Measuring, exploring and estimating biodiversity

Jonathan Dushoff, McMaster University

McMaster CSE Seminar, March 2022

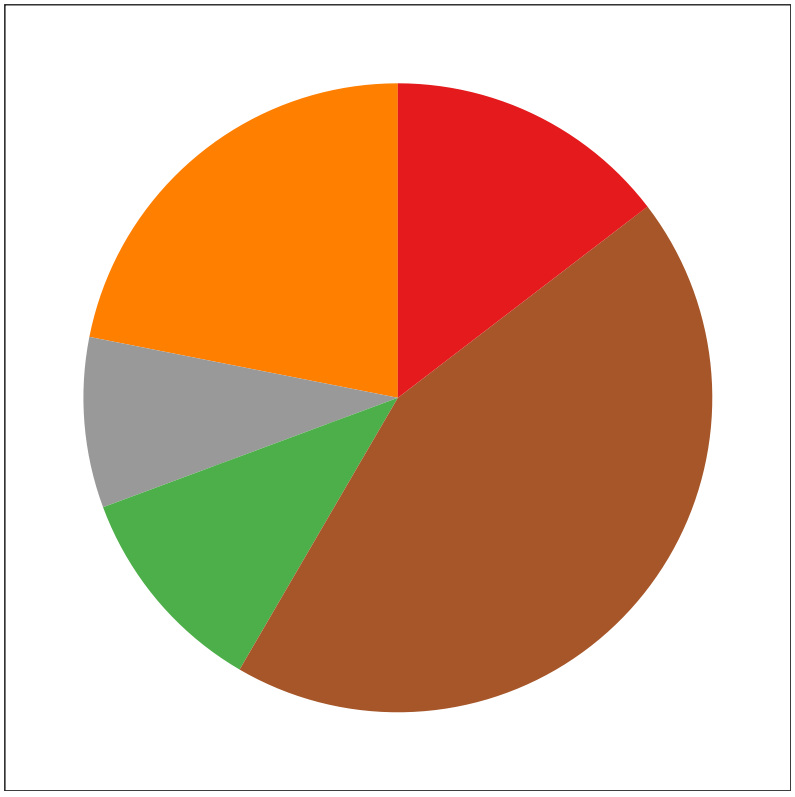
Why do we measure biodiversity?

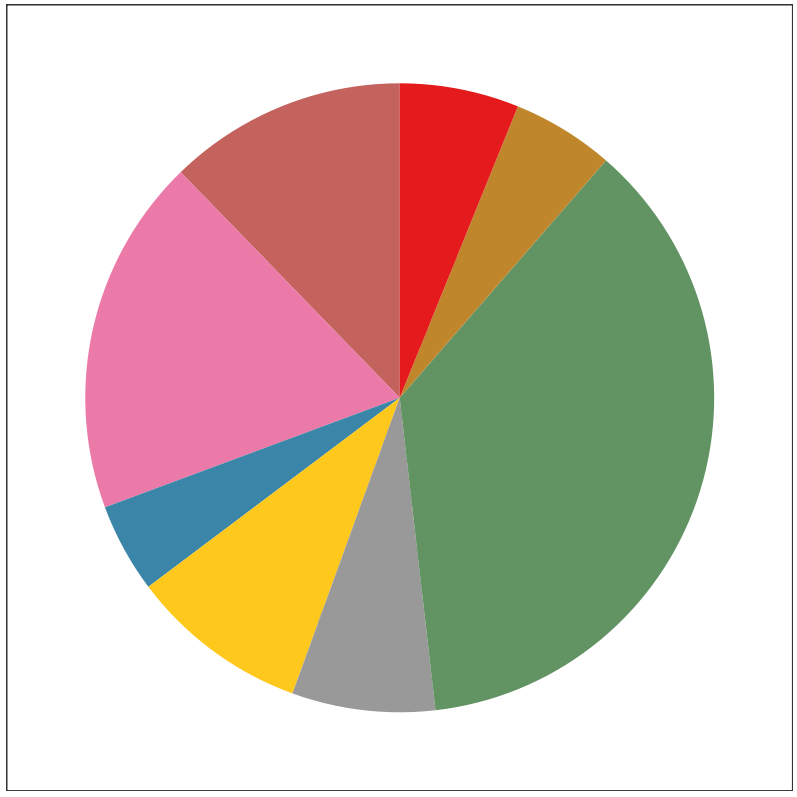
- It seems practically important:
  - How does it correlate with ecosystem function?
  - What can we do to protect it?
- It seems scientifically important
  - Maybe we can understand the world better even if we can't protect it
- But there's no consensus on how best to measure it, or how to compare the metrics themselves

