Introducing Microbiome Bioinformatics

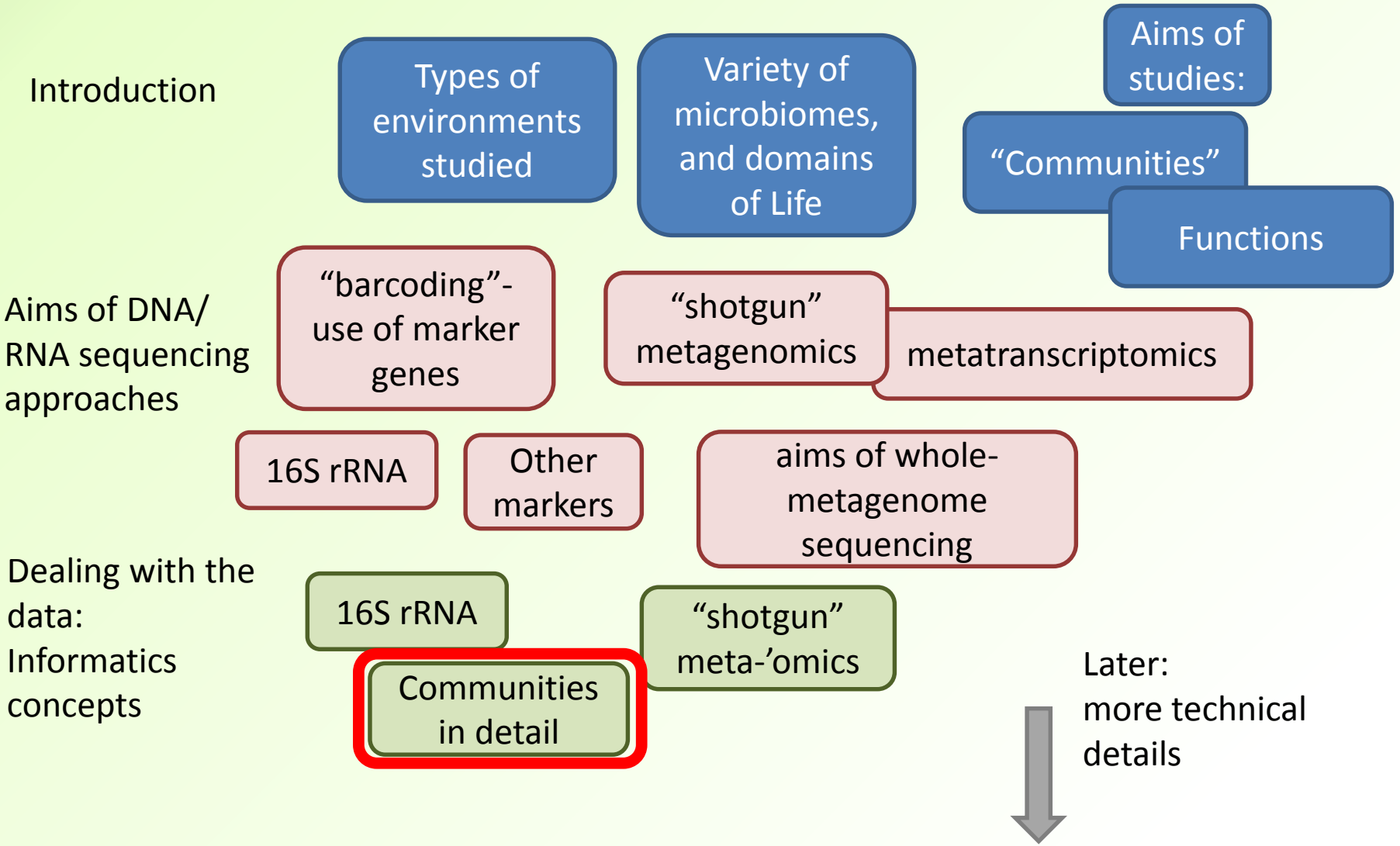
Part 8.

Microbial ecology - Diversity

Recap: Aims

- **Microbiome analysis**
  - with particular regard to **sequence informatics concepts**
- "Top down" – putting analysis tools and resources in context
- No highly detailed technicalities (yet)
  - No instructions on how to run particular programs
- Why you are using the bioinformatics approaches you use; pros, cons; alternatives

Topics, top-down



Series of talks

- 7 so far
- Open ended… as long there is demand
- Expected to be every 2 weeks
  - Notwithstanding some larger gaps for various reasons…
  - all dates will be confirmed in advance
  - *Please refer to:* ***Bite-size bioinformatics mailing list***
- Informal and flexible
  - Please interrupt and ask questions
  - **Suggestions for topics for further focus**

Series of talks

- Part 1: 27/1/2017
  - "Biological and Experimental Stuff that a microbiome bioinformatician needs to know"
  - Overview of marker gene sequencing for community analysis
- Part 2: 10/2/2017
  - Overview of whole-metagenome sequencing
- Part 3: 24/2/2017
  - Focus on metatranscriptomics
- Part 4: 10/3/2017
  - Different bioinformatics approaches to processing 16S read data
- Part 5: 24/3/2017
  - *De novo* OTU clustering: sequence identities and how thresholds have been determined historically; relationships to taxonomic levels
- Part 6: 7/4/2017
  - The clustering problem: different approaches, and what can go wrong; the influence of amplification artefacts, sequencing errors and sequence lengths; computational OTUs versus species
- Part 7: 21/4/2017
  - Introducing microbial ecology: using observed abundances of OTUs (or species, or functions) to estimate the richness of the community (number of different OTUs, species etc)
- Part 8: today – continuing microbial ecology: community diversity
- Slideshows
  - http://ghfs1.ifr.ac.uk/ghfs/

Future talk(s)

- 16th June Barton
- 30th June Barton

The story so far….

- Probing a microbiome community to assess:
  - Abundances of different types of **organisms**
    - 16S, 18S/ITS, metagenomics, metatranscriptomics
    - ….especially with 16S (still a bit more on this, later)
    - More details of other meta-'omics to follow
  - Abundances of different types of gene **functions**
    - Metagenomics, metatranscriptomics
- Last time: microbial ecology
  - Overview of richness and diversity
  - and **Richness** in some detail

You have a table like this:

SAMPLES …. ***…..***

OTUs

or ***species***

…. or other 'phylo-types'

…. or gene functions

	#1	#2	#3	#4	#5	#6	#7	#8
a								
b								
c								
d								
e								
f								
g								
h								
i								
j								
k								

(relative) frequencies….

This could result from 16S rRNA gene sequence (16S rDNA) analysis, **or** metagenomics sequence analysis;

and from OTU-based approaches, and non-OTU based

“Amounts of different things”

- “Things”: different –
  - Species
  - OTUs
  - Some other taxonomic unit
  - Phenotypes
  - Molecular functions
  - Pathways

phylotypes (Species, OTUs, Some other taxonomic unit)

types of organism (Species, OTUs, Some other taxonomic unit, Phenotypes)

types of gene (Molecular functions, Pathways)

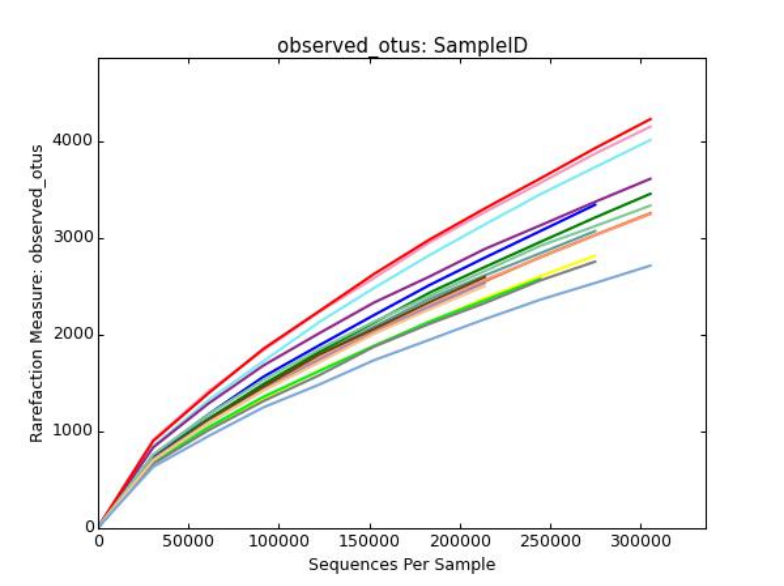
- Whichever we are interested in, we will benefit from a **simple metric**, instead of a large table
- Enables easy and direct comparison between samples
  - Disease/health states
  - Genotypes
  - Different time points for the same subject

Recap: metrics

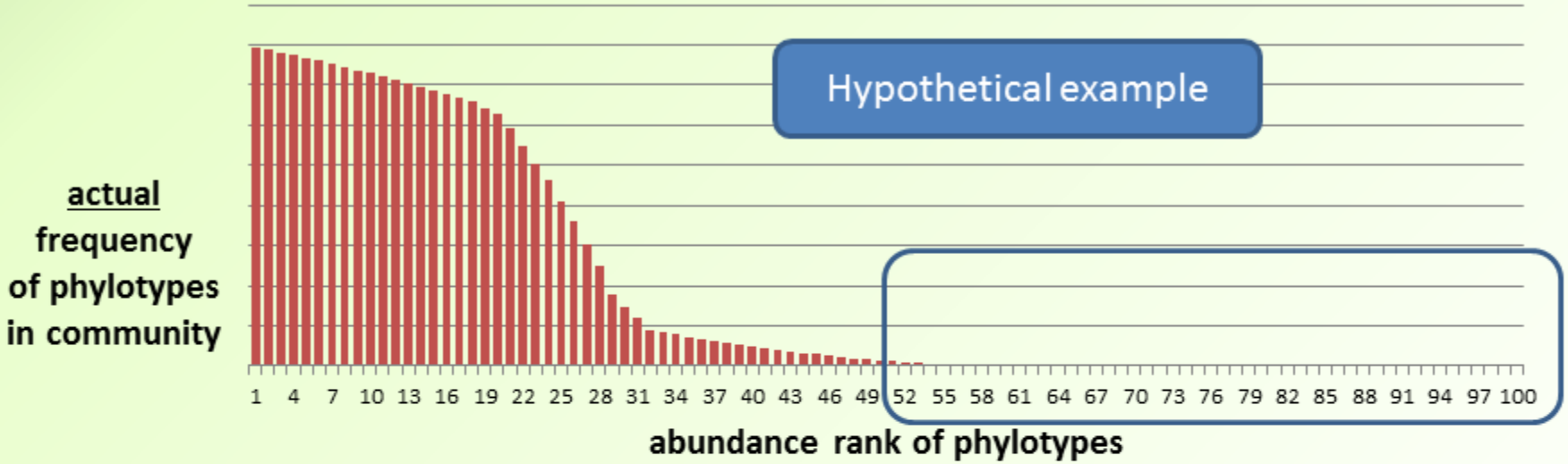
- So which metrics are available?
- **Richness**: number of **different species** (or OTUs, functions, etc)
- Richness is of limited use
- Huge **range of abundances** between different species (or OTUs…), which means:
  - It only tells a very small part of the story
    - And probably a not useful one, at that
- It is very difficult to estimate reliably
  - Abundance distribution may not be known *a priori*
  - But often, there are a large number of species which have extremely low abundance
  - Not helped by the impact of sequencing errors

