STATISTICS BOOTCAMP

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Associate Professor

Sarah Teichman — @sarah_teichman — PhD Candidate

STATISTICS

- Sticky note exercise
  - What is your #1 most pressing statistics question? (Pink)

LEARNING OBJECTIVES

1. Introduce key concepts from statistics in a microbial setting
    - population, parameter, estimate, model…
2. Discuss estimation and properties of estimators
    - bias & variance
3. Introduce hypothesis testing
4. Discuss model misspecification and its implications for estimation and testing
5. Learn about statistical transparency and ethics
6. Draw connections with your previous statistics exposure (if any)

STATISTICS

- ~~Two~~ Three different types of statistics
  - Exploratory statistics
    - What can you say about your data?
  - Inferential statistics
    - What can you say about the population your data represents?
  - Predictive modeling
    - What can you learn from your data to make predictions about future data?

RETURNING TO OUR STICKIES…

- Recall your sticky:
  - What is your #1 most pressing statistics question? (Pink)
- Can you guess if your question is about
  - Exploratory statistics? — What can you say about your data?
  - Inferential statistics? — What can you say about the population your data represents?
  - Predictive modeling? — What can you learn from your data to make predictions about future data?

EXPLORATORY STATISTICS

- What's going on with your data?
- How do we show what it says?
- How do we visualize our data?
- *Descriptive statistics*

EXPLORATORY STATISTICS

- What's going on with your data?
- How do we show what it says?
- How do we visualize our data?
- *Descriptive statistics*

How do you do exploratory statistics?

EXPLORATORY STATISTICS

- What's going on with your data?
- How do we show what it says?
- How do we visualize our data?
- *Descriptive statistics*

How do you do exploratory statistics?

However you want!

INFERENTIAL STATISTICS

- Inferential statistics
  - What can you say about the population your data represents?

POPULATION

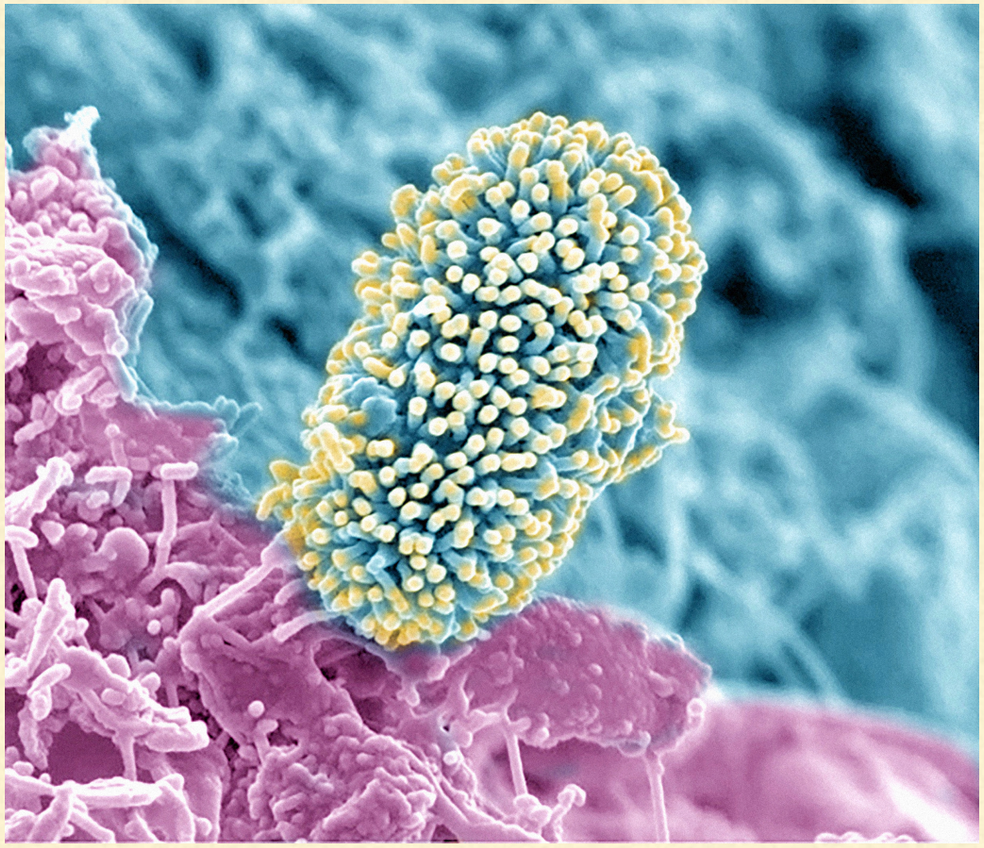
- Stat101

- "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

- A microbiome is a collection of microbes, and their genes and metabolites



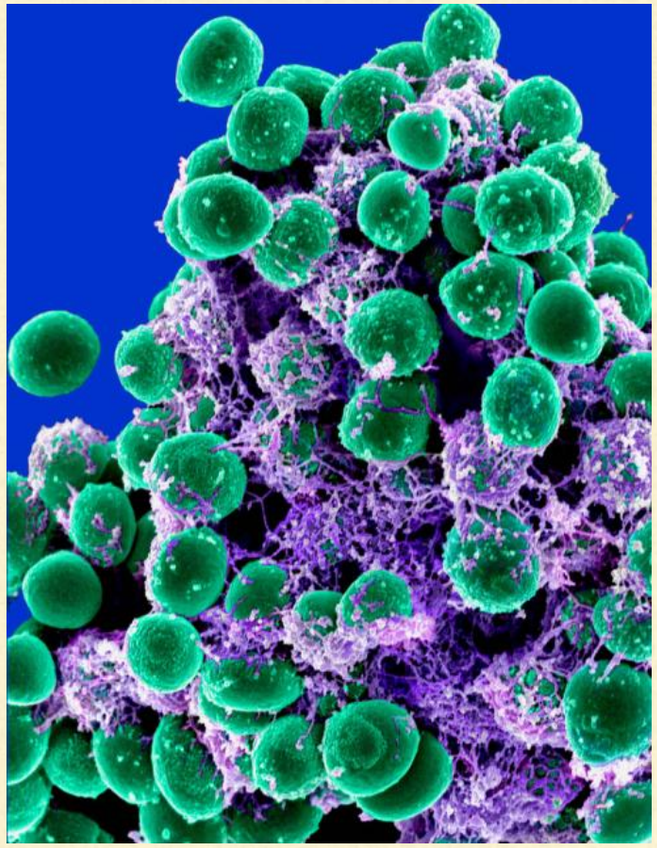


Photo credits: Pacific Northwest National Laboratory; NIAID

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 assay?

