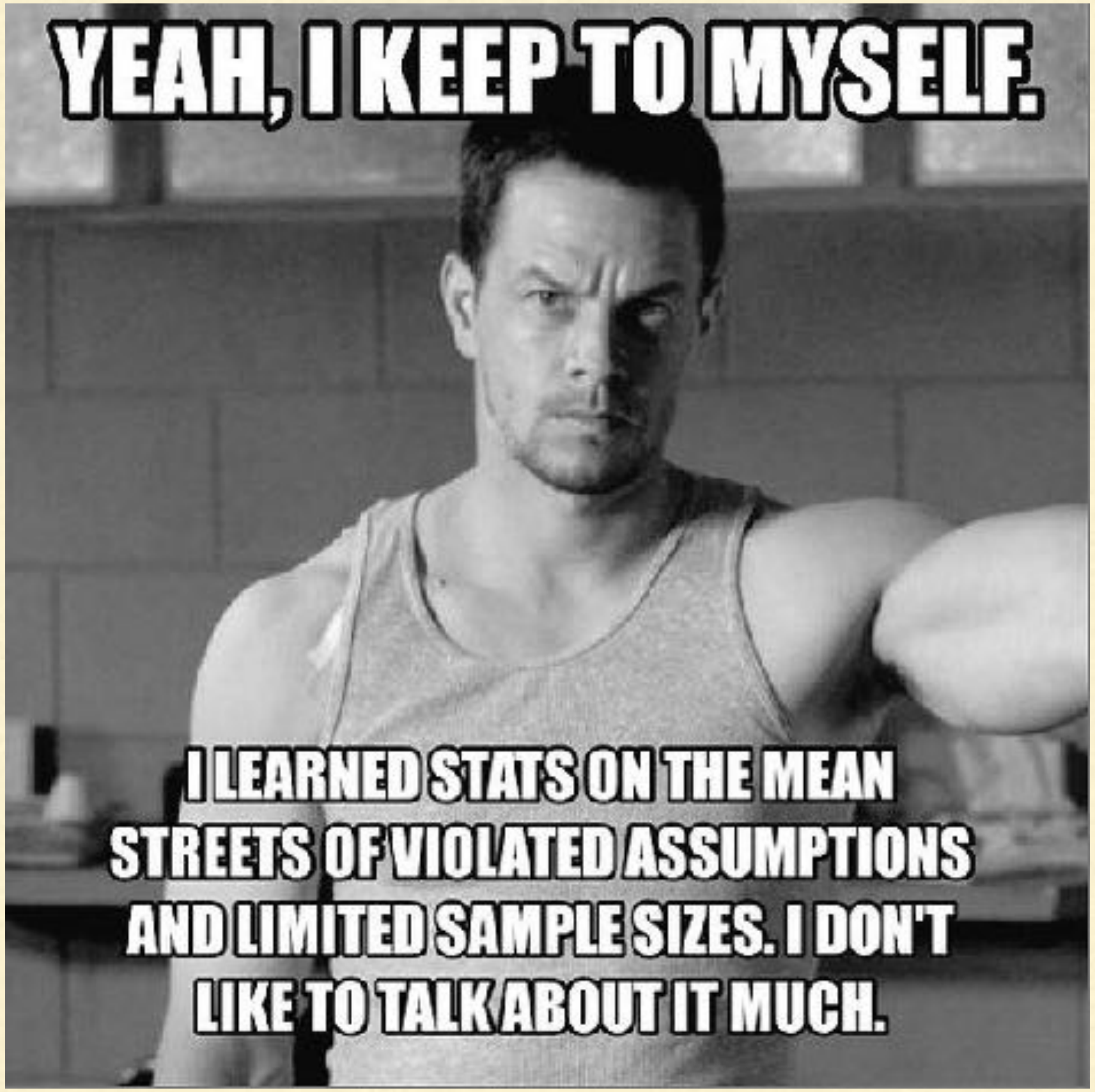

YEAH, I KEEP TO MYSELF.
I LEARNED STATS ON THE MEAN STREETS OF VIOLATED ASSUMPTIONS AND LIMITED SAMPLE SIZES. I DON'T LIKE TO TALK ABOUT IT MUCH.

STATISTICAL THINKING

Research Group: Statistical Diversity Lab

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

@AmyDWillis adwillis@uw.edu

STATISTICS

- Who wants to know more statistics? Why?
- Two different types of statistics
  - Inferential statistics
  - Exploratory statistics

STATISTICS

exploratory statistics	inferential statistics

hypothesis testing

estimation

plotting

ordination

modelling

PCA

p-values

FDR control

STATISTICS

exploratory statistics	inferential statistics
plotting ordination PCA	hypothesis testing p-values FDR control estimation modelling

INFERENTIAL STATISTICS

- Inference of parameters
- Prediction
- Estimation
- Uncertainty
- Reproducibility

EXPLORATORY STATISTICS

- What does your data say?
- How do we show what it says?
- How to visualise it?
- *Descriptive statistics*

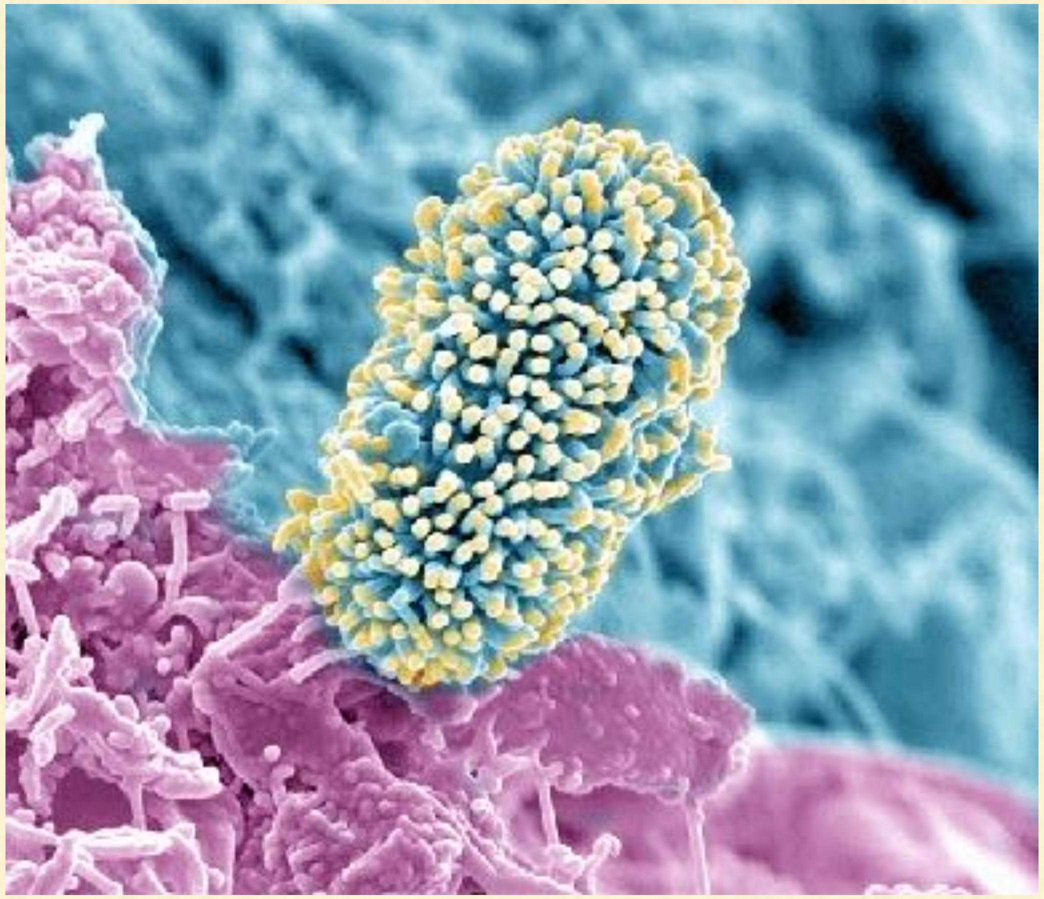
POPULATION

- Stat101
    - "The population of women with breast cancer"
    - "The population of American citizens with graduate degrees"
    - "The voting population of Massachusetts"

What is the population in microbial ecology?

MICROBIAL POPULATIONS

- The (A) microbiome is the (a) collection of microbes, and their genes and metabolites



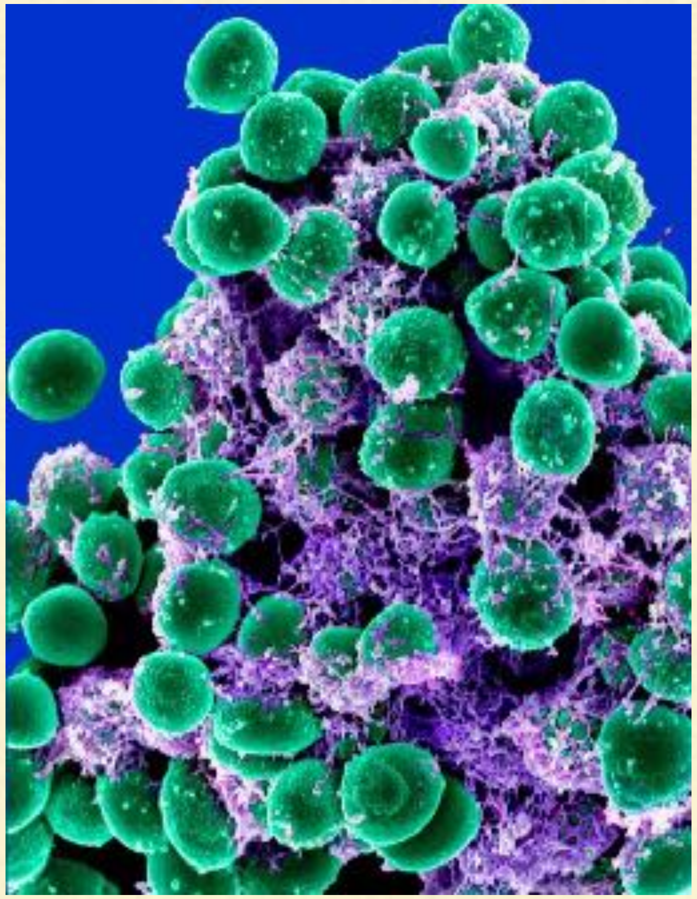


Photo credits: Pacific Northwest National Laboratory; NIAID

MICROBIAL POPULATIONS

- But which microbiome?
  - Are you interested in the gut microbiome of all folks with IBD?
    - ...white men 25-45y.o. with a clinical diagnosis?
    - ...who also live in the city that your study was conducted in?
    - ...Or are you only interested in their poop?

MICROBIAL POPULATIONS

- Are you interested in microbes living in the ocean?
  - Which ocean?
  - At what depth?
  - What time of year and day?
    - Or only those you can detect with your primers?

The population that you want to study may not be the population that you get to study

MICROBIAL POPULATIONS

- The 4 W’s: **Who/What? Where? When? Why?**
    - **Who? What?** …the poop of white men 25-45 y.o. with a clinical diagnosis?
    - **Where?** … who also live in the city that your study was conducted in?
    - **When?** … between January 2018-March 2018?
  - Such observations can help us answer **why** certain patterns exist
    - and why certain patterns don't....

EXPERIMENTAL DESIGN

The population that you want to study may not be the population that you get to study

- Before undertaking a microbiome study, think carefully about
  - the question you want to answer,
  - the data you have access to, and
  - the questions you can answer with the data that you have access to

MICROBIAL POPULATIONS

- Group exercise: (2 minutes)
    - Come up with a microbiome-related question that ***you want to answer*** considering the following questions:
        - **Who/What? Where? When? Why?**
    - Come up with a microbiome-related question that ***you could study***
        - *How do (sequencing) technology and (bioinformatics) tools influence what populations you can study?*

POPULATIONS VERSUS SAMPLES

- The difference between a *population* and *a sample from it* is fundamental in statistics
- Deductive logic: the evidence must imply the conclusion
- Inductive logic: the conclusion could be implied by the evidence

POPULATIONS VERSUS SAMPLES

- Inferential statistics: using information about the sample to infer something about the population
  - Use the observed data to estimate the parameters
  - Inductive not deductive logic

"SOMETHING ABOUT THE POPULATION"

- Statisticians have a formal concept of this
- "Parameter": a numerical characteristic of a probability model
  - Can give any examples of a parameter of interest in a microbiome experiment?

PARAMETERS

- The genus-level relative abundance of Streptococcus in your saliva right now
- The proportion of your S. aureus that are methicillin-resistant
- The fraction of #STAMPS18 attendees carrying MRSA in any abundance
- The phylum-level diversity of microbes on your hands

AN IMPORTANT DISTINCTION

- The genus-level relative abundance of Streptococcus in your saliva right now is not the same as the relative abundance of 16S copies from Streptococcus obtained from a sample
  - Does adjusting for copy number fix this?
    - Why not?

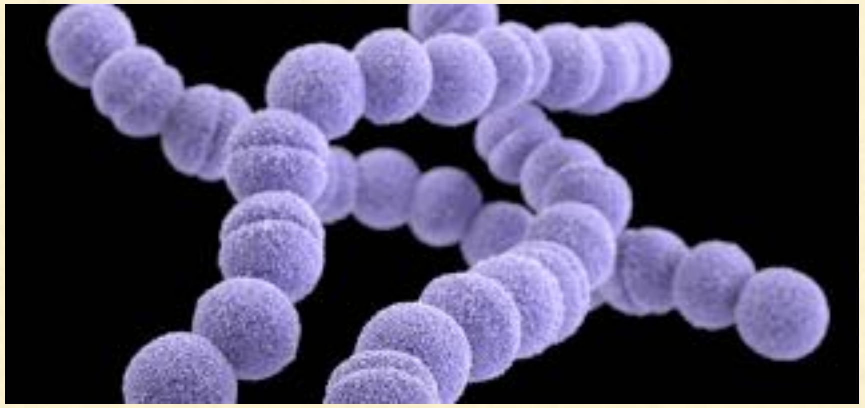
"INFORMATION ABOUT THE SAMPLE"

- Statisticians have a formal concept of this too
- "Estimates": some function of your data
    - Can give any examples of an estimate in a microbiome experiment?

ESTIMATES ESTIMATE PARAMETERS

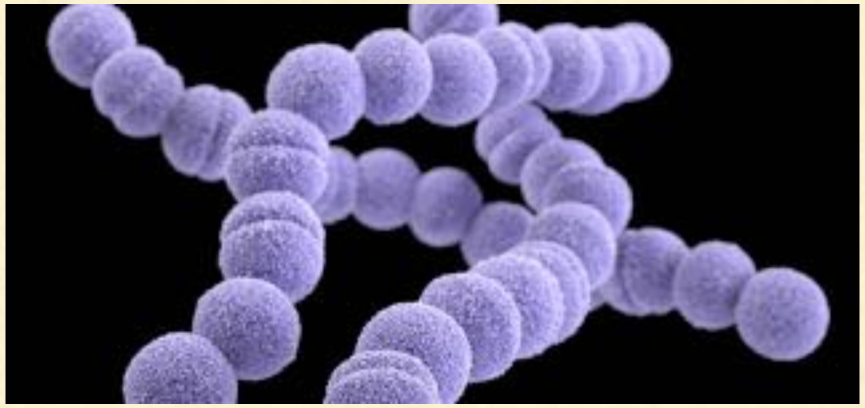
- Estimates (n, pl) estimate (v) parameters (n)
- e.g. What is an example of a parameter in a microbiome study, and what is an example of an estimate of it?

EXAMPLE



- **Motivation**: Estimate the genus-level relative abundance of 16S copies from Streptococcus in your saliva
- Relative abundance is commonly estimated by the observed relative abundance of 16S copies from Streptococcus
- Is that the only estimate? Why does it seem like a good one?

EXAMPLE



- **Motivation**: Estimate the genus-level relative abundance of 16S copies from Streptococcus in *a group of people*
  - What if we have 10 people in our study?
  - What does relative abundance of Streptococcus mean now?

