

POPULATION AND COMMUNITY STRUCTURE ANALYSES

“The search for order in nature is a central theme in ecological research. In community ecology, this often amounts to a search for non-random patterns in the species composition of naturally-occurring assemblages, and for the ecological processes responsible for those patterns.” (Poulin, 2005).

In parasitology, this equates to analyses of the prevalence, abundance and distribution of parasites within their hosts. Parasites are strictly dependent on their hosts for essential resources (food, shelter, transport, etc.); thus their geographical distribution, biodiversity and evolutionary history must, in some part, be influenced by that of their hosts. Host-parasite interactions may lead to the development of strict host-specificity and subsequently to tight host-parasite co-evolution.

Definitions (infra-population, infra-community, component-community, compound community)

Parasite population structure in single host species population

Parasite population structure in multiple populations of same host species

Parasite population structure in multiple populations of different host species

Types of Analyses

- I. Status of distribution (core-satellite)
- II. Dominance (ranks)
- III. Relationships between prevalence and abundance
- IV. Dispersion (aggregation)
- V. Species-abundance distributions
- VI. Diversity
 - A. Richness (number of species)
 - B. Evenness (relative abundance)
 - C. Taxonomic indices
- VII. Functional diversity
- VIII. Host specificity
 - Phylospecificity
 - Beta-specificity
 - Specificity matrix
 - Ecological niche
 - Host specificity
 - Host phylogenetic position
 - Inter-specific associations
 - Ecological fitting
- IX. Correlations
- X. Equilibrium or Non-equilibrium
- XI. Principal Components Analysis
- XII. Correspondence analyses
- XIII. Cluster Analysis
- XIV. Within-host parasite community interaction network
- XV. Statistical Tests

Requirements

Want to analyse variations in parasite population and community structures within and between host species. Analyses pertinent to:

- parasite biodiversity (species richness, relative abundance, species diversity)
- host specificity (sympatry/allopatry/parapatry, physiology/diet, castes/social behaviours, nest types, phylogeny/co-evolution/host-switching)
- biogeography (spatial and temporal distribution and abundance)

Was there variation in the type and numbers of parasites:

- within an individual host population (i.e. between hosts from one population)
e.g. one termite colony, one geographic location
- between populations of the same host species
e.g. colonies of same termite species, different geographic locations
- between populations of different host species
e.g. colonies of different termite species, different geographic locations

Have conducted studies to obtain measures of:

- species identity (usually using morphotypic/biological diagnostic characters)
- parasite occurrence (presence/absence)
 - calculate prevalence (= number infected/number hosts)
- parasite numbers (count per host, concentration per unit volume or weight, often on log scale)
 - calculate mean abundance
(= total number of parasites /total hosts (infected + uninfected))
 - calculate relative abundance
(ratio or percentage of each compared to total present)
 - calculate mean intensity
(= total number of parasites/total number of infected hosts)

Analyses within a single population

- within individuals
 - infra-population (= all parasite individuals of same species within a host individual)
 - occurrence (presence/absence)
 - numbers (abundance, intensity)
 - infra-community (= all parasite individuals of all species within a host individual)
 - mean prevalence
 - mean abundance
 - mean intensity
- between individuals
 - component-community (= all parasite individuals of all species within a host population)
 - traditional descriptors (species richness, mean intensity, mean abundance)
 - community similarity

Analyses between different populations

- between different populations of the same host species
 - guild (= all different parasite species within a particular host species that have a similar niche and that acquire resources in a similar way)
- between populations of different host species
 - compound parasite communities (= array of parasite species inhabiting an array of host species in a given area)

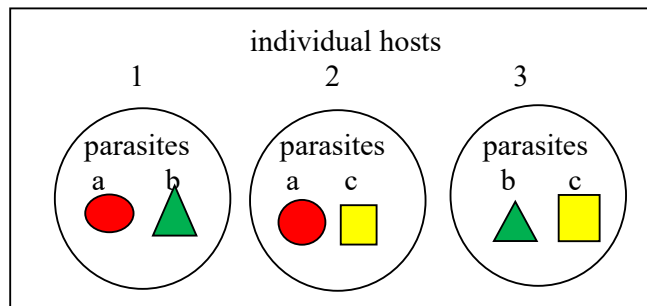
DEFINITIONS

Parasite distribution within a host population may be measured in terms of:

- parasite prevalence (number or percentage infected hosts)
[a, b and c given as presence/absence data]
- parasite abundance (number of parasites per host)
[a, b and c given as numeric values]

HOST POPULATION A

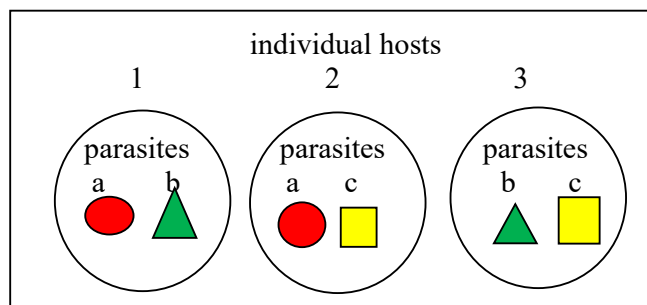
(discrete habitat, largely independent dynamics)(e.g. termite colony, fish school)



- PARASITE INFRA-POPULATION
(= all parasite individuals of same species within a host individual)
[e.g. 1a]
- PARASITE INFRA-COMMUNITY
(= all parasite individuals of all species within a host individual)
[e.g. (1a+1b)]
- PARASITE COMPONENT-COMMUNITY
(= all parasite individuals of all species within a host population)
[e.g. (1a+1b)+(2a+2c)+(3b+3c)]

HOST POPULATION B

(discrete habitat, largely independent dynamics)(e.g. termite colony, fish school)



METAPOPULATION (INTER-POPULATION) ANALYSES

(study of population of populations)[e.g. study of multiple termite colonies A, B...]

- multiple populations of same host species
- multiple populations of different host species (compound parasite communities)
(array of parasite species inhabiting an array of host species in a given area)

PARASITE POPULATION STRUCTURE IN SINGLE HOST SPECIES POPULATION

Let us consider a range of parasite species (a, b, c...)
found in a number of individual hosts of the same species (1,2,3...)

Their occurrence can be displayed in a host-parasite matrix:

- either qualitatively as present (+) or absent (-);
- or quantitatively as numbers per host (allowing mean intensity /abundance to be calculated)
[log transformation of abundance data converts it into normally distributed (Gaussian) data and reduces influence of dominant species (but not as much as conversion to presence/absence values)]
[numbers ranked onto log scale where: 1 = rare (1-2/host); 2 = few (around 10/host); 3 = medium (around 100/host); 4 = numerous (around 1,000/host); 5 = prolific (around 10,000/host), etc.]

		Individual host number				Sum	
		1	2	3	4	4	
Parasite species	a	4	4	3	0	3	prevalence = 3/4 mean intensity = 11/3 mean abundance = 11/4
	b	3	0	2	0	2	prevalence = 2/4 mean intensity = 5/2 mean abundance = 5/4
	c	2	1	0	0	2	prevalence = 2/4 mean intensity = 3/2 mean abundance = 3/4
	d	0	0	0	2	1	prevalence = 1/4 mean intensity = 2/1 mean abundance = 2/4
Sum	4	3	2	2	1		
		high species richness	medium species richness	medium species richness	low species richness		

Infra-population (= all parasite individuals of same species within a host individual)
[e.g. 1a]

Infra-community (= all parasite individuals of all species within a host individual)
[e.g. (1a+1b+1c+1d)]

Component community (= all parasite individuals of all species within a host population)
[e.g. (1a+1b+1c+1d)+(2a+2b+2c+2d)+(3a+3b+3c+3d)+(4a+4b+4c+4d)]

This matrix depicts

- parasite species richness for each host,
- parasite prevalence (ubiquitous, prevalent, rare),
- variation in numbers (light/heavy intensity, low/high abundance), and
- type of distribution (uniform, random, aggregated, or patchy)

It also indicates preliminary community structure:

- all individuals have parasites,
- ranging from 1-3 species,
- most with 2 (not necessarily the same 2)
- total of 4 parasites detected,
- ranging in prevalence from 25-75% (a>b=c>d);
- ranging in intensity from 1.5-3.6 (i.e. from ~50-600) (a>b>d>c),
- ranging in abundance from 0.5-2.75 (i.e. from ~1-80) (a>b>c>d)

Assess status of distribution

Assess dominance

Assess distribution

Is abundance of each parasite species normally distributed (or skewed)?

- one sample t-test for normally distributed
(test difference from mean)
- Wilcoxon's signed rank test for skewed distribution
(test difference between medians)

Compare distribution of pairs of parasite species (are they independent?)

- two sample t-test for normally distributed
(test equality of means)
- Wilcoxon's rank sum test (= Mann-Whitney U test) for skewed distribution
(test equality of medians)

Compare distribution of more than two parasite species (are they independent?)

- analysis of variance for normally distributed
(test equality of means)
- Kruskal-Wallis test for skewed distribution
(test equality of medians)

Analyse data for correlations (Spearman's correlation coefficient) (are they dependent?)

- host castes
- host sexes
- host sizes
- host ages

PARASITE POPULATION STRUCTURE IN MULTIPLE POPULATIONS OF SAME HOST SPECIES (HETEROGENEITY)

Now let us consider a range of parasite species (a, b, c...) found in different populations of the same host species, e.g. from several locations or colonies (1,2,3...)

Their occurrence can be displayed in a host-parasite matrix:

- either qualitatively as present (+) or absent (-);
- or quantitatively as mean numbers (intensity/abundance) per host

		Different populations (colonies)				Sum	
		1	2	3	4		
Parasite species	a	4	4	3	0	3	prevalence = 3/4 mean intensity = 11/3 mean abundance = 11/4 prevalence = 2/4 mean intensity = 5/2 mean abundance = 5/4 prevalence = 2/4 mean intensity = 3/2 mean abundance = 3/4 prevalence = 1/4 mean intensity = 2/1 mean abundance = 2/4
	b	3	0	2	0	2	
	c	2	1	0	0	2	
	d	0	0	0	2	1	
Sum	4	3	2	2	1		

high species richness
medium species richness
medium species richness
low species richness

This matrix depicts

- parasite species richness for each population,
- parasite prevalence for each population (ubiquitous, prevalent, rare),
- variation in numbers between populations (light/heavy intensity, low/high abundance), and
- type of distribution (uniform, random, aggregated, or patchy)

It also indicates preliminary community structure:

- all populations have parasites,
- ranging from 1-3 species,
- most with 2 (not necessarily the same 2)
- total of 4 parasites detected,
- ranging in prevalence from 25-75% (a>b=c>d);
- ranging in intensity from 1.5-3.6 (i.e. from ~50-600) (a>b>d>c),
- ranging in abundance from 0.5-2.75 (i.e. from ~1-80) (a>b>c>d)

Confirm status of distribution (core-satellite hypothesis)

are they similar?

Confirm dominance:

are they similar?

Compare distributions

can data be pooled?

PARASITE COMMUNITY STRUCTURE IN MULTIPLE POPULATIONS OF DIFFERENT HOST SPECIES

COMPOUND PARASITE COMMUNITIES

(= array of parasite species inhabiting an array of host species in a given area)

Now let us consider a range of parasite species (a, b, c...) found in a range of host species (A, B, C...)

Their occurrence can be displayed in a host-parasite matrix:

- either qualitatively as present (+) or absent (-);
- or quantitatively as mean numbers (intensity/abundance) per host

		Host species				Sum	
		A	B	C	D	4	
Parasite species	a	4	4	3	0	3	broad host specificity prevalence = 3/4 mean intensity = 11/3 mean abundance = 11/4 medium host specificity prevalence = 2/4 mean intensity = 5/2 mean abundance = 5/4 medium host specificity prevalence = 2/4 mean intensity = 3/2 mean abundance = 3/4 narrow host specificity prevalence = 1/4 mean intensity = 2/1 mean abundance = 2/4
	b	3	0	2	0	2	
	c	2	1	0	0	2	
	d	0	0	0	2	1	
Sum	4	3	2	2	1		
		high species diversity	medium species diversity	medium species diversity	low species diversity		

This matrix depicts

- parasite biodiversity (species richness, relative abundance) for each host species,
 - total of 4 parasites detected,
 - ranging in prevalence from 25-75% (a>b=c>d)
 - ranging in intensity from 1.5-3.6 (i.e. from ~50-600) (a>b>d>c)
 - ranging in abundance from 0.5-2.75 (i.e. from ~1-80) (a>b>c>d)
- type of distribution (uniform, random, aggregated, or patchy)
- all host species have parasites,
- ranging from 1-3 species (host species specificity)
- most with 2 (not necessarily the same 2)

assess host specificity (are they similar?)

- random distribution between hosts
- associative distribution (pairs of parasites)
- nested distribution (togetherness)

assess community structure (do they cluster?)

- evenness
- clusters

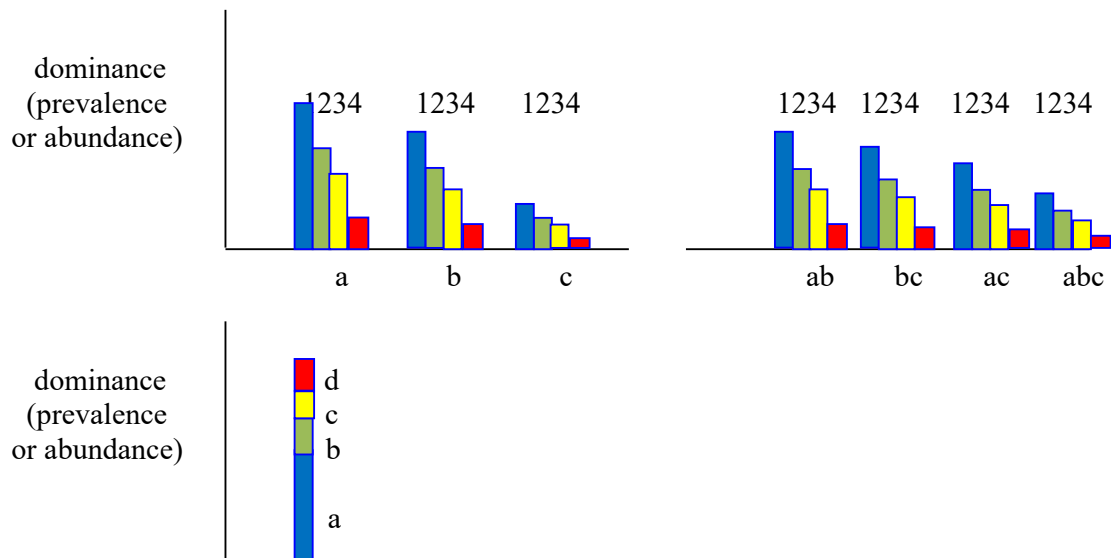
ANALYSES

I. Status of distribution (core-secondary-satellite hypothesis)

- core (central) species = regionally common (> 66% prevalence), locally abundant, consistent occurrence
- secondary species = regionally common (33-66% prevalence), moderately abundant
- satellite species = regionally uncommon (<33% prevalence), locally rare, sporadic occurrence

II. Dominance (ranks)

- Which parasite species is most prevalent?
bar graph prevalence of each species, and combinations of species
- Which parasite species is most abundant?
bar graph numeric values for each species, and combinations of species?
- Which species is most frequent?
bar graph rank of each species as percentage of total number of parasites



The relative prevalence or abundance of each species can be expressed as a ratio (totalling 1) or as a percentage (totalling 100%) of the total found. These values can be statistically analysed by arcsine-transformation of the ratio data and then conducting a two-way multivariate analysis of variance (MANOVA).

III. Relationships between prevalence and abundance.

The parasite distribution within a host population is generally measured by parasite abundance (number of parasite individuals per host) or prevalence (percentage of infected hosts).

- Transforming data changes their properties to be more amenable to statistical analyses. In geometric terms, it shifts the relative positions of points in multivariate space in order to reveal obscured patterns, impose a desired pattern, or hide an undesired pattern. In exploratory statistical analyses, revealing and obscuring patterns are opposite sides of the same coin. Data transformations are generally necessary in order to obtain interpretable results, so knowing how they influence results is critical for interpreting and evaluating multivariate analyses.

- Converting numerical abundances to presence/absence data. This transformation makes all species equally important in characterizing a sample, regardless of their abundance.

$$\text{If } x_{ai} > 0, \text{ then } x^*_{ai} = 1, \text{ else } x^*_{ai} = 0$$

where x_{ai} = abundance of species i in sample a

- Logarithmic transformation. This transformation converts log-normal abundance data into normally distributed (Gaussian) data. It reduces the influence of dominant species, but not as much as conversion to presence/absence data.

$$x^*_{ai} = \log_b (x_{ai} + k)$$

where b = base of logarithm (typically 2, 10 or e)

k = a constant (necessary to prevent undefined log values when $x_{ai} = 0$)

The need for a constant is not a desirable property because it has a disproportionately large influence on the contribution of rare species (adding 1 to 1 is 100% change whereas adding 1 to 10,000 is a 0.01% change).

A log-based transformation that gets around the need to alter abundance values is:

$$\text{if } x_{ai} = 0, x^*_{ai} = 0; \text{ else } x^*_{ai} = \log_b (x^*_{ai}) + 1$$

This function has the advantage of rescaling that makes the classic log transformation so useful, but it maintains the correct relationship between x_{ai} values when they are subtracted from one another (as they are in most measures of similarity) regardless of their rarity.

- Root transformation. Like all transformations, the root transform decreases the influence of dominant taxa. A 'double square root' (i.e. $n=4$) transform is quite common in ecology.

$$x^*_{ai} = \sqrt[n]{x_{ai}}$$

An arcsine-squareroot transformation increases the importance of low abundance taxa and decreases the importance of high abundance taxa. Changes $(x_{ai})/(\sum x_{ai})$ values ranging from 0 to (x^*_{ai}) values that range from 0 to 1.571

$$x^*_{ai} = \arcsin \sqrt{[(x_{ai})/(\sum x_{ai})]}$$

- Standardizing data. Standardization weights samples or taxa so that they contribute to a statistical analysis more equally; i.e. without standardizing data, large samples or abundant taxa can overwhelm a subtle pattern.

- Standardization to total. When applied to a sample, each taxon is represented by its proportion and every sample sums to 1; i.e. what is referred to as relative abundance data (often multiplied by 100 to obtain percentages). If applied to each taxon, it emphasizes rare taxa and diminishes common taxa.

$$y_{ai} = (x_{ai})/\Sigma(x_{ai})$$

- Standardization to maximum. When applied to a sample, the most common taxon is given a value of 1 and all the other taxa are scaled to it. The largest taxon in every collection is

then equal. If applied to each taxon, it equalizes the influence of rare and common taxa (maximum abundance of every taxon equals 1).

$$y_{ai} = (x_{ai})/\max(x_{ai})$$

- Standardization to vector length. If a site is considered a vector (i.e. the point representing it in a space defined by species axes is the head of a vector starting at the origin), this standardization gives the vector a length of 1 (all sample points lie on a spheroid with radius equal to 1).

$$y_{ai} = (x_{ai})/\sqrt{\sum(x_{ai}^2)}$$

- z-transformation. This transform is a common calculation in classical statistics (subtract the mean and divide by the standard deviation).

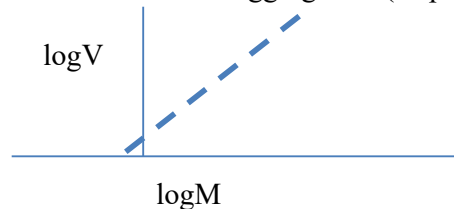
$$z_{ai} = (x_{ai} - \bar{x}_a) / \sigma_a$$

The result is that the mean value of every sample is 0 (it is centred) and the standard deviation of its taxon abundances is 1 (its unit length). This transformation is implicit when Principal Components Analysis (PCA) is applied to a correlation matrix. [Note the similarity in the form of the z-transform to the vector length equation – in effect, a z-transform scales a centred data series to be a vector of length 1.

- Two-way transformation. It is a common procedure to standardize both taxa and samples. A common approach is to standardize taxa to their minimum and samples to their totals.

- Taylor's power law is often found as the relationship between mean abundance (M) and its variance (V): $\log(V) = b \cdot \log(M) + \log(a)$

where a = constant (intercept),
b = index of aggregation (slope)



From basic epidemiological models (Anderson & May, 1985), the prevalence of infection can be linked to the mean abundance of parasites at any time during infection dynamics according to:

$$P = 1 - [1 + (M/k)]^{-k}$$

where P = prevalence;

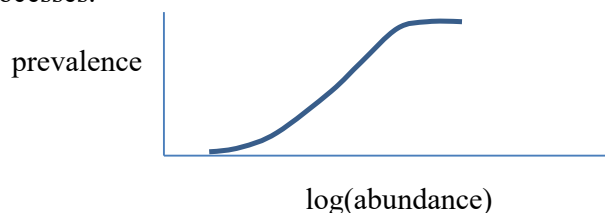
k = aggregation parameter of the negative binomial distribution.