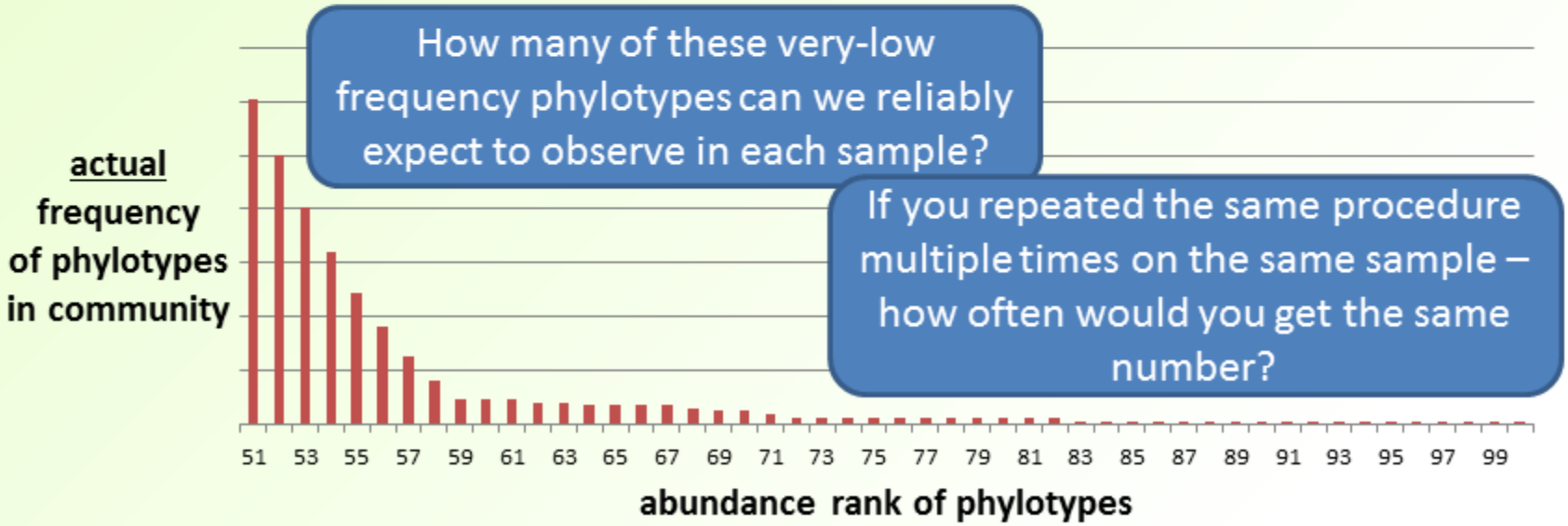
Recap: richness

- In principle, **rarefaction** can help to estimate the number of species (OTUs…) in the community
- But the short story is: **this is likely to be unreliable**



Abundance versus rank: What shape is the tail? How long is it?





Assessment of estimators

- Numerous in the literature
- Haegeman *et al.* (2013):
- "Species richness cannot be estimated from sample data alone"
- "We claim that sample data is always consistent with very different community structures"
- "computation shows that the rarefaction curves do not depend on the abundance distribution of the rare species"
- "We have shown that the number of species in a community cannot be reliably estimated from sample data"
- For anyone who has analysed many sets of 16S-sequenced samples from many experiments, it may be a relief to hear all this…

However… even if we could determine richness with 100% accuracy, every time…

How useful would these numbers actually be?

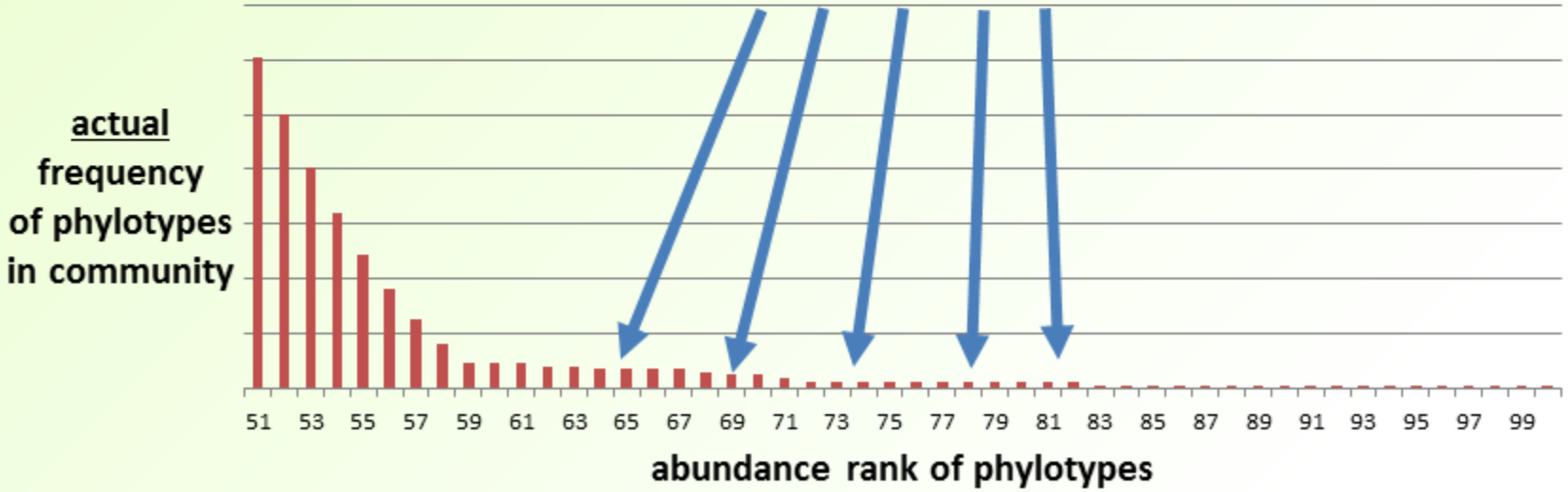
Richness is intrinsically limited

How miniscule does an **actual abundance in the original community** need to be, in order for us to treat it **the same as if it was zero?**

Influenced by considerations of both:

- reliability of determining richness
- lack of importance assumed for very low-abundance types

In general, **how much do you actually care** about these very low-abundance phylotypes?



A

relative abundance

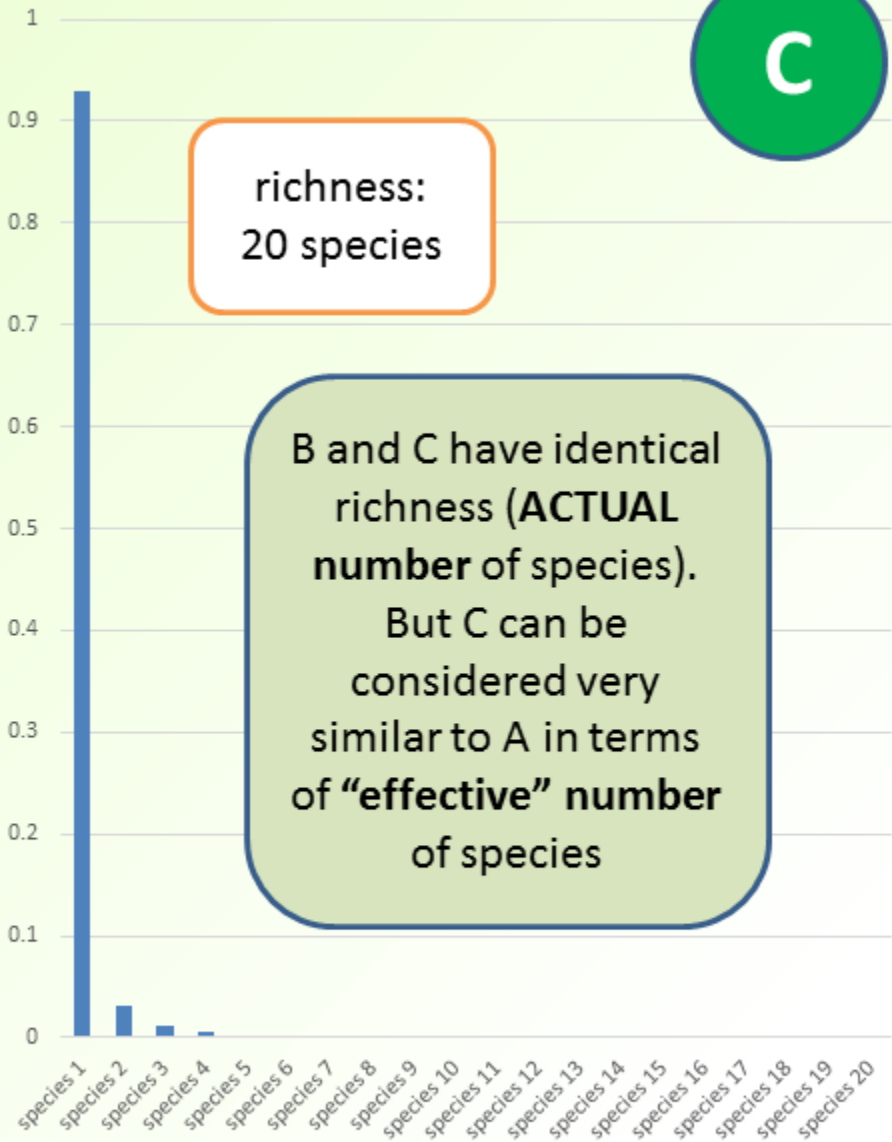
richness:
1 species

B

richness:
20 species

C

relative abundance



richness:
20 species

B and C have identical richness (**ACTUAL number** of species). But C can be considered very similar to A in terms of **"effective" number** of species

- But the *great* thing about Richness is that it is very easy to understand:
- “The number of different species”
- Extremely simple, and the units are obvious

- So what about Diversity ….?

An important notion:

The **effective number** of phylotypes (species, OTUs) or whatever

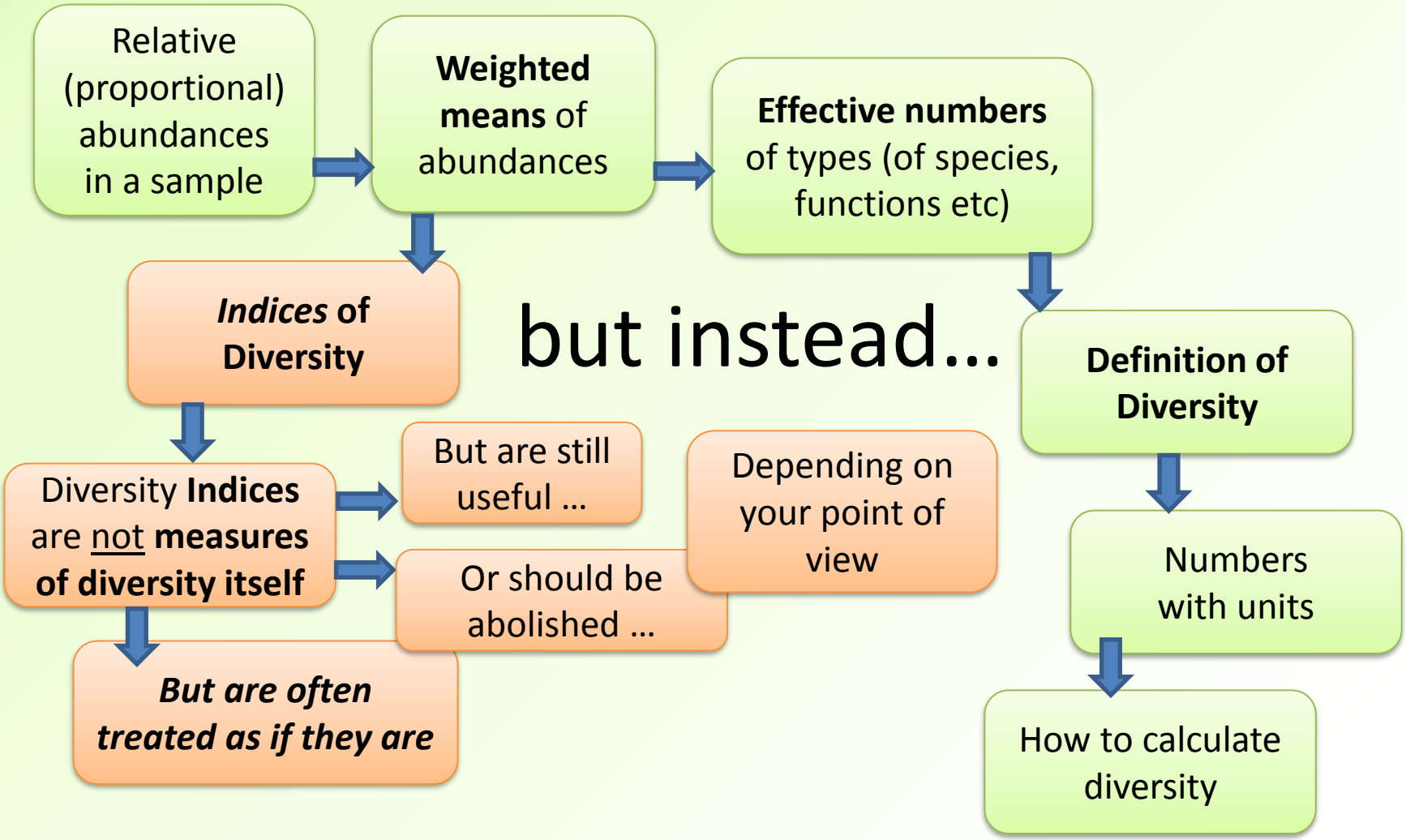
There follows an attempt at a qualitative description….

