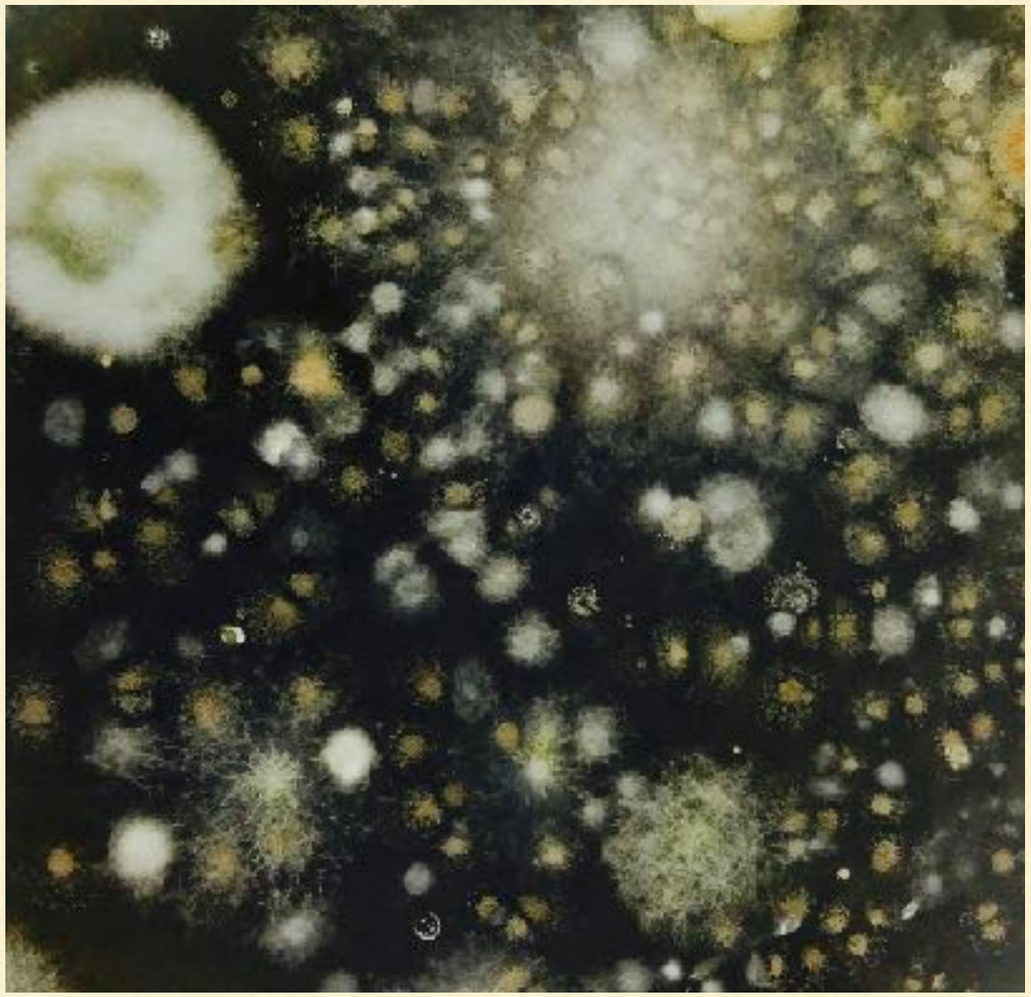


DIVERSITY

Research Group: Statistical Diversity Lab

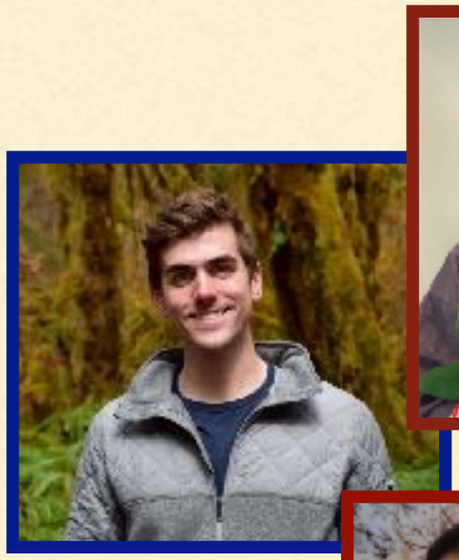
PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

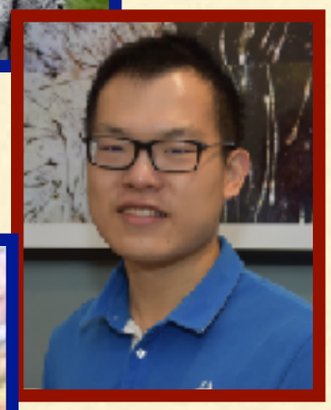
@AmyDWwillis adwillis@uw.edu

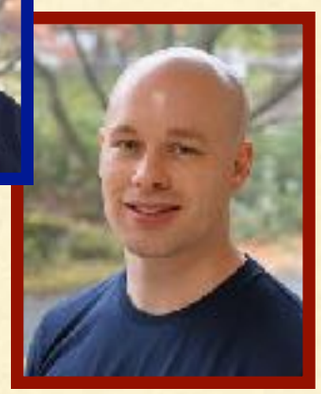
Photo credit: T.D. Berry, Whitman lab, UW Madison

PEOPLE

statistical diversity lab







Pauline Trinh — bioinformatics, human-animal-microbe interactions

Bryan Martin — relative abundance, compositional data

Kendrick Li — networks, taxonomy

David Clausen — replication and data quality in microbiome science

Alex Paynter — diversity and regularised ML

WHAT IS DIVERSITY?

- Low dimensional summaries of entire communities
  - α-diversity: one community
  - β-diversity: multiple communities
  - gamma diversity: totally meaningless

ALPHA DIVERSITY

	A	B	C	D	E	F	G	
1		Sample01	Sample02	Sample03	Sample04	Sample05	Sample06	Sam
2	Microbe01	1	1	2	2	0	0	
3	Microbe02	0	0	0	0	0	0	
4	Microbe03	0	0	3	1	2	0	
5	Microbe04	1	1	0	6	2	0	
6	Microbe05	0	0	0	0	0	1	
7	Microbe06	2	0	0	1	0	1	
8	Microbe07	28	7	24	12	15	100	
9	Microbe08	0	0	0	1	0	0	
10	Microbe09	0	0	0	1	0	0	
11	Microbe10	860	224	120	1107	440	1089	
12	Microbe11	1	0	0	0	0	2	
13	Microbe12	22	4	4	9	7	8	
14	Microbe13	3	0	0	0	0	1	
15	Microbe14	1	0	0	0	0	0	
16	Microbe15	0	0	1	0	0	1	
17	Microbe16	3	1	1	0	1	0	
18	Microbe17	0	0	0	0	0	0	
19	Microbe18	0	0	0	0	0	1	
20	Microbe19	0	26	1	2	0	0	

DIVERSITY & STATISTICAL THINKING

- On Tuesday we learnt that statistics concerns the microbial community
  - Computational biology/exploratory approaches concern the sample
- Key distinctions
    - community versus sample
    - parameter versus data

DIVERSITY & PARAMETERS

- There are multiple choices to make when talking about diversity
- Not just Shannon vs Simpson, UniFrac vs Jaccard...
- Most people talk about strain-level diversity (even without strain-level data resolution)

DIVERSITY & PARAMETERS

- There are multiple choices to make when talking about diversity
  - Which taxonomic level? (strain/species/genus...)
  - Which diversity index?
  - Which estimate of the diversity index?

DIVERSITY & PARAMETERS

- There are multiple choices to make when talking about diversity
  - Which taxonomic level? (strain/species/genus...)
  - **Which diversity index?**
  - **Which estimate of the diversity index?**

STRUCTURE OF THE LECTURE

- α-diversity, parameters, how to estimate them
- Hypothesis testing for α-diversity
- β-diversity
  - Some inference discussed in this session
  - More commonly used for exploratory analyses; will be covered in *Exploratory multivariate visualization*

ALPHA DIVERSITY

- Any function that maps from
  - composition vectors, OR
  - composition vectors and trees
- is a valid α-diversity measure ("index")

REALITY CHECK

- How can you summarise 10,000-dimensional compositional data in a meaningful way?

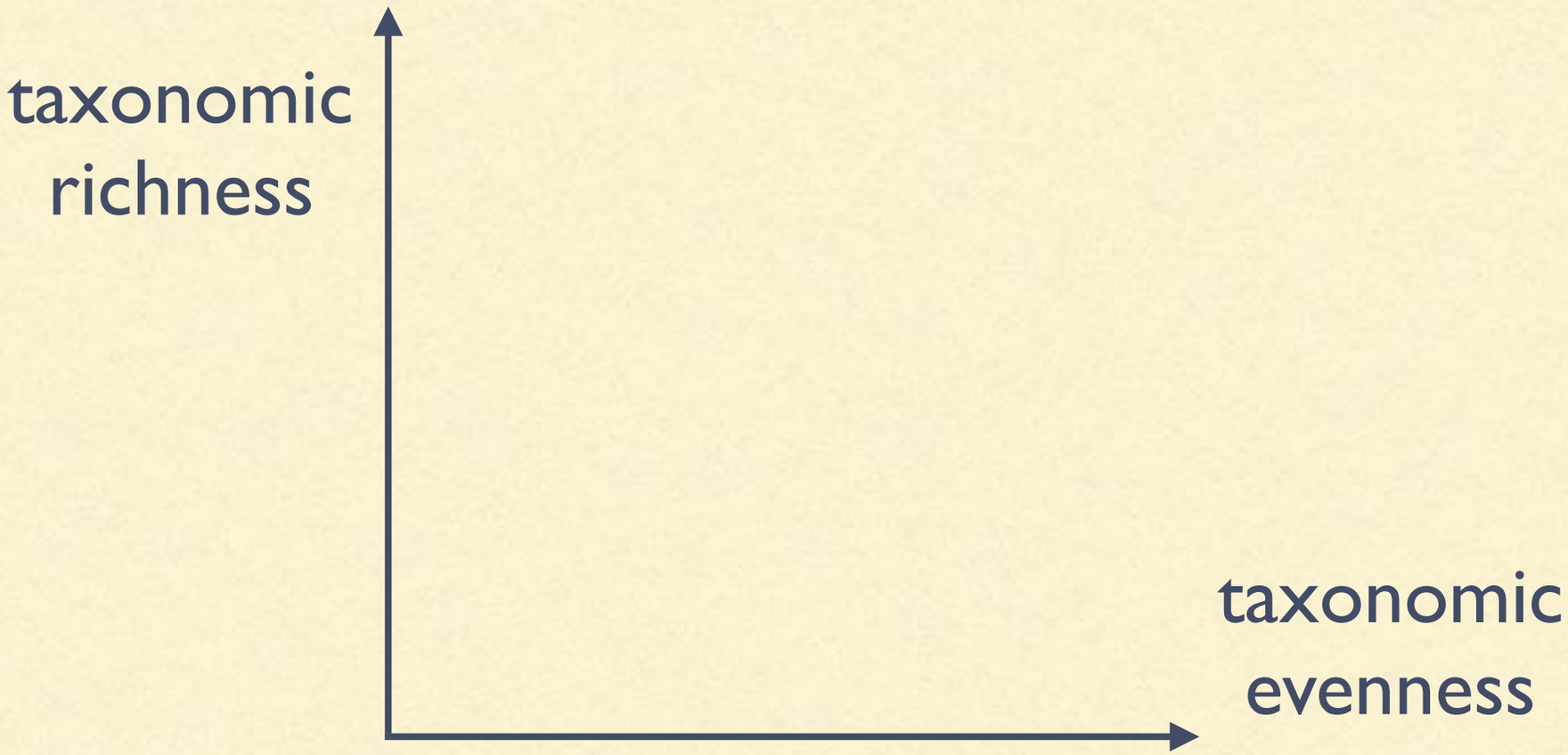
- Shannon, Simpson, Inverse Simpson, Chao1, sample richness…

- "But we need taxonomy!" BWPD…

> How can you summarise 10,000-dimensional compositional data with a single number without losing any information?

YOUR CHOICE

- Think: What difference do you want to highlight?



YOUR CHOICE

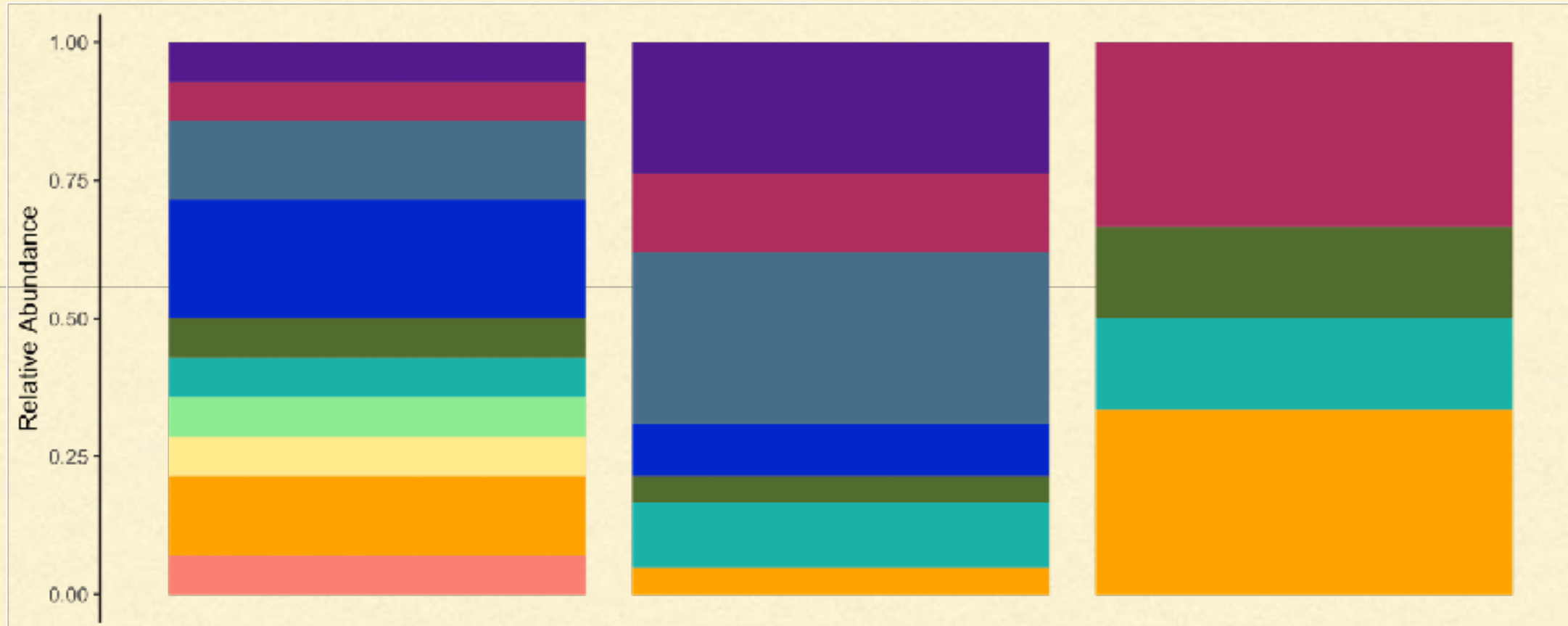

taxonomic
richness
species
richness
Shannon
Simpson
Shannon's E
taxonomic
evenness

YOUR CHOICE

This is a question of *parameter choice:*
Which parameter highlights the differences I care about?

