Compositional data and Simpson's paradox

V. Pawlowsky-Glahn[1] **, J. J. Egozcue**[2]
[1]Dep. Informàtica i Matemàtica Aplicada, Universitat de Girona, Girona, Spain;
vera.pawlowsky@udg.edu
[2]Dep. Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona, Spain;
juan.jose.egozcue@upc.edu

Abstract

Simpson's paradox, also known as amalgamation or aggregation paradox, appears when dealing with proportions. Proportions are by construction parts of a whole, which can be interpreted as compositions assuming they only carry relative information. The Aitchison inner product space structure of the simplex, the sample space of compositions, explains the appearance of the paradox, given that amalgamation is a non-linear operation within that structure. Here we propose to use balances, which are specific elements of this structure, to analyse situations where the paradox might appear. With the proposed approach we obtain that the centre of the tables analysed is a natural way to compare them, which avoids by construction the possibility of a paradox.

Key words: Aitchison geometry, geometric mean, orthogonal projection.

1 Introduction

Simpson's paradox, also known as amalgamation (Good and Mittal, 1987) or aggregation (Haunsperger, 2003) paradox, can be traced back to Pearson (1899) and Yule (1903). It is related to the issue of collapsibility in a multiway contingency table with strongly correlated factors (Aitkin, 1998) and has been recently described in the context of the Kruskal-Wallis nonparametric test (Haunsperger, 2003) and of survival analysis (DiSerio et al., 2007). The essence of the paradox is succinctly described in the following paragraph (Wikipedia, 2006):

> Simpson's paradox (or the Yule-Simpson effect) is a statistical paradox described by E. H. Simpson in 1951 and G. U. Yule in 1903, in which the successes of several groups seem to be reversed when the groups are combined. This seemingly impossible result is encountered surprisingly often ...

In practice, the paradox appears when analysing e.g. the rate of success of two treatments, and the question to be answered refers both to the rate of success in each of several subpopulations and to the *overall rate of success*; or when analysing the behaviour (success/failure) of males and females in several different situations and the question refers to their behaviour in each of the several situations and to their *overall behaviour*. These apparently simple questions have a point of ambiguity, namely what shall be understood under *overall rate of success* or under *overall behaviour* or, in other terms, how shall we *combine* the groups.

Here we briefly review the classic approach (Section 2), and use the fact that the paradox deals with proportions to propose both an explanation and an alternative based on the methodology developed in recent years for the statistical analysis of compositional data (Section 3). Finally, we illustrate the similarities and differences between both approaches using real data (Section 4).

2 Classic approach

As presented by Good and Mittal (1987), formally the paradox can be stated in the following terms. Consider a population divided into n mutually exclusive subpopulations. For each of those subpopulations consider a two-by-two contingency table whose entries in reading order are a_i, b_i, c_i, d_i, with $a_i + b_i + c_i + d_i = N_i$, the corresponding sample size. Denote the tables by $\mathbf{a_i} = [a_i, b_i; c_i, d_i]$, and suppose $a_i b_i c_i d_i \neq 0$ and N_i so large that sampling variation can be ignored. Let $N = \sum_{i=1}^{n} N_i$ denote the total sample size for the whole population. If the n tables are added together (amalgamated), a new table is obtained,

$$\mathbf{A} = [A, B; C, D] = \left[\sum_{i=1}^{n} a_i, \sum_{i=1}^{n} b_i; \sum_{i=1}^{n} c_i, \sum_{i=1}^{n} d_i \right] .$$

Of course, $N = A + B + C + D$ and $ABCD \neq 0$. For each subpopulation, N_i is either assumed to be proportional to the fraction of the population corresponding to it or the table has been scaled to force it. Consider furthermore a measure of association, i.e. a function of a_i, b_i, c_i, d_i, respectively A, B, C, D, denoted by $\alpha(\mathbf{a_i})$, respectively $\alpha(\mathbf{A})$. Then, the paradox occurs if

$$\max_i \alpha(\mathbf{a_i}) < \alpha(\mathbf{A}) \quad \text{or} \quad \alpha(\mathbf{A}) < \min_i \alpha(\mathbf{a_i}) \,, \tag{1}$$

i.e. if $\alpha(\mathbf{A})$ falls out of the range of the $\alpha(\mathbf{a_i})$.

