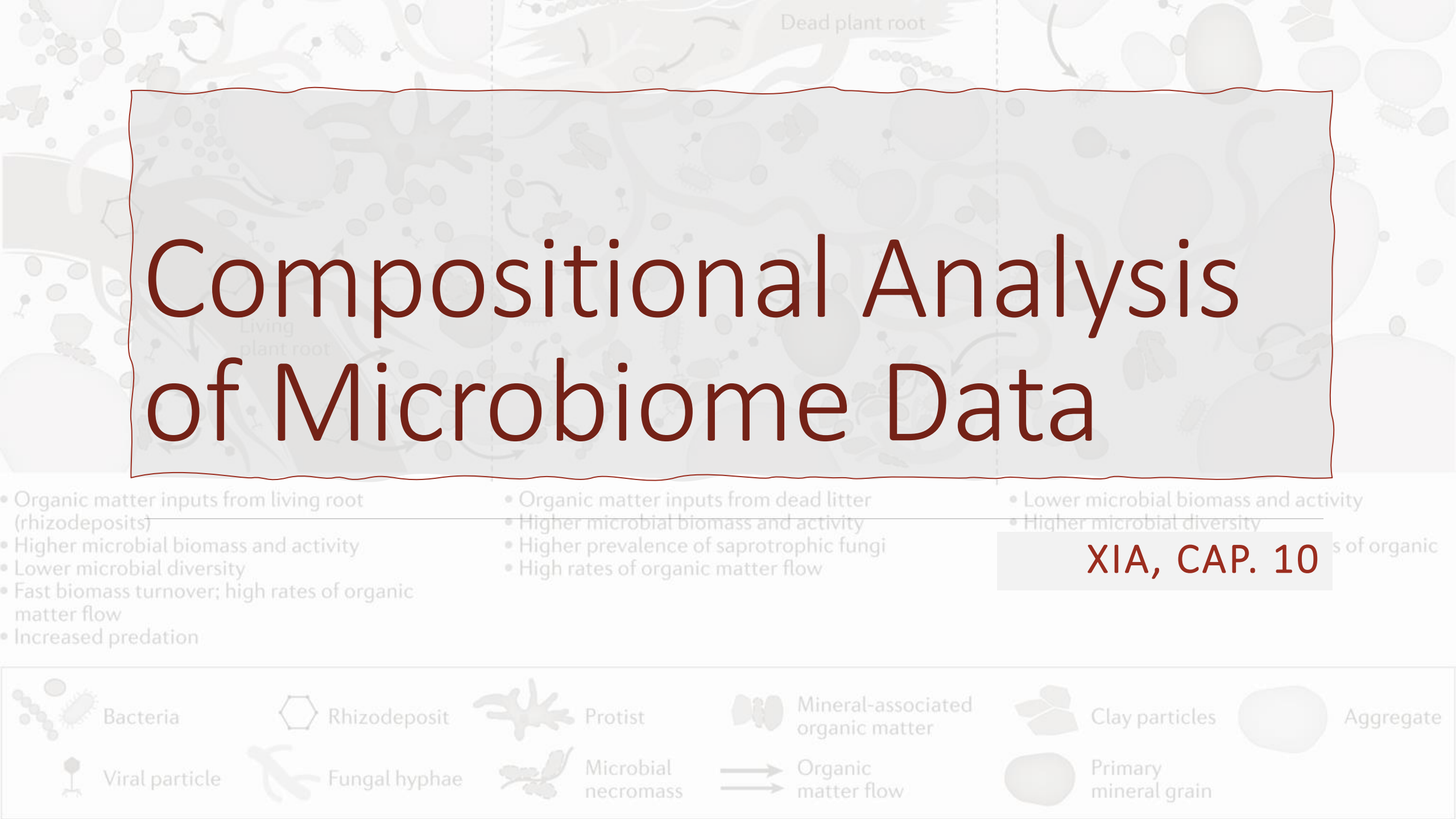
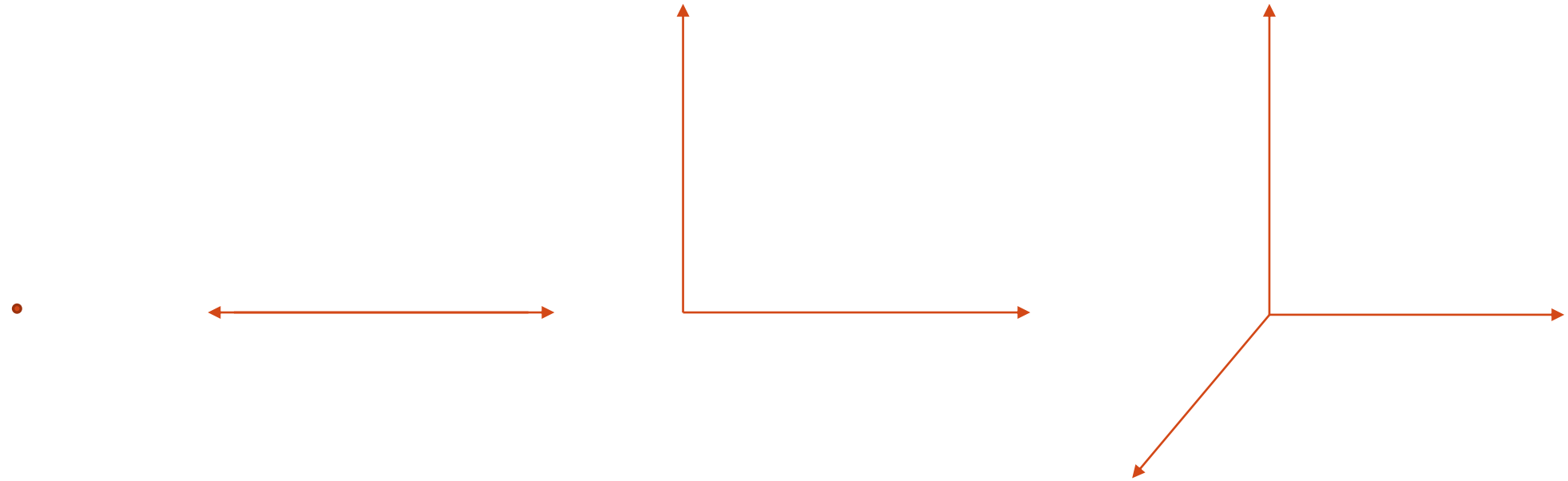
Compositional Analysis of Microbiome Data

XIA, CAP. 10



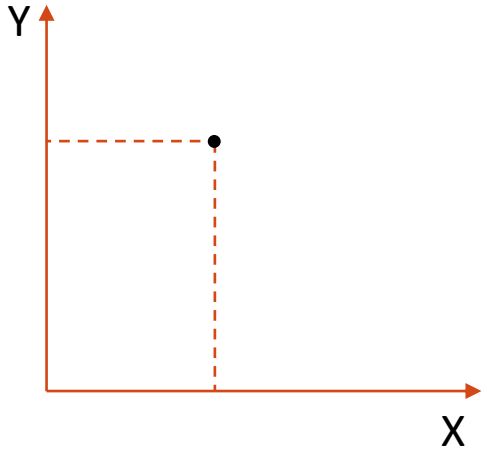
$\mathbb{R}^n$

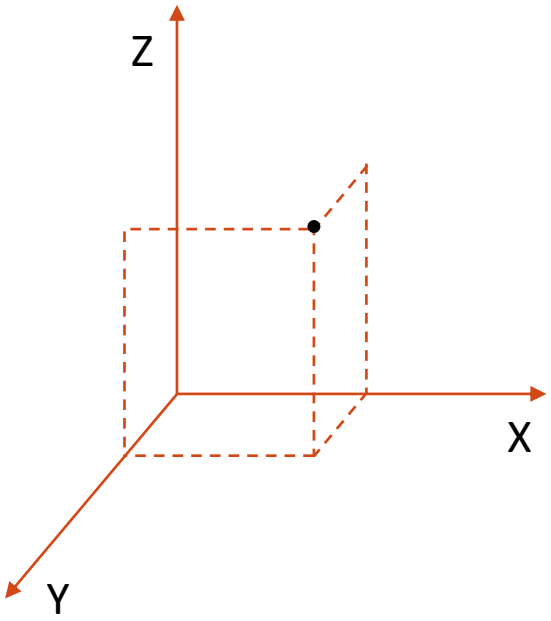
$\mathbb{R}^0$ $\mathbb{R}^1 = \mathbb{R}$ $\mathbb{R}^2$ $\mathbb{R}^3$ $\cdots$



$\mathbb{R}^n$

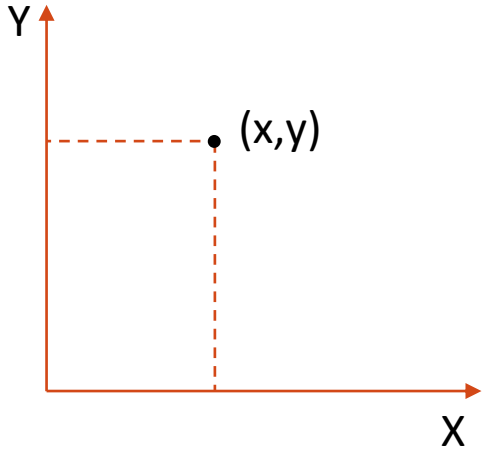
$\mathbb{R}^0$ $\mathbb{R}^1 = \mathbb{R}$ $\mathbb{R}^2$ $\mathbb{R}^3$ • • •

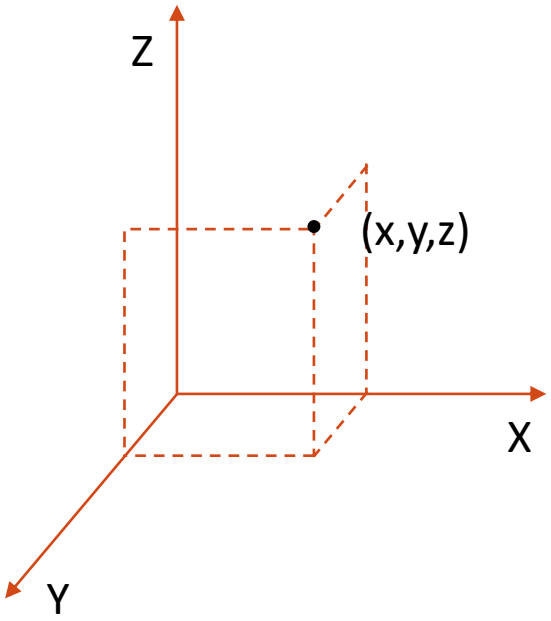




$\mathbb{R}^n$

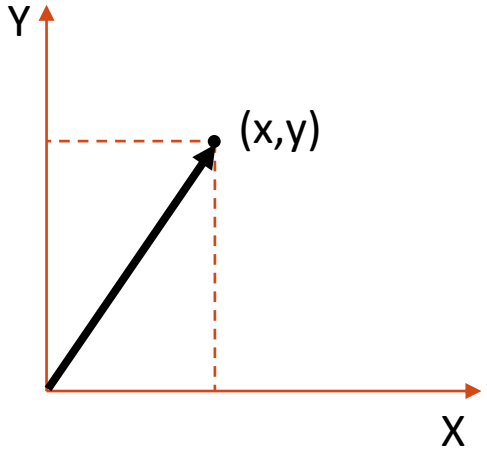
$\mathbb{R}^0$ $\mathbb{R}^1 = \mathbb{R}$ $\mathbb{R}^2$ $\mathbb{R}^3$ . . .