The parameter k is related to the parameters a and b of Taylor's power law by:

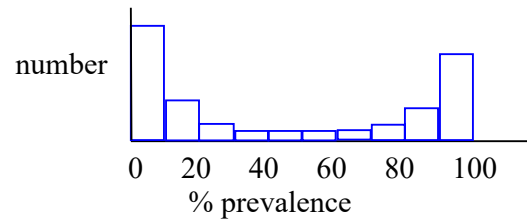
$$1/k = aM^{(b-2)} - (1/M)$$

Values for these parameters have been calculated for several parasite assemblages and they generally occur in the following ranges: $b = 1.7 \pm 0.05$; $a = 0.6 \pm 0.05$

- Correlation between prevalence and abundance. A positive relationship is often observed between the prevalence and abundance of parasites, possibly as the result of epidemiological processes.



- Frequency distribution. The distribution of parasite prevalence shows a bimodal pattern for many parasites, possibly due to demographic explanations. Differences between observed and expected patterns can be tested using Monte-Carlo simulations in epidemiological models.



The core-satellite hypothesis was incorporated into meta-population dynamic models to explain the positive relationship between local abundance and prevalence of infection as it predicts a bimodal distribution of organisms in their environment (i.e. most species are present in either most patches or only in a small fraction of patches). The hypothesis is not based on competition but on the ability of species to recolonize empty patches after extinction (i.e. rescue efforts). Parasites very often show an overdispersed distribution described by a negative binomial distribution (i.e. most hosts have a few or no parasites, a few hosts have many parasites).

IV. Dispersion (aggregation measures)

- Index of Dispersion (DI) (=VMR) (also known as coefficient of dispersion, relative variance, or variance-to-mean ratio (VMR)) is a normalized measure of the dispersion of a probability distribution (used in statistics and probability theory). Studies use the variance-to-mean ratio (statistical 'D') when data do not follow a theoretical Poisson distribution. DI is a measure used to quantify whether a set of observed occurrences are clustered or dispersed compared to a standard statistical model. Under a random distribution of points, DI is expected to equal 1. DI is calculated as the ratio of the variance (σ^2) to the mean (μ),

$$D = \sigma^2 / \mu$$

[aggregated when $d > 1.96$; regular when $d < -1.96$; random when $d < 1.96$]

level of aggregation associated with:

- heterogeneity of host behaviour
 - spatial aggregation of infective stages
 - variation in host susceptibility
- Index of Cluster Size (ICS) (also known as the Index of Clumping (IC) is a direct function of the Index of Dispersion. Under a random distribution of points, ICS is expected to equal 0. Positive values indicate a clumped distribution; negative values a regular distribution.

$$ICS = (s^2 / \bar{x}) - 1 \quad \{ = D - 1 \}$$

- Green's Dispersion Index (GI) is a modification of the Index of Cluster Size that is independent of n. It varies between 0 for random distributions and 1 for maximally clumped distributions.

$$GI = [(s^2 / \bar{x}) - 1] / (n - 1) \quad \{ = ICS / (n - 1) \}$$

- Index of Cluster Frequency (ICF) is a measure of aggregation and is equal to k of the negative binomial distribution. ICF is proportional to the quadrat area and is related to the Index of Cluster Size.

$$ICF = [\bar{x} / (s^2 / \bar{x}) - 1] \quad \{ = \bar{x} / ICS \}$$

- Index of Mean Crowding (IMC) is the average number of other points contained in the quadrat that contains a randomly chosen point. It is related to the Index of Cluster Size.

$$IMC = \bar{x} + (s^2 / \bar{x}) - 1 \quad \{ = \bar{x} + ICS \}$$

- Index of Patchiness (IP) is related to the Index of Cluster Frequency and the Index of Mean Crowding, and is similar to Morisita's Index. It is a measure of pattern intensity that is unaffected by thinning (the random removal of points).

$$IP = [\bar{x} + (s^2 / \bar{x}) - 1] / \bar{x} \quad \{ = IMC / \bar{x} \} \quad \{ = 1 + (1 / ICF) \}$$

- Morisita's Index (I_M) is related to the Index of Patchiness. It is the scaled probability that two points chosen at random from the whole population are in the same quadrat. The higher the value, the more clumped the distribution [\bar{x} = mean]

$$I_M = [n \sum x(x-1)] / [n \bar{x}(n \bar{x} - 1)] \quad \{ = n \bar{x} IP / (n \bar{x} - 1) \}$$

- Aggregation Model of Co-existence In metapopulation studies, it is important to examine the nature of the interactions between different species exploiting the same patchy resource. Communities are classified as non-interactive or interactive depending on whether interactions

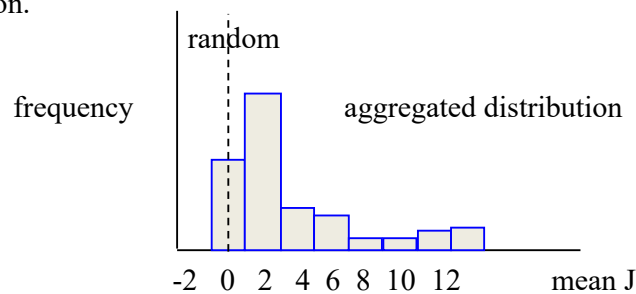
are important or not among the residents of a local habitat. Interactive communities are of two types: niche heterogeneity types (where co-existence is favoured when species differ in one or more important dimensions of their niche) or spatiotemporal heterogeneity types (where species differ in spatial or temporal occurrence). Species can co-exist by reducing the overall intensity of competition via aggregated utilization of fragmented resources, formalized as the 'aggregation model of co-existence'. This model postulates that co-existence is facilitated when the distribution of species leads to the reduction of interspecific aggregation relative to intraspecific aggregation. However, the model assumes saturation of ecological communities with species, which means that there is saturation of local species richness (in individual hosts) independent of the size of the regional pool of species (component parasite species). [Morand & Simkova, 2005: p.306 in Rohde (2005)]

- Intraspecific aggregation (J) can be measured as the proportionate increase in the number of the same parasite species experienced by a random host relative to a random distribution:

$$J = [\Sigma (n(n-1)/m) - m]/m = [(V/m) - 1] / m$$

where n = number in host of parasite species,
 m = mean numbers, and
 V = variance

When $J = 0$, individuals are randomly distributed; $J = 0.5$, indicates a 50% increase in the number of parasite individuals expected in a given host compared to the random distribution.

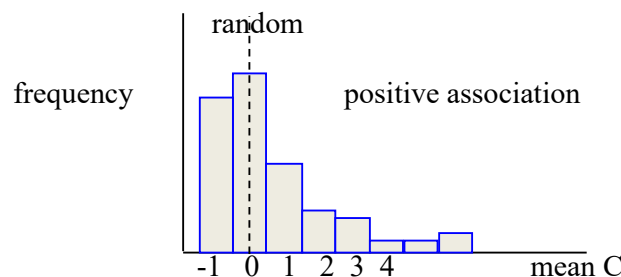


- Interspecific aggregation (C) can be measured as the proportionate increase in the number of different parasite species relative to a random association.

$$C = [\Sigma (n_1.n_2)/(m_1.P) / m_2 = \text{Cov}_{1,2} / (m_1.m_2)$$

where n_1 and n_2 = number of species 1 and species 2 in host;
 m_1 and m_2 = mean numbers of species 1 and species 2 per host;
 P = number of hosts; and
 $\text{Cov}_{1,2}$ = covariance between a pair of parasite species

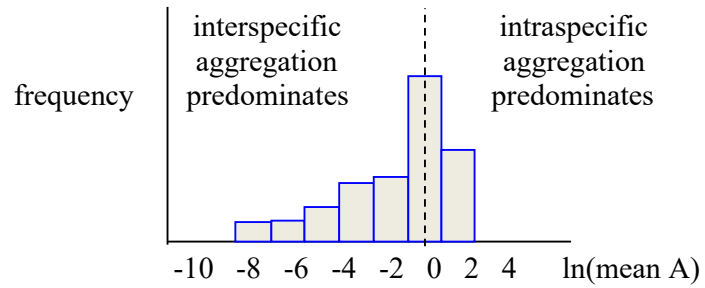
The two parasite species are positively associated when $C > 0$, and negatively associated when $C < 0$



- Comparative index (A) quantifies the reduction in competition caused by intraspecific aggregation; i.e. the relative strength of intraspecific aggregation versus interspecific aggregation.

$$A_{1,2} = [(J_1 + 1)(J_2 + 1)] / (C_{1,2} + 1)^2$$

When $A > 1$, intraspecific aggregation is stronger than interspecific aggregation.



V. Species-abundance distributions

Nearly all diversity and evenness indices are based on the relative abundance of species, thus on estimates of p_i in which:

$$p_i = N_i / N$$

where N_i = the abundance of the i -th species in the sample and

$$N = \sum_{i=1}^S N_i$$

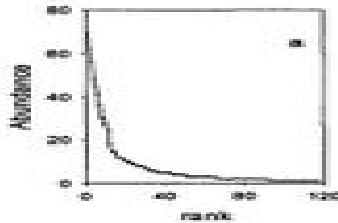
where S = the total number of species in the sample.

If one records the abundance of different species in a sample, it is invariably found that some species are rare, whereas others are more abundant. This feature of ecological communities is found independent of the taxonomic group or the area investigated. An important goal of ecology is to describe these consistent patterns in different communities, and explain them in terms of interactions with the biotic and abiotic environment. A community can be defined as the total set of organisms in an ecological unit (biotope), but the definition should always be qualified by stating limits or boundaries: e.g.

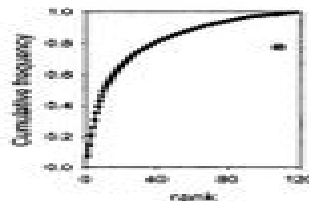
- spatial boundaries of area or volume
- sensitivity/specificity/limits of detection of sampling methods
- time limits spanning observations
- set of species (taxocene) treated as constituting/representing the community

Species-abundance distribution data may be presented in different ways.

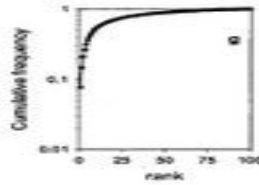
- The rank/abundance plot (ranked species abundance curve) is one of the best known and most informative method. Species are ranked in sequence from most to least abundant along the horizontal (x) axis. Abundances are displayed on the vertical (y) axis, either as numeric values, but more typically subject to logarithmic (\log_{10} or \ln) transformation so that species whose abundances span several orders of magnitude can be accommodated on the same graph. Proportional or percentage abundances are often used.



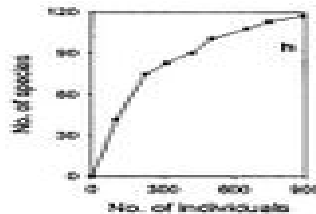
- The k-dominance plot shows the cumulative percentage (the percentage of the k -th most dominant plus all more dominant species) in relation to species (k) rank or log species (k) rank.



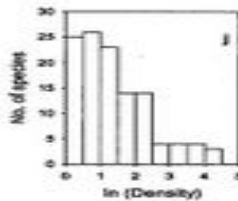
- The Lorenzen curve is based on the k-dominance plot but the species rank k is transformed to $(k/S) \times 100$ to facilitate comparison between communities with different numbers of species.



- The collector's curve addresses a different problem. When one increases the sampling effort, and thus the number of the animals N caught, new species will appear in the collection. A collector's curve expresses the number of species as a function of the number of specimens caught. As more specimens are caught, a collector's curve can reach an asymptotic value but they often do not due to the vague boundaries of ecological communities: as sampling effort increases, also the number of different patches increases.



- The species-abundance distribution plots the number of species that are represented by $r = 0, 1, 2, \dots$ individuals against the abundance r (e.g. 25 species with 1 individual, 20 species with 2 individuals, etc.). This can only be drawn if the collection is large and contains many species. More often than not the species are grouped in logarithmic density classes ($\log_e (= \ln)$ favoured over \log_{10} or \log_2).



- Species-Abundance models. A diverse range of models has also been developed to describe species abundance data. Two kinds of models have been devised: historical 'resource apportioning' models, which make assumptions about the division of some limiting resource; and contemporary 'statistical' models, which make assumptions about probability distributions of the numbers in the several species within the community.
 - Niche preemption model (geometric series ranked abundance list). This resource apportioning model assumes that a species preempts a fraction k of a limiting resource, a second species the same fraction k of the remainder and so on. If the abundances are proportional to their share of the resource, the ranked abundances list is given by geometric series:

$$k, k(1-k), \dots, k(1-k)^{(S-2)}, k(1-k)^{(S-1)}$$
 where S = the number of the species in the community.

The geometric model gives a straight line on a plot of log abundance against rank (species sequence). It is not very often found in nature, only in early successional stages or in species poor environments.

- Negative exponential distribution (broken-stick model). This statistical model is given by the probability density function: $\psi(y) = Se^{-Sy}$. This function can be arrived at via the 'broken-stick' model where a limiting resource is compared with a stick, broken in S parts at $S-1$ randomly located points. The length of the parts is taken as representative for the density of the S species subdividing the limiting resource. If the species are ranked according to abundance, the expected abundance of species i , N_i is given by:

$$E(N_i) = (1/S) \sum_{x=1}^S (1/x)$$

The negative exponential distribution is not often found in nature. It describes a too even distribution of individuals over species to be a good representation of natural communities.

- Log-series distribution (Fisher's logarithmic series) describes the relationship between the number of species and the number of individuals in those species. The expected number of species with r individuals, E_r , is given by:

$$E_r = \alpha (X^r / r)$$

$$r = 1, 2, 3, \dots$$

α (>0) = parameter independent of sample size, for which X ($0 < X < 1$) is representative. The parameters α and X can be estimated by maximum likelihood but are conveniently estimated as solutions of:

$$S = -\alpha \ln(1-X) \quad \text{and}$$

$$N = \alpha X / (1-X)$$

- Log-normal distribution may be expected when a large number of independent environmental factors act multiplicatively on the abundances of species. When species-abundance distribution is log-normal, the probability density function of y (the abundance of species) is given by:

$$\psi(y) = 1 / [y\sqrt{2\pi V_z}] \cdot \exp [-(\ln y - \mu_z)^2 / 2V_z]$$

The mean and variance of y are:

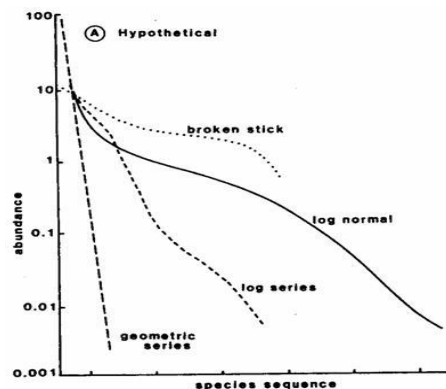
$$\mu_y = \exp [\mu_z + (V_z/2)]$$

$$V_y = (\exp V_z - 1) \cdot \exp (2\mu_z + V_z)$$

where μ_z and V_z are the mean and variance of $z = \ln y$

In a limited sampling, a certain number of species will be unrepresented so the log-normal distribution will be truncated (certain species are hidden behind a veil line).

The four main species- abundance models



VI. Diversity.

The diversity of species in a particular area depends not only on the number of species found, but also in their numbers. Diversity is a function of both the number of species in a community (species richness) and their relative abundances (species evenness). [A game reserve with one antelope and one zebra when compared with another with one antelope and ten zebra, therefore, have the same species richness but different species evenness.] Species diversity is a measure of community complexity. Larger numbers of species and more even abundances of species leads to higher species diversity.

Numerous methods are available to compare communities, either based on univariate or multivariate indices. These indices are based on quantitative data (the percentage of each parasite species in the community). Using presence-absence data, rather than frequency data, corresponds to a loss of information but has the advantage of not conferring artificial weight on frequent species. Classification methods (such as the unweighted pair-group method of arithmetic averages (UPGMA) based on the Jaccard index) or correspondence analysis are tools commonly used to analyse such data. These phenetic analyses are based on the distance matrix and not directly on characters. Conversely, in phylogenetic methods, each character is polarised (in this case the species presence or absence of a species) and is used directly in the analysis. The indices measure diversity within a sample (alpha diversity), within a region (gamma diversity) or along a physical gradient (beta diversity); viz:

- Alpha diversity refers to diversity within a uniform habitat (patch), and is usually measured by counting the number of species present
- Beta diversity is species diversity between ecosystems; the rate and extent of change in species composition from one habitat to another. The analysis of beta-diversity, the extent of change in community structure among sites, has been shown to be a powerful tool in the analysis of biogeographical patterns. One of the most widely used analyses to evaluate beta-diversity is distance decay of similarity, which describes how similarity in community structure varies with increasing distance between localities (similarity decreases with increasing distance).
- Gamma diversity is a measure of the overall diversity for all habitats within a geographical area.

Diversity measurement is based on three assumptions:

- All species are equal: this means that richness measurement makes no distinctions amongst species and treat the species that are exceptionally abundant in the same way as those that are extremely rare species. The relative abundance of species in an assemblage is the only factor that determines its importance in a diversity measure.
- All individuals are equal: this means that there is no distinction between the largest and the smallest individual, in practice however the smallest animals can often escape for example by sampling with nets. Taxonomic and functional diversity measures, however, do not necessarily treat all species and individuals as equal.
- Species abundance has been recorded in using appropriate and comparable units. It is clearly unwise to use different types of abundance measure, such as the number of individuals and the biomass, in the same investigation. Diversity estimates based on different units are not directly comparable.

Measurements of biodiversity: A variety of objective measures have been created in order to empirically measure biodiversity. The basic idea of a diversity index is to obtain a quantitative estimate of biological variability that can be used to compare biological entities, composed of

direct components, in space or in time. It is important to distinguish ‘richness’ from ‘diversity’, as diversity usually implies a measure of both species number and ‘equitability’ (or ‘evenness’).

There have been numerous attempts to create compound indices that combine measures of richness and abundance. Foremost among these are the Shannon’s diversity (H') and Simpson’s diversity ($D1$) indices, which differ in their theoretical foundation and interpretation. H' has its foundations in information theory and represents the uncertainty about the identity of an unknown individual. In a highly diverse (and evenly distributed) system, an unknown individual could belong to any species, leading to a high uncertainty in predictions of its identity. In a less diverse system dominated by one or a few species, it is easier to predict the identity of unknown individuals and there is less uncertainty in the system. This metric is common in the ecological literature, despite its abstract conceptualization. $D1$ is the complement of Simpson’s original index and represents the probability that two randomly chosen individuals belong to different species. $D2$ is closely related to $D1$, being the inverse of Simpson’s original index. Both of these transformations serve to make the index increase as diversity intuitively increases, and although both are used, $D2$ is more common. Finally, evenness represents the degree to which individuals are split among species with low values indicating that one or a few species dominate, and high values indicating that relatively equal numbers of individuals belong to each species. Evenness is not calculated independently, but rather is derived from compound diversity measures such as H' , $D1$, and $D2$, as they inherently contain richness and evenness components. However, evenness as calculated from H' (J') is of only limited use predictively because it mathematically correlates with H' . E , calculated from $D2$, is mathematically independent of $D1$ and therefore a more useful measure of evenness in many contexts. Strong correlations between diversity measures should not be surprising as they represent aspects of the same phenomenon. In fact, most of the measures analyzed can be derived from the same basic generalized entropy formula

$$N_a = (\sum_{i=1}^S P_i^a)^{1/(1-a)}$$

where N_a is the effective species number,
 S is total species number,
 P_i^a is the proportional abundance of species i , and
 a is the power.

H' is equally sensitive to rare and abundant species; sensitivity to rare species increases as a decreases from 1, and sensitivity to abundant species increases as a increases from 1. Therefore, S is sensitive to rare species, $D1$ and $D2$ are sensitive to abundant species, and BP is sensitive to only the most abundant species. As all the N_a ’s have species as the unit, the range of values can be interpreted as a continuum from effective number of the most rare species to effective number of the most abundant species. Formulas to calculate diversity include richness (S), Shannon’s diversity (H'), Berger–Parker dominance (BP), Simpson’s diversity ($D1$), Simpson’s dominance ($D2$), and Simpson’s evenness (E).

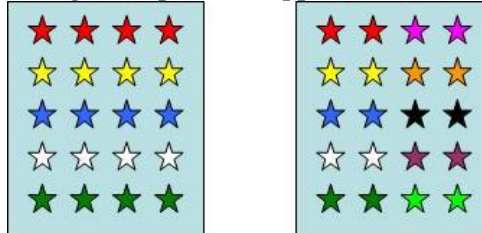
Metric	Traditional formula	Surrogate in Hill’s Series, Hill’s power
Richness (S)	Number of species	$S, 0$
Shannon’s diversity (H')	$-\sum P_i \ln(P_i)$	$\exp(H'), 1$
Simpson’s diversity ($D1$)	$1 - \sum P_i^2$	$D_2, 2$
Simpson’s dominance ($D2$)	$1/\sum P_i^2$	$D_2, 2$
Simpson’s evenness (E)	D_2/S	–
Berger–Parker dominance (BP)	P_{\max}	BP^{-1}, ∞

p_i is the proportion of individuals belonging to species i

p_{\max} is the proportion of individuals belonging to the most abundant species

Basically, three types of indices can be distinguished: species richness indices; evenness indices; and taxonomic indices.