Good and Mittal (1987) claim that such a situation only can arise if not enough care is used in the design of an experiment, and state—for several meaningful measures of association—what the sampling design should be in order to avoid it.

This approach is in itself a specification of how to understand the terms *overall* or *combining* the groups mentioned in the introduction, as they consider amalgamation to be the proper operation

for answering the question. But amalgamation leads to difficult situations. To illustrate what we mean, let us consider two possible probabilistic models.

(i) N_i individuals of each subpopulation $\mathbf{a_i}$ are sampled in a multinomial experiment with parameters (probabilities) $p_{ai}, p_{bi}, p_{ci}, p_{di}$, $i = 1, 2, \ldots, n$, with $p_{ai} + p_{bi} + p_{ci} + p_{di} = 1$. Then the samples in the subpopulations are added to obtain the frequencies A, B, C, D.

(ii) $N = \sum_i N_i$ individuals of the overall population are sampled in a multinomial experiment with parameters (probabilities) $p_{\kappa i}$, $\kappa = a, b, c, d$, $i = 1, 2, \ldots, n$, with $\sum_\alpha \sum_i p_{\alpha i} = 1$.

Case (i) corresponds to the situation in which Simpson's paradox appears frequently. The counts in the overall population A, B, C, D constitute a sample of a sum of multinomial variables. A sum of independent multinomial variables is only again multinomial whenever their probabilities are equal, i.e. $p_{\kappa 1} = p_{\kappa 2} = \cdots = p_{\kappa n}$ for $\kappa = a, b, c, d$. But this trivial case is rarely encountered in practice. Contrarily, when the probabilities are not equal, the sum of multinomial variables is no longer multinomial but a distribution where $p_{\kappa i}$, $\kappa = a, b, c, d$, $i = 1, 2, \ldots, n$, are parameters. These parameters are not related to the probabilities p_A, p_B, p_C, p_D of an individual of the overall population to be identified in the respective category. Moreover, to obtain a realisation of the corresponding random variable we need to select at least one individual within each subpopulation, otherwise the sum cannot be performed. The approximation of the distribution of the sum of multinomial variables by a single multinomial variable with parameters $\hat{p}_A = \sum_i a_i / \sum_i N_i$ is certainly possible, but there is no guarantee of being accurate and it may produce the paradoxical results.

To analyse case (ii), let a_i^e, …, d_i^e denote the events that produce the counts in a_i, …, d_i, and $A^e = \{\bigcup_{i=1}^n a_i^e\}$, …, $D^e = \{\bigcup_{i=1}^n d_i^e\}$ the compound events. The total probability theorem gives

$$p_{A^e} = \sum_{i=1}^{n} \frac{p_{ai}}{P[\mathbf{a_i}]} \cdot P[\mathbf{a_i}] = \sum_{i=1}^{n} p_{ai} \quad , \quad P[\mathbf{a_i}] = p_{ai} + p_{bi} + p_{ci} + p_{di} \ ,$$

and similarly for p_{B^e},p_{C^e}, p_{D^e}. Therefore, the marginal is a multinomial with parameters p_{A^e}, p_{B^e},p_{C^e}, p_{D^e}. Estimates of these probabilities are given by the frequencies

$$\hat{p}_{A^e} = \sum_{i=1}^{n} \frac{a_i}{a_i + b_i + c_i + d_i} \frac{a_i + b_i + c_i + d_i}{N} = \frac{1}{N} \sum_{i=1}^{n} a_i \ ,$$

and similar expressions for B^e, C^e, D^e. These probabilities and their estimated values are weighted averages of the probabilities, respectively of the estimates, in the subpopulations. The paradoxical situations are obviously avoided, but the assumptions of this sampling scheme seldom hold.