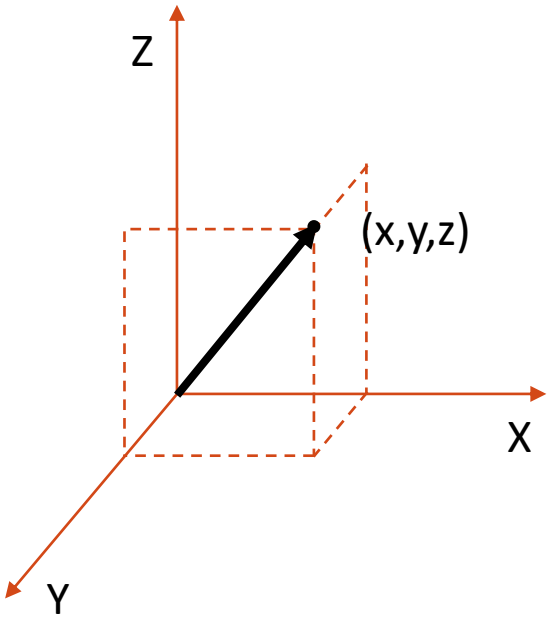




$\mathbb{R}^n$

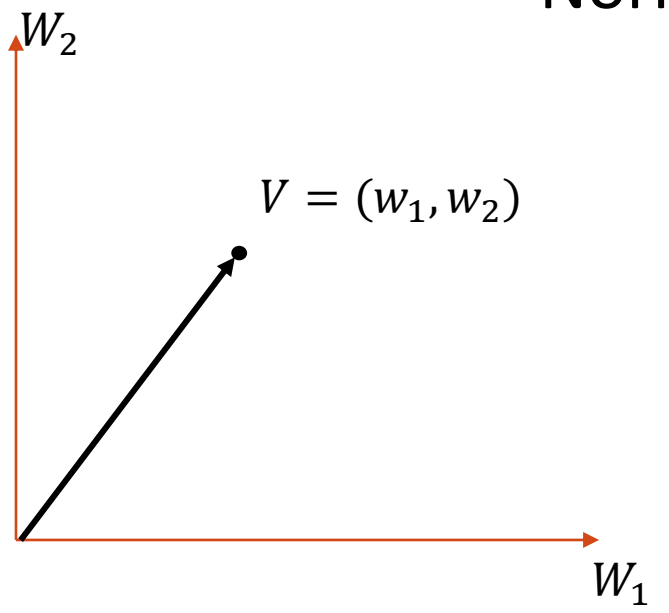
$\mathbb{R}^0$ $\mathbb{R}^1 = \mathbb{R}$ $\mathbb{R}^2$ $\mathbb{R}^3$ . . .



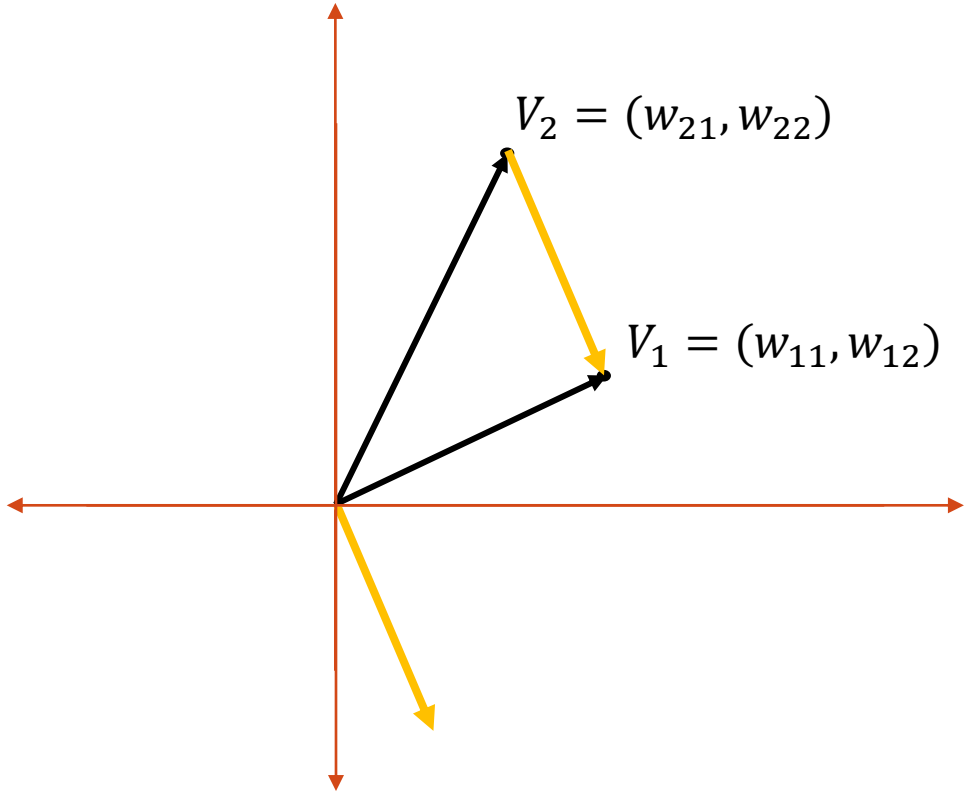


Métricas

Normas



Distancias



Manhattan $\quad \left|\left|V\right|\right|_1 = |w_1| + |w_2|$

$$d_1(V_1, V_2) = \left|\left|V_1 - V_2\right|\right|_1 = |w_{11} - w_{21}| + |w_{21} - w_{22}|$$

Euclidiana $\quad \left|\left|V\right|\right| = \sqrt{w_1^2 + w_2^2}$

$$d(V_1, V_2) = \left|\left|V_1 - V_2\right|\right| = \sqrt{(w_{11} - w_{21})^2 + (w_{12} - w_{22})^2}$$

Introduction to Compositional Analysis

10.1

Análisis de datos de composición de microbioma

Composición es “el acto de unir partes o elementos para formar un todo”, o es “la forma en que dichas partes se combinan o relacionan: la forma en que se constituyen”.

Los datos de composición describen cuantitativamente las partes del todo y proporcionan solo información relativa entre sus componentes.

Aitchison Simplex

Mathematically, a data is defined as compositional, if it contains D multiple parts of nonnegative numbers whose sum is 1 or any constant-sum constraint. It can be formally stated as:

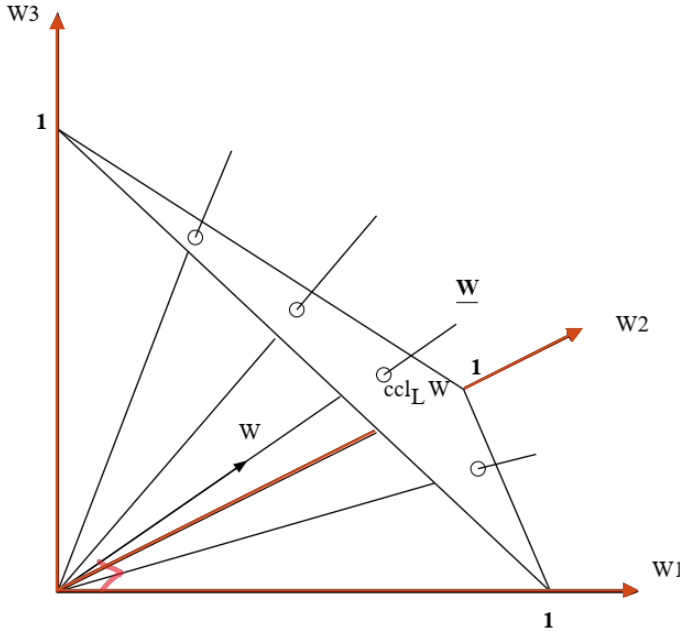
$$S^D = \left\{ X = (x_1, x_2, \ldots, x_D) \mid x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}$$

- Compositional data can be represented by constant sum real vectors with positive components.
- This defines the sample space of compositional data as a hyperplane, called the **simplex**.
- κ is arbitrary. Depending on the units of measurement or rescaling, frequent values are 1 (per unit, proportions), 100 (percent, %), 10^6 (ppm, parts per million), and 10^9(ppb, parts per billion).

Análisis de datos de composición de microbioma

The simplex

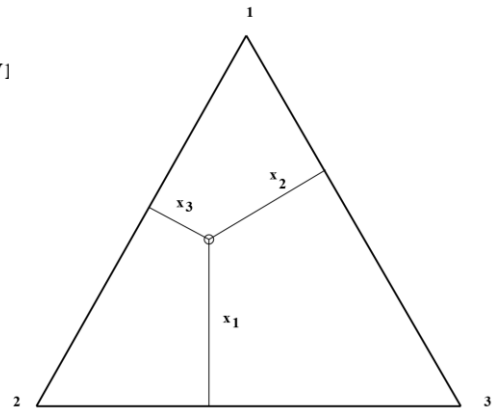
- a 0-dimensional simplex is a point,
- a 1-dimensional simplex is a line segment,
- a 2-dimensional simplex is a triangle,
- a 3-dimensional simplex is a tetrahedron,



The set $\mathcal{C}^{D-1}$ of all D-compositions will be called the $(D-1)$-dimensional compositional space.

The compositional closure mapping from $\mathbb{R}_+^D$ to $\mathcal{C}^{D-1}$-denoted by ccl- is defined by

$$\mathrm{ccl}\mathbf{w} = \underline{\mathbf{w}} \quad \left(\mathbf{w} \in \mathbb{R}_+^D\right).$$



Análisis de datos de composición de microbioma

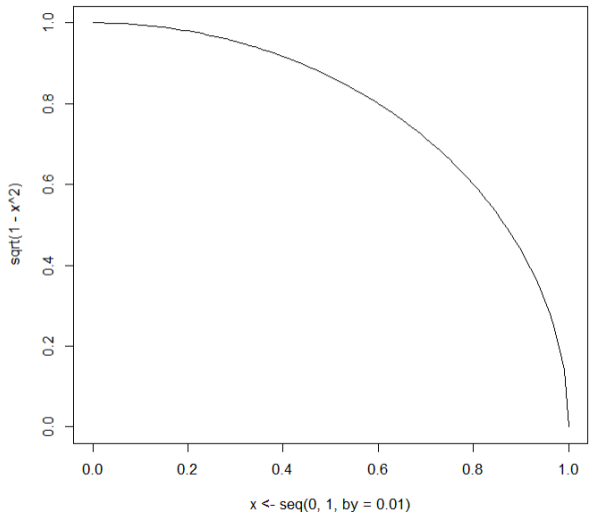
Spherical criterion

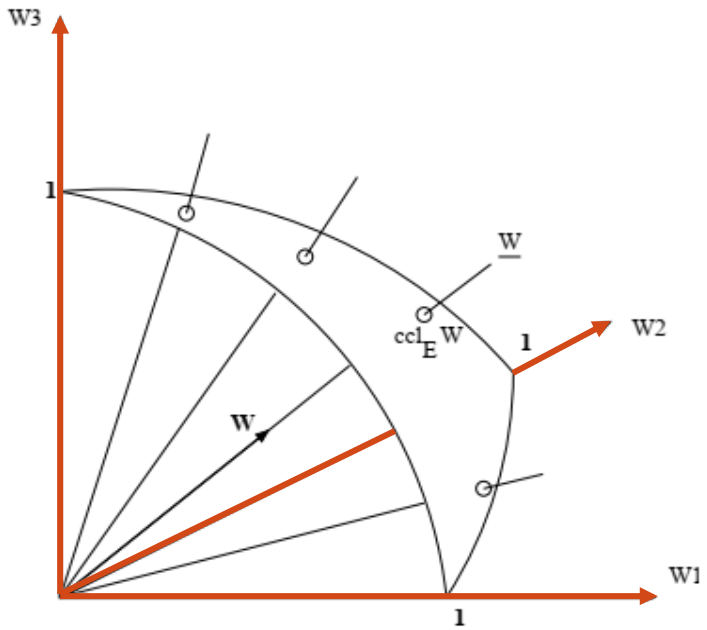
Hyperbolic criterion

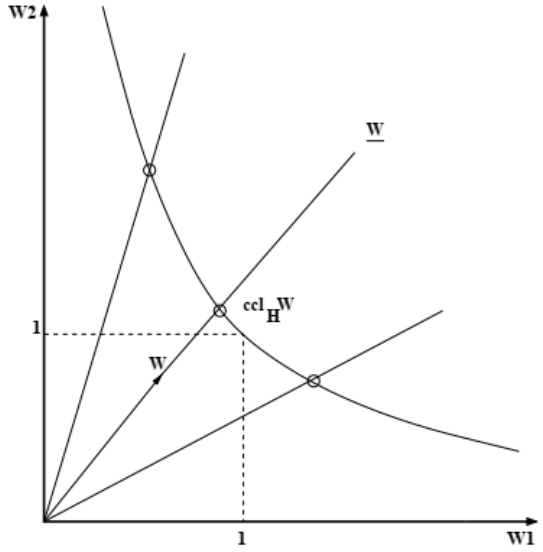
Ejemplos:

$$S^D = \left\{ X = (x_1, x_2, \ldots, x_D) \mid x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i^2 = \kappa \right\}$$

$$S^D = \left\{ X = (x_1, x_2, \ldots, x_D) \mid x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} \sqrt{x_i} = \kappa \right\}$$