….and also some maths

Effective number of types

- The **effective number** of phylotypes results from a consideration of "**dominance**" versus "**evenness**", and can be quantified (by various methods).
- It is also has a (very) close relationship to **diversity**
- It also relates to our ability to reliably and **reproducibly estimate** the number of phylotypes by sampling
  - The **effective number is more reproducible** than the actual number

How today's session *could* have been…



Why?

- It is important to understand **diversity indices**
- Because they are used **a lot** in the literature
- Importance of understanding the difference between a diversity index and diversity itself
  - Apparently many authors do not appreciate this distinction
- Appreciate that some **indices** could be used as a reasonable **definition** of **actual diversity** itself
  - (but the purists say, No)
- For next time:
  - How some commonly used indices of diversity are simply related to actual diversity
  - (if they were not, they would be rubbish indices of it….)

What's all the fuss?

- Some **extremely** useful papers by Hanna Tuomisto
- "A consistent terminology for quantifying species diversity? Yes, it does exist" Tuomisto (2010) *Oecologia* **164** 853-860

"The term 'diversity' has been used in **at least four conceptually different ways** in the ecological literature, primarily **because indices of diversity have been equated with diversity itself**"

"**true diversity** seems to have been **much less used** in the ecological literature than **the diversity indices**"

- "I discuss the conceptual definition of diversity and show that a **logical terminology for diversity studies is already available**"
  - Tuomisto (2010)
- More from her next time, when we have a look at even more diversity of definitions of diversity in another context (α, β, γ)

The remainder of this session

- An important distinction between:
  - **defining what something actually is,** and
  - using a number as an **index of that thing**
  - (worth knowing to avoid confusion from the literature)
- Using weights to usefully describe uneven distributions
  - The basis of oft-used indices of diversity
- More about **effective numbers of species** (or numbers of OTUs, functions, whatever)
- A quantitative definition of diversity

Next session:

- Details of the quantitative definition of diversity: the Hill numbers
- Phylogenetic diversity
- α-diversity, β-diversity, γ-diversity

But first…….

<u>the important distinction…</u>

Always remember that these two concepts are NOT THE SAME

1. The **definition** of a **Thing**
   - This definition also necessarily specifies the **units** by which that Thing is measured
2. An '**Index**' of that Thing
   - The 'Index' can be thought of as a utilitarian indicator of the Thing
   - The Index may be easier to work with than the strict definition
   - The Index may well be measured in different units

- *(Yes, but what on earth does this mean…)*

Example: '**the size of a room**'

- Room sizes:
- Ormesby A 9
- Ormesby B 9
- Ranworth 16
- Barton 24
- Rollesby 24
- Lecture Theatre 126

Example: **‘the size of a room’**

- We can **define** the **size of a room** as the volume enclosed by its walls, floor, ceiling
- → Units: cubic metres (or some other proportional unit)
- Useful **indices** of the size of a room (don’t involve a tape-measure…)
  - *“How large is the Barton Room? Large enough for a microbiome talk?”*
  - *“It only takes 24 people!”*
  - *“No good! He’s talking about microbiome diversity today, you know! So how large is the lecture theatre?”*
  - *“126 people…. might just work?”*
- This (**audience capacity**) is probably a more practical way of dealing with the size of a room, than discussing cubic metres
- But note, “people” is not a unit of volume!
- “People” is a unit of audience capacity, our index

A Thing, and indices of a Thing

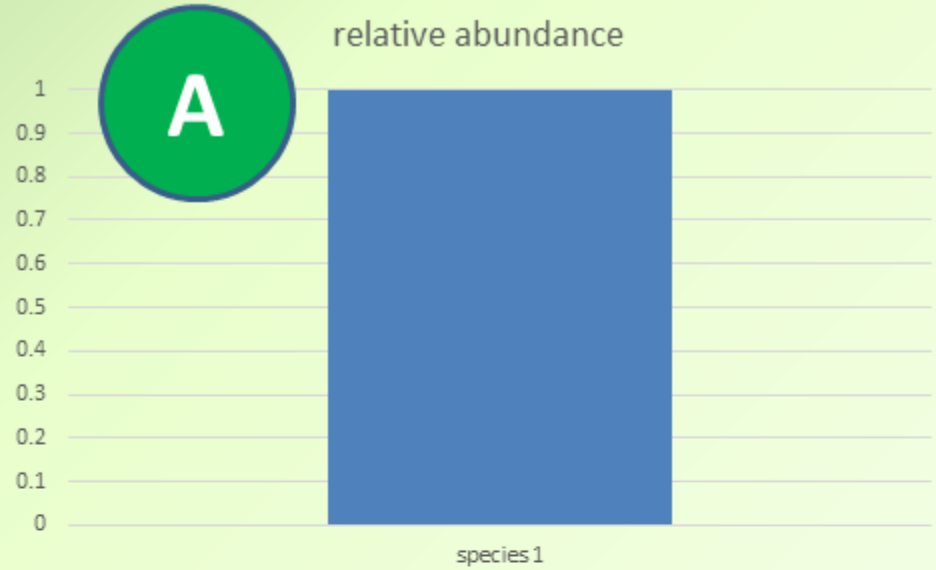
- Other possible indices of the size of a room might be:
  - Numbers of items of furniture it contains
    - (probably less useful than the **audience capacity**)
  - Number of windows it has (much cruder)
  - The no. of people the room would *literally* contain (sardines)
- Note that alternative **definitions** of 'the size of a room' could be argued
- E.g. we might formally define the size of a room as:
  - The literal volume of the room
  - The surface area of the floor of the room
- The choice is ours; note that the above have different units
- But the same **indices** (e.g. audience capacity) could be used, irrespective of the definition

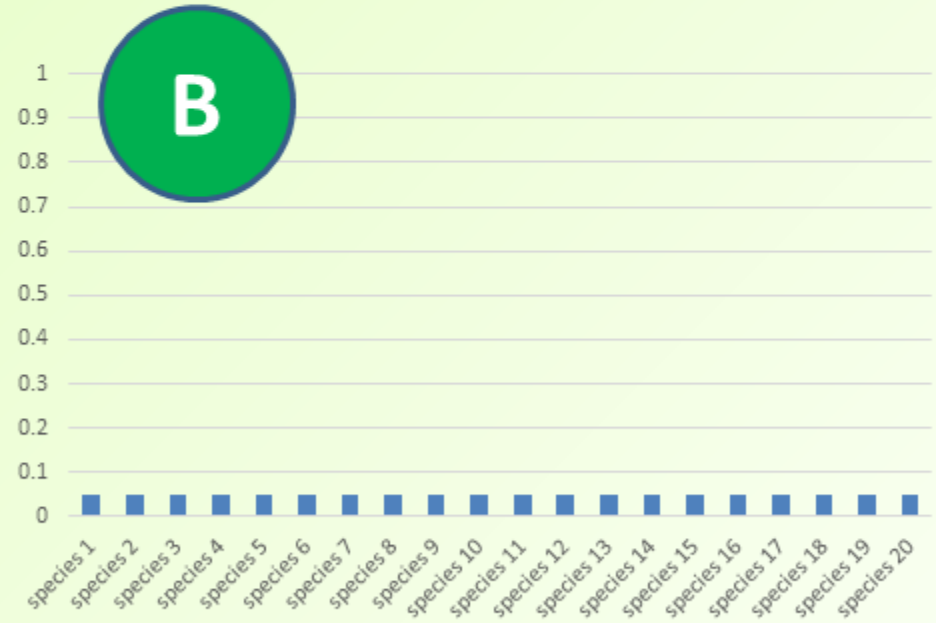
Diversity, and Indices of Diversity

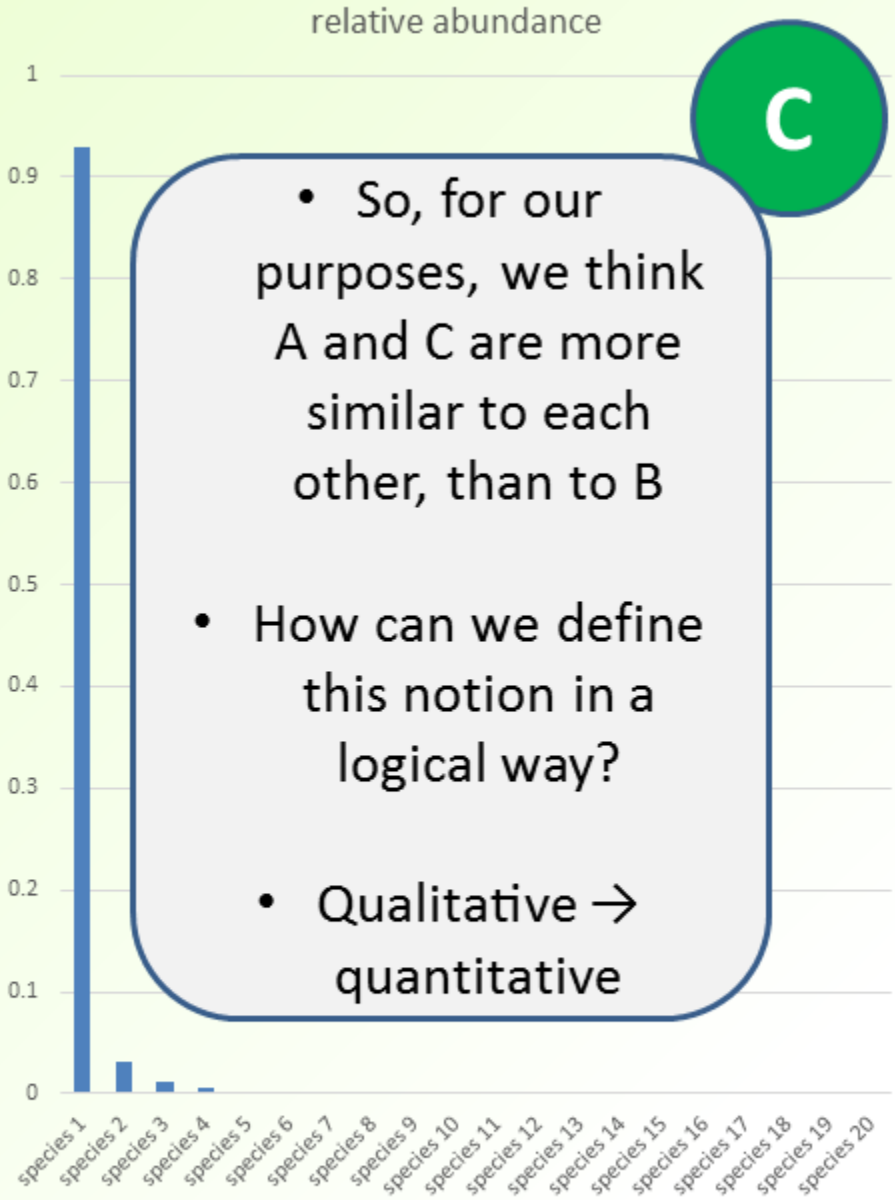
- Defining ‘the size of a room’ seems pretty clear
- Defining ‘diversity’ less so
- This has led to lots of definitions of diversity
- Some are much more widely used than others
- Some seem to make more sense than others
- Indices of diversity are not the same as diversity
- Often they are discussed as if they are
- **Lots of confusion in the literature**