- A. Species richness indices: Species richness is a measure for the total number of the species in a community. However, complete inventories of all species present at a certain location, is an almost unattainable goal in practical applications.



A visualization of species richness with 5 and 10 species respectively.

○ Species richness

- The simplest measure of biodiversity is a count (S) of the total number of different species present in a given area. It does not take into account the proportion and distribution of each species within the community. This measure is strongly dependent on sampling size and effort. Two species richness indices try to account for this problem:

- Margalef's diversity index:

$$DMg = (S-1) / \ln N$$

- Menhinick's diversity index:

$$DMn = S / \sqrt{N}$$

where N = the total number of individuals in the sample,
and S = the number of species recorded.

Despite the attempt to correct for sample size, both measures remain strongly influenced by sampling effort. Nonetheless they are intuitively meaningful indices and can play a useful role in investigations of biological diversity.

- The Smith & van Belle equation can be used to estimate species richness by bootstrap methods B(S) using computer simulated subsampling of data,:

$$B(S) = S + \sum(1-p_i)^n$$

where S = observed number of species

p_i = proportion of the n bootstrap quadrats that have species I present.

The simulation is repeated 100 times to get the mean estimate and its SD.

- Existence of Association (binary coefficients of association)
[not based on abundance data, but on presence/absence data]

- Contingency Table of Co-occurrence

		Sample B	
		1	0
Sample A	1	a (= no. taxa occurring in both samples)	b (= no. taxa occurring in A but not B)
	0	c (= no. taxa occurring in B but not A)	d (= no. taxa absent in both A and B)

- Simple Matching Coefficient

$$S_{SM} = (a+d) / (a+b+c+d)$$

Note that mutual absences contribute to similarity in this coefficient

- Jaccard's Coefficient (also known as Jaccard's number, Jaccard's index, similarity coefficient, species identity) compares members for two sets to see which members are shared and which are distinct. It is calculated as the number of mutual presences divided by the total number of taxa present in the two samples being compared. That is, the number in both sets / the number in either set.

$$S_J = a / (a+b+c) \quad \text{or in notation form: } J(X,Y) = |X \cap Y| / |X \cup Y|$$

It is a measure of similarity for the two sets of data, with a range from 0 to 1 (or 0-100%). The higher the number (or percentage), the more similar the two populations. Two sets that share all members would be 100% similar. The closer to 100%, the more similarity. If they share no members, they are 0% similar. The midway point (50%) means that the two sets share half of the members.

A simple example using set notation:

How similar are these two sets? $A = \{0,1,2,5,6\}$; $B = \{0,2,3,4,5,7,9\}$

$$\begin{aligned} \text{Solution: } J(A,B) &= |A \cap B| / |A \cup B| \\ &= |\{0,2,5\}| / |\{0,1,2,3,4,5,6,7,9\}| \\ &= 3/9 \\ &= 0.33 \text{ (x100 = 33\%)} \end{aligned}$$

- Sorensen's Coefficient (also known as Sorensen-Dice index, F1 score or Dice Similarity Coefficient) is a statistic used for comparing the similarity of two samples. It is calculated as the number of mutual presences divided by average number of taxa in the two samples being compared.

$$\begin{aligned} S_S &= 2a / (2a+b+c) \\ &= 2S_J / (1+S_J) \end{aligned}$$

or alternatively: Sørensen's original formula was intended to be applied to presence/absence data, and is given as the $|X|$ and $|Y|$ are the numbers of elements in the two samples. ranges between 0 and 1 and provides a similarity measure over sets. It differs from Jaccard's index which only counts true positives once in both the numerator and denominator.

- Quantitative Coefficients of Association
[incorporate abundance data]

- Similarity Ratio

$$S_{SR} = \sum(x_{ai} x_{bi}) / [\sum x_{ai}^2 + \sum x_{bi}^2 - \sum(x_{ai} x_{bi})]$$

If used with presence/absence data, this reverts to Jaccard's coefficient.

- Percentage Similarity

$$S_{PS} = 2\sum \min(x_{ai} x_{bi}) / [\sum x_{ai} + \sum x_{bi}]$$

If used with presence/absence data, this reverts to Sorensen's coefficient.

- Percentage Difference. The complement of percentage similarity (Sorensen's similarity index) is the percentage difference (called Bray-Curtis dissimilarity

or Lance-Williams metric). It quantifies the compositional dissimilarity between two different sites based on counts at each site.

$$BC = \frac{\sum |x_{ai} - x_{bi}|}{[\sum x_{ai} + \sum x_{bi}]} \quad \text{which also} = [1 - S_{PS}]$$

An alternative notation is as follows:

$$BC_{ij} = 1 - [2C_{ij} / (S_i + S_j)]$$

where C_{ij} = sum of the lesser values for only those species in common between both sites;

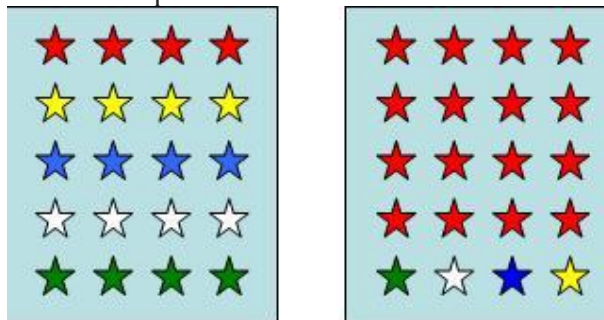
S_i and S_j = numbers of specimens counted at each site.

This index reduces to $1 - 2C/2 = 1 - C$ where abundances at each site are expressed as a percentage.

The BC index is regarded as very good in retaining underlying ecological patterns. Values range from 0 to 1, where 0 means that the two sites have the same composition (they share all the species) and 1 means that the two sites do not share any species.

[Bray-Curtis and Jaccard indices are rank-order similar, but Jaccard's index is metric and probably should be preferred instead of the default Bray-Curtis which is semi-metric].

- B. Evenness/heterogeneity/equitability indices (relative abundance). The distribution of individuals over species is called evenness. It makes sense to consider species richness and species evenness as two independent characteristics of biological communities that together constitute its diversity. Evenness expresses how evenly the individuals in a community are distributed among the different species.



A visualization of the evenness of 5 species.

Species evenness refers to how close in numbers each species in an environment is. Mathematically it is defined as a diversity index, a measure of biodiversity which quantifies how equal the community is numerically. There are several measures of richness and evenness. The famous Simpson index is basically a measure of richness, whereas the Shannon index includes also a measure of evenness. Evenness refers to the similarity of frequencies of the different units making up a population (or sample). It is complementary to richness, which is the number of different units relative to population (or sample) size.

An example is probably more telling than those definitions. Assume you get three population samples (each with $N=10$ individuals) from three different sites. Sample A contains 5 individuals of genotype 1, 2 of genotype 2, 2 of genotype 3 and 1 of genotype 4. Sample B contains 5 individuals of genotype 1 and 5 of genotype 2. Sample C contains 3 individuals each of genotypes 1, 2 and 3, and 1 individual of genotype 4.

	Sample A	Sample B	Sample C
genotype 1	5	5	3
genotype 2	2	5	3

genotype 3	2		3
genotype 4	1		1

Richness in these samples is 0.2 in sample B (2 genotypes out of 10 individuals), and 0.4 (4 genotypes out of 10 individuals) in samples A and C. However, although their richness is the same, evenness in sample C is higher than in sample A, because genotype frequencies are more similar (or more 'even'); they range from 0.1 to 0.3 in sample C vs 0.1 to 0.5 in sample A. Therefore, overall, the population from which sample C comes is assumed to be more diverse than that from which sample A comes.

Several equations have been proposed to calculate evenness from diversity measures:

- Evenness index: The most frequent index used, which converges for large samples, is:

$$E = [I - I_{\min}] / [I_{\max} - I_{\min}] \quad \text{and}$$

$$E = I / I_{\max}$$

where I = a diversity index, and

I_{\min} and I_{\max} = the lowest and highest values of this index for the given number of species and the sample size.

- Pielou's evenness index (J) given by:

$$J = H' / H'_{\max} = H' / \log S$$

The condition of independence of evenness measures from richness measures is not fulfilled for the most frequently used evenness indices. Such measures depend on a correct estimation of S^* , the number of species in the community, which is nearly impossible to do.

Substituting S , the number of species in the sample, makes the evenness index highly dependent on sample size. It also becomes very sensitive to the near random inclusion or exclusion of rare species in the sample.

- Hill's evenness ratios. Hill proposed to use ratios as evenness indices:

$$E_{a:b} = N_a / N_b$$

where N_a and N_b = diversity numbers of order a and b respectively.

[Note that $H' - H'_{\max} = \ln(N_1/N_0)$ belongs to this class, but $J = H' / H'_{\max}$ does not].

Hill showed that in an idealised community, where the hypothesised number of species is infinite and the sampling is perfectly random, $E_{1:0}$ is always dependent on sample size. $E_{2:1}$ stabilises, with increasing sample size, to a true community value. However, in practice all measures depend on sample size.

$$E_{1:0} = e^H / S$$

- Heip's evenness index (E_h). Heip proposed to change the evenness index to

$$E'_{1:0} = (e^H - 1) / (S - 1)$$

In this way the index tends to 0 as the evenness decreases in species-poor communities. Due to a generally observed correlation between evenness and number of species in a sample, $E_{1:0}$ tends to 1 as both $e^H \rightarrow 1$ and $S \rightarrow 1$. However this index falls into the same category as J , being dependent on an estimate of S .

Heterogeneity measures are those that combine the richness and the evenness component of diversity. Heterogeneity measures fall into two categories: parametric indices, which are

based on a parameter of a species abundance model, and nonparametric indices, that make no assumptions about the underlying distributions of species abundances.

○ Parametric

- log series index α is a parameter of the log series model. The parameter is independent of sample size and describes the way in which the individuals are divided among the species, which is a measure of diversity. The attractive properties of this diversity index are that it provides a good discrimination between sites, it is not very sensitive to density fluctuations and it is normally distributed, in this way confidence limits can be attached to α .

The log series takes the form:

$$\alpha x, (\alpha x^2)/2, (\alpha x^3)/3, \dots (\alpha x^n)/n$$

where αx = the number of species to have one individual,
 $(\alpha x^2)/2$ = number of species with two individuals, and so on.

Since $0 < x < 1$ and α and x are presumed to be constant, the expected number of species will be the highest in the first abundance class.

$$x \text{ is calculated iteratively from } S/N = [(1-x)/x] \cdot \ln [1/(1-x)]$$

and α from the equation: $\alpha = [N(1-x)]/x$

○ Non-parametric

- Shannon's Diversity index (H') (also known as Shannon-Wiener, Shannon-Weaver Diversity Index) is similar to Simpson's index, but takes into account species richness and proportion of each species within the community. The Shannon-Wiener diversity index is the most widely used diversity index in the ecological literature. It assumes that individuals are randomly sampled from an infinitely large community, and that all species are represented in the sample. The Shannon index is calculated from the equation:

$$H = \sum_{i=1}^s - (P_i * \ln P_i)$$

where:

H = the Shannon diversity index

P_i = fraction of the entire population made up of species i

S = numbers of species encountered

\sum = sum from species 1 to species s

Note: The power to which the base e ($e = 2.718281828\dots$) must be raised to obtain a number is called the natural logarithm (\ln) of the number.

To calculate the index:

- Divide the number of individuals of species #1 you found in your sample by the total number of individuals of all species. This is P_i
- Multiply the fraction by its natural log ($P_i * \ln P_i$)
- Repeat this for all of the different species that you have.
- Sum all the $-(P_i * \ln P_i)$ products to get the value of H

For example:

Birds	N_i	P_i	$\ln P_i$	$-(P_i * \ln P_i)$
Pigeon	96	.96	-.041	.039
Robin	1	.01	-4.61	.046
Starling	1	.01	-4.61	.046
Crow	1	.01	-4.61	.046

Sparrow	1	.01	-4.61	.046
---------	---	-----	-------	------

H = 0.223

High values of H would be representative of more diverse communities. A community with only one species would have an H value of 0 because P_i would equal 1 and be multiplied by $\ln P_i$ which would equal zero. If the species are evenly distributed then the H value would be high. So the H value allows us to know not only the number of species but how the abundance of the species is distributed among all the species in the community.

- Brillouin index (H) is the appropriated form of the information index where randomness cannot be guaranteed, for example when certain species are preferentially sampled. It is calculated as follows:

$$H = (1/N) \cdot \ln[N! / (\pi N_i)]$$

where $\pi N_i = N_1 \cdot N_2 \cdot N_3 \dots N_i$

N_i = number of individuals in species i and

N = total number of individuals in the community.

- Simpson's index (γ) is one of the best known and earliest evenness measures, given by:

$$\gamma = \sum p_i^2$$

where p_i = proportion of individuals found in the ith species.

This index is used for large, sampled communities. Simpson's index expresses the probability that any two individuals drawn at random from an infinitely large community belong to the same species.

Alternative notation is as follows:

$$D = [\sum n(n-1)] / [N(N-1)]$$

where D = Simpson's diversity index

N = total number of organisms of all species found

n = number of individuals of a particular species

$\uparrow D \equiv$ stable or ancient site, $\downarrow D \equiv$ recent colonization or perturbation

Can also include Jackknife estimate to account for probability of missing some species during sampling)

$$S = n + [(n-1)/n]^k$$

where S = species richness

n = total number of species in sample population

K = number of unique species (of which only one organism was found in sample)

- Simpson's diversity index (D_1 , D or SDI) is a measure of diversity which takes into account the number of species present, as well as the relative abundance of each species. The index assumes that the proportion of individuals in an area indicate their importance to diversity. As species richness and evenness increase, so diversity increase. SDI measures community diversity. The range is from 0 to 1, where high scores (close to 1) indicate high diversity, and low scores (close to 0) indicate low diversity.

$$D = 1 - [\sum n(n-1)] / [N(N-1)]$$

where n = number of individuals of each species,

and N = total number of individuals of all species

EXAMPLE: What is Simpson's Diversity Index for the following 5 species?

Species	# of individuals
1	81
2	2
3	2
4	2
5	2
Total:	89

Calculate answer: Sum the total number in the set ($N=89$) and calculate $N(N-1) = 89(89-1) = 7832$.

Calculate $n(n-1)$ for each species:

Species	# of individuals	(n - 1)	n(n - 1)
1	81	80	6480
2	2	1	2
3	2	1	2
4	2	1	2
5	2	1	2
			6488

Calculate $D = 1 - (6488 / 7832) = 0.17$

- Simpson's dominance index (D_2 , or λ)

$$\lambda = \sum_i p_i^2$$

For large sampled communities, if two individuals are sampled at random and without replacement, this index expresses the probability that they belong to the same species. In small fully censused communities, Simpson's diversity index is used.

- Evenness indices. A whole series of evenness indices can be derived from Simpson's dominance index λ . Since the maximum value of λ is $1/S$ (S = number of species), an evenness index can be written as:

$$E = (1/\lambda)/S$$

This corresponds to Hill's $E_{2:1}$ ratio

$$E_{2:1} = (1/\lambda)/e^H$$

which was modified by Alatalo to

$$E_{2:1} = [(1/\lambda)-1]/[e^H - 1]$$

If the criterion of independence of measures for species richness and evenness is accepted, the choice of indices becomes restricted. It was suggested to use the variation in species abundance in indices. If one uses Hill's number $H_2 = 1/\lambda$, a simple statistic is the weighted mean-square deviation from the proportional abundances that would be expected for H_2 equally abundant species. A measure of evenness is then:

$$D_{MS} = [\sum w_i (p_i - \lambda)^2] / \sum w_i$$

where MS = mean square,

λ = Simpson's index, and

$w_i = p_i$

Hill showed that the expected mean and variance of the relative abundance p_i are given by:

$$E(p_i) = \lambda$$

$$\text{Var}(p_i) = D_{MS}$$

A measure of the shape of the species abundance relation is given by

$$D^*_{MS} = D_{MS} / \lambda^2$$

and a measure of evenness by:

$$E_{MS} = 1/(1 + D^*_{MS})$$

In general, species-abundance distributions show more information about the evenness than any single index. On the other hand, statistics describing these distributions can also be used as measures of evenness.

- Simpson's evenness (E) This index is defined as Simpson's dominance index divided by the number of species:

$$E = D_2/S$$

- Berger-Parker dominance (BP) This index is simply defined as the proportion of individuals belonging to the most abundant species:

$$BP = P_{\max}$$

- Hill diversity numbers (H_0, H_1, H_2) show the relation between the species-richness indices and the evenness-indices. Hill defined a set of diversity number of different order. The diversity number of order a is defined as:

$$H_a = [\sum p_i^a]^{1/(1-a)}$$

where P_i = the proportional abundance of species i in the sample and
 a = the order in which the index is dependent of rare species.

For $a = 0$, H_0 can be seen to equal S , the number of species in the sample.

For $a = 1$, H_1 is undefined by the equation, but defining $H_1 = \lim_{a \rightarrow 1} (H_a)$ gives $H_1 = \exp(H')$ where H' is the well-known Shannon-Weiner diversity index (the most widely used index in ecology).

The next diversity number H_2 is the reciprocal of Simpson's dominance index λ for large sampled communities; i.e. $H_2 = 1/\lambda$

Hill's diversity numbers of different orders probe different aspects of the community. The number of order $+\infty$ only takes into account the commonest species. At the other extreme, $H_{-\infty}$ is the reciprocal of the proportional abundance of the rarest species, ignoring the more common ones. The numbers H_0, H_1 , and H_2 are in between in this spectrum. H_2 gives more weight to the abundance of common species (and is, thus, less influenced by the addition or deletion of some rare species) than H_1 . This, in turn, gives less weight to the rare species than H_0 , which, in fact, weighs all species equally, independent of their abundance.

It is good practice to give diversity numbers of different order when characterising a community. Moreover, these numbers are useful in calculating evenness.

- Metrics based on geometry. Species abundances in a sample can be thought of a x, y, z , etc. coordinates of a point in a multidimensional space; the sample is depicted as a point and the distances between points are related to their similarity/differences.

- Euclidean Distance. This metric is based on the Pythagorean Theorem.

$$D_{AB} = \sqrt{\sum (x_{ai} - x_{bi})^2}$$

D_{AB} has several important properties:

- > $D_{AB} \geq 0$ (positive)
- > $D_{AB} = D_{BA}$ (symmetrical)
- > $D_{Ac} \leq D_{AB} + D_{BC}$ (conforms to triangular inequality)
- > If $A = B$, $D_{AB} = 0$; if $A \neq B$, $D_{AB} > 0$

The value of Euclidean distance is dependent on number of taxa, so one way of scaling it is to divide by the total number of taxa:

$$D_{AB} = \sqrt{(1/p)\sum(x_{ai} - x_{bi})^2}$$

Euclidean distance is one of a general set of distance metrics called Minkowski Metrics:

$$D = \sqrt[z]{(1/p)\sum(x_{ai} - x_{bi})^z}$$

If $z = 1$, then the Manhattan or City Block distance metric is obtained:

$$MCD = (1/p)\sum|x_{ai} - x_{bi}|$$

Overall, Euclidean distance (and more generally Minkowski family of distance metrics) is not good for analysing sparse data, which are typical of ecological data sets. However, this measure is fundamental to methods like Polar Ordination and Non-Metric Multidimensional Scaling (MDS) and sees wide application in geometric morphometrics. In addition, several coefficients (e.g. chord distance) that have proven to be very useful in ecology are closely related to Euclidean distance.

- Cos-theta or Ochiai Coefficient. Rather than evaluate taxonomic difference using multivariate distance, we can examine the angle between two sample points in a multidimensional space:

$$\cos\theta_{AB} = \sum(x_{ai} x_{bi}) / \sqrt{\sum(x_{ai}^2 x_{bi}^2)}$$

when $\theta = 0$, $\cos\theta = 1$; when $\theta = 90$, $\cos\theta = 0$; when $\theta = 180$, $\cos\theta = -1$

If data are z-transformed, then this metric reverts to Pearson's r (i.e. a geometric interpretation of r is as an angle between two vectors in ordination space). This metric has automatic vector length standardization. Like Euclidean distance, it is susceptible to sparse data.

- Chord distance. This metric combines Euclidean distance and angles between points; it is equal to comparing samples standardized to unit vector length using Euclidean distance.

$$CD = \sqrt{\sum \{ [x_{ai} / \sqrt{\sum(x_{ai}^2)}] - [x_{bi} / \sqrt{\sum(x_{bi}^2)}] \}^2}$$

C. Taxonomic indices. These indices take into account the taxonomic relation between different organisms in a community. Taxonomic diversity, for example, reflects the average taxonomic distance between any two organisms, chosen at random from a sample. The distance can be seen as the length of the path connecting these two organisms along the branches of a phylogenetic tree. If two data-sets have identical numbers of species and equivalent patterns of species abundance, but differ in the diversity of taxa to which the species belong, it seems intuitively appropriate that the most taxonomically varied data-set is the more diverse. As long as the phylogeny of the data-set of interest is reasonably well resolved, measures of taxonomic diversity are possible.