RELATIVE ABUNDANCE

- Suppose…
  - **n** = samples, indexed by **i** = 1, …, n
  - $\mathbf{p_i}$ = the relative abundance in each subject
  - $\mathbf{W_i}$ = # of observed sequenced copies from Strep
  - $\mathbf{M_i}$ = total # of sequenced copies
- Most common estimate of p_i is W_i/M_i

RELATIVE ABUNDANCE

- Why?
  - (Seems reasonable)
  - Under a model where each observed copy of the 16S gene is from Strep with probability p_i, and all copies are independent, this estimate is
    - consistent, normally distributed, efficient, unbiased, minimum variance out of all unbiased estimates…

PARAMETERS

- Two key concepts for evaluating estimates of parameters
  - bias: how far?
  - variance: how stable?
- Suppose we have a parameter θ and an estimate $\hat{\theta}$

ESTIMATES: NOTATION

- The parameter *Amy*:

ESTIMATES: NOTATION

- An estimate of the parameter *Amy*:

BIAS

- If you care about a parameter θ, then the bias is the expected difference between the parameter any estimate

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

where

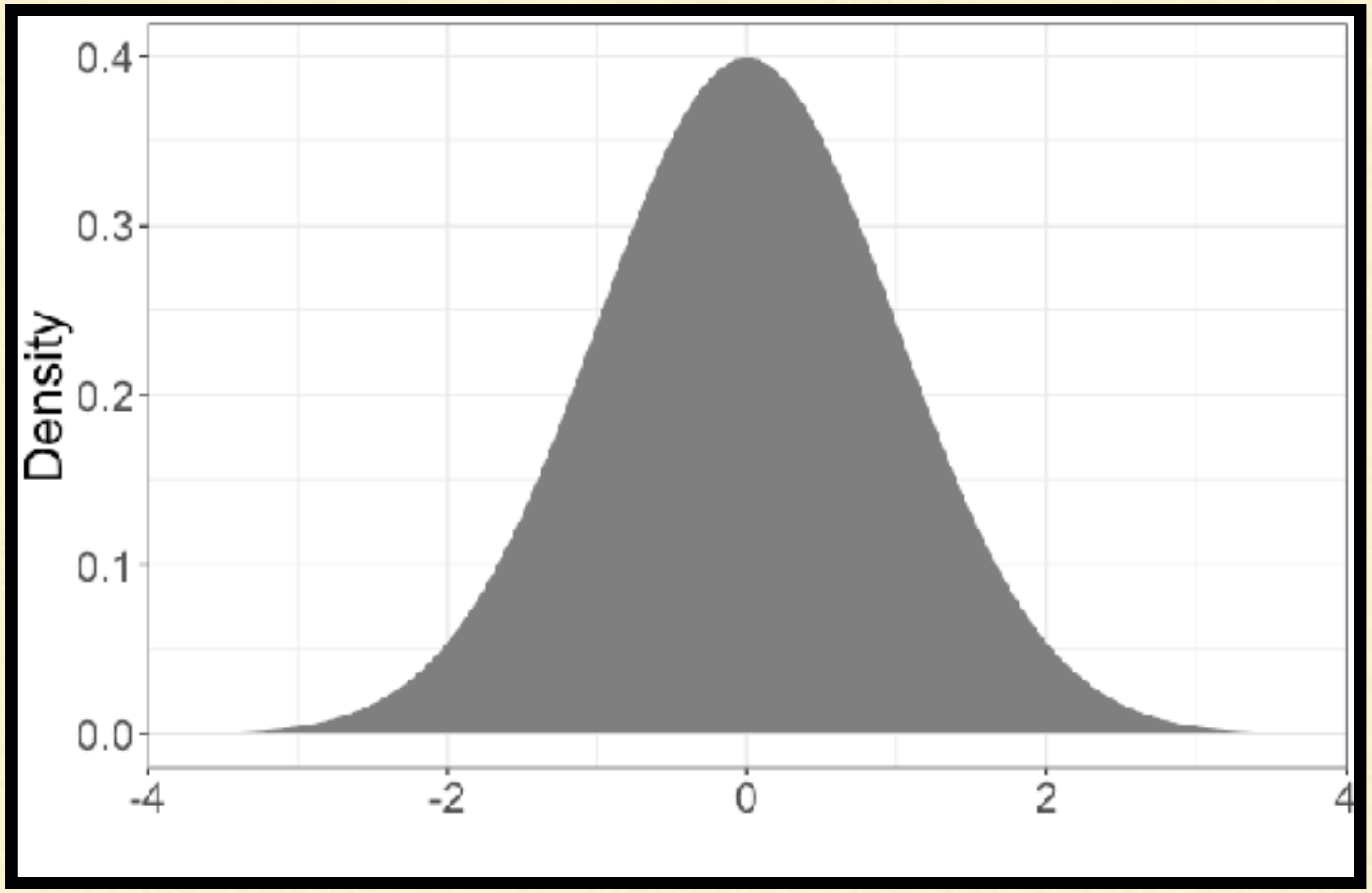
$$\mathbb{E}\hat{\theta} = \text{“expected value” of } \hat{\theta}$$

EXPECTED VALUE

- Expected value comes from the concept of a “distribution”
- It is the “middle” of the distribution

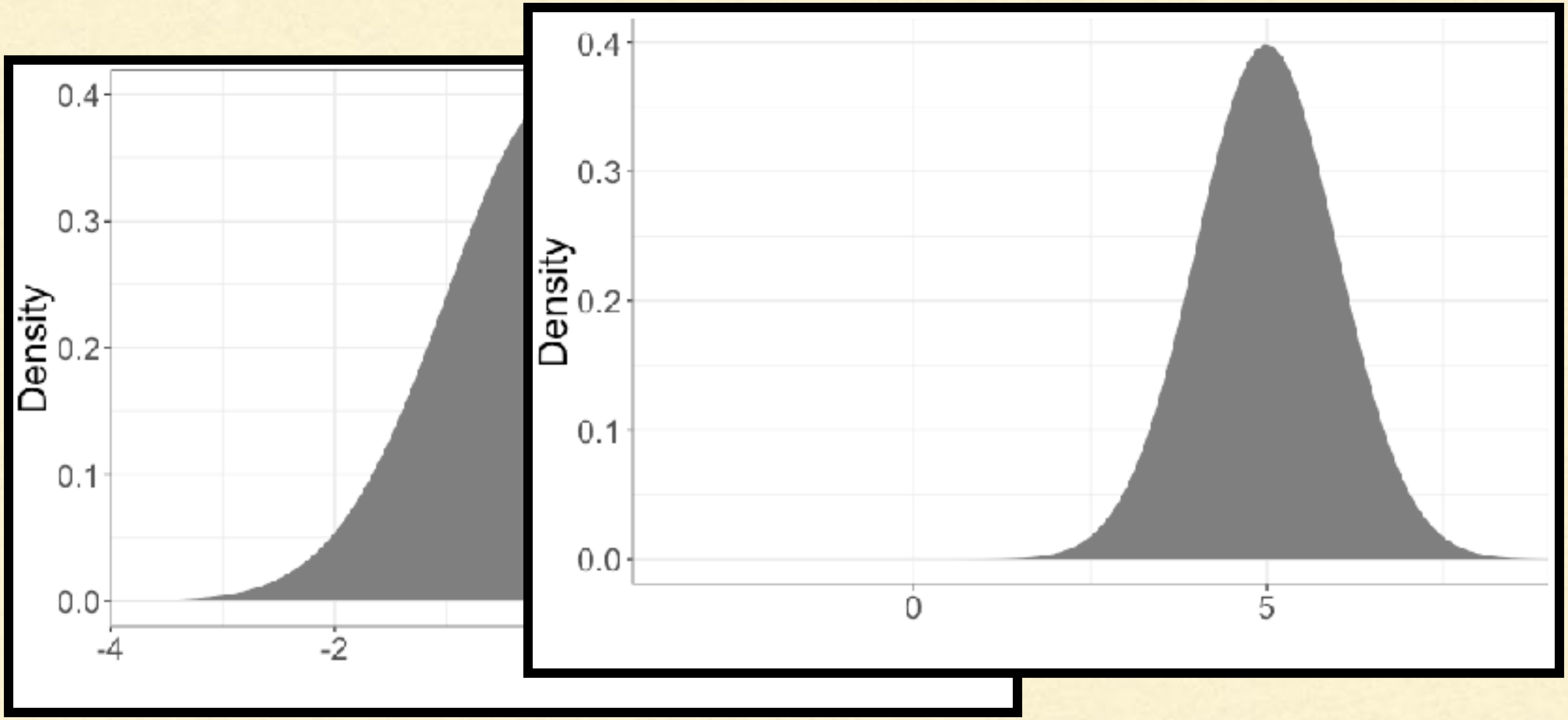
EXPECTED VALUE

- What you think is the expected value of these distributions?



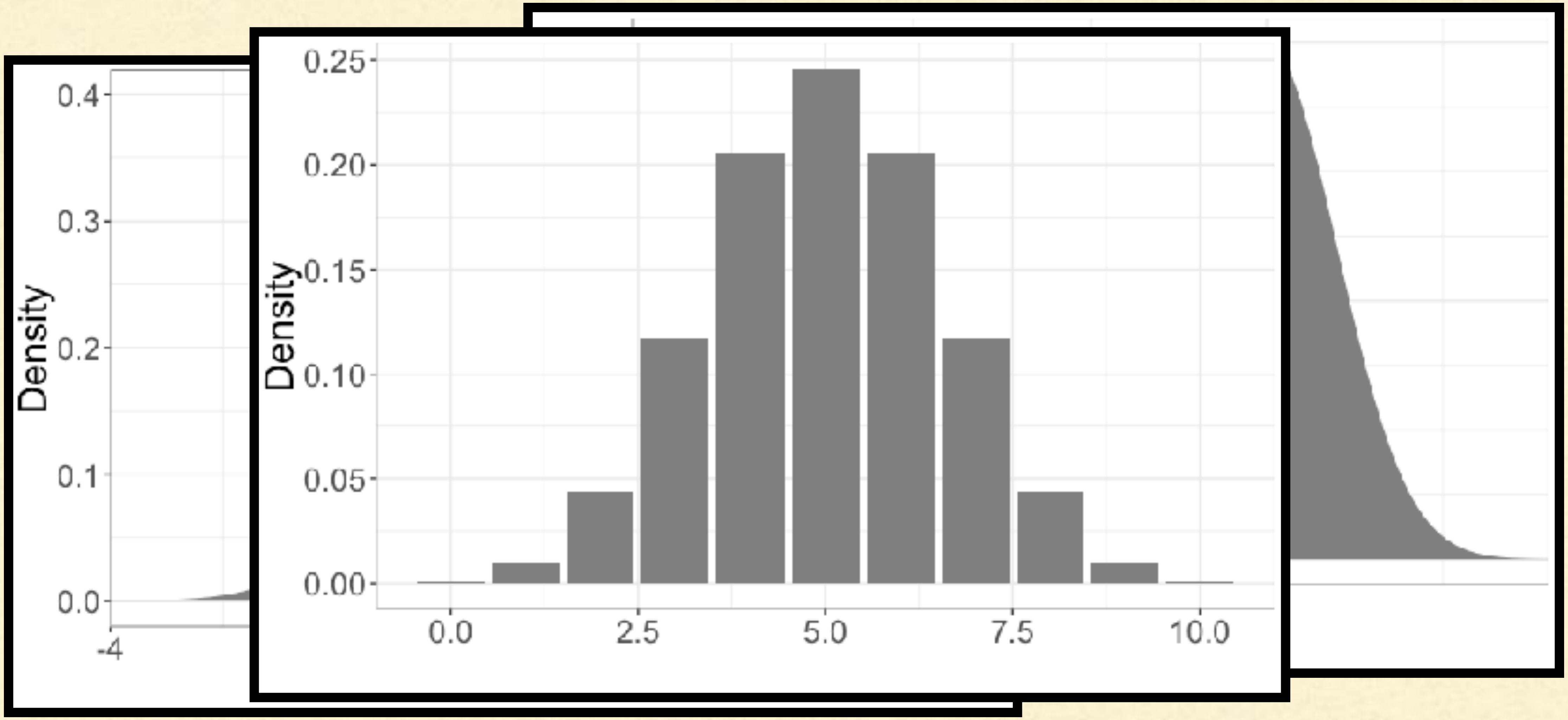
EXPECTED VALUE

- What you think is the expected value of these distributions?



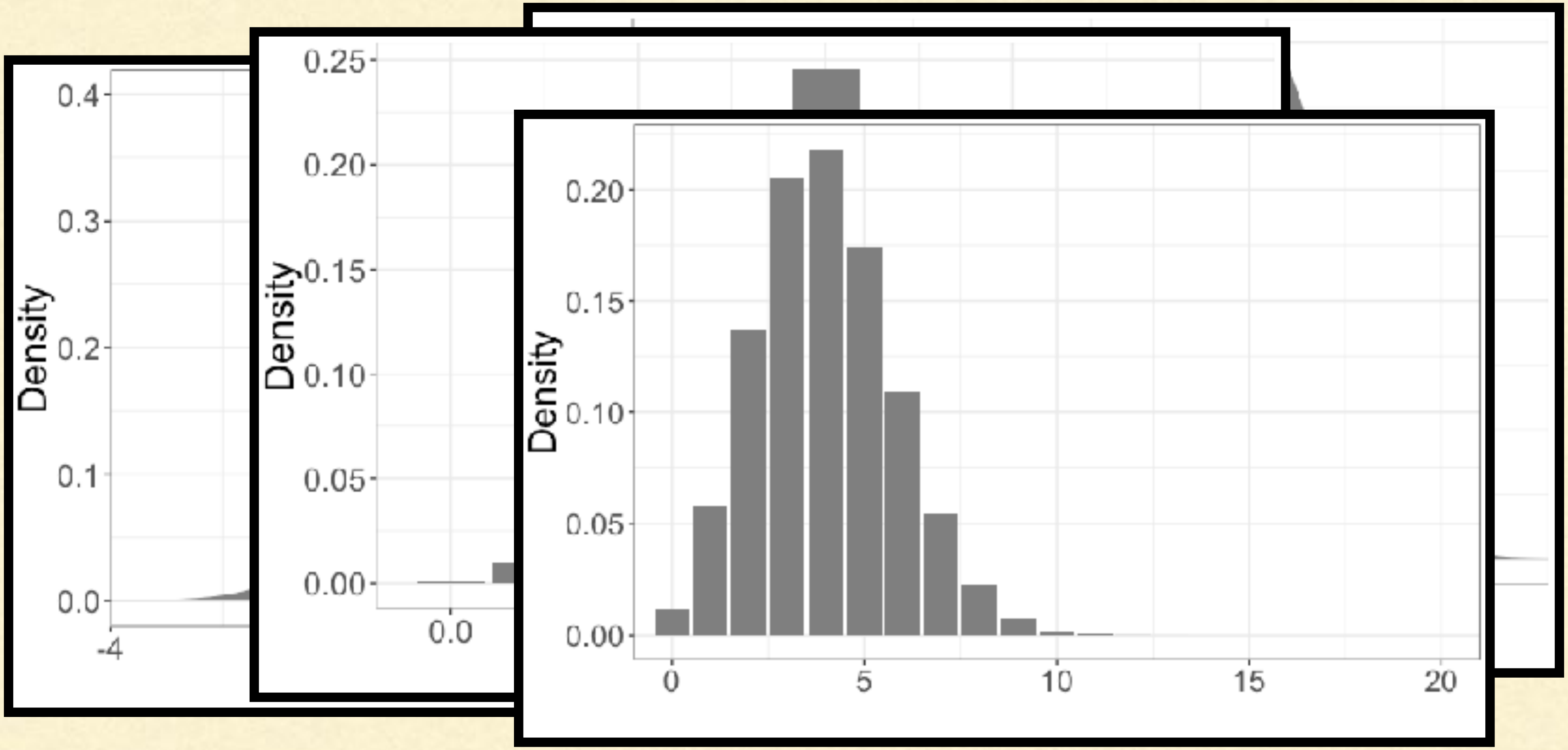
EXPECTED VALUE

- What you think is the expected value of these distributions?



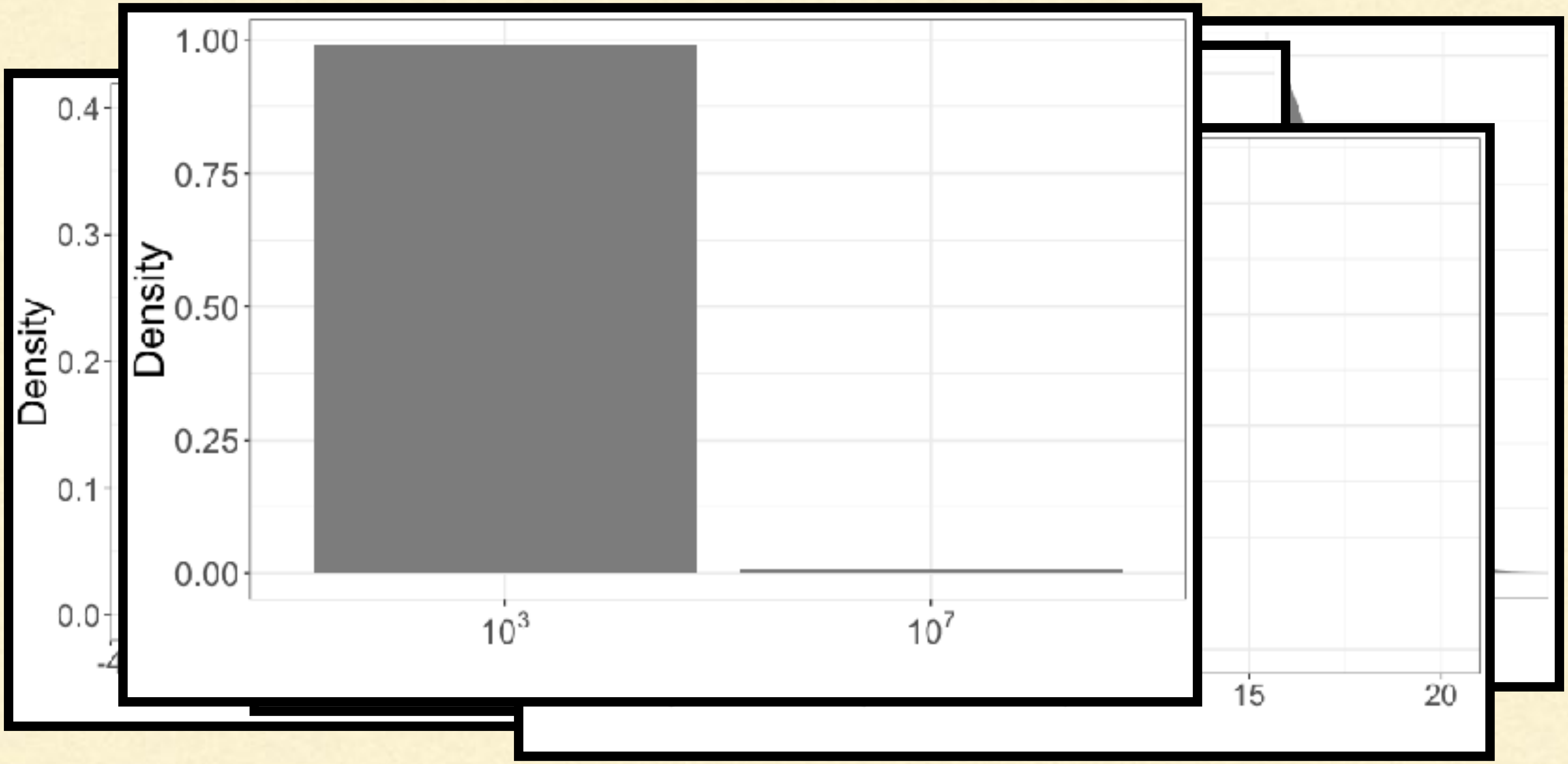
EXPECTED VALUE

- What you think is the expected value of these distributions?