Another distinction

- The goal of **defining** something (like Diversity) is not the same as:
- Solving the problem of **reliably estimating** it from your sample data
- Again, see Tuomisto (2010) for a discussion of how these two things are sometimes conflated in the literature
- As for estimation: as ever, you can calculate the diversity indices (and the actual diversity) from the sequence reads in your sample
- And use the result as an estimator of what is in the original community



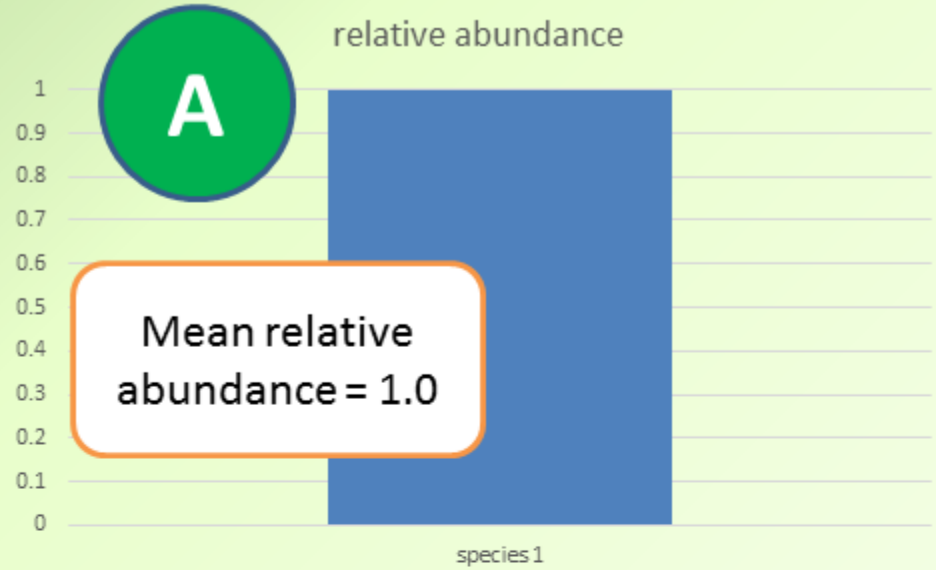


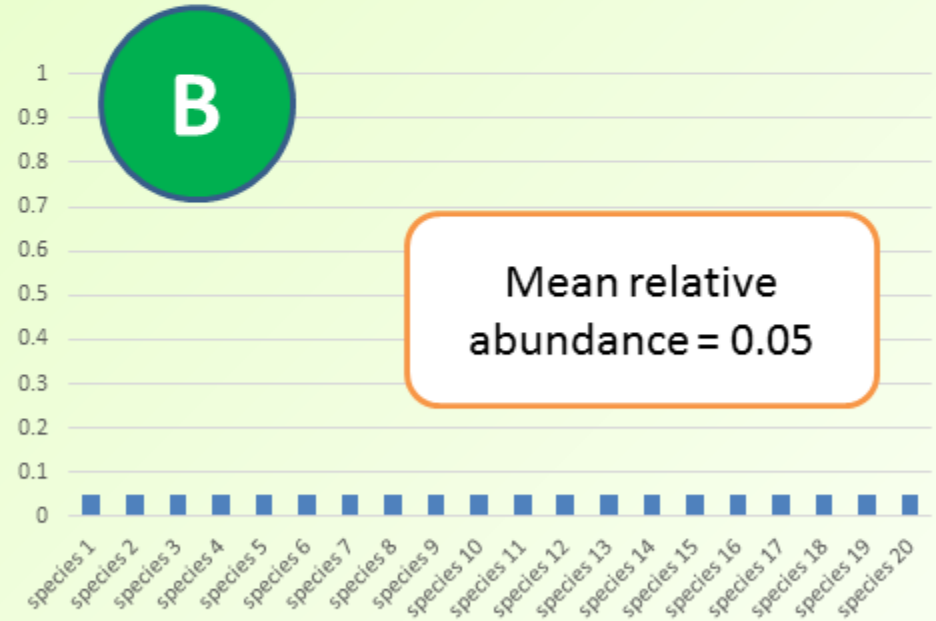


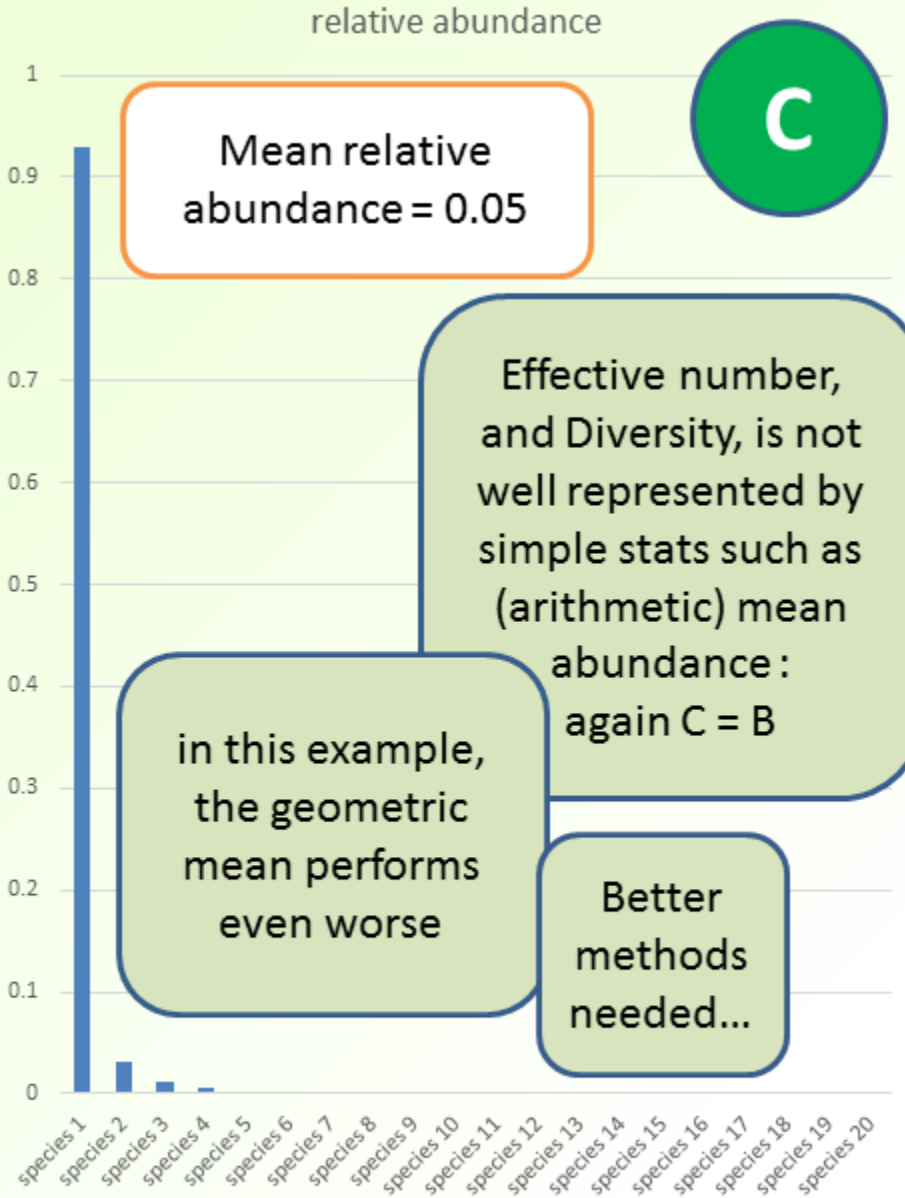
- So, for our purposes, we think A and C are more similar to each other, than to B
- How can we define this notion in a logical way?
- Qualitative → quantitative

How do we quantify this?

- One useful way of regarding the relative abundances is as a **probability distribution**
  - If you picked one observation (sequence read) at random, what is the probability that it would be assigned to each species?
  - More on this later…
- But how do we deal with the numbers?
- First, let's look at our **means**
  - First, simple (unweighted) means
  - but these don't help at all….


A
relative abundance
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean relative abundance = 1.0
species 1


B
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean relative abundance = 0.05
species 1
species 2
species 3
species 4
species 5
species 6
species 7
species 8
species 9
species 10
species 11
species 12
species 13
species 14
species 15
species 16
species 17
species 18
species 19
species 20


relative abundance
C
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean relative abundance = 0.05
Effective number, and Diversity, is not well represented by simple stats such as (arithmetic) mean abundance : again C = B
in this example, the geometric mean performs even worse
Better methods needed...
species 1
species 2
species 3
species 4
species 5
species 6
species 7
species 8
species 9
species 10
species 11
species 12
species 13
species 14
species 15
species 16
species 17
species 18
species 19
species 20

Sums, Means and Weights

- Use arithmetic means as an example
- *N* observations: $x_1 , x_2 , x_3 , x_4 , x_5 \ldots , x_N$
  - For **relative abundances** of species (or whatever), these will all be $0 < x_i \leq 1$
- Unweighted sum:
  - $\sum = x_1 + x_2 + x_3 + x_4 + x_5 \ldots + x_N = 1$
- Unweighted arithmetic mean = $\boldsymbol{\sum / N} = 1 / N$
  - (see previous slide)
- This arithmetic mean is simply a weighted sum, where all the weights are the **same** ($w_i = 1/N$)
  - Mean = $\boldsymbol{w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 \ldots + w_N x_N}$

Sums, Means and Weights

- What if we **weighted** the observed values before adding them – each is weighted to indicate its relative size?
- Use a **non-uniform** value of w_i
  - The size of w_i is influenced by the size of x_i
- (Weighted) mean

  $= \boldsymbol{w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 \ldots + w_N x_N}$
- What might be suitable?

Simpson Index

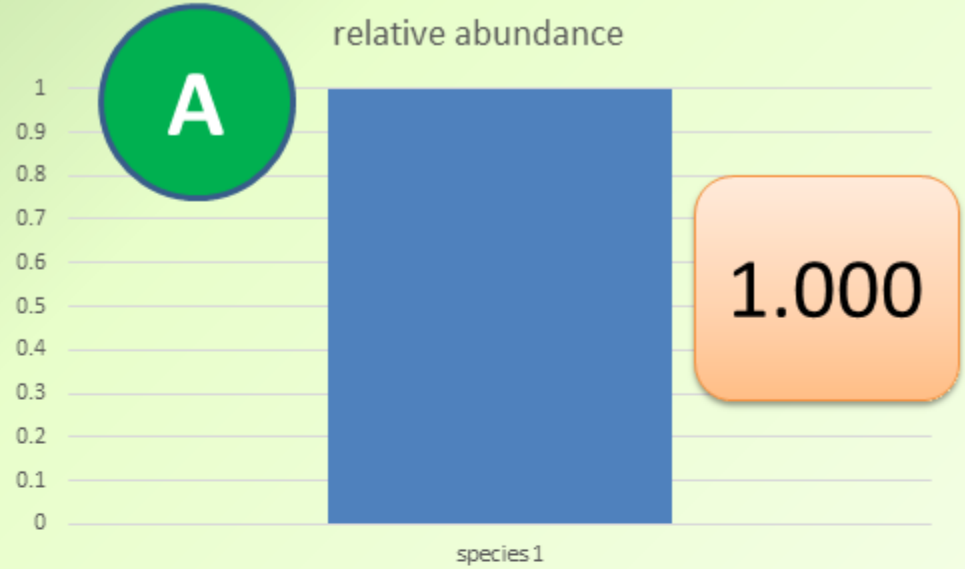
- $w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 \ldots + w_N x_N$
- Simply use $w_i = x_i$
- I.e., each proportional abundance is simply weighted by **itself**

  Simpson index = $x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 \ldots + x_N^2$

