Bayesian tools for *count zeros* in compositional data sets

Josep Daunis-i-Estadella[1], Josep Antoni Martín-Fernández[2], and Javier Palarea-Albaladejo[3]

[1]Dept. Informàtica i Matemàtica Aplicada, UdG. Campus Montilivi Edif. P-4 E-17071 Girona; *josep.daunis@udg.edu*

[2]Dept. Informàtica i Matemàtica Aplicada, UdG.

[3]Dep. Informática de Sistemas, Univ. Católica San Antonio, Murcia

Abstract

The log-ratio methodology makes available powerful tools for analyzing compositional data. Nevertheless, the use of this methodology is only possible for those data sets without null values. Consequently, in those data sets where the zeros are present, a previous treatment becomes necessary. Last advances in the treatment of compositional zeros have been centered especially in the zeros of structural nature and in the rounded zeros. These tools do not contemplate the particular case of count compositional data sets with null values. In this work we deal with "count zeros" and we introduce a treatment based on a mixed Bayesian-multiplicative estimation. We use the Dirichlet probability distribution as a prior and we estimate the posterior probabilities. Then we apply a multiplicative modification for the non-zero values. We present a case study where this new methodology is applied.

Key words: count data, multiplicative replacement, composition, log-ratio analysis

1 Introduction

It is very important to realize that the well-known 'zeros problem' in compositional data is inherent in the nature of the data rather than in the log-ratio methodology (Aitchison, 1982). Certainly, it is obvious that if some sample has null values nor ratios neither logarithms can be formed. Therefore, it does not take into account the nature of the null value, it may be preferable to apply classical multivariate methods based on Euclidean distance because this methodology has not the zeros problem. Nevertheless, that kind of reasoning is too much simple and incomplete.

Martín-Fernández et al. (2000; 2003) proposed take into account the nature of the null value, and the first question is to decide if the zero value is a true value or not. If it is considered as a true value then it is informative by itself. Therefore, this null value means the absolute absence of the part in the observation, i.e. the null value is an **essential or structural zero**. On the other hand, if the null value indicates the presence of a component, but below the detection limit, then this zero represents a missing small value, i.e. null values are **rounded zeros**. Since the nature of the two kinds of zeros is different the treatment should be different. Note that this different treatment is a consequence of the nature of the zero rather than of the statistical methodology (Euclidean, log-ratio, ...).

1.1 Existing types and treatment of zeros

In the structural zeros case, two initial questions must be answered:

1. Is the number of parts too large for the goals of the study?
2. Is the presence in a part of an essential zero an indication that the composition belongs to a different group or population?

An affirmative answer to the first question is related to the sampling or measuring step and suggests amalgamating some related parts (Aitchison, 1986). This amalgamation procedure reduces the dimensionality and probably the amount of null values. The answer to the second question is related to the true information of the null value. For example, an affirmative answer could indicate that it is sensible to divide the sample, and then a statistical analysis of any kind would be applied to each sub-sample separately. After both questions have been solved, and after data have been closed, the statistical analysis can be applied. But, now the question is: which one? Nowadays, in the context of compositional studies, most scientists apply either Euclidean or log-ratio statistical methods. Obviously, when a scientist wrongly chooses the Euclidean option then there are no problems with the remaining structural zeros but he don't is taking into account the compositional nature of the data.

On the other hand, selection of the log-ratio methodology involves a weakness with the null values. Fortunately, recent works (Aitchison and Kay, 2003; Bacon-Shone, 2003) offer new strategies for the application of the log-ratio methodology combined with conditional models and subcompositional analysis.