MY CHOICE

- Deeply personal decision
- My favourite combinations are
  - (richness, Shannon's E)... in an ideal world
  - (Shannon, Shannon's E)... with questionable rare taxa
- Why? Richness and evenness are 2 completely different things, and both are important



Richness	10	7	4
Shannon	2.21	1.75	1.33
Evenness	0.96	0.90	0.96
Simpson's	0.88	0.80	0.72
Inverse Simpson's	8.17	4.98	3.60

THE PROBLEM

- If we repeated our sequencing experiment to infinite effort, we would find
  - the true taxonomic richness C
  - the proportions of each taxon: $p_1, p_2, \ldots, p_C$
- In practice, we don't observe the entire community, just a sample from it
  - we don't know C or $p_1, p_2, \ldots, p_C$

DEFINITIONS

- If we repeated our sequencing experiment to infinite effort, we would find

- the true taxonomic richness C

- the proportions of each taxon: $p_1, p_2, \ldots, p_C$

DEFINITIONS

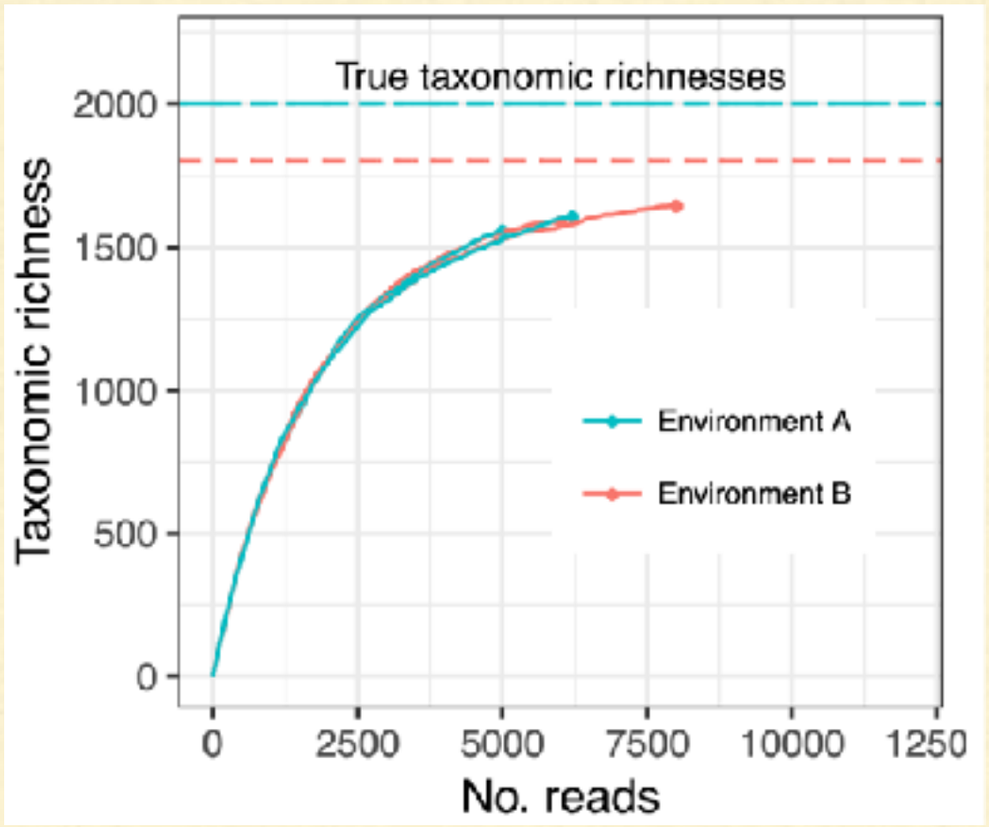
- Some examples of α-diversity measures include
  - Species richness: C
  - Simpsons index: $\sum_{i=1}^{C} p_i^2$
  - Shannon diversity: $-\sum_{i=1}^{C} p_i \ln p_i$
  - Shannon's E: $\frac{-\sum_{i=1}^{C} p_i \ln p_i}{\ln C}$

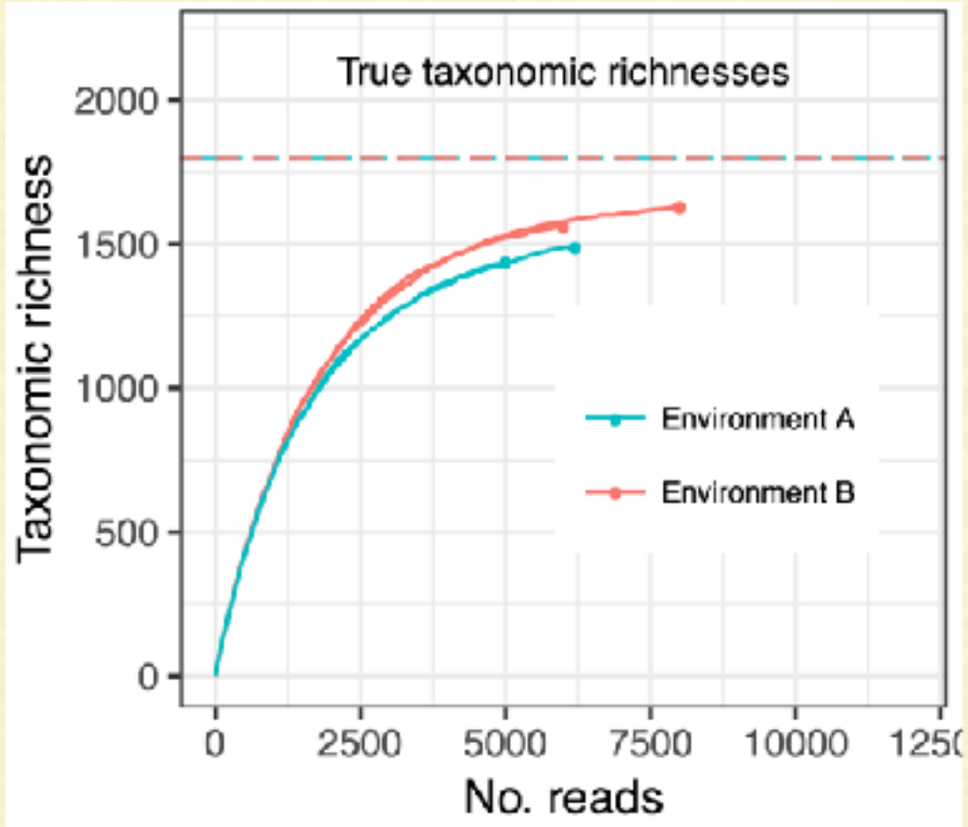
THE "CLASSICAL" APPROACH

- Substitute the observed abundances $\hat{p}_1, \ldots, \hat{p}_c$ for the unknown, true abundances $p_1, p_2, \ldots, p_c$ and pretend nothing happened
  - e.g. Estimate the richness with: $c = \#\{i : \hat{p}_i \neq 0\}$
  - e.g. Estimate the Simpsons index: $\sum_{i=1}^{c} \hat{p}_i^2$

ONE PROBLEM (OF MANY)

- In the first of these cases, the estimate *underestimates* the true parameter of interest; in the second, the estimate *overestimates*





- There is nothing wrong with the parameters themselves, but we are choosing a bad way to estimate them

A FIX

- There are 2 things we need to fix here
  - The bias (under/overestimation)
  - The variance (how big are the error bars)
- The best way to fix these depends on the setting
  - We'll discuss *species richness* and *not-species-richness* separately

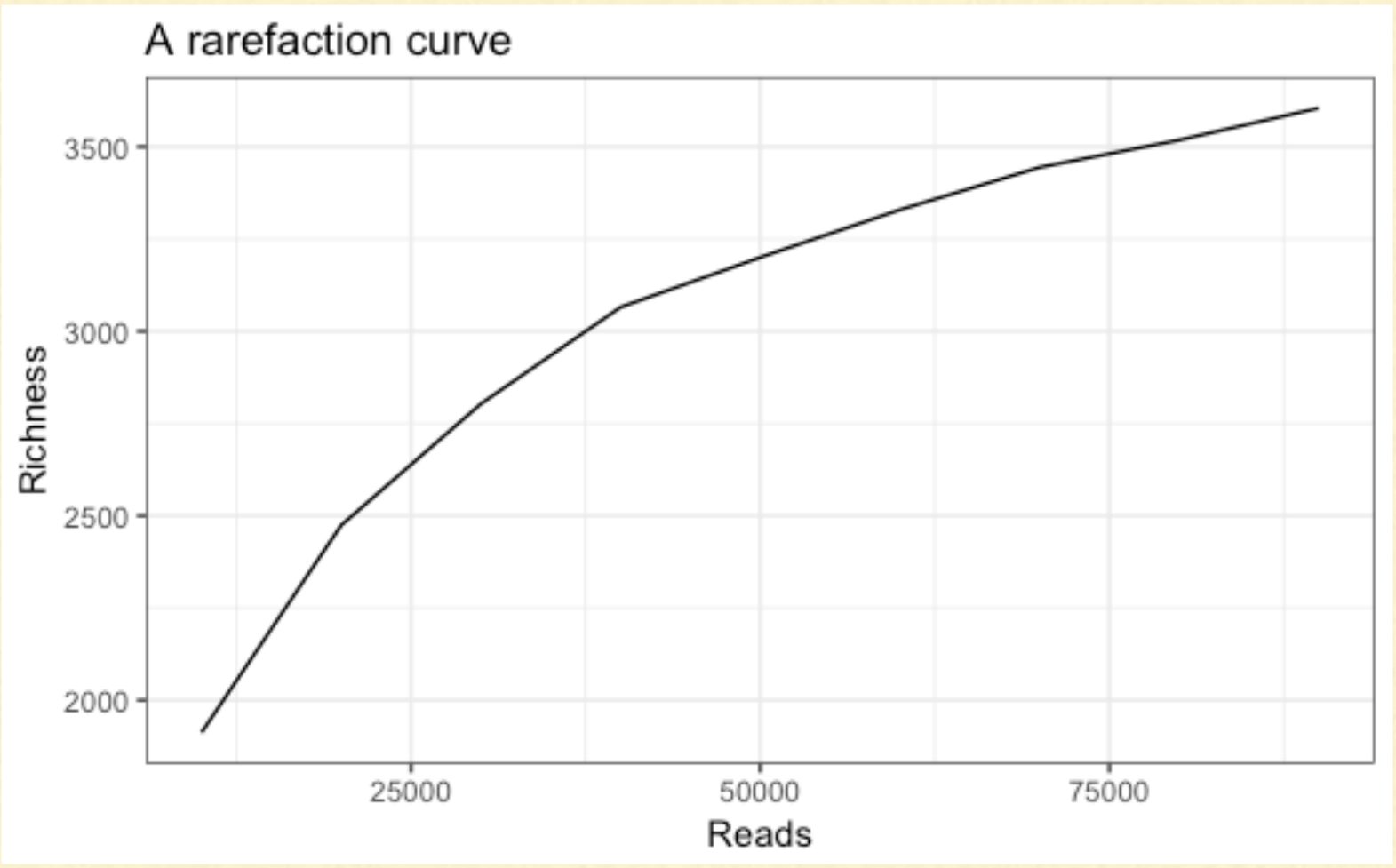
SPECIES RICHNESS

- The "species problem": how many species were missing from the sample
- Two main avenues
  - Comp Bio approach: rarefy
  - Mixed-Poisson: Chao1, Chao-Bunge, CatchAll...
  - Not MP: Kemp models ("breakaway")

SPECIES RICHNESS ESTIMATES

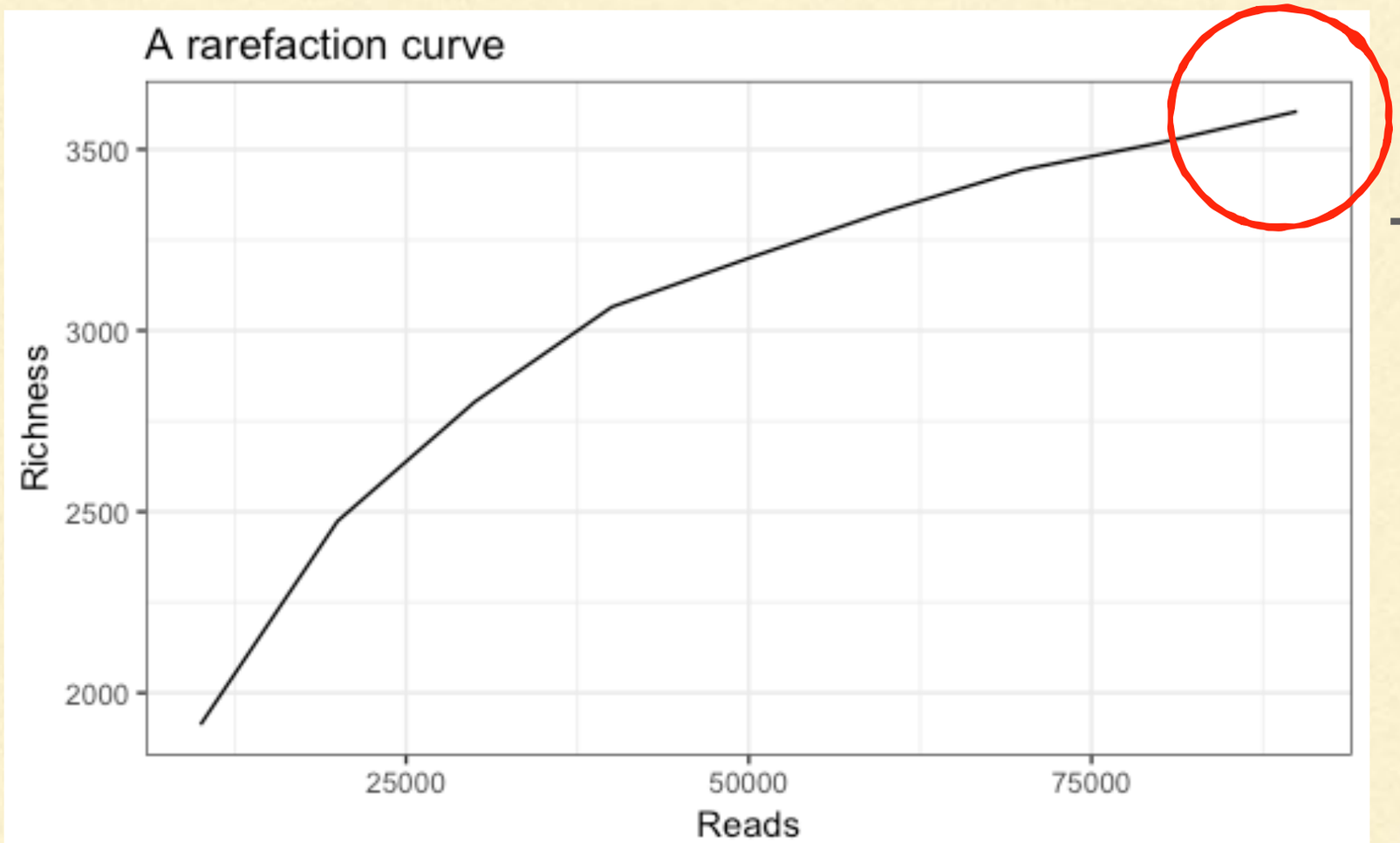
- C = species richness of entire community
    - parameter
- c = species richness of sample
    - estimate (a bad one)
- **Is the bias of c (for estimating C) positive or negative?**