  $= \sum x_i^2$

  - This is also the **probability** that if two observations (e.g. sequence reads) are drawn at random from the data set, they are from the same species (or OTU, or function etc)
- Simpson (1949)

A



1.000

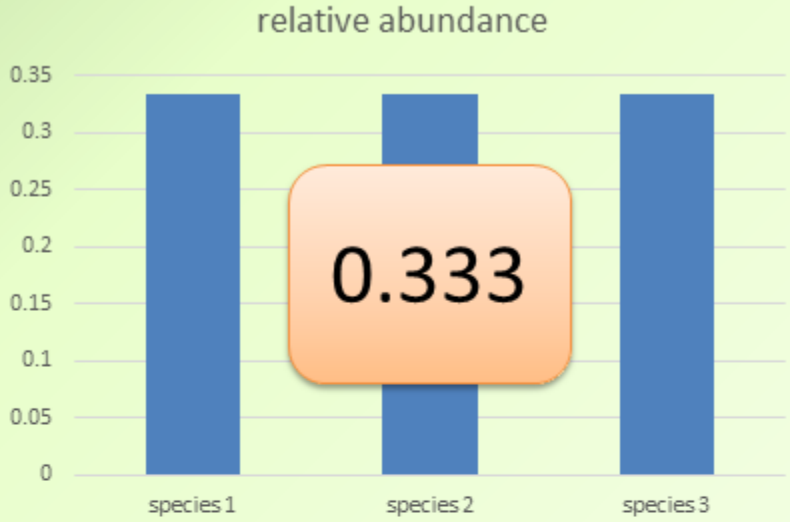
B

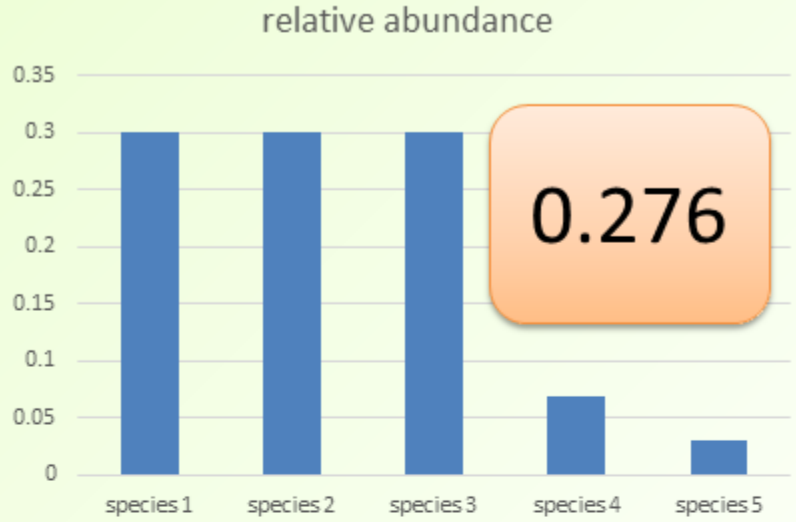
0.050

C

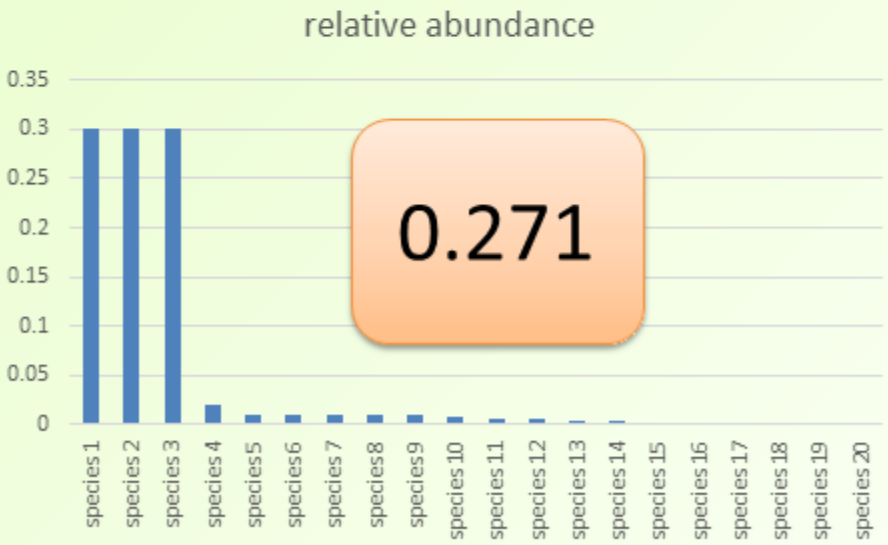
Simpson Index

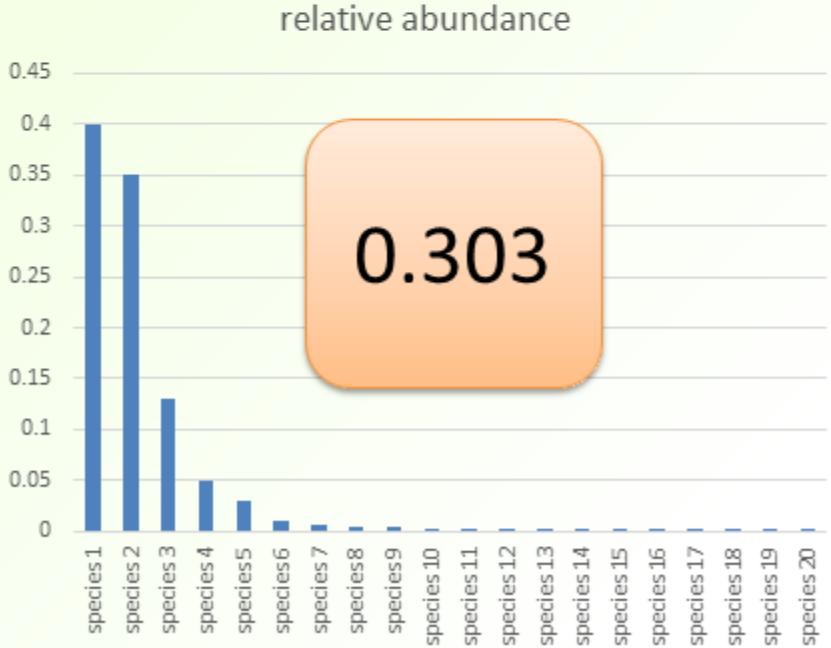
0.866





Robustness to long tails:



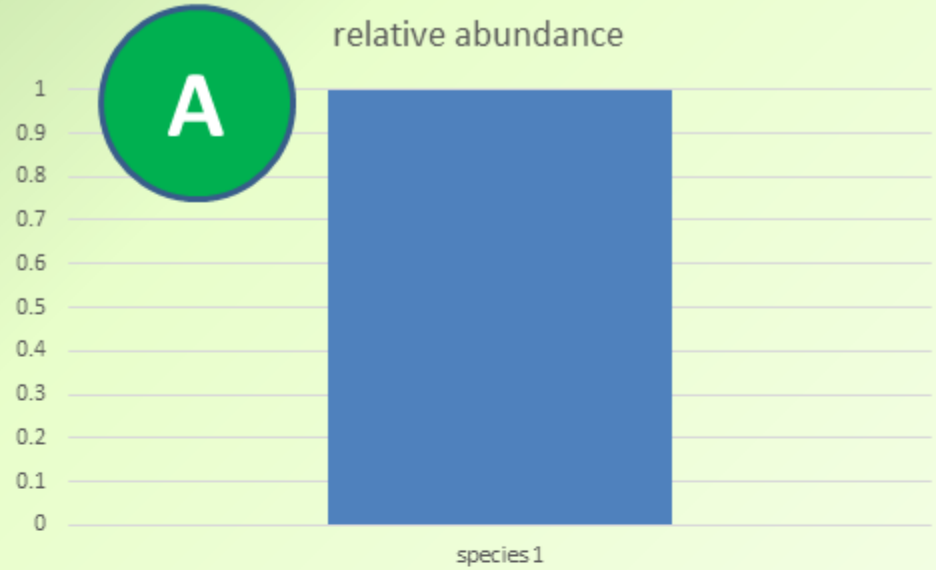


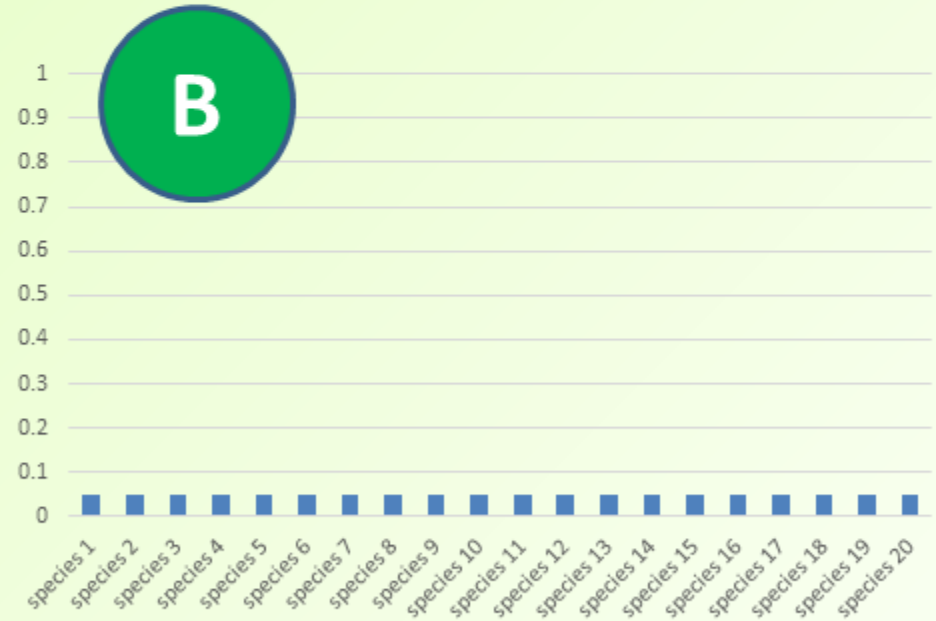
Simpson variants

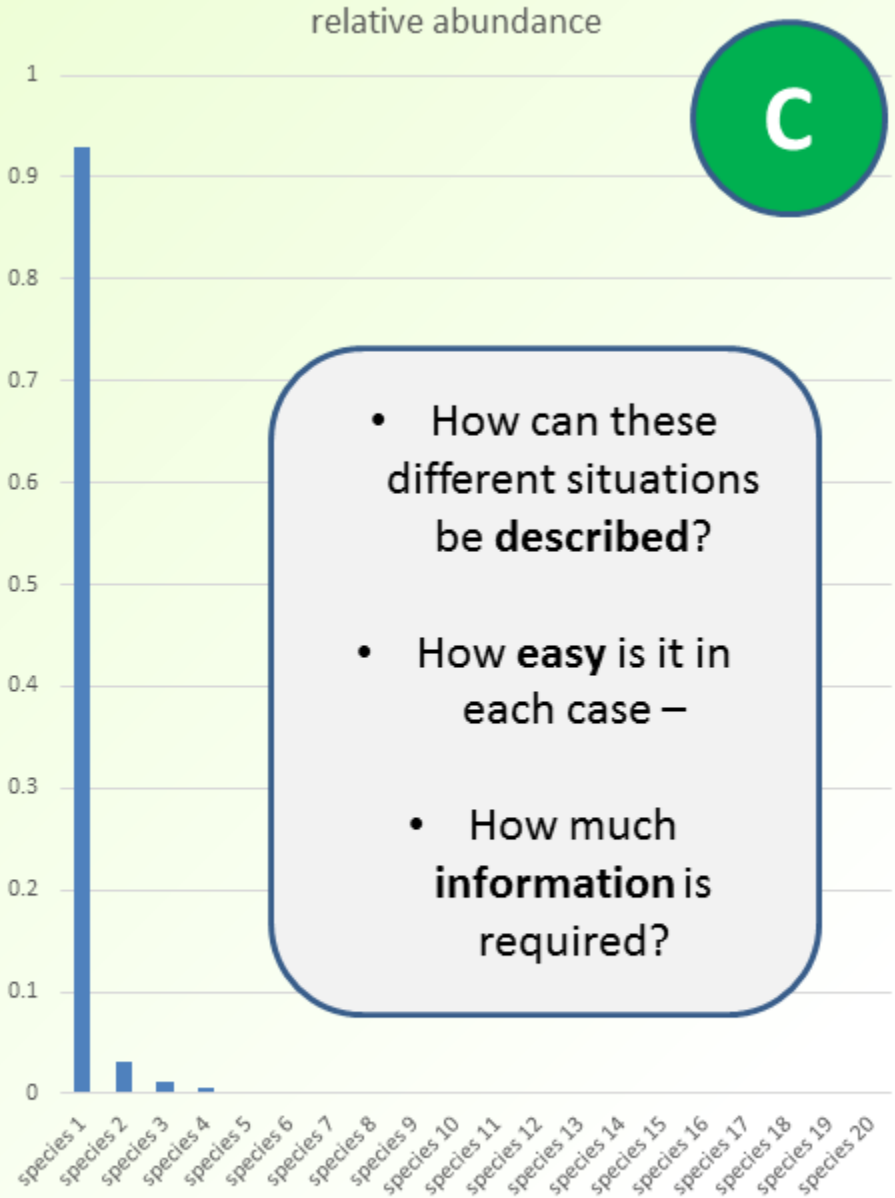
- The Simpson index is larger (closer to 1.0) for the least complex (lowest diversity) communities
- Maximum diversity could be considered to consist of a very large number (*N*) of equally-abundant types
  - As *N* increases, the Simpson index gets closer to 0
- To indicate diversity, we might prefer the opposite trend : the more diverse, the larger the number
  - The Gini-Simpson index is simply: **1 – Simpson**
  - Alternatively, the **inverse of the Simpson** is also sometimes used
    - More on this later! (when we talk about True Diversity)

