

Analysing Activities in a Classroom - Remembrances of Professor John Aitchison in Hong Kong with Applications to a Service Provider

Shirley Coleman
Newcastle University

Abstract

Compositional data analysis formed a main focus of statistical activities at Hong Kong University when Professor John Aitchison was head of the Statistics department. It was part of a new Master's degree in Statistics that he set up, and as this was the first such post graduate degree to be offered in Hong Kong, it attracted many gifted statisticians from the Government Statistical Service and other employments making it a very lively program.

Acknowledging the constrained nature of many types of data led to a new way of looking at proportions and percentages of components making up data items. Professor Aitchison's seminal book *The statistical analysis of compositional data* contains background theory and many examples of data arising from a wide variety of applications such as geology, economics and human behaviour. One example was an analysis of the daily activities of a statistician. This prompted an analysis of classroom activities in a range of classes and schools encountered during teacher training within the Professional Educational Studies department of Hong Kong University. It was found that the nature of the target class and the school level affected the pattern of lesson activities with more listening carried out in the higher target classes and higher level schools. More time was spent dealing with educational equipment in lower level schools.

Data analytics is increasingly popular in all walks of life and many small and medium enterprises are realising the benefits. Compositional data forms a large part of internal company operational data and its analysis can provide useful insights. For example, the change in proportions of different activities undertaken over time is important information for a service provider. In addition to biplots and coordinate representations of isometric log ratios, using ternary diagrams to illustrate proportions is an informative way to share findings with company staff.

Keywords: business analytics, inspirational leader, statistical consultancy.

1. Introduction

Staff and students alike were delighted to embrace Professor Aitchison's erudite and stately leadership of the Statistics department at Hong Kong University from the time he started work there in 1976 to his retirement in 1989. His seminal book *The statistical analysis of*

compositional data contained background theory and many examples of data arising from applications including geology, economics and human behaviour encountered during his consulting projects with a wide range of colleagues (Aitchison 1986).

Compositional data analysis formed a main focus of statistical activities at the University. It was part of a new Master's degree in Statistics that Professor Aitchison set up, and as this was the first such post graduate degree to be offered in Hong Kong, it attracted many gifted statisticians from the Government Statistical Service and other employments making it a very lively program.

The idea of scale invariance was very important to Professor Aitchison, and thus the associated idea of compositional data as equivalence classes. Awareness of the constrained nature of many types of commonly encountered data led his colleagues to look at proportions and percentages in a new way. The example of an analysis of daily activities of statisticians in Professor Aitchison's book prompted consideration of possible applications in many other fields.

This paper applauds Professor Aitchison's work and in the next section gives some examples of compositional data analysis and describes an analysis of classroom activities carried out in Hong Kong. Data analytics is increasingly popular in all walks of life and many small and medium enterprises are realising the benefits (Coleman 2016). Compositional data forms a large part of internal company operational data and its analysis can provide useful insights (Ahlemeyer-Stubb and Coleman 2018). The third section gives an application in a healthcare setting. The final section is a conclusion and meditation on the benefits of working with charismatic people.

2. Compositional data analysis examples

Professor Aitchison wrote a suite of computer programs to accompany his book. The CODA programs and data used in the book enabled readers and students to replicate the analyses in the book and apply the techniques to any data that they might encounter from research and consulting projects. The analysis of the activity patterns of a statistician over 20 days was one such dataset. Figure 1 shows a hand drawing of the ternary plot of the activities summarised into work, sleep and other.

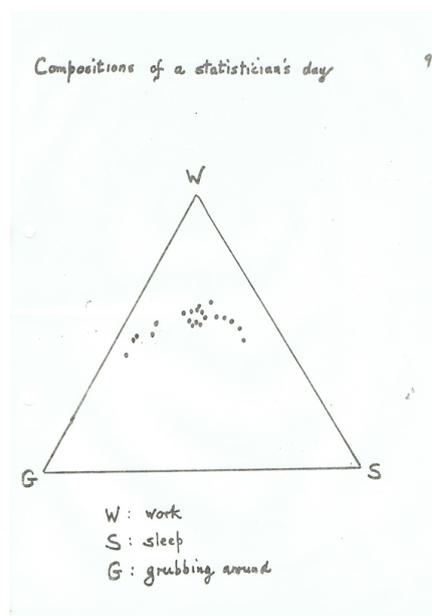


Figure 1: Ternary plot of activities of a statistician.

The ternary plot shows that a disagreeably large proportion of each day was spent working. This application prompted an analysis by the author of the activities taking place in a classroom during a particular type of lesson.

Listening is an important part of teaching English language and was the focus of an analysis of classroom activities in a range of classes and schools encountered during teacher training within the Professional Educational Studies department of Hong Kong University (Coleman and Lee 1988), a survey was carried out amongst student teachers of English in the Department of Professional Educational Studies and the Institute of Language in Education. The survey questionnaire contained general questions about the school, the listening equipment and the teachers' opinions on teaching listening. One item asked for the time spent on different activities and is thus a compositional data example. Responses were obtained from 99 student teachers.