When the null values denote that no quantifiable proportion could be recorded due to the accuracy of the measurement process, then this kind of zero is usually understood as 'a trace too small to measure' or **rounded zero**. Note that these null values actually are missing values rather than zeros (Martín-Fernández et al., 2003). Consequently, it seems reasonable to apply specific techniques for a statistical analysis of incomplete multivariate data (Little and Rubin, 2002). Following this point of view, Martín-Fernández et al. (2003) analyzed one multiplicative replacement and suggested that when the proportion of these null values is not large (less than 10% of the values in data matrix) a replacement method which uses an imputation value equal to 65% of the threshold value can be used. Then after this imputation a multiplicative modification of the non-zero values is applied. This kind of modification was suggested independently by Fry et al. (2000) and

Martín-Fernández et al. (2000). In Martín- Fernández et al. (2003) multiplicative replacement was analysed in depth.

1.2 A new type of zeros: zeros in count data sets

Let the sample space Ω be an exhaustive set of categories ω_j that are mutually exclusive. We assume that any event of interest can be identified with a subset of Ω, and that each observation falls into exactly one category ω_j. Let we assume the standard multinomial model: N observations are independently chosen from Ω with an identical probability distribution $P(\omega_j) = \theta_j$ for $j = 1, 2, \ldots, k$, where each $\theta_j \geq 0$ and $\sum_{j=1}^{k} \theta_j = 1$. Let n_j denote the number of observations of category ω_j in the N trials, so that n_j is a non-negative integer and $\sum_{j=1}^{k} n_j = N$. To simplify the notation, we write $\omega = (\omega_1, \omega_2, \ldots, \omega_k)$ and $n = (n_1, n_2, \ldots, n_k)$.

When we deal with count data we are assigning each one of the individuals to each one of the previously defined categories or parts ω_j and after we count the total number of individuals in each class or category. Then, when we have a problem related to a zero, and we try to answer the question that if it is an essential or structural zero, we can decide that nor structural neither essential. We can also assume that no detection limit exists and that it is possible to take values in this part, even more if we take a sample with more observations (N). Then, when we have zeros in count data sets, we have a new type of zero related to a sampling problem: parts are unobserved due to the limited size of the sample.

So, our question consists on estimating the values of θ_j parameters, in order to have the true values of $n_j = N\theta_j$, breaking down the sample limitations.

2 Proportion estimation

We can deal with the estimation of the true values of θ_j, which is equivalent to n_j, under the two main philosophies: the frequentist philosophy and the bayesian one. But in this case the classical frequentist perspective doesn't help us to solve the problem, because the estimation of the parameter θ_j is again 0, and it persists the log-ratio problems related to zeros. Then our solution (Daunis-i-Estadella and others, 2008) may be found on the field of bayesian methodology, that we reference as classical in order to distinguish with our proposal based on bayesian estimation but according to the multiplicative replacement (Martín-Fernández et al., 2003)

2.1 Bayesian estimation

The classical bayesian estimation of the probabilities is based in the likelihood funcion and their conjugate prior. The Dirichlet distribution is the conjugate prior of the multinomial distribution, and it is a multivariate generalization of the beta distribution.

Prior uncertainty about θ is expressed by

$$\theta \sim Diri(st),$$

the Dirichlet prior, with hyper-parameters: s, the total prior strength; and the prior expectation of θ, $t = (t_1, t_2, ..., t_k)$, with $t_j > 0$, $\sum_j t_j = 1$. Note that t belongs to the simplex S^k. We denote $\alpha_j = st_j$ the prior strength of class j. It can be easily shown that the prior expectations of θ_j is $\mathrm{E}(\theta_j) = t_j$.

Then from Bayes' theorem, the posterior distribution is calculated from the prior $p(\theta)$ and the likelihood function:

$$L(\theta, n) \propto \prod_{j=1}^{k} \theta_j^{n_j}$$

The Dirichlet posterior uncertainty about $\theta|x$ is expressed by

$$\theta|x \sim Diri(x + st)$$

Therefore, the posterior expectation is

$$E(\theta_j|x) = \frac{x_j + \alpha_j}{N + s} = \frac{x_j + st_j}{N + s}$$

Almost all proposed priors (Bernard, 2005) for fixed N are symmetric Dirichlet priors, i.e. $t_j = 1/k$:

Table 1: Proposed Dirichlet priors

Haldane	$s = 0$	$\alpha_j = 0$
Perks	$s = 1$	$\alpha_j = \frac{1}{k}$
Jeffreys	$s = \frac{k}{2}$	$\alpha_j = \frac{1}{2}$
Bayes-Laplace	$s = k$	$\alpha_j = 1$

There are other priors, like the Berger-Bernardo reference priors, or based on the imprecise Dirichlet model. Several priors are proposed for prior ignorance, but none of them satisfies all desirable principles (Bernard, 2005):

- Inferences often depend on Ω and/or k
- Some solutions violate the Likelihood principle (LP) where inferences should depend on the data through the likelihood function only.
- Inferences about various derived parameters can be incoherent
- Prior uncertainty should depend on refinements or coarsenings of categories.

In addition, from the compositional perspective, there is another problem:

- Doesn't preserve ratios between non-zero parts.

Moreover, when the sample size is small (N small), these changes in the ratios may be large ones. For example, consider that we have a sample size $N = 5$ with $k = 3$ classes, with $(3, 0, 2)$ observations in each class. The ratio between the first and third elements equal to 1.5. After the classical bayesian estimation is applied, we have:

$$\left(\frac{3 + \alpha_j}{5 + s}, \frac{0 + \alpha_j}{5 + s}, \frac{2 + \alpha_j}{5 + s}\right).$$

Observe that the ratio between the first and the third elements may be fewer than 1.5 (e.g. 1.33 with the Bayes-Laplace prior), but not necessary equal to 1.5.

Then, the bayesian prior technique must be improved in order to achieve the ratio preservation.

2.2 New proposal of estimation

Our proposal (Daunis-i-Estadella and others, 2008) is based on the multiplicative replacement analyzed in depth by Martín- Fernández (2003). Furthermore, Martín- Fernández and Thió-Henestrosa (2006) revisited this approach from a theoretical point of view and the authors compared its properties in relation to the properties of the additive and simple replacements. Using this methodology

combined with the bayesian estimation one can preserve the ratios between non-zero parts. This strategy consists on replace the zero and no zero percentages as follows:

$$\begin{cases} x_j^* = \frac{\alpha_j}{N+s} & x_j = 0, \\ x_j^* = x_j(1 - \sum_{x_k=0} \frac{\alpha_k}{N+s}) = x_j(1 - \frac{s}{N+s} \sum_{x_k=0} t_k) & x_j > 0. \end{cases} \tag{1}$$

In Equation (1) we can see that all zero percentages are replaced by its posterior expectation and the non-zero percentages are multiplied by a factor according to the number of zero counts.

In our example, with sample size $N = 5$ and $k = 3$ classes, with $(3, 0, 2)$ observations in each class or that is equivalent

$$\left(\frac{3}{5}, 0, \frac{2}{5}\right).$$

With the multiplicative replacement estimation and Bayes-Laplace prior ($s = 3, \alpha_j = 1$), we have:

$$\left(\frac{3}{5}(1 - \frac{3}{8}\frac{1}{3}), \frac{1}{8}, \frac{2}{5}(1 - \frac{3}{8}\frac{1}{3})\right) = \left(\frac{21}{40}, \frac{5}{40}, \frac{14}{40}\right),$$

where the ratio between the first and third elements is preserved: $\frac{21}{40}/\frac{14}{40} = 1.5$.

3 Case study

The largest amount of information about the welfare of animals, in this case are sows, is obtained from the measures of behaviour, particularly measures of activity and stereotypies. Stereotypies are related to poor welfare because they are developed in situation of stress, frustration or lack of control. They reflect a past or present difficulty to cope with the environment. Therefore, the decrease in stereotypies level in group-housing systems could already be considered as a welfare improvement.

The data used in this study are obtained from the comparison among two different commercial housing and feeding systems (trickle feeding and electronic sow feeder) and conventional stalls for pregnant sows. One hundred and eighty pregnant sows were selected on a commercial farm and used in three different replicas (60 sows per replica). In each replica, 20 sows were housed in conventional stalls (Stall), 20 sows were observed using the trickle feeding system (Trick) and 20 sows more using the electronic sow feeder system (Fitmix).