Another index….

“information”



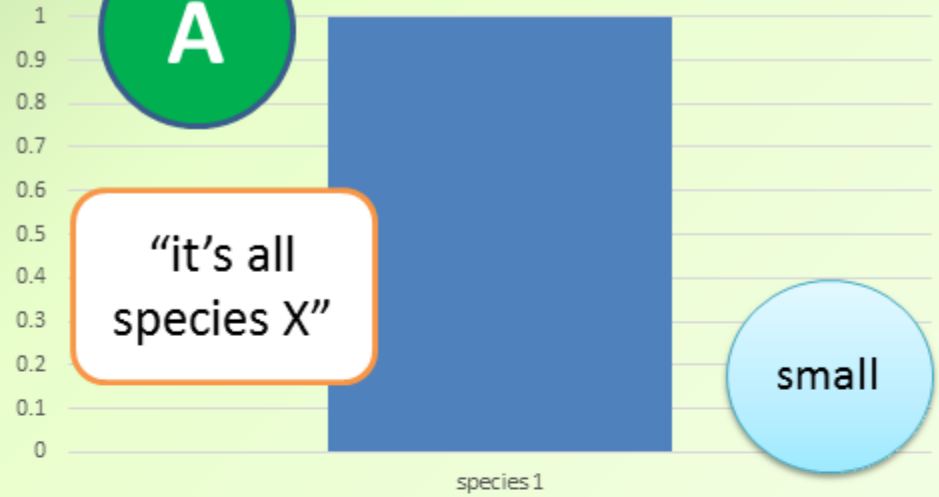




- How can these different situations be **described**?
- How **easy** is it in each case –
- How much **information** is required?

A

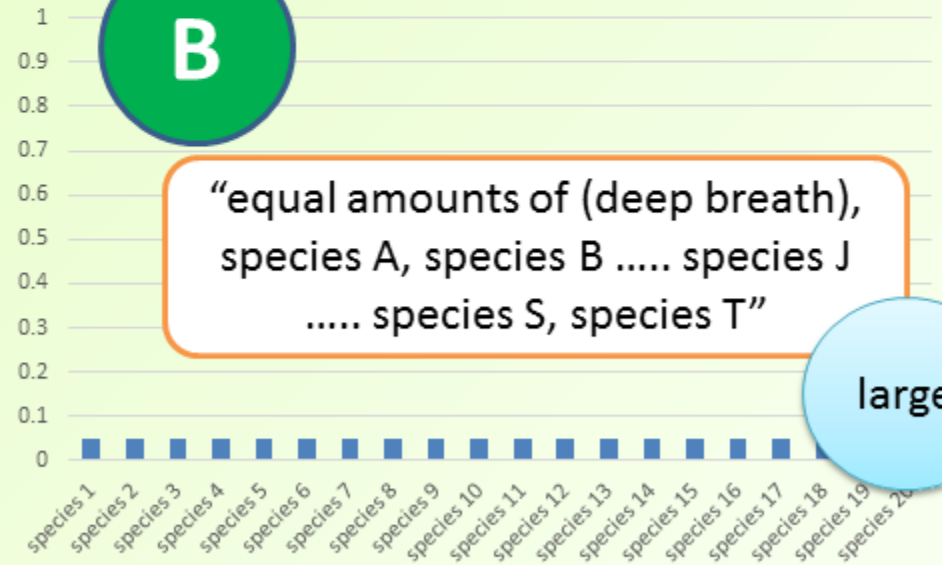
relative abundance



"it's all species X"

small

B

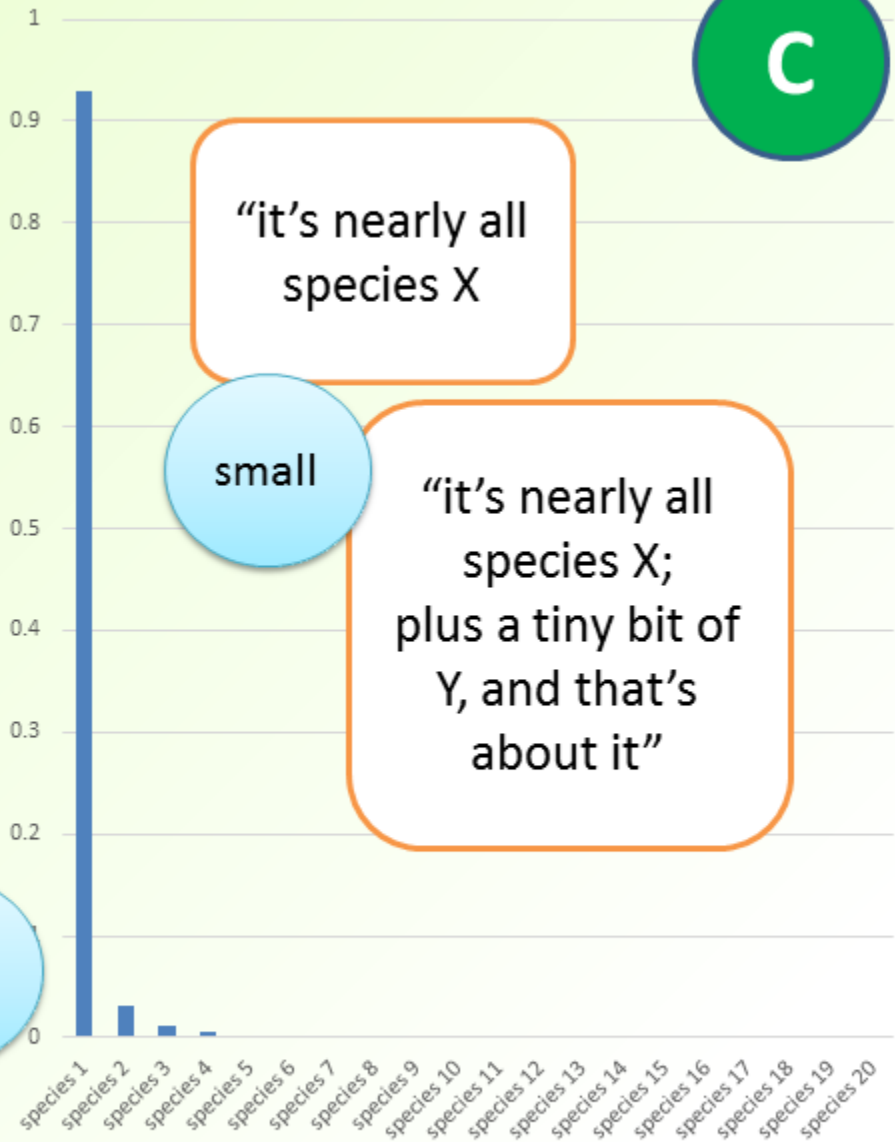


"equal amounts of (deep breath), species A, species B ..... species J ..... species S, species T"

large

C

relative abundance



"it's nearly all species X

small

"it's nearly all species X; plus a tiny bit of Y, and that's about it"

Shannon entropy

- Formally, an **amount of information is usually quantified** as **Shannon entropy** (usually denoted *H)*
  - Like the Simpson index, this is determined by the **probability distribution of 'events'** (species, OTUs, functions, etc)
  - uses **logs**, due to their additive properties with independent events

$$\boldsymbol{H} = -\sum p(A_i) \log_b(p(A_i)) \qquad \text{for all events } A_i$$

- Different bases (*b*) may be used; typically:
  - Natural logarithms ('ln') use base *e* ; information unit is '***nat***'
  - If base of 2 is used, units are '***bits***'
  - converting between Shannon entropy values calculated with different bases is trivial

Shannon entropy

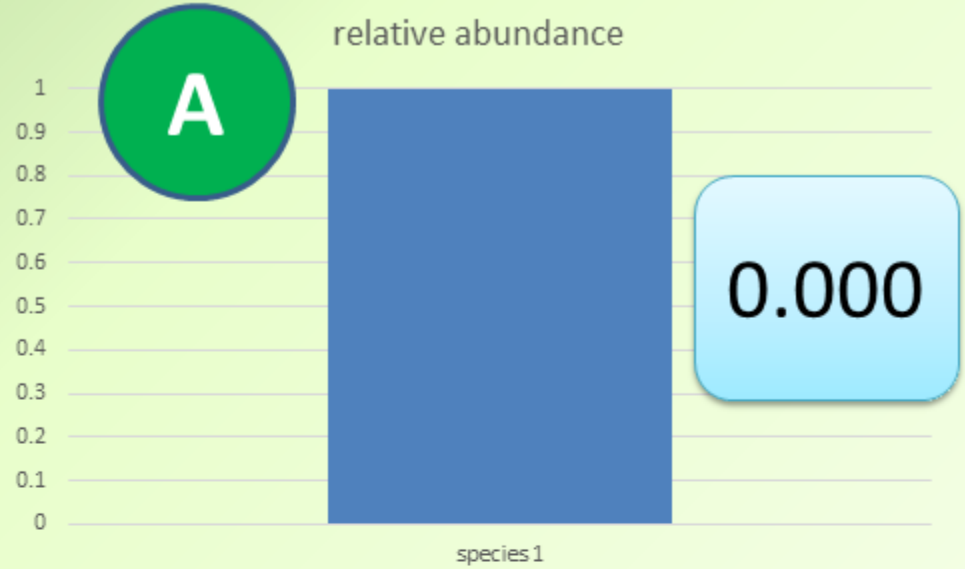
- Shannon (1948a, 1948b)
- Shannon entropy is used a lot in many fields of science and engineering
  - including in various fields of bioinformatics
  - E.g. information content of a DNA or protein sequence
- Shannon entropy is **mathematically** very similar to **thermodynamic entropy**
- In both cases, entropy quantifies **uncertainty** about something
  - A system of molecules
  - An ecosystem
    - Or the data set of sequence reads which represents it

Shannon index

- $\boldsymbol{H} = -\sum p(A_i) \log_b(p(A_i))$

- For ecological diversity, $p(A_i)$ is simply the proportional abundance, x_i
- Natural logarithms are used
  - $\boldsymbol{H} = -\sum x_i \ln(x_i)$

