1

Introduction

Compositional data describe parts of some whole. They are commonly presented as vectors of proportions, percentages, concentrations, or frequencies. As proportions are expressed as real numbers, one is tempted to interpret, or even analyze, them as real multivariate data. This practice can lead to paradoxes and/or misinterpretations, some of them well known even a century ago, but mostly forgotten and neglected over the years. Some simple examples illustrate the anomalous behavior of proportions when analyzed without taking into account the special characteristics of compositional data.

Example 1.1 (Intervals covering negative proportions).
Daily measurements of an air pollutant are reported as $3 \pm 5\ \mu g/m^3$. The given interval of concentration covers a nonsensical range of concentrations that includes negative values. It is probably generated by an average of concentrations which contain some values much higher than $3\ \mu g/m^3$. For instance, the following is a set of rounded random percentages: $1, 1, 2, 3, 4, 4, 7, 13, 29, 37$. Their mean is 10.1%, while their standard deviation is 12.7%. Thus a typical $2s$-interval for the mean value would be an interval covering negative proportions, namely, $(-15.3\%; 35.5\%)$. A frequent procedure is to cut this interval at zero, but then the question arises on what happens to the probability assigned to the eliminated part of the interval, $(-15.3\%; 0\%)$, and to the probability assigned to the retained part, $(0\%; 35.5\%)$. ◇

Modeling and Analysis of Compositional Data, First Edition.
Vera Pawlowsky-Glahn, Juan José Egozcue and Raimon Tolosana-Delgado.

Example 1.2 (Small proportions: Are they important?).
Frequently, when some components or parts of a composition are very small, they are eliminated, with the argument that they are negligible. In such a case, it is important to think about *the salt in a soup*. Consider a soup that is perfectly seasoned to your taste, and imagine somebody adds to the soup the same amount of salt you used, thinking that it was not yet seasoned. Probably, doubling the amount of salt will spoil it completely. To our understanding, this is a perfect example on how important a small proportion can be and why a relative scale gives you better information in this case than an absolute one. Sometimes, small proportions are added to other parts, for example, salt and other spices, but that leads to a loss of information, making the recipe insufficiently specified. ◇

Example 1.3 (Reporting changes in proportions).
In the 1998 election to the German Bundestag, the German Liberal Party (FDP) obtained 6.2% of the votes. Eleven years later, in the 2009 elections, they obtained a share of 14.6%. This could be reported as an increment of 8.4 percentage points. We are more used to reading that FDP increased its proportion of votes a 135% ($6.2 + 6.2 \times 135/100 = 14.6$). In the following election, just 4 years later, the party decreased its votes by a significant 67%, but still half of the increment that occurred between 1998 and 2009. Nevertheless, that meant that the FDP was not anymore represented in the Bundestag, because its share ($14.6 - 14.6 \times 67/100 = 4.8$) dropped below the threshold of 5% required by the German electoral law. How can it be that increasing 135% and decreasing 67% gives a negative balance? Perhaps this is a bad way of reporting changes in proportions (data extracted from Wikipedia (2014)).

Reporting increments of shares in differences of percentage points have also disappointing properties, as the relative scale of proportions is ignored. In fact, an increment of 8.4 percentage points represents a very important change from the 1998 result of FDP (6.2%). It would be not so important if the previous 1998 result were, for instance, 30%. ◇

Example 1.4 (The scale of proportions).
In a given year, the annual proportion of rainy days in a desert region is 0.1%, and near a mountain range it is 20.0%. Some years later, these proportions have changed to 0.2% and 20.1%, respectively. To summarize the situation, one can assert that the rainy days in both regions have increased by 0.1%. Such a statement suggests the idea of a homogeneous change in the two different regions, ignoring that the rainy days in the desert have been doubled, while in the mountain range the proportion is almost the same. Using the increment of ratios typical of election results or economic reports, the rainy days would have increased a critical 100% in the desert, and a slightly relevant 5% in the mountains.