RAREFACTION: DON'T



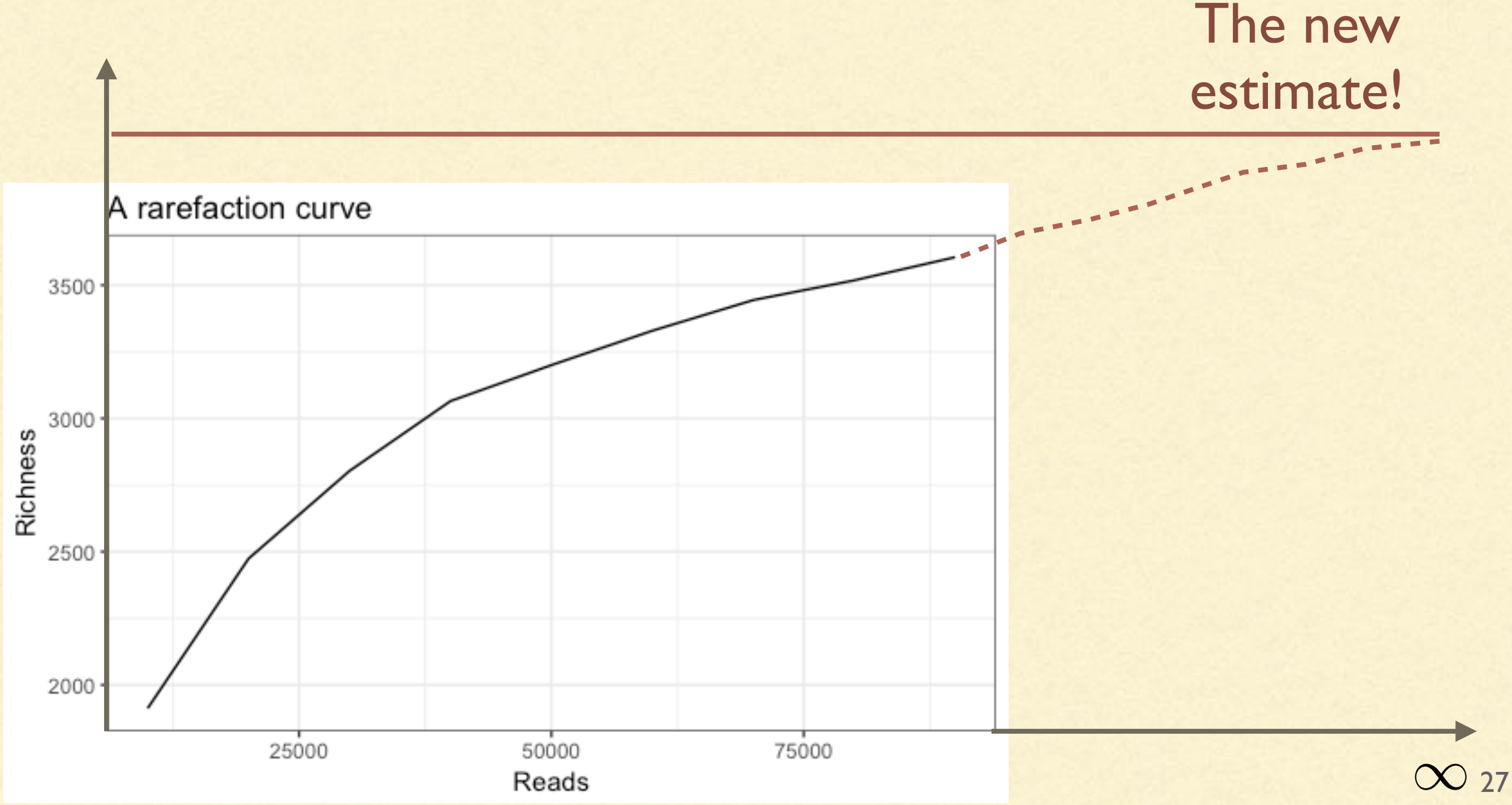
RAREFACTION: DON'T

```
  [1]  350 1796  280   33    0   63  141  420   63   26   52   40   71    9 1091   56  464
 [18]  825  206   64  200  119   25   63   76 1062   98   68   35   86  261  692  814  248
 [35]   28  308 1191  306   91  218   84   30  142   26  272    7   76  686   85    0  559
 [52]   84   96  320  120   31    0   25   35   55   27  792    0   81    4   21   76  255
 [69]   10   16  713  474   91  405  270  263   60    1  207    7   11  150  259   69   40
```



The only real data point

RAREFACTION: DON'T



RULES BY STATISTICIANS

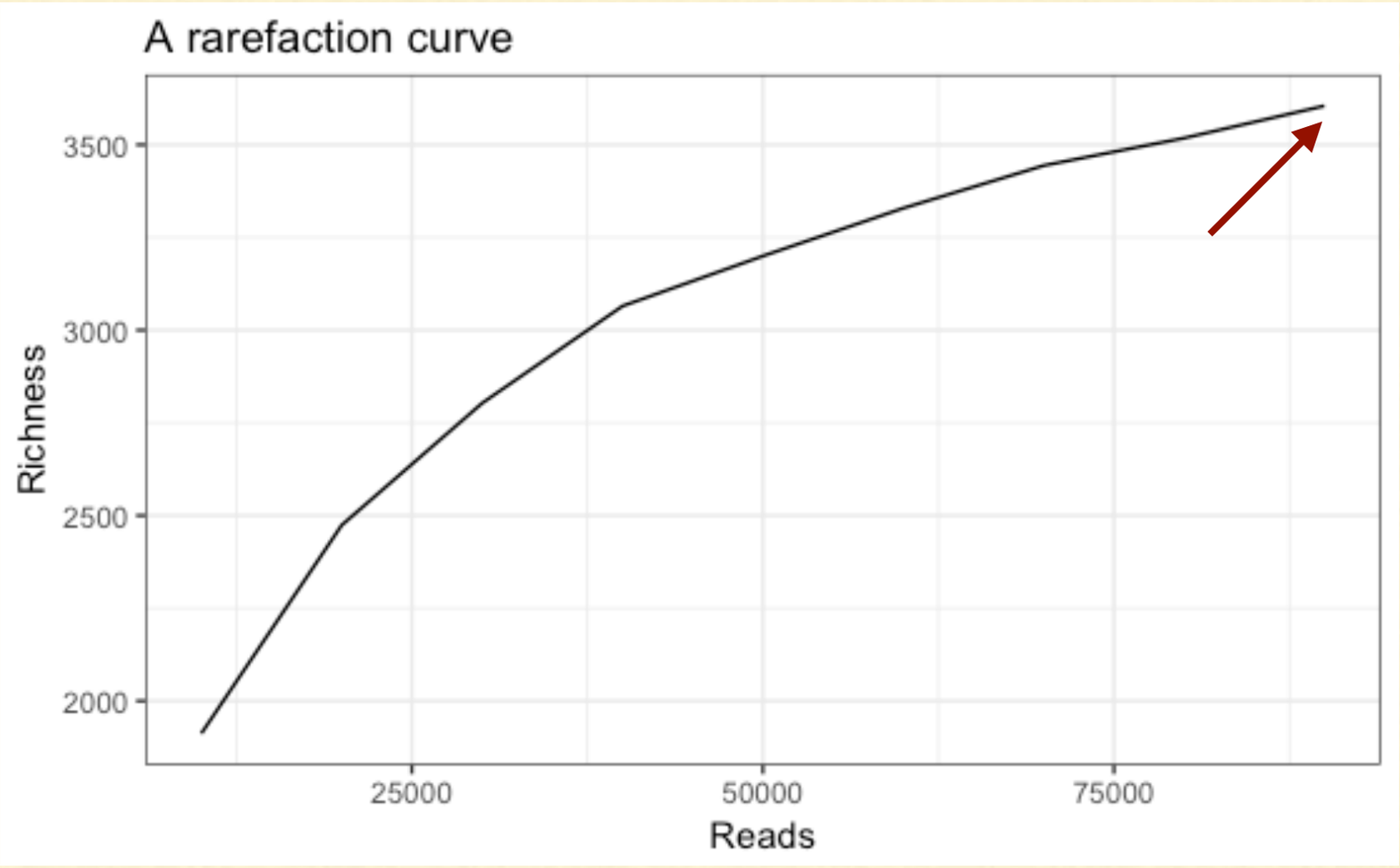
No extrapolating abundance curves! Use species richness estimates instead!

- Willis (2017), Extrapolating abundance curves has no predictive power…

  No rarefying to compare across samples! Use reasonable precision-adjusting models instead!

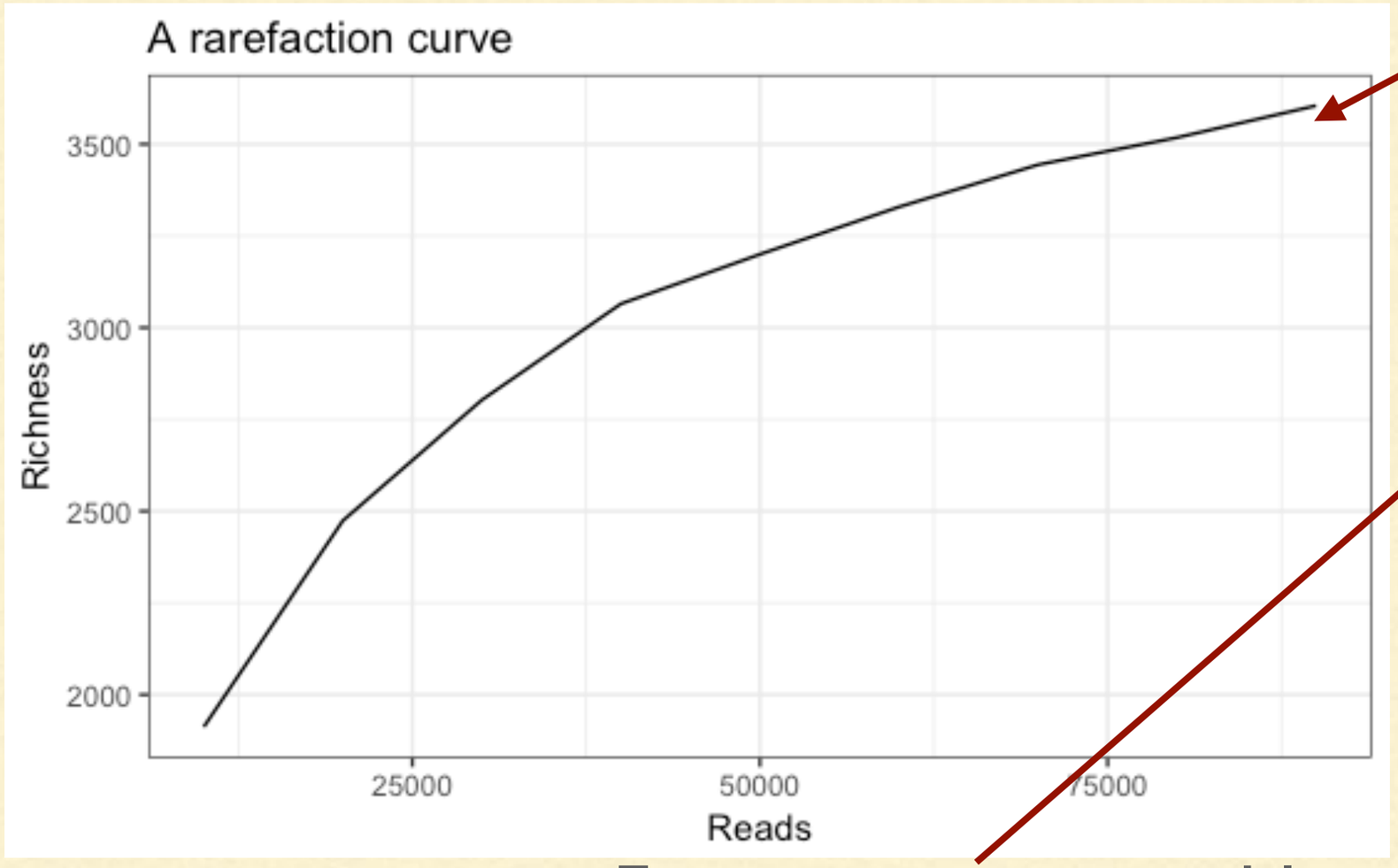
- McMurdie & Holmes (2014), Waste Not Want Not

WHAT TO DO INSTEAD?



- At this point, you have the maximum amount of data
- This data has structure
- You can use this structure to upscale better

WHAT TO DO INSTEAD?



We have more structure here than just a richness of 3590!

1	516
2	472
3	334
4	252
5	208
6	139
7	109
8	104
9	96

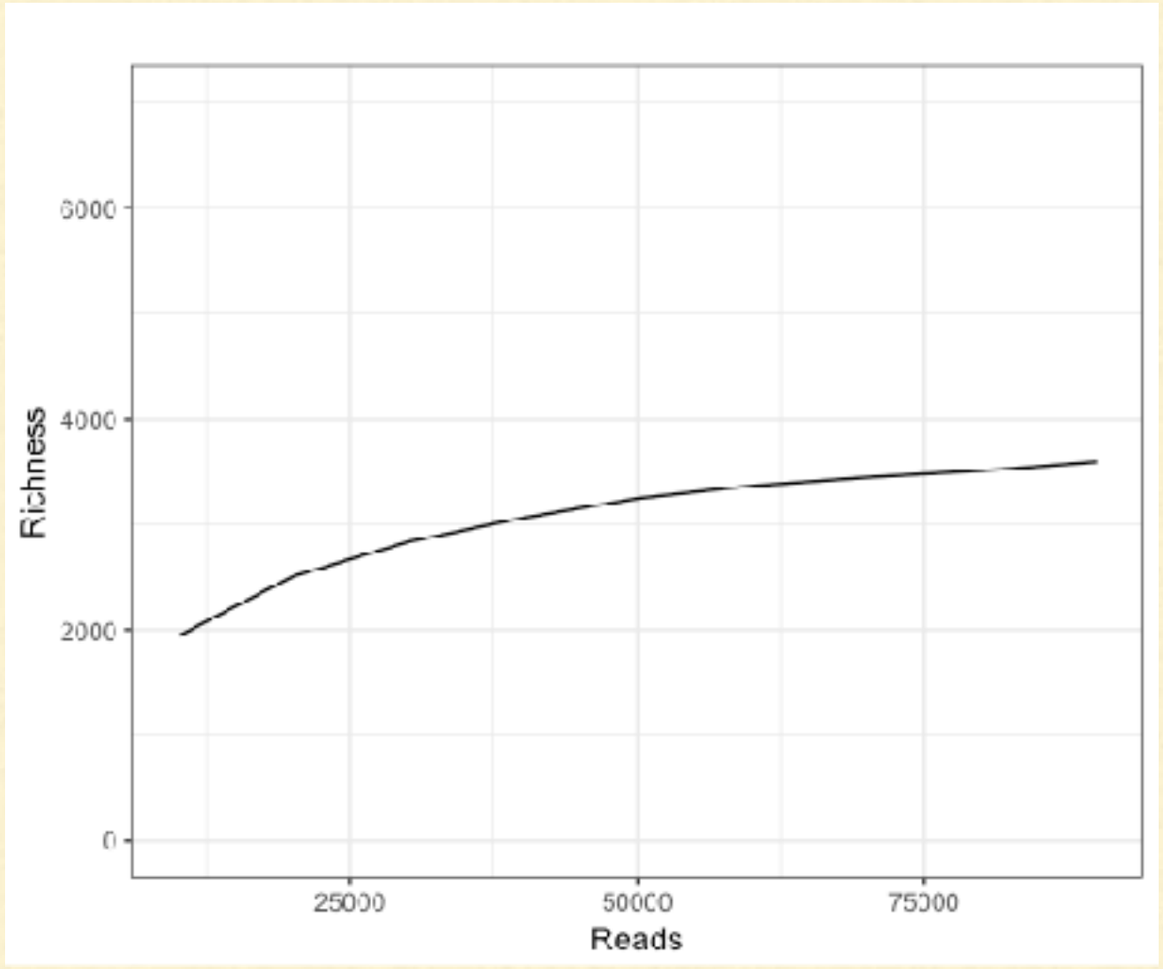
Model this data to get an estimate of richness