EXPECTED VALUE

- What you think is the expected value of these distributions?



EXPECTED VALUE

- Suppose we have a distribution with discrete support on $x_1, ..., x_n$, and the probability of selecting each point is $p(x_1), ..., p(x_n)$. Then

$$\text{Expected value} = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n)$$

- Suppose we have a distribution with continuous support on x_{lower} to x_{upper}, and the density of a point x is f(x). Then

$$\text{Expected value} = \int_{x_{\text{lower}}}^{x_{\text{upper}}} x f(x) dx$$

BIAS

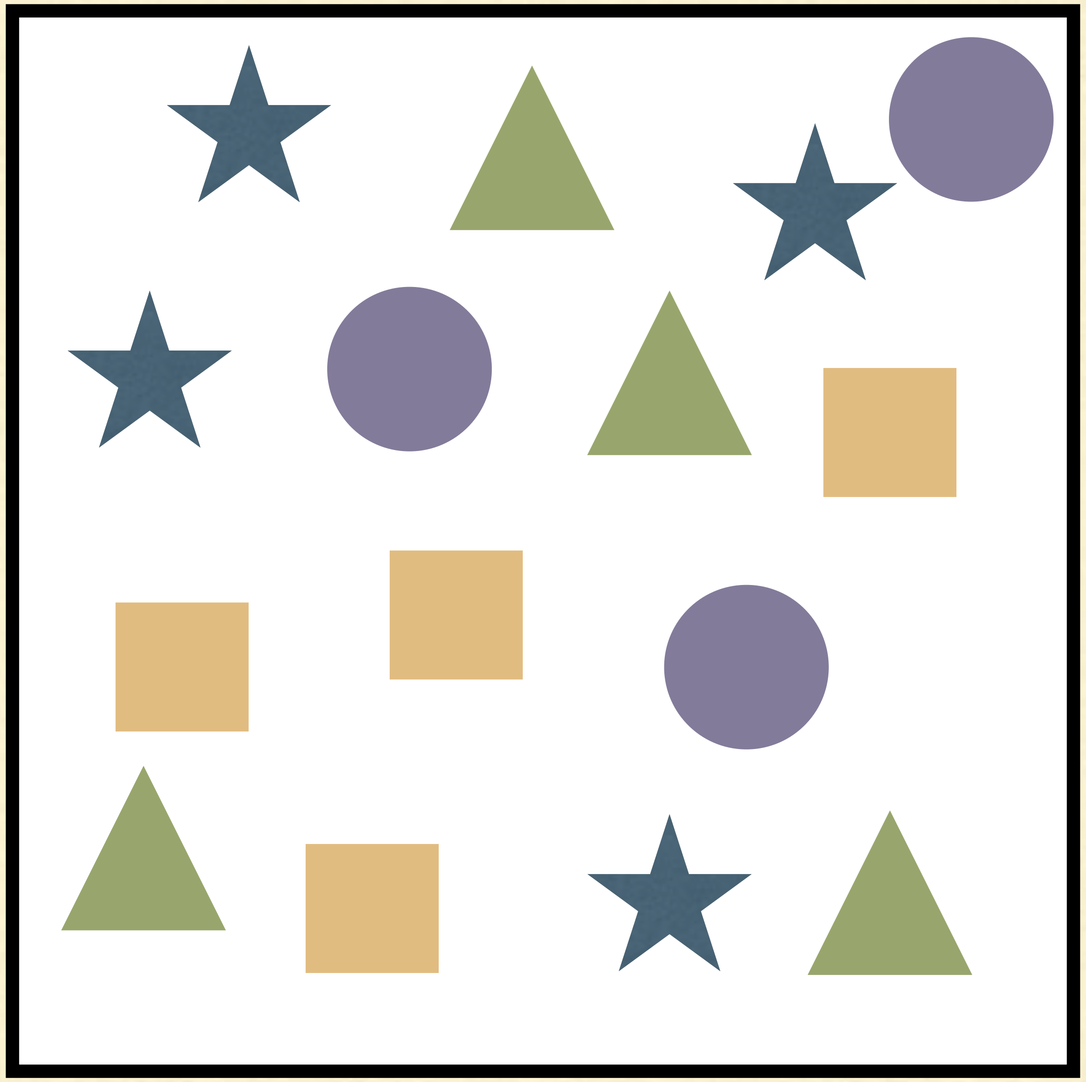
- An estimate is unbiased if its bias is zero
- The fine print:
  - An estimate of a parameter
    - is unbiased if
      - its bias is zero under the model
- The distribution of the estimate depends on the distribution of the data, and thus, on the model!

BIAS

- Be careful — this word is used frivolously
- Before being impressed, ask yourself
    - What is the estimate?
    - What is the model?
    - Is the model reasonable?
    - Why do they think its unbiased?

EXAMPLE

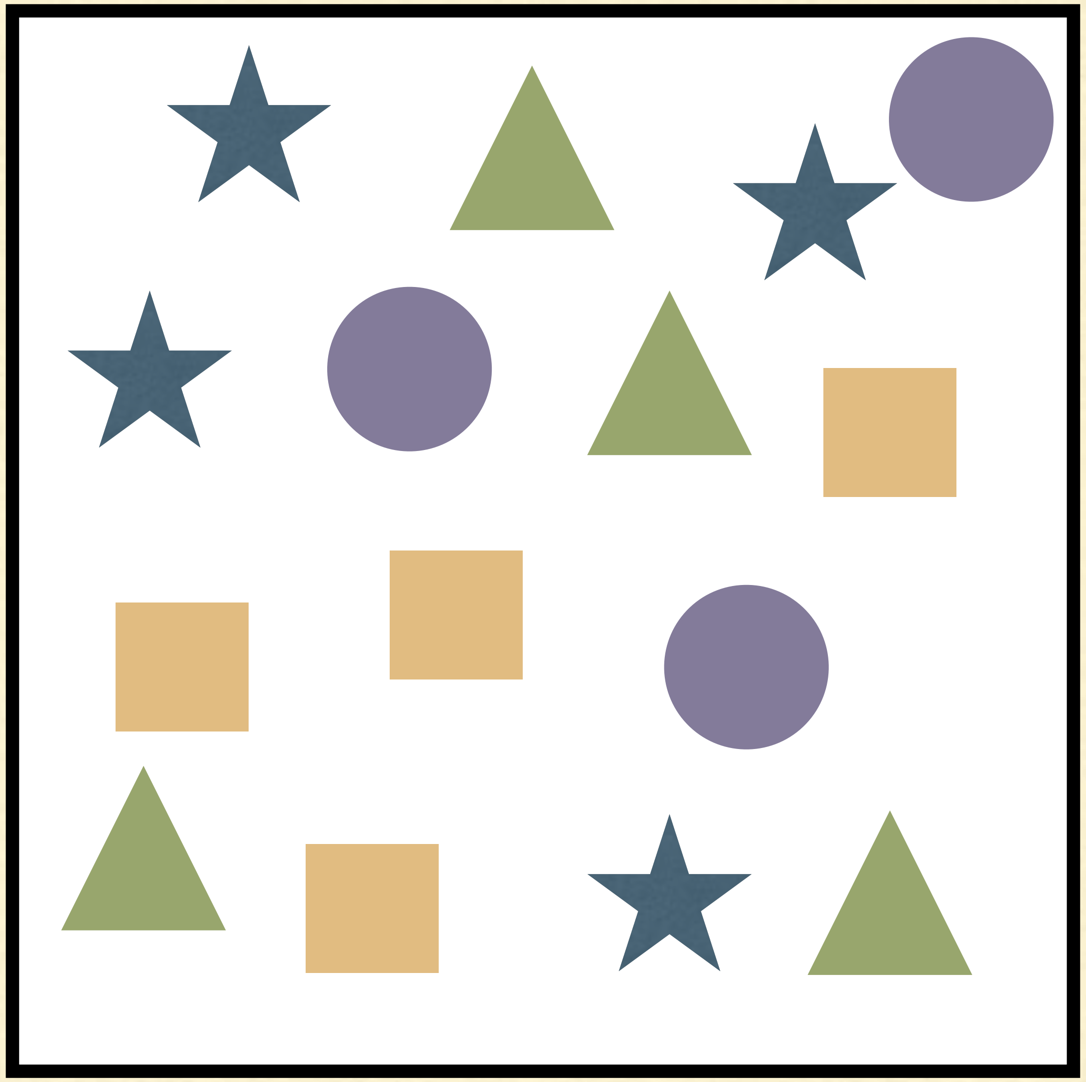
- (1 minute)
  - What are the true relative abundances in the community?



EXAMPLE

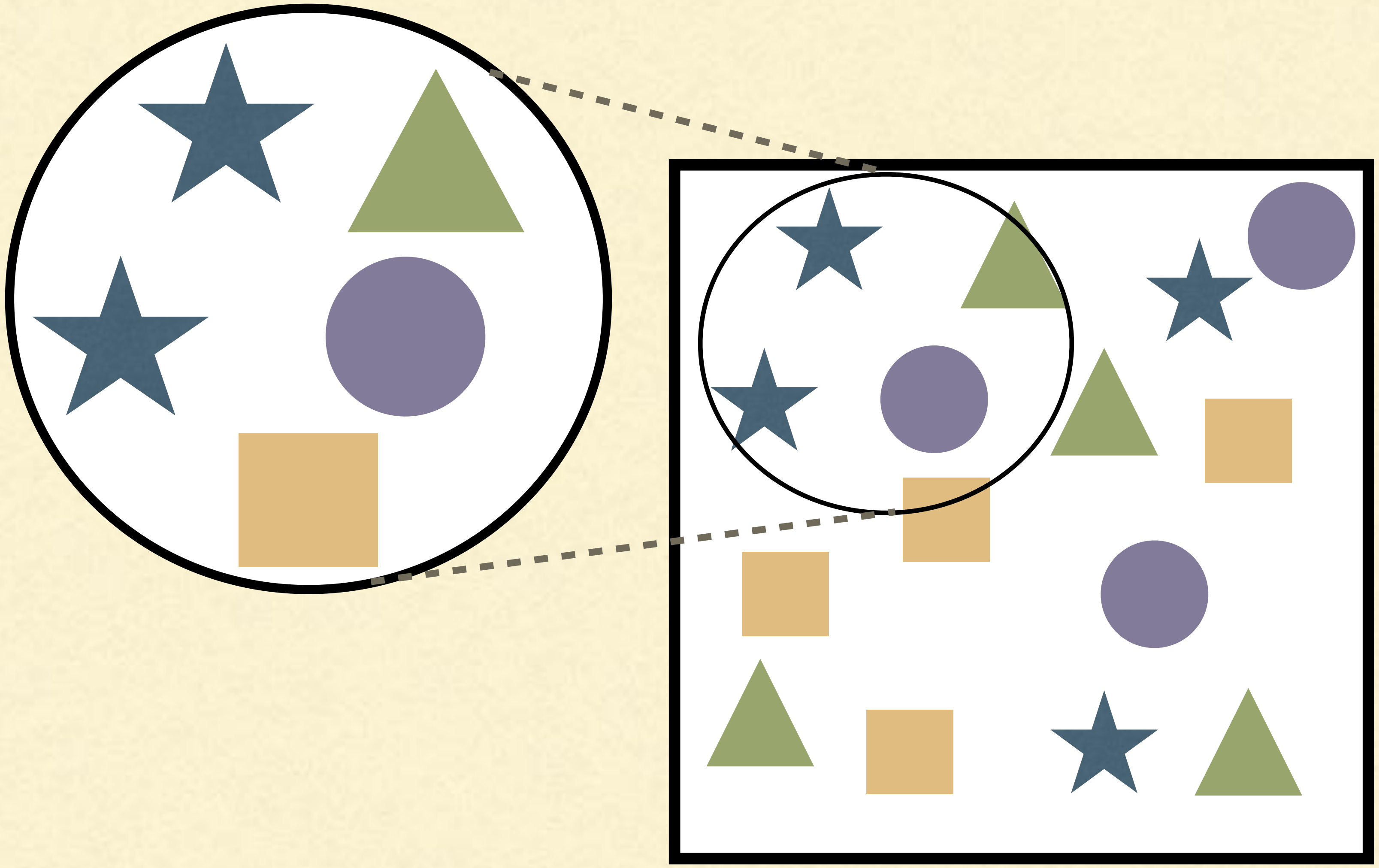
True abundances:

- ★ = 4/15
- ● = 3/15
- ▲ = 4/15
- ■ = 4/15



EXAMPLE

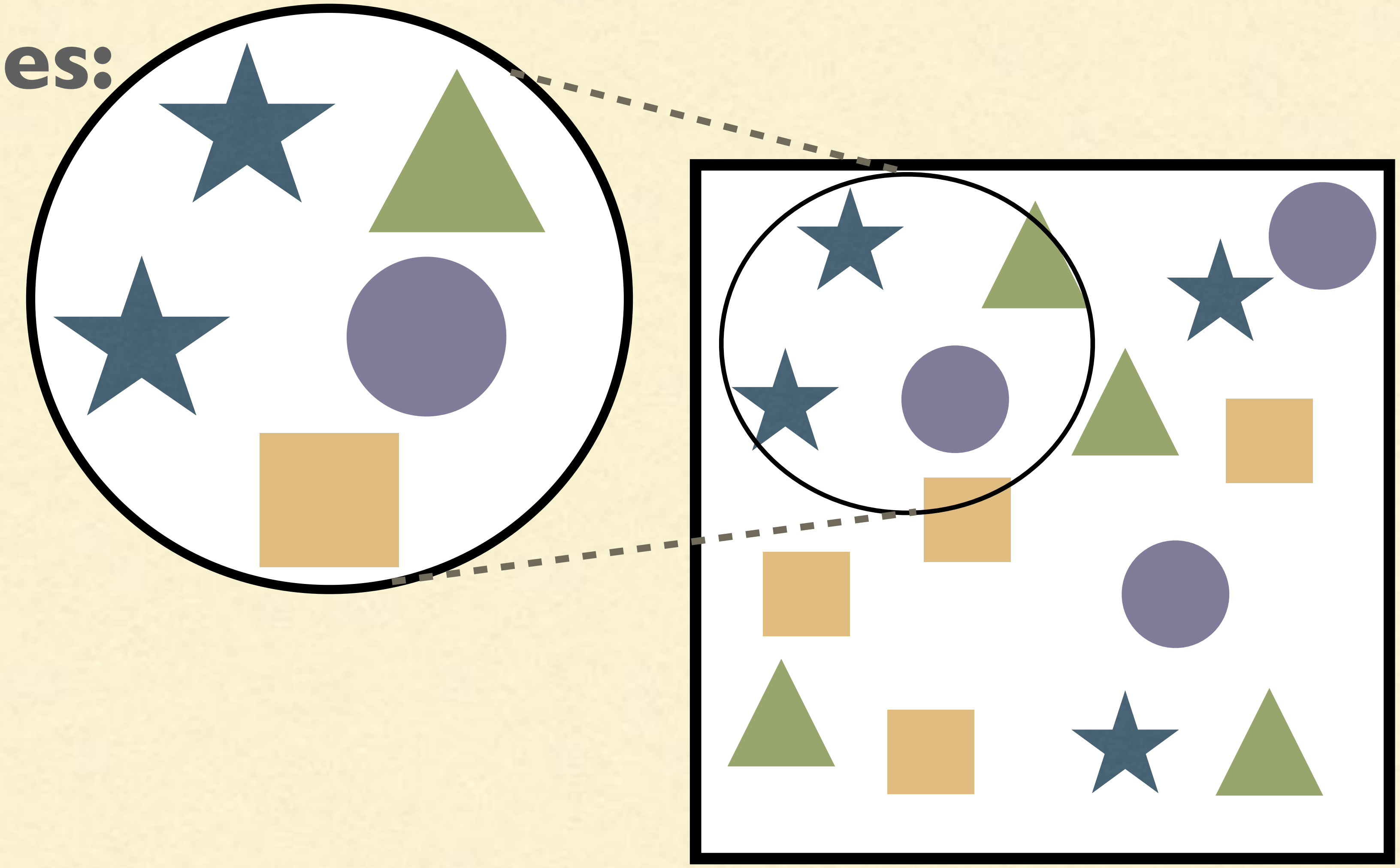
- (3 minutes)
  - Draw some nets. What are the observed relative abundances?



