1. Introduction to compositional data

Compositional data are data where the elements are parts of a whole. Whenever a fraction is measured, whether it is as a percentage, a concentration, ppm, or a frequency, it is a part of a composition. Probabilities are also compositional, making compositional data analysis particularly useful in the field of statistics. Compositional data are always positive, and the sum of a composition is a constant which is usually, but not necessarily, 1. The sum is often irrelevant and only the *relative* magnitude of the parts has meaning.

Compositional data are encountered in many aspects of daily life and is dealt with implicitly. Political surveys, election polls, nutrition declarations on packaged food, and many types of probabilistic games are all examples of everyday compositional data. They are also commonly encountered in most sciences, in particular chemistry, geology, and life science.

In this chapter, we will learn how to identify compositional data and how to distinguish between compositional data and ordinary real multivariate data. Compositional data need to be analyzed using compositional methods, and we will show how failure to do so can lead to meaningless or even paradoxical results.

1.1 Identifying compositional data

A geologist studies meteorites and she wants to compare two meteorites which she has found. The smaller of the two contains 100 g of FeNi alloys which amounts to 95% of the total weight of the meteorite. The remaining 5% are silicates. The larger meteorite contains 400 g of iron which is 23% of the total weight with the remaining 77% consisting of silicates. Despite the fact that the second meteorite contains more iron (by weight), she correctly identifies it as a chondrite and the smaller one as an iron meteorite. The larger meteorite contains more iron because it is bigger and therefore, the absolute amount of iron has no significance. The only thing that matters is the amount of iron relative to the amount of silicates. She reports the content as percentages because the weight depends on

the size of the meteorites and is therefore irrelevant.

In another example, a meteorologist (who does not study meteoritics) is out to compare the weather at two different locations. He measures the temperature, pressure, and wind speeds. If he measures a higher temperature at one location he will correctly conclude that it is warmer there than at the other location, irregardless of the measured pressure and wind speeds. These quantities can in principle take any positive value and they are not constrained by one another and therefore the sum of these numbers is meaningless. Meteorological data are not compositional, but rather what is called real multivariate data.

1.2 Relevance for metagenomics

When studying genetic samples containing DNA pooled from a (large) number of organisms it is called metagenomics as opposed to ordinary WGS genomics which done using cultivated clonal isolates from a single organism. The practical process of extracting, preparing and sequencing the DNA is beyond the scope of this course, but in the end, ideally, metagenomic high throughput sequencing produces compositional data, by identifying which organism all the DNA fragments belong to and counting the number of fragments associated with a certain organism. Effectively, this is not always the case and the outputted data is pseudo-compositional for reasons discussed later. In this course we will discuss more aspects of compositional data analysis than what is relevant for metagenomics, but emphasis will be placed on the techniques that apply to genomic data.

1.2.1 Genomic data

Genomic data are generally a set of short nucleotide sequences (reads) which are packed into a so called *FASTQ*-file. At a first glance, it can be difficult to see how such data are compositional. It is crucial to understand what the reads represent in order to accept their compositional nature. A bottle of sea water contains cells from hundreds or maybe thousands of different organisms: a lot bacteria, possibly parasites and viruses, plant matter and algae, fungi, archaea, and maybe remnants of fish or even mammals. There are billions of cells in the sample, each containing a strand of DNA, but only a random fraction of these cells will be opened during the process of DNA extraction. The process may be biased towards bacterial cells or eukaryotic cells, depending on the kit used, but in the end, it is a random process. The DNA gets fragmented and loaded onto the flow cell in the sequencing machine, where it is read and decoded. Again, only a random subset of all the fragments gets sequenced. In the end, we may end up with 127 reads belonging to a certain species. Does this number mean anything? Absolutely not. It is the same as walking to a highway bridge and for five minutes, count all the white cars driving underneath. Maybe it is 85. Is that a lot? That depends on how many non-white cars passed during the same five minutes. Only by counting some other part, blue cars, all cars, or motor bikes, can we conclude anything about the observation of 85 white cars. It is the proportion that matters, not the absolute number. Therefore, the reads in the *FASTQ*-files are parts of a composition and they can be grouped (amalgamated) in various ways, using bioinformatic techniques, such as mapping, assembly, and binning. But it is important to remember that the only meaningful quantity is the ratio between parts and not the absolute magnitude of each part, which is based on arbitrary properties, such as sample size, sample storage conditions, DNA extraction efficacy, sequencing depth, etc.