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

MICROBIAL POPULATIONS

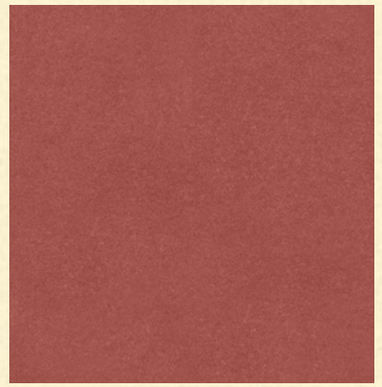
- The 4 W's: **Who/What? Where? When? Why?**
  - **Who? What?** …the poop of female 25-60 y.o.'s with a clinical IBD diagnosis, and with no IBD diagnosis…
  - **Where?** … who also live in your city & have access to healthcare…
  - **When?** … between January 2023-March 2023?
- Such observations can help us answer **why** certain patterns exist…
  - and why others 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: (5 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

- Inferential statistics: using information about the sample to infer something about the population

PARAMETERS

- Statisticians have a formal concept of "something about the population"
- Statistical *parameters* are a way to connect reality to your data
- *You* need to decide on a reasonable model for reality

PARAMETERS

- Example of a model:
  - There are microbes in your saliva, and they all have a taxonomic label at the genus level
- Example of a parameter:
  - The genus-level relative abundance of *Prevotella* in your saliva right now

PARAMETERS

- Example of a model:
  - Every #STAMPS2022 attendee carries a member of the *Prevotella* genus in their oral cavity, or they don't
- Example of a parameter of this model
  - The fraction of #STAMPS2022 attendees carrying *Prevotella* in their oral cavity in any abundance

PARAMETERS

- Other possible parameters of interest
  - The proportion of people in US hospitals carrying *S. aureus* that are methicillin-resistant
  - The mean phylum-level diversity of microbes on the hands of employees in the dining hall
- Who is the implicit population here? When?

PARAMETERS

- Other possible parameters of interest
  - The proportion of people in US hospitals carrying *S. aureus* that are methicillin-resistant
  - The mean phylum-level diversity of microbes on the hands of employees in the dining hall
- Who is the implicit population here? When?

PARAMETERS

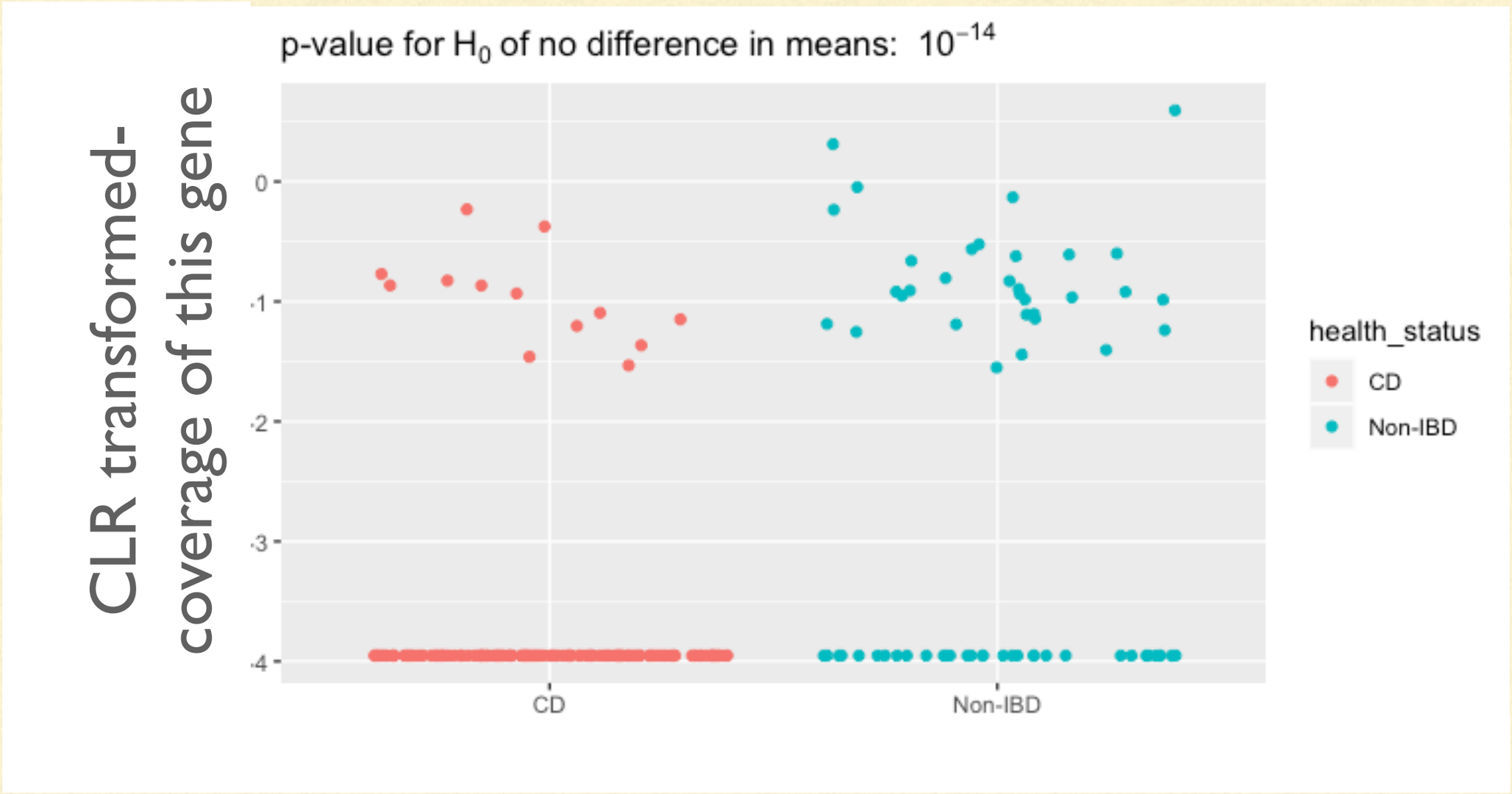
- Many common parameters are *averages*
  - Averages are summaries across a population
    - Inferential statistics: need a population
- Results about *means* say nothing about *individuals*!
  - Common misconception: "Taxon X was highly significantly differentially abundant when comparing [disease group] to [non-disease group], and is therefore a promising diagnostic…"

INFERENCE VS PREDICTION

- How exciting is a p-value of 10^{-14}?