Contingency tables are not only obtained as a result of a probabilistic approach, with the corresponding sampling design, but frequently also related to the extraction of information contained in a census. In any case the paradox might appear, and therefore it can be important to have a different look at the data, as described in the next section.

3 Compositional approach

Given that the paradox appears when dealing with proportions, and that proportions are by construction parts of a whole, it is clear that they can be interpreted as compositions assuming they only carry relative information. Recall that the usual representation of compositions as adding up to some constant corresponds to the idea of selecting a representant of an equivalence class (Barceló-Vidal et al., 2001), and that the Aitchison geometry of the simplex is scale invariant (Pawlowsky-Glahn and Egozcue, 2001). Consequently, we can consider the proportions in each of the n subpopulation as compositions. Moreover, we can consider the proportions in each of the

n subpopulations to be subcompositions of a composition with $4n$ parts with randomly observed subpopulations. From the proportions, estimations of the corresponding multinomial probabilities, p_{ai}, p_{bi}, p_{ci}, p_{di}, can be obtained. Zeroes can be avoided using for instance $\tilde{p}_{ai} = (a_i+1/2)/(N_i+2)$; i.e. taking $\tilde{a}_i = a_i + 1/2$, $\tilde{b}_i = b_i + 1/2$, $\tilde{c}_i = c_i + 1/2$, and $\tilde{d}_i = d_i + 1/2$.

For a systematic representation of the compositional approach, consider the previous information as represented in table 1. For the sake of simplicity, given that for each subpopulation we have a 2×2 table, one dimension is called the *treatment* dimension, labelled T_1 and T_2, and the other dimension *success*, S, and *failure*, F. Each combination of subpopulation, treatment and success or failure is called a *category*.

Table 1: Representation of contingency tables as a single composition. For each subpopulation we have a table $\mathbf{a_i}$, whose categories correspond to two treatments T_j, $j = 1, 2$, for which we have either success S or failure F.

$\mathbf{a_1}$				$\mathbf{a_2}$				...	$\mathbf{a_n}$			
T_1		T_2		T_1		T_2		...	T_1		T_2	
S	F	S	F	S	F	S	F	...	S	F	S	F
$\tilde{a}_1$	$\tilde{b}_1$	$\tilde{c}_1$	$\tilde{d}_1$	$\tilde{a}_2$	$\tilde{b}_2$	$\tilde{c}_2$	$\tilde{d}_2$	...	$\tilde{a}_n$	$\tilde{b}_n$	$\tilde{c}_n$	$\tilde{d}_n$

In the following we are going to use the fact that the simplex, the sample space of compositional data, has a Euclidean space structure (Barceló-Vidal et al., 2001; Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). The operations which underly such a structure—perturbation, powering, Aitchison inner product—reflect the relative character of the information contained in compositional data. With their use it is relatively simple to show that amalgamation is not a linear operation in the simplex (Egozcue and Pawlowsky-Glahn, 2005), which explains the Simpson paradox appearing in so many different situations.

To analyse data in a Euclidean space, the best is to represent them with respect to some basis. If that basis is orthonormal, we can use standard operations in real space on the coordinates in a straightforward way. The most convenient way to build such a basis in the simplex such that it is not only orthonormal, but also interpretable, is to use balances as described by Egozcue and Pawlowsky-Glahn (2005). They require a sequential binary partition (SBP), based on previous information, which we describe below for different questions related to the paradox. For example, to answer the question on how the pattern of success and failure for each treatment can be compared within each subpopulation and between subpopulations, we would use an extended version of the SBP given in Table 2, where only two subpopulations have been considered for illustration. The

Table 2: Sequential binary partition to study the pattern of success and failure for each treatment within each subpopulation and between subpopulations (see text for details).