Community diversity

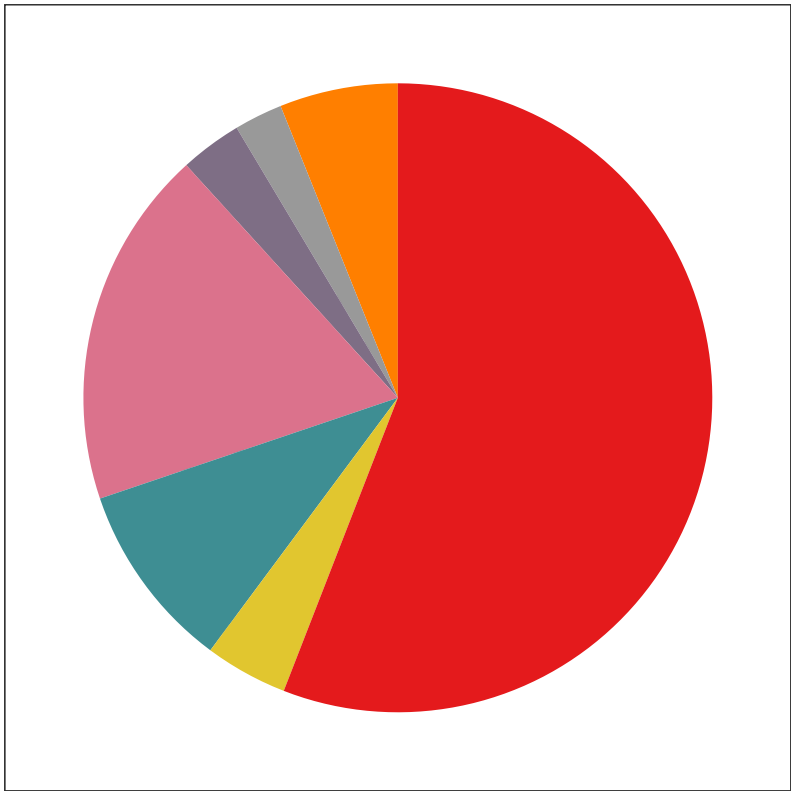
- I am going to be talking about community diversity, often called α-diversity
  - How different are individuals in a given sampling unit?
- I am not going to be talking about diversity of units themselves, often called β-diversity
  - How different are two sampling units from each other?
- When I compare two different sampling units, I will be asking which one is more diverse, not how different they are