- Clarke and Warwick's taxonomic distinctness index which describes the average taxonomic distance – simply the “path length” between two randomly chosen organisms through the phylogeny of all the species in a data-set – has different forms: taxonomic diversity and taxonomic distinctness.
 - Taxonomic diversity (Δ) reflects the average taxonomic distance between any two organisms, chosen at random from a sample. The distance can be seen as the length of the path connecting these two organisms through a phylogenetic tree or a Linnean

classification. This index includes aspects of taxonomic relatedness and evenness. It is calculated as:

$$[\sum_{i < j} \omega_{ij} x_i x_j] / [(N(N-1))/2]$$

- Taxonomic distinctness (Δ^*) is the average path length between two randomly chosen but taxonomically different organisms. This measure is measure of pure taxonomic relatedness. It is calculated as:

$$[\sum_{i < j} \omega_{ij} x_i x_j] / [\sum_{i < j} x_i x_j]$$

- When only presence/absence data is considered both Δ and Δ^* converge to the same statistic Δ^+ , which can be seen as the average taxonomic path length between any two randomly chosen species. It is calculated as:

$$[\sum_{i < j} \omega_{ij}] / [(S(S-1))/2]$$

- VII. Functional diversity. The positive relationship between ecosystem functioning and species richness is often attributed to the greater number of functional groups found in richer assemblages. Petchey and Gaston proposed a method for quantifying functional diversity. It is based on total branch length of a dendrogram, which is constructed from species trait values. One important consideration is that only those traits linked to the ecosystem process of interest are used. Thus a study focusing on bird-mediated seed dispersal would exclude traits such as plumage color that are not related to this function, but traits such as beak size and shape should be included. With standard clustering algorithms a dendrogram is then constructed. The method makes sense. For example a community with five species with different traits will have a higher functional diversity than a community of equal richness but where the species are functional similar.

VIII. Host specificity

A parasite that is specific for a single host species is said to be oioxenous, one that parasitizes closely-related hosts is stenoxenous, while one that parasitizes unrelated hosts is euryxenous. Host-specificity is determined by a complex of factors, some obvious and others still obscure:

- ecological specificity (prospective host shares its environment with the parasite)
- ethological specificity (host behaviour must expose it to the parasite)
- physiological specificity (recognize appropriate cues in permissive host)

Need to define differences between host range, host specificity and host preference:

- host range (= number of hosts that can be infected by a particular parasite species)
- host specificity (usually same as host range, but there are deeper dimensions; e.g.
 - structural specificity (hosts exhibit different population structures)
 - phylospecificity (phylogenetic host specificity) (hosts closely-related or not)
 - geographic host specificity (β -specificity) (host populations structured differently)
- host preference (preferred hosts with greater prevalence)(variations in dominance)

Various techniques have been developed to illustrate similarities/differences in the hosts utilized by parasites: including tables (matrices), diagrams (dendrograms, tanglegrams), quantitative measures (maths models) and statistics (probability).

Host specificity is inversely proportional to ‘generalism’ in host use, and it is the extent to which parasites are generalists that is measured by most indices. Low index values correspond to high host specificity, and vice versa.

- Basic and structural host specificity. Traditionally, host specificity is simply estimated as the number of host species (S) used by a parasite species, but this value can also be corrected for biases arising from the under-sampling of rare hosts using Chao indices. When data on prevalence or abundance can be incorporated into the measurement of structural specificity, composite indices can be used, such as the Shannon index, or else a ‘pure’ evenness index, Σ independent from the number of host species, such as the Bulla index.
- Phylogenetic host specificity (phylospecificity). Because host species are phylogenetically related, we can estimate the phylogenetic host specificity of a parasite, PS_i , as the phylogenetic diversity of its hosts, which is equivalent to the measure PD_i of the biodiversity literature. Here, PS_i represents the total length of branches linking the host species of parasite i along the phylogenetic tree. Because PS_i is not totally independent from the number of host species used by a parasite and thus provides information redundant with S, two options are possible.
 - Estimate the standardized effect size of PS_i , or SPS_i , using random subsets of potential host species drawn from the regional pool to determine whether the hosts actually used by the parasite are more or less closely related than expected by chance, and thus whether the phylospecificity of parasite i is high or low for a given value of S using:

$$SPS_i = (PS_i - \underline{PS}_{sim}) / [SD(PS_{sim})]$$
 where PS_i = observed phylospecificity of parasite i ,
 \underline{PS}_{sim} = mean phylospecificity of all random host subsets, and
 $SD(PS_{sim})$ = standard deviation of all randomized phylospecificity values.

- Estimate phylospecificity as the average phylogenetic distinctiveness, SPD_i , between all pairs of host species, which is independent from how many host species are used by a parasite:

$$SPD_i = 2 \{ \sum \sum_{j < k} \omega_{jk} / [S(S-1)] \}$$

where ω_{jk} = the phylogenetic distance between host species j and k used by parasite i ,
[or, when the phylogeny is not fully resolved, the number of taxonomic steps required to reach a node common to both]

The double summation is over the set $\{k = 1, \dots, S; j = 1, \dots, S, \text{ such that } j < k\}$
in order to consider all host species pairs.

- Geographic host specificity (β -specificity). Measuring the turnover of host species used by a parasite among different localities, in other words β -specificity or BS_i , involves estimating the dissimilarity in host species identities between localities. Most dissimilarity or β -specificity indices have been designed for two samples only. Ideally, estimates of β -specificity across space should include several samples (i.e. data from different localities). We suggest using the extension of the Sørensen dissimilarity index for multiple-sites to measure β -specificity:

$$BS_i = \{1 - [T / (T - 1)].[1 - (S_T / \sum_t S_t)]\}$$

where T = the number of samples or localities,

S_t = the number of host species used in locality t , and

S_T = the total number of host species used by parasite species i across all T localities (i.e. the regional host pool).

If parasite i exploits the same host species across all localities, then $S_t = S_T$ and $BS_i = 0$.

If parasite i uses totally different host species from one locality to the next, then $S_T = \sum_t S_t$ and $BS_i = 1$.

- Combining phylogenetic and geographic specificity (phylobetaspecificity). Information about the phylogenetic relatedness of host species and their different use across localities can be combined into a single index of phylogenetic β -specificity, or PBS_i . This corresponds to the phylogenetic turnover of host species used by parasite i over geographic space. For this, we can use an extension of the Sørensen index to branches instead of species following the principle underlying the construction of the Phylosor index:

$$PBS_i = \{1 - [T / (T - 1)].[1 - (PD_T / \sum_t PD_t)]\}$$

where T = the number of samples or localities,

PD_t = the phylogenetic diversity of host species used by the parasite in locality t , and

PD_T = the phylogenetic diversity of all host species used by parasite species i across all T localities.

If parasite i exploits the same host species over all localities, $PD_t = PD_T$ and $PBS_i = 0$.

If parasite i uses different host species from one locality to the next, then the less phylogenetically related those hosts are, the higher the PBS_i value.

The Sørensen index is used as a common statistical index because it can cope with multiple localities and can incorporate phylogenetic diversity. If prevalences are known for each host species used by a parasite, an alternative framework could be developed to estimate specificity over geographic and phylogenetic space while also incorporating structural host specificity, based for instance on the Rao index or the Shannon entropy index

Two R packages can be used to compute the above-named indices.

First, the package ‘vegan’ allows one to estimate:

- Basic host specificity, i.e. the number of host species used by a parasite in a locality, with the function `specnumber`. The function `estimateR` then allows a correction for biases arising from the undersampling of rare species that could escape detection as hosts.

- Structural host specificity, with a composite index such as Shannon or Simpson using the function diversity.
- Phylogenetic distinctiveness among host species, or SPDi, with the function taxondive which requires the taxonomic (or the phylogenetic) distances between all host species pairs as an input file.

Second, the package 'picante' allows one to estimate:

- The phylogenetic diversity of the host species exploited by parasite i (PDi) using the function pd. The phylogenetic tree required as input needs to be ultrametric.
 - The standardized effect size of the phylogenetic diversity of hosts exploited by parasite i (SPSi) using the function ses.pd and shuffling taxa labels in the host phylogenetic tree. The phylogenetic tree also needs to be ultrametric.
 - The Rao index, through the function raoD, which allows inclusion of prevalence data to compute specificity indices over space and phylogeny.
- Specificity matrix. A $n \times n$ matrix (table) can be developed to show:
 - parasite species richness (shown along diagonal)
 - similarity scores (the numbers of shared parasite species above diagonal)
 - difference/dissimilarity scores (numbers of non-shared species below diagonal)

		Host species			
		A	B	C	D
Host species	A	3	2	2	0
	B	1	2	1	0
	C	1	2	2	0
	D	4	3	3	1

Interpretation:

- 3 parasite species detected in host species A, 2 shared with B, 2 with C and none with D
- 2 parasite species detected in host species B, 1 shared with A, 1 with C
- 2 parasite species detected in host species C, 1 shared with A, 1 with B
- 1 parasite species detected in host species D, not shared with others (= host specific)

- Similarity matrix. A square symmetrical matrix with the similarity value of every pair of samples (if Q-mode) or species (if R-mode) in the data matrix. The similarity matrix is the basis for all multivariate techniques depicting relationships among community samples or taxa, so the choices made at the initial stage of an analysis will strongly influence the results at the final stage. Similarity matrices are required for cluster analysis.
- Dissimilarity matrix. All similarity matrices can be converted to dissimilarity (difference) metrics by subtracting them from their maximum value. Difference matrices are required for ordination analysis. Analyses convert abundance data to presence/absence data to make all species equally important in characterizing a sample, regardless of their abundance. Compare dissimilarity of parasite assemblages, and evolutionary distance between host species. Dissimilarity matrix created using Sorensen's Index of Similarity for each host pair, then visualized in ordination graph [non-metric multi-dimensional scaling (MDS) ordination] [similar approach to that of Nipperess et al. 2012 (examined plant phylogeny and insect assemblages)]. Evolutionary distance = age of MRCA (MYA) from molecular clock studies.

- Ecological niche is defined as a multidimensional hypervolume determined by environmental (biotic and abiotic) variables within which a species can exist. Such dimensions include host range (host specificity), microhabitats, macrohabitats of the host, geographical range, sex and age of the host, season, food, hyperparasites, etc. Niches are restricted to varying degrees along all dimensions, particularly for parasites. Some infect a wide range of hosts and others are restricted to a single or a few hosts. Some exhibit high tissue tropism (select microhabitats), others may infect many tissues. Niches are also not static, but vary over time. Measures of niche width include: [Rohde & Rohde (2005) in Rohde (2005) p.286]
 - Levin's niche width (B) $B = 1 / (\sum p^2)$
 - Shannon-Weiner measure (H') $H' = -\sum p \cdot \log p$
 - Smith' measure (FT) $FT = \sum \sqrt{p \cdot a}$

where p = proportion of individuals found
 a = proportion of total resources

The number of hosts utilized is particularly important for parasites. The degree of host specificity observed may be an artifact of sampling effort (often indicated by strong positive correlation between number of host species and number of time a parasite has been recorded).

- Host specificity. Rohde distinguished between host range and host specificity.
 - Host range is the total number of host species found to harbour a certain parasites species, irrespective of prevalence and intensity.
 - Host specificity considers prevalence (percentage infected) and/or intensity (number of parasites per host individual). The most commonly used measure for specificity is Rohde's specificity index (S_i)

$$S_i = \sum_j [x_{ij} / (n_{ij} \cdot h_{ij})] / \sum_j (x_{ij} / n_{ij})$$

If intensity is considered,

x_{ij} = number of parasite individuals of i -th species in j -th host species;

n_{ij} = number of host individuals of j -th species examined;

h_{ij} = rank of host species j (species with greatest intensity has rank 1); and

x_{ij}/n_{ij} = intensity of infection

If prevalence (frequency) is considered,

x_{ij} = number of host individuals of j -th species infected with parasite species i ;

n_{ij} = number of host individuals of j -th species examined;

h_{ij} = rank of host species j (species with highest frequency has rank 1); and

x_{ij}/n_{ij} = mean frequency of infection.

Numerical values vary between 0 and 1. The closer to 1, the higher the degree of host specificity. However, the minimum value for S_i depends on the number of host species used for calculating the index. Since most parasites infect fewer than 10 hosts (for which the minimum S_i is about 0.2), the high value of S_i does not necessarily mean a strong preference for a single or a few host species (thus making comparisons unreliable).

Rohde's index can also be modified to remove the problem of sensitivity to the number of host species infected

$$S_i = \{[\sum_j [x_{ij} / (n_{ij} \cdot h_{ij})] / \sum_j (x_{ij} / n_{ij})] - [1 / j \sum_j (1/j)]\} / [1 - (1/j) \sum_j (1/j)]$$

where $(1/j) \sum_j (1/j)$ = minimum possible S_i (used to normalise S_i);

j = number of host species; and n_{ij} = number of individuals of host species j infected with parasite species i .

This modified equation is based on the assumption that for minimum possible S_i all species have approximately the same value of x_{ij}/n_{ij} and maintain unique ranks.

- Phylogenetic information can be added to this modified equation, by adding codes for the phyla, classes, orders, families and genera infected (P.C.O.F.G): as follows:
 1.1.1.1.1 = 1 phylum, 1 class, 1 order, 1 family, 1 genus infected
 1.2.5.6.7 = 1 phylum, 2 classes, 5 orders, 6 families, 7 genera infected
 For example, $S_i = 1.1.1.2.3.099$ implies that 3 genera in 2 families in 1 order are infected but that almost all parasites are concentrated in a single host species; while $S_i = 2.2.2.2.3.02$ implies that 2 phyla, 2 classes, 2 orders, 2 families and 3 genera are infected, and that infections are spread more or less evenly over all taxa.

However, this does not consider the unevenness of infection between different taxa above the species level (which is included in Poulin & Mouillot's S_{TD} index (below).

- Host phylogenetic position. Many parasites infect several host species. If those host species are closely-related, the parasites should be considered more host specific than those infecting the same or smaller numbers of hosts belonging to different higher taxa. The phylogenetic position of the hosts can be considered using alternative measures than S_i with phylogenetic information added (above).
 - Decimalized index (S.G.F.O.C) (HS). It was proposed that host specificity could be indicated by a decimalized system showing the breadth of the host taxonomic ranks (species, genus, family, order, class). For example, a species restricted to a single host species has the rank 1.1.1.1.1 = 1 while a species infecting 1,000 species, 500 genera, 150 families, 75 orders and 5 classes (the maximum number in each taxon permitted by this index) has the rank 11 795 988 501. Rank values are calculated by enumerating all combinations of S, G, F, O and C. Using the log values of the ranks, index (HS) values of between 0 and 10 are reached (program available, Cairns 2003).

The index does not consider the prevalence or intensity of infection. It considers both the number of higher taxa involved and the number of species infected (e.g. a parasite found in a single host species has HS of 0, one found in 1000 species all in one genus has a HS of 3). However, the index is not sensitive to an uneven distribution of host species among higher taxa (e.g. HS is the same for 1 parasite infecting 3 host species in each of 10 genera; and for 1 parasite infecting 10 genera but 21 species in 1 genus and 1 each in the other 9). The following modification was proposed.

- Poulin & Mouillot's index (S_{TD}) The host specificity of a parasite is not merely a function of how many host species it can exploit, but also of how closely related these host species are to each other. This index of host specificity takes into account the average taxonomic or phylogenetic distance between pairs of host species used by a parasite. The index is derived from measures of taxonomic distinctness used in biodiversity studies. When these host species are placed within a taxonomic hierarchy, based on the Linnean classification into phyla, classes, orders, families, genera and species, the average taxonomic distinctness is simply the mean number of steps up the hierarchy that must be taken to reach a taxon common to 2 host species, computed across all possible pairs of host species. Thus, if 2 host species are congeners, 1 step (species-to-genus) is necessary to reach a common node in the taxonomic tree; if the 2 species belong to different genera but the same family, 2 steps will be necessary (species-to-genus, and genus-to-family); and so on, with these numbers of steps averaged across all host species pairs. For any given species pair, the number of steps corresponds to half the path length connecting two species in the taxonomic tree, with

equal step lengths of one being postulated between each level in the taxonomic hierarchy.

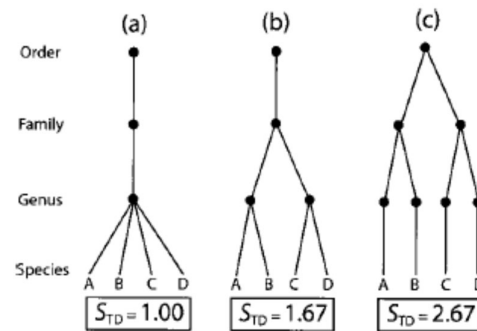


Fig. 1. Hypothetical taxonomic trees for the hosts of 3 species of parasites. Each parasite species is found on 4 host species, A–D. In (a), all 4 host species are congeners, and thus they average a single step toward a common node i.e. $S_{TD}=1$. In (b), the 4 hosts belong to 2 different genera but to the same family, with some species pairs, such as A–B, requiring 1 step to reach a common node, and others, such as A–C, requiring 2 steps. Example (c) shows even greater taxonomic distinctness, with the 4 hosts belonging to different families within the same order. Note that the maximum possible value of S_{TD} in these examples is 3, because there are 3 taxonomic levels above species.

The greater the taxonomic distinctness between host species, the higher the number of steps needed, and the higher the value of the index S_{TD} : thus it is actually inversely proportional to specificity. It is based on presence/absence data and does not consider the prevalence or intensity of infection.

$$S_{TD} = 2 \left[\sum \sum_{i < j} \omega_{ij} \right] / [s(s-1)]$$

where s = number of host species used by a parasite;

double summation is over the set $i=1 \dots s; j=1 \dots s$, such that $i < j$), and

ω_{ij} = taxonomic distinctness between host species i and j

(or the number of taxonomic steps required to reach a node common to both)

The greater the number of steps, the larger the value of S_{TD} . Using the standard 5 taxonomic levels above species i.e. genus, family, order, class and phylum, the maximum value that the index S_{TD} can take (when all host species belong to different classes) is 5, and its lowest value (when all host species are congeners) is 1. The index measures the average taxonomic distinctness between host species, but not all the asymmetries in the taxonomic distribution of host species across higher taxa. The following modification was proposed.

- Variance of distribution ($VarS_{TD}$) considers asymmetries in the taxonomic distribution of host species.

$$VarS_{TD} = \sum \sum_{i \neq j} (\omega_{ij} - \underline{\omega})^2 / [s(s-1)]$$

where $\underline{\omega}$ = average taxonomic distinctness, or S_{TD} .

The variance can only be computed when a parasite uses no fewer than 3 host species. It also does not consider the number of species in a genus infected (e.g. $S_{TD} (=1)$ for species found in 20 congeneric host species is same as one found in 5).

- Inter-specific associations. The distribution of parasites among individuals in a host population (infracommunity) can range from random to structured (with the distribution of different

parasite species associated negatively or positively). The null hypothesis is that the probability of occurrence of any parasite species in a host individual is equal to its prevalence in the host population, and is completely independent of the presence of other parasite species. Analyses can be conducted on associations between two parasite species (pairwise) or multiple parasite species (nestedness).

- Covariance (associations between pairs of parasites). When considering all possible pairwise associations, the number of positive covariances should equal the number of negative covariances if infra-communities are assembled randomly. However, positive covariances often outnumber negative ones for parasites. Pairwise associations can be analysed using either prevalence or abundance data (although the two methods usually give congruent results).
- Nestedness (associations between multiple parasite species). Rather than analyse data one pair at a time, it is possible to analyse data sets comprising multiple parasite species.
 - presence/absence matrix (parasite species rows, individual hosts columns) test for presence of nested species subsets (nestedness) (rare species often found only in species-rich hosts)
 - test for departure from 'null' model based on each species' prevalence, using Monte-Carlo simulations
 - nestedness often associated with older (bigger) hosts as they accumulate infections with time
 - repeatability of community structure in space (other locations) or time (seasons). Are they similar (strong consistent structuring processes) or different (too dynamic for any structuring forces)
 - processes may include:
 - interspecific competition (niche apportionment, food..)
 - transmission (similar, linked, independent)(same life-cycles, same alternate hosts)
 - host food preferences, foraging behaviours

Patterns of community structure often differ between different populations of a single host species suggesting they are transient properties and that parasite communities are shaped by multiple forces acting simultaneously. Analyses may be profoundly affected by small sample sizes.

The RANDOM1 algorithm (Patterson & Atmar, 1986) can be used to compute an index of nestedness for all locations in all host species (indicating whether parasite species followed a nested pattern within samples). Only parasite component communities involving at least 3 different parasite species are included (nestedness is meaningless for communities of 1 or 2 species). The index of nestedness N was computed for each component community. N corresponds to the sum, among all parasite species, of the instances in which a parasite species is absent from infra-communities richer than the most species-poor infra-community in which it occurs. For each component community, the observed N value is compared with N values of 1,000 randomly generated presence/absence matrices produced using RANDOM1. In these Monte Carlo simulations, the probability of each species being included in an infra-community was set as equal to its observed prevalence in the studied host sample. The proportion of simulated N values that were lower than or equal to the observed N value provided the RANDOM1 P value, which was used as a measure of departure from the structure expected from random assembly. When the RANDOM1 P value is ≤ 0.05 , the infra-communities are significantly nested; when the P value is ≥ 0.95 , they have a significant

anti-nested pattern. Either significant nestedness or anti-nestedness is considered a departure from a random assemblage.