1.3 Proportions

We are taught to work with proportions already in primary school, and to most people, dealing with (simple) compositional data is intuitive. However, as the following quotation shows, not everybody has an intuitive comprehension of proportional data and the mathematical rules that apply to them. On November 16, 1999, member of the Danish parliament, Folketinget, Aase D. Madsen spoke the following words during a debate about usage of public libraries:

> "[...] Og med hensyn til, hvem der kommer på bibliotekerne, og hvem der ikke kommer, er der en tabel 18 med en gruppe delt ind efter alder, og dér står, at 39 pct. af den mandlige del af befolkningen aldrig kommer på bibliotekerne, og at 30 pct. af den kvindelige del af befolkningen, altså fordelt gennemsnitligt over alder, aldrig kommer der. Og når jeg lægger mænd og kvinder sammen - det skal man være lidt forsigtig med, men på det her område tør jeg godt - så giver 39 pct. af mændene og 30 pct. af kvinderne i befolkningen tilsammen, og det må være 69 pct. Tager jeg fejl?"

It is easy to laugh about this, but once the compositions become bigger and more complex, mathematical errors are easily made, even by people who find the quotation above amusing.

Proportional data behave differently from real data, which is illustrated in the following three examples.

1.3.1 Negative proportions

A proportion is the ratio between two real and positive numbers, while a fraction is the ratio between two real, but not necessarily positive, numbers. It is important to understand the difference. Fractions can be negative while proportions are strictly positive. You cannot ask someone to cut minus one quarter of a pizza. The following example shows what can happen when applying normal statistics to proportional data.

Before the elections, 8 opinion polls are conducted in order to predict the success of a certain political party. The resulting 8 percentages of people who claim they will vote for the party are {2%,2%,2%,3%,3%,4%,11%,21%}. Calculating the mean and the standard deviation, assuming a normal distribution, gives $6\% \pm 6.3$, suggesting that there is a 17% probability that the party will receive a negative proportion of the votes, which is clearly nonsense.

A similar problem arises when food producers want to test if their product is free of a certain ingredient. This could, for instance be a brewery making alcohol-free beer or a dairy plant making lactose-free milk. These products are rarely 100% free of the ingredient they claim to be free of, but contains trace amounts, which is allowed as long as no single item contains more than a fixed (and small) proportion. For beer, this is typically around 0.03% alcohol. In order to comply with these rules, companies sample their product and measure how much alcohol or lactose it contains. However, because the proportions are so small, the ordinary standard deviation will typically extend into the range of negative numbers, making their statistics invalid if they do not take the compositionality of the data into account.

1.3.2 Small proportions

Oftentimes small proportions are important but hard to quantify accurately in absolute terms. Cooking recipes are usually not given as proportions, but rather as absolute amounts (weight or volume). However, in many cases, the exact amount of salt is not given explicitly, because the ratio of salt to water in, say, a soup, is very small, and even a small measuring error in the amount of salt could render the soup inedible. If a dish requires 1 g of salt, and by accident you add 2 g, then you have doubled the amount of salt and spoiled the dish even if you only added a single gram too much. A small absolute change to a small proportion can can lead to a large proportional change.

1.3.3 Proportional changes

Reporting changes as a proportion can lead to false impressions of the significance of the change. If on a give day, a stock on the market has the value of \$100 per stock, and on the following day its value has increased to \$220, we could report the change as a 120% increase. On the third day the value decreases by 70%, which at a glance appears to be a smaller decrease than the increase the day before. However, the value of the stock on day 3 will be \$66 which is much less than the \$100 that it started out with.