order	$\tilde{a}_1$	$\tilde{b}_1$	$\tilde{c}_1$	$\tilde{d}_1$	$\tilde{a}_2$	$\tilde{b}_2$	$\tilde{c}_2$	$\tilde{d}_2$
1	+	+	+	+	−	−	−	−
2	+	+	−	−	0	0	0	0
3	0	0	0	0	+	+	−	−
4	+	−	0	0	0	0	0	0
5	0	0	+	−	0	0	0	0
6	0	0	0	0	+	−	0	0
7	0	0	0	0	0	0	+	−

strategy consists in separating first the subpopulations in a sequential way. This requires $n-1$ steps for n subpopulations (order 1 in Table 2). Then, within each subpopulation, treatment T_1 is separated from treatment T_2. This requires n steps (order 2 and 3 in Table 2). Finally, within each treatment, success (S) is separated from failure (F). This requires $2n$ steps (order 4 to 7 in Table 2). Thus, the total number of steps required is, as expected $4n-1$ ($4 \cdot 2 - 1 = 7$ in Table 2). Note that the $+$ and $-$ signs in Table 2 indicate which parts are involved, while the 0 indicates that the corresponding part does not play any role in the considered order of partition (Egozcue and Pawlowsky-Glahn, 2005). The coordinates or balances which are of interest to the question posed are the $2n$ last ones, that is

$$\alpha(\mathbf{a_1};T_1) = \frac{1}{\sqrt{2}} \ln \frac{\tilde{a}_1}{\tilde{b}_1} \; ; \; \alpha(\mathbf{a_1};T_2) = \frac{1}{\sqrt{2}} \ln \frac{\tilde{c}_1}{\tilde{d}_1} \; ; \; \alpha(\mathbf{a_2};T_1) = \frac{1}{\sqrt{2}} \ln \frac{\tilde{a}_2}{\tilde{b}_2} \; ; \; \alpha(\mathbf{a_2};T_2) = \frac{1}{\sqrt{2}} \ln \frac{\tilde{c}_2}{\tilde{d}_2} \; ,$$

which can directly be understood as measures of association for each treatment within each subpopulation, and hence the notation. They allow a straightforward comparison of rates of success within the subpopulations to check for differences between treatments, or between subpopulations to check for the same or different treatments. In the traditional version of the Simpson's paradox success and failure are compared for treatments T_1 and T_2, for instance, observing that $\tilde{a}_i/\tilde{b}_i < \tilde{c}_i/\tilde{d}_i$ in each subpopulation $\mathbf{a_i}$. This can be readily expressed as

$$\alpha(\mathbf{a_1};T_1) < \alpha(\mathbf{a_1};T_2) \; ; \; \alpha(\mathbf{a_2};T_1) < \alpha(\mathbf{a_2};T_2) \; . \tag{2}$$

These inequalities are intended to be compared to a similar one for the overall population. A measure of association over the whole population can be obtained as a balance corresponding to the SBP represented in Table 3. In Table 3, first the two treatments are separated, and then,

Table 3: Sequential binary partition to study the overall pattern of success and failure separated by treatment (see text for details).

order	$\tilde{a}_1$	$\tilde{b}_1$	$\tilde{c}_1$	$\tilde{d}_1$	$\tilde{a}_2$	$\tilde{b}_2$	$\tilde{c}_2$	$\tilde{d}_2$
1	$+$	$+$	$-$	$-$	$+$	$+$	$-$	$-$
2	$+$	$-$	0	0	$+$	$-$	0	0
3	0	0	$+$	$-$	0	0	$+$	$-$
4	$+$	0	0	0	$-$	0	0	0
5	0	$+$	0	0	0	$-$	0	0
6	0	0	$+$	0	0	0	$-$	0
7	0	0	0	$+$	0	0	0	$-$

within each treatment, success is separated from failure, leading to the balances or measures of association of interest,

$$\alpha(T_1) = \frac{1}{2} \ln \frac{\tilde{a}_1 \tilde{a}_2}{\tilde{b}_1 \tilde{b}_2} \; ; \quad \alpha(T_2) = \frac{1}{2} \ln \frac{\tilde{c}_1 \tilde{c}_2}{\tilde{d}_1 \tilde{d}_2} \; .$$

