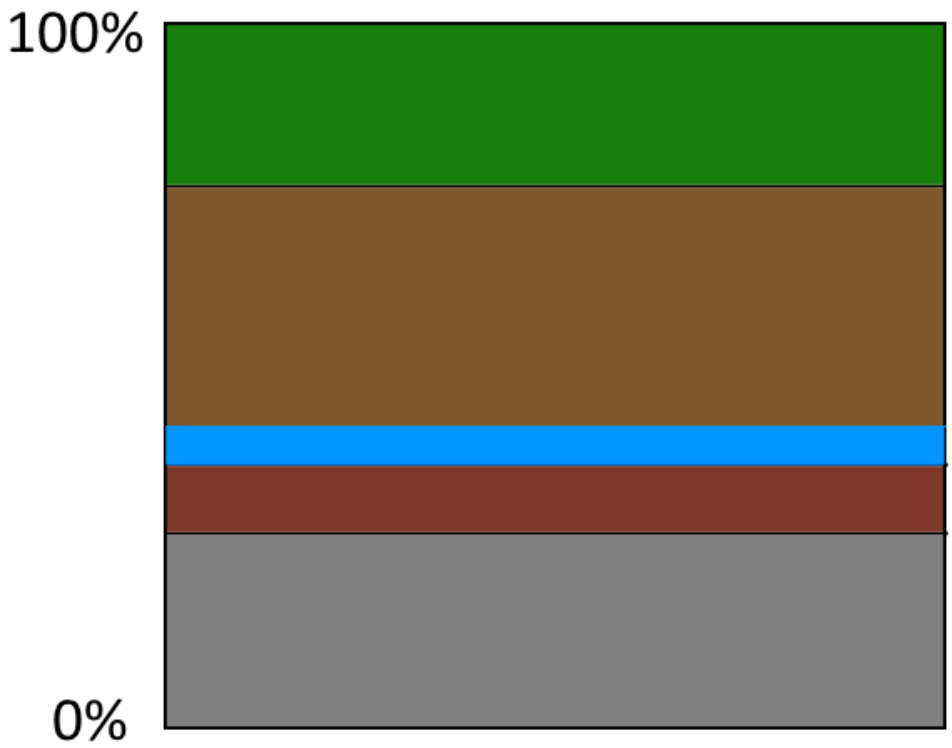
Zusammengesetzte Daten (compositional data)

Benjamin Schlegel

23. Mai 2016

Zusammengesetzte Daten sind Daten, bei denen der Anteil des einen vom Anteil eines oder mehreren anderen abhängt.

Ein einfaches Beispiel (von John Aitchison) ist folgendes: Ich habe einen Topf mit Wasser, Erde und Samen auf der Terrasse. Am Abend messe ich den Anteil dieser drei Stoffe. Am nächsten Morgen messe ich nochmals und stelle fest, dass sich der Anteil Wasser im Topf erhöht hat. Was bedeutet das nun? Eine Möglichkeit wäre, dass es in der Nacht geregnet hat und es damit mehr Wasser im Topf hat. Es könnte aber ebenso gut sein, dass es in der Nacht stark gewindet hat und der Wind Erde und Samen aus dem Topf fortgeblasen hat. Wir können es nicht wissen mit nur den Anteilen. Die Daten sind voneinander abhängig.



Auch in der Politikwissenschaft kann es solche Daten geben. Bei Majorzwahlen macht ein Kandidat prozentual weniger Stimmen, wenn ein anderer Kandidat mehr Stimmen macht. Wenn ein Kandidat schlecht abschneidet, kann das bedeuten, dass er einfach für die Wähler nicht wählbar war. Es kann aber auch sein, dass er sehr starke Gegner hatte. Das gleich gilt auch für den umgekehrten Fall. Schneidet eine Kandidatin sehr gut ab, kann es sein, dass sie besonders gut ist oder dass die anderen Kandidaten noch schlechter waren als sie und sie damit als kleinstes Übel gewählt wurde.

Tritt diese Abhängigkeit bei der abhängigen Variable einer Regression ein, kann dies zu Problemen führen. Es gibt jedoch eine Lösung: das logarithmierte Verhältnis (logratio). Dabei nimmt man die Fälle von 1 bis n-1 und teilt sie durch den nten-Fall. Aus diesem Quotient berechnet man den natürlichen Logarithmus, um das Logratio zu erhalten. Der nte-Fall fällt aus der Analyse raus.

$$y_i = ln(\frac{x_i}{x_N}), i : 1 \; bis \; N-1$$

Mit den berechneten Werten kann nun eine Regression gerechnet werden. Die Koeffizienten müssen anschliessend zurückgerechnet werden. Dazu kann folgende Formal verwendet werden:

$$x_i = e^{y_i}$$

Da auch die Varianz-Kovarianzmatrix auf der Logratio-Skala liegt, muss auch diese in die ursprüngliche Skala zurückgerechnet werden. Die kann mit folgender Formel gemacht werden:

$$vcov = \left(e^X\right)^2 \cdot vcov_{logratio}$$

weiterführende Links

A concise guide to compositional data analysis (Englisch)