- Host-parasite co-speciation, host-switching, ecological fitting (Brooks et al., 2006). There are two approaches to studying the evolution of host-parasite associations.
 - Co-speciation is based upon comparing host-parasite phylogenies and identifying points of congruence. There are no assumptions about underlying processes, nor is there an expectation of complete congruence. Incongruent portions of host-parasite phylogenies require further investigation into the influence of other factors (e.g., dispersal and host switching). For example, parasites might diverge more rapidly than their hosts via sympatric speciation, producing sister species inhabiting the same host (lineage duplication), or ecological or immunological evolution in the host lineage could cause parasite extinction (lineage sorting or missing the boat).
 - Maximum (synchronous) co-speciation assumes that hosts and their parasites share such a specialized and exclusive evolutionary association and that speciation in one lineage causes speciation in the other. Host parasite phylogenies are thus expected to be completely congruent, with departures from congruence explained by invoking extinction in one lineage or the other.

The two approaches differ with respect to the importance of host switching during the evolution of host-parasite associations. It is assumed that hosts and parasites share a specialized exclusive evolutionary association, making it extremely unlikely that a parasite could change host species. This assumption, however, arises from believing that it is the host species, not a biological characteristic or combination of characteristics of the host, that is important to the parasite. Once researchers began thinking in terms of traits rather than taxonomy, it became evident that parasites might be able to switch hosts if the trait they were tracking was shared among two or more hosts. The fact that present-day associations might be shaped in part by the distribution of phylogenetically conservative traits is called ecological fitting.

There are many macro-evolutionary manifestations of ecological fitting. For example, any given parasite species might be a resource specialist, but also might share that specialist trait with one or more close relatives. That is, specialization on a particular resource can be plesiomorphic within a group. On the other hand, the resource itself might be at once very specific and taxonomically and geographically widespread if it is a persistent plesiomorphic trait in the hosts. The evolutionary basis for ecological fitting is thus deceptively simple, yet powerful. If specific cues/resources are widespread, or if traits can have multiple functions (or both), then the stage is set for the appearance of ecological specialization and close (co) evolutionary tracking as well as host switching. Ecological fitting thus explains how a parasite can be ecologically specialized and still switch hosts: if the resource is widespread across many host species, then the parasite can take advantage of an opportunity to establish a "new" specialized association without the cost of evolving novel abilities. Just because a resource is widespread does mean that it is automatically available. The geographic distribution of the parasite might not coincide with the geographic distribution of all hosts having the resource, or some other aspect of host biology might make the resource inaccessible to the parasite. For example, if host species A bearing resource x is highly abundant in a community, then less-abundant host species B and C, which also bear x might not be "apparent" to a parasite specializing on that resource. Such density-dependent factors provide the appearance of close ecological tracking between the parasite and species A at time T₀. If some environmental stressor later decreases the abundance of species A, and C becomes relatively more apparent, then the parasite will become associated with C at time T. This manifestation of ecological fitting could explain seemingly rapid and virtually unconstrained evolution of novel specialized host associations. Finally, a parasite might have a hierarchy of host preferences, even though it is tracking the same resource. The hierarchy arises because the costs of accessing the resource

might not be identical across all host species or even across individuals in the same species. Such costs will depend on many different factors, including concentration of the resource, host density, and difficulty in extracting the resource. Overall, parasites accessing a plesiomorphically (or, less often, homoplasiously) distributed resource are "faux generalists": specialists whose host range appears large, but who are in reality using the same resource. If a parasite species evolves the ability to utilize a novel resource, a second and more complicated type of host preference hierarchy can arise if the parasite also retains sufficient information to use the plesiomorphic resource.

Ecological fitting is generally investigated in insect plant systems, because researchers can reconstruct phylogenetic patterns of association between the two clades, then examine the processes underlying those patterns by (1) identifying the resource being tracked by the insect, (2) determining the distribution of that resource among host plants, and (3) delineating the host preference hierarchy of the insects. Currently, we do not have this degree of detailed information for any host-parasite system. It is possible, however, to take advantage of "natural experiments", or even to make inferences based on contemporary patterns of host-parasite association, if hosts vary in their use of a habitat to which parasite species are constrained. The associations between anurans and their plathyhelminth parasites provide a model system for such an investigation, because the majority of helminths require water for the development and transmission of infective stages, while most, but not all, major groups of anurans have a sexual and developmental tie to aquatic habitats. It was suggested that species richness in anuran parasite communities was directly related to the amount of time the host spent in or near water. A shared plesiomorphic requirement for an aquatic habitat, coupled with a gradient of adult anuran preferences ranging from aquatic to arboreal, suggests that ecological fitting as a determinant of the parasites associated with a given anuran taxon should be evidenced as a nested-subset structure of host-parasite associations across anuran taxa. At one extreme, if all the host-parasite associations are the result of ecological fitting, then all host taxa are interchangeable from the point of suitability for the parasites, and associations will be determined solely by the habitats the host utilizes and its feeding preferences. The shared requirement of tadpoles for aquatic habitats should thus provide a baseline assemblage of parasites that infect the tadpole stage, while the parasites of adult anurans should accumulate in anuran host species as a function of the time they spend in aquatic habitats as adults. If specialized coevolutionary processes dominate, sympatry between anurans and the infective parasitic stages will result in parasitism of only appropriate hosts, producing idiosyncratic (i.e., "unexpected") presences absences in the matrix of host-parasite associations.

Examination of the nested-subset structure of parasite genera within the pooled anuran genera across all six localities was conducted using the nestedness temperature calculator, which calculates the temperature of the matrix (a measure of order, with lower temperatures indicating a higher degree of order) and idiosyncratic host and parasite temperatures, which indicate host species and parasite species contributing disproportionately to the lack of order in the matrix.

IX. Correlations

- Spearman's rank correlation coefficient (r_s)
- Pearson product-moment correlation coefficient (r). The basic measure of correlation in classical statistics is Pearson's product-moment correlation coefficient (r). This is the covariance scaled by products of the standard deviations of the two variables.

$$r = \frac{\sum(X_{ai} - \bar{X}_{ai})(X_{bi} - \bar{X}_{bi})}{\sqrt{[\sum(X_{ai} - \bar{X}_{ai})^2](\sum(X_{bi} - \bar{X}_{bi})^2)}}$$

This coefficient plays critical role in many techniques of parametric ordination, such as Principal Components Analysis (PCA), but as a measure of ecological similarity, it is very susceptible to sparse data (i.e. data matrices with many zeros indicating that many species do not occur in many samples). [Note that r is the dot product of two z-transformed vectors of data].

Look for correlations between:

- prevalence and abundance/intensity

Then look for correlations between parasite prevalence/abundance with:

- host phylogeny (species)
- biogeography (distribution)
 - spatial (location, colonies, etc)
 - temporal (seasons)
- host population demographics
 - size/age
 - gender
 - castes
 - diets
 - behaviours
- parasite population dynamics
 - size
 - diet
 - motility/cytoskeleton
 - organelles (mitochondria/hydrogenosomes)
 - symbiotic bacteria (ecto/endo, cytoplasmic/nuclear)

X. Equilibrium or Non-equilibrium

Do species live under equilibrium or non-equilibrium conditions?

Are populations in balance? Are dynamical changes balanced (e.g. births/deaths, predator/prey, emigration/immigration, etc.).

- Equilibrium assumptions (Hubbell's neutral theory of biodiversity)
 - communities saturated with individuals leading to zero-sum game
 - saturation with species
 - open communities
- Contrary paradigm, non-equilibrium prevails. Few populations reach equilibrium because:
 - repeated strong disturbances (storms, fires, droughts, predation..)
 - many habitats not saturated with species
 - interspecific competition occurs

Parasites often live in non-saturated non-equilibrium assemblages. Evident when:

- prevalence/abundance is very low
- empty niches common
- mating hypothesis of niche restriction
- nestedness uncommon
- rare hyperparasites
- spatial scaling laws do not apply
- little evidence for non-random co-occurrences
- assemblages log series or log normal distributions

- XI. Principal Components Analysis (PCA). This statistical method used to reduce the number of variables in a dataset by lumping highly correlated variables together (this comes at the expense of accuracy). However, if you have 50 variables and realize that 40 of them are highly correlated, you trade-off accuracy for simplicity.

Principal component analysis (PCA) is a powerful yet simple method widely used for analyzing high dimensional datasets. When dealing with datasets such as gene expression measurements, some of the biggest challenges stem from the size of the data itself. Transcriptome wide gene expression data usually have 10,000+ measurements per sample, and commonly used sequence variation datasets have around 600,000 - 900,000 measurements per sample. The high dimensionality not only makes it difficult to perform statistical analyses on the data, but also makes the visualization and exploration of the data challenging. These datasets are typically never fully visualized because they contain many more datapoints than you have pixels on your monitor. PCA can be used as a data reduction tool, enabling you to represent your data with fewer dimensions, and to visualize it in 2-D or 3-D where some patterns that might be hidden in the high dimensions may become apparent.

Another way to think about PCA is in terms of removing redundancy between measurements in a given dataset. What we do is getting rid of the redundancy in the data by grouping the related measurements together into a dimension. In PCA these dependencies, or relationships between measurements are assessed by calculating the covariance between all the measurements. So how do we calculate covariance? First, let us begin with the variance of a single variable which is calculated by:

$$E[(x - \bar{x})^2] = s^2 = \Sigma(x_i - \bar{x})^2 / (n - 1)$$

Covariance can simply be thought of as variance with 2 variables, which takes this form:

$$E[(x - \bar{x})(y - \bar{y})] = \text{cov}(x,y) = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / (n - 1)$$

If this looks familiar you are not mistaken. Correlation between two variables is just a scaled form of covariance, which only takes values between -1 and 1.

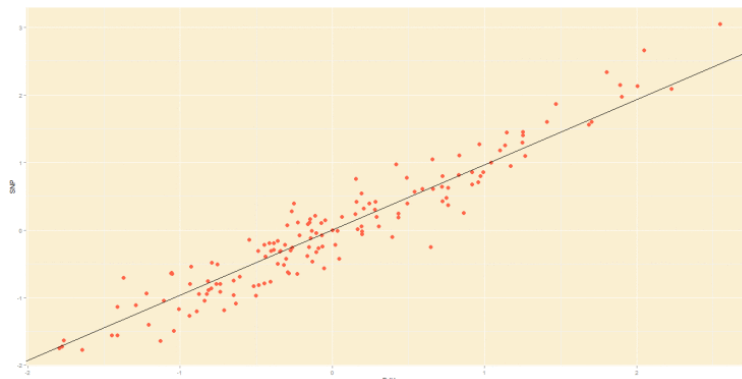
$$\text{cor}(x; y) = \text{cov}(x,y) / \sigma_x \sigma_y$$

So just like correlation, covariance will be close to 0 if the two variables are independent and will take a positive value if they tend to move in the same direction and a negative value if the opposite is true.

A simple example involves a dataset of two variables: e.g.

- D = Dow Jones Industrial Average (a stock market index), and
- S = S&P 500 aggregate (another market index)

Not surprisingly, the D and S are highly correlated (their daily % readings move together). The paired data-points are represented in 2 axes: X and Y. PCA allows the data to be represented along one axis (called the principal component) (represented by line):



Reducing the data to a single axis will reduce accuracy because the data varies about the axis. In PCA, the following steps are conducted to transform the data.

- 1. Standardize the scale of the data. This was done by transforming the data into daily % change (now both D and S occur on a 0-100 scale).
- 2. Calculate covariance matrix for data. The covariance between D and S is a measure of how the two variables move together.

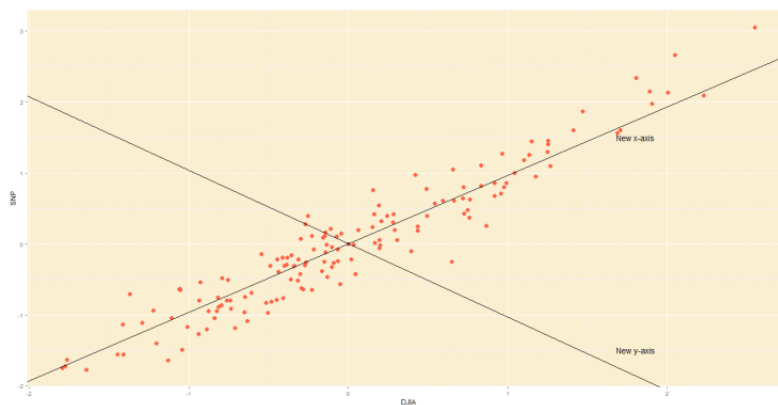
$$\text{cor}(X,Y) = \text{cov}(X,Y) / (\sigma_X \cdot \sigma_Y)$$

The covariance matrix looks like:

$$\begin{bmatrix} \text{Cov}(\text{DJIA}, \text{DJIA}) & \text{Cov}(\text{DJIA}, \text{S\&P}) \\ \text{Cov}(\text{S\&P}, \text{DJIA}) & \text{Cov}(\text{S\&P}, \text{S\&P}) \end{bmatrix} = \begin{bmatrix} \text{Var}(\text{DJIA}) & \text{Cov}(\text{DJIA}, \text{S\&P}) \\ \text{Cov}(\text{S\&P}, \text{DJIA}) & \text{Var}(\text{S\&P}) \end{bmatrix}$$

$$= \begin{bmatrix} 0.7846 & 0.8012 \\ 0.8012 & 0.8970 \end{bmatrix}$$

- 3. Deduce eigens: Eigenvectors and values exist in pairs: every eigenvector has a corresponding eigenvalue. An eigenvector is a direction (vertical, horizontal, 45 degrees etc.) and an eigenvalue is a number representing the amount of variance in the data in that direction (i.e. telling us how spread out the data is on the line). The eigenvector with the highest eigenvalue is the main principal component. This becomes the new X-axis and a line perpendicular to it becomes the new Y axis (the other principal component).



The data are now rotated to fit the new axes. The coordinates of the rotated data are calculated by converting the data by multiplying them by eigenvectors, which indicate the direction of the new axes (principal components). First, the eigenvectors (one per axis) are deduced where each eigenvector correspond to an eigenvalue, whose magnitude indicates how much of the data's variability is explained by its eigenvector. The definition of eigenvalue and eigenvector is:

$$[\text{Covariance matrix}] \cdot [\text{Eigenvector}] = [\text{eigenvalue}] \cdot [\text{eigenvector}]$$

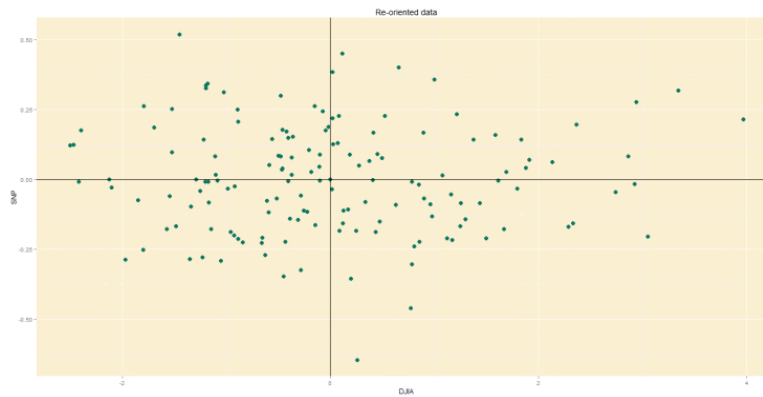
This equation and the covariance matrix are used to yield the eigenvalues (e) and eigenvectors (E):

$$e_1 = 1.644 \quad E_1 = \begin{bmatrix} 0.6819 \\ -0.7314 \end{bmatrix} \quad e_2 = 0.0376 \quad E_2 = \begin{bmatrix} -0.7314 \\ 0.6819 \end{bmatrix}$$

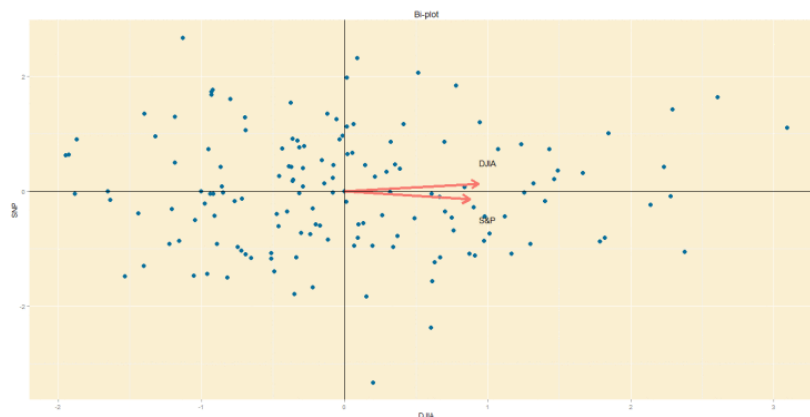
- 4. Re-orient data. Since the eigenvectors indicates the direction of the principal components (new axes), the original data is multiplied by the eigenvectors to re-orient the data (which is now called a score, Sc).

$$\text{Sc} = [\text{orig.data}] \cdot [\text{eigenvectors}] = \begin{bmatrix} \text{DJIA}_1 & \text{S\&P}_1 \\ \text{DJIA}_n & \text{S\&P}_n \end{bmatrix} \times \begin{bmatrix} 0.6819 & 0.7314 \\ -0.7314 & 0.6819 \end{bmatrix}$$

- 5. Plot re-oriented data. The rotated data (scores) are now plotted.

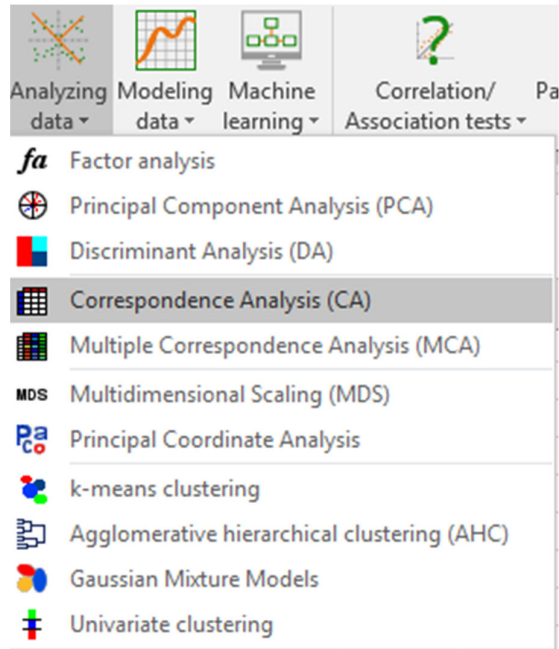


- 6. Bi-plot. A PCA would not be complete without a bi-plot. This is basically the plot above, except the axes are standardized on the same scale, and arrows are added to depict the original variables.
 - Axes: In the bi-plot, the X and Y axes are the principal components.
 - Points: These are the dat points (D and S) re-oriented to the new axes.
 - Arrows: The arrows point in the direction of increasing values for each original variable. For example, points in the top right quadrant will have higher D readings than points in the bottom left quadrant. The closeness of the arrows means that the two variables are highly correlated.



PCA was performed using R, with help of ggplot2 package for graphs. Applying PCA in R is very simple once you get familiar with the outputs of the functions that perform PCA. Similarity patterns in parasite community structure between host individuals can be assessed by PCA using arcsin-transformed population ratio data for different colonies. The first principal component (PC1) score can be subjected to F-tests to compare the degrees of community structure variation between castes and colonies. All analyses can be carried out with R version 2.9.2 (<http://www.r-project.org/>) using a function “princomp2” by Aoki (<http://aoki2.si.gunma-u.ac.jp/R/pca.html>). Hierarchical cluster analyses (“hclust” package for R) using the Bray-Curtis distance may be used to identify geographic regions with similar parasite assemblages. The unweighted pair group method with arithmetic mean (UPMGA) agglomerative algorithm chosen after analysis of cophenetic correlation coefficients (Pearson correlation between cophenetic distances calculated on cluster branches and the parasite dissimilarity matrix).

Analyses can also be conducted in Excel using XLSTAT software (see on-line tutorials).



XII. Correspondence analyses

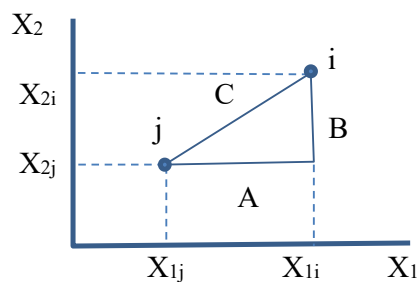
Canonical Correspondence Analysis (CCA) was used to characterize associations between lake environmental variables and characteristics of the parasite component communities. CCA is a common type of multivariate analysis that is used to 1) infer species-environment relationships, and 2) detect patterns that are best explained by a particular set of environmental variables. CCA was performed on the statistical package PC-ORD (Version 2.0). The highest canonical coefficients indicated those environmental variables with the most influence on parasite community structure.