Note that the above measures of association satisfy the following properties:

$$\begin{aligned} \alpha(T_1) &= \frac{1}{\sqrt{2}} \left(\alpha(\mathbf{a_1};T_1) + \alpha(\mathbf{a_2};T_1) \right) \; ; \\ \alpha(T_2) &= \frac{1}{\sqrt{2}} \left(\alpha(\mathbf{a_1};T_2) + \alpha(\mathbf{a_2};T_2) \right) \; . \end{aligned}$$

Whenever inequalities (2) hold, then

$$\alpha(T_1) < \alpha(T_2) \; ,$$

thus avoiding any paradoxical result. This situation is related to the principle of subcompositional dominance—a property which is characteristic of compositional data analysis (Aitchison, 1986)—extended to orthogonal projections in general (Egozcue and Pawlowsky-Glahn, 2005).

Similarly, success and failure in subpopulations can be compared to overall success and failure when the treatment is not taken into account.

$$\alpha(\mathbf{a_1}) = \frac{1}{2}\ln\frac{\tilde{a}_1\tilde{c}_1}{\tilde{b}_1\tilde{d}_1}\ ; \quad \alpha(\mathbf{a_2}) = \frac{1}{2}\ln\frac{\tilde{a}_2\tilde{c}_2}{\tilde{b}_2\tilde{d}_2}\ .$$

As before, they can be interpreted as measures of association, as they represent the overall rates of success within each subpopulation. These measures of association are balances corresponding to the SBP in Table 4. To compare with the overall population consider the alternative SBP in Table 5 which gives the corresponding measures of association over the whole population. In Table 5

Table 4: Sequential binary partition to study the pattern of success and failure within each subpopulation for comparison between subpopulations independently of the treatment (see text for details).

order	$\tilde{a}_1$	$\tilde{b}_1$	$\tilde{c}_1$	$\tilde{d}_1$	$\tilde{a}_2$	$\tilde{b}_2$	$\tilde{c}_2$	$\tilde{d}_2$
1	+	+	+	+	−	−	−	−
2	+	−	+	−	0	0	0	0
3	0	0	0	0	+	−	+	−
4	+	0	−	0	0	0	0	0
5	0	+	0	−	0	0	0	0
6	0	0	0	0	+	0	−	0
7	0	0	0	0	0	+	0	−

Table 5: Sequential binary partition to study the overall pattern of success and failure independently of treatment (see text for details).

order	$\tilde{a}_1$	$\tilde{b}_1$	$\tilde{c}_1$	$\tilde{d}_1$	$\tilde{a}_2$	$\tilde{b}_2$	$\tilde{c}_2$	$\tilde{d}_2$
1	+	−	+	−	+	−	+	−
2	+	0	+	0	−	0	−	0
3	0	+	0	+	0	−	0	−
4	+	0	−	0	0	0	0	0
5	0	0	0	0	+	0	−	0
6	0	+	0	−	0	0	0	0
7	0	0	0	0	0	+	0	−

partition of order one responds to the question posed, as it separates success from failure, leading to the only balance or measure of association of interest,

$$\alpha(\mathbf{a}) = \frac{1}{2\sqrt{2}}\ln\frac{\tilde{a}_1\tilde{c}_1\tilde{a}_2\tilde{c}_2}{\tilde{b}_1\tilde{d}_1\tilde{b}_2\tilde{d}_2}\ ,$$

which satisfies

$$\alpha(\mathbf{a}) = \frac{1}{\sqrt{2}}\left(\alpha(\mathbf{a_1}) + \alpha(\mathbf{a_2})\right)\ ,$$

similarly to the previous case. This shows the possibilities of using balances to compare compositional characteristics in subpopulations to the corresponding overall population.