Furthermore, if some analysis of the evolution of the rainy days is made in both regions, it should be guaranteed that equivalent results are obtained if the nonrainy days are analyzed. In the desert region, the annual proportion of nonrainy days has changed from 99.9% to 99.8% and near the mountain range from 80.0% to 79.9%. That represents that nonrainy days have decreased, respectively, 0.001% and 0.00125%, which suggests almost no difference between the mountain and the desert. How can it then be that rainy days change so dramatically in the desert and nonrainy days do not change at all? A proper analysis should assure that no paradoxical results are obtained when analyzing one type of days and its complementary. ◇

Example 1.5 (The Simpson's paradox).
The lectures on statistics started very early this morning. Students (men and women) are divided into two classrooms. Some of them arrived on time and some of them were late. Academia was interested in knowing about punctuality according to the gender of the students. Therefore, data were collected this morning during the statistics lectures. The data set is reported in Table 1.1. The paradoxical result is that, for both classrooms, the proportion of women arriving on time is greater than that of men. On the contrary, if the individuals of both classrooms are joined in a single population, the proportion of punctual men is larger than that of the women. This kind of paradoxical results are known as Simpson's paradox (Simpson, 1951; Julious and Mullee, 1994; Zee Ma, 2009). The paradox can be viewed from different points of view. The simplest one, the arithmetic perspective, is to look at the way in which proportions are aggregated: to find the proportion of on-time women in the joint population, the per classroom proportions a_1/W_1, a_2/W_2 are *averaged* as $(a_1 + a_2)/(W_1 + W_2)$, where a_i is the number of on-time women in the classroom i and W_i is the corresponding

Table 1.1 Number of students of two classrooms, arriving on time and being late, classified by gender. Proportions are reported under the number of students. The largest proportion of arriving on-time men and women are in boldface for easy comparison.

	Classroom 1		Classroom 2		Total	
	On time	Late	On time	Late	On time	Late
Men	53	9	12	6	65	15
	0.855	0.145	0.667	0.333	**0.813**	0.188
Women	20	2	50	18	70	20
	0.909	0.091	**0.735**	0.265	0.778	0.222

total of women. This kind of average is ill-behaved for proportions as shown by Simpson's paradox.

A second point of view is to look at the total proportion of on-time women as a mean value of this proportion in the two classrooms. Each classroom is treated as a sample individual and $(a_1 + a_2)/(W_1 + W_2)$ is taken as the sample mean of the proportions. The paradoxical result suggests that mean values of proportions should be redefined carefully to get consistent results. ◇

Example 1.6 (Spurious correlation).
The Spanish Government publishes the number of affiliations to the Social Security on a monthly basis, which is classified into the following categories depending on the type of company: agricultural, industrial, construction, and service. The 144 data, corresponding to a monthly series going from 1997 to 2008, were downloaded from the corresponding web site (Gobierno de España, 2014). A version, prepared for processing, is available in (www.wiley.com/go/glahn /practical). First, to obtain proportions between the different types of company, the data were normalized to add to 1 in the full composition comprising the four categories. Then, the correlation matrix was computed (see Table 1.2). Next, to analyze the behavior of the companies excluding *construction*, a subcomposition of three categories was obtained, suppressing the category *construction* and converting the three-part vector to proportions, so that the three components add up to 1. Again, the correlation matrix was computed (see Table 1.3). When analyzing correlations in the full composition with four parts and the subcomposition with three parts, the correlation between the proportion of agricultural and industrial companies only changed slightly, actually from −0.9808 to −0.9887, whereas the correlation between the service companies and either agricultural or industrial companies changed dramatically, from 0.1699 to 0.9863 in the first case and from −0.0723 to −0.9999 in the second. This is a typical effect when analyzing a set of parts adding up to a constant, or a subset of the same parts, closed to any constant.

Table 1.2 Correlation of proportion of affiliations to social security in Spain according to the type of company (four-part composition: agricultural, industrial, construction, and service).

	Agricultural	Industrial	Construction	Service
Agricultural	1.0000	−0.9808	0.9201	0.1699
Industrial	−0.9808	1.0000	−0.9663	−0.0723
Construction	0.9201	−0.9663	1.0000	−0.1867
Service	0.1699	−0.0723	−0.1867	1.0000

Table 1.3 Correlation of proportion of affiliations to social security in Spain according to the type of company (three-part subcomposition: agricultural, industrial, and service).