Correspondence analysis (CA) or reciprocal averaging is a multivariate statistical technique conceptually similar to principal component analysis, but applies to categorical rather than continuous data. In a similar manner to principal component analysis, it provides a means of displaying or summarising a set of data in two-dimensional graphical form. All data should be nonnegative and on the same scale for CA to be applicable, keeping in mind that the method treats rows and columns equivalently. It is traditionally applied to contingency tables — CA decomposes the Chi-squared statistic associated with this table into orthogonal factors. Because CA is a descriptive technique, it can be applied to tables whether or not the Chi-squared statistic is appropriate.

XIII. Cluster Analysis

Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. The objective is to assign observations to groups (clusters) so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves stand apart from one another. In other words, the objective is to divide the observations into homogeneous and distinct groups. In contrast to the classification problem where each observation is known to belong to one of a number of groups and the objective is to predict the group to which a new observation belongs, cluster analysis seeks to discover the number and composition of the groups. However, cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. The choice of variables included in a cluster analysis must be underpinned by conceptual considerations. This is very important because the clusters formed can be very dependent on the variables included.

The data used in cluster analysis can be interval, ordinal or categorical. A number of different measures have been proposed to measure 'distance' for binary and categorical data. For interval data the most common distance measure used is the Euclidean distance. If you have two variables X_1 and X_2 measured on a sample of n subjects, the observed data for subject i can be denoted by x_{1i} and x_{2i} , and the observed data for subject j by x_{1j} and x_{2j} .



The Euclidean distance between two subjects is given by calculating the length of the hypotenuse of the triangle ABC using Pythagoras theorem (i.e. the distance between the two points i and j is calculated using their coordinates to indicate rise B and run A):

$$D_{ij} = \sqrt{A^2 + B^2} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2}$$

Alternative measures are the squared Euclidean distance D_2 :

$$D_{2ij} = A^2 + B^2 = (x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2$$

or the city block distance (D_3) (the long way round the block, i.e. the sum of A and B):

$$D_{3ij} = |A| + |B| = |x_{1i} - x_{1j}| + |x_{2i} - x_{2j}|$$

The distance measures can be extended to more than two variables. If you have p variables X_1, X_2, \dots, X_p measured on a sample of n subjects, the observed data for subject i can be denoted by $x_{1i}, x_{2i}, \dots, x_{pi}$ and the observed data for subject j by $x_{1j}, x_{2j}, \dots, x_{pj}$. and the Euclidean distance is:

$$D_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \dots + (x_{pi} - x_{pj})^2}$$

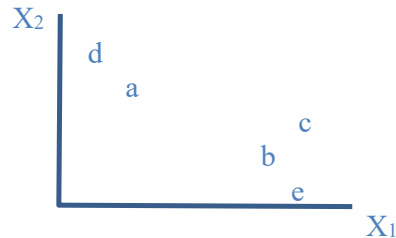
When using Euclidean distance, the scale of measurement of the variables under consideration is an issue, as changing the scale will obviously effect the distance between subjects. If one variable has a much wider range than others, then this variable will tend to dominate. To get around this problem, each variable can be standardised (converted to z-scores). However, this in itself presents a problem as it tends to reduce the variability (distance) between clusters. Many textbooks recommend standardisation, although if in doubt, one strategy would be to carry out the cluster analysis twice — once without standardising and once with — to see how much difference this makes to the resulting clusters.

Once the cluster analysis has been carried out, it is necessary to select the 'best' cluster solution. There are a number of ways in which this can be done, some rather informal and subjective, and some more formal. When carrying out a hierarchical cluster analysis, the process can be represented on a diagram known as a dendrogram. This diagram illustrates which clusters have been joined at each stage of the analysis and the distance between clusters at the time of joining. If there is a large jump in the distance between clusters from one stage to another then this suggests that at one stage clusters that are relatively close together were joined whereas, at the following stage, the clusters that were joined were relatively far apart. This implies that the optimum number of clusters may be the number present just before that large jump in distance.

The following example is used to show how to carry out variations of cluster analyses. For this example, take a set of data where observations are made on 2 variables for 5 individuals:

Host	X1	X2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

It can be plotted as:



The Euclidean distances between pairs are calculated and tabulated in a difference matrix ($n \times n = 5 \times 5$); e.g. the distance between a and b = $\sqrt{[(2-8)^2 + (4-2)^2]} = \sqrt{(36+4)} = 6.325$; etc.

Cluster	a	b	c	d	e
a	0	6.325	7.071	1.414	7.159
b		0	1.414	7.616	1.118
c			0	8.246	2.062
d				0	8.500
e					0

There are two main approaches to cluster analysis:

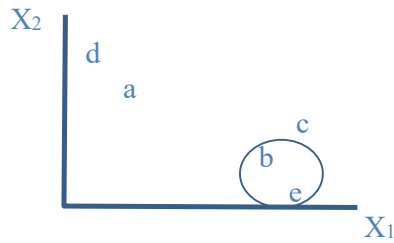
- Hierarchical methods (step-wise)
 - Hierarchical agglomerative methods in which subjects start in their own separate cluster. The two 'closest' (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. It is a hierarchical agglomerative method because all clusters formed consist of mergers of previously formed clusters. At the end, the optimum number of clusters is then chosen out of all cluster solutions. Within this approach to cluster analysis there are a number of different methods used to determine which clusters should be joined at each stage:
 - Nearest neighbour method (single linkage method). In this method the distance between two clusters is defined to be the distance between the two closest members, or neighbours. This method is relatively simple but is often criticised because it doesn't take account of cluster structure and can result in a problem

called chaining whereby clusters end up being long and straggly. However, it is better than the other methods when the natural clusters are not spherical or elliptical in shape.

The smallest distances between pairs is selected to form the first cluster (here comprising b and e):

Cluster	a	b	c	d	e
a	0	6.325	7.071	1.414	7.159
b		0	1.414	7.616	1.118
c			0	8.246	2.062
d				0	8.500
e					0

The cluster can be indicated as:

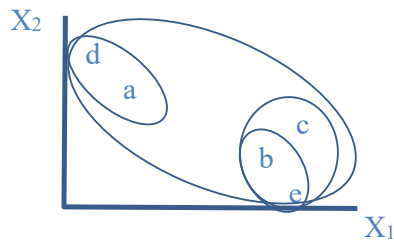


The difference matrix is then collapsed to a (n-1) x (n-1) matrix where b and e are clustered together. In the nearest neighbour method, the smallest distance between cluster (be) and other observations is chosen:

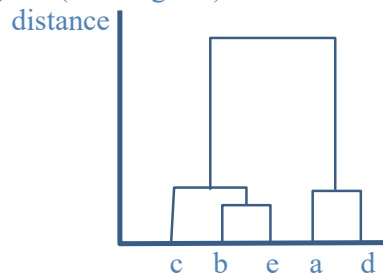
e.g. $D(be,a) = \min\{D(b,a), D(e,a)\} = \min\{6.325, 7.159\} = 6.325$; etc

Cluster	be	a	c	d
be	0	6.325	1.414	7.616
a		0	7.071	1.414
c			0	8.246
d				0

The next smallest distance is selected to indicate the next cluster, in this case (ad). These steps are repeated reiteratively, until everything is collapsed into one cluster.



The Euclidean distances between observations are also used to construct a tree diagram (dendrogram)



- Furthest neighbour method (complete linkage method). In this hierarchical method the distance between two clusters is defined to be the maximum distance between members — i.e. the distance between the two subjects that are furthest apart. This method tends to produce compact clusters of similar size but, as for the nearest neighbour method, does not take account of cluster structure. It is also quite sensitive to outliers.

$$\text{In this case, } D(b,c) = \max\{D(b,a), D(c,a)\}$$

- Average linkage method (sometimes referred to as UPGMA). The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. This hierarchical method is considered to be a fairly robust.

$$\text{In this case, } D(b,c) = \text{average}\{[D(b,a)+D(c,a)]/2\}$$

An SPSS program can be used to run the average linkage method based on the Euclidean dissimilarity coefficient matrix.

- Centroid method. Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged. This method is also fairly robust.
 - Ward's method. In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.
- Hierarchical divisive methods in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster. Agglomerative methods are used more often than divisive methods.

The following SPSS program can be used for hierarchical cluster analysis:

- Analyze
- Classify
- Hierarchical cluster
- Select the variables you want the cluster analysis to be based on and move them into the Variable(s) box.
- In the Method window select the clustering method you want to use. Under Measure select the distance measure you want to use and, under Transform values, specify whether you want all variables to be standardised (e.g. to z-scores) or not.
- In the Statistics window you can specify whether you want to see the Proximity Matrix (this will give the distance between all observations in the data set — only really recommended for relatively small data sets!). You can also specify whether you want the output to include details of cluster membership — either for a fixed number of clusters or for a range of cluster solutions (e.g. 2 to 5 clusters).
- In the Save window you can specify whether you want SPSS to save details of cluster membership — again, either for a fixed number of clusters or for a range of cluster solutions (e.g. 2 to 5 clusters). If you ask it to do this, this information will be included as additional variables at the end of the data set.
- In the Plots window you can specify which plots you would like included in the output.
- OK

Non-hierarchical methods (notable example being the k-means clustering method, often employed for quick clustering by some statistical programs). In these methods the desired number of clusters is specified in advance and the 'best' solution is chosen. Non-hierarchical cluster analysis tends to be used when large data sets are involved. It is sometimes preferred because it allows subjects to move from one cluster to another (this is not possible in hierarchical cluster analysis where a subject, once assigned, cannot move to a different cluster). Two disadvantages of non-hierarchical cluster analysis are that it is often difficult to know how many clusters you are likely to have and therefore the analysis may have to be repeated several times, and that it can be very sensitive to the choice of initial cluster centres. The steps in such a method are as follows:

- Choose initial cluster centres (essentially this is a set of observations that are far apart — each subject forms a cluster of one and its centre is the value of the variables for that subject).
- Assign each subject to its 'nearest' cluster, defined in terms of the distance to the centroid.
- Find the centroids of the clusters that have been formed.
- Re-calculate the distance from each subject to each centroid and move observations that are not in the cluster that they are closest to.
- Continue until the centroids remain relatively stable.

The following SPSS program can be used for K-means cluster analysis

- Analyze
- Classify
- K-means cluster
- Select the variables you want the cluster analysis to be based on and move them into the Variable(s) box.
- Under Method, ensure that Iterate and Classify is selected (this is the default).
- In the Iterate window you can specify how many iterations you would like SPSS to perform before stopping. The default is ten. It might be worth leaving it as ten to start with and then increasing this if convergence doesn't occur (i.e. a stable cluster solution is not reached) within ten iterations.
- In the Save window you can specify whether you want SPSS to save details of cluster membership and distance to the cluster centre for each subject (observation).
- OK

XIV. Within-host parasite community interaction network (Pedersen & Fenton, 2006)

Parasite community ecology has been highly descriptive, driven by pattern-based analyses at the host population level. Broadly, two main approaches have been adopted to examine parasite communities, although these are not mutually exclusive. The first classifies parasite communities based on patterns of species occurrence (presence and absence data) and tests for community structuring by comparing observed species distributions against null models. The second approach quantifies pairwise associations between species, inferring interspecific interactions from correlations in species abundance or more complex models that control for biotic and abiotic factors. However, although these approaches provide a basic description of parasite communities at the host population level, they provide little mechanistic insight into the within-host processes shaping these patterns.

To obtain a mechanistic understanding of parasite communities, we need to consider the network of interactions (both direct and indirect) that occurs between parasite species within an individual host. The most common interaction networks in community ecology are food webs, which incorporate explicit trophic structure and directionality such that primary producers (basal level) are consumed by species at the intermediate level, which are in turn, consumed by predators higher up the network. This can be illustrated with a hypothetical within-host parasite network comprising three trophic levels: host resources, the parasite community and the host immune system.

- Level 1: host resources: The basal level is defined by host resources, which can be a specific component that parasites feed on (e.g. blood), or the physical space available (e.g. within the gastrointestinal tract). Parasite feeding or growth depletes resources and debilitates the host, indirectly affecting other parasites within the community.
- Level 2: the parasite community. The second level includes all parasites (both micro- and macro-parasites) that infect the host. Where possible, parasite species should be placed into functional guilds of similar species. Defining guilds can be controversial but we suggest they should be based on functional similarity of species rather than on taxonomic classifications. In particular, parasite guilds can be defined in terms of a shared niche, where species differentiate themselves along three major axes:
 - a resource axis (e.g. what resources do the parasites feed on?);
 - a location axis (e.g. where do the parasites occur within the host?); and
 - an immunological axis (e.g. what components of the immune response of the host do the parasites stimulate?).

The location of a parasite along each of these axes defines its niche and parasites occupying similar niches (i.e. occupying similar locations, consuming similar resources and stimulating comparable host immune responses) can be placed in the same guild. However, there is a degree of subjectivity in the definition of guilds and it should be seen as a simplifying approach. Frequently, individual species will occupy their own unique guild.

- Level 3: host immune system. The third level comprises the immune system of the host, which is analogous to a predator trophic level in community ecology food webs. This predator–prey analogy of host immunity–parasites is frequently adopted for modelling the within-host dynamics of single pathogen species, where the immune response ‘consumes’ the pathogen. This trophic level can be divided into different components of the immune system (i.e. cellular response, humoral response and T-helper cell types), akin to a suite of generalist and specialist predators, with potential interactions between them.

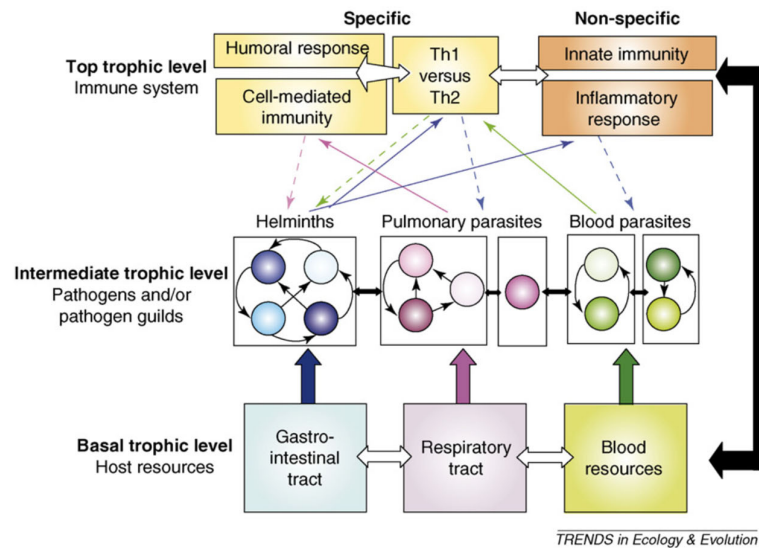


Figure. A hypothetical within-host parasite community interaction network defined with three levels of explicit trophic structure; given that parasites consume resources of the host for development, reproduction and transmission, and the immune system acts as a predator destroying the infecting pathogens. The basal level is defined by the host resources, analogous to the primary producers in a typical free-living food web. However, by contrast, host resources are inextricably linked to each other (white arrows) because the fitness and survival of the host depends on all resource components. The intermediate level comprises the parasites (colored circles) and parasite guilds that infect the host. Pathogens that consume similar resources, share a locality within the host and are attacked by the same components of the immune system can be considered parasite guilds (boxes), in which direct interactions between parasites are most probable (unidirectional arrows). Parasite guilds can comprise a single species. The vertical arrows represent the flux of energy from host to pathogen. The top trophic level represents the diverse responses of the immune system that vary in their degree of specificity. Here, we highlight a few common components (boxes), and use solid colored arrows to represent the aspects of the immune system that target each parasite or parasite guild, whereas the dashed arrows represent the top-down indirect interactions of co-infection parasites, mediated by the immune system.

Parasites can substantially affect host populations and community structure by influencing host mortality, fecundity, growth, nutritional status, energetic requirements, and behavior. Such host–parasite interactions may shape components of an ecological community other than the host population, particularly if the host is abundant or ecologically influential. For example, parasites may weaken competitively dominant hosts, altering the outcome of competition between the host and its competitors. Parasites are also known to alter rates of predation, and hence, the feeding ecology of predators and population dynamics of prey.

The traditional method for detecting community structure in biological networks is hierarchical clustering. One first calculates a weight W_{ij} for every pair i, j of vertices in the network, which represents in some sense how closely connected the vertices are. Then one takes the n vertices in the network, with no edges between them, and adds edges between pairs one by one in order of their weights, starting with the pair with the strongest weight and progressing to the weakest.

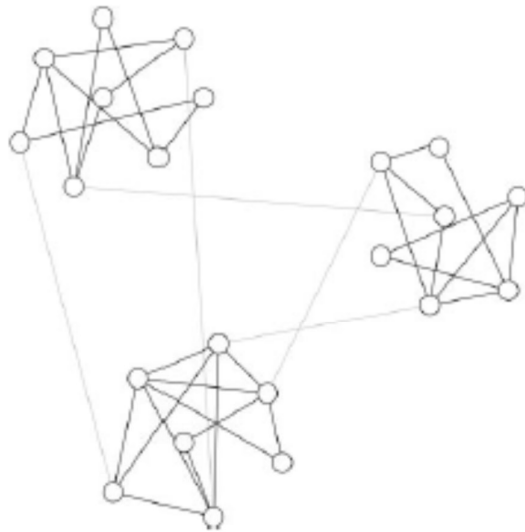


Fig. 1. A schematic representation of a network with community structure. In this network there are three communities of densely connected vertices (circles with solid lines), with a much lower density of connections (gray lines) between them.

As edges are added, the resulting graph shows a nested set of increasingly large components (connected subsets of vertices), which are taken to be the communities. Because the components are properly nested, they all can be represented by using a tree of the type shown in Fig. 2, in which the lowest level at which two vertices are connected represents the strength of the edge that resulted in their first becoming members of the same community.

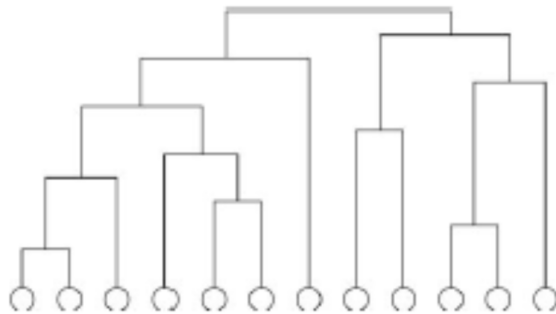


Fig. 2. An example of a small hierarchical clustering tree. The circles at the bottom represent the vertices in the network, and the tree shows the order in which they join together to form communities for a given definition of the weight W_{ij} of connections between vertex pairs.

A “slice” through this tree at any level gives the communities that existed just before an edge of the corresponding weight was added. Trees of this type are sometimes called dendrograms in the sociological literature. Many different weights have been proposed for use with hierarchical clustering algorithms.

- One possible definition of the weight is the number of node-independent paths between vertices. Two paths that connect the same pair of vertices are said to be node-independent if they share none of the same vertices other than their initial and final vertices. One can similarly also count edge-independent paths. It is known that the number of node-independent (edge-independent) paths between two vertices i and j in a graph is equal to the minimum number of vertices (edges) that must be removed from the graph to disconnect i and j from one another. Thus these numbers are in a sense a measure of the robustness of the network to deletion of nodes (edges). Numbers of

independent paths can be computed quickly by using polynomial-time “max-flow” algorithms such as the augmenting path algorithm.

- Another possible way to define weights between vertices is to count the total number of paths that run between them (all paths, not just those that are node- or edge-independent). However, because the number of paths between any two vertices is infinite (unless it is zero), one typically weights paths of length L by a factor α^L with α small, so that the weighted count of the number of paths converges. Thus long paths contribute exponentially less weight than those that are short. If \mathbf{A} is the adjacency matrix of the network, such that A_{ij} is 1 if there is an edge between vertices i and j and 0 otherwise, then the weights in this definition are given by the elements of the matrix:

$$W = \sum_{L=0}^{\infty} (\alpha A)^L = [1 - \alpha A]^{-1}$$

For the sum to converge, we must choose α smaller than the reciprocal of the largest eigenvalue of \mathbf{A} .

Both of these definitions of the weights give reasonable results for community structure in some cases. In other cases they are less successful. In particular, both have a tendency to separate single peripheral vertices from the communities to which they should rightly belong. If a vertex is, for example, connected to the rest of a network by only a single edge then, to the extent that it belongs to any community, it should clearly be considered to belong to the community at the other end of that edge. Unfortunately, both the numbers of independent paths and the weighted path counts for such vertices are small and hence single nodes often remain isolated from the network when the communities are constructed. This and other pathologies, along with poor results from these methods in some networks where the community structure is well known from other studies, make the hierarchical clustering method, although useful, far from perfect.

Edge “Betweenness” and Community Structure. To sidestep the shortcomings of the hierarchical clustering method, we here propose an alternative approach to the detection of communities. Instead of trying to construct a measure that tells us which edges are most central to communities, we focus instead on those edges that are least central, the edges that are most “between” communities. Rather than constructing communities by adding the strongest edges to an initially empty vertex set, we construct them by progressively removing edges from the original graph. Vertex betweenness has been studied in the past as a measure of the centrality and influence of nodes in networks.