Problems with Standard Statistical Methods

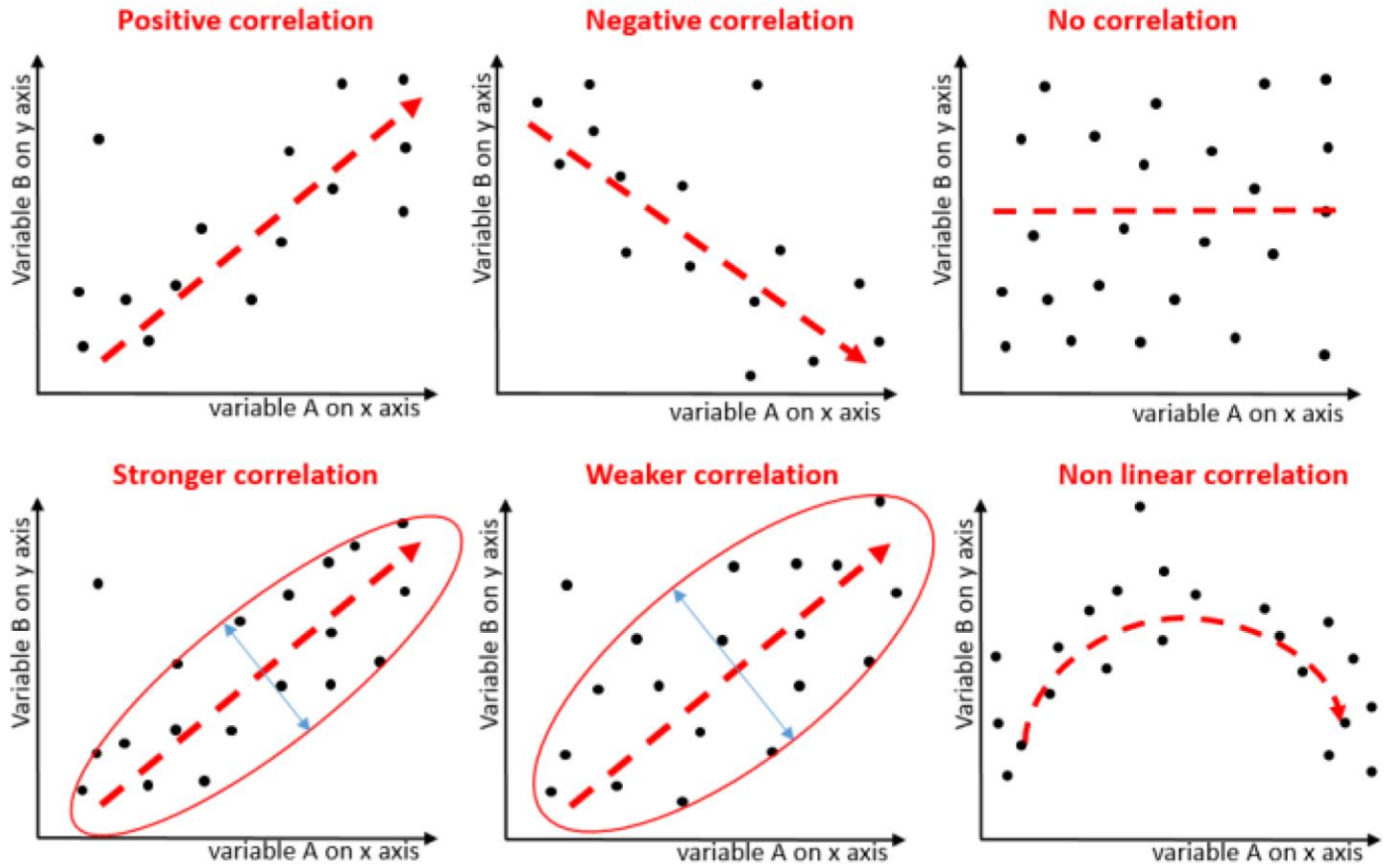
Applying them to compositional data may yield misleading results because the compositional data represent the special properties of the sample space, the simplex.

In “The Statistical Analysis of Compositional Data” (Aitchison 1986), John Aitchison reviewed and discussed some challenging problems in compositional data analysis.

- **Correlation**
- **Bias**
- **High dimensionality**
- **Constant-sum problem**
- **Parametric modelling**

Problems with Standard Statistical Methods: Spurious Correlation

Standard data analysis techniques, such as correlation analysis, rely on the assumption of the Euclidean geometry in real space (Eaton 1983).



$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{s_x^2 {s_y}^2}}$$

$$\mathrm{Cov}(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1}$$

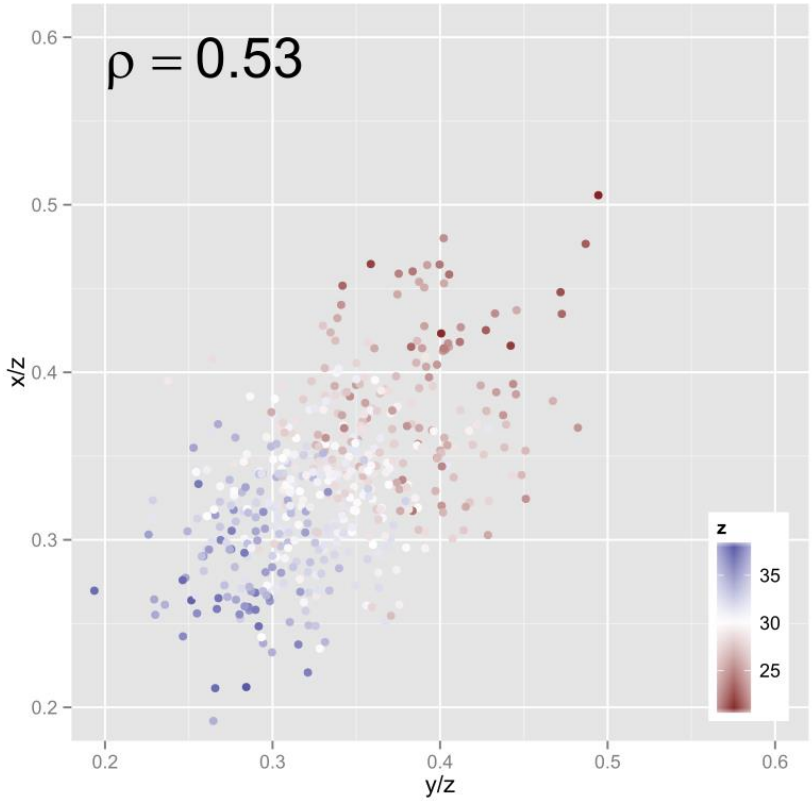
$$s_x^2 = \frac{\sum \left(X_i - \bar{X}\right)^2}{n-1}$$

Problems with Standard Statistical Methods: Spurious Correlation

Pearson (1897) "If $u = f(x, y)$ and $v = g(z, y)$ be two functions of three variables x, y, z, and these variables be selected at random so that there exists no correlation between x and y, y and z, or z and x, there will still be found to exist correlation between u and v. . . . That is likely to occur when u and v are indices with the same denominator".

Simple example of spurious correlation:

Let $x, y \sim N(10, 1)$ and $z \sim N(30, 3)$ be independent. We simulate some values x/z and y/z for each triplet, and correlation will be found between these indices.



The problem: negative bias & spurious correlation

Example: scientists A and B record the composition of aliquots of soil samples; A records (animal, vegetable, mineral, water) compositions, B records (animal, vegetable, mineral) after drying the sample; both are absolutely accurate (adapted from Aitchison, 2005)

sample A	x_1	x_2	x_3	x_4
1	0.1	0.2	0.1	0.6
2	0.2	0.1	0.2	0.5
3	0.3	0.3	0.1	0.3

sample B	x'_1	x'_2	x'_3
1	0.25	0.50	0.25
2	0.40	0.20	0.40
3	0.43	0.43	0.14

corr A	x_1	x_2	x_3	x_4
x_1	1.00	**0.50**	0.00	**-0.98**
x_2		1.00	**-0.87**	**-0.65**
x_3			1.00	0.19
x_4				1.00

corr B	x'_1	x'_2	x'_3
x'_1	1.00	**-0.57**	**-0.05**
x'_2		1.00	**-0.79**
x'_3			1.00

Problems with Standard Statistical Methods: Negative bias difficulty

Let us consider the general case of a D-part composition $\mathbf{x}$, subject to the now familiar unit-sum constraint, $x_1 + \cdots + x_D = 1$. Since

$$\operatorname{cov}(x_1, x_1 + \cdots + x_D) = 0$$

we have

$$\operatorname{cov}(x_1, x_2) + \cdots + \operatorname{cov}(x_1, x_D) = -\operatorname{var}(x_1).$$

is negative except for the trivial situation where the first component is constant.

Thus at least one of the covariances on the left must be negative or, equivalently, there must be at least one negative element in the first row of the crude covariance matrix K.

The same negative bias must similarly occur in each of the other rows of K so that at least D of its entries must be negative.

These negative properties are equally displayed by estimated covariance matrices.

Hence correlations are not free to range over the usual interval $(-1, 1)$ subject only to the non-negative definiteness of the covariance or correlation matrix, and there are bound to be problems of interpretation.

$$\operatorname{cov}(x_1, x_1 + \cdots + x_D) = 0$$

$Cov(X,Y) = E\left[(X-E(X))(Y-E(Y))\right] = E(XY) - E(X)E(Y)$

Ahora, sea $X = x_1$ y $Y = x_1 + x_2 + \cdots + x_D$

Queremos calcular $Cov(X,Y)$ pero

$Y = 1$ (es decir, es constante)

Entonces $Cov(X,Y) = Cov(X,1) = E(X \cdot 1) - E(X)E(1)$

$= 1 \times E(X) - E(X) = 0$

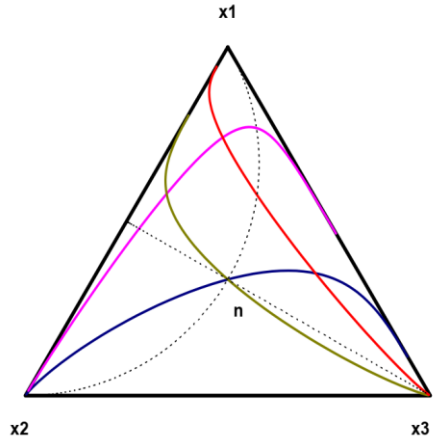
→ 1

para cualquier v.a. X.

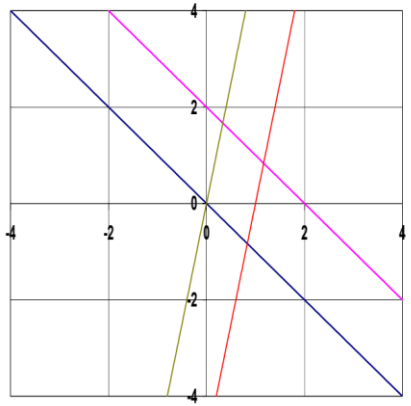
Problems with Standard Statistical Methods: High Dimensionality and constant sum restriction

This problem is due to distortions of the multivariate pattern of variability.

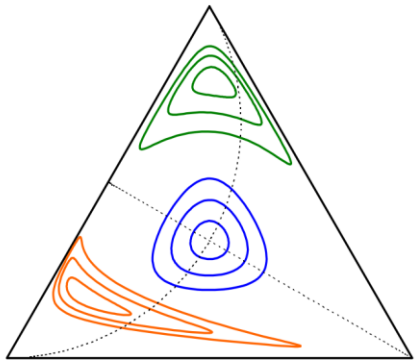
- When the analysis is restricted to a selection of subcompositions rather than the compositions as a whole, then it projects a partial analysis and loses a picture of the multivariate pattern of variability.
- Due to unit-sum constraint confines compositional vectors to a simplex, graphical distortions happened: graphical pattern seen is no guarantee as the same in familiar space such as $\mathbb{R}^2$.