EXAMPLE

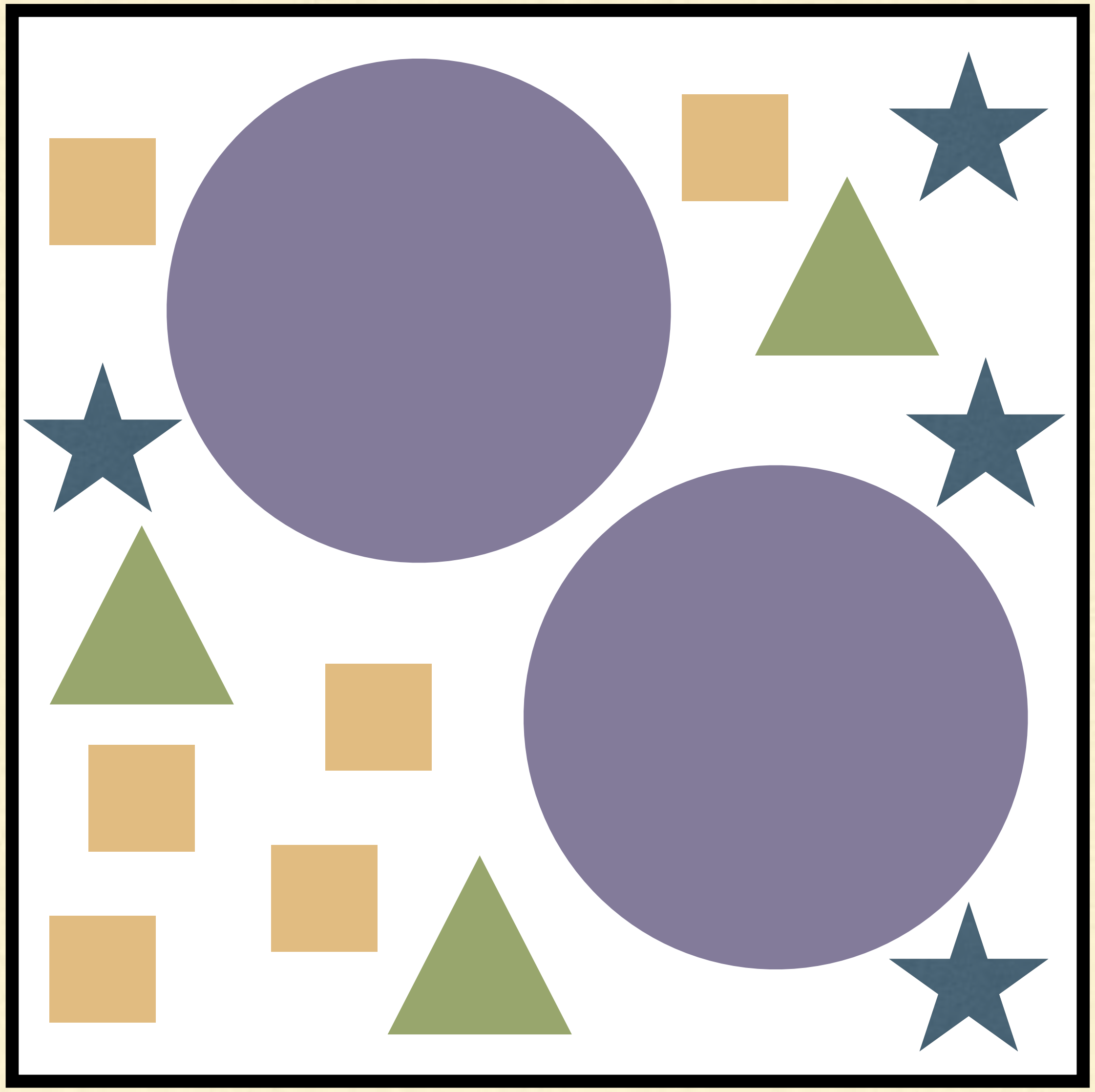
Observed abundances:

- ★ = 2/5
- ● = 1/5
- ▲ = 1/5
- ■ = 1/5



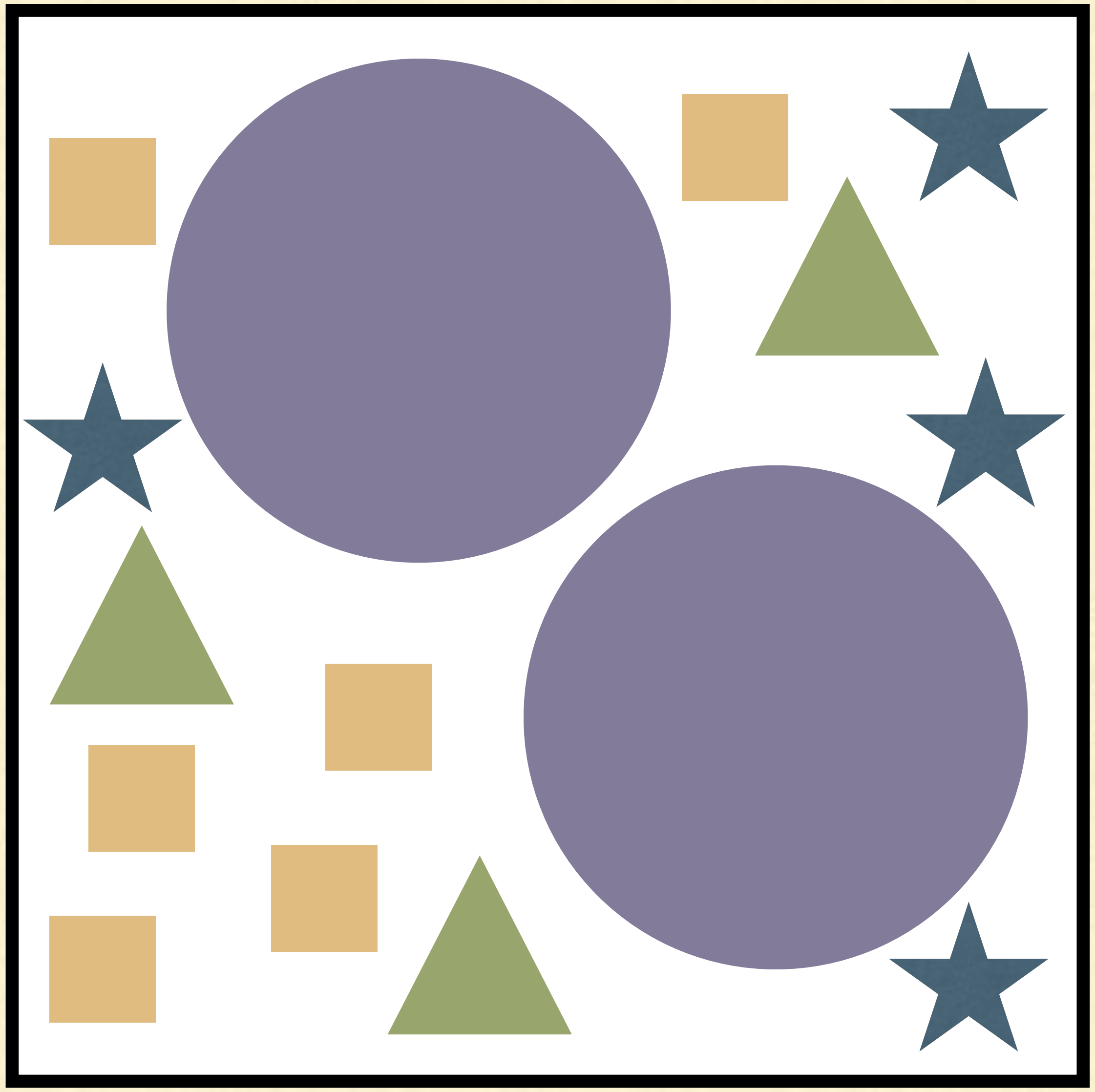
BIAS

- (1 minute)
  - What are the true relative abundances in the community?



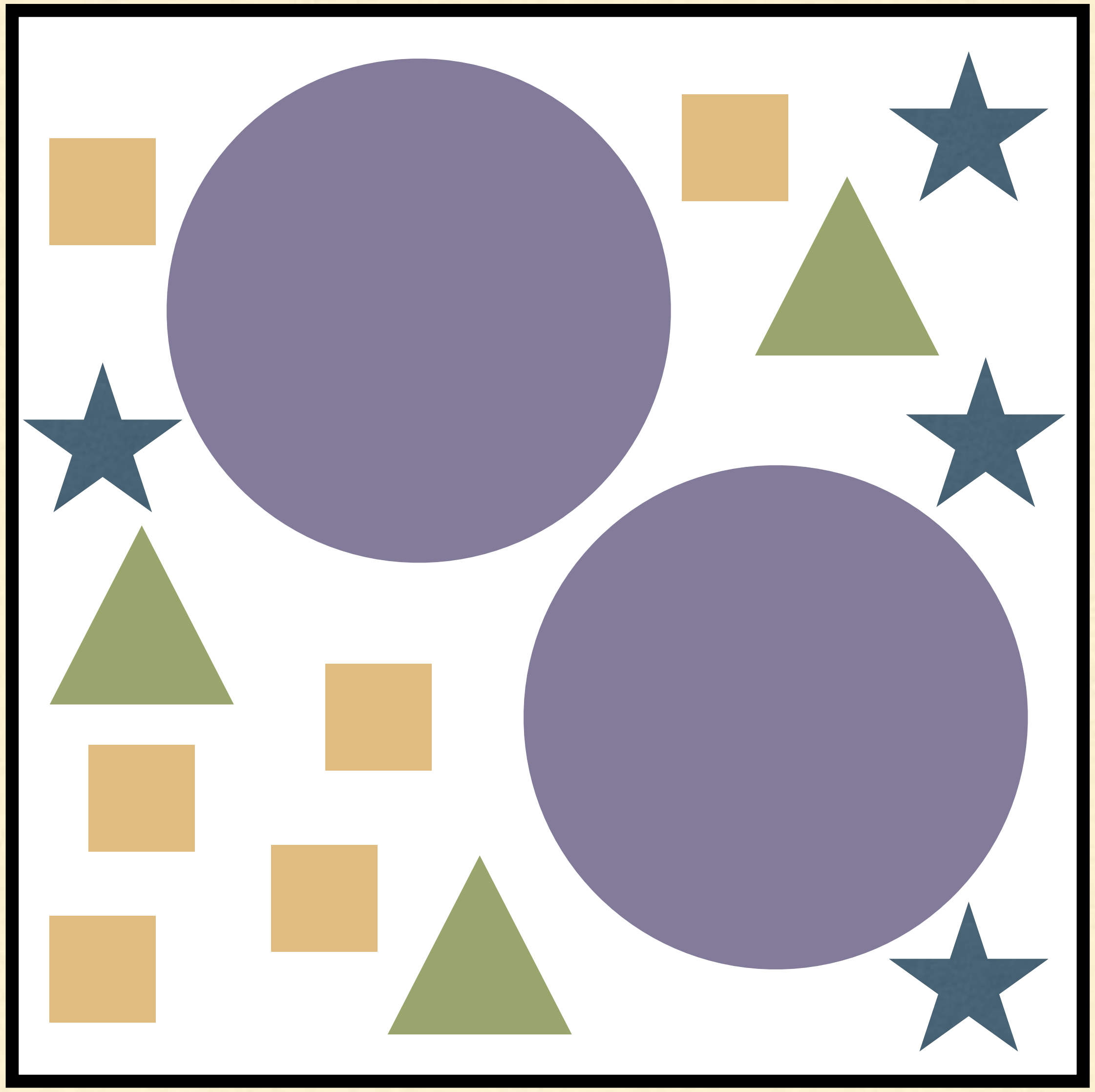
BIAS

- ★ = 4/15
- ● = 2/15
- ▲ = 1/5
- ■ = 2/5



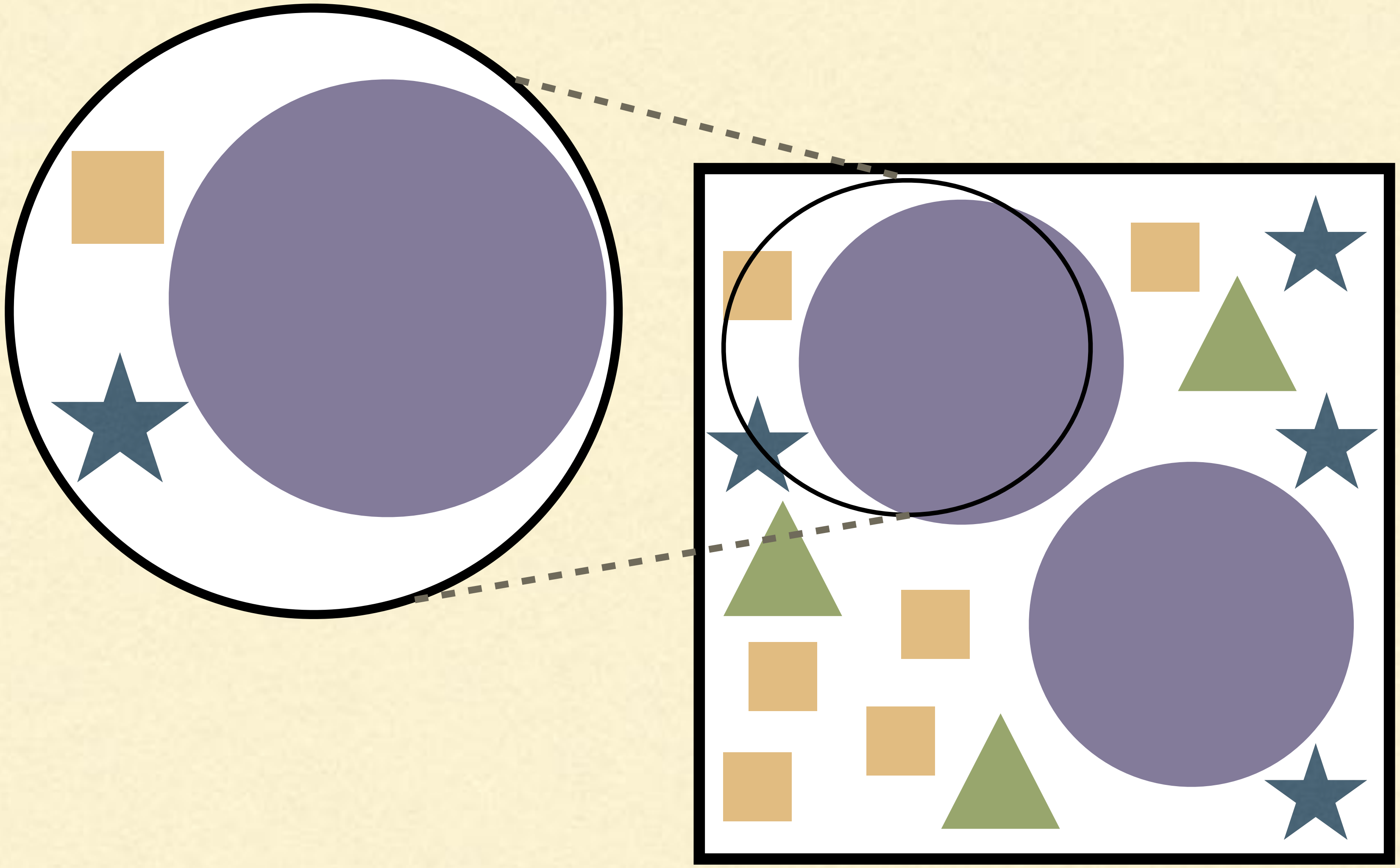
BIAS

- (3 minutes)
  - Draw some nets. What are the observed relative abundances?



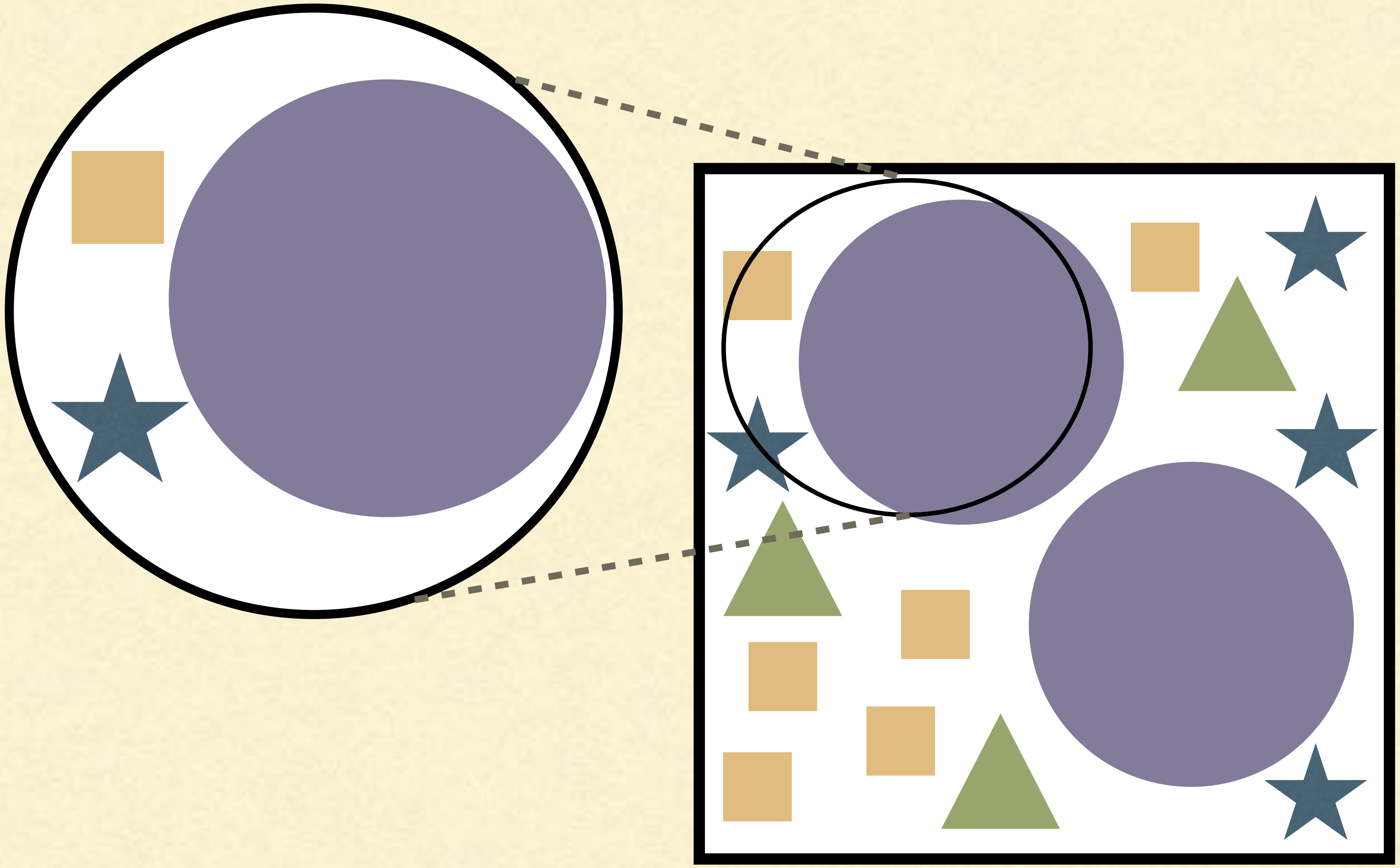
BIAS

- ★ = 1/3
- ● = 1/3
- ■ = 1/3



BIAS

- ★ = 5/15
- ● = 5/15
- ■ = 5/15



Truth:

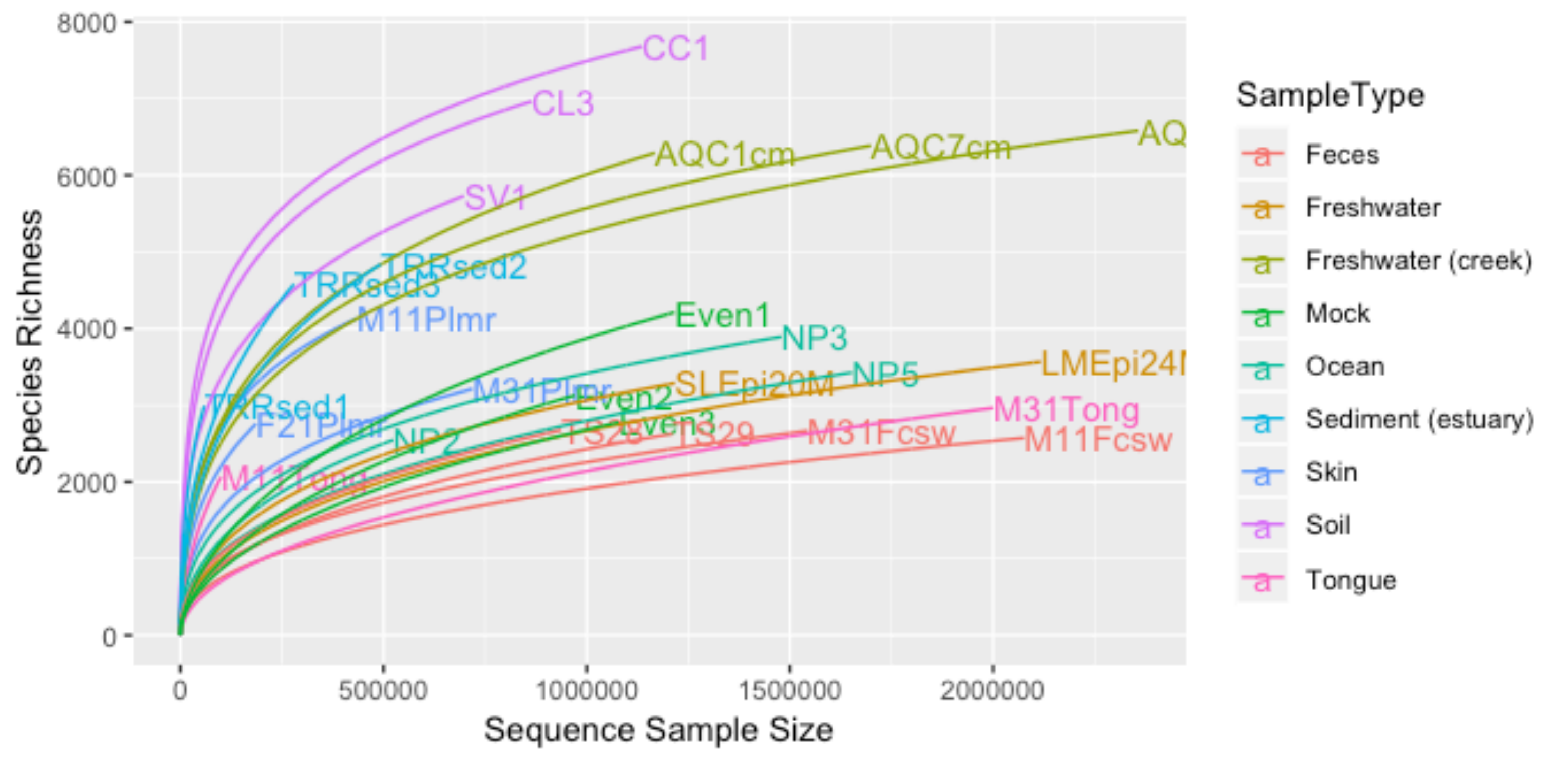
★ = 4/15 ● = 2/15 ▲ = 3/15 ■ = 6/15

BIAS AND DIVERSITY

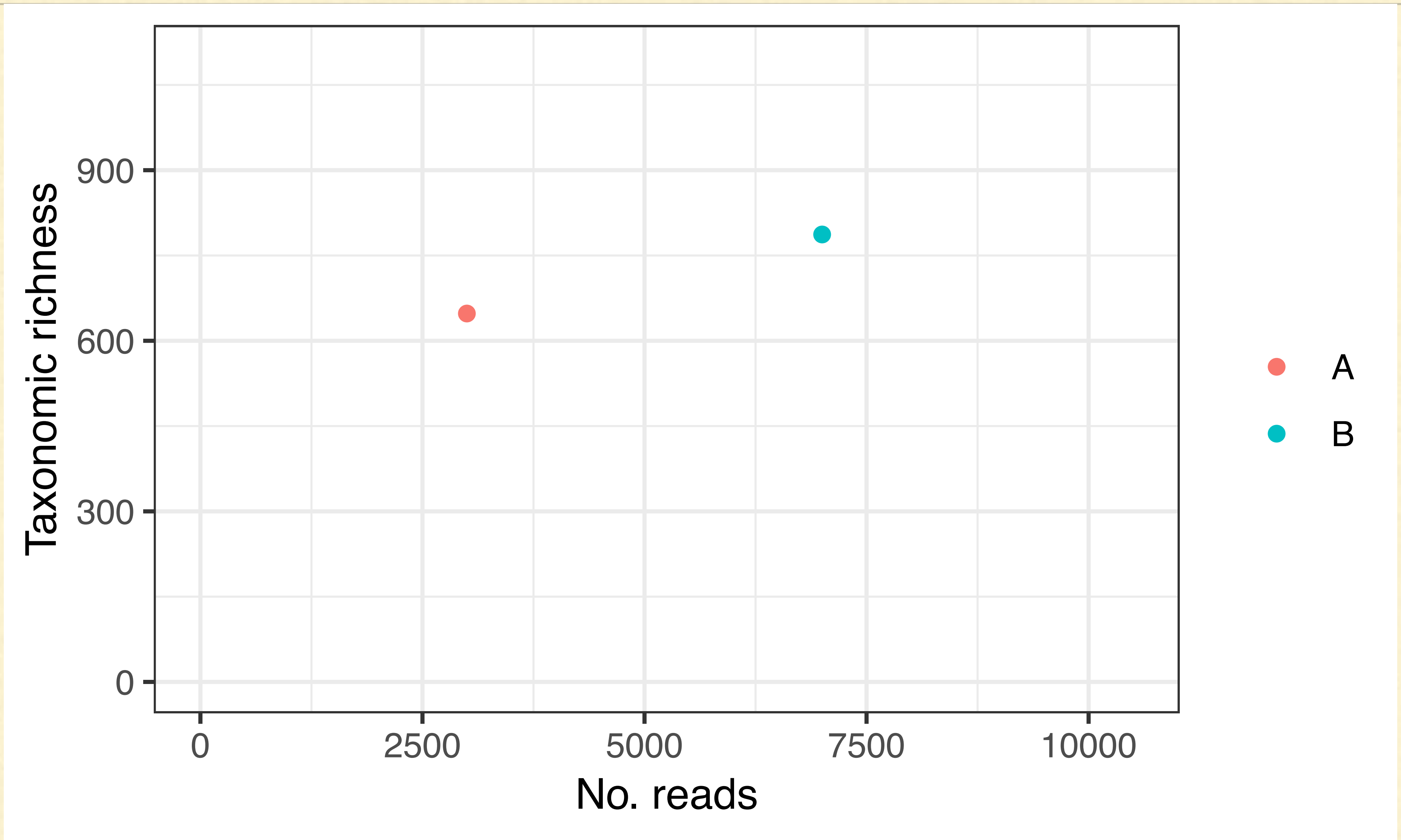
- Rarefaction curves are a fantastic illustration of what happens when people who don't understand statistics invent methods

- Rarefying, also called normalising, is a "method" for throwing away data to account for different levels of bias

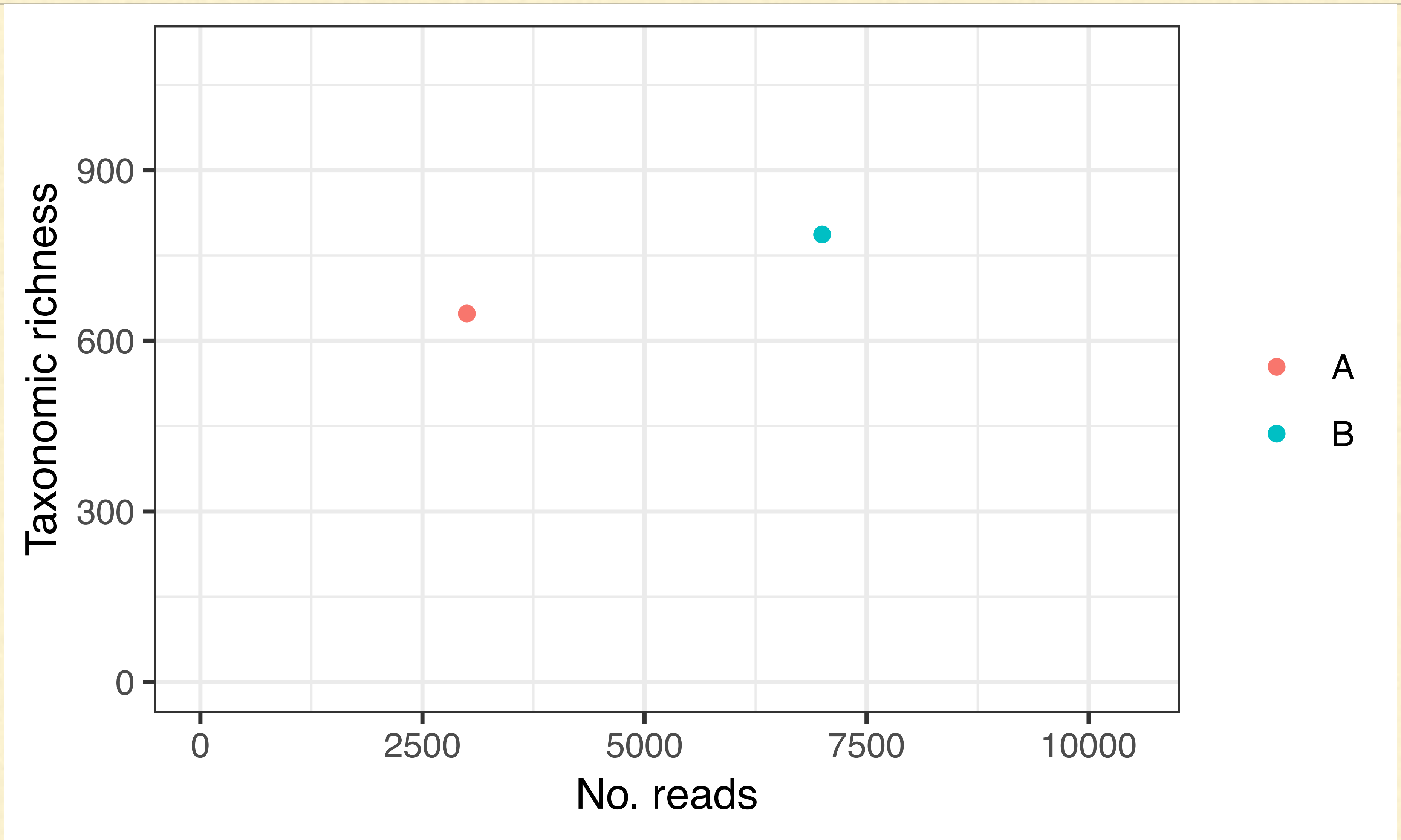
RAREFACTION



RAREFACTION

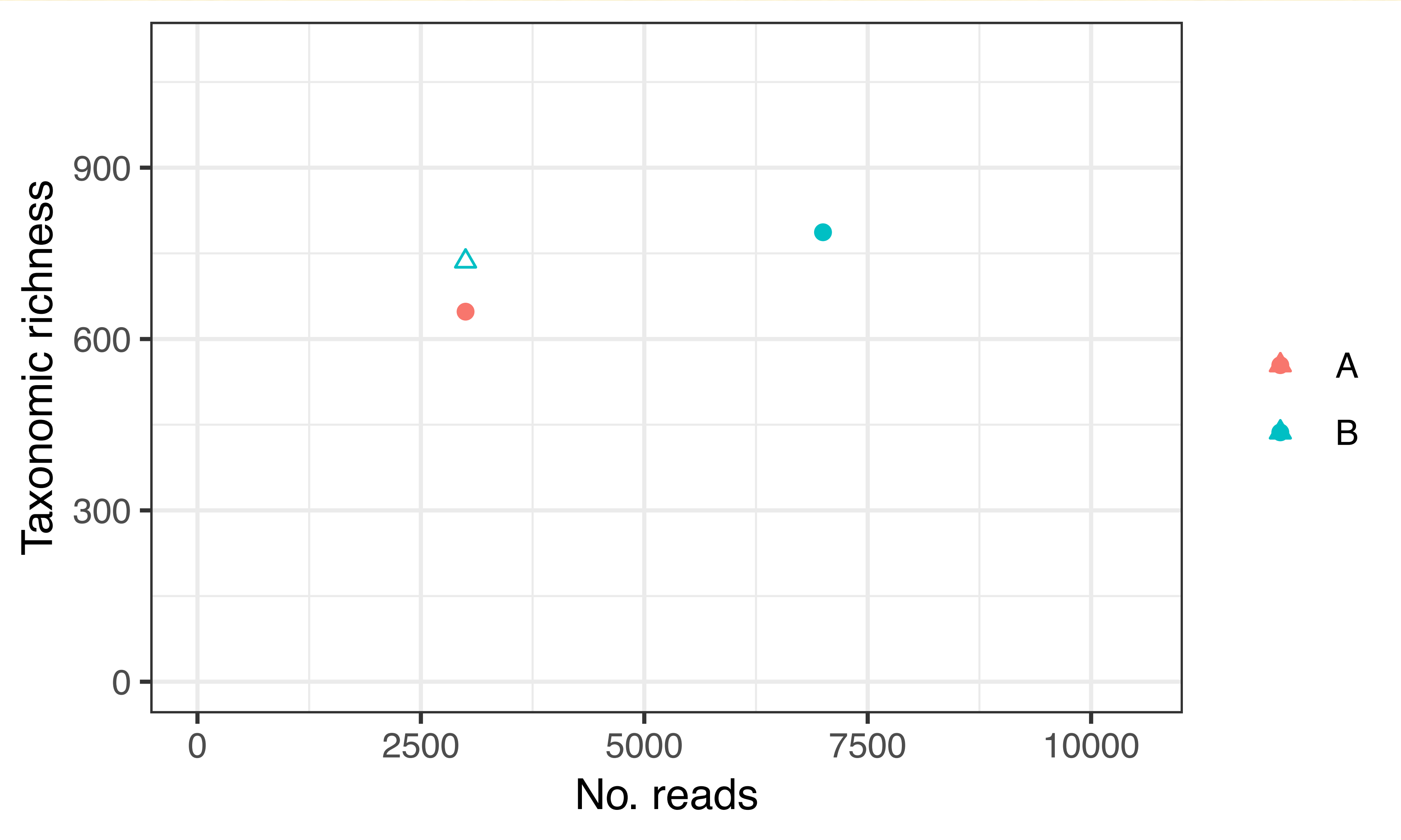


RAREFACTION



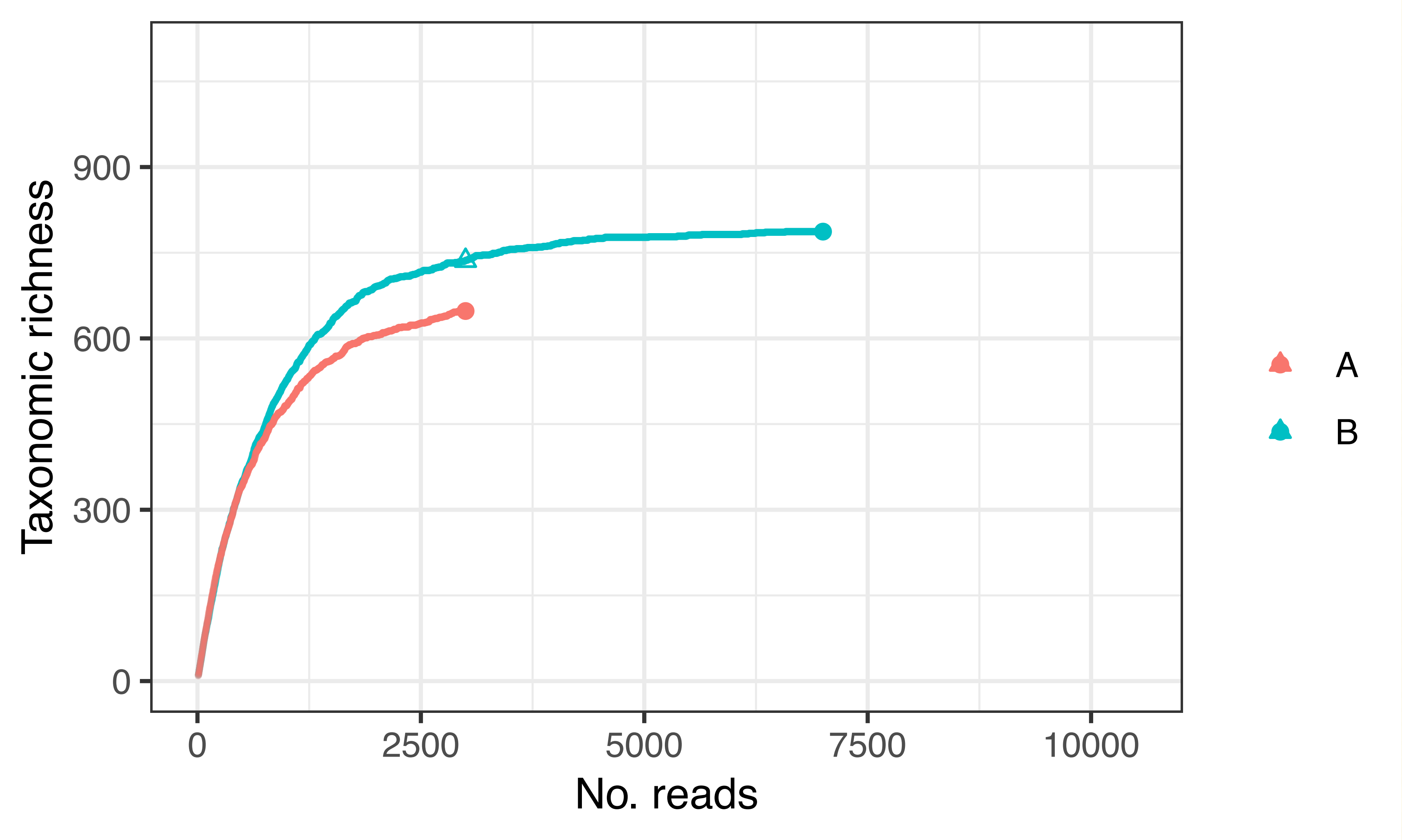
What's the parameter?
What's the estimate?