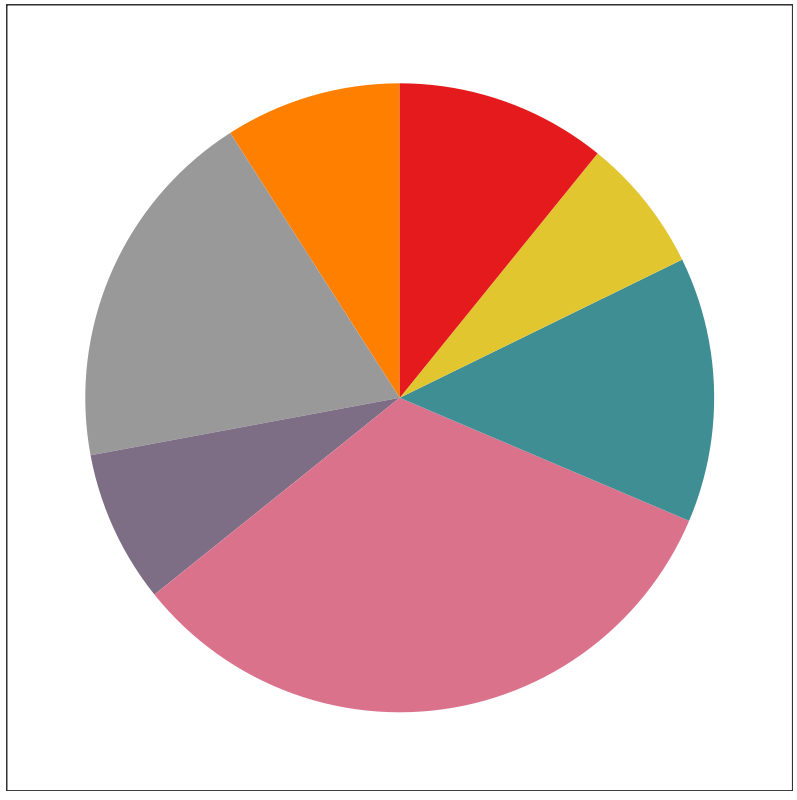
Comparing diversity



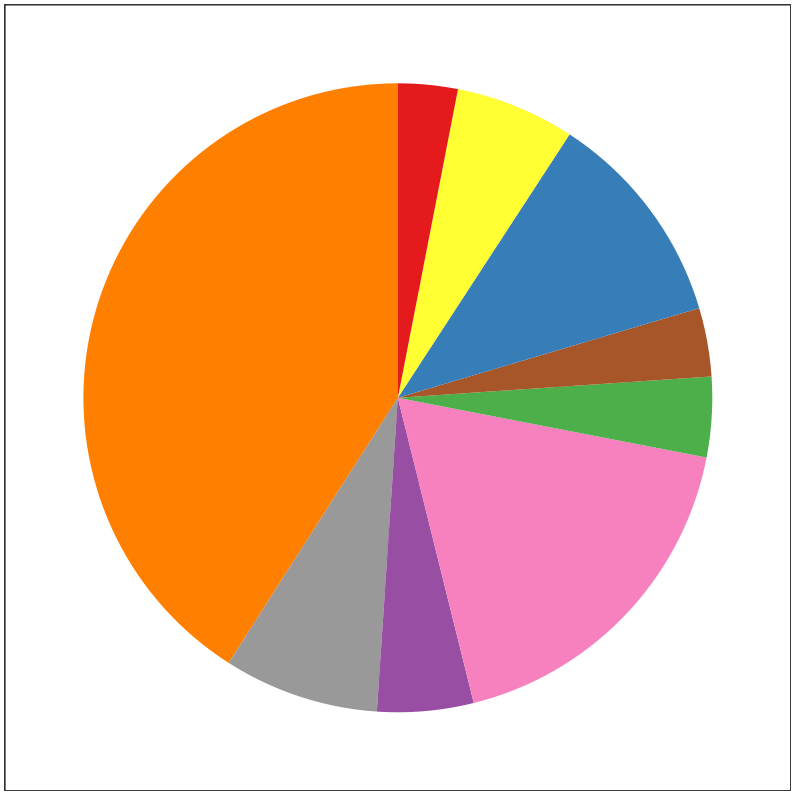


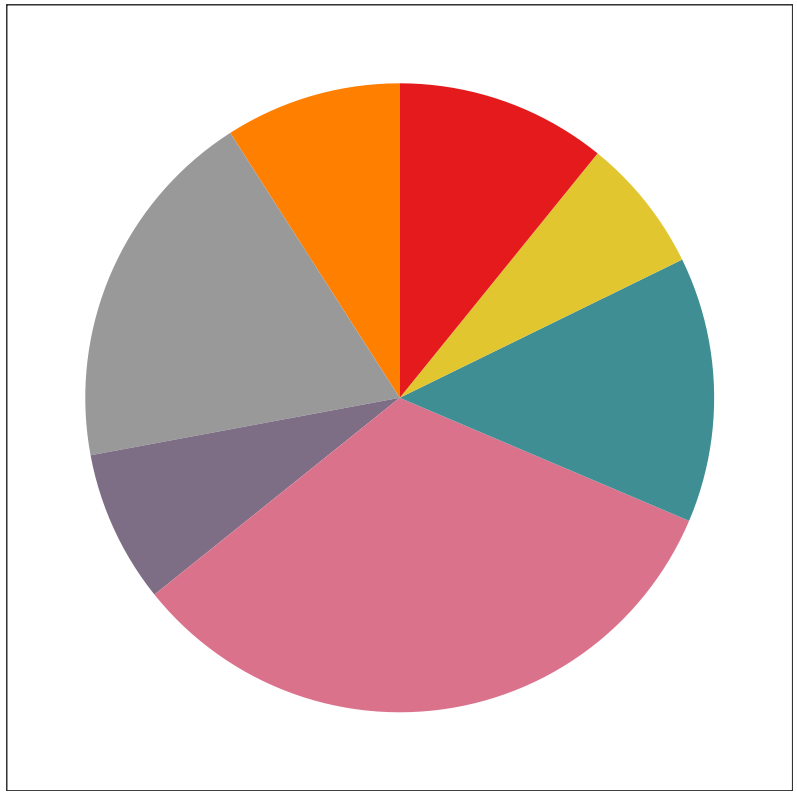
Comparing diversity





Comparing diversity





Richness

- How many species are there?
  - This is apparently what aliens want to know
- Can we ever know?
  - Other "diversity" measures combine diversity with relative abundance
  - ... but so does this one!
  - In practice

Simpson diversity

- What is the probability that two different individuals are from different species?
- $$1 - \sum p_i^2$$

Shannon diversity

- How much Shannon information is there in a particular observation?
- How many distinct ways of combining individuals?
- $$\sum -p \log(p)$$

Richness is unstable

- If we assume an effectively infinite population, it's impossible to get good estimates of population richness
  - We don't know what we don't know

Simpson diversity *seems* stable

- We can get an unbiased estimate of Simpson diversity from two individuals!
- Recall: unbiased means that the expected value of the estimate is equal to the true value of the estimand
- This is equivalent to saying we get an "unbiased" estimate of an unfair coin from a single flip
- Of course, we'll always have more than 2 individuals, but we can always get an unbiased estimate

Effective number of species

- The value of Simpson or Shannon diversity is hard to interpret.
  - Or at least, easy to misinterpret.
- We can ask how do they map to an effective number of species?
  - On a given scale, the observed diversity is equivalent to E *evenly distributed* species

Hill diversity

$$H_q = \left(\sum p_i^q \right)^{1/1-q}$$

- Take limits as necessary
- Spans from richness ($q = 0$) to Hill-Shannon ($q = 1$) to Hill-Simpson
- Gives an effective species number
- Decreases with q