A



0.000

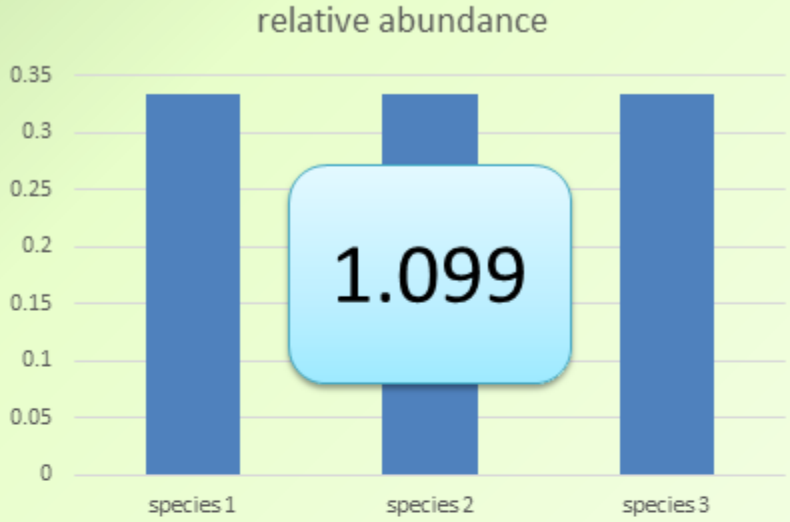
B

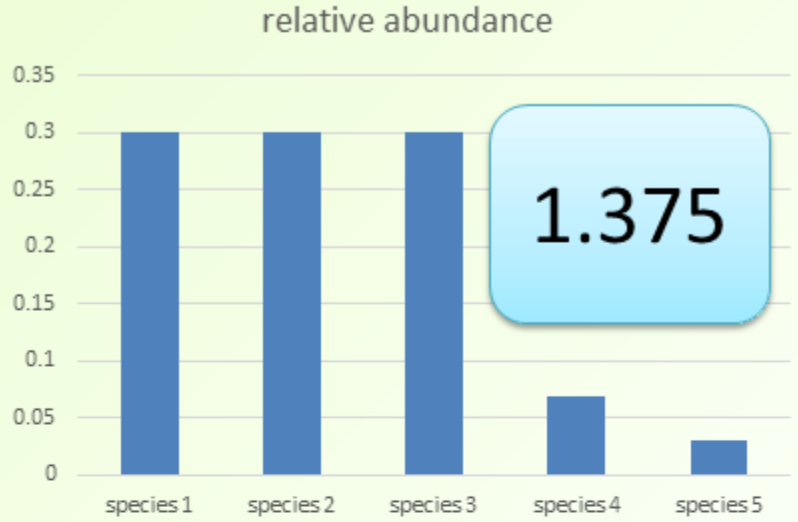
2.996

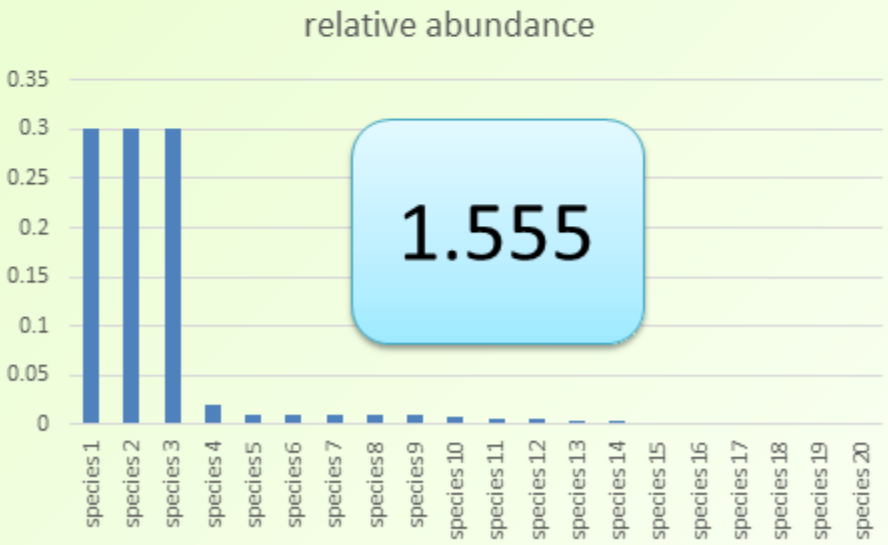
C

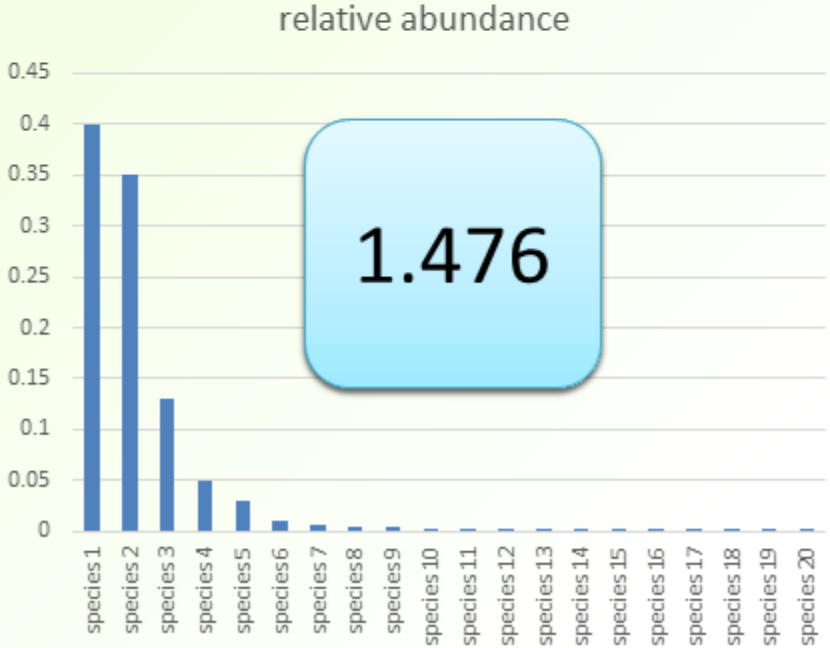
Shannon Index

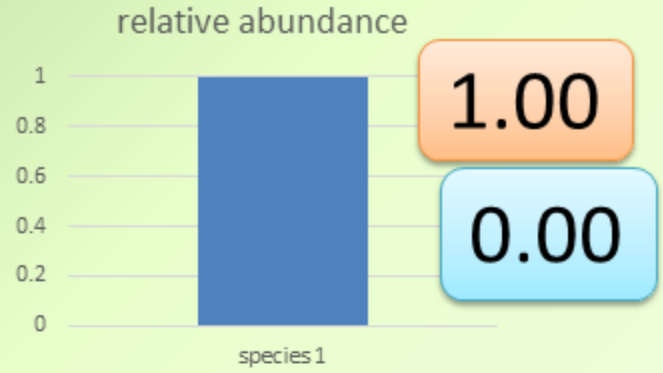
0.395

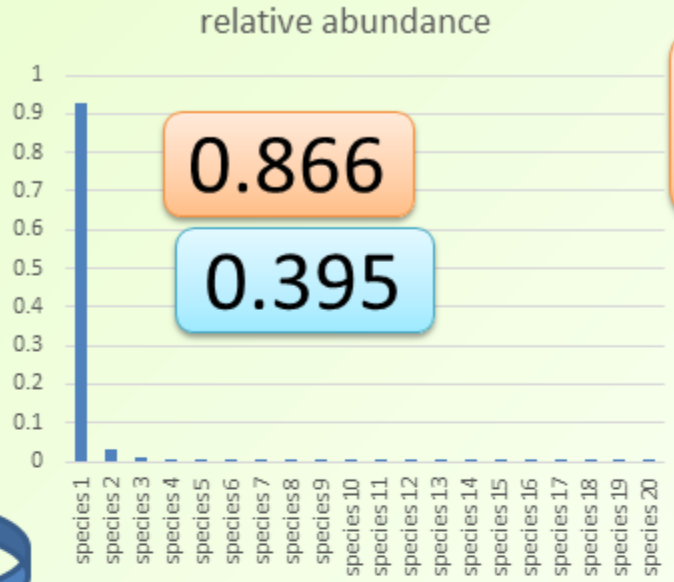

relative abundance
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
species 1
species 2
species 3
1.099


relative abundance
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
species 1
species 2
species 3
species 4
species 5
1.375


relative abundance
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
species 1
species 2
species 3
species 4
species 5
species 6
species 7
species 8
species 9
species 10
species 11
species 12
species 13
species 14
species 15
species 16
species 17
species 18
species 19
species 20
1.555


relative abundance
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
species 1
species 2
species 3
species 4
species 5
species 6
species 7
species 8
species 9
species 10
species 11
species 12
species 13
species 14
species 15
species 16
species 17
species 18
species 19
species 20
1.476

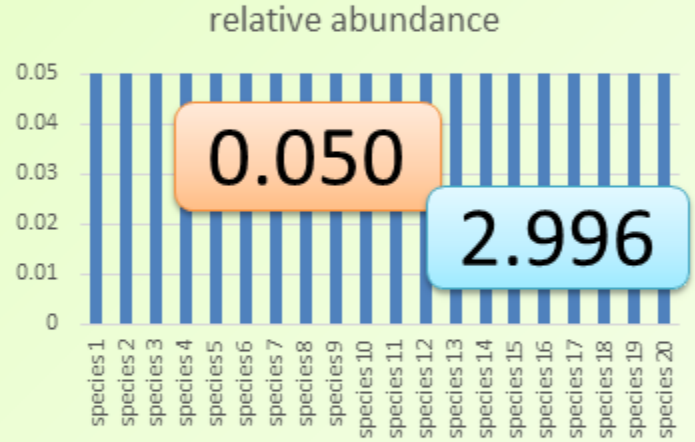


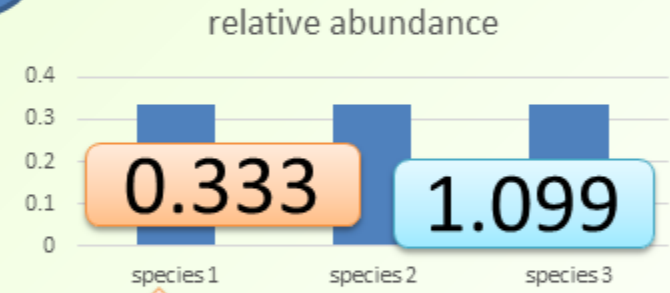


Simpson Index

high

low

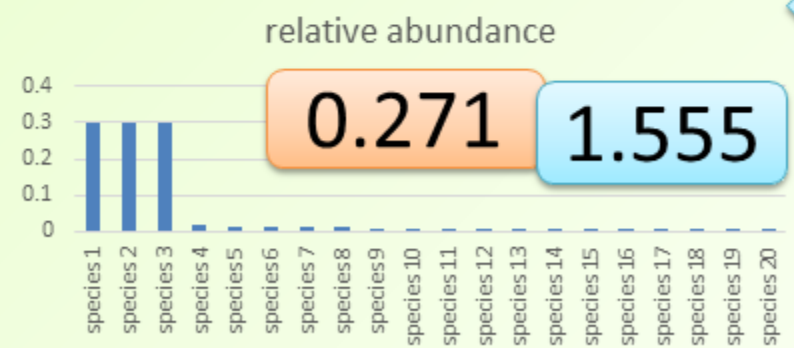


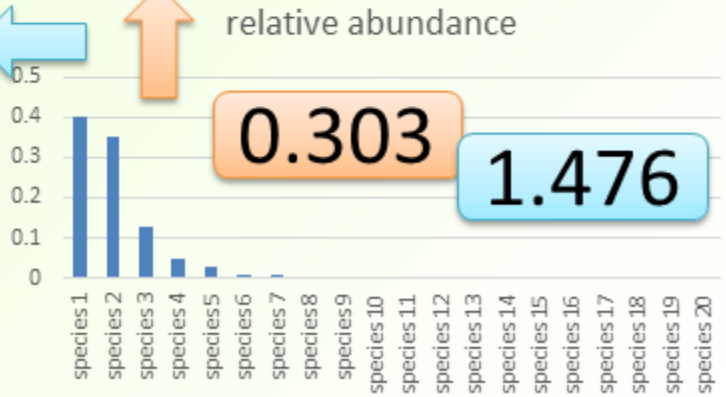


Shannon Index

low

high





- This has illustrated two commonly used indices of diversity
- Numerous others are available
- The Simpson and Shannon indices show basically similar trends
  - Albeit the 'directions' of the two indices are opposite
- The reasons for the different 'closeness' of some pairs of distributions include the **different emphases they place on the more-abundant and less abundant species**
- And we have a very neat, **logical choice** of what emphasis – or emphases - we use: more next time

Indices, or actual Diversity?