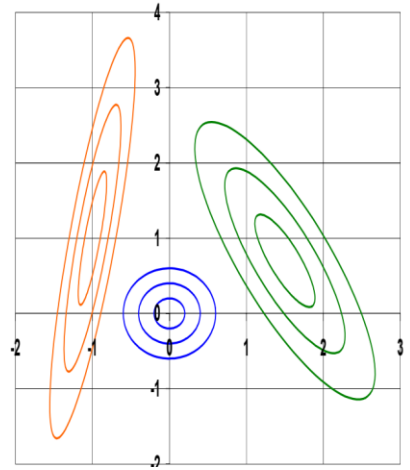


in $\mathcal{S}^3$



coordinate representation





Problems with Standard Statistical Methods: Parametric modelling

It is difficult to see how analysis of the patterns of variability can ever be wholly successful in the absence of a rich enough parametric class of distributions over the appropriate sample space.

For example, the multinormal class and its transformed classes such as the multivariate lognormal class have proved themselves to be flexible instruments in the analysis of data in $\mathbb{R}^d$ and $\mathbb{R}_+^d$. For the simplex sample space such classes of distributions have been extremely slow to emerge, with only the Dirichlet class and a few simple generalizations being considered until recently.

Unfortunately, these Dirichlet classes turn out to be totally inadequate for the description of the variability of compositional data.

Análisis Estadístico de Datos de Composición

Principles of Compositional Data: Scaling invariance

Aitchison proposed three fundamental principles for the analysis of compositional data.

The principles are all rooted in the definition of compositional data: only ratios of components carry information.

1. **Scaling invariance.** statistical inferences about compositional data should not depend on the scale used.

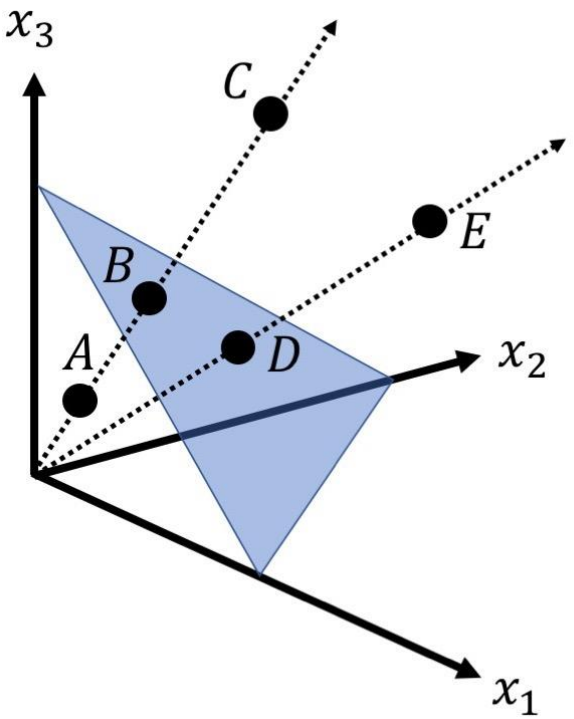
   For example, the vectors

   a = [11, 2, 5],

   b = [110, 20,50], and

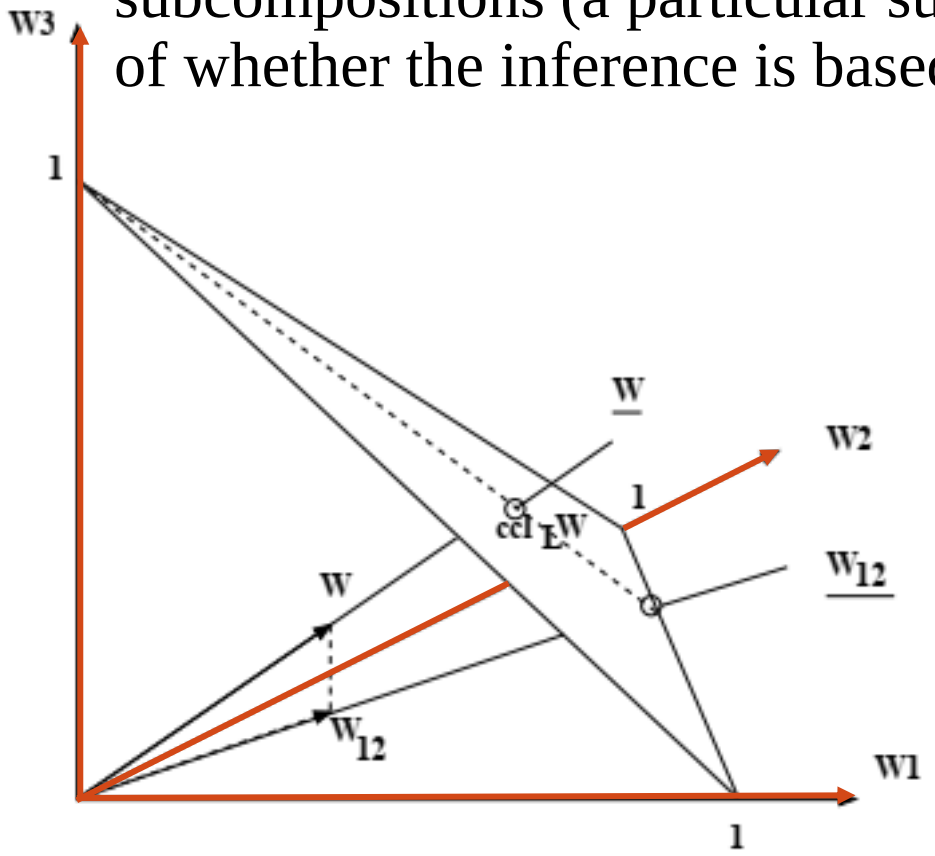
   c = [1100, 200, 500]

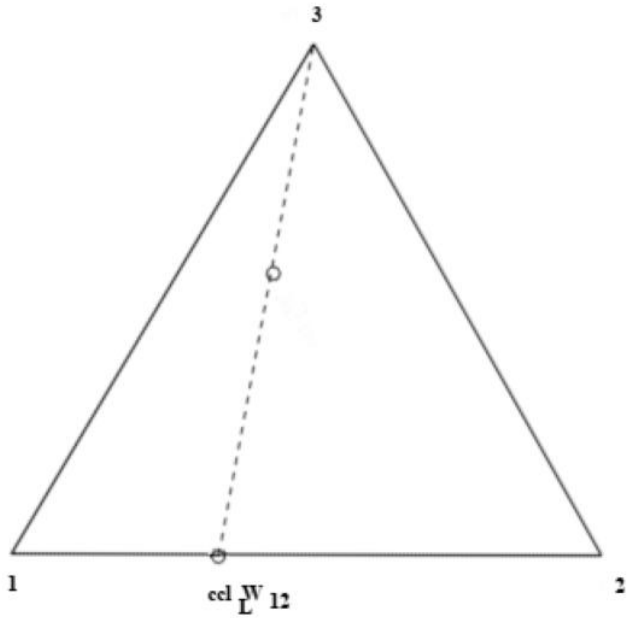
   represent all the same composition.



Principles of Compositional Data: Sub compositional coherence

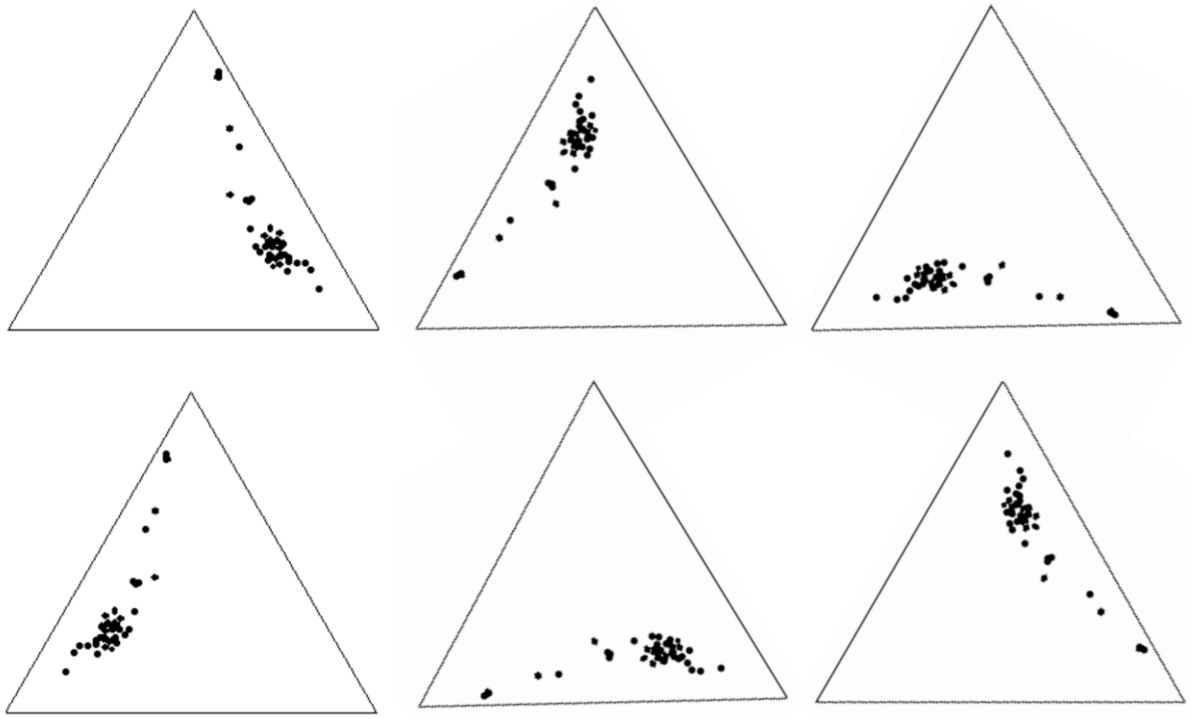
2. **Sub compositional coherence.** Analyses should depend only on data about components (or parts) within that subset and statistical inferences about subcompositions (a particular subset of components) should be consistent, regardless of whether the inference is based on the subcomposition or the full composition.





Principles of Compositional Data: Permutation Invariance

3. **Permutation invariance.** Conclusions of a compositional analysis should not depend on the order (the sequence) of the components (the parts).



A Family of Log-Ratio Transformations

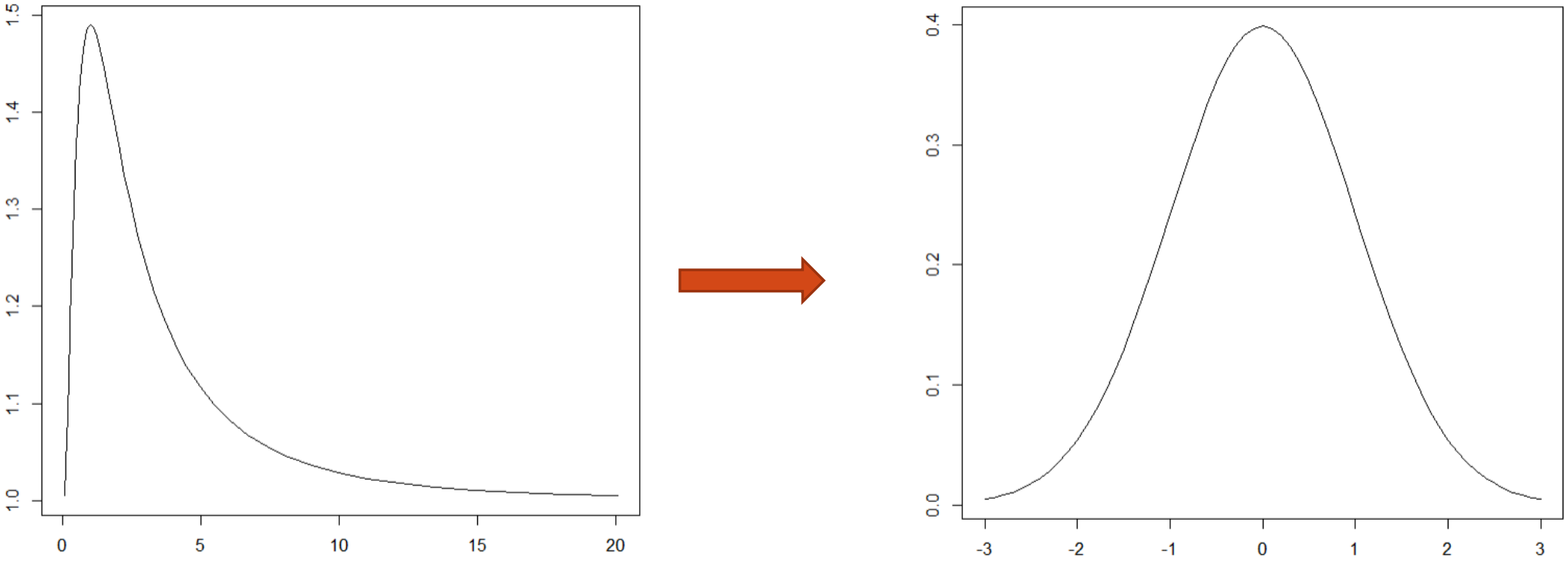
The major problem with compositional data is that the data points do not map to Euclidean space, but instead to the Aitchison simplex (Aitchison 1986). The question is:

How to analyze compositional data? Should we move or stay with the simplex?