The betweenness centrality of a vertex i is defined as the number of shortest paths between pairs of other vertices that run through i . It is a measure of the influence of a node over the flow of information between other nodes, especially in cases where information flow over a network primarily follows the shortest available path. To find which edges in a network are most between other pairs of vertices, we generalize Freeman’s betweenness centrality to edges and define the edge betweenness of an edge as the number of shortest paths between pairs of vertices that run along it. If there is more than one shortest path between a pair of vertices, each path is given equal weight such that the total weight of all of the paths is unity. If a network contains communities or groups that are only loosely connected by a few intergroup edges, then all shortest paths between different communities must go along one of these few edges. Thus, the edges connecting communities will have high edge betweenness. By removing these edges, we separate groups from one another and so reveal the underlying community structure of the graph.

The algorithm we propose for identifying communities is simply stated as follows:

1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.

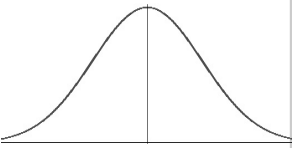
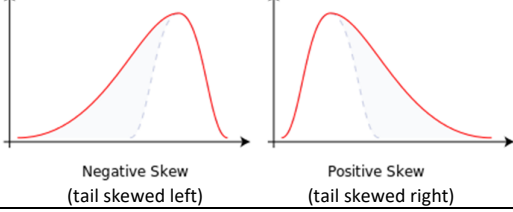
3. Recalculate betweennesses for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.

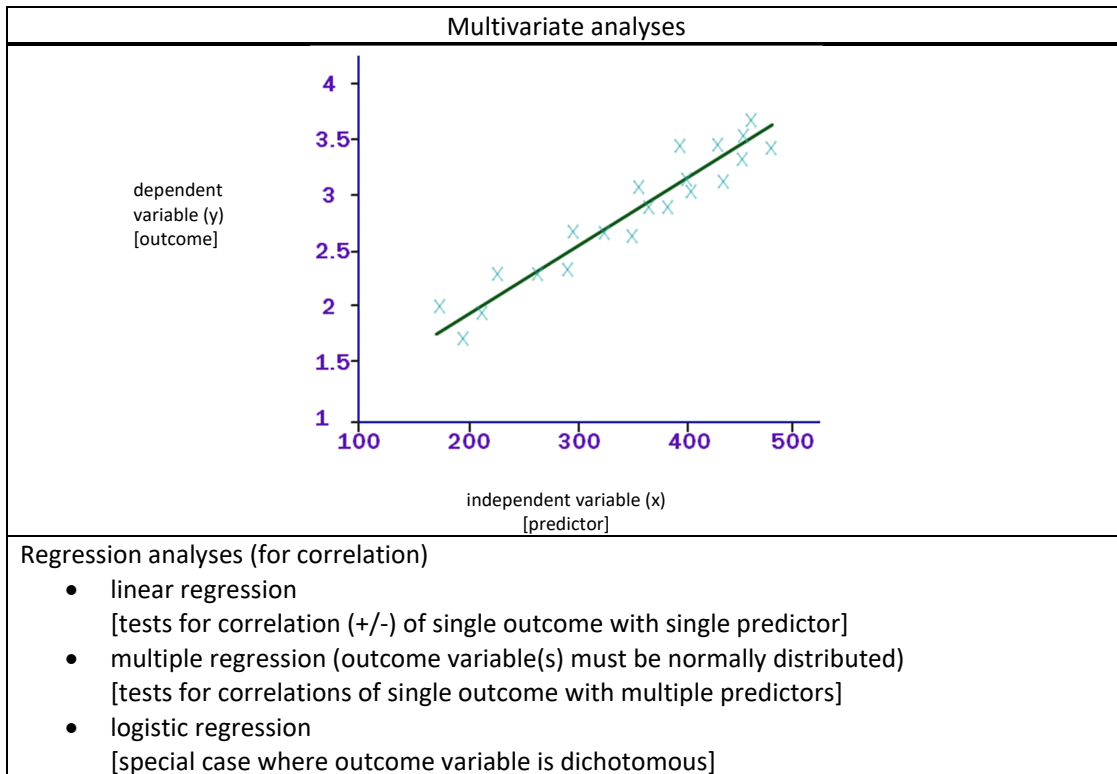
As a practical matter, we calculate the betweennesses by using the fast algorithm of Newman, which calculates betweenness for all m edges in a graph of n vertices in time $O(mn)$. Because this calculation has to be repeated once for the removal of each edge, the entire algorithm runs in worst-case time $O(m^2n)$. However, after the removal of each edge, we only have to recalculate the betweennesses of those edges that were affected by the removal, which is at most only those in the same component as the removed edge. This means that running time may be better than worst-case for networks with strong community structure (those that rapidly break up into separate components after the first few iterations of the algorithm).

XV. Statistical Tests

DATA SETS	
Categorical data (qualitative)	<ul style="list-style-type: none"> dichotomous (binary): e.g. presence/absence nominal (no order): e.g. red/blue/green ordinal (ordered): e.g. weak/medium/strong
Quantitative data	<ul style="list-style-type: none"> counts (whole numbers): e.g. 1,2,3... continuous (any value): e.g. 21.35, 23.67...

SIGNIFICANCE	abbreviation	probability (p) values	percentage
not significant	ns	$p > 0.05$	< 95%
significant	*	$0.01 < p < 0.05$	95-99%
significant	**	$0.001 < p < 0.01$	99-99.9%
highly significant	***	$p < 0.001$	>99.9%

STATISTICAL TESTS	
NORMAL DISTRIBUTION	SKEWED DISTRIBUTION
	
measures of spread: <ul style="list-style-type: none"> mean (average) standard deviation 95% confidence intervals [range = mean \pm 1.96 SD] [95% of observations clustered within 1.96 SD of mean] 	measures of spread: <ul style="list-style-type: none"> median (middle value) interquartile range [25-75% quartiles] [box-whisker plots]
Parametric statistical tests <ul style="list-style-type: none"> t-tests (one sample, paired sample, two sample) [compares means and variance] analysis of variance (ANOVA) F-value [compares variance within groups to variance between groups] 	Non-parametric statistical tests <ul style="list-style-type: none"> Wilcoxon Rank Sum test (= Mann-Whitney U test) [compares sums of ranked data] Kruskal-Wallis test (= Kruskal-Wallis one-way ANOVA) [compares ranks]



COMPARISON	Data normally distributed	Data not normally distributed
sample group v. population	one-sample t-test	Wilcoxon's Signed Rank test (1)
matched pairs (categorical data)	paired t-test	Wilcoxon's Signed Rank test (2)
two populations	two means; two-sample t-test*	two medians; Wilcoxon Rank Sum test two proportions >5/cell: Chi-squared test <5/cell: Fischers exact test
more than two populations	means; ANOVA*	medians; Kruskal-Wallis test
predictor(s) v. continuous outcome	linear regression multiple regression*	GEE with transformation
predictor v. dichotomous outcome	logistic regression	

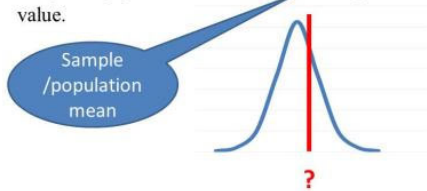
*all general linear models (GLM)

Power and sample size	
Type I errors [accepting true when actually false] [= false positive]	Type II errors [accepting false when actually true] [= false negative]
usually set α at 0.05 or 0.01	usually set β at 0.2
	giving power of study = $(1-\beta) = 0.8$ (80%)

Visual guide

One sample t-test

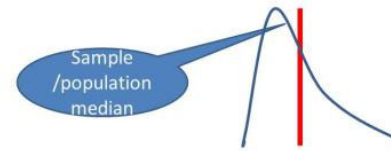
Used where it is required to test whether the mean of a sample or population is different from a particular value.



- ❖ data must be normally distributed
- ❖ one group

Wilcoxon's Signed Rank Test (1)

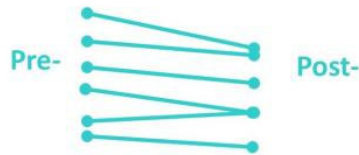
Used to test the difference between a sample/population median and expected median.



- ❖ data need not be normally distributed
- ❖ non-parametric equivalent of one sample t-test

Paired t-test

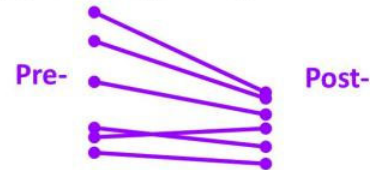
Used to test equality of the means of two samples/populations, when the observations arise as paired samples.



- ❖ differences between the pairs must be normally distributed
- ❖ two groups

Wilcoxon's Signed Rank Test (2)

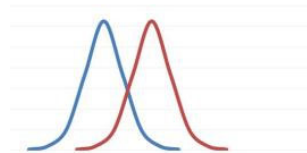
Used to test the difference between two samples/populations using matched pairs.



- ❖ data need not be normally distributed
- ❖ two groups which are paired
- ❖ non-parametric equivalent of paired t-test

Two sample t-test

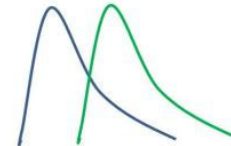
Used to test equality of the means of two populations.



- ❖ data must be normally distributed
- ❖ two groups

Wilcoxon's Rank Sum Test (Mann-Whitney U test)

Used to test equality of the medians of two samples/populations.



- ❖ data need not be normally distributed
- ❖ two groups
- ❖ non-parametric counterpart of two sample t-test

Analysis of Variance (ANOVA)

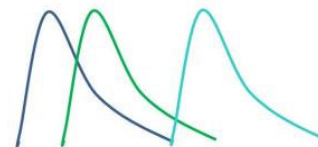
Used to test equality of the means of two or more populations.



- ❖ data must be normally distributed
- ❖ groups must have equal variance (homoscedasticity)
- ❖ two or more groups

Kruskal-Wallis test

Used to test equality of the medians of two or more samples/populations.



- ❖ data need not be normally distributed
- ❖ two or more groups

Chi-square test (χ^2 test)

Used to assess the independence of the two variables forming a contingency tables with r rows and c columns.
The chi-squared statistic is derived from the difference between observed and expected frequencies of cells in the table if under independence.

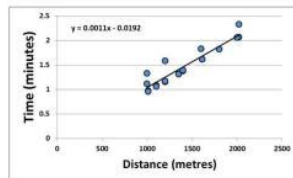
	Disease + Lung Cancer	Disease - No Lung Cancer	
Exposure + Smoker	80	20	100
Exposure - Non-smoker	10	90	100
	90	110	

80% vs 10%

- comparing two or more proportions (eg 80% vs 10% have lung cancer)
- comparing a sample proportion to an expected proportion (goodness of fit test)(eg if 40% babies born this month are males is this really different to 50% males we expected?)

Multiple Regression

A form of regression analysis where the outcome variable is a continuous variable.



Time = race distance + sex

- ❖ One or more explanatory variables (eg risk/protective factors) can be included in analysis
- ❖ Explanatory variables can be continuous or categorical
- ❖ Outcome variable must be normally distributed

Logistic Regression

A form of regression analysis where the outcome variable is a binary (dichotomous – eg yes/no sick/well) variable.

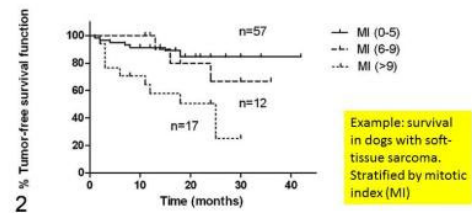
example

(Odds) Heart disease (yes/no) = age(years) + family hx (yes/no) + smoking (years) + diabetes

- ❖ Many explanatory variables (eg risk/protective factors) can be included in analysis
- ❖ Outcome of analysis is expressed as **Odds Ratio**, can be converted to **Probability**
- ❖ Explanatory variables can be continuous or categorical
- ❖ **Most common analysis method in the biomedical sciences**

Kaplan-Meier curve with Log-Rank test

Used to describe and compare survival times.



- ❖ accounts for individuals who drop out (censored observations)
- ❖ one or more groups
- ❖ Log-Rank test tests whether survival is different between two or more groups

Cox Proportional Hazards Regression

Used to describe and compare survival times where there are one or more explanatory variables.

Example:

Death from heart disease (yes/no) = age(years) + family hx (yes/no) + smoking (years) + diabetes (years)

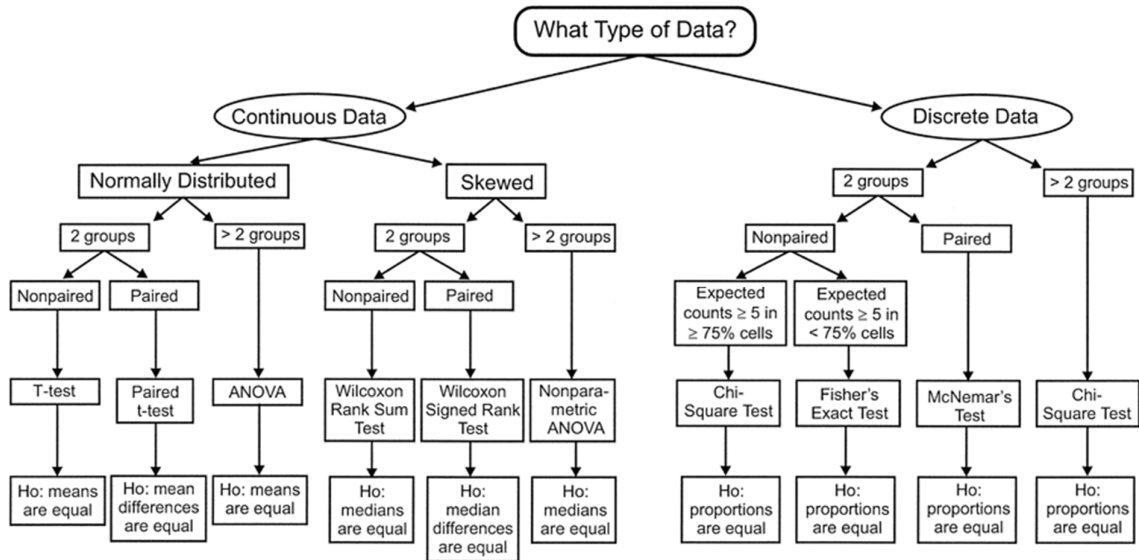
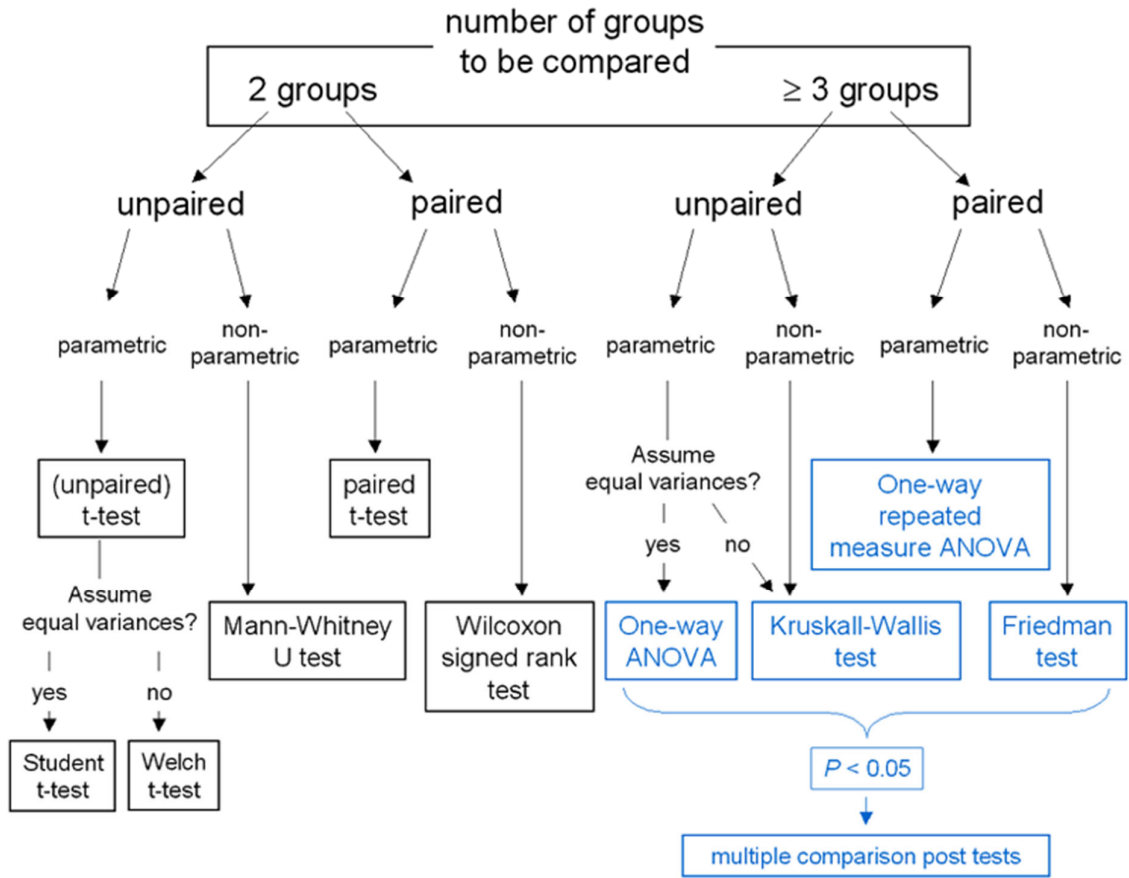
- ❖ a bit like logistic regression but **uses additional data** because it takes survival time into account
- ❖ outcome of analysis is expressed as a **Hazard Ratio**, which is interpreted like an Odds Ratio

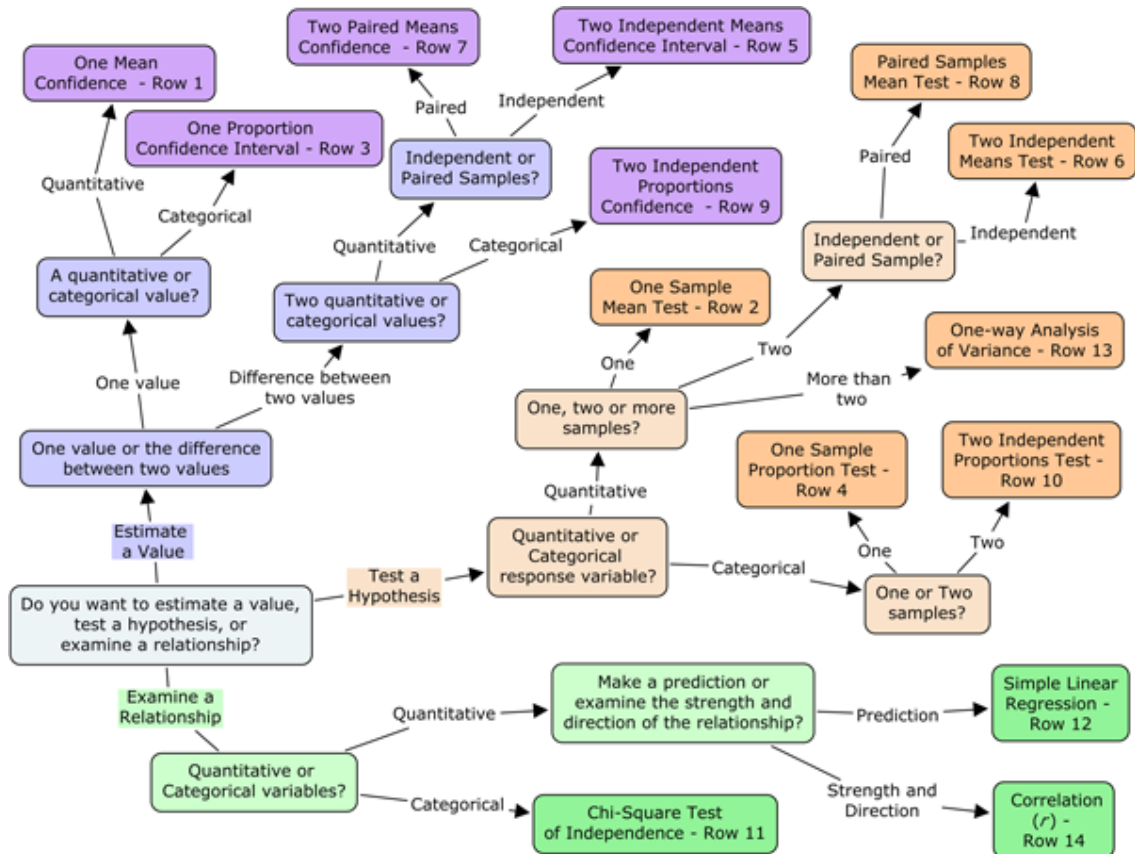
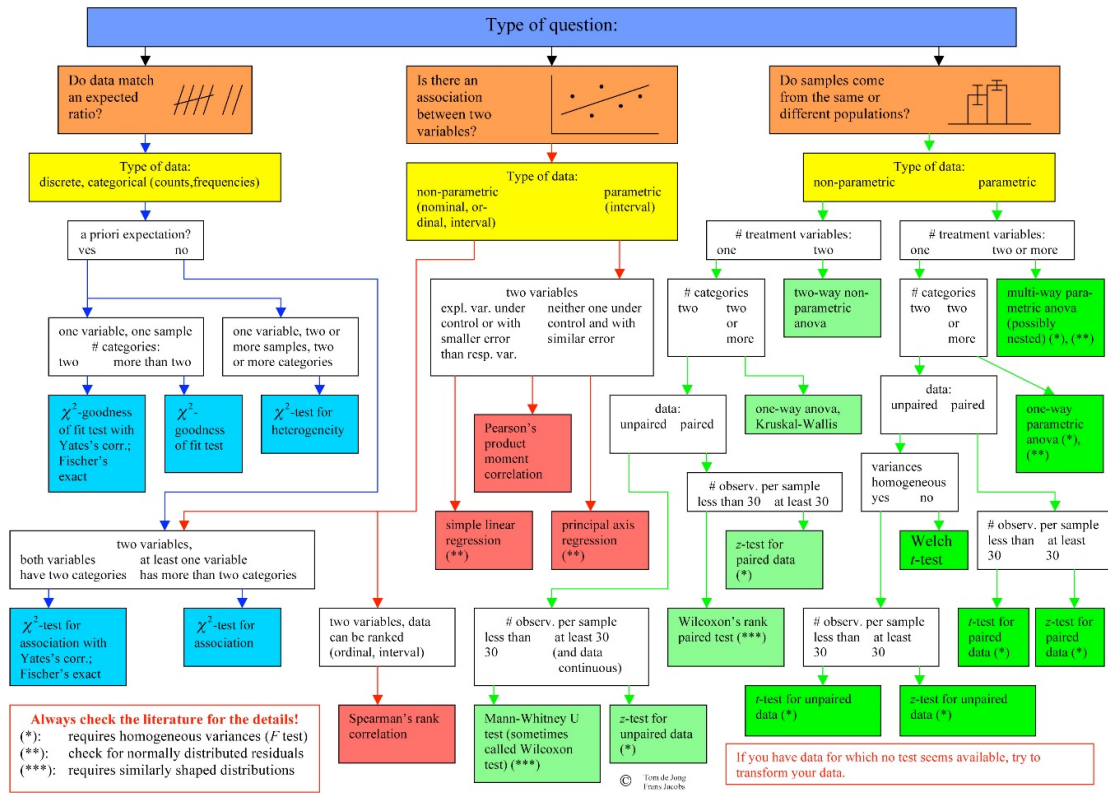
Summary Table of Statistical Tests