1.4 Simpson's paradox

A pitfall often encountered when working with compositional data is the amalgamation paradox, or Simpson's paradox, named after mathematician Edward Simpson who first described the effect in 1951. The effect is illustrated in the table 1.1, where the success rates of two treatments for kidney stone, A and B, are compared. The data shows that the success rate of treatment B is higher when all cases are considered, but when the cases are split into two parts, small and large kidney stones, treatment A works better in both situation.

	Treatment A	**Treament B**
Small kidney stone	(81/87) 93%	(234/270) 87%
Large kidney stone	(192/263) 73%	(55/80) 69%
Total	(273/350) 78%	(289/350) 83%

Table 1.1: The success rates of two different treatments for kidney stone, when calculated from the total number of cases and when the cases are split into parts.

The problem lies in the fact that there are many more cases of large kidney stones being treated with method A and small kidney stones being treated with B than vice versa, and the total represents a weighted arithmetic average of the cases, which is a property that is ill-behaved when applied to proportions.

1.5 Correlations

A question that often arises in metagenomics is if certain organisms within a number of samples are correlated in some manner. For instance, in a study of gut microbiomes, it

could be relevant to look for bacterial species that are in high abundance whenever another (set of) species is abundant. Correlations in compositional data are, however, not as easily interpretable as they are for real data, and very often, if correlations show up, they are artifacts and not real.

1.5.1 Spurious correlations

Consider three randomly distributed variables, x, y, and z. By construction, these are all uncorrelated, but it turns out that the ratios x/z and y/z can be highly correlated, giving the false impression that there is a relationship between x and y. This situation is commonly encountered in metagenomics, where gene counts are often "normalized" by organism counts. If, for example, x and y represents the counts of two different bacterial genes across a number of samples and z represents the total number of bacteria, then it is common practice to express the gene abundance as x/z and y/z, a ratio sometimes referred to as *FPKM*, in which case it may seem like the occurrence of gene 1 and gene 2 is highly correlated even if that is not the case at all. An example is shown in Fig. 1.1.

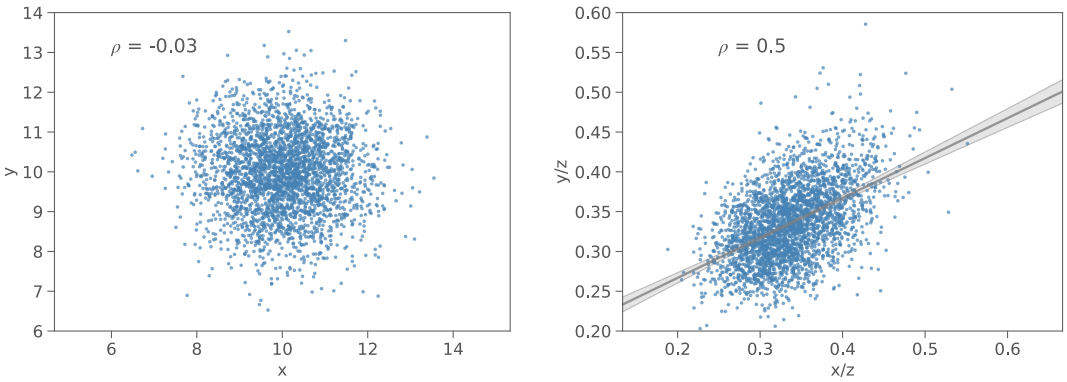

Figure 1.1: Two random and normal distributed variables x and y are shown in the left panel to be uncorrelated. By dividing both with a third random variable z, the ratios are seen to be positively correlated in the right panel.

The spurious correlation occurs because the two ratios share a common denominator and since compositional data are, by construction, made up by proportions, spurious correlations are widespread in genomics.