INFERENCE VS PREDICTION

- How exciting is a p-value of 10^{-14}?



More on this later…

POPULATIONS VERSUS SAMPLES

- Key idea: *the sample is not the population*

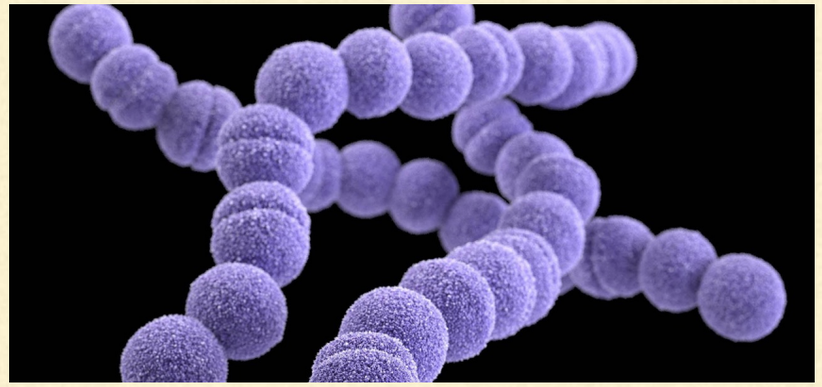
- Inferential statistics: using information about the sample to infer something about the population

  - Use the observed data to estimate the parameters

"INFORMATION ABOUT THE SAMPLE"

- Estimators (n, pl) estimate (v) parameters (n)
- "Estimate": the number calculated from your data
  - “An estimate of the phylum-level relative abundance of Bacteroidetes in my gut right now is 30%”
- "Estimator": a function of your data
  - “My estimator of relative abundance is plug-in 16S relative abundance…”

EXAMPLE



- *Estimate the genus-level relative abundance of Streptococcus in your saliva using 16S data*

- 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?

RELATIVE ABUNDANCE

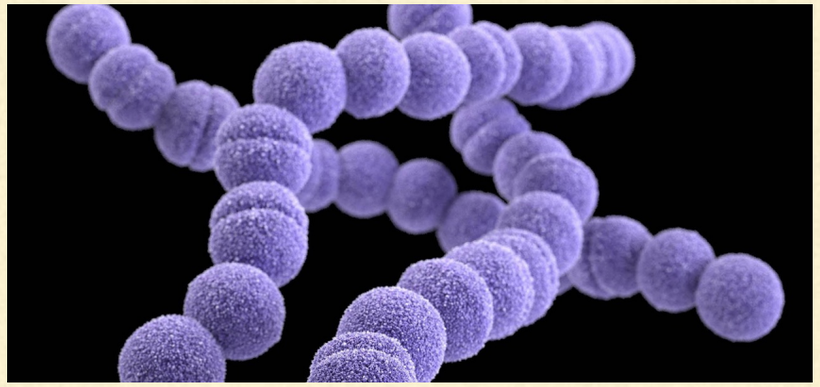
- Suppose…
  - n = samples, indexed by $i = 1, \ldots, n$
  - p_i = the relative abundance in each subject
  - W_i = # of observed sequenced copies from Strep
  - M_i = total # of sequenced copies
- Most common estimate of p_i is $\frac{W_i}{M_i}$

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…

EXAMPLE



- **Motivation**: Estimate the average 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?

COMPARING ESTIMATORS

- **Motivation**: "Estimate the mean genus-level relative abundance of Strep in a population"
  - W_i = # observed sequenced reads mapping to Strep in person i
  - M_i = total # reads sequenced from person i
- *Consider the following two estimators*
  - Take everyone's relative abundance $\left(\frac{W_1}{M_1}, \ldots, \frac{W_n}{M_n}\right)$, then average them
  - Add up everyone's Strep counts $W_{total} = W_1 + \ldots + W_n$, then add up everyone's total counts $M_{total} = M_1 + \ldots + M_n$, and divide the two: W_{total}/M_{total}

PARAMETERS

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

ESTIMATION: NOTATION

- The parameter *Amy*:

ESTIMATION: NOTATION

- An estimator of the parameter *Amy*:

BIAS

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

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

where

$$\mathbb{E}\hat{\theta} = \text{value of } \hat{\theta} \text{ on average}$$

- Your data is random, so $\hat{\theta}$ is random, and it has an average

BIAS

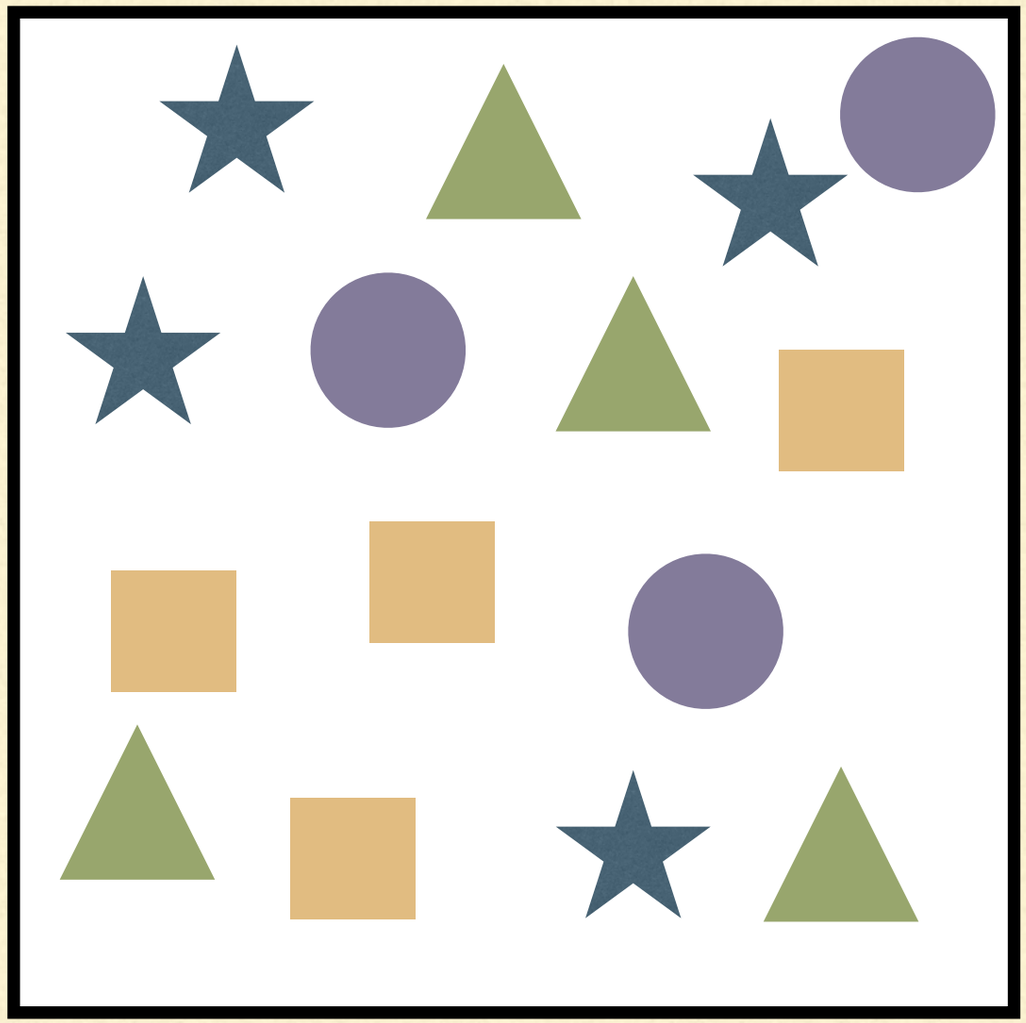
- An estimator of a parameter
  - is unbiased if
    - its bias is zero under the model
- The distribution of the estimator depends on the distribution of the data...
  - i.e., your model

BIAS

- **Data** isn't biased
- **Estimators** can have bias

EXAMPLE

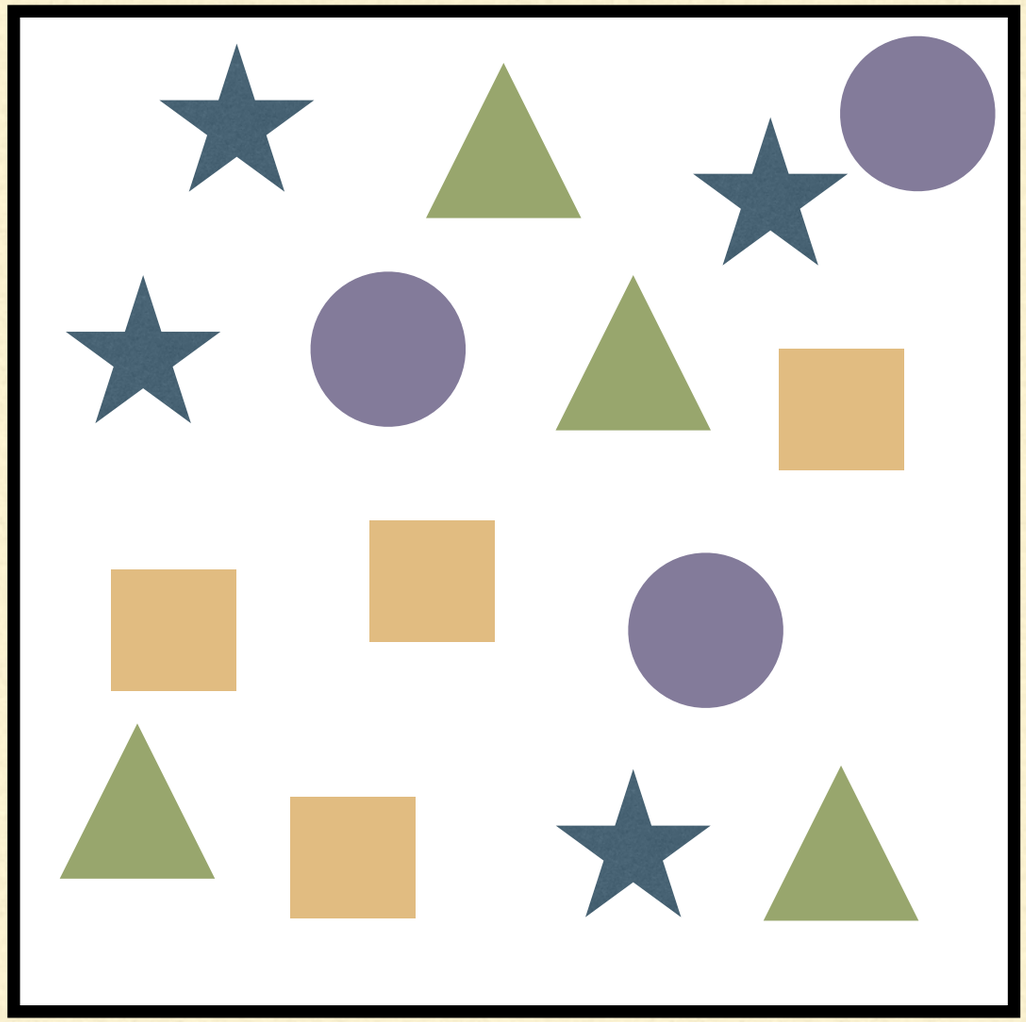
- What are the true relative abundances in the community?
- What's the (true) richness?
- Shannon diversity?



