Use of ratio variables in the analyses of compositional data: a simulation study

Georgia Tomova, PhD

University of Leeds, UK

25 Sep 2024

WCE

WORLD CONGRESS OF EPIDEMIOLOGY 2024

Background

Compositional data describes data containing parts that together sum to a larger whole/total.
e.g. protein + fat + carbohydrates = total energy intake
e.g. sedentary behaviour + sleep + physical activity = total time in a day

There exist different approaches to the analysis of compositional data, and they have varying levels of complexity.

One of the intuitive ways in which compositional data is often considered is by using **ratio variables:**
e.g. % energy intake from fat (fat/total energy)
e.g. % time spent sedentary (sedentary behaviour/total time)

However, the use of ratio variables in regression has been discouraged since 19th century…

WCE
WORLD CONGRESS OF EPIDEMIOLOGY 2024

Background

1897

Karl Pearson
(1857-1936)

Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs

Karl Pearson

Published: 01 January 1897 https://doi.org/10.1098/rspl.1896.0076

$$\begin{matrix} x \perp y \\ x \perp z \\ y \perp z \end{matrix} \longrightarrow \frac{x}{y} \sim \frac{z}{y} \longrightarrow r_{\frac{x}{y'}\frac{z}{y}} \approx 0.5$$

random. I term this a spurious organic correlation, or simply a spurious correlation. I understand by this phrase the amount of correlation which would still exist between the indices, were the absolute lengths on which they depend distributed at random.

WCE
WORLD CONGRESS OF EPIDEMIOLOGY 2024

Background

A causal inference perspective on the analysis of compositional data

Kellyn F Arnold, [1,2*†] Laurie Berrie, [1,2†] Peter WG Tennant [1,2,3‡] and Mark S Gilthorpe [1,2,3‡]

[1]Leeds Institute for Data Analytics, University of Leeds, Leeds, UK, [2]School of Medicine, University of Leeds, Leeds, UK and [3]The Alan Turing Institute, London, UK

However, compositional data comes in two main forms:
The total is **fixed** (a constant), **e.g. total time in a day,**
The total **varies** (a variable), **e.g. total energy intake**

Since ratios involve **division**, whether or not their use is appropriate may depend on the type of compositional data.

Methods

To explore how ratio variables perform in compositional data with fixed and variable totals:

- We simulated dietary (variable totals) and physical activity (fixed totals) data, and fasting plasma glucose as an illustrative outcome;
- 10,000 simulations with 1,000 observations in each;
- DagSim (Al Hajj et al., 2023 PLoS ONE) package in Python;

Carbs + fat + protein + alcohol = total energy

$$FPG \sim \hat{\alpha}_0 + \widehat{\alpha_1}\frac{Carbs}{Total} + \widehat{\alpha_2}\frac{Fat}{Total} + \widehat{\alpha_3}\frac{Alcohol}{Total}$$

substitution between carbs and protein

Sleep + sedentary + light PA + moderate-to-vigorous PA = total time

$$FPG \sim \hat{\beta}_0 + \widehat{\beta_1}\frac{MVPA}{Total} + \widehat{\beta_2}\frac{SB}{Total} + \widehat{\beta_3}\frac{Sleep}{Total}$$

substitution between MVPA and LPA

- Compared model estimates to simulated true effects

Results

Note: these results are part of a larger study looking at a variety of possible relationships

- 10-min substitution between moderate-to-vigorous physical activity and light physical activity **(fixed totals)**

Ground truth	Model estimate (95% SI)
-0.200 mmol/l	-0.200 (-0.201, -0.199) mmol/l

- 100-kcal substitution between carbohydrates and protein **(variable totals)**

Ground truth	Model estimate (95% SI)
0.400 mmol/l	0.475 (0.419, 0.533) mmol/l

One way to attempt and resolve this is through adjusting for the total…

This offers some improvement – how much will depend on the exact structure of the data.