Because standard statistical methods cannot solve the compositional data problems in simplex, the critical step towards compositional data analysis is to provide an approach for a one-to-one mapping onto a real space.

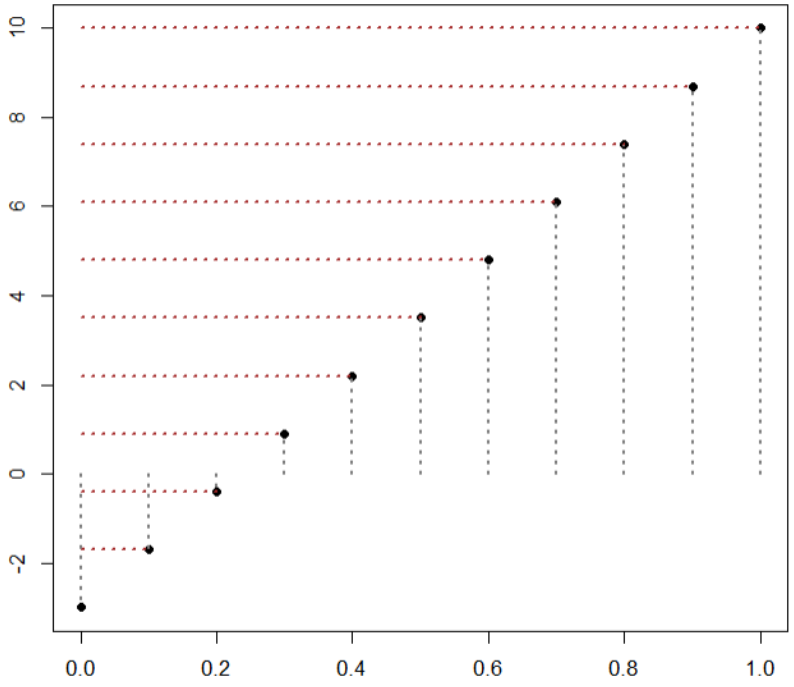
1. first transform compositions into real space using a log-ratio transformation,
2. then to apply standard statistical methods to the transformed data,
3. finally, return to the simplex by using the inverse log-ratio transformation.

Transformaciones

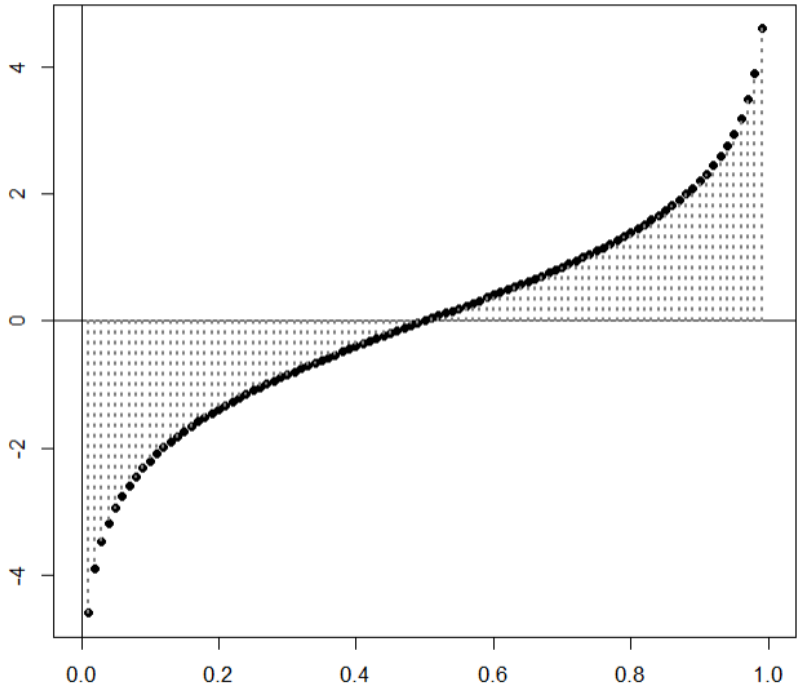


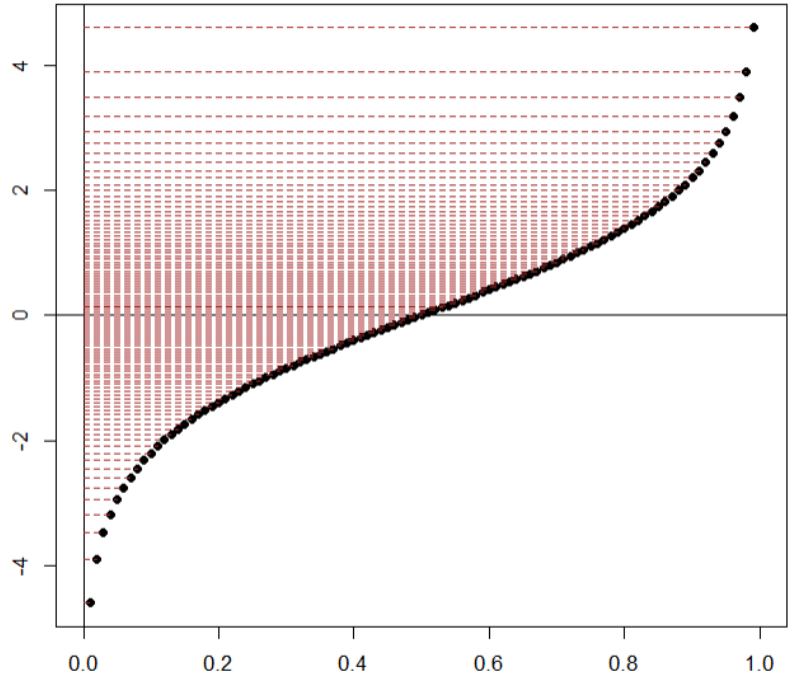
Transformaciones

[0,1] -> [-3,10] Usamos y=13x-3



[0,1] -> [-inf,inf] Usamos log(x/(1-x))





Additive log-ratio transformation

The additive log-ratio (alr) transformation is the simplest one which chooses one component as a reference.

The original approach proposed in Aitchison (1986) for the compositional data analysis was based on the additive log-ratio (alr) transformation. It is defined as:

$$\mathrm{alr}(x) = \left[\ln\left(\frac{x_1}{x_D}\right), \ldots, \ln\left(\frac{x_i}{x_D}\right), \ldots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right]$$

The distinguishing feature of this formula is to map a composition in the D-part Aitchison simplex none isometrically to a D-1 dimensional Euclidean vector.

We can also consider

$$\mathrm{alr}_j(x) = \left[\ln\left(\frac{x_1}{x_j}\right), \ldots, \ln\left(\frac{x_{j-1}}{x_j}\right), \ln\left(\frac{x_{j+1}}{x_j}\right) \ldots, \ln\left(\frac{x_D}{x_j}\right)\right]$$

Centered log-ratio

Centered log-ratio (clr) transformation maps a composition in the D-part Aitchison simplex isometrically to a D dimensional Euclidean vector. The clr representation of composition $x = (x_1, \ldots, x_i, \ldots, x_D)$ is defined as the logarithm of the components after dividing by the geometric mean of x :

$$\mathrm{clr}(x) = \left[\ln\left(\frac{x_1}{g_m(x)}\right), \ldots, \ln\left(\frac{x_i}{g_m(x)}\right), \ldots, \ln\left(\frac{x_D}{g_m(x)}\right)\right],$$

with $g_m(x) = \sqrt[D]{x_1 \cdot x_2 \cdots x_D}$ ensuring that the sum of the elements of clr(x) is zero.

1. clr avoids subjectivity of the choice of the denominator
2. clr ends up with D components instead of only D − 1 for alr.
3. **these D components sum up to zero.**

From a geometrical point of view, there is one more composition than necessary to form the basis in the Aitchison geometry, being just of dimension D−1.

The D components of clr(x) sum up to zero.

$$clr(\underline{x}) = \left[\ln\left(\frac{x_1}{g_m(\underline{x})}\right), \ln\left(\frac{x_2}{g_m(\underline{x})}\right), \ldots, \ln\left(\frac{x_D}{g_m(\underline{x})}\right)\right]$$

$$\sum_{i=1}^{n} \ln \frac{x_i}{g_m(\underline{x})} = \sum_{i=1}^{n}\left[\ln(x_i) - \ln g_m(\underline{x})\right]$$

Pero $\ln g_m(\underline{x}) = \ln\left(x_1 \cdot x_2 \cdots x_D\right)^{1/D} = \frac{1}{D}\ln(x_1 \cdot x_2 \cdots x_D)$

$$= \frac{\ln x_1 + \ln x_2 + \cdots + \ln x_D}{D} = \sum_{j=1}^{n} \frac{\ln x_j}{D}$$

Entonces

$$\sum_{i=1}^{n} \ln \frac{x_i}{g_m(\underline{x})} = \sum_{i=1}^{n} \ln(x_i) - \sum_{i=1}^{n} \frac{1}{D}\left[\sum_{j=1}^{n} \ln x_j\right]$$

$$= \sum_{i=1}^{n} \ln(x_i) - \frac{1}{D} D \left[\sum_{j=1}^{n} \ln x_j\right]$$

$$= \sum_{i=1}^{n} \ln(x_i) - \sum_{j=1}^{n} \ln x_j = 0$$

Isometric Log-Ratio transformation

The isometric log-ratio (ilr) transformation was defined by Egozcue et al. (2003) as below:

$$y = \text{ilr}(x) = (y_i, \ldots, y_{D-1}) \in \mathbb{R}^{D-1}.$$

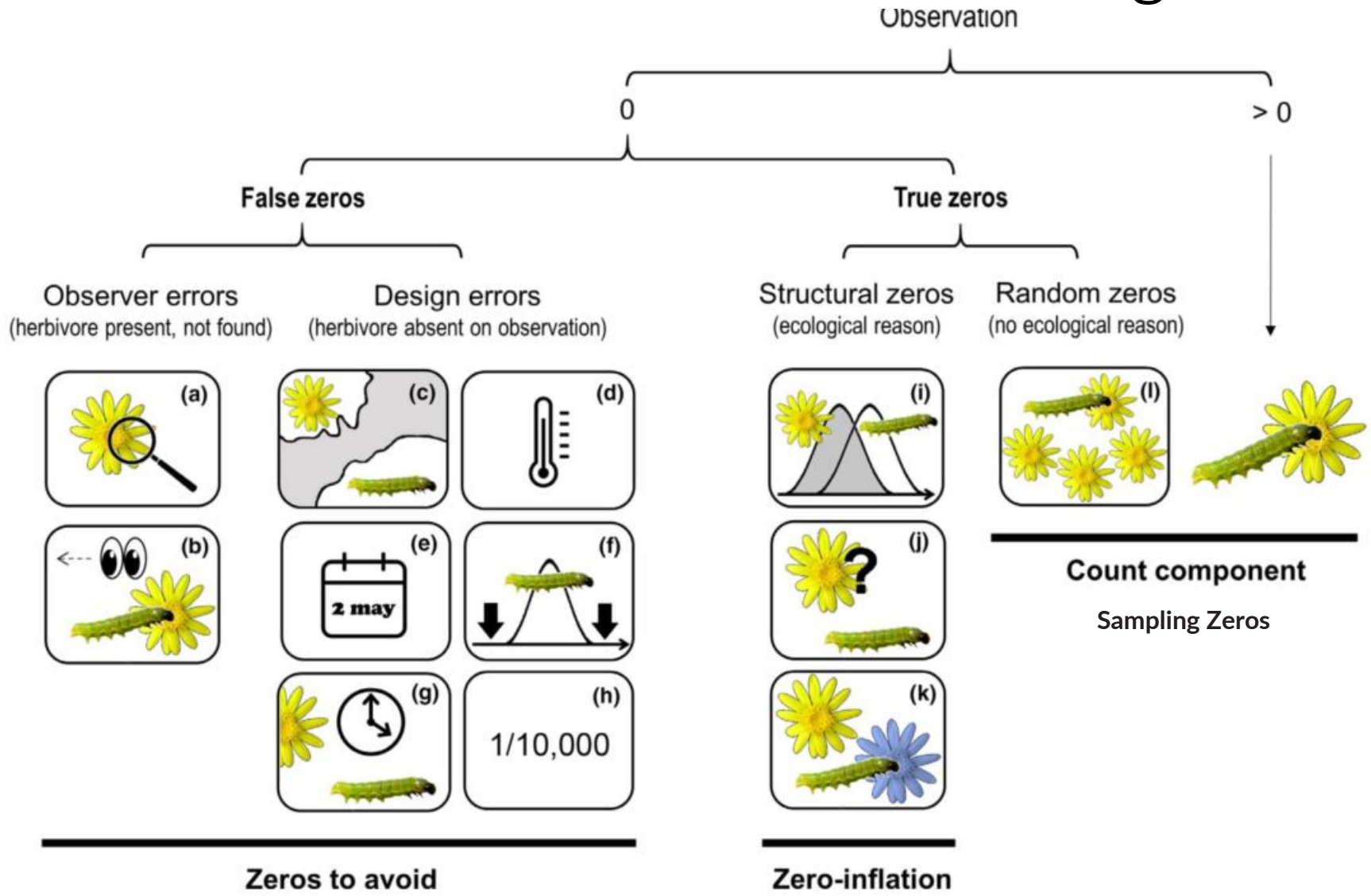
where, $y_i = \frac{1}{\sqrt{i(i+1)}} \ln \left[\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})i} \right]$.

1. Like the clr, the ilr transformation maps a composition in the D-part Aitchison simplex isometrically to a D-1 dimensional Euclidian vector.
2. The ilr transformation is the product of the clr and the transpose of a matrix which consists of elements. The elements are clr-transformed components of an **orthonormal basis**.
3. There are infinitely many possibilities to define such an orthonormal basis system. For this reason, ilr is considered as a class of coordinates.

Visualizaciones

https://rpubs.com/DavidLovell/SimplexToILR

Different sources of zeros that could emerge in count data



Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, *10*(7), 949-959

Zeros in Compositional Data Analysis

Because the logarithm of zero is not defined, log and log-ratio transformations require non-zero elements in the data matrix; as a consequence, compositional data analysis must be preceded by a treatment of the zeros.

The zeros are caused by **many** complicated reasons and currently, no simple general treatment strategy exists.

Rounded Zeros.

Most approaches treat them as a particular NMAR (Not Missing At Random) case, and deal with them by using both

- Nonparametric multiplicative replacement
- Other model/based replacement parametric methods such as EM, to replace them with a small nonzero value.

Zeros in Compositional Data Analysis

Sampling Zeros.

Sampling zeros are assumed to be a consequence of the **sampling process,** *not genuine zeros*.

To address the sampling zero problem, a Bayesian-multiplicative (BM) treatment combining with the Dirichlet distribution has been proposed.

It involves Bayesian inference on the zero values and a multiplicative modification of the non-zero values in the vector of counts. A zero value is replaced by its posterior Bayesian estimate. The multiplicative modification preserves the original ration between parts.

Structural Zeros.

Although currently there is no general method for dealing with the structural zero, it is clear that strategies for replacing it by a small value are not appropriate

Software

Statistical Tools for Compositional Data Analysis (CoDA)

- CoDaPack 3d (http://www.compositionaldata.com/codapack.php). Exploratory.
- **Compositions** (R package, Aitchison 1986 is available) . The package provides functions for the consistent analysis of compositional data.
- robCompositions
- **zCompositions.** Methods for imputing zeros in compositional count data sets
- CCREPE (Compositionality Corrected by REnormalizaion and PErmutation)
- SparCC
- SpiecEasi
- ANCOM (Analysis of Composition of Microbiomes)
- **ALDEx2** (R package). Uses standard Bayesian techniques to infer the posterior distribution of $(p_1, p_2, \ldots p_D)$ as the product of the multinomial likelihood with a Dirichlet $\left(\frac{1}{2}, \frac{1}{2} \cdots, \frac{1}{2}\right)$ prior.

Exploratory Compositional Data Analysis

10.3

Código en R

Comparisons of a Taxon of Interest Between Two Groups

8.2

Comparisons of Diversities Between Two Groups

In our Vdr−/− mouse study, one of the purposes is to test the difference of diversities between two groups (Vdr−/− and wild type mices) in fecal and cecal sites.

In Chap. 6, we calculated the Shannon diversity using the fecal samples.

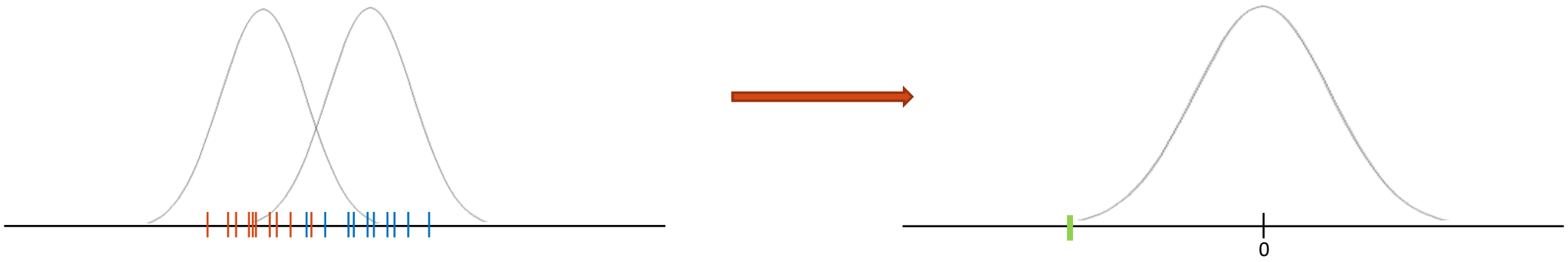
Here, to illustrate univariate community analysis, we shall compare the calculated Shannon diversity using various testing statistics.

Two-Sample t-Test

A two-sample t-test is used to test the means of two populations are equal. It is most commonly applied when the test statistic would follow a normal distribution. If the two groups have the same variance, the t statistic can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (\sim t_{n_1+n_2-2})$$

where, $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ is an estimator of the pooled standard deviation of the two samples.



Samples of two populations

Test Statistic

Welch's t Test

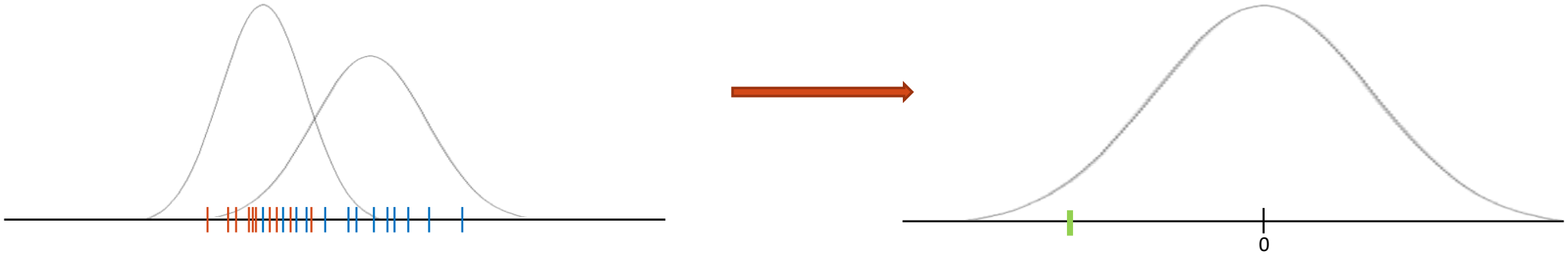
Welch's t-test or unequal variances t-test is adapted from t-test. The **Welch's t-test statistic** is given:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

where, $s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$; s_1^2 and s_2^2 are the unbiased estimator of the variance of samples 1 and 2 , respectively.

$$s_1^2 = \frac{\sum (X_i - \bar{X}_1)^2}{n_1 - 1}, \qquad s_2^2 = \frac{\sum (X_i - \bar{X}_2)^2}{n_2 - 1}$$

When the two samples have unequal variances and unequal sample sizes, Welch's t-test is considered as more reliable (Ruxton 2006). Thus, here we use Welch's t-test to our Vdr^{-1-} mouse data.



Samples of two populations

Test Statistic

Welch's t Test

The **Welch's t-test statistic** is given:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

has distribution t with ν degrees of freedom:

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Welch’s t Test

- Welch's t-test is more robust than Student's t-test
- maintains type I error rates close to nominal for unequal variances and for unequal sample sizes under normality.
- The power of Welch's t-test comes close to that of Student's t-test, even when the population variances are equal and sample sizes are balanced.
- Welch's t-test can be generalized to more than 2-samples.

It is not recommended to pre-test for equal variances and then choose between Student's t-test or Welch's t-test. Welch's t-test can be applied directly and without any substantial disadvantages to Student's t-test as noted above.

Wilcoxon Rank Sum Test

It is a **nonparametric** alternative to the two sample t-test that uses ranks of two independent sample data to test the **null hypothesis**: the two independent samples come from populations with the **same distribution** (that is, the two populations are identical).

Unlike the t-test, Wilcoxon rank sum test **does not require the assumption of normal distributions.**

Step 1. Assign ranks to all the observations, the smallest value gets a rank of 1. Where values are tied, assign the mean of the ranks involved in the tie.

Step 2. Sum the ranks for either one of the two samples. The sum of ranks in another sample can be determinated since the sum of all the ranks equals $N\,(N\,+\,1)/2$, where N is the total number of observations.

If the two testing populations have the same distribution, then the rank R has

$$\text{mean of } \mu_R = \frac{n_1(n_1+n_2+1)}{2}\text{ , and standard deviation of } \sigma_R = \sqrt{\frac{n_1\,n_2(n_1+n_2+1)}{12}}.$$

The Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions when the rank sum R is far from its mean.

Wilcoxon Rank Sum Test

The rank sum statistic becomes approximately normal as the two sample sizes increase. We can form the statistic by standardizing rank sum.

Step 3. Calculate the value of the z test statistic using the formula below:

$$z = \frac{R - \mu_R}{\sigma_R},$$

where

R sum of ranks of the sample with number n_1

n_1 the sample size for which the rank sum R is found (such as sample 1)

n_2 the other sample size (such as sample 2).

Wilcoxon Rank Sum Test

	A	B	C	D	E	F	G	H	I
1	**Wilcoxon Rank-Sum Test**								
2									
3	Original data			Ranks					
4								Control	Drug
5	Control	Drug		Control	Drug		count	12	12
6	11	34		4	22.5		rank sum	119.5	180.5
7	15	31		10	20.5				
8	9	35		2	24		α	0.05	
9	4	29		1	17.5		tails	2	
10	34	28		22.5	16		W	119.5	
11	17	12		11	5.5		W-crit	115	
12	18	18		12.5	12.5		sig	no	
13	14	30		8.5	19				
14	12	14		5.5	8.5				
15	13	22		7	14				
16	26	10		15	3				
17	31	29		20.5	17.5				
18	17	24.33333		119.5	180.5				

Código en R

Proportionality Analysis

10.5.2 AND 10.5.3

Proportionality metrics

Lovell et al. 2015 proposed the **proportionality measure** for analyzing relative abundances data and think that proportionality obeys all three principles: scaling invariance, subcompositional coherence, and permutation invariance.

Consider a matrix of D taxon count values measured across N samples subjected to condition, treatment status or time. The condition could be a binary or continuous event.

Let $\{A_i\}$ be the are log-ratio transformed vectors of the original sample vectors $\{X_i\}$.

The centered log-ratio transformation (clr), $\mathrm{clr}(X) = \left[\ln\left(\frac{x_1}{g_m(x)}\right), \ldots, \ln\left(\frac{x_i}{g_m(x)}\right), \ldots, \ln\left(\frac{x_D}{g_m(x)}\right)\right]$ is used by default to transform the sample vectors $\{X_i\}$.