Although these relationships between balances guarantee that no paradoxical situations appear, we can go one step further and collapse the sample space in such a way that the new composition

represents directly the question to be answered, in an analogous manner to amalgamation. This consists in an orthogonal projection onto the appropriate subspace. According to Egozcue and Pawlowsky-Glahn (2005) this implies in the case corresponding to Table 3 the substitution of the initial 8-part composition by

$$\mathcal{C}\left[(\tilde{a}_1\tilde{a}_2)^{1/2},(\tilde{b}_1\tilde{b}_2)^{1/2},(\tilde{c}_1\tilde{c}_2)^{1/2},(\tilde{d}_1\tilde{d}_2)^{1/2},(\tilde{a}_1\tilde{a}_2)^{1/2},(\tilde{b}_1\tilde{b}_2)^{1/2},(\tilde{c}_1\tilde{c}_2)^{1/2},(\tilde{d}_1\tilde{d}_2)^{1/2}\right] ,$$

where $\mathcal{C}[.]$ stands for the closure operation (Aitchison, 1986). Particularly, this is obtained as a composition whose three first balances (order in Table 3) are maintained and the other ones are null. This representation can be simplified taking the appropriate subcomposition, obtained by elimination of the repeated parts,

$$[\mathbf{T_1},\mathbf{T_2}] = \mathcal{C}\left[(\tilde{a}_1\tilde{a}_2)^{1/2},(\tilde{b}_1\tilde{b}_2)^{1/2},(\tilde{c}_1\tilde{c}_2)^{1/2},(\tilde{d}_1\tilde{d}_2)^{1/2}\right] . \tag{3}$$

Now, the corresponding measures of association of interest are

$$\alpha(\mathbf{T_1}) = \frac{1}{\sqrt{2}}\ln\frac{(\tilde{a}_1\tilde{a}_2)^{1/2}}{(\tilde{b}_1\tilde{b}_2)^{1/2}} , \quad \alpha(\mathbf{T_2}) = \frac{1}{\sqrt{2}}\ln\frac{(\tilde{c}_1\tilde{c}_2)^{1/2}}{(\tilde{d}_1\tilde{d}_2)^{1/2}} ,$$

which satisfy

$$\begin{aligned}\alpha(\mathbf{T_1}) = \frac{1}{\sqrt{2}}\alpha(T_1) &= \frac{1}{2}\left(\alpha(\mathbf{a_1};T_1)+\alpha(\mathbf{a_2};T_1)\right) ,\\ \alpha(\mathbf{T_2}) = \frac{1}{\sqrt{2}}\alpha(T_2) &= \frac{1}{2}\left(\alpha(\mathbf{a_1};T_2)+\alpha(\mathbf{a_2};T_2)\right) ,\end{aligned}$$

and thus, being an arithmetic mean, the conditions required to avoid Simpson's paradox are satisfied by construction.

Analogously, for the case described in Table 4, we would obtain

$$[\mathbf{A_1},\mathbf{A_2}] = \mathcal{C}\left[(\tilde{a}_1\tilde{c}_1)^{1/2},(\tilde{b}_1\tilde{d}_1)^{1/2},(\tilde{a}_2\tilde{c}_2)^{1/2},(\tilde{b}_2\tilde{d}_2)^{1/2}\right] ,$$

and in the case corresponding to Table 5

$$\mathbf{A} = \mathcal{C}\left[(\tilde{a}_1\tilde{c}_1\tilde{a}_2\tilde{c}_2)^{1/4},(\tilde{b}_1\tilde{d}_1\tilde{b}_2\tilde{d}_2)^{1/4}\right] .$$

The corresponding measures of association of interest are, respectively,

$$\alpha(\mathbf{A_1}) = \frac{1}{\sqrt{2}}\ln\frac{(\tilde{a}_1\tilde{c}_1)^{1/2}}{(\tilde{b}_1\tilde{d}_1)^{1/2}} , \quad \alpha(\mathbf{A_2}) = \frac{1}{\sqrt{2}}\ln\frac{(\tilde{a}_2\tilde{c}_2)^{1/2}}{(\tilde{b}_2\tilde{d}_2)^{1/2}} ,$$

$$\alpha(\mathbf{A}) = \frac{1}{\sqrt{2}}\ln\frac{(\tilde{a}_1\tilde{c}_1\tilde{a}_2\tilde{c}_2)^{1/4}}{(\tilde{b}_1\tilde{d}_1\tilde{b}_2\tilde{d}_2)^{1/4}} ,$$

and again, we obtain that the global measures of association are the arithmetic mean of those in the subpopulations, i.e.