Rarity perspective

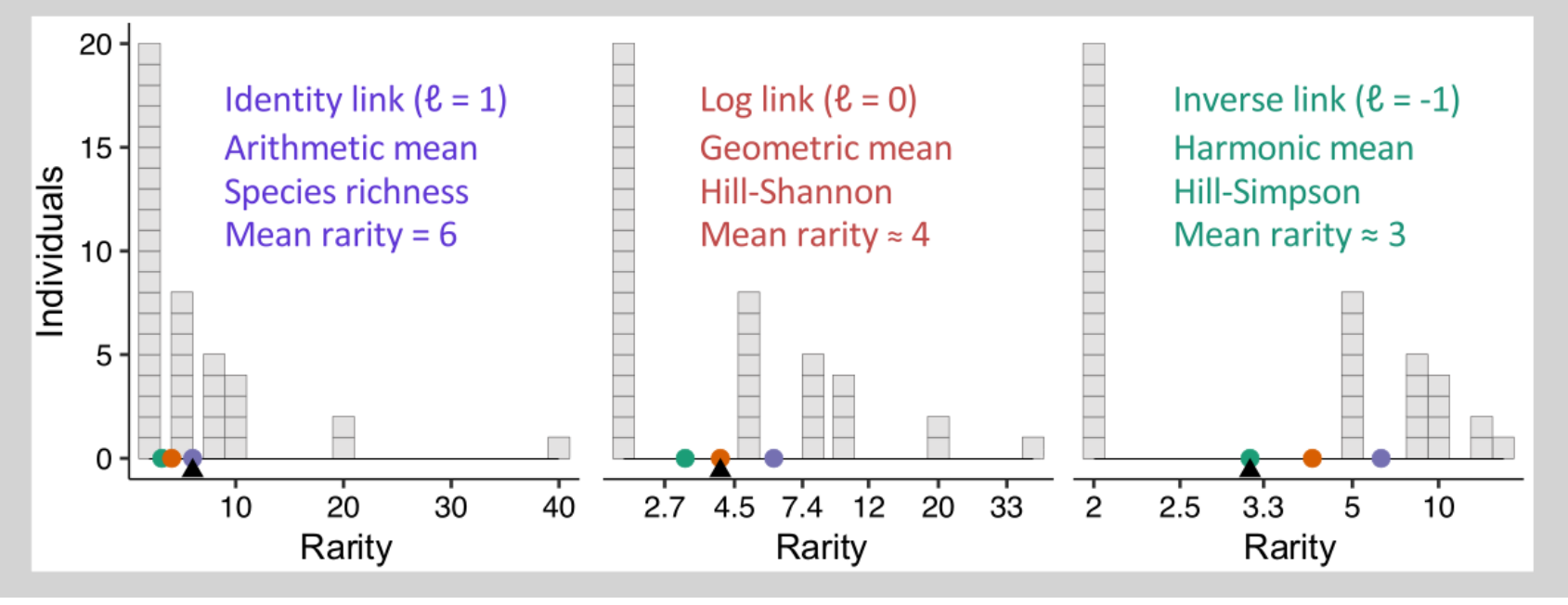
$$D_\ell = \left(\sum p_i (r_i)^\ell \right)^{1/\ell},$$

- where $r_i = 1/p_i$ is the rarity
- $\ell = 1 - q$ has a natural interpretation in terms of power means:
  - $\ell = 1$ is the arithmetic mean, corresponds to richness
  - $\ell = 0$ is the geometric mean, corresponds to Hill-Shannon
  - $\ell = -1$ is the harmonic mean, corresponds to Hill-Simpson

Leverage of rare species

- The higher values of ℓ are often described as giving more *weight* to rare species
- But this doesn't correspond to an actual weighting
- We use "leverage" to describe the increasing importance of rare species
- And illustrate with "seesaw" plots