RAREFACTION



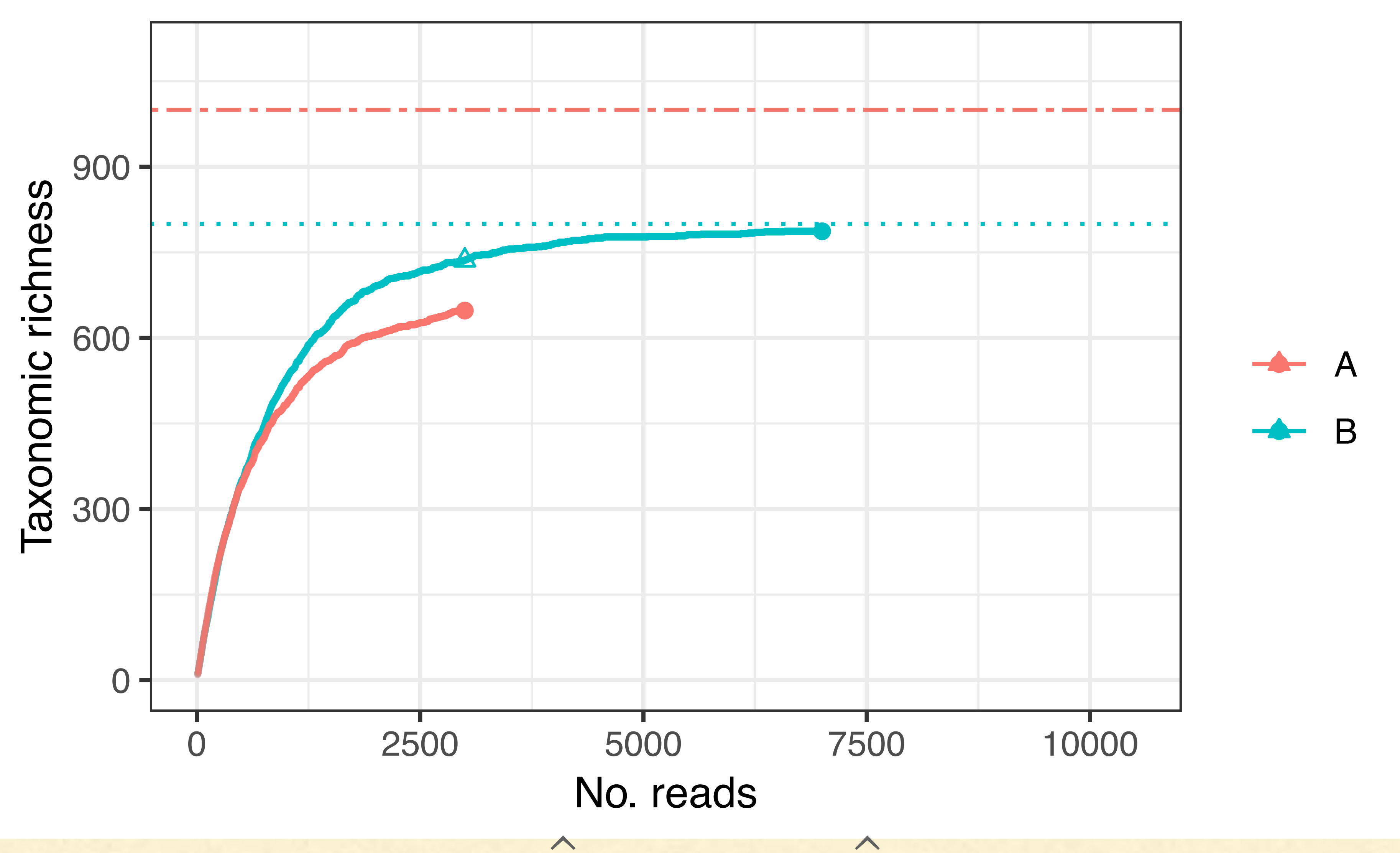
Now what's the parameter?
Now what's the estimate?

RAREFACTION



RAREFACTION

What's the bias of each estimate? Which is more biased?



$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

BIAS

- Some estimates are biased
- Biased estimates are not good, especially if the bias depends on sample size
- There is a solution (and it's not rarefying)
  - More discussion on Thursday when we talk about species richness

BREAK

☕☕☕☕☕☕☕☕🍪☕

☕☕☕☕☕☕☕☕☕☕

☕☕☕☕☕☕☕☕☕☕

VARIANCE

- Variance is a property of the estimate
- It describes how much the estimate varies
- Variance actually isn't about the parameter
- Definition:

$$\text{Variance}(\hat{\theta}) = \mathbb{E}\left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right)^2$$

VARIANCE IN REAL LIFE

- Definition:

$$\text{Variance}(\hat{\theta}) = \mathbb{E}\left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right)^2$$

- What does this expectation mean *in terms of your experiment*?

VARIANCE

- The variance reflects how far apart the repeated estimates are
- If your estimates (from repeated experiments) are
  - 12, 12, 12, 12, 12… => variance is 0
  - 12, 12, 12, 13, 12… => variance is 0.167
  - 12, 12, 12, 13013, 12… => variance is 28171000
- A large change in the estimates equals a large variance

VARIANCE

- Repeat the experiment, calculate the estimate => $\hat{\theta}_1$
- Repeat the experiment again, calculate the same estimate => $\hat{\theta}_2$
- $\ldots$
- Let $\hat{\theta}_j$ be your estimate from the j-th time you do the experiment

$$\text{Variance} = \text{limit of average}\left(\hat{\theta}_{\text{repeat } j} - \text{average } \hat{\theta}\right)^2$$

VARIANCE

- Repeating our experiment is expensive, and so we use models to estimate the variance
  - the variance we would get if we repeated the experiment again and again and again...

$$\text{Variance} = \text{limit of average}\left(\hat{\theta}_{\text{repeat } j} - \text{average } \hat{\theta}\right)^2$$

STANDARD DEVIATION

- The standard deviation is the square root of the variance
  - Sometimes it's more convenient to work on the original scale of the data
- *What is a standard error?*

ESTIMATING THE VARIANCE

- The variance and standard deviation *of an estimate* are not known!
  - The data is random
  - The estimate is a function of your data => estimate is random
  - Distribution of the data is not known
    - That's why we're estimating a parameters
  - Distribution of the estimate is not known
  - The variance is therefore not known

STANDARD ERROR

- **standard error = estimate of standard deviation**
- Distinction:
    - If you have one-dimensional data, it has a standard deviation

      🎩
    - If you have a one-dimensional estimate, it has a standard error

STANDARD ERRORS

- "Model-based standard errors"
  - are based on models!
  - Your standard error is only as good as your model!

A MODEL

- Model: you have a community of Q microbes, which have relative abundances $p_1, ..., p_Q$
- You observe them independently
- The probability of observing microbe i on any draw is p_i

A MODEL

- You observe M_i individuals, W_i from group i

MICROBE i	1	2	3	4	5
Wi	5	1	1	2	1

A MODEL

MICROBE	1	2	3	4	5
WI	5	1	1	2	1

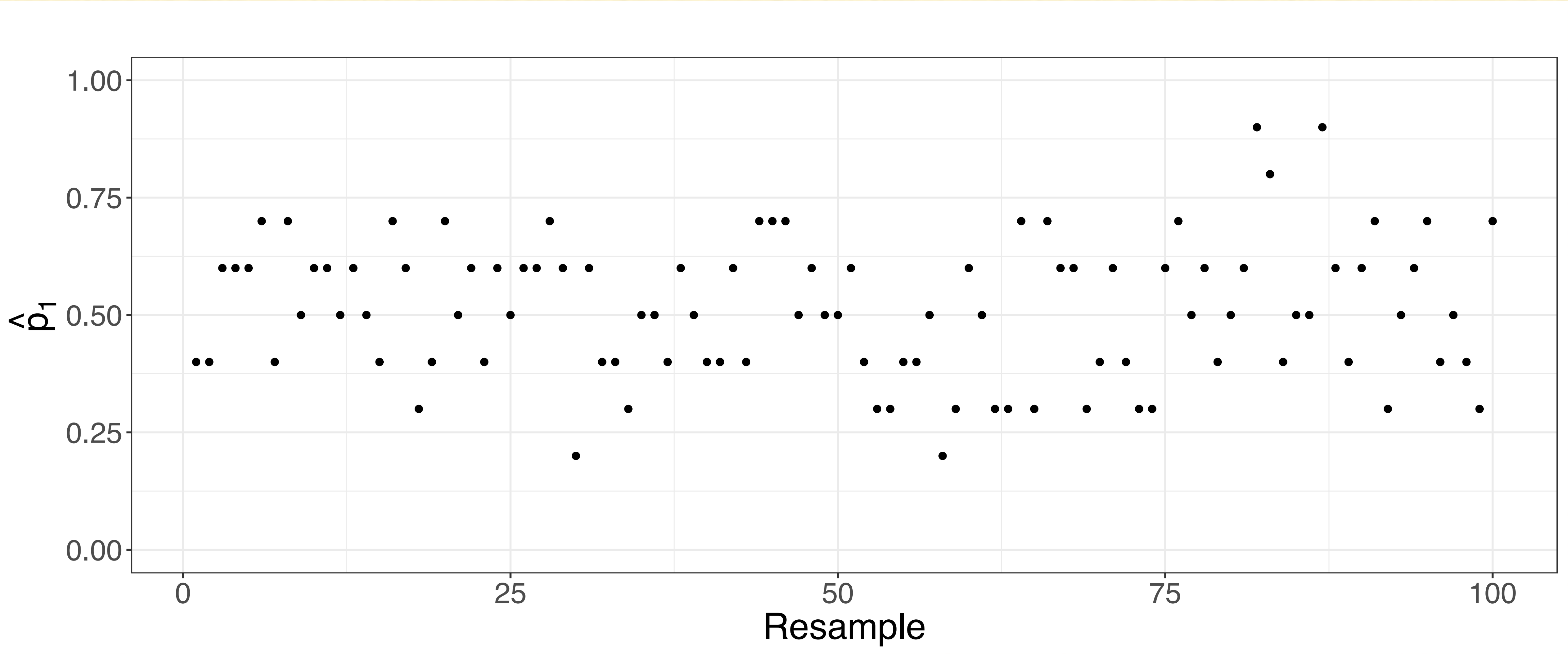
- Invent.... (4 minutes)
  - ...an estimate of Q
  - ...an estimate of p_1 and p_2 (probability of observing microbe 1&2)
  - ...a standard error for your estimate of Q
  - ...standard errors for your estimates of p_1 and p_2

A PROPOSAL

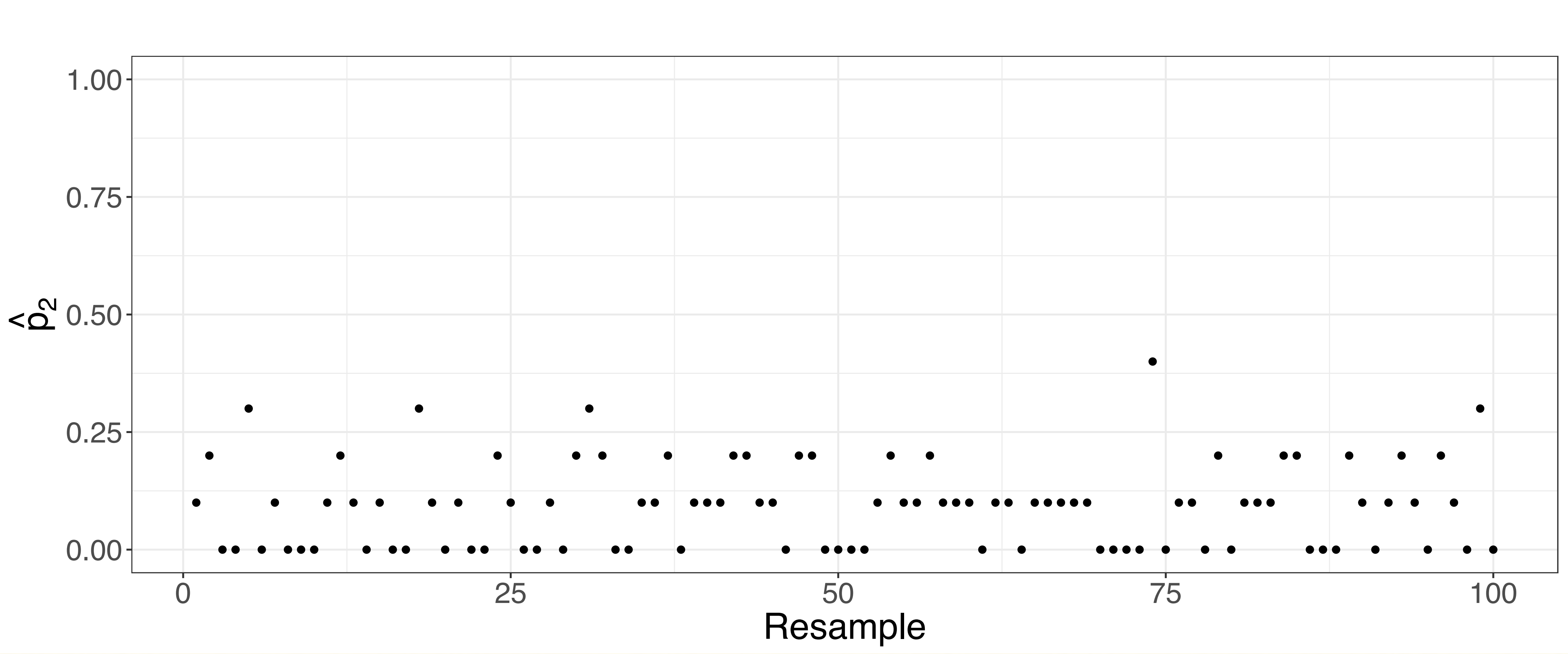
- (to illustrate; don't do this!)
  - You could take your sample and randomly sample Mi individuals from it