Frequency count table:
516 taxa seen once, 472 seen twice, …

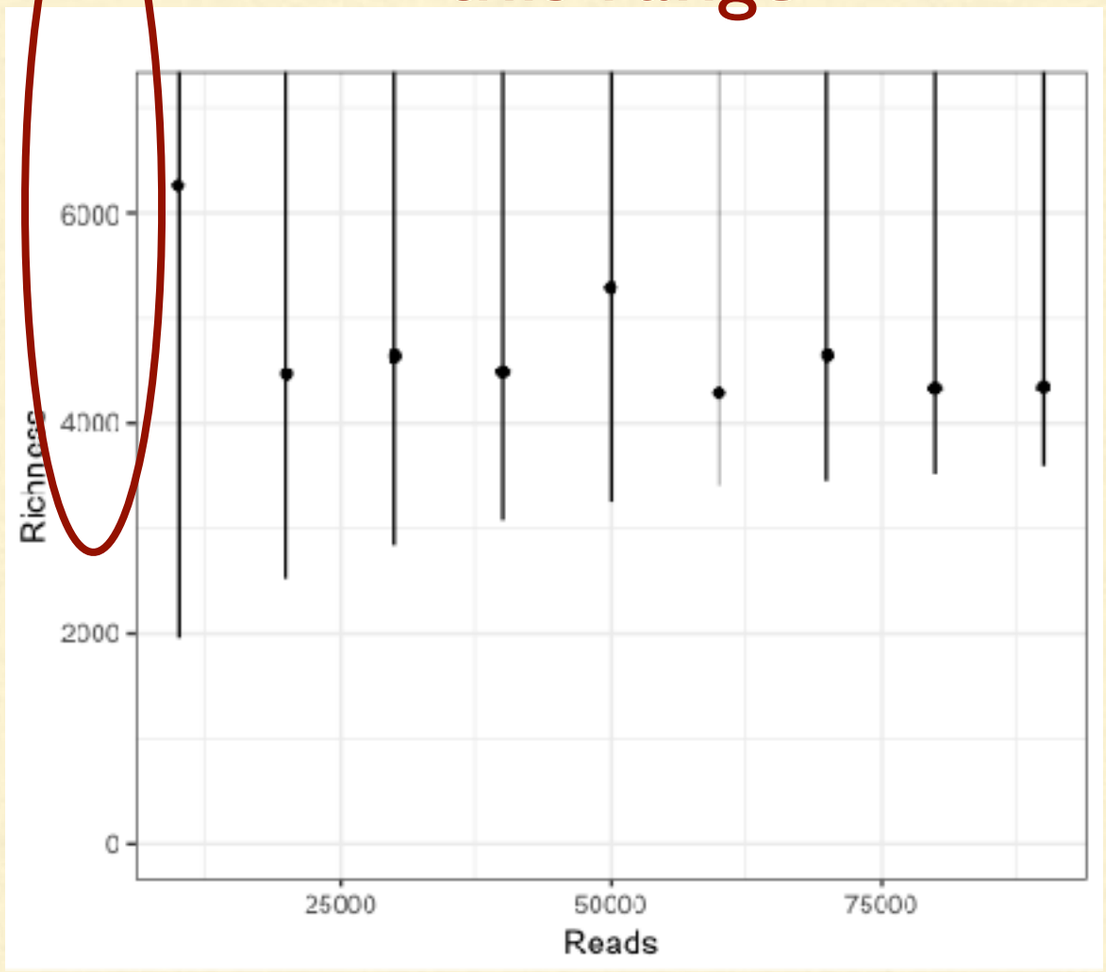
WHY SHOULD THESE APPROACHES BE DIFFERENT?

- Extrapolating beyond the range of the data is hard!
- Extrapolation gets harder as you move further from the data
  - Species richness estimation: extrapolating to 0
  - Extrapolating abundance curves: extrapolating to “infinity”
    - Also discarding structure

RESOLVING THE PARADOX



True richness probably in this range



RAREFYING

- True α diversity parameters are not sensitive to sample size
- Observed species richness is sensitive to sample size
- We should use better estimates of species richness
  - (same is true for other diversity indices)

SPECIES RICHNESS ESTIMATES

- The necessary data for richness is the frequency counts

- f_j = number of species observed j times

- f_1 = singletons,

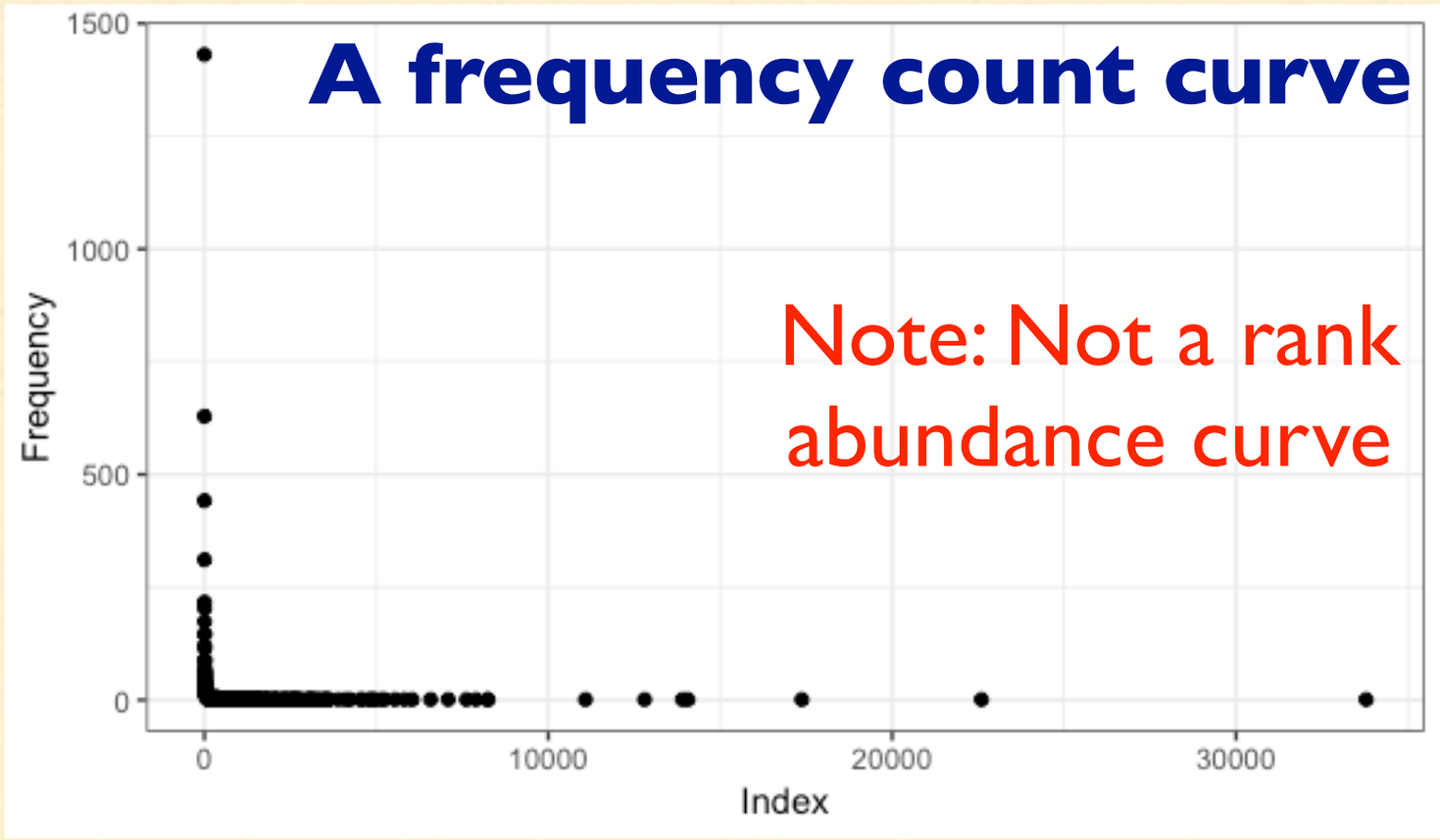
- f_2 = doubletons, …

- *e.g. 1431 strains observed once*

```
> library(phyloseq)
> library(magrittr)
> library(breakaway)
> data("GlobalPatterns")
> GlobalPatterns %>%
+   otu_table %>%
+   build_frequency_count_tables %>%
+   head(1)
$CL3
      Index Frequency
 [1,]     1      1431
 [2,]     2       629
 [3,]     3       442
 [4,]     4       311
 [5,]     5       217
 [6,]     6       203
 [7,]     7       174
 [8,]     8       146
 [9,]     9       146
[10,]    10       121
[11,]    11       114
[12,]    12        89
[13,]    13        74
```

SPECIES RICHNESS ESTIMATES

- Idea: extend the pattern in $f_1, f_2, f_3 \ldots$ to f_0

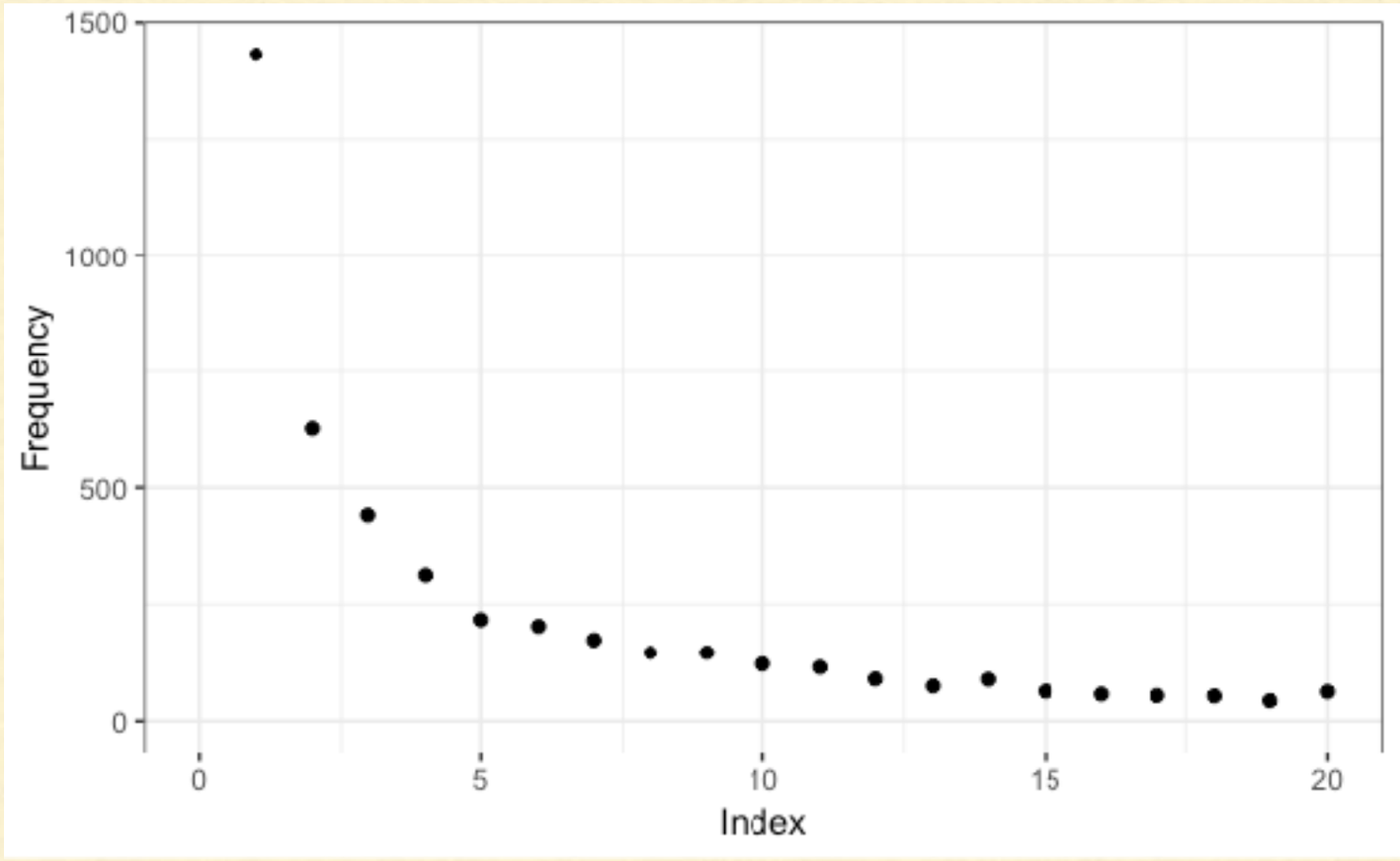


- Total species = unobserved species + observed species

- $C = f_0 + c$

SPECIES RICHNESS ESTIMATES

- Idea: extend the pattern in $f_1, f_2, f_3 \ldots$ to f_0



- Rare taxa are most informative for missing taxa

MIXED-POISSON MODELS

- Number of species observed k times is a draw from a zero-truncated mixed-Poisson distribution:

$$P(X = x) = \int \frac{\lambda^x}{x!} e^{-\lambda} dF(\lambda)$$

- F is the stochastic abundance distribution
- $\hat{f}_0 = n\hat{P}(X = 0)$

MIXED-POISSON MODELS: CHAO 1

- "Chao1 diversity index" is not an index -- it's an estimate of species richness, and it's based on the questionable assumption that

all species have the same abundance

- Large negative bias; very high variance
- Should not be used

MIXED-POISSON MODELS: CHAO-BUNGE

- Assumes stochastic abundance distribution is Gamma distributed
  - Negative binomial model
- Better than many, not as good as some
- Fast, sometimes unstable
- breakaway::chao_bunge()

MIXED-POISSON MODELS: CATCHALL

- "Fit lots of models and choose the best"
  - Robust, reasonable
- Executable: northeastern.edu/catchall/
  - Easy on Windows; painful on anything else
- R: github.com/adw96/CatchAll (under development!)