DiversitySeesaw

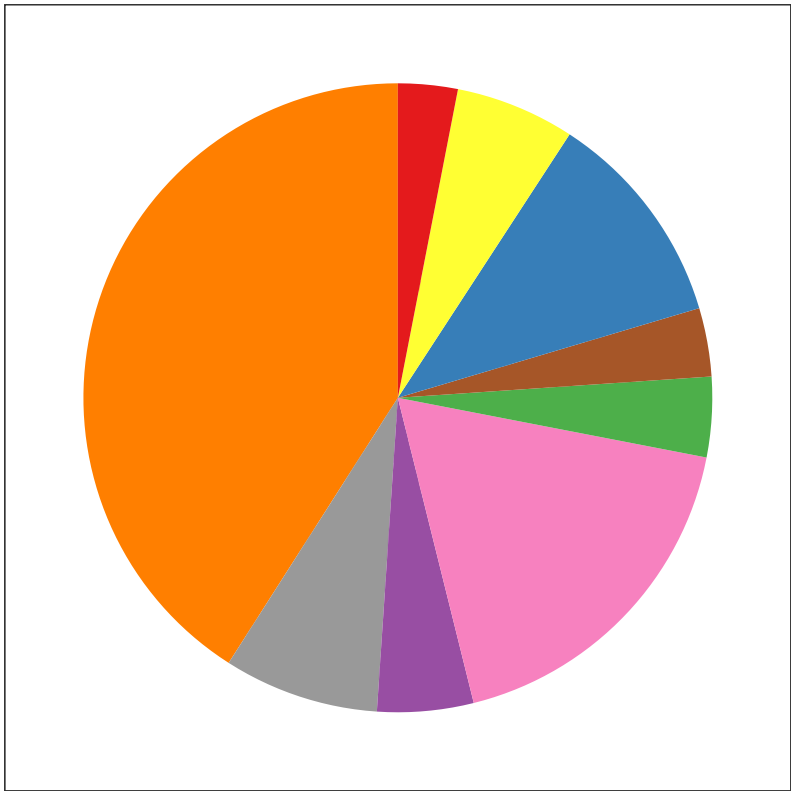


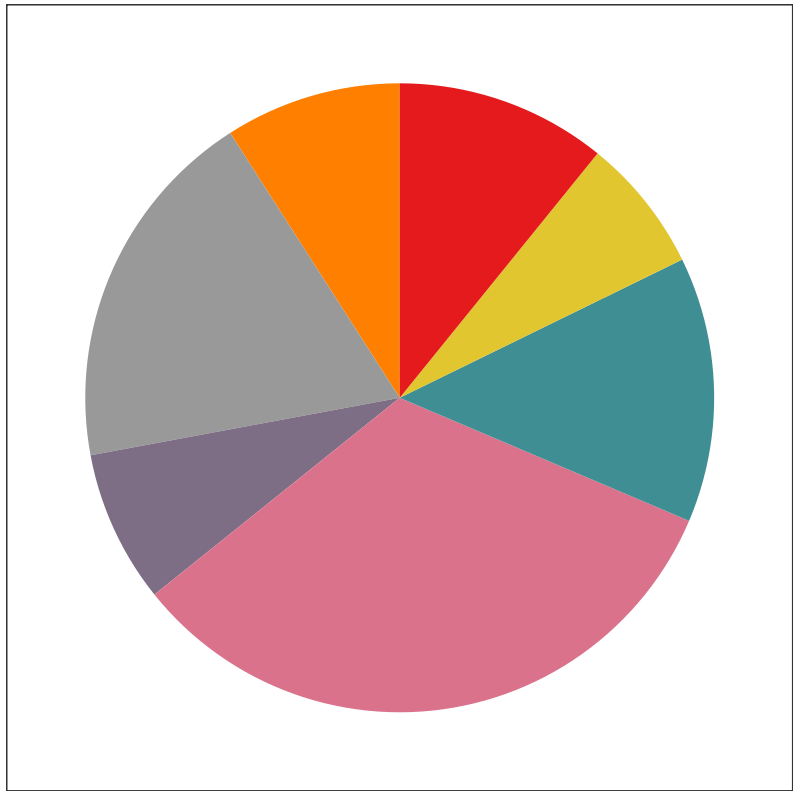
`https://github.com/mikeroswell/MeanRarity`

Is Hill-Shannon just better?

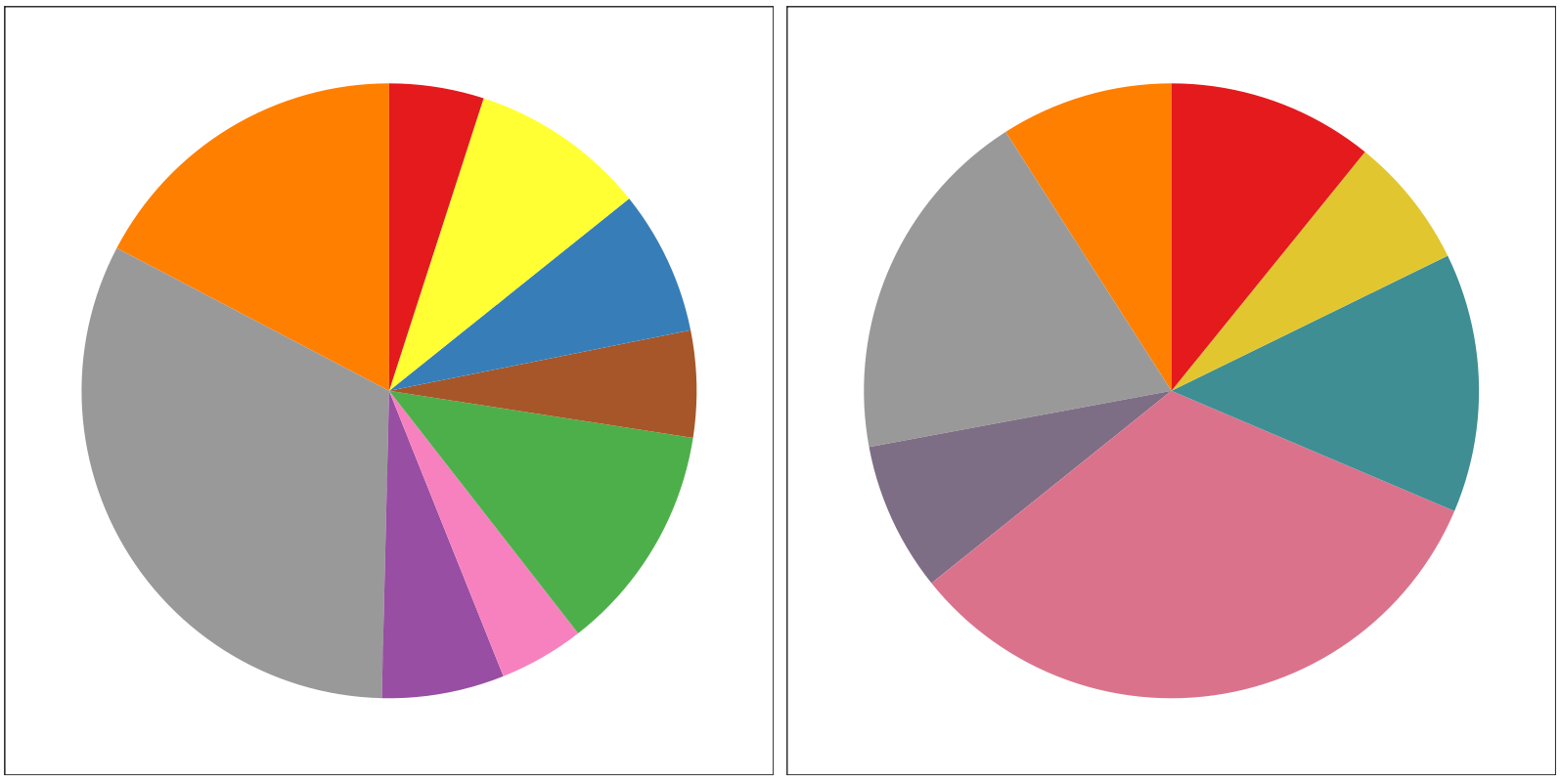
- Geometric mean corresponds to ratio-based distance: 1:5 as 5:25
- E.g., The difference between 20% abundance (rarity of 5) and 10% abundance
  - Hill-Simpson: the same as the difference between 10% and 0% (∞ rarity)!
  - Richness: the same as the difference between 10% and 6.7%
    - or 2% and 1.8%
  - Hill-Shannon: the same as the difference between 10% and 5%