Proportionality measure, ϕ

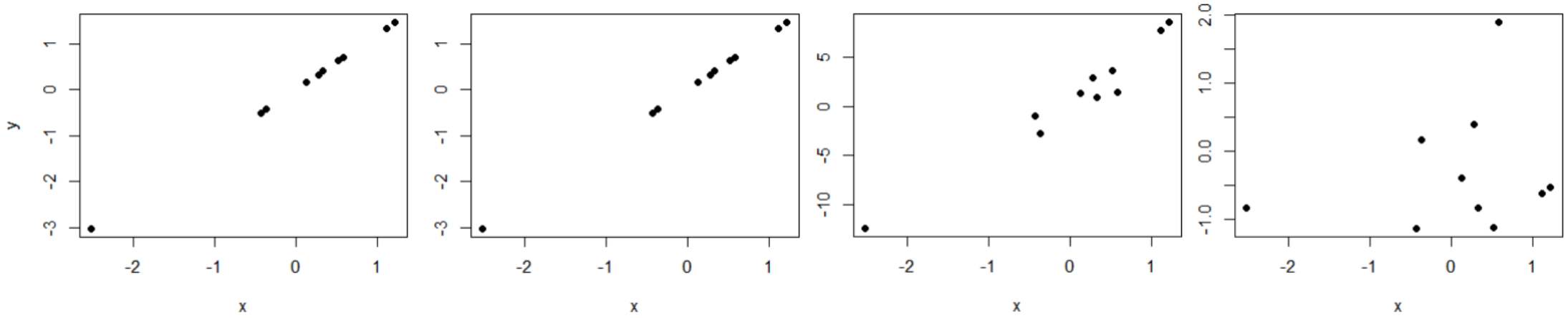
The proposed **proportionality measure** is statistic ϕ, which is used to describe the strength of proportionality between two variables: to assess the extent to which a pair of random variables (x, y) are proportional.

$$\phi\left(A_i, A_j\right) = \frac{\operatorname{var}\left(A_i - A_j\right)}{\operatorname{var}\left(A_i\right)}$$

$\phi \in [0, \infty)$ phit()

Lower values of ϕ indicates more proportionality (the closer ϕ is to zero, the stronger the proportionality)

ϕ = 0.0278 ϕ = 0.64 ϕ = 0. 679 ϕ = 1.87



Partial proportionality, ρ

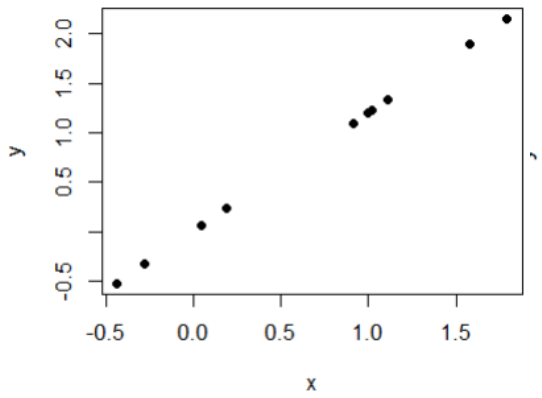
The **partial proportionality** ρ, adopted from partial correlations (Erb and Notredame 2016) has some advantages over ϕ statistic in that it is symmetric, has a limited range, can also detect reciprocality and allows for the definition of a partial coefficient.

$$\rho\left(A_i, A_j\right) = 1 - \frac{\operatorname{var}\left(A_i - A_j\right)}{\operatorname{var}\left(A_i\right) + \operatorname{var}\left(A_j\right)}$$

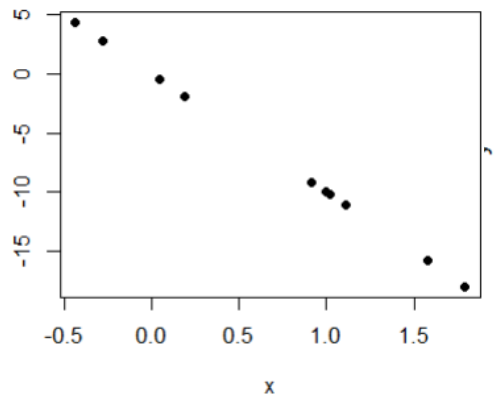
$\rho \in [-1,1]$ perb()

The greater $|\rho|$ values indicates more proportionality with negative ρ values indicating inverse proportionality.

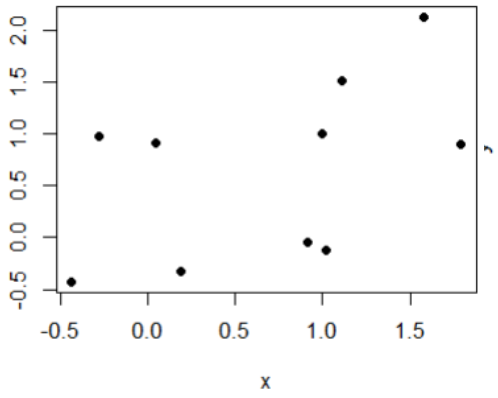
ρ = 0.983



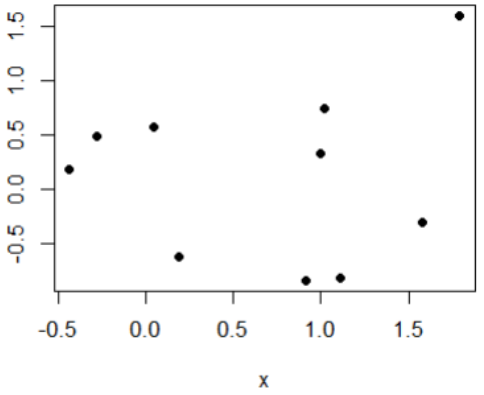
ρ = -0.198



ρ = 0.4843



ρ = 0.104



Symmetric variant of ϕ, ϕ_s

The naturally symmetric variant of ϕ, labeled as ϕ_s, defines the function phis() below:

$$\phi_s\left(A_i, A_j\right) = \frac{\operatorname{var}\left(A_i - A_j\right)}{\operatorname{var}\left(A_i + A_j\right)}$$

$\phi_s \in [0, \infty)$ phis()

Lower values of ϕ indicates more proportionality (the closer ϕ is to zero, the stronger the proportionality)

R code

Modeling Over-dispersed Microbiome Data

XIA, CAP. 11

Introduction

In the previous chapter we treated microbiome abundance data as compositional.

When doing so, we treated the “really discrete count data” as continuous.

However, count data is not purely relative—the count pair (1, 2) carries different information than counts of (1000, 2000) even though the relative amounts of the two components are the same.

Furthermore, simulation studies suggest that count models give more statistical power to detect differential expression than approximate normal models.

In next two chapters we treat the count as discrete variables.

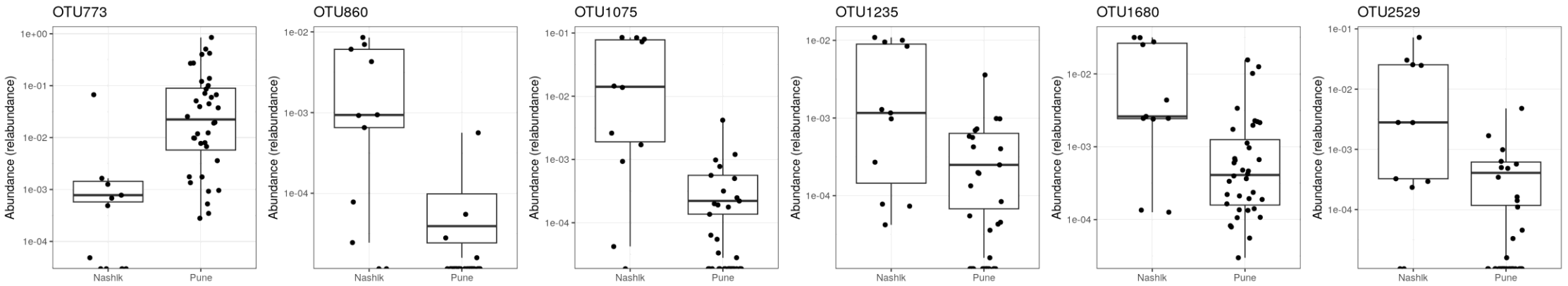
Count-Based Differential Abundance Analysis of Microbiome Data

In microbiome study:

after the composition of the microbiome is estimated at a given taxonomic level, researchers are often interested in identifying the taxa that show differential abundance between two or more groups.

In the **simplest case**, the aim is to compare taxa differential abundance between **two conditions**, e.g., treated versus untreated or mutant versus wild type.

A species/OTU is considered differentially abundant if its mean proportion is significantly different between two or more sample classes in the experimental design.



nature communications

Explore content | About the journal | Publish with us

nature > nature communications > articles > article

Article | Open Access | Published: 17 January 2022

Microbiome differential abundance methods produce different results across 38 datasets

Jacob T. Nearing, Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M. A. Jones, Robyn J. Wright, Akhilesh S. Dhanani, André M. Comeau & Morgan G. I. Langille

Nature Communications **13**, Article number: 342 (2022) | Cite this article

48k Accesses | **82** Citations | **530** Altmetric | Metrics

An Author Correction to this article was published on 03 February 2022

https://www.nature.com/articles/s41467-022-28034-z

In relation to models, methods and software

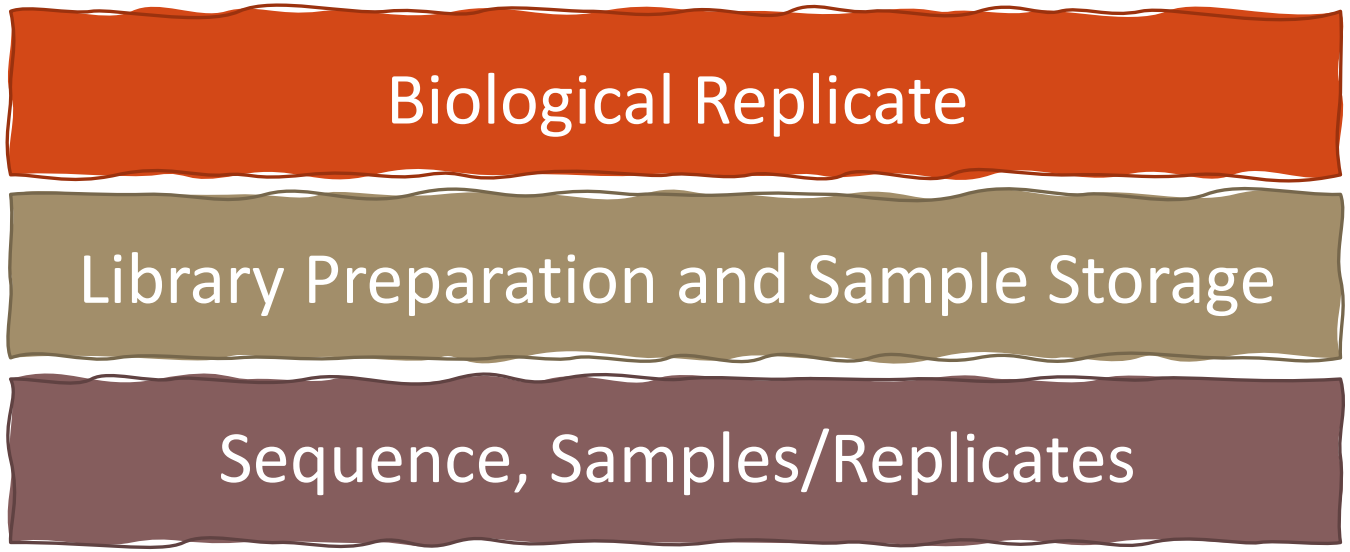
- DNA sequencing-based microbiome investigations not only have same questions to ask, but also use the same sequencing machines and represent the processed sequence data in the same manner as RNA-Seq analysis

- Statistical tools that were originally developed for differential analysis of RNAseq data, were suggested for directly use to identify differentially abundant OTUs .

- Thus, the statistical tools that were originally developed for differential analysis of RNAseq data, such as the packages **edgeR** and **DESeq, DESeq2** were suggested for directly use to identify differentially abundant OTUs (McMurdie and Holmes 2014).