MIXED-POISSON MODELS: CATCHALL

	A	B	C	D	E	F	G	H
1	Total Number of Observed Species = 828	Model	Tau	Observed Sp	Estimated Tc	SE	Lower CB	Upper CB
2	Best Parm Model	TwoMixedExp	96	823	1194.2	30.9	1138.4	1260
3	Parm Model 2a	TwoMixedExp	97	824	1194.2	30.9	1138.4	1260
4	Parm Model 2b	TwoMixedExp	98	825	1194.2	30.9	1138.4	1260
5	Parm Model 2c	TwoMixedExp	101	828	1194.2	30.9	1138.4	1259.9
6	WLRM							
7	Parm Max Tau	TwoMixedExp	101	828	1194.2	30.9	1138.4	1259.9
8	WLRM Max Tau							
9	Best Discounted	SingleExp	96	145	153.7	3.9	150.6	169.7
10								

MIXED-POISSON MODELS: CATCHALL

	A	B	C	D	E	F	G	H
1	Total Number of Observed Species = 828	Model	Tau	Observed Sp	Estimated Tc	SE	Lower CB	Upper CB
2	Best Parm Model	TwoMixedExp	96	823	1194.2	30.9	1138.4	1260
3	Parm Model 2a	TwoMixedExp	97	824	1194.2	30.9	1138.4	1260
4	Parm Model 2b	TwoMixedExp	98	825	1194.2	30.9	1138.4	1260
5	Parm Model 2c	TwoMixedExp	101	828	1194.2	30.9	1138.4	1259.9
6	WLRM							
7	Parm Max Tau	TwoMixedExp	101	828	1194.2	30.9	1138.4	1259.9
8	WLRM Max Tau							
9	Best Discounted	SingleExp	96	145	153.7	3.9	150.6	169.7
10								

MIXED-POISSON MODELS: CATCHALL

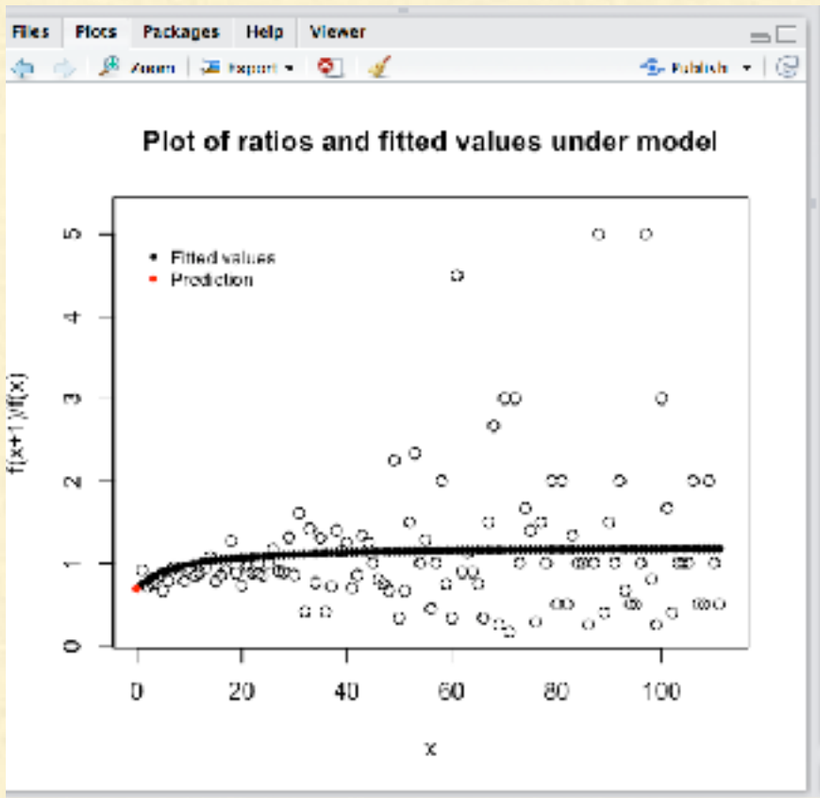
	A	B	C	D	E	F	G	H
1	Total Number of Observed Species = 828	Model	Tau	Observed Sp	Estimated Tc	SE	Lower CB	Upper CB
2	Best Parm Model	TwoMixedExp	96	823	1194.2	30.9	1138.4	1260
3	Parm Model 2a	TwoMixedExp	97	824	1194.2	30.9	1138.4	1260
4	Parm Model 2b	TwoMixedExp	98	825	1194.2	30.9	1138.4	1260
5	Parm Model 2c	TwoMixedExp	101	828	1194.2	30.9	1138.4	1259.9
6	WLRM							
7	Parm Max Tau	TwoMixedExp	101	828	1194.2	30.9	1138.4	1259.9
8	WLRM Max Tau							
9	Best Discounted	SingleExp	96	145	153.7	3.9	150.6	169.7
10								

NON MIXED-POISSON MODELS

- Non mixed-Poisson model: number of species observed k times is **not** a draw from a zero-truncated mixed-Poisson distribution
- Appealing because
  - Gives reasonable estimates for microbiome data
  - More reasonable assumptions
- Unappealing because
  - Harder to interpret model

NON MIXED-POISSON MODELS: KEMP

- R: github.com/adw96/breakaway
  - Function breakaway::kemp()




```
> par(mfrow=c(1,1))
> breakaway(frequency_counts)
################## breakaway ##################
        The best estimate of total diversity is 4341
         with std error 1297
        The model employed was model_1_1
        The function selected was
          f_{x+1}/f_{x} ~ (beta0+beta1*(x-xbar))/(1+alpha1*
(x-xbar))
        Coef estimates Coef std errors
beta0       1.15113208       0.14057373
beta1       0.01905826       0.01393881
alpha1      0.01567839       0.01417407
xbar                   56
>
>
>
>
>
>
>
>
```

Environment History

A COMPROMISE: BREAKAWAY

- R: github.com/adw96/breakaway
  - Function breakaway::breakaway()
  - A wrapper for many species richness estimates
  - Saves you from choosing!

Species richness is a hard problem! Confidence intervals may be huge!

Do not trust anyone who tells you otherwise

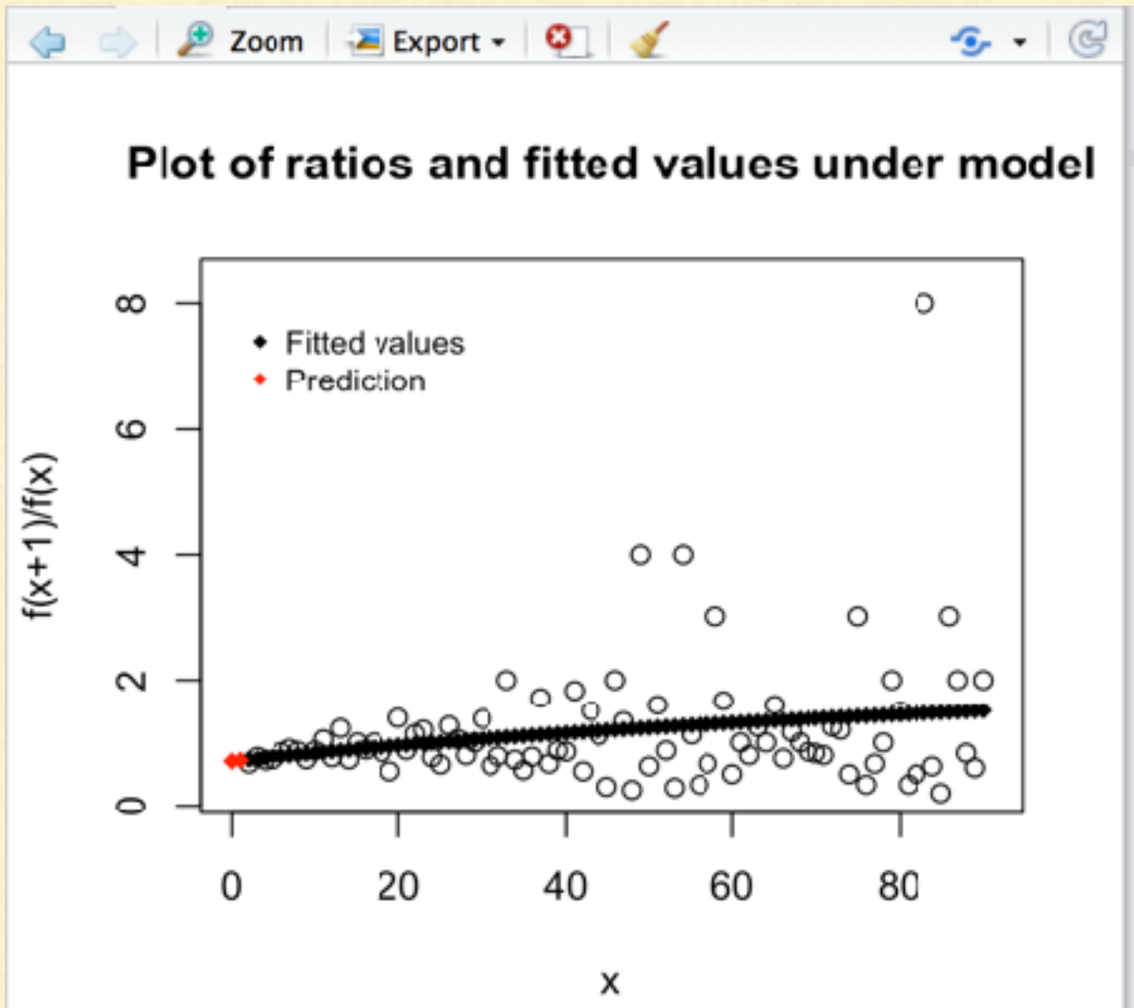
SPECIES RICHNESS ESTIMATION

- Good options
  - breakaway::breakaway()
  - breakaway::chao_bunge()
  - breakaway:: objective_bayes_*()
- Bad options
  - breakaway::chao1
  - vegan::... (more on this later)
  - scikitbio....

UNBELIEVABLE SINGLETONS

- All species richness estimates are based on *zero-truncated models*
  - Species exist in your population that were not observed in your sample
- All estimates sensitive to singleton count, f_1
  - Nearest to f_0
- If you don't trust your singleton count, fit *singleton-truncated models*

BREAKAWAY_NOF1




```
> breakaway_nof1(frequency_counts[-1, ], print = T)
Iterative reweighting didn't produce any outcomes after
the first iteration, so we use 1/x
################### breakaway ###################
        The best estimate of total diversity is 4913
         with std error 516
        The model employed was model_1_1
        The function selected was
          f_{x+1}/f_{x} ~ (beta0+beta1*(x-xbar))/(1+alp
ha1*(x-xbar))
       Coef estimates Coef std errors
beta0      1.196002878       0.10916519
beta1      0.014157602       0.01152216
alpha1     0.004786584       0.01205283
xbar                            45.5
>
```

Environment History

Large standard errors reflect more difficult problem

NEVER FORGET

rubbish in, rubbish out

ALPHA DIVERSITY ESTIMATION

- Species richness is special because it is the sum of discontinuous functions
- Most other alpha diversity indices are the sum of continuous functions, i.e., use proportions
- This makes them easier to estimate in general
  - Simpson: $\sum_{i=1}^{C} p_i^2$
  - Shannon: $-\sum_{i=1}^{C} p_i \ln p_i$
  - Evenness: $\frac{-\sum_{i=1}^{C} p_i \ln p_i}{\ln C}$

THE "CLASSICAL" APPROACH

- Substitute the observed abundances $\hat{p}_1, \ldots, \hat{p}_c$ for the unknown, true abundances $p_1, p_2, \ldots, p_c$ and pretend nothing happened
  - e.g. Shannon index $$-\sum_{i=1}^{c} \hat{p}_i \ln \hat{p}_i$$
  - e.g. Simpsons index $$\sum_{i=1}^{c} \hat{p}_i^2$$

DON'T DO THIS!

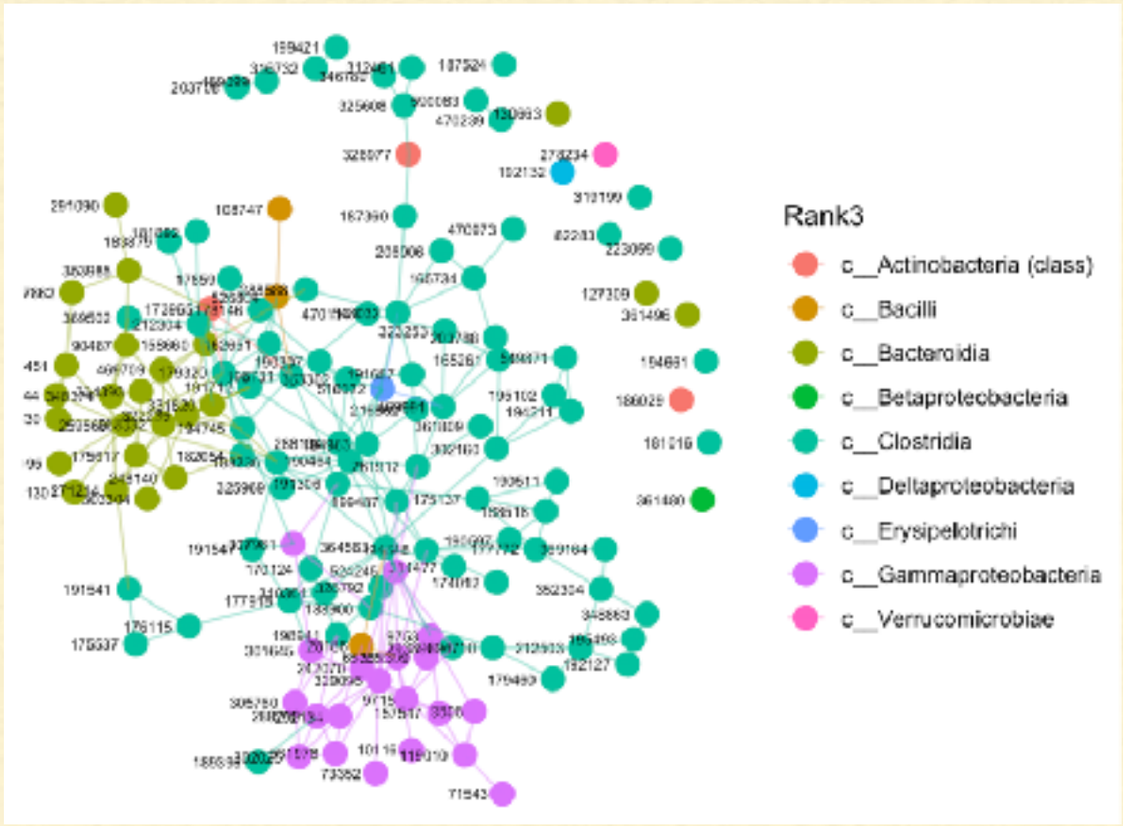
- Why not?
  - Your sample is not your population
  - Your estimate needs to be adjusted for missing taxa (fix bias)
  - You need error bars (state variance)

ERROR BARS FROM SUB-SAMPLING

- The most common way to put error bars is with subsampling
  - Resample individual microbes and recompute estimate based on the sample
- Subsampling understates the actual variability
  - Common misperception: nonparametric methods are model free
  - Subsampling in this way actually implies a very different structure (recall Tuesday's lecture)

DIVNET

- DivNet is a method that leverages microbial networks to improve diversity estimation
  - and more importantly, hypothesis testing for diversity



DIVNET

- How do you impose cooccurrence structure on compositional data?