MICROBE	1	2	3	4	5
ORIGINAL	5	1	1	2	1
RESAMPLE 1	4	1	1	1	3
RESAMPLE 2	4	2	2	1	1
RESAMPLE 3	6	0	0	4	0
RESAMPLE 4	6	0	2	1	1

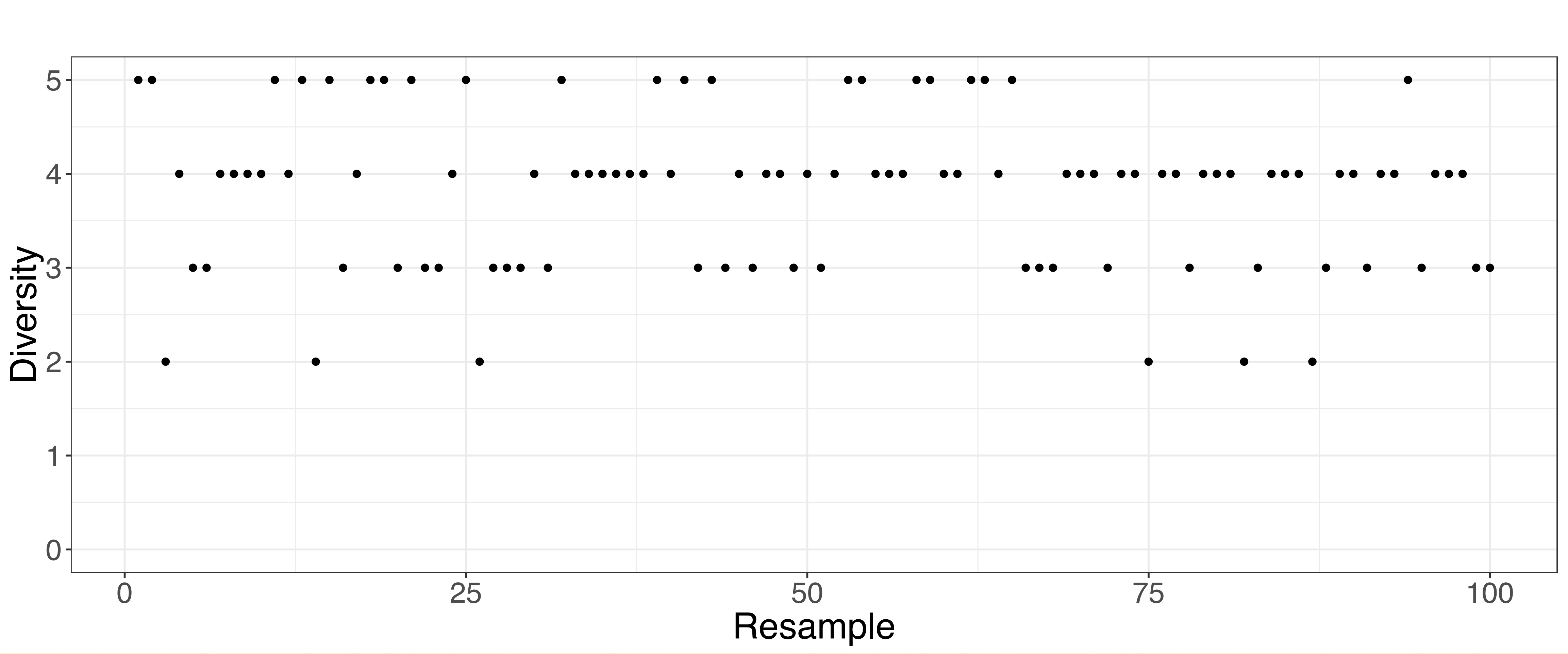
A PROPOSAL



A PROPOSAL



A PROPOSAL

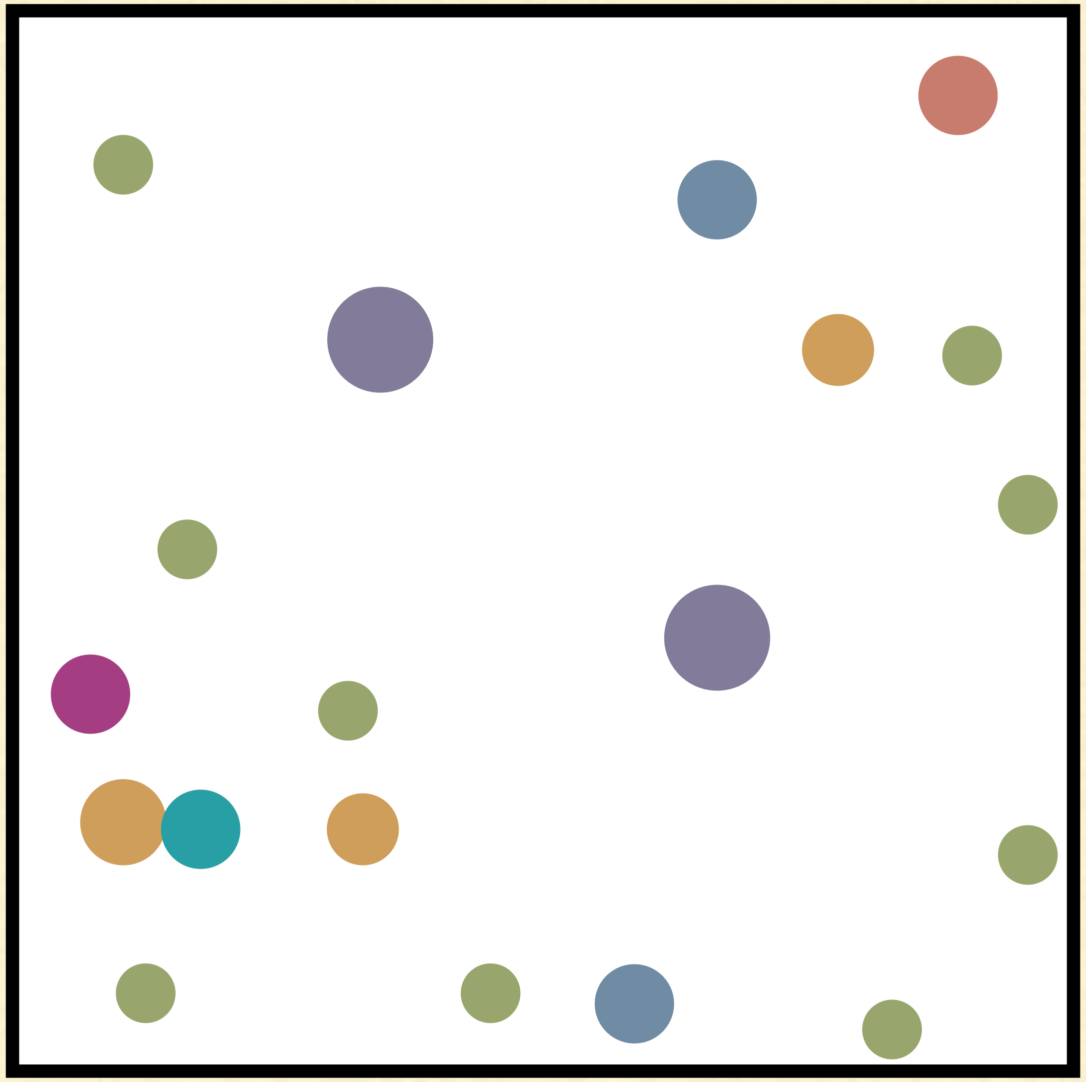


IS THIS A REASONABLE MODEL?

- This is a reasonable way to generate estimates and standard errors if this is a reasonable model
  - this is (one type of) the bootstrap
- Is this a reasonable model?

REASONABLE MODELS

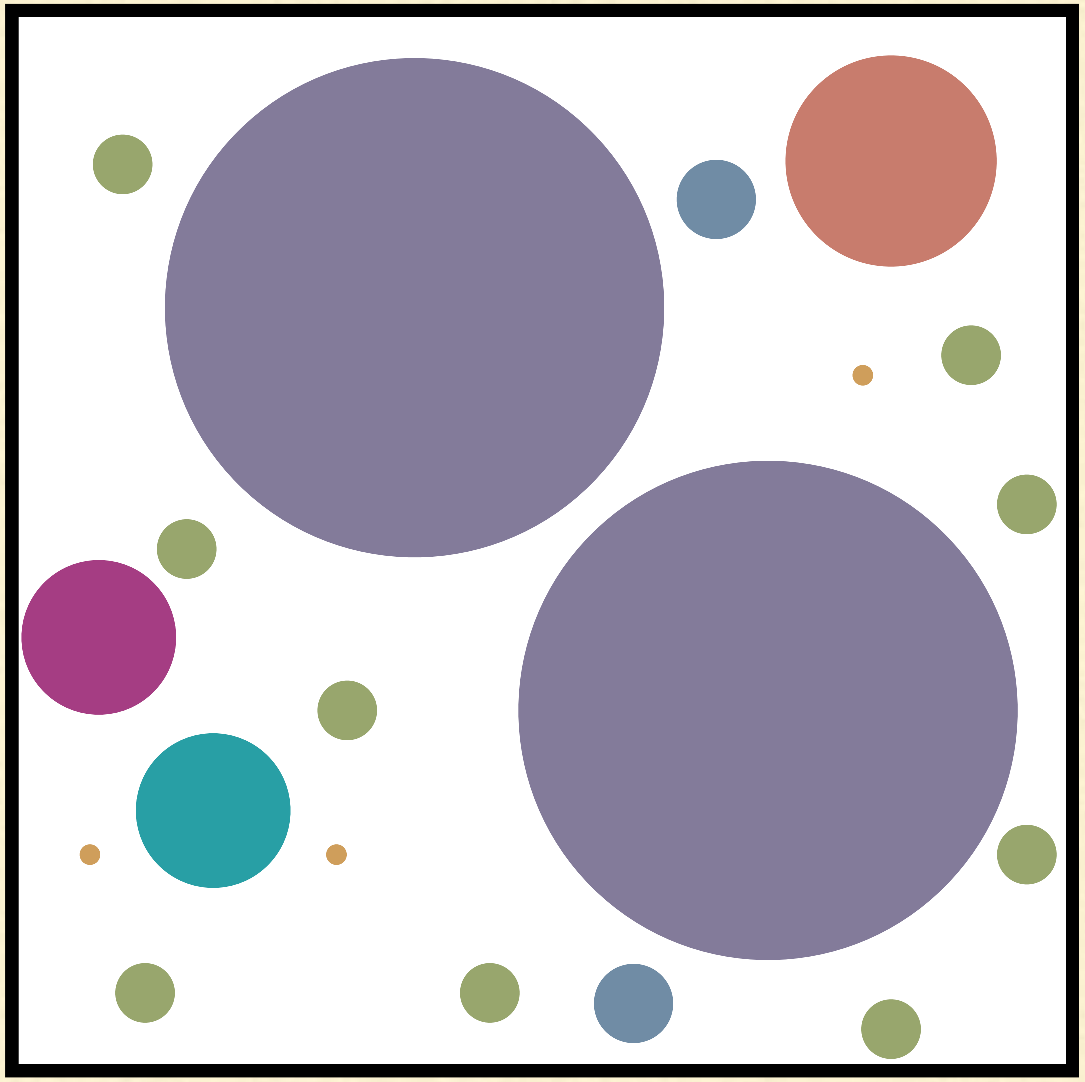
- Is subsampling reasonable here?



REASONABLE MODELS

- What about here?

- Does the parameter that you care about change your answer?



UNREASONABLE MODELS

- Our model from earlier is called a multinomial model
  - microbes observed independently
  - microbes observed in their abundances
  - which of these assumptions doesn't hold if we have a probability-proportional-to-size model?