Sows were observed for 11 non-consecutive days during 4 hours per day. General activity and stereotypies were measured by scan-sampling observation (10-min intervals) in all the systems. The final data set contain information about 177 individuals as 3 of the 180 initially selected sows cannot conclude the study.

Our data are a 5-part vectors containing the observed frequencies of 4 oronasofacial behaviours: interacion with the equipment (E), floor manipulation (T), drinking (D), sham-chewing (S) and one residual part (h). This data are previously studied in Chapinal (2006) and Daunis-i-Estadella and others (2006a).

We have sample sizes between 270 and 305 observations varying in each sow, with 150 complete data, 24 compositions with on zero in one part and 3 data with doble zeros. Zeros belong mainly to D part (23), 4 zeros to BCI part and 3 zeros to T part. Parts S and h are zero free.

The zero pattern is resumed in the Table 2.

We apply the mixed multiplicative bayesian estimation with $s = 5$ and $\alpha_j = 1$. After that, we apply

Table 2: Tabulated zero pattern

Count	D	BCI	T	S	h
150					
21		0			
2	0				
1			0		
1	0	0			
1	0		0		
1		0	0		

log-ratio methodology for describing the proportion spent in each activity (Daunis-i-Estadella and others, 2006b).

We start the multivariate analysis of the behaviour data with the compositional biplot (Aitchison and Greenacre, 2000). Biplot shows the variables and the sample pattern in the same plot which facilitates the interpretation of sample clusters in relation to variables.

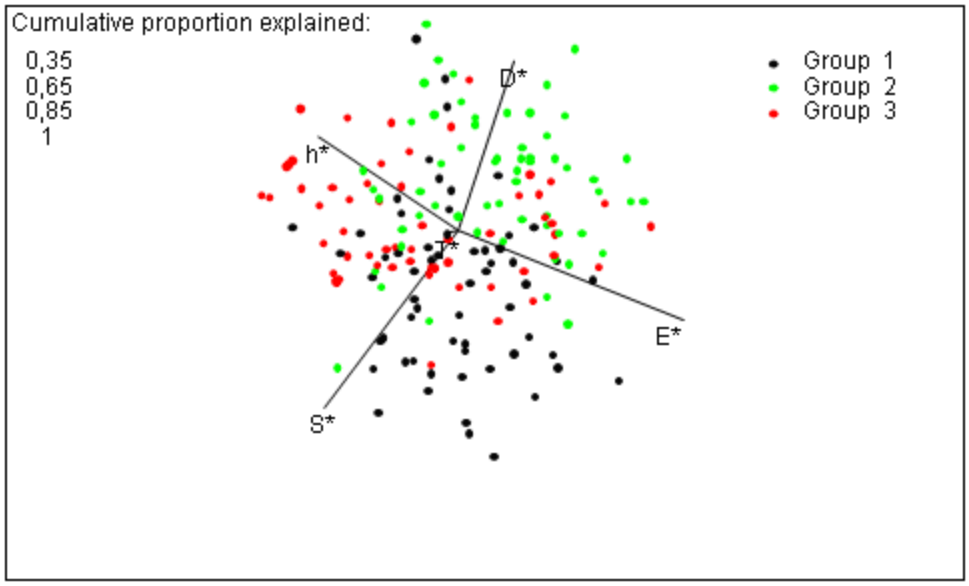

Figure 1: Biplot ot the behaviour data set

The variation array of the complete data set (Table 3) provides us information of the relative variation and the variation of the clr-transformed data. It shows near equal clr-variances for all the parts which are related to the length of the vectors projected in the biplot. None of the variances of the logratios of the parts is close to zero which indicates that these none of the components are proportional.

The total variability $2,6778$ is decomposed and visualized into the biplot of the data set and the quality or percentage of variation explained is near 65%.