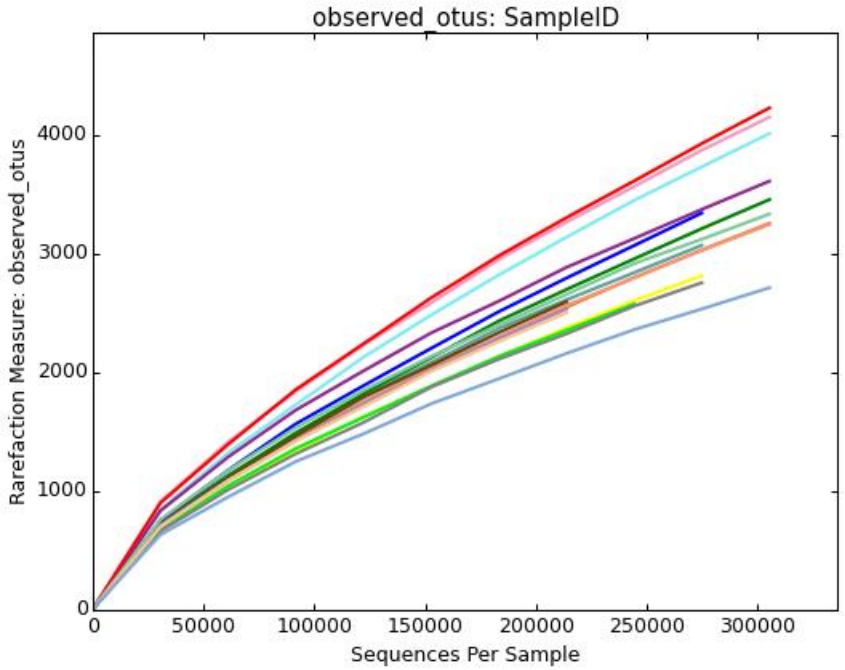
- Q: could either of these meet a **definition of what diversity <u>actually is</u>?**
  - Shannon index looks promising
  - The more diverse (uncertain) the community gets, the larger the Shannon index becomes
- A: Shannon entropy would not be a particularly bad definition (IMO)
  - You will find some literature which treats Shannon entropy *as being* diversity
  - But…. the classic definition of diversity….
    - (and most meaningful and logical IMO)
    - (but not necessarily the most widely used)
  - …. is **not** Shannon entropy (not quite….)
- Remember, "people" is not a unit of volume
- The Shannon index is literally a unit of entropy – not diversity

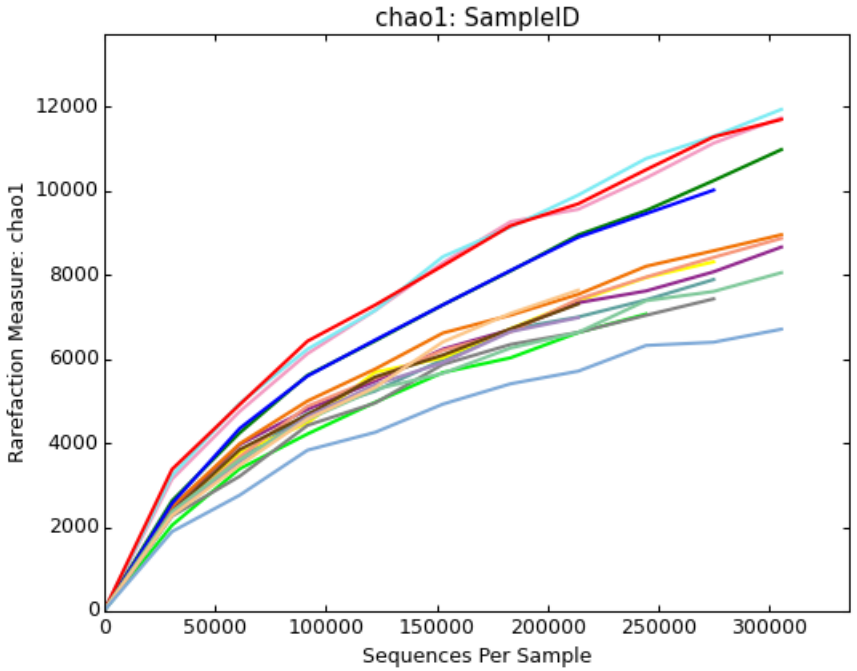
Tuomisto (2010) again…

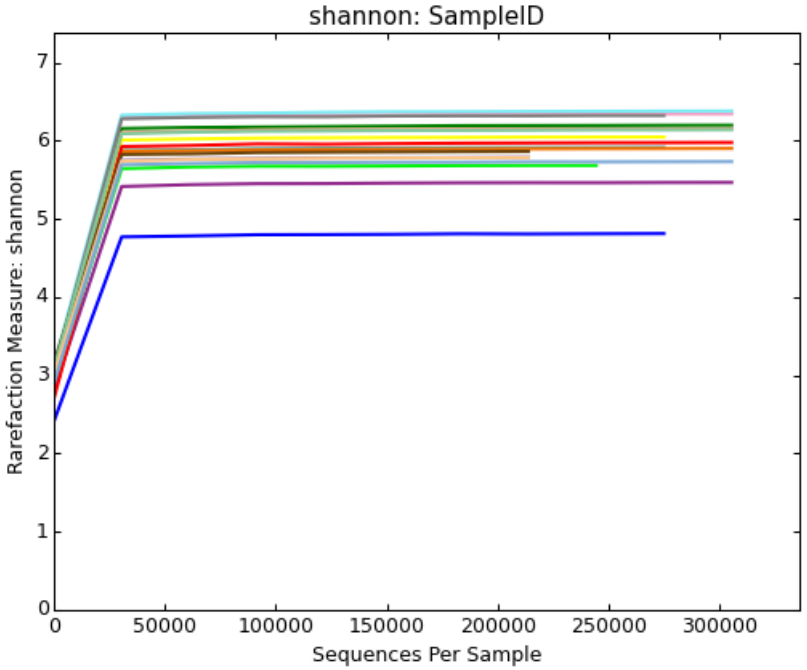
- "Jost has made a great contribution to the field by evaluating the properties of **true diversity** and pointing out how easily **diversity index** values are misinterpreted.
- Jost argued that **traditional diversity indices are superfluous**.
- I would take the argument one step farther: …**they are actually counter-productive**."

Diversity indices and rarefaction

- Recap: rarefaction involves repeated **resampling** of **subsets** of our actual sample
- Metrics can be calculated on each of these smaller subsamples, such as:
  - Number of species (or OTUs etc) observed
  - **Estimates** of the actual number of species, inferred from the observed number
  - I.e. estimation of true richness
  - Diversity indices
  - (and true Diversity… see next session)



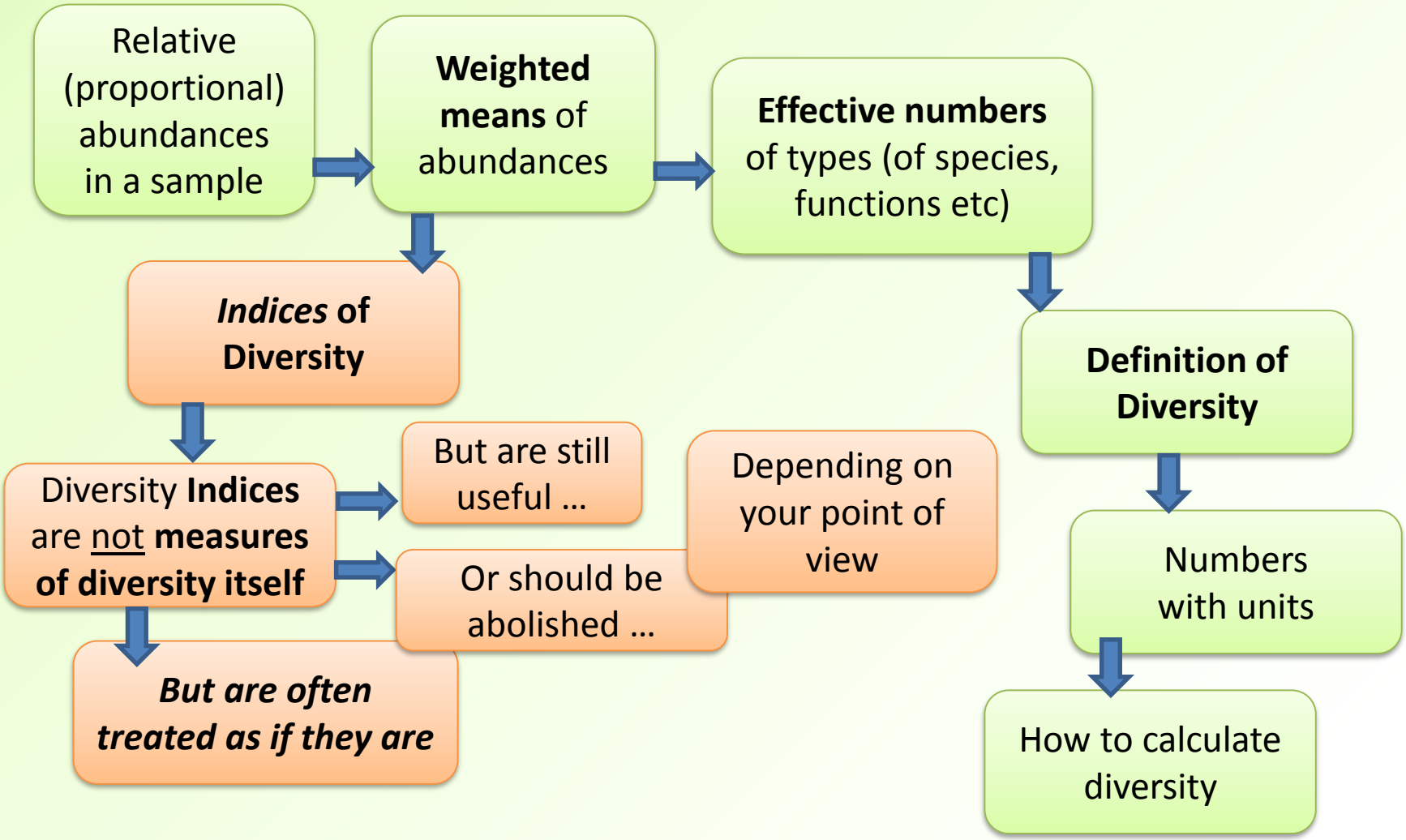




Indices such as Simpson and Shannon are robust to long tails

Tend to level off after the most abundant types have been accounted for

So that's the Diversity Index detour



References

- Haegeman B., Hamelin J., Moriarty J., Neal P., Dushoff J. and Weitz J.S. (2013) Robust estimation of microbial diversity in theory and in practice *ISME J.* **7:** 1092-1101
- Shannon C.E. (1948a) A Mathematical Theory of Communication *Bell System Technical Journal* **27** (3): 379-423
- Shannon C.E. (1948b) A Mathematical Theory of Communication *Bell System Technical Journal* **27** (4): 623-656
- Simpson E.H. (1949) Measurement of Diversity *Nature* **163:** 688
- Tuomisto H. (2010) A consistent terminology for quantifying species diversity? Yes, it does exist, *Oecologia* **164:** 853-860