NETWORK MODELS

- Let W_{iq} be the number of times taxon/gene q is observed in sample i
- Let M_i be the number of reads in sample i
- Let $X_i \in R^p$ be a vector of information about sample i

$$(W_{i1}, \ldots, W_{iq}) \sim \mathcal{N}_q(X_i\beta, \Sigma)$$

- $\beta \in R^{p \times q}$ is a matrix of the effect of each covariate on the abundance of each taxon

NETWORK MODELS

- Let W_{iq} be the number of times taxon/gene q is observed in sample i
- Let M_i be the number of reads in sample i
- Let $X_i \in R^p$ be a vector of information about sample i

terrible idea!

~~$(W_{i1}, \ldots, W_{iq}) \sim \mathcal{N}_q(X_i\beta, \Sigma)$~~

counts are discrete; model is continuous

- $\beta \in R^{p \times q}$ is a matrix of the effect of each covariate on the abundance of each taxon

NETWORK MODELS

$$\left(\frac{W_{i1}}{M_i}, \ldots, \frac{W_{iq}}{M_i}\right) \sim \mathcal{N}_q(X_i\beta, \Sigma)$$

NETWORK MODELS

terrible idea!

$$\left(\frac{W_{i1}}{M_i}, \ldots, \frac{W_{iq}}{M_i}\right) \sim \mathcal{N}_q(X_i\beta, \Sigma)$$

proportions have to be positive!

NETWORK MODELS

$$\left(\frac{W_{i1}}{W_{iD}}, \ldots, \frac{W_{iq}}{W_{iD}}\right) \sim \mathcal{N}_{q-1}(X_i\beta, \Sigma)$$

Pick "baseline" taxon D

NETWORK MODELS

terrible idea!

$$\left(\frac{W_{i1}}{W_{iD}}, \ldots, \frac{W_{iq}}{W_{iD}}\right) \sim \mathcal{N}_{q-1}(X_i\beta, \Sigma)$$

abundances are heavy tailed!

Pick "baseline" taxon D

NETWORK MODELS

$$\left(\log\left(\frac{W_{i1}}{W_{iD}}\right),\ldots,\log\left(\frac{W_{iq}}{W_{iD}}\right)\right) \sim \mathcal{N}_{q-1}(X_i\beta,\Sigma)$$

NETWORK MODELS

not bad...

$$\left(\log\left(\frac{W_{i1}}{W_{iD}}\right),\ldots,\log\left(\frac{W_{iq}}{W_{iD}}\right)\right) \sim \mathcal{N}_{q-1}(X_i\beta,\Sigma)$$

But shouldn't larger samples be more heavily emphasised?

DIVNET

true abundances in environment

abundances observed in sample

$$(W_{i1}, \ldots, W_{iq}) \sim Multinomial(M_i, (Z_{i1}, \ldots, Z_{iq}))$$

$$\left(\log \left(\frac{Z_{i1}}{Z_{iD}} \right), \ldots, \log \left(\frac{Z_{iq}}{Z_{iD}} \right) \right) \sim \mathcal{N}_{q-1}(X_i \beta, \Sigma)$$

network model at environment level

NETWORK MODELS

true abundances in environment

abundances observed in sample

$$(W_{i1}, \ldots, W_{iq}) \sim Multinomial(M_i, (Z_{i1}, \ldots, Z_{iq}))$$

$$\left(\log\left(\frac{Z_{i1}}{Z_{iD}}\right), \ldots, \log\left(\frac{Z_{iq}}{Z_{iD}}\right)\right) \sim \mathcal{N}_{q-1}(X_i\beta, \Sigma)$$

- proportions > 0 and sum to 1
- larger samples have more emphasis
- network permitted

NESTED MODELS

- As the largest element of Σ gets closer to 0, model converges to model without network
- "No network" is permitted, but so is "network"

$$(W_{i1}, \ldots, W_{iq}) \sim Multinomial(M_i, (Z_{i1}, \ldots, Z_{iq}))$$

$$\left(\log\left(\frac{Z_{i1}}{Z_{iD}}\right), \ldots, \log\left(\frac{Z_{iq}}{Z_{iD}}\right) \right) \sim \mathcal{N}_{q-1}(X_i\beta, \Sigma)$$

DIVERSITY ESTIMATION

- Fit model to obtain $(\hat{\beta}, \hat{\Sigma})$
- Calculate fitted proportions for experimental condition i

$$\hat{Z}_{ik} \propto e^{X_i \beta_k}, \qquad k \neq D$$

$$\hat{Z}_{ik} \propto 1, \qquad k = D$$

- Estimate diversity index by, e.g.,

$$\hat{\alpha}_{i,Shannon} = -\sum_{k=1}^{q} \hat{Z}_{ik} \log \hat{Z}_{ik} \qquad \hat{\beta}_{ij,Bray-Curtis} = 1 - \sum_{k=1}^{q} \min(\hat{Z}_{ik}, \hat{Z}_{jk})$$

- Estimate variance of estimates using $\hat{\Sigma}$

Willis & Martin, 2018, biorxiv

DIVNET

- This idea works for estimating any diversity index (**α or β**) that is a function of relative abundances

- It can also be used to estimate any diversity index that is a function of the tree

github.com/adw96/DivNet

Coming soon...

Willis & Martin, 2018, biorxiv

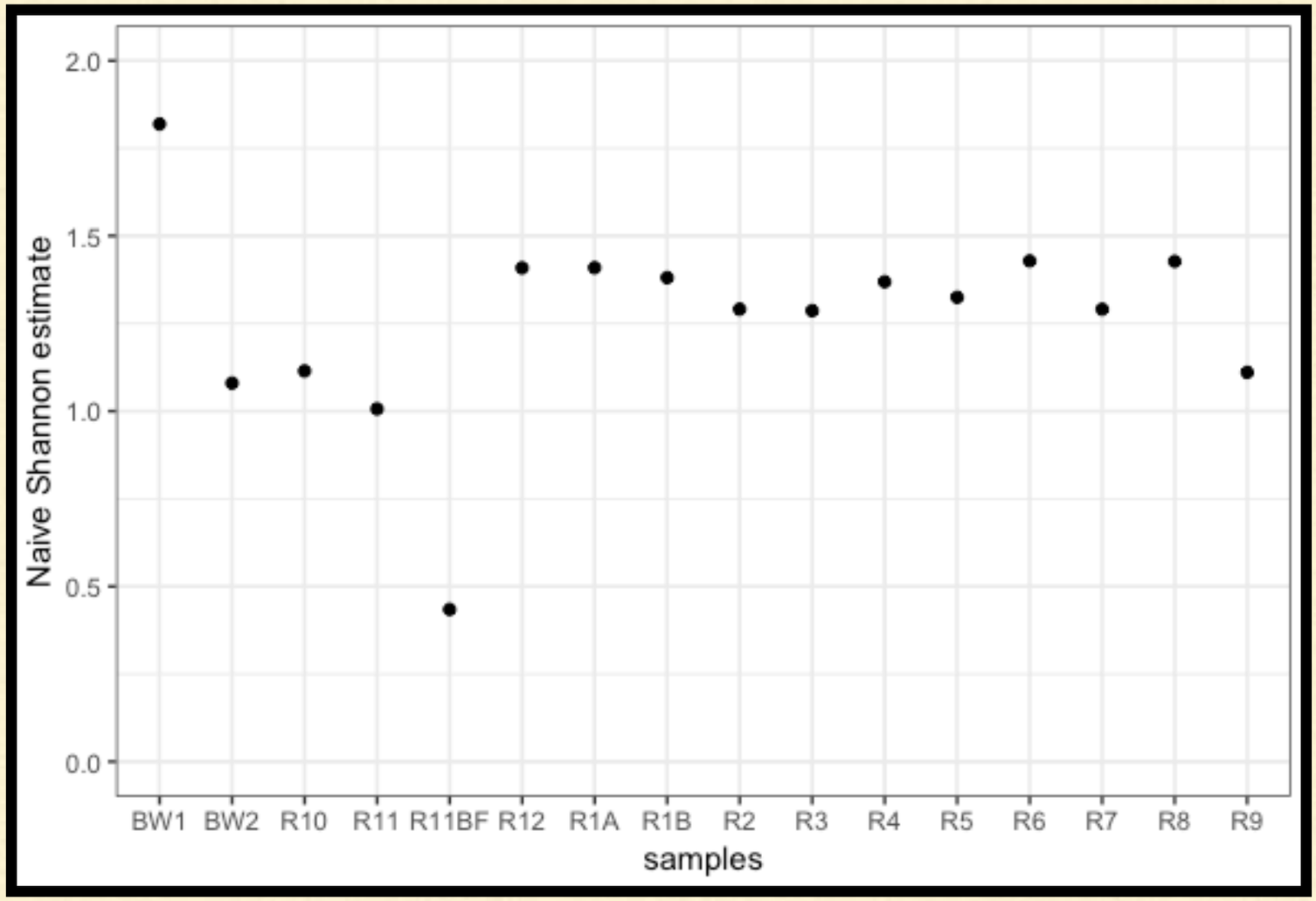
STRUCTURE: ABUNDANCES

- These datasets have the same mean structure, but different covariance/network structures

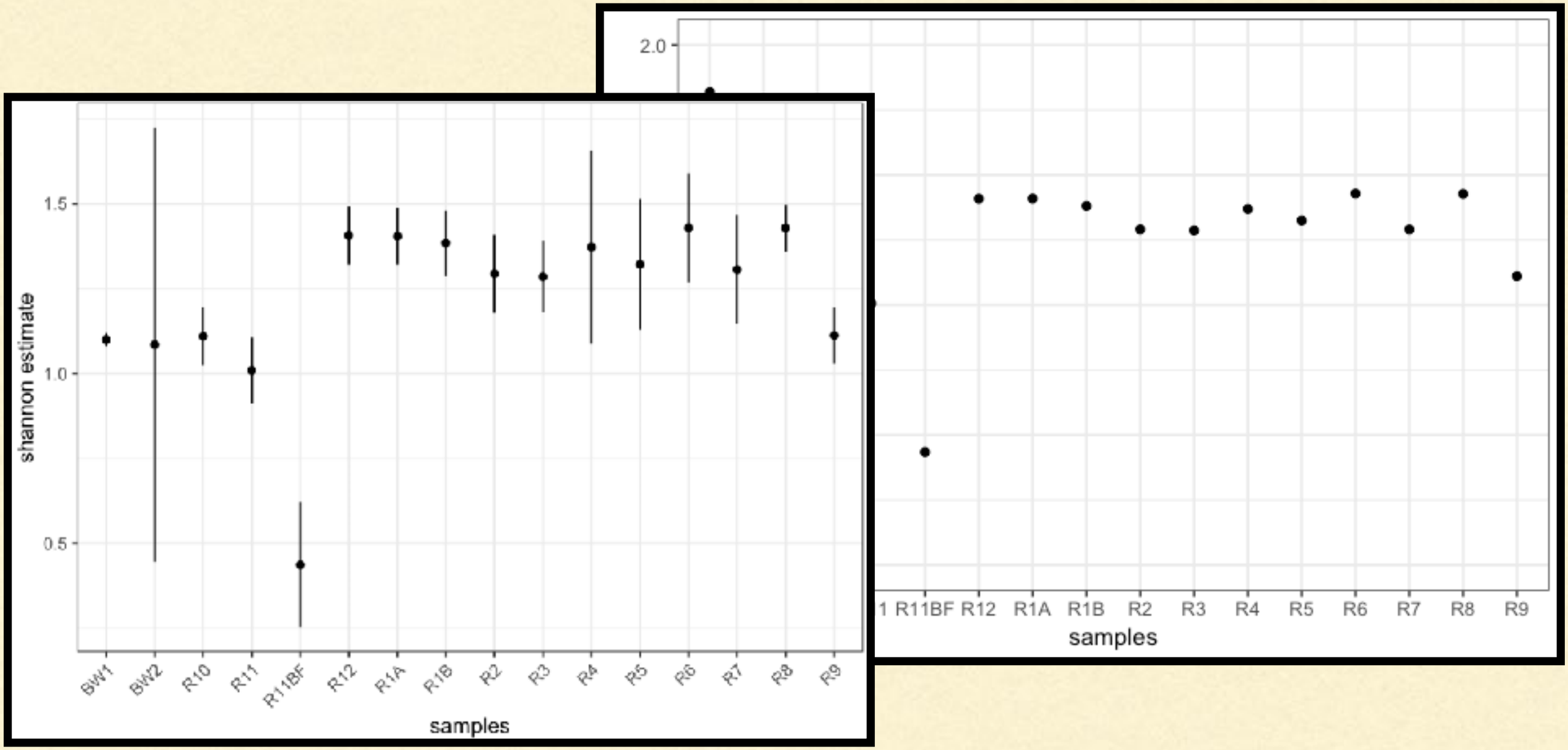
	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5	Taxon 6	Taxon 7
Sample 1	21	26	31	32	26	46	18
Sample 2	21	25	31	26	24	42	31
Sample 3	24	40	20	31	27	32	26
Sample 4	27	33	26	28	24	40	22

	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5	Taxon 6	Taxon 7
Sample 1	3	12	12	75	30	48	20
Sample 2	0	0	189	0	1	0	10
Sample 3	142	16	0	24	6	12	0
Sample 4	4	2	11	110	9	48	16

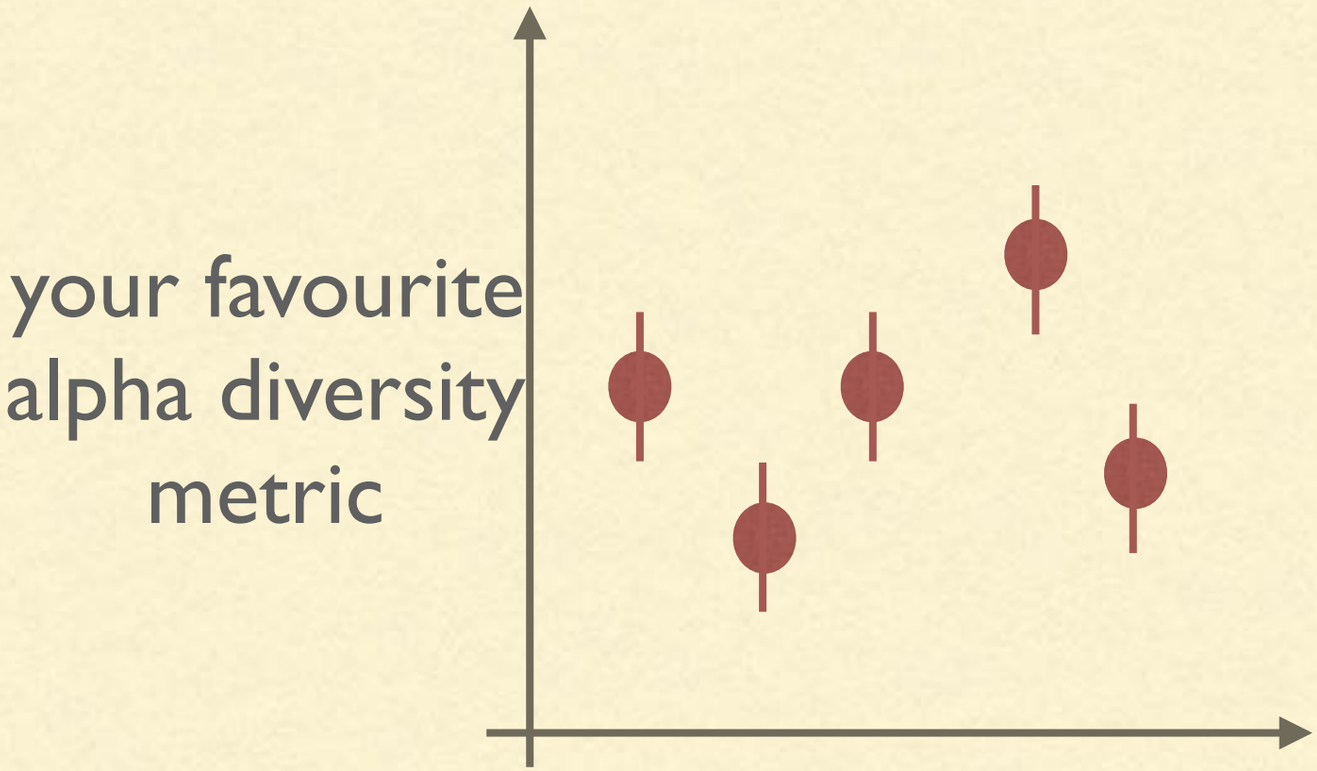
STRUCTURE: ESTIMATES



STRUCTURE: ESTIMATES



ALPHA DIVERSITY



1. What estimate do we use?

Where do we draw the dots?

