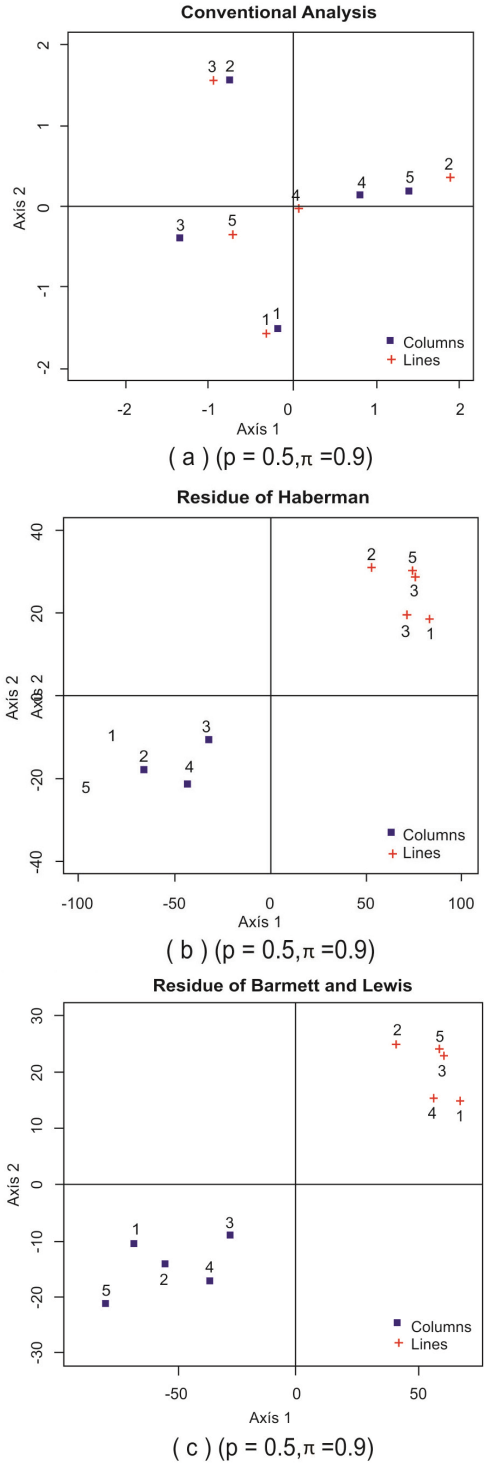


Figure 1. Perceptual Maps generated from the coordinates of conventional correspondence analysis and the incorporation of residues in 5x5dimension tables.

For dimension tables where they obtained better results with the incorporation of residues, the proportion of explanation is higher than in the 10x10 dimension tables where conventional analysis provided better results.

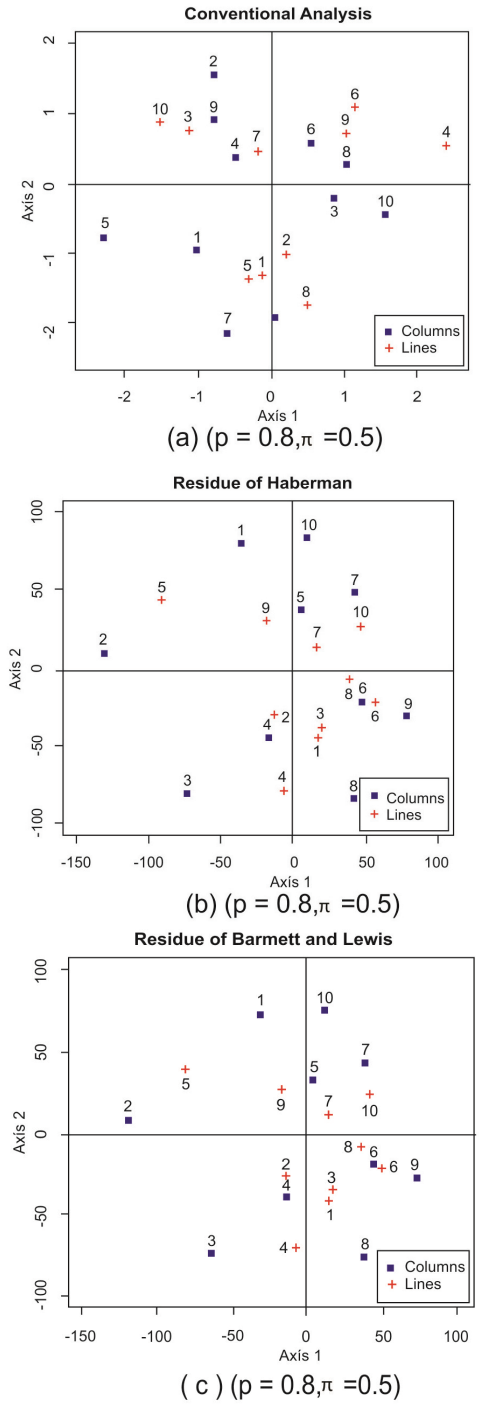


Figure 2. Perceptual maps generated from the coordinates of conventional correspondence analysis, and the incorporation of the residues into size 10x10 table.