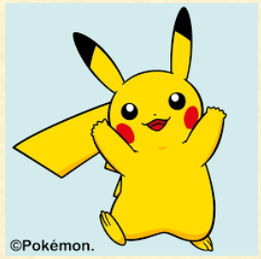
EXAMPLE

True abundances:

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

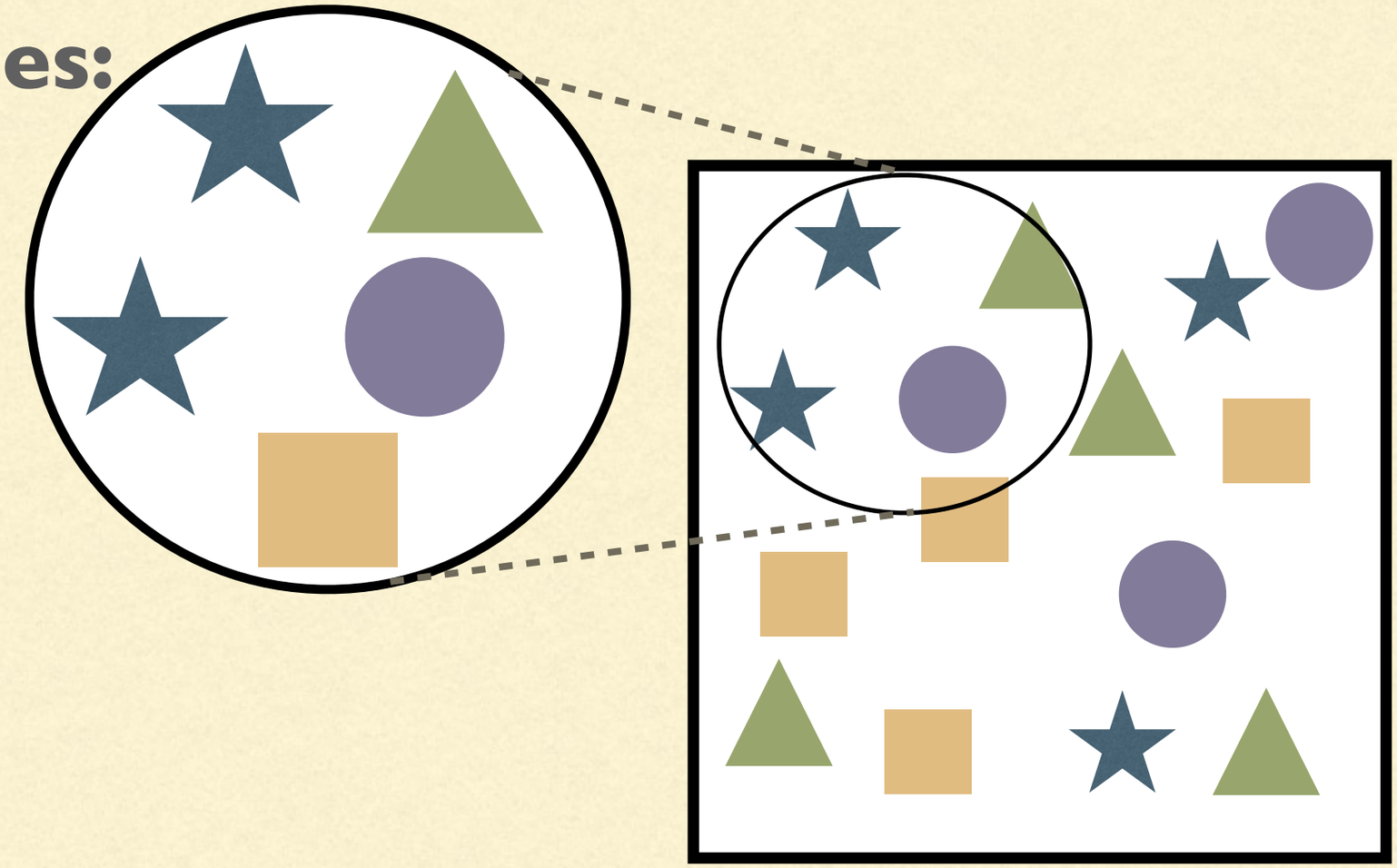


EXAMPLE



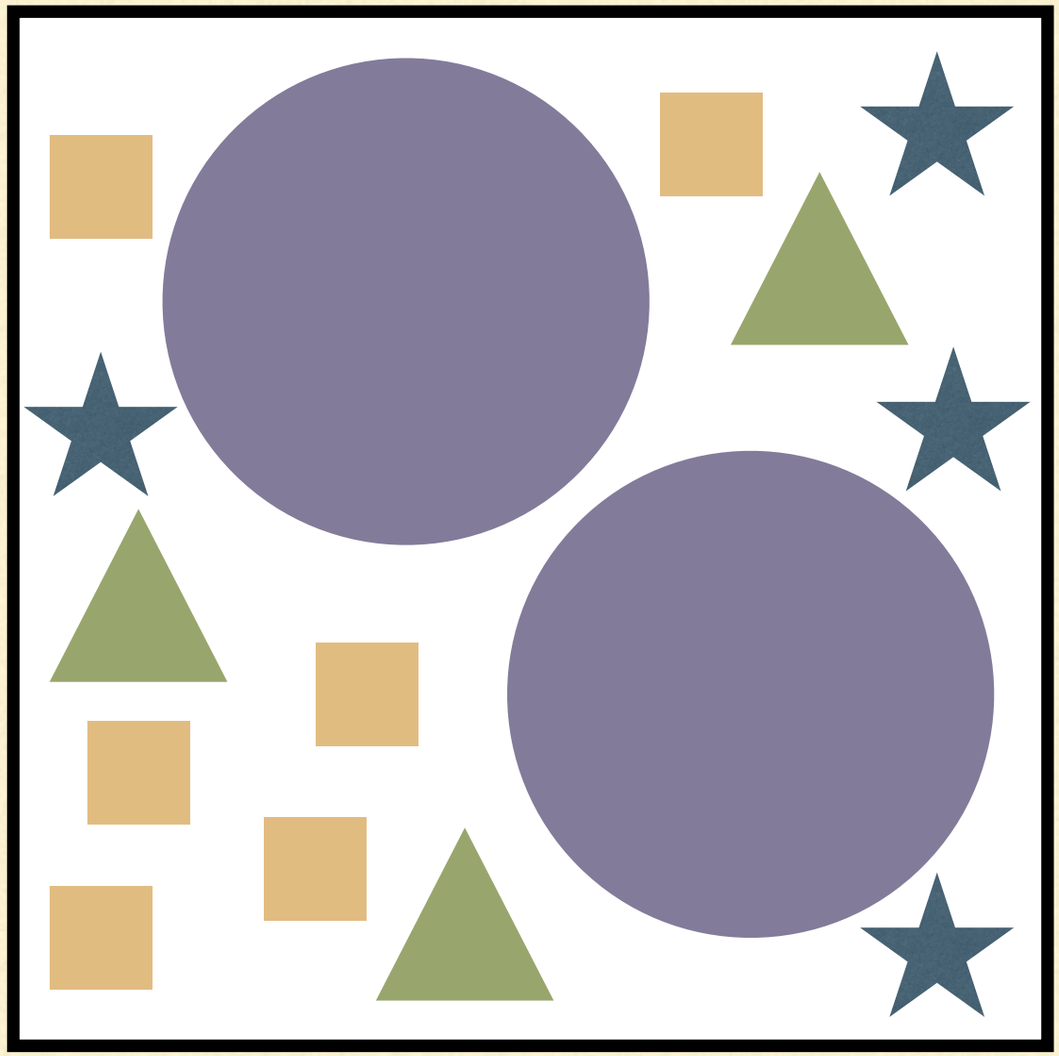
Observed abundances:

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



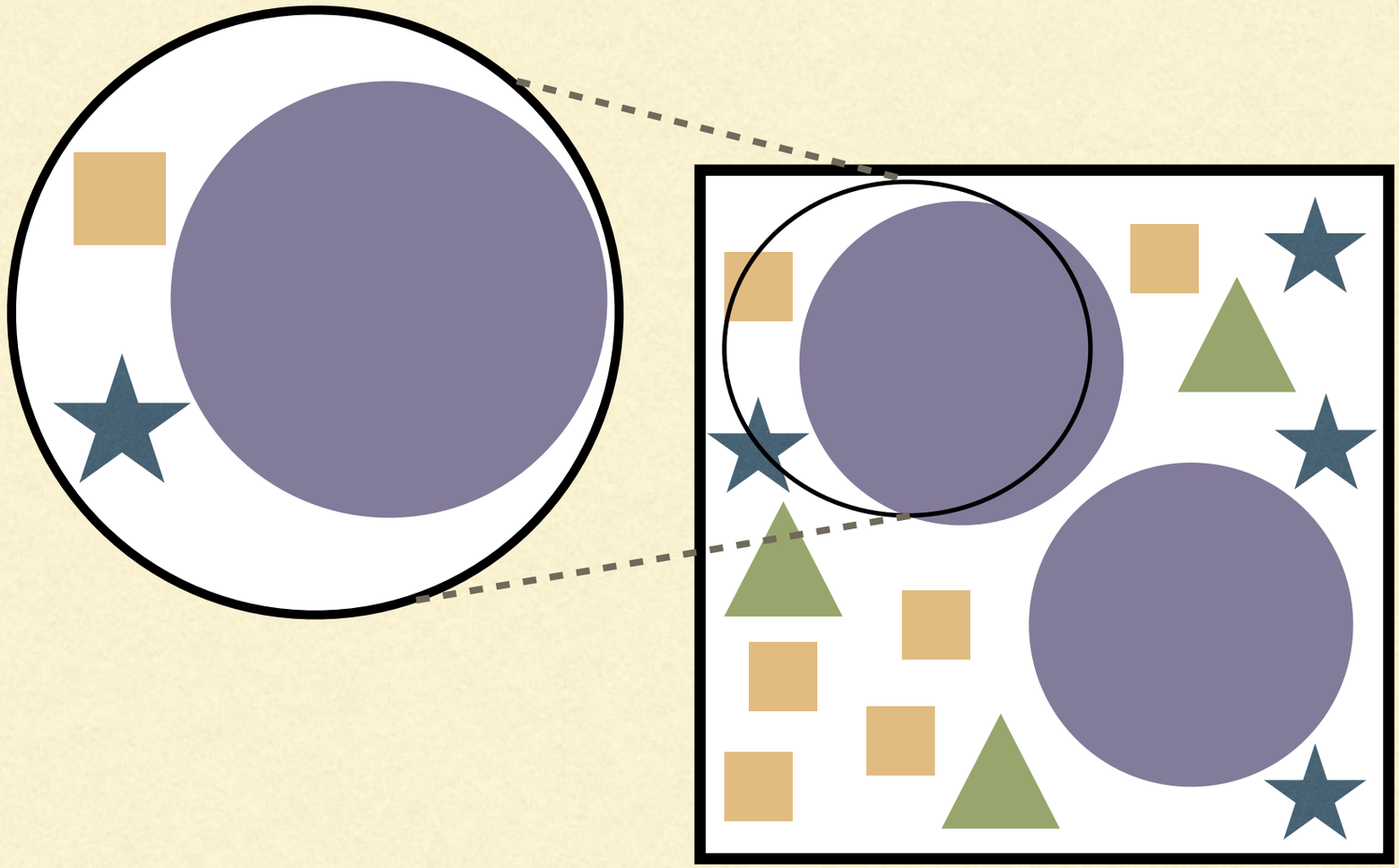
BIAS

- (4 minute)
  - What are the true relative abundances in the community?
  - Draw some nets. What are the observed relative abundances?