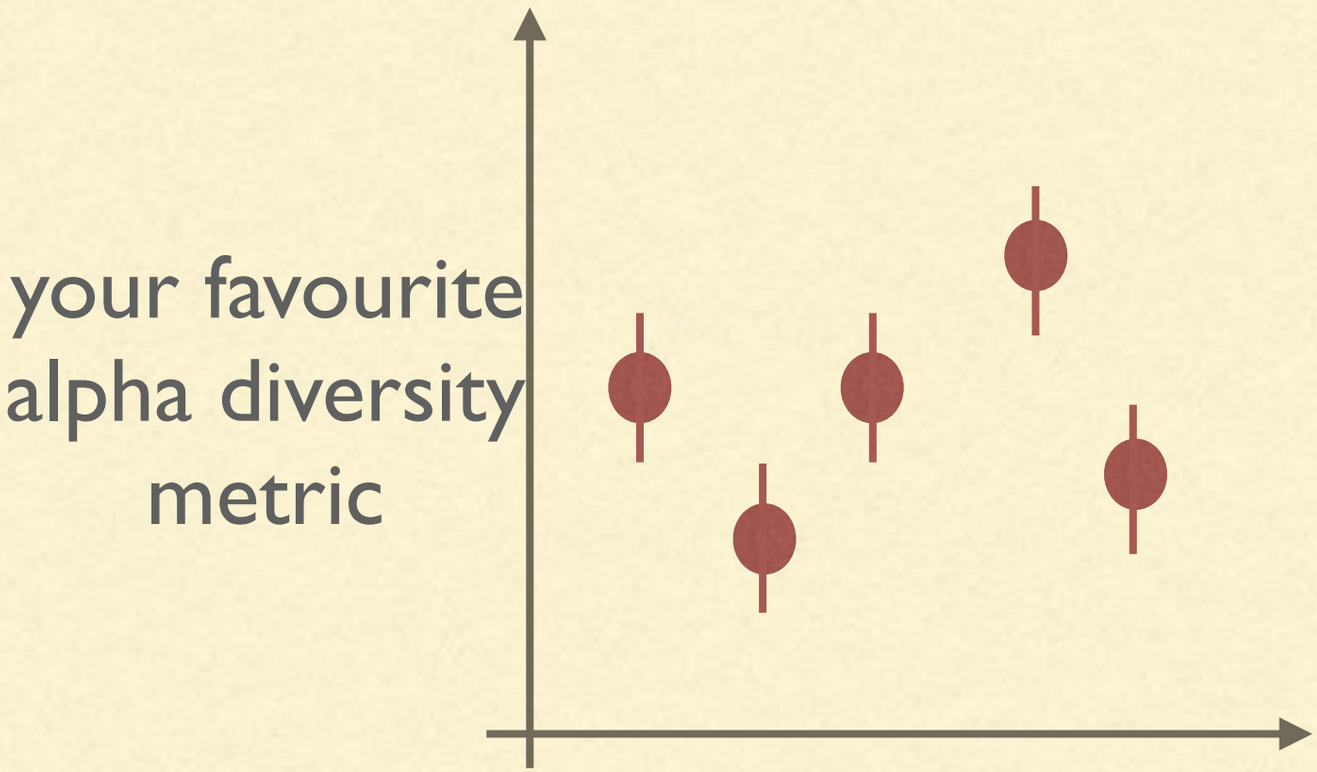
2. What is the standard error?

How long are the lines?

3. How do I model a change?

What line can I draw?

ALPHA DIVERSITY



~~1. What estimate do we use?~~

~~*Where do we draw the dots?*~~

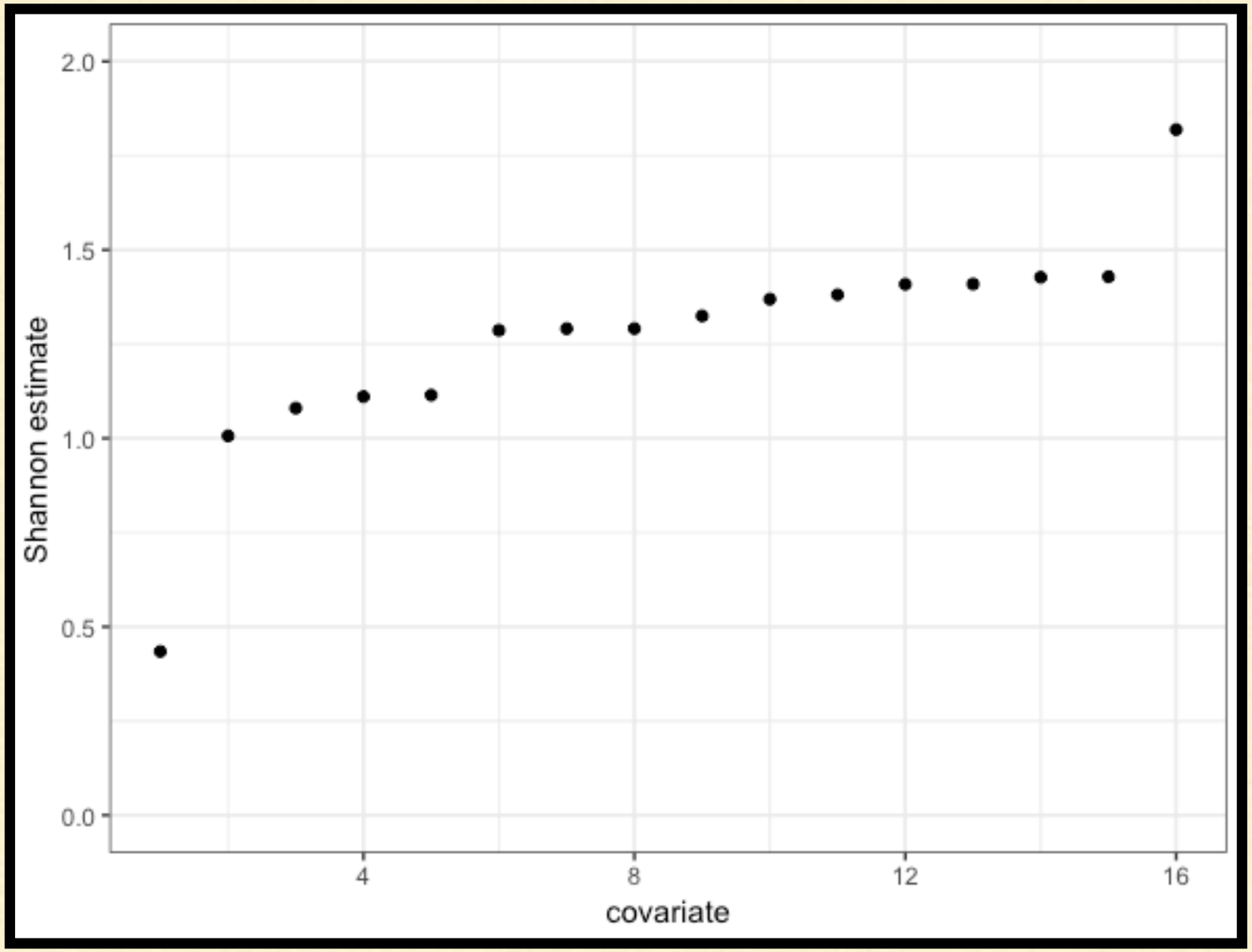
~~2. What is the standard error?~~

~~*How long are the lines?*~~

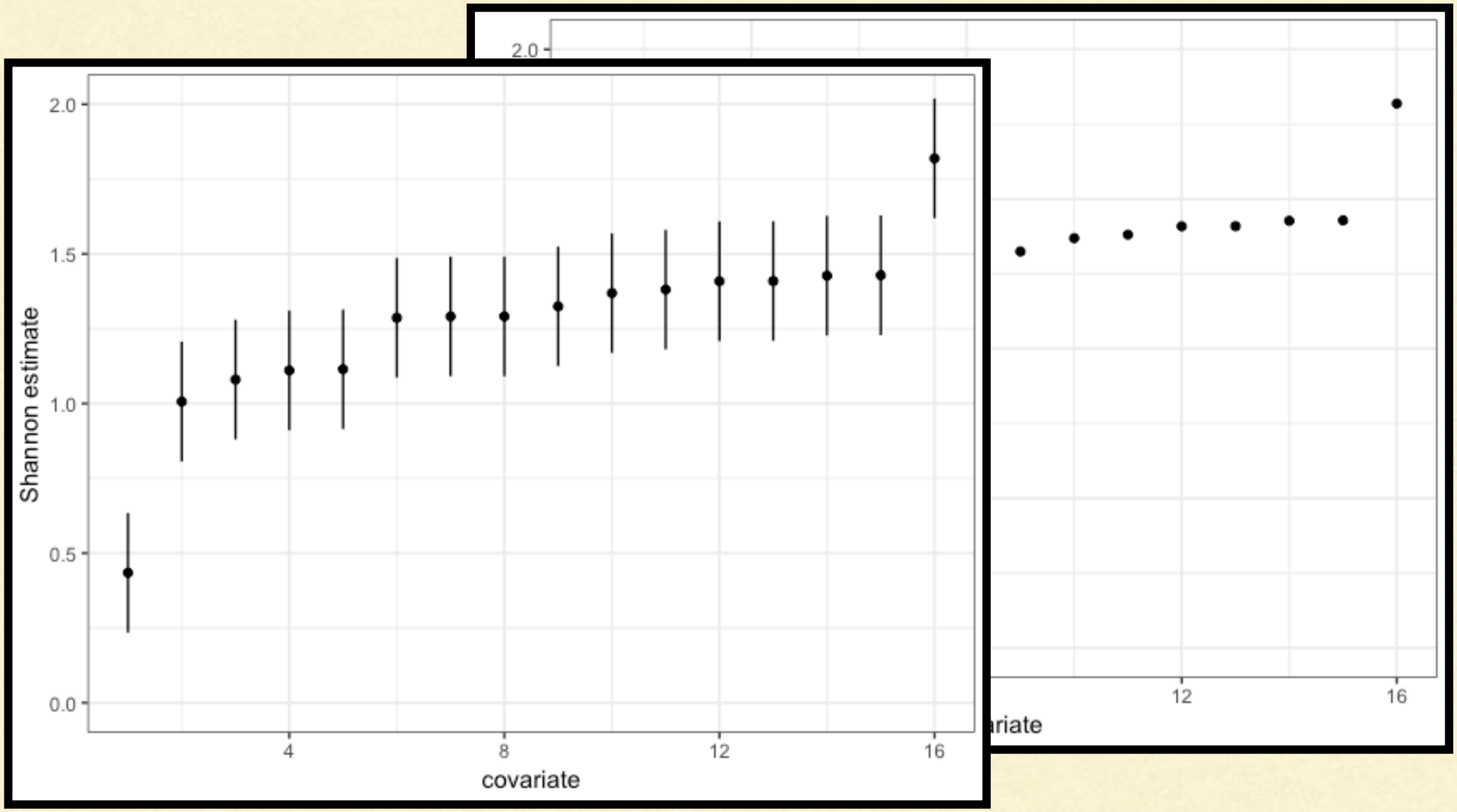
3. How do I model a change?

What line can I draw?