INDEPENDENCE

- Now let's play with the independence assumption of a multinomial model

INDEPENDENCE

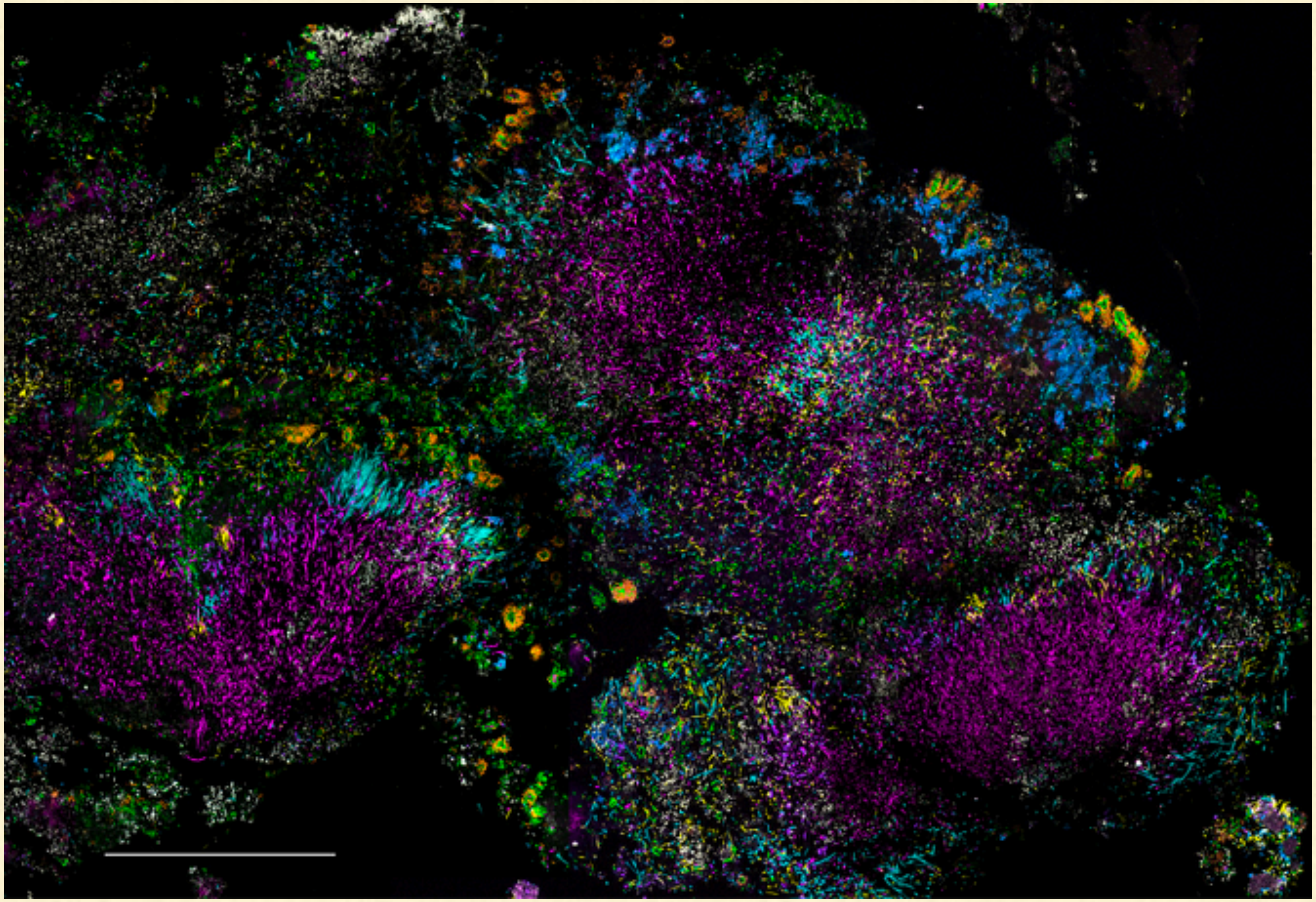


Photo credits: Mark Welch et al

INDEPENDENCE

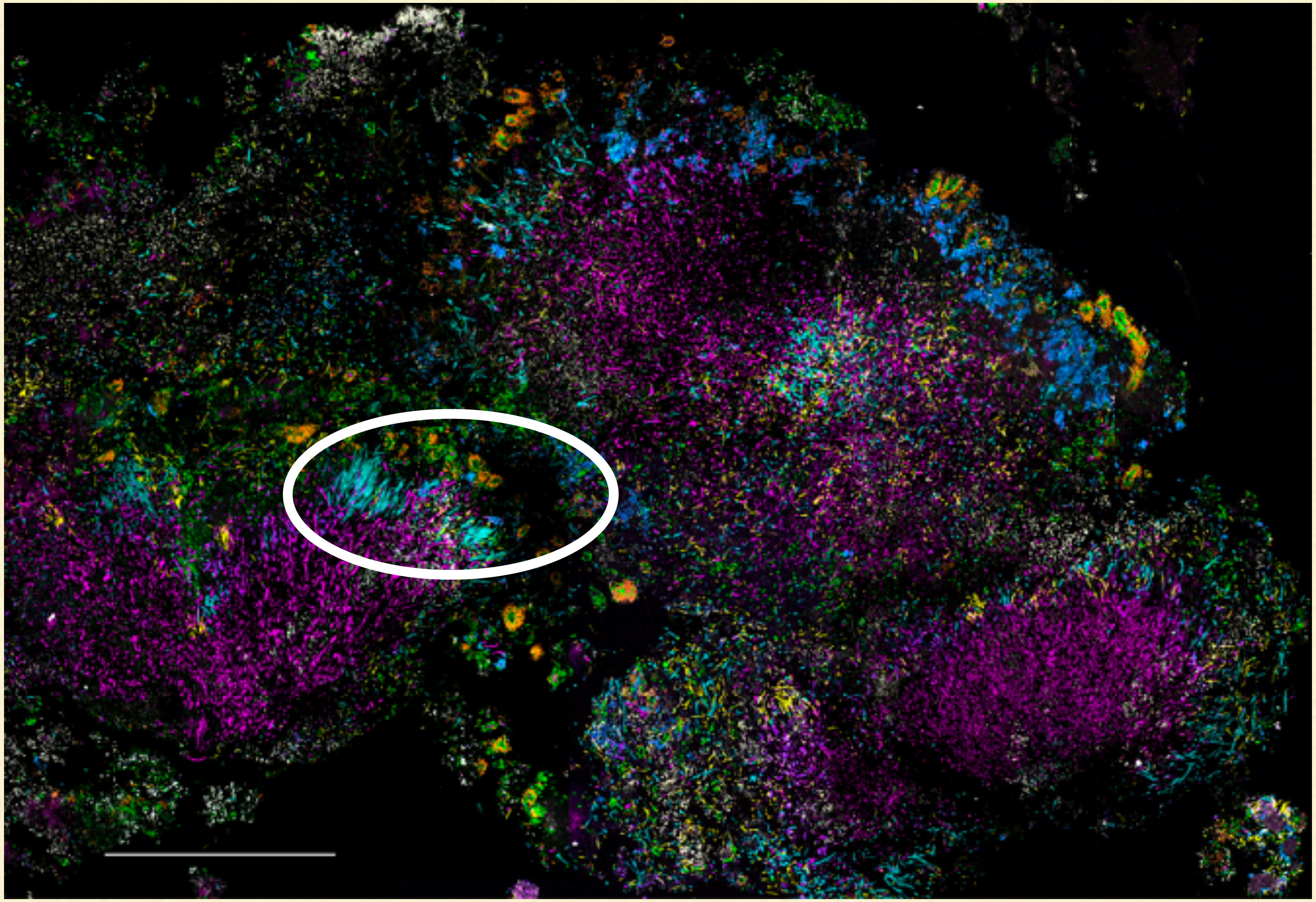


Photo credits: Mark Welch et al

INDEPENDENCE

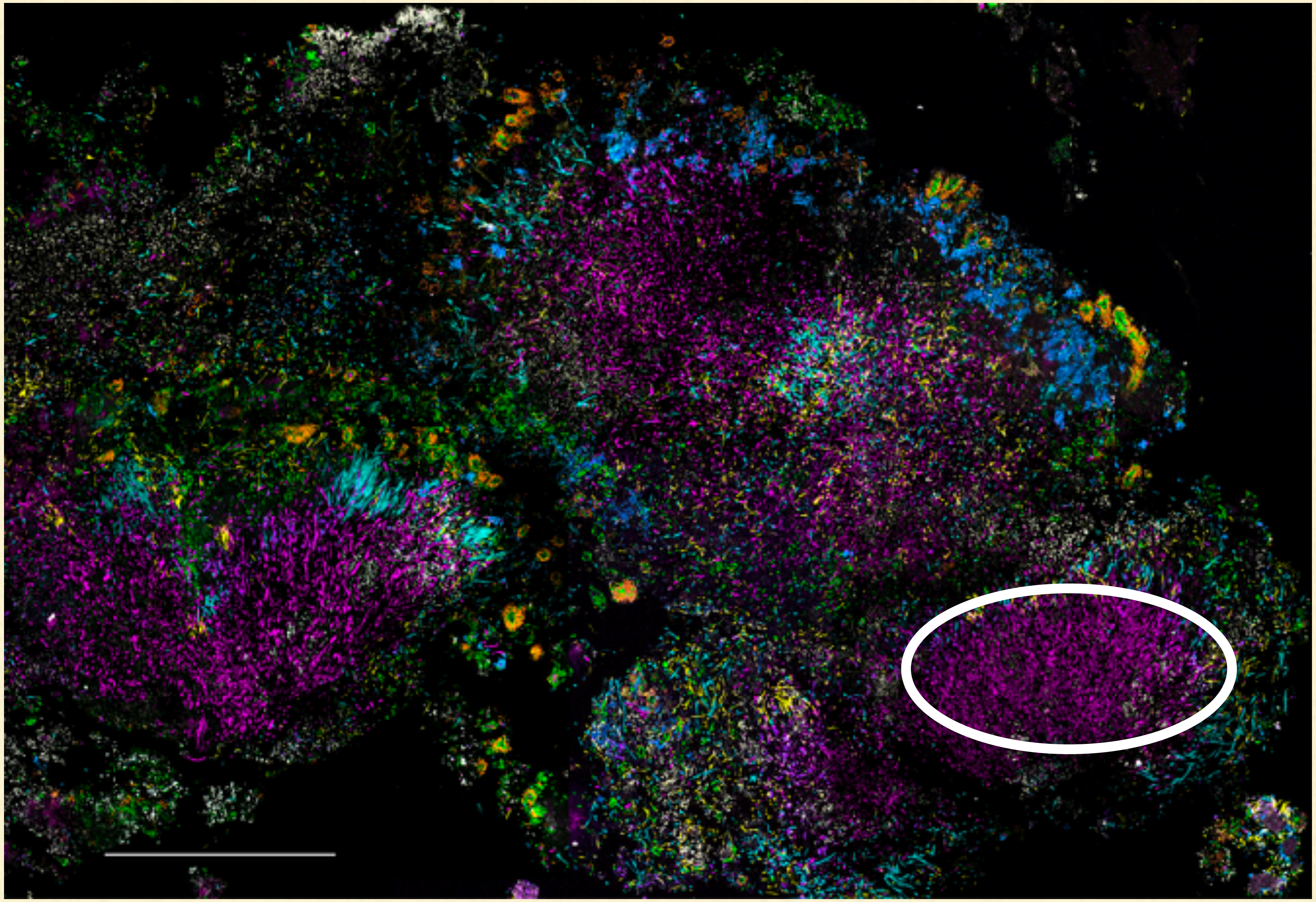
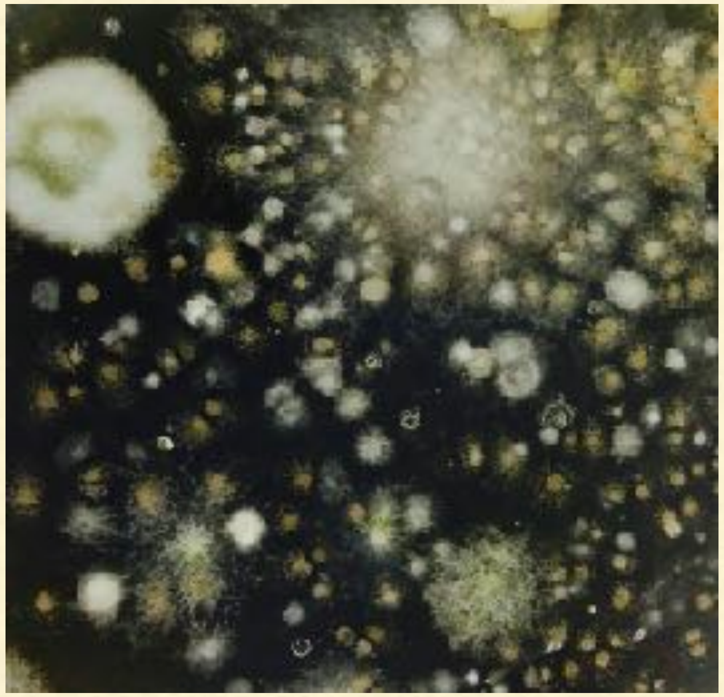


Photo credits: Mark Welch et al

INDEPENDENCE

- Cooccurrence of microbes, community dynamics, spatial structures all lead to non-independence

- Think about your ecology before deciding on a model



MULTINOMIAL MODEL

- Unfortunately the multinomial model has been used almost universally in microbiome data analysis
  - Subsampling
  - Rarefying
  - Bootstrapping
    - are all fancy ways of using getting "model-based standard errors" from the multinomial model

DECEPTION

- Reiterating, seemingly "nonparametric" approaches to variance estimates are highly parametric
- The variance estimates that you get drastically understate the true variance -- the variance if you repeated the experiment
  - This is why every signal appears significant in microbiome science

VARIANCE AND HYPOTHESIS TESTS

- Why is estimating variance important?
- Hypothesis testing
- Most hypothesis tests take the form

$$\frac{\text{estimate}}{\text{standard error}} \sim N(0,1)$$

VARIANCE AND HYPOTHESIS TESTS

- Wald test statistic

$$\frac{\text{estimate}}{\text{standard error}} \sim N(0,1)$$

- Suppose your variance is half what it should be
- What happens to p-values?

VARIANCE AND HYPOTHESIS TESTS

- If your estimate was 1, and the (true) standard deviation is 1…

STANDARD ERROR	1	0.5	0.33	0.25
P-VALUE	0.318	0.046	0.002	<0.001

P VALUES AND CONFIDENCE INTERVALS

- Common adage: don't quote p-values, give confidence intervals
  - (I actually agree with this)
  - BUT confidence intervals almost never overlap

REPLICATION WITH YOUR EXPERIMENT

- Amy's recommendations for how to deal with this
    - Use the most reasonable models you can
        - We will discuss in more detail on Thursday
    - Be skeptical
        - Don't be sucked in by flashy math or machine learning or methods
        - If you don't understand it, it may not make sense
    - Take biological replicates, and use them effectively

BIOLOGICAL REPLICATES

- Using biological replicates effectively
  - Validate your own findings before someone else can't

BIOLOGICAL REPLICATES

- Validation using biological replicates involves
  - Carefully considering the parameter you care about
  - Splitting your data into 2 sections
  - Constructing a confidence interval for your parameter using 1 section
  - Repeating with the second section
  - Confirming that your interval estimates are at least close

IN PRACTICE

- Example:
  - 12 patients, before/after antibiotics, strain-level diversity
  - Use 8 to estimate the difference in before/after, construct a confidence interval
  - Use the remaining 4 to construct another confidence interval
    - If they overlap, your model seems reasonable
    - If they don't, your model may not be reasonable

REALISTICALLY

- No, unfortunately this is not how papers get published
- Even if you use all 12 patients for your paper, as a responsible scientist you should be doing some sort of validation of your results

STATISTICAL MODELS

- Statisticians generally don't believe their models
  - Models are simple approximations to complex realities
  - They turn unsolvable problems into solvable ones
- Parameters can be useful descriptions of biology
  - e.g., diversity and relative abundance

3 BIG QUESTIONS

1. What is a reasonable model?

2. How do we estimate the parameters?

3. How reasonable are those estimates?

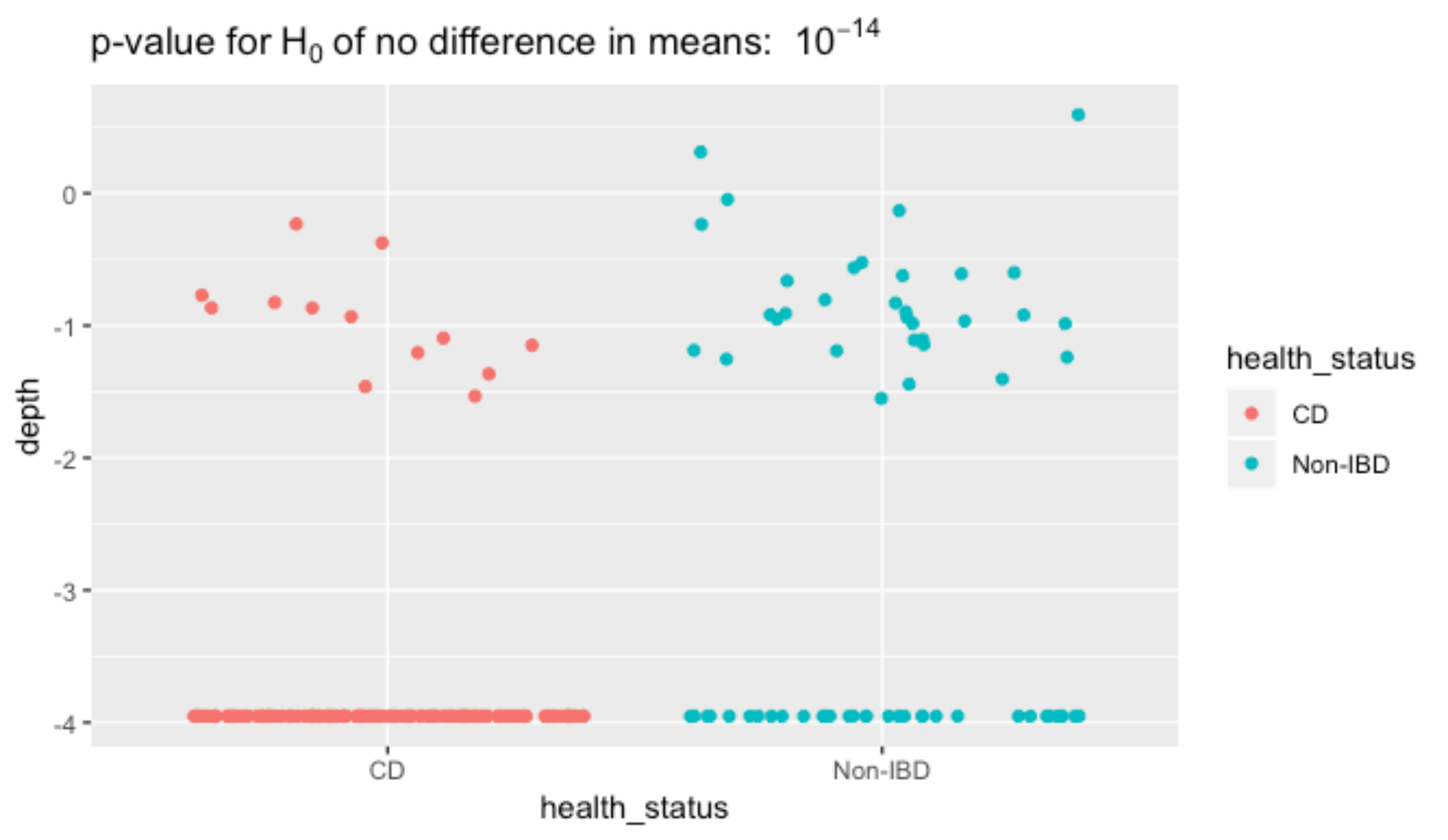
Today's lecture was intended to give you some high-level understanding of how to evaluate models and estimates

REPLICATES

- Technical replicates help you assess technical variation (**important!**), but are useless for assessing biological variation
  - Samples from different patients help you understand patient-to-patient variability
  - Samples from different sites help you understand within-patient variability
  - Samples from different instances of the same protocol help you understand within-protocol variability
- Technical variability is only one source of variability

REMINDERS ABOUT P VALUES

- If there is no accompanying plot, it's probably not interesting



REMINDERS ABOUT P VALUES

- Small p-values tell you about "statistical significance," and nothing about "biological significance"
  - Prediction is a totally different problem
    - which we currently have no ability to solve

WHAT CAN WE DO?

- Take replicates
  - Independently repeating the experiment is the gold standard for confirming the study is reproducible
    - independently = in a different lab
    - Dependence is induced by using the same lab
- Think critically
- Use plots, not p-values

A NOTE ON WORDS

- Efficient, optimal, uninformative, admissible, best, unbiased…
- These words have extremely precise meanings. Do not mislead your with statistical words!

WHAT ELSE CAN WE DO?

- Be honest
  - Keep all analyses that you ran, not just the final one
- Write down all of the hypotheses that you care about
  - Before doing the experiment
  - Before doing the analysis
- Your university might house a statistician; try to involve them...
  - ...in the entire process, not just calculating p-values

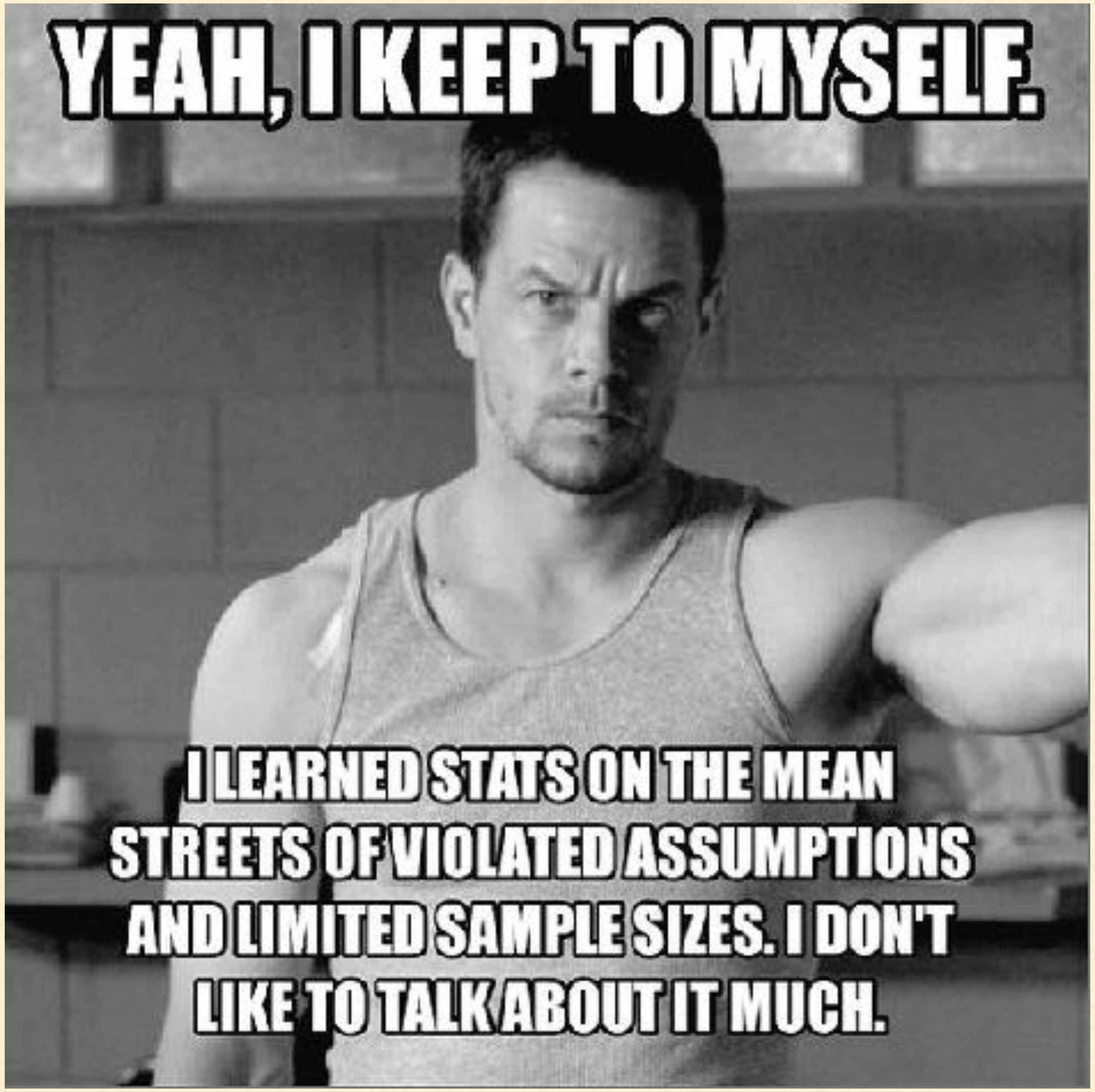
MANY THANKS

- Pauline Trinh and Bryan Martin TAs @ #STAMPS2018

  - For their help with this presentation!!!
- Tracy & Mihai
  - For all of their organisation of this wonderful workshop
- YOU!
  - For engaging in reproducible and ethical science

ACTIVITY & PRESENTATIONS

- Pick a microbiome paper where a sequencing experiment was performed to make a claim about the microbiology/ecology
  - Read the abstract/intro and write down what parameters the authors were interested in
  - Read the experimental design and data collection and write down whether this was reasonable and why
  - Were the tools/methods/claims convincing? Why/not?
- Explain the idea of the paper to the person next to you, and talk through your answers to the above questions.


YEAH, I KEEP TO MYSELF.
I LEARNED STATS ON THE MEAN STREETS OF VIOLATED ASSUMPTIONS AND LIMITED SAMPLE SIZES. I DON'T LIKE TO TALK ABOUT IT MUCH.

STATISTICAL THINKING

Research Group: Statistical Diversity Lab

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

@AmyDWillis adwillis@uw.edu