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	χ^2 or binomial	χ^2	Macnarmar's χ^2	χ^2	Cochran's Q	
Rank or Ordinal		Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friendman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r
		Factorial (2 way) ANOVA				

Summary Table of Statistical Tests

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	χ^2 or binomial	χ^2	Macnarmar's χ^2	χ^2	Cochran's Q	
Rank or Ordinal	χ^2	Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r
		Factorial (2 way) ANOVA				



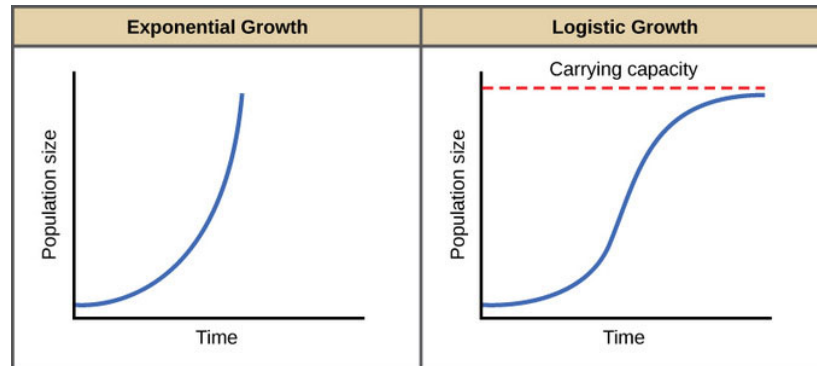


XVI. r/K selection theory [reproduction-development trade-off (quantity versus quality)]

The fitness of a species can be fundamentally enhanced by two different strategies: fast reproduction (multiplication, replication) or prolonged development (long-life, stability). These strategies are not independent, but form a continuum from one extreme to another for different species. How much one species invests in one strategy over the other depends on the selective environment, and in biology is called r-K selection. The mathematical principles are derived from logistic (constrained) growth model (Verhulst equation) of population biology which indicates the traits that favour either quantity or quality of offspring in a species (MacArthur & Wilson, 1967). That is, r-selected species invest in reproduction (quantity) while K-selected species invest in prolonged development and long-life (quality).

$$\frac{dN}{dt} = N' = rN \left[1 - \left(\frac{N}{K} \right) \right]$$

where r = growth rate (reproduction) (quantity)
 and K = carrying capacity (development) (quality)
 N = population size
 t = time



r-selected species	K-selected species
population grows exponentially (abundant resources) but never reaches carrying capacity (predators, droughts, etc)	population approaching carrying capacity, slow growth due to competition for limited resources
colonizers (opportunists)	competitors
unstable environments	stable environments
exploit less-crowded ecological niches, produce many offspring, each of which has a relatively low probability of surviving to adulthood	strong competitors in crowded niches, invest heavily in fewer offspring, each of which has a relatively high probability of surviving to adulthood
small organisms	large organisms
short-lived	long-lived
weak competitors	strong competitors
numerous offspring	few offspring
little parental care	significant parental care
fast maturation	slow maturation
rapid dispersal	slow dispersal
broad range	territorial
bacteria	terns
diatoms	whales
insects	elephants
weeds	trees
rodents	humans

Topics to cover in Discussion

ubiquity

- occurrence
- prevalence
- intensity
- abundance

biodiversity

- species richness
- relative abundance

host specificity

- clusters

species identification

- morphotypic characters
- novelty

taxonomic characters

- groupings (classification)
- relationships (phylogeny)

dietary specialization

- flagellate diet
- host diet
- symbiotic bacteria

distribution

- spatial (geographic)
- temporal (seasonal)

want to analyse parasite population structure in:

- one individual termite [infra-population]
 - qualitative (presence/absence)
 - quantitative (number = intensity)
- ten individual termites of same species from one colony (inter-termite variation but intra-colony variation)
 - quantitative (prevalence)
 - quantitative (abundance)
- termites of same species from different colonies (inter-colony variation)
- termites of different species from different colonies (inter-species variation)
 - host specificity

want to analyse parasite dynamics for correlations with:

- host castes
- host habitat (type of nest)
- host diet
- host phylogeny
- host biogeography

want to identify differences in morphology of flagellate species with respect to:

- type of food (wood, starch....)
- presence of hydrogenosomes
- presence of symbiotic bacteria (ectosymbiotic, endosymbiotic cytoplasmic/nuclear)

Previous observations

numbers of protozoa in lower termites: workers > soldiers > alates

trichomonads (do not contain wood particles) in drywood termites (eat sound wood)

hypermastigids (contain wood particles) in subterranean termites (eat degrading wood)

parabasalids (amitochondriate, with hydrogenosomes, few endosymbiotic bacteria)

metamonads (amitochondriate, no hydrogenosomes, numerous symbiotic bacteria (ecto + endo))

	hypermastigids	trichomonads	oxymonads
rhinotermitids (degrading wood)	+++	++	+
kalotermitids (sound wood)	±	+++	+
mastrotermitids			
termopsids			
higher termites (humus, grass, litter)			
hydrogenosomes	numerous	few	none
ectosymbiotic bacteria	none	none	numerous
endosymbiotic bacteria (cytoplasmic)	many	few	many
endosymbiotic bacteria (nuclear)	numerous	none	none

inverse relationship between presence of hydrogenosomes and symbiotic bacteria

XLSTAT (software for Excel)

LIST OF TUTORIALS

Statistical guides

- Which statistical method should you choose and other learning resources
- What is a statistical test?
- Which statistical test should you use?
- What is the difference between a two-tailed and a one-tailed test?
- What is the difference between paired and independent samples tests?
- What is the difference between a parametric and a nonparametric test?
- What is statistical modeling?
- Which statistical model should you choose?
- Which multivariate data analysis method to choose?
- Which descriptive statistics tool should you choose?

Managing data

- Aggregating data in Excel
- Filtering observations in Excel
- Stack / unstack data in Excel
- Merging tables in Excel
- Removing duplicates in Excel

Preparing data

- Coding and recoding data in Excel
- Discretizing a continuous variable in Excel
- Box-Cox transformation tutorial in Excel
- Missing data imputation using NIPALS in Excel
- Cross-tab or contingency table in Excel
- Create a disjunctive table in Excel
- Raking a survey sample, tutorial in Excel
- Stratified data sampling tutorial in Excel

Describing data

- Mean, median, standard deviation & more in Excel
- Skewness and Kurtosis in Excel
- Frequencies, mode & bar charts in Excel
- Distribution sampling & normality tests in Excel
- Variables characterization tutorial in Excel
- Quantiles or percentiles computation in Excel
- Bootstrap statistics tutorial in Excel
- Biserial correlations tutorial in Excel
- Create an intelligent pivot table in Excel
- Reliability analysis in Excel

Visualizing data

- Scatter plot tutorial in Excel
- Scatter plot with confidence ellipses in Excel
- Box plot tutorial in Excel
- Notched box plots tutorial in Excel
- Histograms and distribution fitting tutorial in Excel
- Dynamic histograms tutorial in Excel
- Parallel Coordinates Visualization in Excel tutorial
- Ternary diagram in Excel tutorial
- Chart with error bars in Excel with just two clicks
- Motion charts tutorial in Excel
- Adding a curve on an Excel chart

- Creating and customizing a plot
- Customize a PCA chart for an easier interpretation

Analyzing data

- Principal component analysis (PCA) in Excel
- Correspondence Analysis (CA) from a contingency table
- Correspondence Analysis from raw data with 3D charts
- Multiple Correspondence Analysis (MCA) in Excel
- Principal Coordinate Analysis in Excel tutorial
- Multidimensional Scaling (MDS) in Excel tutorial
- Factor analysis in Excel tutorial
- Discriminant Analysis in Excel tutorial
- Agglomerative Hierarchical Clustering (AHC) in Excel
- k-means clustering in Excel tutorial
- Clustering big datasets using k-means then AHC
- Gaussian mixture model clustering in Excel tutorial
- Filtering observations and variables in PCA charts
- Filtering observations within a PCA

Modeling data

- Fitting a distribution to a sample of data in Excel
- Simple linear regression in Excel tutorial
- Multiple Linear Regression in Excel tutorial
- One-way ANOVA & multiple comparisons in Excel tutorial
- Contrast analysis after a one-way ANOVA in Excel
- Two-way unbalanced ANOVA with interactions in Excel
- Pairwise multiple comparisons after a multi-way ANOVA
- What is the difference between LS Means and Observed Means?
- How to interpret contradictory results between ANOVA and multiple pairwise comparisons?
- ANCOVA analysis in Excel tutorial
- Running a logistic regression with XLSTAT
- Ordinal logit model in Excel tutorial
- Multinomial logit model in Excel tutorial
- Log-linear regression (Poisson) in Excel tutorial
- Quantile regression in Excel tutorial
- Cubic spline in Excel tutorial
- Nonparametric regression (kernel & Lowess) tutorial
- Nonlinear regression in Excel tutorial
- Nonlinear multiple regression in Excel tutorial
- Partial Least Squares PLS regression in Excel
- Partial least squares discriminant analysis PLSDA tutorial
- Repeated measures ANOVA in Excel tutorial
- Run repeated measures ANOVA using mixed models
- Random components mixed model in Excel tutorial
- Two-stage least squares regression 2SLS in Excel

Machine Learning

- Classification tree in Excel tutorial
- Association rules for market basket analysis tutorial
- K Nearest Neighbors KNN in Excel tutorial
- Naive Bayes classification in Excel tutorial
- Training a Support Vector Machine (SVM) in Excel

Correlation/Association tests

- Spearman correlation coefficient in Excel tutorial
- RV coefficient test in Excel tutorial
- Run Chi-square and Fisher's exact tests in Excel

- Mantel test in Excel tutorial
- Cochran-Armitage trend test in Excel tutorial
- Pearson correlation coefficient in Excel

Parametric tests

- One proportion test in Excel tutorial
- Compare two proportions in Excel tutorial
- Compare k proportions in Excel tutorial
- Multinomial goodness of fit test in Excel tutorial
- One sample t-test or z-test in Excel tutorial
- Student's t test on two independent samples tutorial
- Student's t-test on two paired samples tutorial
- One sample variance test in Excel tutorial
- Fisher's F-test to compare two variances in Excel
- Levene and Bartlett tests on variances in Excel
- Compare ≥ 2 samples described by several variables
- Testing equivalence with TOST in Excel tutorial
- Two sample t-test using XLSTAT spreadsheet functions

Non parametric tests

- Mann-Whitney test in Excel tutorial
- Wilcoxon signed rank test in Excel tutorial
- Kruskal-Wallis test in Excel tutorial
- Friedman non parametric test in Excel tutorial
- Kolmogorov-Smirnov test in Excel tutorial
- Page non parametric test in Excel tutorial
- McNemar test in Excel tutorial
- Cochran's Q test in Excel tutorial
- Durbin, Skillings-Mack test in Excel tutorial
- Cochran-Mantel-Haenszel CMH test in Excel tutorial

Testing for outliers

- Grubbs test to detect outliers in Excel tutorial
- Dixon test to detect outliers in Excel tutorial
- Cochran C test to detect outlying variances tutorial
- Run Mandel's h and k statistics to detect outliers

Sensory data analysis

- Preference Mapping in Excel tutorial
- Semantic Differential Chart in Excel tutorial
- Penalty analysis in Excel tutorial
- Sensory product characterization in Excel tutorial
- TURF analysis in Excel tutorial
- Sensory panel analysis in Excel tutorial
- Bradley-Terry model in Excel tutorial
- Sensory shelf life analysis in Excel tutorial
- Run sensory discrimination triangle test in Excel
- CATA Check-All-That-Apply analysis tutorial in Excel
- Design an experiment for sensory analysis in Excel
- Multiple Factor Analysis (MFA) in Excel tutorial
- Temporal Dominance of Sensations (TDS) in Excel
- Sensory wheel tutorial in Excel
- Time-Intensity analysis in Excel

Conjoint analysis

- Conjoint analysis in Excel tutorial
- Choice Based Conjoint (CBC) in Excel tutorial
- MaxDiff analysis in Excel tutorial

- Choice based conjoint analysis with Hierarchical Bayes (CBC HB)
- Run Monotone regression / MONANOVA in Excel
- Conditional logit model tutorial in Excel

Time series analysis

- Using differencing to obtain a stationary time series
- Holt-Winters seasonal multiplicative model in Excel
- Fit an ARIMA model to a time series in Excel
- Spectral analysis in Excel tutorial
- Mann-Kendall trend test in Excel tutorial
- Time series homogeneity test in Excel tutorial
- Cochran-Orcutt estimation in Excel tutorial
- Durbin-Watson test in Excel tutorial
- Unit root (Dickey-Fuller) and stationarity tests on time series
- Cointegration test on time series in Excel tutorial

Monte Carlo simulations

- Running a simple simulation model with XLSTAT
- Simulation model with scenario variables tutorial
- Run a simulation model with correlations between distributions and compute SPC (process ca...)
- Generating many distributions in a simulation model by copying

Power analysis

- Sample size & statistical power of a mean comparison test tutorial
- Sample size & statistical power in a multiple regression tutorial
- Sample size for a clinical trial tutorial in Excel

Statistical Process Control

- Individual control chart in Excel tutorial
- Subgroup control chart in Excel tutorial
- Attribute control chart in Excel tutorial
- Pareto plot in Excel tutorial
- Time weighted control chart in Excel tutorial
- Gage R&R for quantitative data in Excel tutorial
- Gage R&R for Attributes in Excel tutorial

Design of Experiments

- Factor effect (screening) design in Excel tutorial
- Surface response design in Excel tutorial
- Mixture design in Excel tutorial

Survival analysis

- Life table analysis in Excel tutorial
- Kaplan-Meier survival analysis in Excel tutorial
- Cox proportional hazards model in Excel tutorial
- Sensitivity and specificity in Excel tutorial
- ROC curve analysis in Excel tutorial
- Nelson-Aalen analysis in Excel tutorial
- Cumulative Incidence analysis in Excel tutorial
- Weibull model in Excel tutorial
- Parametric survival curves analysis in Excel tutorial
- ROC curves comparison in Excel tutorial

Method validation

- Bland Altman plot to compare methods in Excel
- Run Passing Bablok regression to compare methods
- Run Deming regression to compare methods in Excel

Dose effect analysis

- Dose effect analysis in Excel tutorial

- Run 4 or 5-parameter logistic regression in Excel

OMICS data analysis

- Heat map (OMICS) in Excel tutorial
- Differential expression (OMICS) in Excel tutorial

Multiblock data analysis

- Run Generalized Procrustes Analysis (GPA) in Excel
- Canonical Correspondence Analysis (CCA) tutorial
- Canonical Correlation analysis in Excel tutorial
- Redundancy Analysis (RDA) in Excel tutorial

PLS Path modeling

- Create & run a basic PLS Path Modeling project
- Create & run a basic PLSPM Project in Excel 2003
- PLS Path Modeling in Excel: group comparison
- PLS Path Modeling in Excel: REBUS classification
- PLS Path Modeling in Excel: moderating effects
- Consumer satisfaction analysis in Excel with PLSPM

XLSTAT-LatentClass

- Latent Class Regression Model in Excel tutorial
- Latent Class Cluster Model in Excel tutorial

XLSTAT-3DPlot

- 3D plot in Excel tutorial
- Save a 3D model to reuse it later or on other data

XLSTAT 365

- Installing XLSTAT 365
- Distribution sampling & normality tests in Excel 365

XLSTAT-Base (the simplest version with the following 100 programs)

PREPARING DATA

- Data sampling
- Distribution Sampling
- Discretization
- Coding
- Coding by ranks
- Presence/Absence coding
- Missing data
- Complete disjunctive tables (Creating dummy variables)
- Create contingency tables
- Variable transformations
- Data management
- Raking survey data

DESCRIBING DATA

- Descriptive statistics (including Box plots and scattergrams)
- Histograms
- Reliability Analysis
- Normality tests
- Contingency table (descriptive statistics)
- Similarity/Dissimilarity matrices (correlation...)
- Multicollinearity statistics
- Quantiles estimation
- Resampled statistics
- Biserial correlation
- Variable characterization
- Pivot table

ANALYZING DATA

- Principal Component Analysis (PCA)
- Correspondence Analysis (CA)
- Multiple Correspondence Analysis (MCA)
- Principal Coordinate Analysis
- Multidimensional Scaling (MDS)
- Factor analysis
- Discriminant Analysis (DA)
- Agglomerative Hierarchical Clustering (AHC)
- k-means clustering
- Univariate clustering
- Gaussian mixture models

VISUALIZING DATA

- Scatter plots
- Histograms
- Parallel coordinates plots
- Ternary diagrams
- Error bars
- Semantic differential charts
- Plot a function
- Univariate plots
- Plot management

- Motion charts

MODELING DATA

- Distribution fitting
- Linear regression
- ANOVA (Analysis of variance)
- Welch and Brown-Forsythe one-way ANOVA
- ANCOVA (Analysis of Covariance)
- Multivariate Analysis of Variance (MANOVA)
- Logistic regression (Binary, Ordinal, Multinomial, ...)
- Ordinal logit model
- Log-linear regression (Poisson regression)
- Quantile regression
- Cubic splines
- Nonparametric regression (Kernel and Lowess)
- Nonlinear regression
- Partial Least Squares regression (PLS)
- PLS discriminant analysis
- Repeated measures Analysis of Variance (ANOVA)
- Mixed models
- Ordinary Least Squares regression (OLS)
- Principal Component Regression (PCR)
- Two-stage least squares regression

CORRELATION/ASSOCIATION TESTS

- Tests on contingency tables
- Correlation tests
- Mantel test
- Cochran-Armitage trend test
- Biserial correlation
- RV coefficient

PARAMETRIC TESTS

- Test for one proportion
- Test for two proportions
- k proportions test
- Multinomial goodness of fit test
- One-sample t-test and z-test
- Two-sample t-test and z-test
- One-sample variance test
- Two-sample comparison of variances
- k-sample comparison of variances
- Multidimensional tests (Mahalanobis, ...)
- TOST (Equivalence test)

NONPARAMETRIC TESTS

- Non parametric tests on two independent samples
- Non parametric tests on two paired samples
- Kruskal-Wallis test
- Friedman test
- Page test
- McNemar's test
- Cochran's Q test

- Durbin and Skillings-Mack tests
- Cochran-Mantel-Haenszel test
- One sample runs test
- Mood test (Median test)

TOOLS

- Export to GIF/JPG/PNG/TIFF
- Manage data (DataFlagger, MinMaxSearch, Remove text values in a selection)
- Manage workbook (Sheets management, Delete hidden sheets, Show hidden sheets)
- Manage the menu bars (Display the main bar, Hide the sub-bars)

TESTING FOR OUTLIERS

- Grubbs' test for outliers
- Dixon test for outliers
- Cochran C test for outlying variances
- Mandel's h and k statistics for outliers

MACHINE LEARNING

- Classification and regression trees
- Association rules
- K Nearest Neighbors (KNN)
- Naive Bayes classifier
- Support Vector Machine
- k-means clustering
- Gaussian mixture models