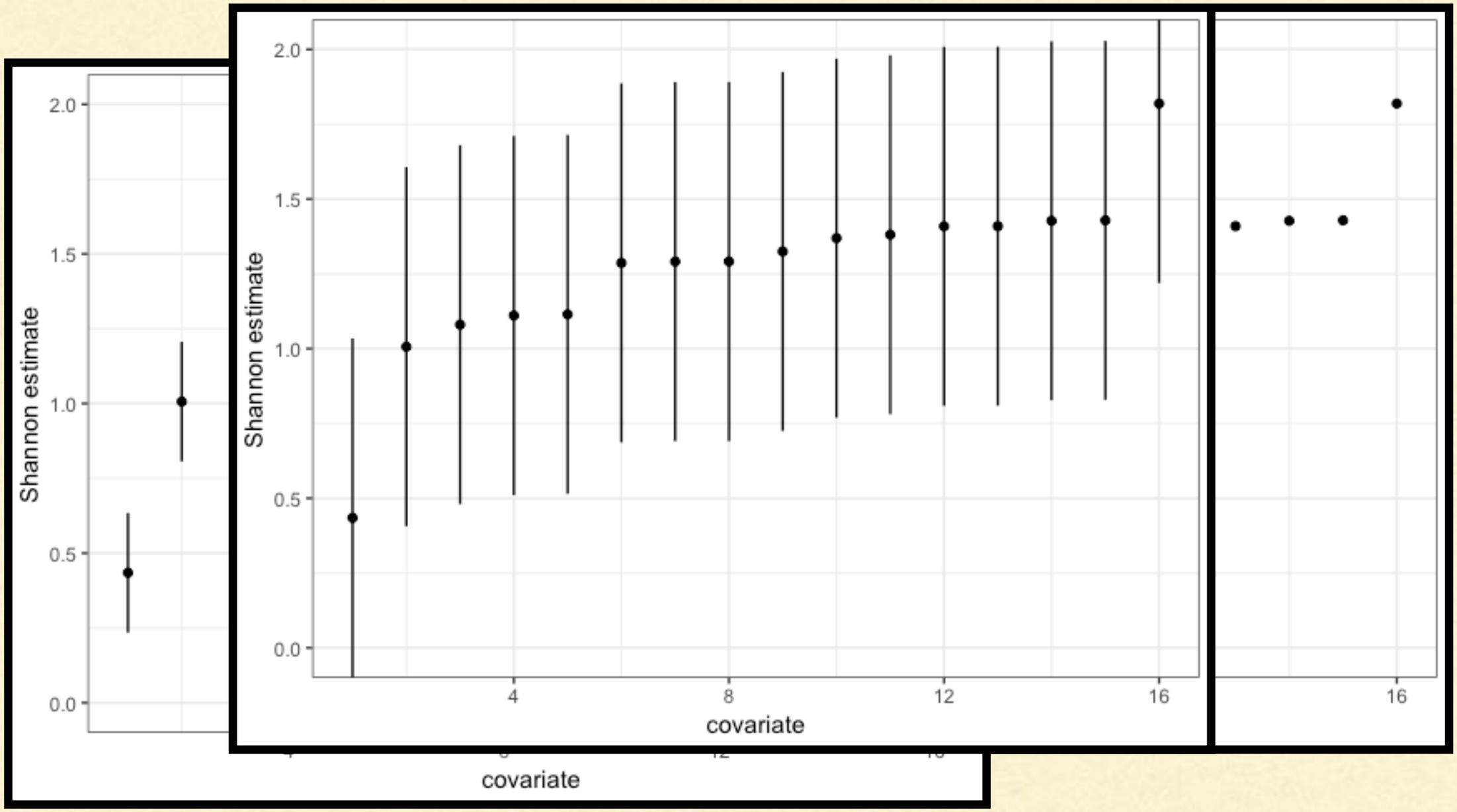
HYPOTHESIS TESTING FOR DIVERSITY



HYPOTHESIS TESTING FOR DIVERSITY



HYPOTHESIS TESTING FOR DIVERSITY

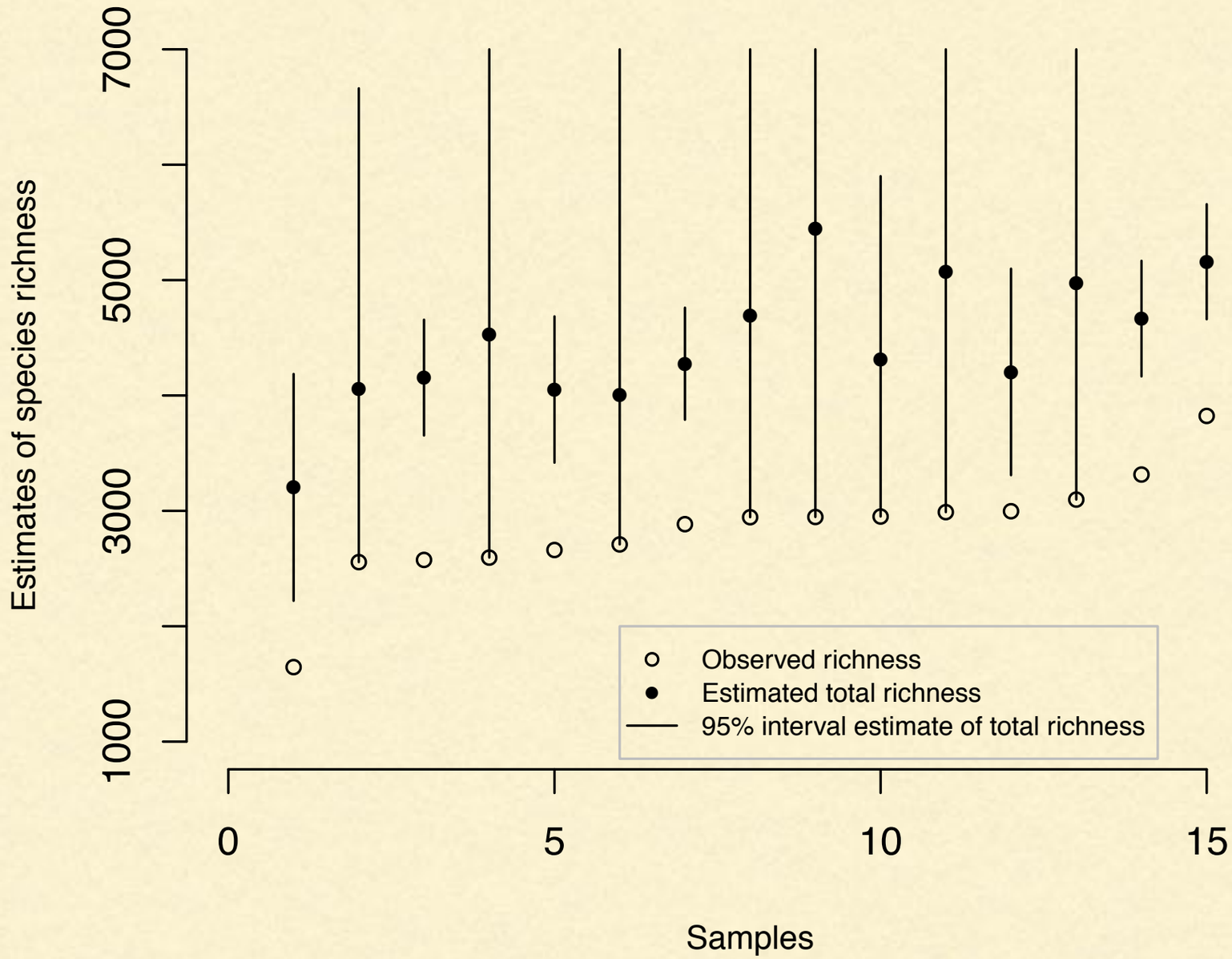


HYPOTHESIS TESTING FOR DIVERSITY

- Critical issue: need to adjust for different resolution when testing
  - Resolution = error bars
- Take your reasonable error bars from your reasonable estimate of alpha diversity and use them to do hypothesis testing
  - breakaway::betta()

Willis, Bunge & Whitman (2017), JRSSC

TESTING DIVERSITY



Willis, Bunge & Whitman (2017), JRSSC

TESTING DIVERSITY

- Function betta() in R package breakaway
- Regression models for α diversity with two error sources
  - Heterogeneity of samples
  - Uncertainty in estimating α diversity parameter

```
betta(ests, ses, my_X)$table
```

```
##           Estimates Standard Errors p-values
## No Amdmt 4680.4121         63.51318    0.000
## Biochar     0.0000        100.79685    1.000
## Biomass  -497.2997        158.99322    0.002
```

"We reject the null hypothesis that fresh biomass additions have no effect on species richness (p=0.002) and conclude an average loss of 497 taxa compared to non-fertilized controls."

BETA DIVERSITY

- β-diversity is typically used for exploratory analyses
  - Discussed more in "Exploratory multivariate visualization"
  - Nice approach: boyuren158/DirFactor
- There exists very little statistical literature on estimating β-diversity
  - Bray-Curtis & Euclidean distances using adw96/DivNet
  - UniFrac coming soon to adw96/PhyloDivNet (stay tuned)
  - Avoid Jaccard -- very difficult to estimate

ADDENDUM

- The statistical problems discussed in this talk assume that your abundance table is correct!
- Are you worried about this? Us too!

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH · UNIVERSITY *of* WASHINGTON

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Recruiting: Postdoctoral research associate
Project: Addressing misclassification in the microbiome: A data-scientific approach to propagating uncertainty in microbial community composition
Institutions: University of Wisconsin–Madison (Dr. Thea Whitman and Dr. Karl Broman) and the University of Washington (Dr. Amy Willis)

We are recruiting a postdoctoral researcher to work on a multi-institutional project developing an automated pipeline for analyzing microbiome data that adjusts for different levels of data quality in the microbial genome sequencing process. Bioinformatics choices have an enormous impact on microbiome data analysis, but the preprocessing and statistical analysis steps are almost always considered as independent steps. The project will develop a method and software to integrate information about data quality (*e.g.*, in taxonomic assignment) into data analysis (including, but not limited to, hypothesis testing). A core goal of the project is improving the reproducibility of microbiome studies.

FACTS

- Diversity is a parameter, which you can estimate using your data

- Reasonable estimates and accurate standard errors facilitate scientific reproducibility

MY OPINIONS

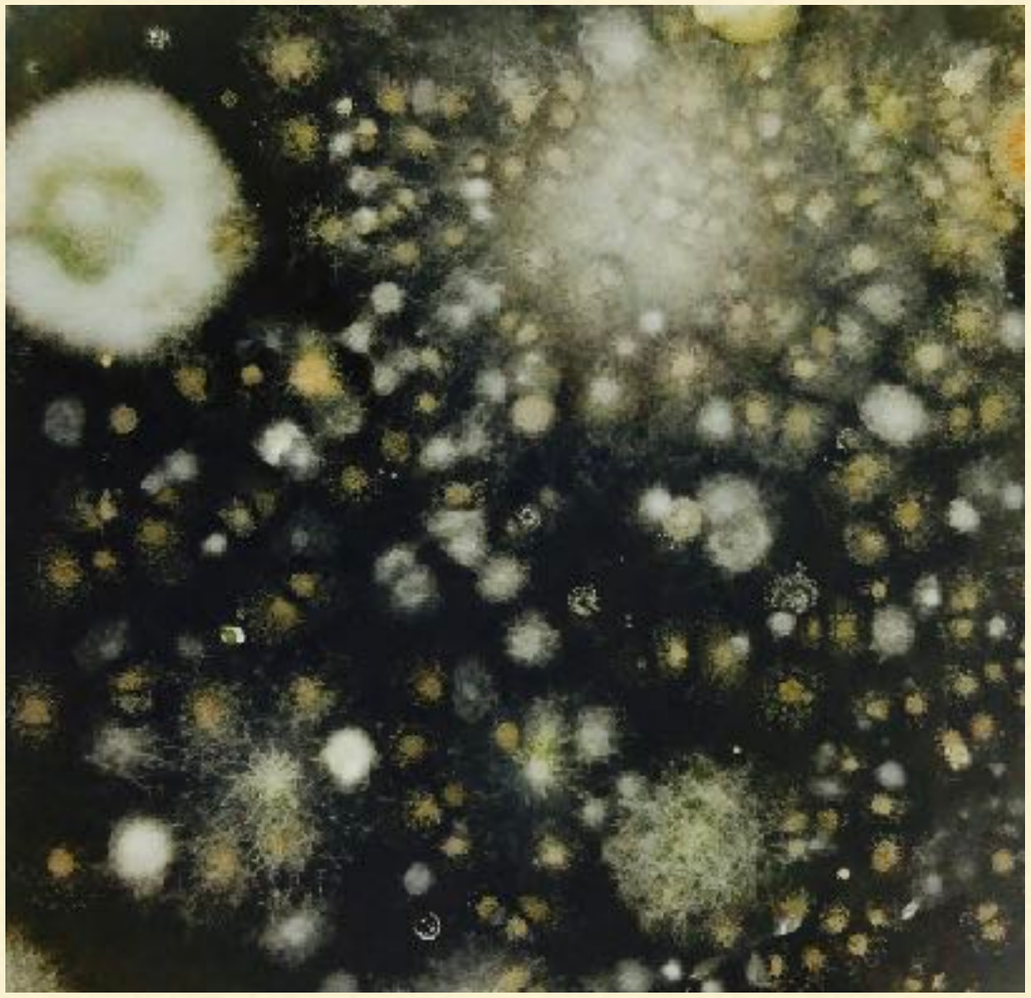
- Every individual diversity estimate should be accompanied by a standard error
- Substituting observed proportions for true proportions in diversity estimates is a bad idea
- No standard errors make your reader think the standard error is zero
- OTU-level diversity is not interesting, since it has no biological meaning (strain-level diversity is meaningful)
- Richness estimation is possible for some microbiomes and not for others

SUMMARY

- How to estimate/plot/hypothesis testing for
  - richness
  - α-diversity
  - β-diversity (briefly -- to be covered more later)
- Questions?
- Lab! 30 minutes

DIVERSITY LAB

- **github.com/adw96/stamps2018/**
- Click on **estimation**
- Click on **diversity-lab.R**
- Click on **raw**
- Copy and paste into an R script and start reading and working through



DIVERSITY

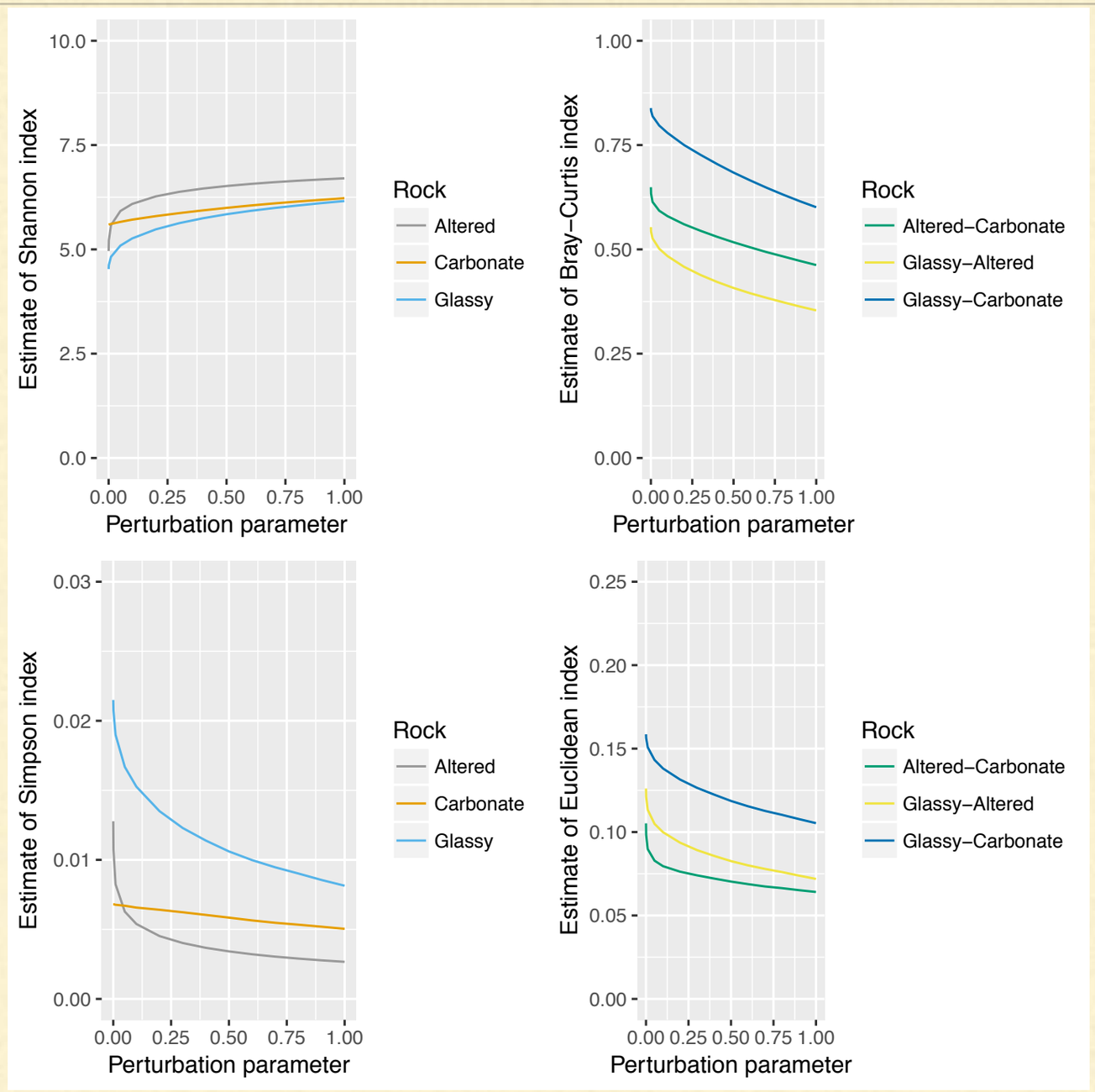
Research Group: Statistical Diversity Lab

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis adwillis@uw.edu

Photo credit: T.D. Berry, Whitman lab, UW Madison

PERTURBATION PARAMETER



VARYING DENOMINATOR TAXON