Table 3. Decomposition of inertia.

Axis	Best performance parameters					
	(π=0.9; ρ=0.5) (5x5)			(π=0.9; ρ=0.8)(10x10)		
	Inertia	Proportion	Cumulative Proportion	Inertia	Proportion	Cumulative proportion
Axis 1	0.0006	0.5081	0.5081	0.2489	0.3483	0.3483
Axis 2	0.0004	0.3594	0.8675	0.1495	0.2092	0.5575
Axis 3	0.0001	0.1143	0.9818	0.1215	0.1700	0.7274
Inertia	0.0012	–	–	0.7148	–	–

Application example

The classification of coffee for defective and types is made by counting faulty grains or impurities contained in a 300 g sample of processed grains. This classification obeys the Brazilian Official Rate table (BOC) (Abrahão, Pereira, Borém, Rezende, & Barbosa, 2009), according to which each type corresponds to a greater or lesser number of defects found in a sample of coffee. Imperfect grains and impurities (intrinsic and extrinsic defects) are considered defects in coffee classification.

The commercialization of coffee is made according to several classifications in force. The most important are the ratings for sieve, by type, by drink among others. For this example, the database contains information on the intrinsic defects which are those contained in the coffee bean, caused by improper use of agricultural and industrial processes and modifications of physiological or genetic origin, such as black beans, flamed grain, brocade grain, green grain, among others (Bandeira, Toci, Trugo, & Farah, 2009).

It is noticed that many coffee farmers do not care about the defect percentage at the time of preparation of coffee for commercialization, which may impact on the prices charged by the market, and the price is one of the decisive factors in the coffee industry (Wachholz & Poyer, 2014; Brighenti & Cirillo, 2018).

A study on the proportion of defective grains screens is important since the coffee beans are sorted on these sieves. Because they are grown-ups and more valuable grains, coffee exporters have greater interest in them, and the lower the proportion of smaller defective coffee beans lower will be the influence on the purchase price from the producer (Pelupessy, 2007). Table 4 contains information on the proportion of defective coffee beans in a 300 g sample, classified by some type of defect found after grooming in sieves (17/18).

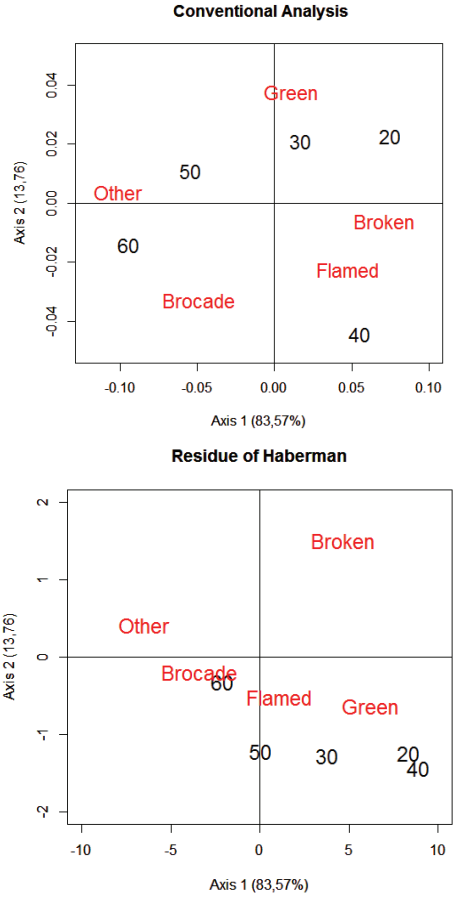
Table 4. Count of grain retained in relation to defects in sieves (17/18).

Proportion of defects in the sample(%)	Grain defect type				
	Green	Flamed	Broken	Brocade	Other
20	18	21	26	21	18
30	11	11	8	8	2
40	25	24	23	18	7
50	7	10	2	11	11
60	3	12	7	22	18

Preliminarily, we applied the chi-square test, in which the hypothesis of independence was rejected (p-value < 0.05). Therefore, we justify an example that violates the basic assumption of correspondence analysis being a situation which suggests that the incorporation of residues can provide some improvement in the identification of associations.

The graphical interpretation of the perceptual maps (Figure 3) is restricted to the first two axes, since they have high inertia values. Regarding the perceptual maps, it is clear that the maps are symmetrical to the centroid and that when residues were used, it has better discrimination between the proportion of grains and their defects, what is perceived to brocades grains, which is strongly related to 60% of the grains and that broken grains which have no association with the ratios studied, in other words are independent proportions.

The broken grain is a defect that occurs by excessive drying coffee bean or also for quick drying and mechanical dryers, which makes it different from other defects that arise usually from genetic or physiological origin and pests. The broken grain is minor compared to other defects, not affecting the quality of coffee. (Schmidt, Miglioranza, & Prudêncio, 2008).