The activities were summarised into 4 categories, being time spent:

- setting up and distributing, collecting and storing listening equipment
- preparing students for the listening activity, establishing and maintaining "class order"
- by students actually doing the listening activity
- checking student performance.

It was found that the nature of the target class and the school level affected the pattern of lesson activities with more listening carried out in the higher target classes and higher level schools. More time was spent dealing with educational equipment in lower level schools.

An interesting set of compositional data that Professor Aitchison used in his teaching was the proportions of different blood genotypes (MM, MN and NN) in 26 samples of people from different ethnic groups. The proportions are analysed by first calculating their geometric mean, g , then taking logs of the ratios of proportions to geometric mean giving a centred log ratio covariance matrix, then finding principal components (PCs). Dividing by the geometric mean introduces a sum to zero constraint on the log ratios but is different in nature to the original sum constraint on the proportions as it is the result of a construction that would induce dependence even if the original set of data points were not part of a zero sum constraint. The new constraint is reflected in the fact that the third PC has zero eigenvalue. Table 1 shows the analytical output.

Table 1: Principal components analysis of genotype proportions.

Eigenvalue	3.186	0.045	0	Variable	PC1	PC2
Proportion	0.986	0.014	0	$\log(MN/g)$	0.012	0.816
Cumulative	0.986	1	1	$\log(MM/g)$	0.701	-0.418
				$\log(NN/g)$	-0.713	-0.398

The first PC represents 0.986 (nearly 99%) of the variation in the data and indicates that the main source of variation between the samples is the contrast between their MM and NN centred log ratios. The variation represented by the second PC is virtually zero (around 1%). Figure 2 is a biplot of the data.

Note that in Figure 2 the scales of the axes are very different; in fact the samples show much smaller variation around the second principal component (on the vertical axis) than around the first principal component (on the horizontal axis).

The fact that the second PC has virtually no variation implies that the expression for the second PC is approximately equal to a constant:

$$0.816 \times \log(MN) - 0.418 \times \log(MM) - 0.398 \times \log(NN) = \text{a constant}$$

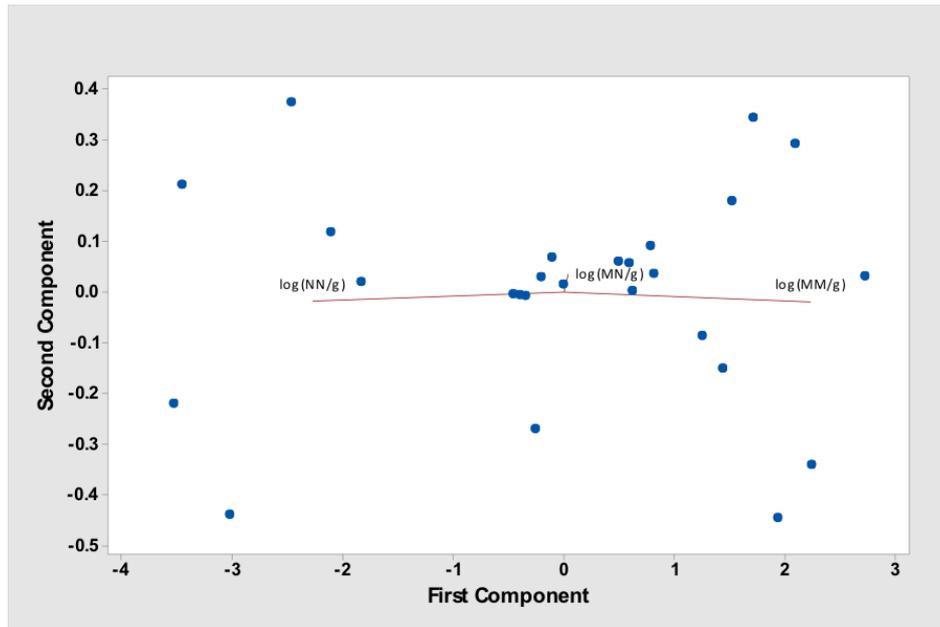


Figure 2: Biplot of genotype data.

which is approximately,

$$2 \times \log(MN) - \log(MM) - \log(NN) = \text{a constant.}$$

This equation can be rearranged to give $MN^2/(MM \times NN) = \text{a constant}$. The value of the constant can be approximated from the data by finding the mean value of $MN^2/(MM \times NN)$ for the 26 samples. Using the raw data (not shown), the constant has a value of just over 4.