$$\begin{aligned}\alpha(\mathbf{A_1}) &= \frac{1}{\sqrt{2}}\alpha(\mathbf{a_1}) = \frac{1}{2}\left(\alpha(\mathbf{a_1};T_1)+\alpha(\mathbf{a_1};T_2)\right) ,\\ \alpha(\mathbf{A_2}) &= \frac{1}{\sqrt{2}}\alpha(\mathbf{a_2}) = \frac{1}{2}\left(\alpha(\mathbf{a_2};T_1)+\alpha(\mathbf{a_2};T_2)\right) ,\\ \alpha(\mathbf{A}) &= \frac{1}{2}\left(\alpha(\mathbf{A_1})+\alpha(\mathbf{A_2})\right) = \frac{1}{4}\left(\alpha(\mathbf{a_1};T_1)+\alpha(\mathbf{a_1};T_2)+\alpha(\mathbf{a_2};T_1)+\alpha(\mathbf{a_2};T_2)\right) .\end{aligned}$$

From a probabilistic point of view the composition (3) can be obtained as a mean value called centre in the compositional framework (Aitchison, 1997). Think of the set of subpopulations as a probability space, from which a subpopulation is randomly sampled. Consider the random composition $\mathbf{x}$ which assigns to each subpopulation $\mathbf{a_i}$ the proportions $\mathcal{C}[\tilde{a}_i, \tilde{b}_i, \tilde{c}_i, \tilde{d}_i]$. The question of how to get a central or representative subpopulation suggest to use the centre of the random composition

$$\text{cen}[\mathbf{x}] = \mathcal{C}\left[\left(\prod_{i=1}^{n}\tilde{a}_i\right)^{1/n}, \left(\prod_{i=1}^{n}\tilde{b}_i\right)^{1/n}, \left(\prod_{i=1}^{n}\tilde{c}_i\right)^{1/n}, \left(\prod_{i=1}^{n}\tilde{d}_i\right)^{1/n}\right],$$

that, in the case of $n = 2$, coincides with (3). It is remarkable that in this expression there is no reference to the number of individuals sampled from each subpopulation. In fact, the question addressed refers to *proportions of individuals* in the subpopulations irrespective of the number of sampled individuals from each subpopulation. Note that *individuals* for the random composition $\mathbf{x}$ are just subpopulations. Here we arrive to the starting point: which is the question to be answered? Essentially, this compositional approach answer the question: *which is a central representative of the subpopulations such that proportions on it (e.g. successes, failures) can be compared with proportions in each subpopulation?*

4 Case study: Kidney stone treatment

To illustrate the approach, consider a real-life example from a medical study (Julious and Mullee, 1994), adapted here from the summary cited in Wikipedia (2006).

The study compared the success rates of two treatments for kidney stones (Table 6). Success rates (successes/total) for treatment T_1 were 0.778, and for treatment T_2 0.824, showing that treatment T_2 is more effective. Including data about kidney stone size (Table 7), however, the

Table 6: Kidney stone treatment; overall population (number of cases and estimated row proportions).

	success	failure
treatment T_1	273	77
row proportion	0.778	0.222
treatment T_2	289	61
row proportion	0.824	0.176

Table 7: Kidney stone treatment; population classified according to stone size (number of cases and estimated row proportions).

small stones	success	failure
treatment T_1	81	6
row proportion	0.926	0.074
treatment T_2	234	36
row proportion	0.865	0.135

large stones	success	failure
treatment T_1	192	71
row proportion	0.729	0.271
treatment T_2	55	25
row proportion	0.685	0.315

same set of treatments revealed a different answer, as for small stone sizes Treatment T_1 overrated Treatment T_2, with 0.926 vs. 0.865 success rates, and the same happened for large stone sizes with, respectively, 0.729 and 0.685 success rates. Now treatment T_1 was seen to be more effective in both

Table 8: Kidney stone treatment (estimated proportions).