Ground truth	Model estimate (95% SI)
0.400 mmol/l	0.383 (0.344, 0.422) mmol/l

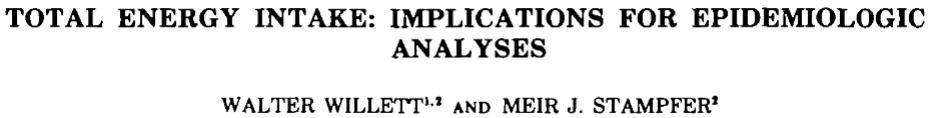
WCE
WORLD CONGRESS OF EPIDEMIOLOGY 2024

Discussion

- Ratio variables perform differently depending on whether the compositional totals are fixed or not.
- In the past (ever since Pearson 1897), the use of ratios in regression or correlation analyses has been discouraged (because of ‘spurious correlations’).
- However, this is only a problem when dividing by a **variable** (e.g. **total energy intake**). When dividing by a **constant** (e.g. **total time in a day**), ratio variables do not introduce ‘spurious correlations’.
- This is logical, but not always fully appreciated in applied research.

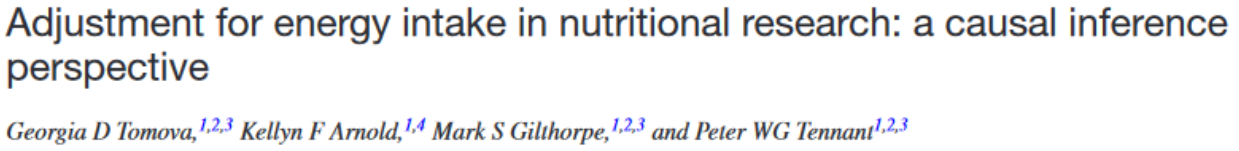
Discussion

- In compositional data with **variable totals (e.g. dietary data)**, ratios *do* introduce bias.
  Why?
  - Division by the total is often used as a way to standardise against / account for the total.
  - However, this is not a valid way to adjust. It simply conflates the numerator and denominator.
  - If the denominator is strongly associated with the outcome, this can lead to sign-reversal.
  - e.g. issues with the nutrient density model in nutrition have been discussed in the past



TOTAL ENERGY INTAKE: IMPLICATIONS FOR EPIDEMIOLOGIC ANALYSES

WALTER WILLETT[1,2] AND MEIR J. STAMPFER[2]



Adjustment for energy intake in nutritional research: a causal inference perspective

Georgia D Tomova,[1,2,3] Kellyn F Arnold,[1,4] Mark S Gilthorpe,[1,2,3] and Peter WG Tennant[1,2,3]

- In the **physical activity** domain (where totals are **fixed**), issues with ratios variables are widely known, and more sophisticated methods are encouraged (e.g. CoDA).
  However, ratio variables do not introduce bias in such data.

WCE
WORLD CONGRESS OF EPIDEMIOLOGY 2024

Summary

- Ratio variables are commonly used in compositional data, even though their use has been discouraged since 1897.

- Compositional data can have totals that are either variable or fixed (e.g. in dietary vs time-use data), which are different in nature.

✓ **Fixed totals:**

- ratios are simply division by a constant
- rescales the estimates but does not introduce spurious associations

✗ **Variable totals:**

- coefficient estimates of ratios conflate the effects of the numerator and denominator
- this may lead to spurious associations (or even sign-reversal)
- conditioning on the total may not fully eliminate the bias but offers improvement

WCE
WORLD CONGRESS OF EPIDEMIOLOGY 2024

Acknowledgments

Dr Peter Tennant
University of Leeds

Prof Michelle Morris
University of Leeds

Dr Rosemary Walmsley
Public Health Wales

Dr Laurie Berrie
University of Edinburgh

Funding

The Alan Turing Institute

Contact details

G.D.TOMOVA@LEEDS.AC.UK

@GeorgiaTomova

WCE
WORLD CONGRESS OF EPIDEMIOLOGY 2024