We also can compute the compositional center of the whole data set:

$$(E^*, D^*, T^*, S^*, h^*) = (0.0564, 0.0150, 0.0421, 0.1202, 0.7663)$$

And also the center of each group in order to describe differences among centers:

Table 3: Variation array of behaviour data: log-ratio variances and means, and clr-variances

	BCI*	D*	T*	S*	h*		CLR Variance
BCI*		1,3598	1,2315	1,6442	1,7758		0,5121
D*	1,3228		1,1775	1,6366	1,1311		0,6011
T*	0,2935	-1,0292		1,2186	0,9812		0,5306
S*	-0,7557	-2,0784	-1,0492		1,2325		0,4609
h*	-2,6082	-3,9310	-2,9018	-1,8526			0,5732
						Tot var	2,6778

Group 1: $(E^*, D^*, T^*, S^*, h^*) = (0.0728, 0.0156, 0.0438, 0.2504, 0.6173)$

Group 2: $(E^*, D^*, T^*, S^*, h^*) = (0.0714, 0.0263, 0.0416, 0.0818, 0.7789)$

Group 3: $(E^*, D^*, T^*, S^*, h^*) = (0.0303, 0.0072, 0.0368, 0.0749, 0.8508)$

A first approach seems to indicate that there are differences related to the behaviour in the three methods.

4 Conclusions

A new typology of zero is introduced in compositional data: count zeros.

Since frequentist estimations and classical bayesian do not preserve the proportions among ratios a new proposal of mixed bayesian-multiplicative estimation is presented based on the bayesian methodology and the multiplicative replacement.

The new proposal preserves the proportions among ratios and it allows the application of the log-ratio techniques.

Acknowledgements and appendices

This work has been supported by the Spanish Ministry of Education and Science under project 'Ingenio Mathematica (i-MATH)' No. CSD2006-00032 (Consolider – Ingenio 2010) and under project MTM2006-03040.

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B*, **44(2)**, 139-177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J., and Greenacre, M. (2002) Biplots of compositional data *Applied Statistics*, **51(4)** 375-392.

Aitchison, J. and Kay, J. W. (2003) Possible solutions of some essential zero problems in compositional data analysis. See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Bacon-Shone, J. (2003) Modelling structural zeros in compositional data. See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Bernard, J.M. (2005) An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning* **39**, Issues 2-3, 123-150

Chapinal, N. (2006). *Effect of the housing and feeding system on the welfare and productivity of pregnant sows.* Ph-D Thesis. Dept. Ciència dels Animals i dels Aliments. UAB-Barcelona

Daunis-i-Estadella, J., Mateu-Figueras, G., Chapinal, N., Manteca, X. and Ruiz de la Torre, J. L.(2006a). Aplicación de técnicas composicionales al estudio del comportamiento de cerdas gestantes. In: *Libro de actas del XXIX Congreso de la SEIO.*

Daunis-i-Estadella, J., Barceló-Vidal, C. and Buccianti, A.(2006b). Exploratory compositional data analysis. In: *Compositional Data Analysis in the Geosciences: From Theory to Practice* (eds: A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn), Geological Society, London, Special Publications, 264, 161–174.

Daunis-i-Estadella, J., Martín-Fernández, J. A., and Palarea-Albaladejo, J., (2008). Compositional count zero: what it is and how to deal with it. *Appl. Economics* (**in prep.**).

Fry, J. M., Fry, T. R. L., and McLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Appl. Economics* **32**, 953-959.

Little, R. J. A. and Rubin, D.B. (2002) *Statistical analysis with missing data* Wiley & Sons, New York 381 p.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2000). Zero replacement in compositional data sets. In H. Kiers, J. Rasson, P. Groenen and M. Shader (eds.) Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000), Berlin, Springer-Verlag, 155-160.

Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2003) Dealing with Zeros and Missing Values in Compositional Data Sets. *Mathematical Geology*, **35(3)**, 253-278.

Martín-Fernández, J. A. and Thió-Henestrosa, S. (2006) Rounded zeros: Some practical aspects. In: Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (eds) Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, Special Publications, 264, 191–201. The Geological Society of London.