Data Variations

The DNA- or RNA-based sequencing experiments has three steps that introduce variations to our observations



Data Variations

Two of the main principal sources of data variations are:

- **Biological,**
- **Technical.**

Biological Variations

Involves in the experimental units and it occurs within the same specimen or within an individual over time and may be influenced by **genetic** or **environmental factors**, as well as by whether the samples are **pooled or individual**.

Technical variation

Is produced by various ligations of adaptors and PCR amplifications involves during the generation of libraries of cDNA fragments.

It measures the variability in a sample subject, e.g., library preparation and sample storage.

There still exists the third source of variation at the bottom layer, beyond the library preparation effect (e.g. lane and flow cell effects).

Among these sources of variation, biological effect is far larger than other effects.

Introducción: Variables Aleatorias

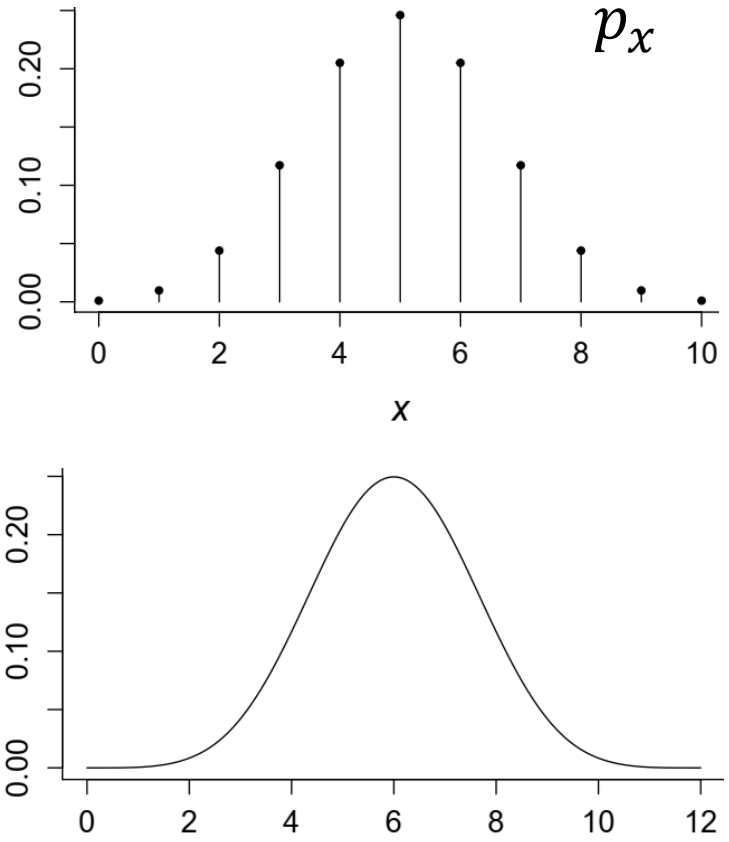
Tipos

- Discretas (Número de águilas en 10 tiros de moneda)
- Continuas
- Mixtas

Introducción: Variables Aleatorias

Tipos **fdp** **fd**

- Discretas $P(X = x)$
- Continuas $f(x)$



Introducción: Variables Aleatorias

Tipos	fdp	fd	$E(X)$
Discretas	$P(X = x)$		$\sum_{x \in X} xP(X = x)$
Continuas		$f(x)$	$\int_{-\infty}^{\infty} xf(x)dx$

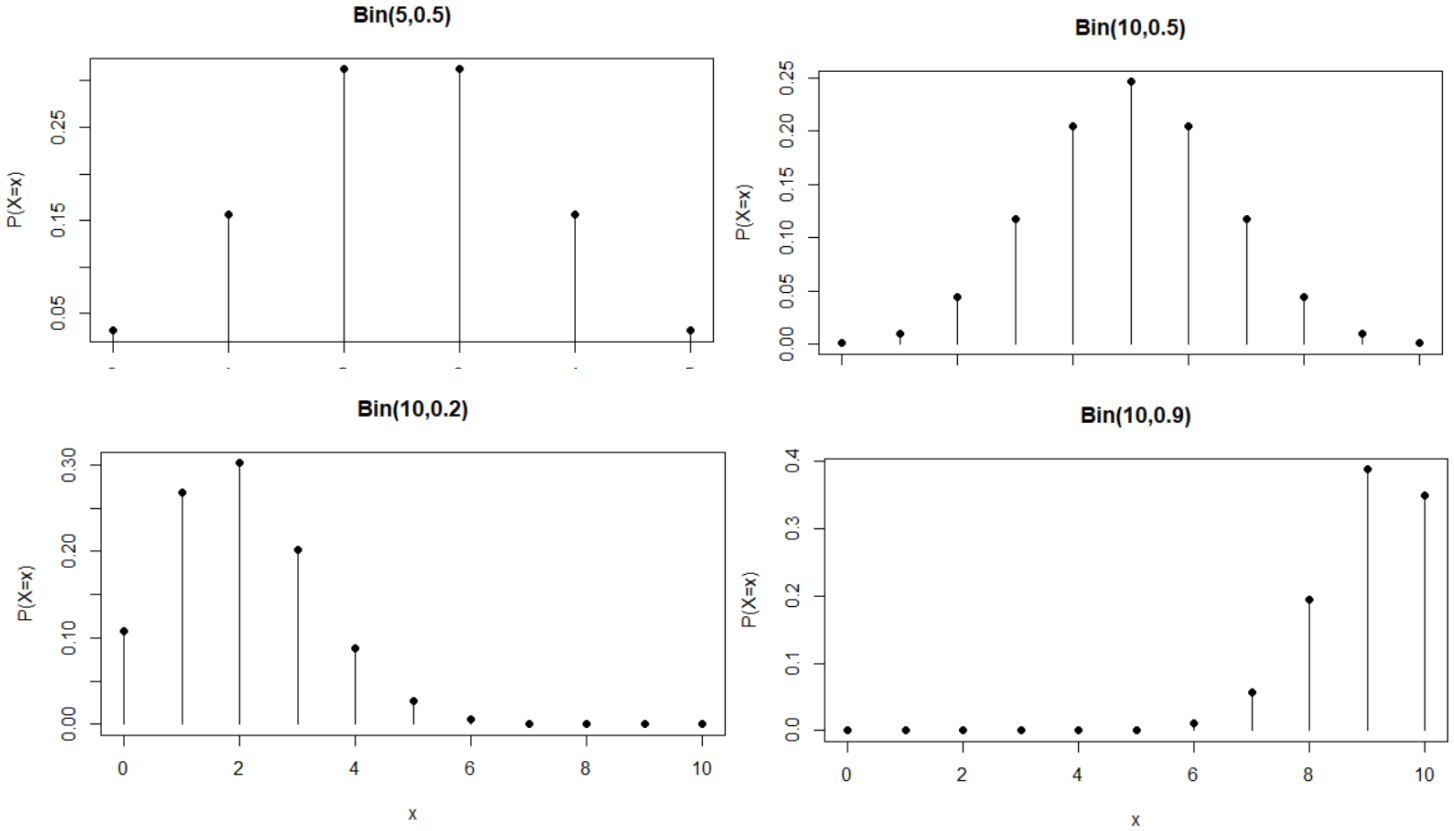
Introducción: Variables Aleatorias

Tipos	fdp	fd	$E(X)$	$F_X(x)$
Discretas	$P(X = x)$		$\sum_{x \in X} xP(X = x)$	$\sum_{i \leq x} P(X = i)$
Continuas		$f(x)$	$\int_{-\infty}^{\infty} xf(x)dx$	$\int_{-\infty}^{x} f(u)du$

Distribución Binomial

$$X \sim Bin(n,p), \quad n \in \mathbb{N}, \quad p \in (0,1)$$

$$p_x = P(X = x) = \binom{n}{p} p^x (1-p)^{n-x}, \quad x \in \{0,1,2,\dots,n\}$$



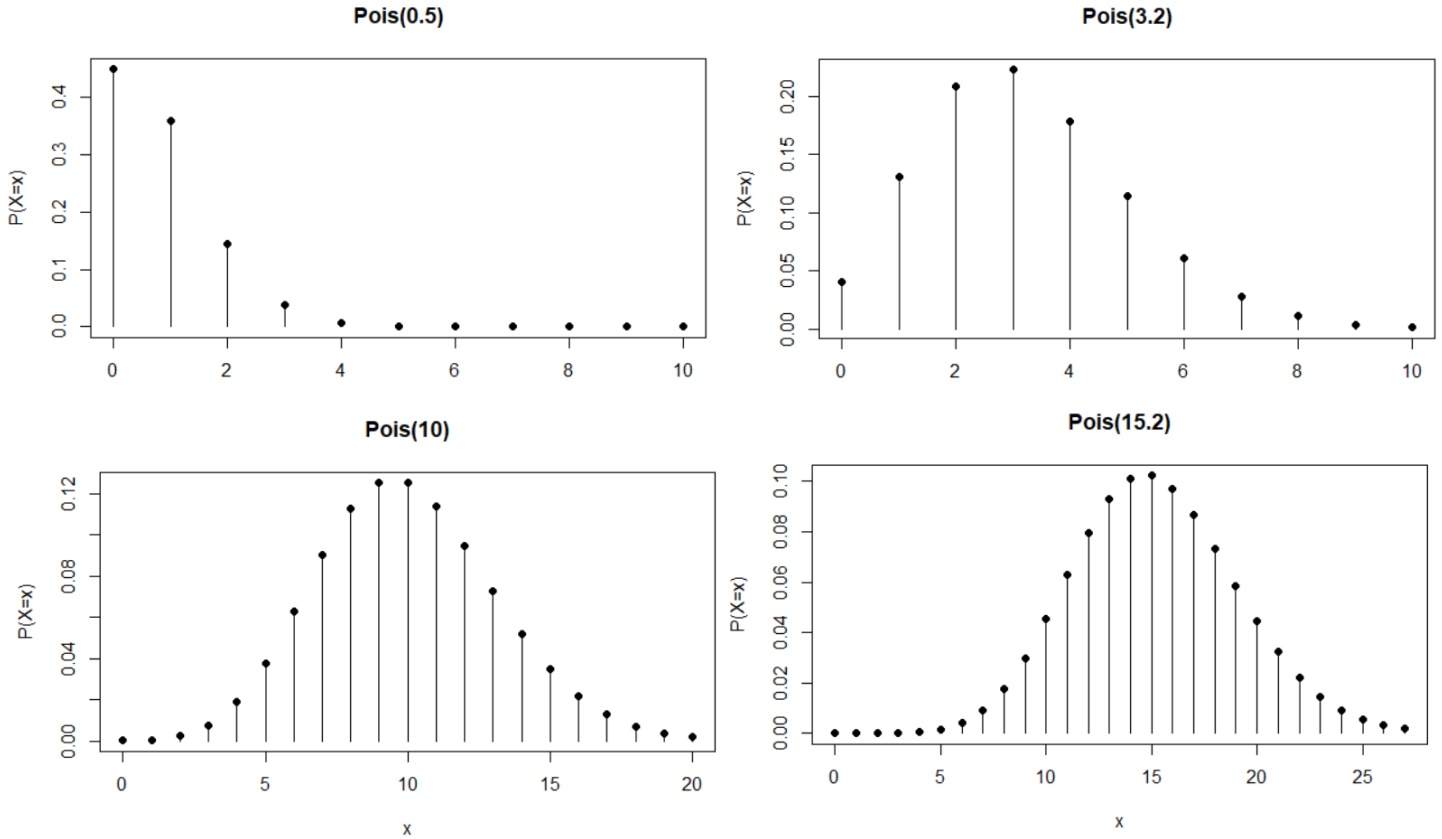
$$\mathrm{E}(\mathrm{X}) = \mathrm{np}$$
$$V(X) = np(1-p)$$

$$V(X) \leq E(X)$$

Distribución Poisson

$$X \sim Pois(\lambda), \quad \lambda > 0$$

$$p_x = P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}, \quad x \in \{0,1,2, \dots\}$$



$$\mathrm{E}(\mathrm{X}) = \lambda = V(\mathrm{X})$$

Se puede mostrar que cuando $\mathrm{X} \sim Bin(n, p)$ con n grande y p pequeña, su fdp se aproxima a la distribución $Pois(\lambda = np)$.