Shannon equivalence

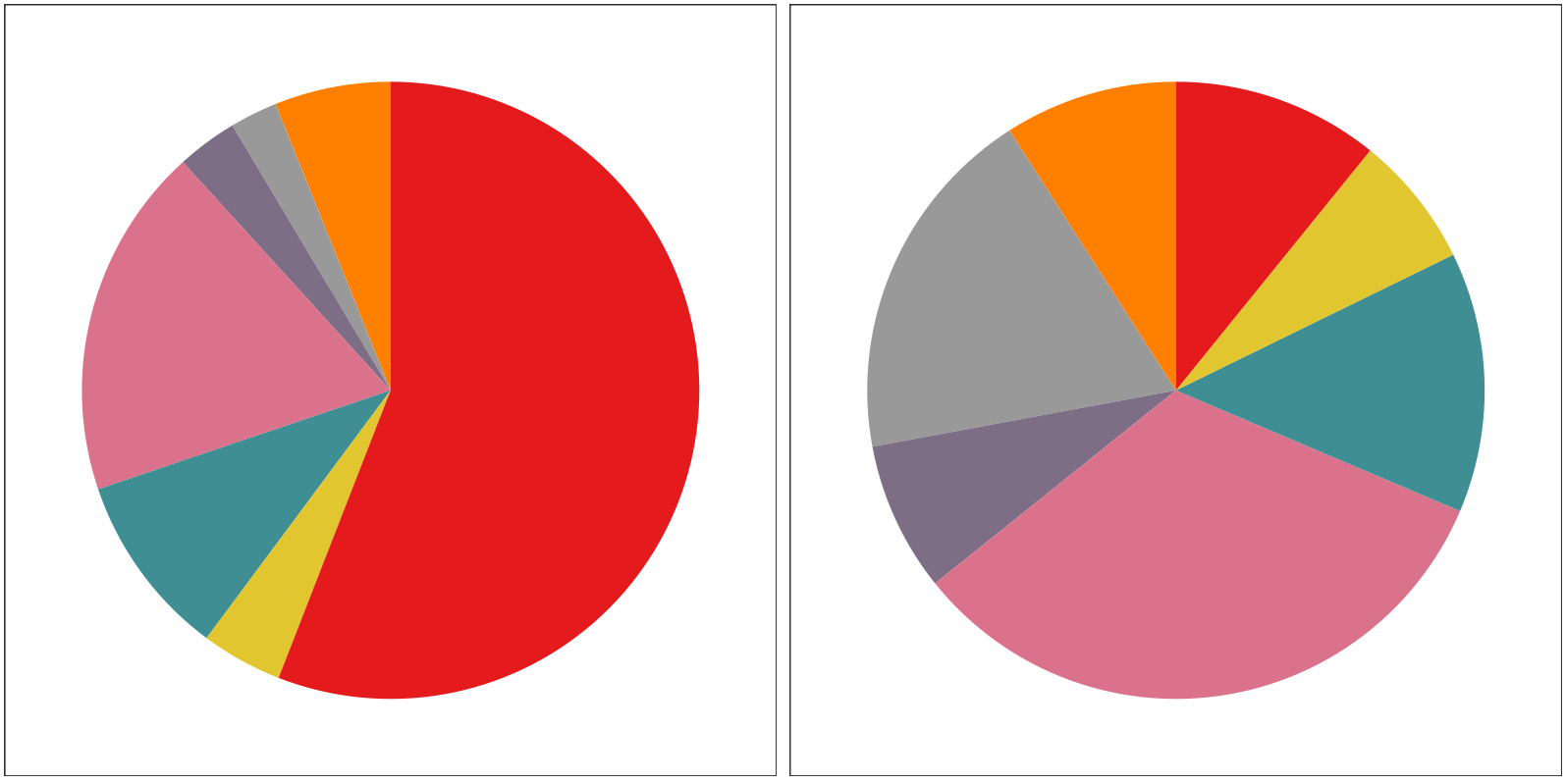




Simpson equivalence



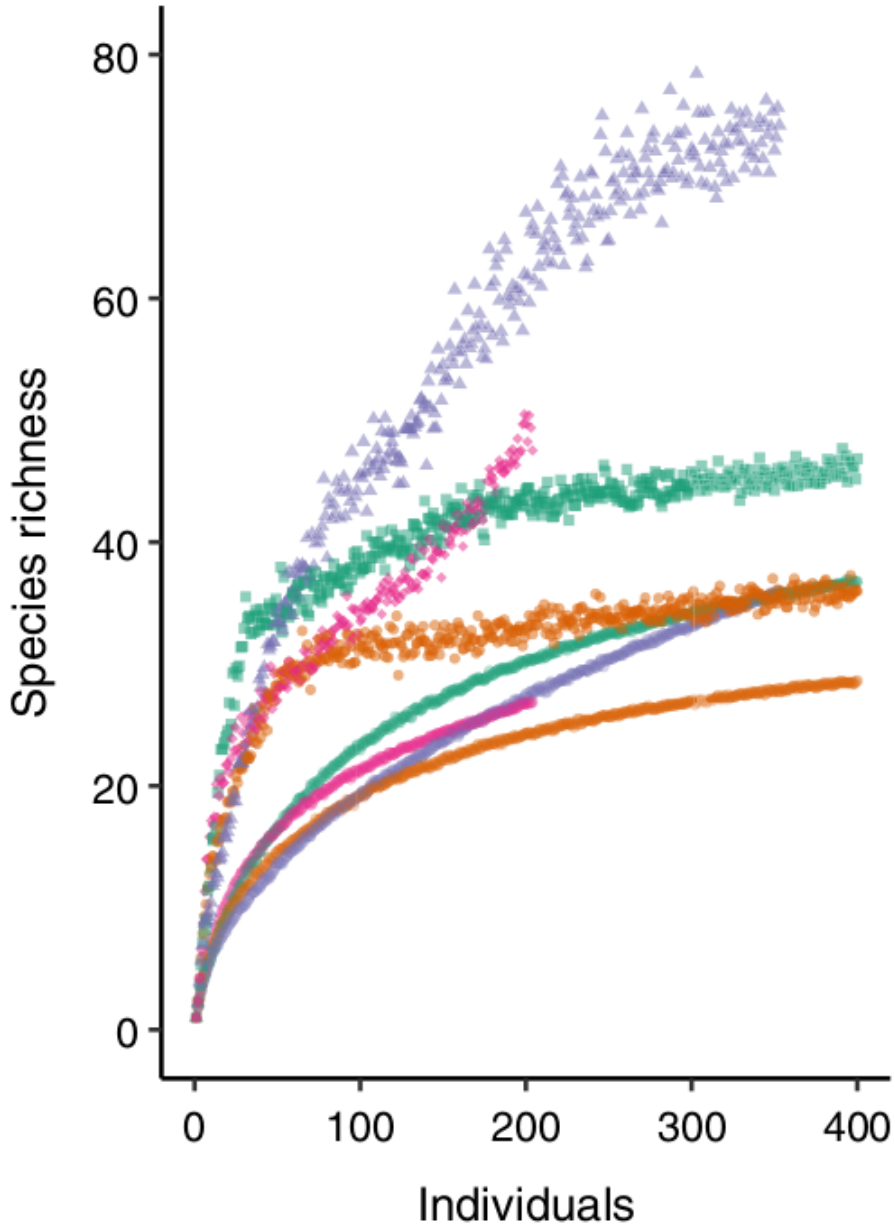
Richness equivalence



How to compare *sampled* communities?

- We don't know the true diversity
- Rarefaction
  - Compare at equal number of individuals sampled
  - Or at equal "coverage"
- Asymptotic estimation
- Much of this goes back to the work of Turing, who was interested (effectively) in diversity metrics on coded messages and ciphertext