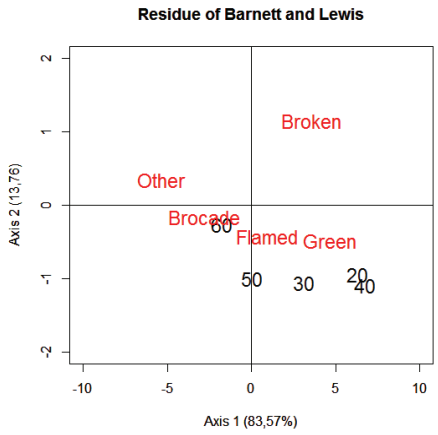


Figure 3. Perceptual Maps generated from the coordinates of conventional correspondence analysis, and the incorporation of residues on the proportion of defective coffee beans in sieves 17/18.

Conclusion

For the simulated scenarios, the results indicated that it is feasible to incorporate the residues in obtaining coordinates, including the situations where that the hypothesis of independence between categories of lines and/or columns were violated. The effect of these residues showed a better discrimination of objects for smaller size table.

The use of the two residues showed similar results, which can be observed by cophenetic correlation coefficients and perceptual maps generated.

The application of simple correspondence analysis modified by the incorporation of residues to real data is feasible because it is perceived, in the studied sample, a better discrimination between the proportion of grains and their defects when compared with conventional simple correspondence analysis.

References

Abrahão, A., Pereira, R., Borém, F., Rezende, J., & Barbosa, J. (2009). Classificação física e composição química do café submetido a diferentes tratamentos fungicidas. *Coffe Science*, 4(2), 100-109. doi 10.25186/cs.v4i2

Bandeira, R. D. C. C., Toci, A. T., Trugo, L., C., & Farah, A. (2009). Composição volátil dos defeitos intrínsecos do café por CG/EM-headspace. *Química Nova, 32*(2), 309-314. doi 10.1590/S0100-40422009000200008

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Chichester, UK: John Wiley & Sons.

Beh, E. J. (2004). Simple correspondence analysis: a bibliographic review. *International Statistical Review, 72*(2), 257-284. doi 10.1111/j.1751-5823.2004.tb00236.x

Beh, E. J. (2012). Simple correspondence analysis using adjusted residuals. *Journal of Statistical Planning and Inference, 142*(4), 965-973. doi 10.1016/j.jspi.2011.11.004

Blasius, J., Greenacre, M., Groenen, P., & Velden, M. V. (2009). Special issue on correspondence analysis and related methods. *Computational Statistics and Data Analysis, 53*(8), 3103-3106. doi 10.1016/j.csda.2008.11.010

Brighenti, C. R. G. & Cirillo, M. A. (2018) Analysis of defects in coffee beans compared to biplots for simultaneous tables *Revista Ciência Agronômica, 49*(1), 62-69. doi 10.5935/1806-6690.20180007

Cirillo, M. A., & Ramos, P. S. (2014). Goodness-of-fit tests for modified multinomial logit model. *Chilean Journal of Statistics, 5*(1), 73-85.

Greenacre, M. (1992). Correspondence analysis in medical research. *Statistical Methods in Medical Research, 1*(1), 97-117. doi 10.1177/096228029200100106

Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton, FL: Chapman and Hall/CRC.

Guedes, T. A., Ivanqui, I. L., Martins, A. B. T., & Cochia, E. B. R. (1999). Seleção de variáveis categóricas utilizando análise de correspondência e análise correspondência e análise procrustes. *Acta Scientiarum. Technology, 21*(4), 861-868. doi 10.4025/actascitechnol.v21i0.3084

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics, 29*(1), 205-220. doi 10.2307/2529686

Lee, A. H., & Yick, J. S. (1999). Theory & methods: a perturbation approach to outlier detection in two-way contingency tables. *Australian & New Zealand Journal of Statistics, 41*(3), 305-315. doi 10.1111/1467-842X.00085

Pelupessy W. (2007). The world behind the world coffee market. *Études rurales, 180*(2), 187-212

Schmidt, C. A. P., Miglioranza, É., & Prudêncio, S. H. (2008). Interação da torra e moagem do café na preferência do consumidor do oeste paranaense. *Ciência Rural, 38*(4), 1111-1117. doi 10.1590/S0103-84782008000400032

Tallis, G. M. (1962). The use of a generalized multinomial distribution in the estimation of correlation in discrete data. *Journal of the Royal Statistical Society. Series B (Methodological), 24*(2), 530-534.

Veloso, M. V., & Cirillo, M. A. (2016). Principal components in the discrimination of outliers: A study in simulation sample data corrected by Pearson's and Yates´s chi-square distance. *Acta Scientiarum. Technology, 38*(2), 193-200. doi 10.4025/actascitechnol.v38i2.26046

Wachholz, L., & Poyer, M. G. (2014). A importância das cooperativas cafeicultoras para os pequenos agricultores na exportação de café na região sul de Minas Gerais. *Revista Gestão & Sustentabilidade Ambiental, 2*(2), 27-44.

Received on January 24, 2017.
Accepted on August 29, 2017.