The equation $MN^2/(MM \times NN) = 4$, is in line with the Hardy-Weinberg equilibrium. In this equilibrium model, if the probability of having the M genotype is p and the probability of having the N genotype is q , where $p + q = 1$, then the probabilities of MM , MN and NN are p^2 , $2pq$ and q^2 respectively. Therefore $MN^2/(MM \times NN) = 4$. The Hardy-Weinberg equilibrium implies that genotype frequencies in a population remain constant from generation to generation in the absence of other evolutionary influences.

The original derivation of the Hardy-Weinberg law was extremely long and tedious; this example shows how the result can be attained using a simple Singular Value Decomposition. It also shows that, whereas attention is usually only focused on PCs with large eigenvalues, looking at PCs with small eigenvalues can be very important especially if there is just one large PC and one small PC. So, a very interesting result was obtainable from a simple set of sample data when analysed taking account of the compositional nature of the data.

3. Compositional data in a healthcare setting

Compositional data forms a large part of internal company operational data and its analysis can provide useful insight. For example, the occurrence of failed and cancelled appointments on different days of the week, or changes in the proportions of different activities undertaken over time are important information for a service provider (Pritchett, Coleman, Campbell, and Pabary 2018). Composite bar charts can be used to present the proportions, but mask the fact that the proportions are constrained to add up to the whole. Figure 3 shows a business oriented ternary diagram illustrating the change in proportions of activities carried out in a healthcare setting. The service provider is keen to move from routine activities to more advanced value added procedures and the diagram illustrates their success in moving closer

towards the advanced practice vertex in quarter 2 compared to a predominance of routine examinations carried out in quarter 1.

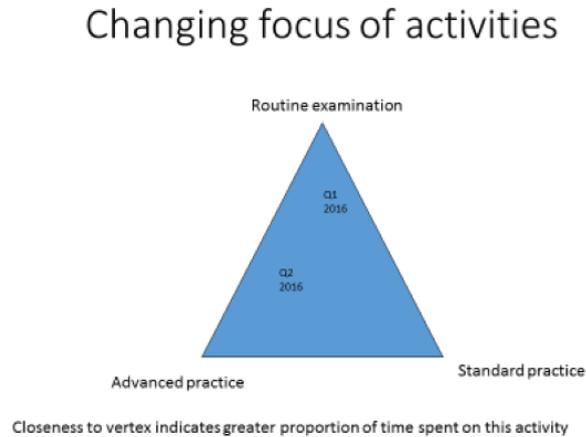


Figure 3: Ternary diagram illustrating changing practice.

Using ternary diagrams to illustrate proportions is an informative way to share data with company staff and can help motivate continual improvement in a healthcare setting.

4. Conclusions

Learning about compositional data analysis provides an important addition to the knowledge of different types of statistical analysis. There are many situations in which compositional data arise and there are applications of compositional data analysis in education, medicine, healthcare and other sectors.

Using ternary diagrams to represent data is an unusual and concise way to share insight and to motivate improvements in education and healthcare where managers and practitioners need to explore the results of their actions in depth.

Professor Aitchison's lectures were rich with applications and also theory, for which he consistently gave full and clear explanations and derivations. His style was affable and self-deprecating. He described an incident in his inaugural lecture at Hong Kong University as a warning about the capricious nature of random variation. He was demonstrating the random nature of sampling by drawing coloured balls from a bag containing a mixture of different colours but as luck would have it he consistently drew exact proportions instead of the random variation he was intending to demonstrate.

Professor Aitchison was a charming man. He had a warm family life and dedicated his compositional data book in memorable style to his wife with the words:

"To M. the constant among many variables"

He conducted the Statistics Department at Hong Kong University in a very calm and inclusive manner encouraging high level research work and a responsible, dedicated, professional focus on teaching the students.

It is a testament to his reach and influence that CODA is such a thriving community (<http://compositionaldata.com>) and that CODAWORKS conferences are so active and well attended.

References

- Ahlemeyer-Stubb A, Coleman S (2018). *Monetising Data - How to Uplift Your Business*. Wiley.
- Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.
- Coleman S (2016). “Data Mining Opportunities for Small to Medium Enterprises from Official Statistics.” *Journal of Official Statistics*, **32**(4), 849–866.
- Coleman S, Lee S (1988). “Compositional Data Analysis of Activities in a Typical Lesson.” *Hong Kong Educational Research Journal*, **3**, 98–103.
- Pritchett R, Coleman S, Campbell J, Pabary S (2018). “Understanding Our Patient Base: an Introduction to Data Analytics in Dental Practice.” *Dental Update*, **45**(3), 236–246.

Affiliation:

Shirley Coleman

Industrial Statistics Research Unit

School of Mathematics, Statistics and Physics

Herschel Building, Newcastle University

E-mail: shirley.coleman@ncl.ac.uk

URL: <https://www.ncl.ac.uk/maths-physics/staff/profile/shirleycoleman.html#background>