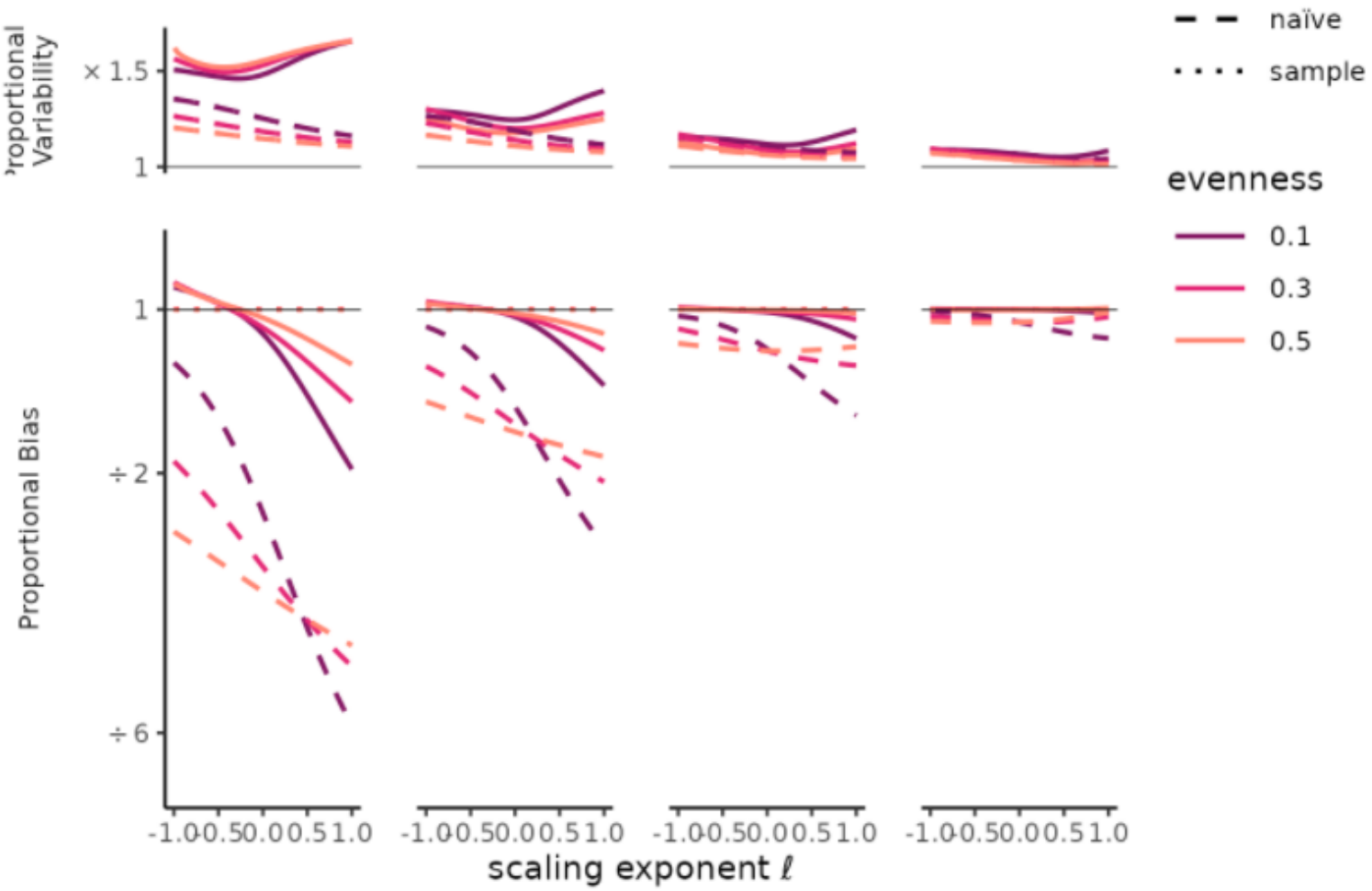
Comparing observed diversity



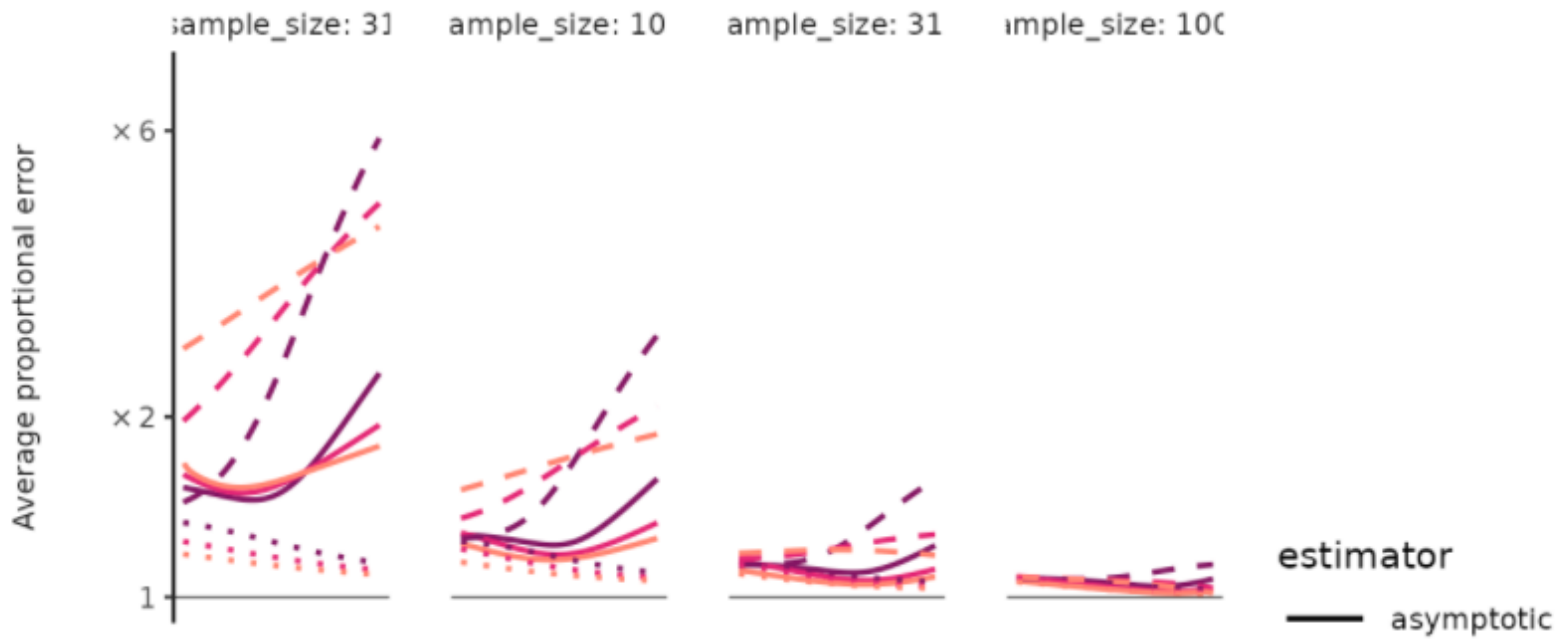
Ongoing work

- How *biased* are different asymptotic or sample-based measures?
- How *variable* are different asymptotic or sample-based measures?
- `https://github.com/mikeroswell/MeanRarity`

Bias-variance tradeoff



Bias-variance tradeoff



Future work

$$\hat{D}_\ell = \left(\sum p_i (\hat{r}_i)^\ell \right)^{1/\ell},$$

- Does the rarity perspective provide insight into estimation methods?
- p and r are typically treated the same, but actually play different roles:
  - if we had good estimates of $\hat{r}$, then $\hat{D}$ becomes an unbiased estimator
  - not *necessarily* a low-variance estimator
- Does the rarity perspective allow us to expand our concept of coverage?

Thanks

- Collaborators: Mike Roswell, Rae Winfree and others
- Organizers
- Audience