**Logic Breakdown**

**Explanation of Compositional Data**

Some data sets have components that will always sum to an arbitrary total that is pre-determined by sampling process being used to generate said data set. When this occurs, the data is referred to as being compositional. Compositional data consist of proportions, not absolute values and therefor the relationships among components cannot be interpreted in the same way as non-compositional data. Most notably the change of the absolute count of one component between sampling sets will alter the associated proportions of all components regards if their absolute count also alters.

This errors like this occur because components of compositional data sets are not absolute counts but rather proportions. For example, consider two data sets for which organism A has a count of 500 in set one and 1500 in set two. Organism B has a count of 500 for both data sets. Looking at total count it is easy to conclude that organism A increased between sets and organism B did not change at all. However, if these data sets became composition with 100 being the arbitrary total the data sets sum to, then proportions would be reported instead of total counts. Now organism A goes from 0.5 (500/1000) to 0.75 (1500/2000) between the data sets and organism B goes from 0.5 (500/1000) to 0.25 (500/2000). Organism B experienced a decrease in proportion even though its count total stayed consistent. If it was incorrectly assumed that these proportions were total counts, the conclusion that would have been drawn is that organism B’s count decreased between the data sets when really just its proportion changed. Therefore, a failure to recognizing when an experiment’s results are produced as compositional data and adjusting analysis accordingly can lead to complete incorrect conclusions being drawn about the actual population in the sample.

Biological count data (relative data) is an example of compositional data with the arbitrary total being the read depth of the sequencing platform. As long as a limit exists to sequencer capabilities, the data produced by them meets the definition of compositional data. A common practice to transform these proportions back to total counts is to use the method of normalization. The method involves subsampling and introduces another level of error into the results. However, techniques for compositional data analysis (CoDA) exists and thus reduce the level error associated with conclusions when used in place of normalization. The reason normalization is more common in literature even though it introduces a higher error rate is that statistical analysis which can be performed on normalized data is usually considered easier than CoDA.

**Explanation of spurious correlation**

Correlations observed in relative data do not correlate with those observed in the absolute data the relative data stemmed from. This is because of the existence of *spurious correlation.* Spurious correlation is the term used to describe the following mathematical phenomenon: if independent values are divided by the same independent value, then correlation is introduced among their quotients. Because the creation of relative data incudes such a process, there is always the possibility that correlation in a relative dataset is spurious rather than true correlation. Therefore, the statement *if there is correlation observed among elements of a relative data set, then the same correlations must also be present among the elements of the parental absolute data se*t is incorrect due to the fact that any correlation classified as spurious correlation will not be present in the absolute data.

This is illustrated by the fact that the correlation coefficients generated from relative data have high p values and low R values when compared to the correlation coefficients of the respective absolute data. (i.e., relative coefficients have low precision and low sensitivity when used as a metric for absolute coefficients). However, just because spurious correlation will be present in relative data, it does not mean that this is the only correlation present in relative data. Not all rectangles are squares, but some are. If you need a square, grabbing the first rectangle you see probably won’t work. But carefully searching a field of many rectangles will most likely find you at least one square (*I wrote this analogy myself and I am proud*). A hypothesis exists that a method could be developed that can detect which of the correlation present in relative data is indicative of the same correlation existing in the absolute data. In other words, it may be possible to develop methods that allow the relative correlation to be used as a proxy for absolute correlation.

Proportionality

The proportional nature of compositional data is both a problem and a solution when trying to connect absolute correlation to relative correlation.

The transition from absolute count data to proportions is actually very significant mathematically. Any time there is a 2D or 3D object that can be defined by points on the x, y, z plane, this is referred to as being in Euclidean space. This definition encapsulates most things encountered in daily life and, therefore, the mathematics associated with these things also fall into Euclidean space. Something that is not a part of the Euclidean space is proportions. Any mathematical operation that transforms data into proportions is also is transforming the data to another mathematical dimension referred to as Atkinson space. Once the quirk of mathematics is realized it is easy to see why trying to directly compare proportions to count data and any other data that belongs to the Euclidean space gets problematic. If you take a procedure developed on Earth and try to execute it step by step on the moon the results will not be what was expected due to the difference in gravity. That being said, because the way gravity and other forces differ on the moon is understood, adapting Earth intended procedures to produce the same result on the moon is a realistic solution. Atkinson space is not a mystery, it too has been described using mathematical rules and theorems, just ones that differ from the familiar ones used for Euclidean space.

In summary, it is known that the errors produced by trying to analyze compositional data using standard statistical methods are caused by not addressing the fact that data can no longer be considered Euclidean space. It is also known that log-ratio transformations can be used as a way to transform non-Euclidean data back into a form our Euclidean centric brains can comprehend. (These two paragraphs might not be 100% correct mathematically speaking regarding word choice but I am confident of the concept it conveys.)

CoDA explores the unique nature of compositional data, one of which being the property of proportionality. Proportionality has been identified as an alternative to correlation. Research shows that proportionality coefficient is the same regardless if a dataset’s relative or absolute data was used to calculate it. It too can be used to describe datasets with terms like proportionality coefficients. E.g., the metric ρ which has a range of [-1,1], where 1 indicates perfect proportionality, conveys similar concepts as Pearson’s correlation coefficient. Developing methods that can link ρ values to more commonly used correlation values allows information present in a relative data set to be linked back to its parent absolute dataset.

propr Package

* Default uses center log-ratio transformations (clr)
* Two metrics, phit Φ and perb ρp
* The log transformation produces matrix A consisting of N rows (samples) with D columns (features)
* Proportionality is then calculated via these columns, both metrics return a D2 matrix
* Paper defines this D2 matrix as “relating each combination of log-ratio transformed feature vectors, Ai and Aj (i, j ∈ D)” which I believe means: feature vector = Column, at column i and column j in D
* The metric phis (Φs) can be calculated from Φ and it is the naturally symmetric variant
* A function exists that relates Φs and ρp
* Also has the capacity to use alternate log-ratio transformations (alr) if a reference feature can be created that will be the same for all samples (e.g., read depth, read percent etc.)
* Alr has more inferencing capabilities than clr and in theory can be used back calculate absolute counts from relative data
* User indicates that the function should switch from clr to alr by assigning said reference to the variable ivar, present in the arguments all three metric functions (phit, perb, or phis)