$\mathbf{a_1}$ = small stones				$\mathbf{a_2}$ = large stones			
T_1		T_2		T_1		T_2	
S_{11}	F_{11}	S_{12}	F_{12}	S_{21}	F_{21}	S_{22}	F_{22}
0.232	0.018	0.545	0.203	0.666	0.104	0.158	0.072

cases, leading to a typical case of Simpson's paradox. Standard interpretation for this to happen is that sizes of the groups which are combined when the lurking variable (stone size) is ignored are very different. Doctors tend to give the severe cases (large stones) the better treatment (T_1), and the milder cases (small stones) the inferior treatment (T_2). Therefore, the totals are dominated by these two groups. The lurking variable has a large effect on the ratios, i.e. the success rate is more strongly influenced by the severity of the case than by the choice of treatment. In other terms: even if apparently the size of the stone has not been a criterium for selecting the patients to be included in the study, the total number of cases in each group dominates the result.

Following the approach described in Section 3, consider now the data in Table 7 as an 8-part composition. The result in estimated proportions is represented in Table 8. If the effectiveness of the treatments has to be compared both considering the size of the stones and independently of it, the approach described above leads to project the composition onto the simplex defined in Eq. (3), resulting in

$$[\mathbf{T_1}, \mathbf{T_2}] = [0.853, 0.147, 0.789, 0.211] \ ,$$

and the measures of association of interest satisfy

$$\begin{aligned} \alpha(\text{small stones}, T_1) = 1.788 &> \alpha(\text{small stones}, T_2) = 1.315 \ ; \\ \alpha(\text{large stones}, T_1) = 0.700 &> \alpha(\text{large stones}, T_2) = 0.550 \ ; \\ \alpha(\mathbf{T_1}) = 1.244 &> \alpha(\mathbf{T_2}) = 0.933 \ . \end{aligned}$$

Thus, we can see that the three required inequalities hold, and that the measure of association of each treatment is inside the range of the corresponding measures within each subpopulation.

5 Conclusions

The compositional approach to analyse proportions in which Simpson's paradox might appear shows that using balances is a natural way to analyse these data which leads to reasonable results. The consequence is that the centre, or closed geometric mean, of the tables to be analysed is a sensible alternative to amalgamation, which is not a linear operation in the Aitchison geometry of the simplex.

Acknowledgements

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

References

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. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology*, Volume I, II and addendum, pp. 3–35. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p.

Aitkin, M. (1998). Simpson's paradox and the bayes factor. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 60*(1), 269–270.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 — The sixth annual conference of the International Association for Mathematical Geology*, pp. 20 p. CD-ROM.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*(456), 1205–1214.

DiSerio, C., Y. Rinott, and M. Scarsini (2007). Simpson's paradox for the cox model. Unpublished paper, 19 p.

Egozcue, J. and V. Pawlowsky-Glahn. Simplicial geometry for compositional data.

Egozcue, J. J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology 37*(7), 795–828.

Good, I. J. and Y. Mittal (1987). The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics 15*(2), 694–711.

Haunsperger, D. B. (2003). Aggregated statistical rankings are arbitrary. *Social Choice and Welfare 20*(2), 261–272.

Julious, S. A. and M. A. Mullee (1994). Confounding and Simpson's paradox. *BMJ 309*(6967), 1480–1481.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA) 15*(5), 384–398.

Pearson, K. (1899). Theory of genetic (reproductive) selection. *Philosophical Transactions of the Royal Society of London Series A 192*, 260–278. (especially pages 277-278, On the spurious correlation produced by forming a mixture of heterogeneous but uncorrelated materials).

Wikipedia (2006). Simpson's paradox — wikipedia, the free encyclopedia. [Online; accessed 19-October-2006].

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika 2*, 121–134.