	Agricultural	Industrial	Service
Agricultural	1.0000	−0.9887	0.9863
Industrial	−0.9887	1.0000	−0.9999
Service	0.9863	−0.9999	1.0000

The problem of spurious correlation is sometimes circumvented by avoiding the closure when considering a subcomposition. This is equivalent to say: the percentages of agricultural, industrial, construction, and service affiliates constitute a composition as the percentages add to 100%; to overcome the compositional intricacies, we can remove one component, for example, service, so that the remaining percentages do not add to 100%. This way, the correlation matrix between the percentages of agricultural, industrial, and construction affiliates are exactly those reported in Table 1.2 in the first three columns and rows. However, a new question arises: what would happen if we start with two additional categories of affiliation closed to 100%? ◇

The awareness of problems related to the statistical analysis of compositional data dates back to a paper by Karl Pearson (1897) the title of which began significantly with the words *"On a form of spurious correlation ... "*. Since then, as stated in Aitchison and Egozcue (2005), the way to deal with this type of data has gone through roughly four phases, which can be summarized as follows:

Phase I: 1897–1960

Karl Pearson, in his paper on spurious correlations, pointed out the problems arising from the use of standard statistical methods with proportions. But his warnings were ignored until around 1960, despite the fact that a compositional vector – with components the parts of some whole – is usually subject to a constant-sum constraint.

Phase II: 1960–1980

Around 1960, the geologist Felix Chayes (1960) took up the problem and warned against the application of standard multivariate analysis to compositional data. He tried to separate what he called the *real* from the *spurious* correlation, in an attempt to avoid the *closure problem*, expressed mainly as a negative bias induced by the constant-sum constraint. Important contributions in geological

applications were made, among others, by Sarmanov and Vistelius (1959), and Mosimann (1962) which drew the attention of biologists. However, as pointed out by Aitchison and Egozcue (2005), *distortion of standard multivariate techniques when applied to compositional data was the main goal of study*.

Phase III: 1980–2000

Aitchison, in the 1980s, realized that compositions provide information about relative, not absolute, values of parts or components. Consequently, every statement about a composition can be stated in terms of ratios of components (Aitchison, 1981, 1982, 1983, 1984). The facts that logratios are easier to handle mathematically than ratios, and that a logratio transformation provides a one-to-one mapping onto a real space, led to the advocacy of a methodology based on a variety of logratio transformations. These transformations allowed the use of standard unconstrained multivariate statistics applied to transformed data, with inferences translatable back into compositional statements. But they were, not without difficulties, derived from the fact that the usual Euclidean geometry and measure were implicitly assumed for the sample space of compositional data.

This phase deserves special attention because transform techniques have been very popular and successful over more than a century; from the Galton-McAlister introduction of the logarithmic transformation for positive data, through variance-stabilizing transformations for sound analysis of variance, to the general Box–Cox transformation (Box and Cox, 1964) and the implied transformations in generalized linear modeling. The logratio transformation principle is based on the fact that there is a one-to-one correspondence between compositional vectors and associated logratio vectors, so that any statement about compositions can be reformulated in terms of logratios, and vice versa. The advantage is that the problem of a constrained sample space, the simplex, is removed. Data are projected into multivariate real space, opening up all available standard multivariate techniques. The original transformations were principally the additive logratio transformation (Aitchison, 1986, p. 113) and the centered logratio transformation (Aitchison, 1986, p. 79). The logratio transformation methodology seemed to be accepted by the statistical community; see, for example, the discussion of Aitchison (1982).

Phase IV: 2000–present

Around 2000, several scientists realized independently that the internal simplicial operation of perturbation, the external operation of powering, and the simplicial metric define a metric vector space (indeed a Hilbert space) (Billheimer et al., 1997, 2001, Pawlowsky-Glahn and Egozcue, 2001). The recognition of

the algebraic-geometric structure of the sample space of compositions led to the staying-in-the-simplex approach for the analysis of compositional problems (Pawlowsky-Glahn, 2003). This approach is essentially based on the *principle of working in coordinates* (Mateu-Figueras et al., 2011). Compositions are represented by orthonormal coordinates, which live in a real Euclidean space. They can be interpreted in themselves or from their representation in the simplex. The sample space of random compositions is represented by the simplex with a simplicial metric and measure, different from the usual Euclidean metric and Lebesgue measure in real space.

The book presented here corresponds to the fourth phase. It summarizes the state of the art in the stay-in-the-simplex approach. Therefore, the first part will be devoted to the algebraic-geometric structure of the simplex, which we call *Aitchison geometry* (Pawlowsky-Glahn and Egozcue, 2001). Although it is a Euclidean geometry, we felt that it is important to distinguish it from the usual geometry in real space.