1.5.2 The negative correlation bias

Apart from spurious correlations, compositional data suffer from a so-called negative correlation bias. Consider the composition of results from N coin tosses. If the number of heads is n, the the number of tails must be $N - n$. The more heads you get, the fewer tails you will get. Heads and tails will always be perfectly negatively correlated. This effect extends to larger compositions: if a set of parts goes up, another set of parts must go down, and this negative correlation is fundamentally indistinguishable from true negative correlations.

1.5.3 Compositional correlations

If two variables, x and y, are correlated, we expect that a linear relationship exists between them, such that $y = \alpha x + \beta$. For real data, α and β are irrelevant. The strength of the correlation is determined by how close the data points follow the relation. This is seen in the left panel of Fig. 1.2, where three sets of variables show equally good correlation.

For compositional data, in order for parts to be correlated, the ratio between the parts must be constant. For the red and the black data sets in Fig. 1.2, the ratios between the variables are constant and equal to 1 and 5 respectively. In the blue data set, however, the ratio is not constant, but equal to $y/x = 1 + 20/x$. If we plot the three data sets in log-log space, as is seen in Fig. 1.2 panel b, it is clear that there is a linear relationship within the red and the black data set, but not within the blue. Thus for parts in compositional data to be correlated, they must follow a linear relationship in log-space of the form $y = x + \beta$, where the intercept β in log-space equals the slope α in Euclidean space (and is irrelevant for the strength of the correlation).

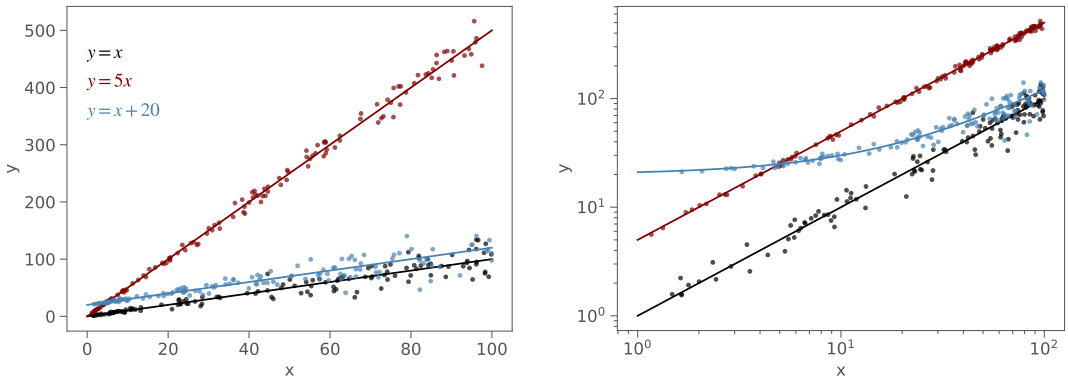

Figure 1.2: Parts in a set of composition correlates if they have a constant ratio, in which case they will have a linear relationship in log-space.

1.6 A brief history of compositional data analysis

The phenomenon of spurious correlations was the main motivation for the development of compositional analysis. Karl Pearson was the first to point out the problems of spurious correlations when applying standard statistical methods to proportions back in 1897. Pearson, and other prominent statisticians of the that time, was worried that conclusions would be drawn from correlations that are artifacts of the analysis method, rather than actual relationships between variables. His warning, however, was largely ignored until 1960. Around this time, Felix Chayes, who was a geologist, noted that standard multivariate analysis should not be applied to compositional data and scientists began to move away from multivariate correlations within the field of geology. Only in the 1980'ies John Aitchison, a Scottish statistician, established the modern methods of compositional data analysis and noted

> It seems surprising that the warnings of three such eminent statistician-scientists as Pearson, Galton and Weldon should have largely gone unheeded

> for so long: even today uncritical applications of inappropriate statistical methods to compositional data with consequent dubious inferences are regularly reported.

Although this quote is almost 40 years old, "normalized" count data and correlations between them are still widely reported in scientific publications today, particularly in the fields of bio- and life-science.