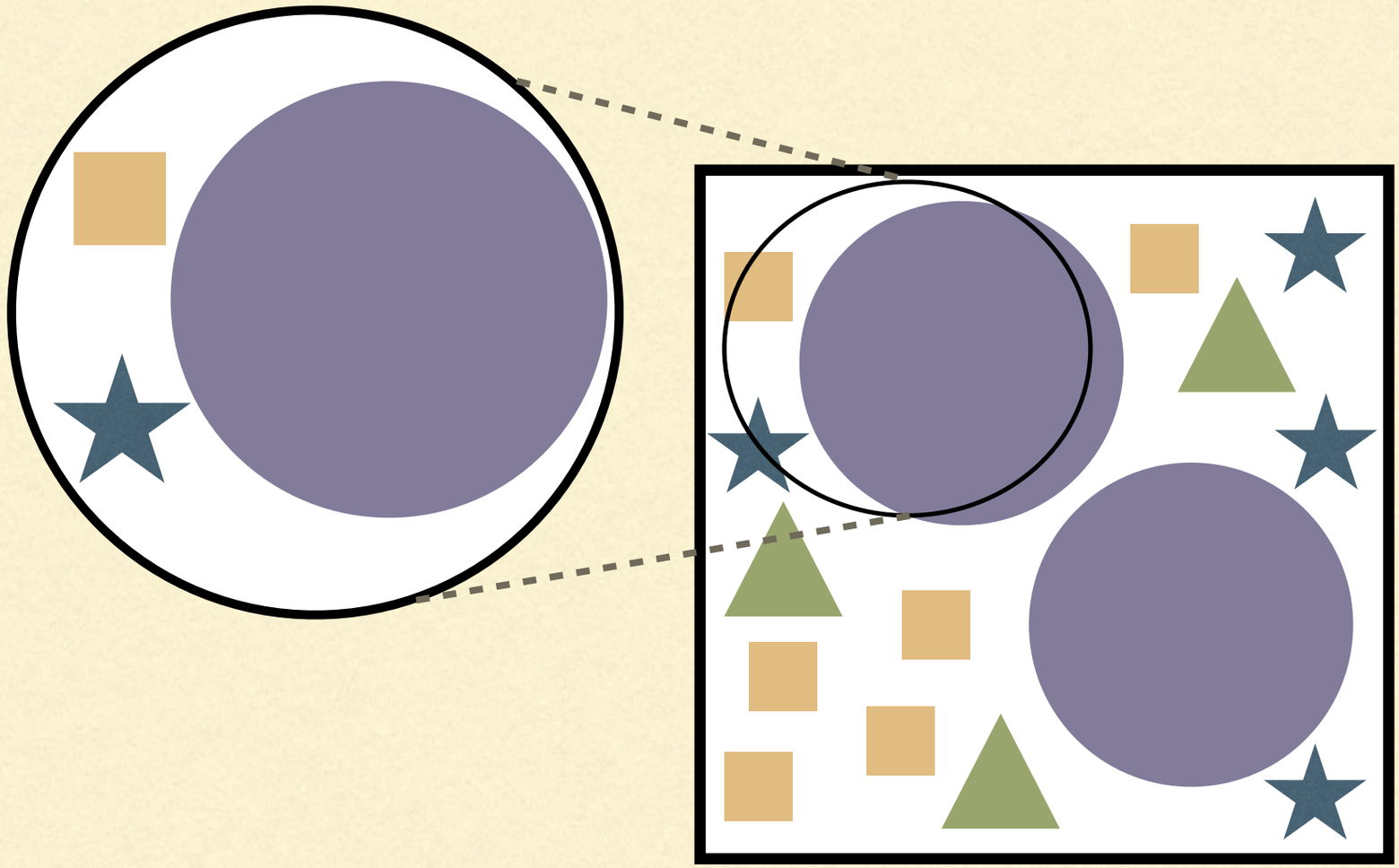
BIAS

- ★ = 1/3
- ● = 1/3
- ■ = 1/3



BIAS

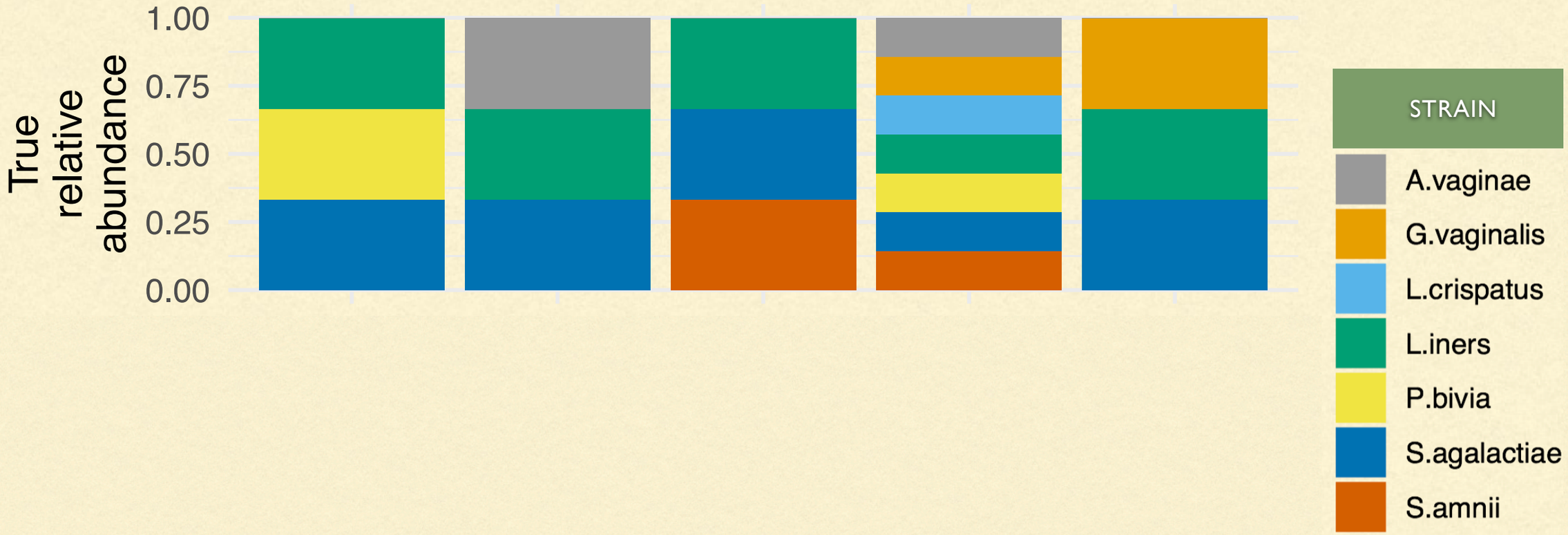
- ★ = 5/15
- ● = 5/15
- ■ = 5/15



Truth:

★ = 4/15 ● = 2/15 ▲ = 3/15 ■ = 6/15

BIAS AND RELATIVE ABUNDANCE

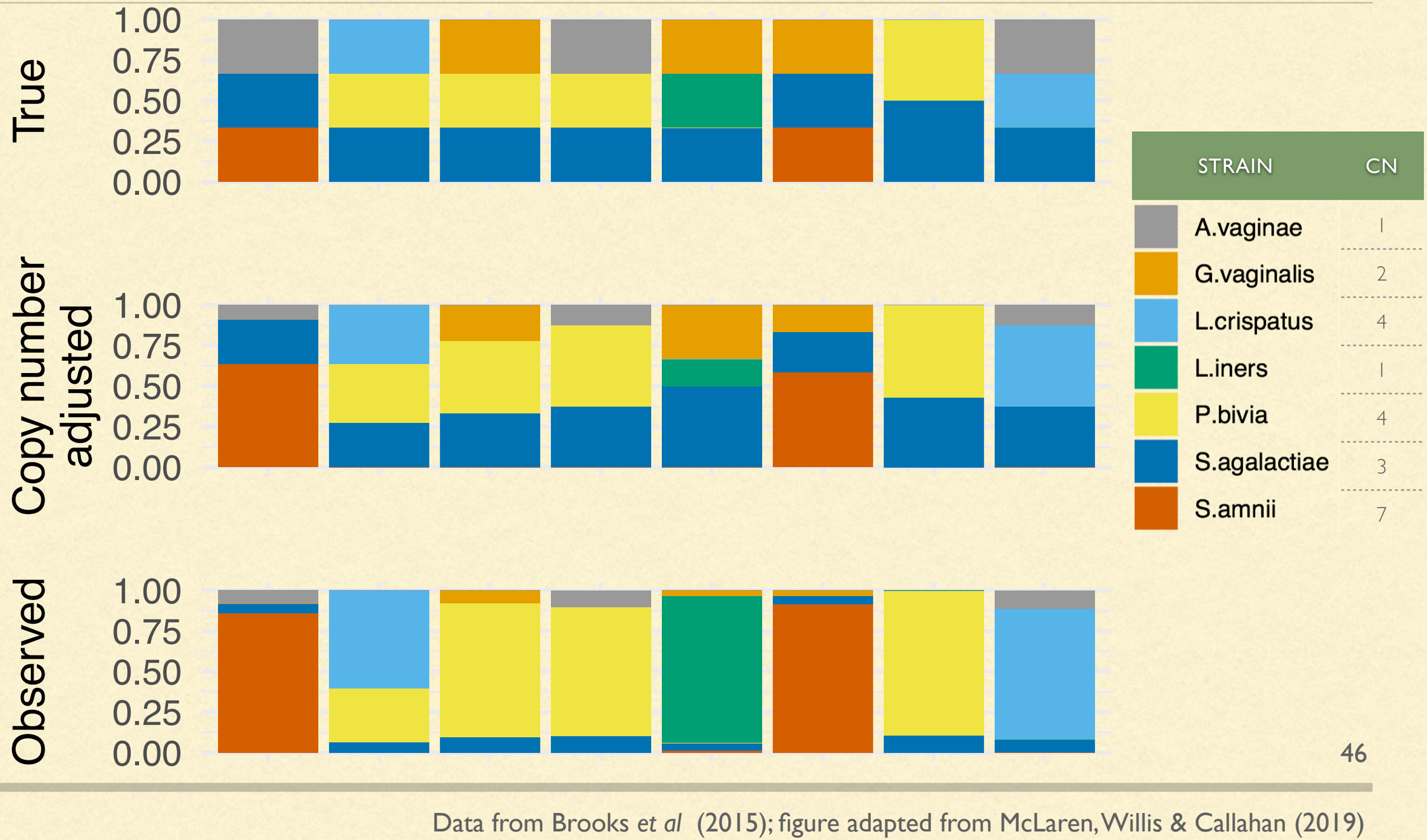


Data from Brooks *et al* (2015); figure adapted from McLaren, Willis & Callahan (2019)

BIAS AND RELATIVE ABUNDANCE

Data from Brooks *et al* (2015); figure adapted from McLaren, Willis & Callahan (2019)

BIAS AND RELATIVE ABUNDANCE



Data from Brooks *et al* (2015); figure adapted from McLaren, Willis & Callahan (2019)

BIAS AND RELATIVE ABUNDANCE

Taxon specific detection efficiencies

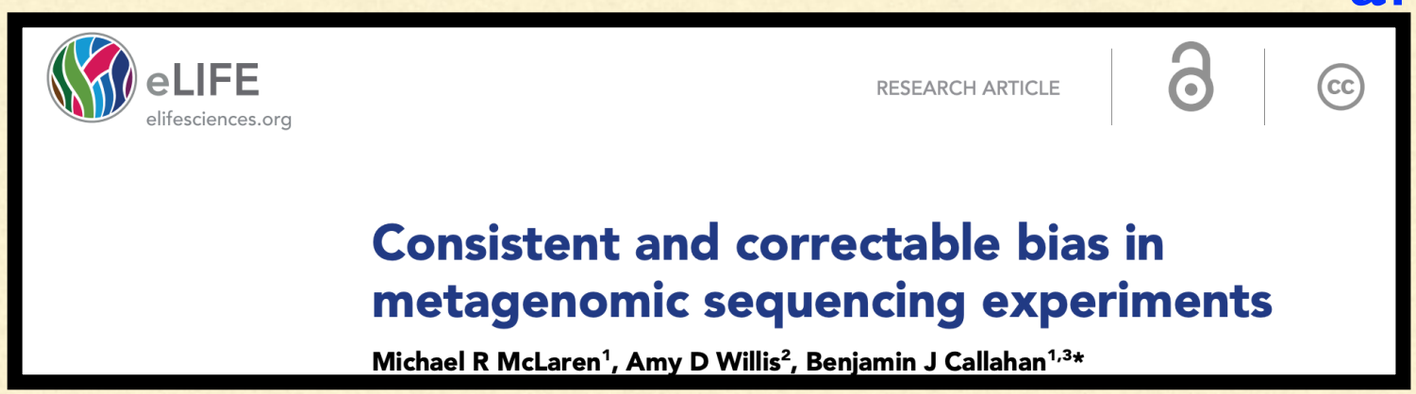
make

sample relative abundance

biased for

actual relative abundance

Ben will talk about this after lunch!!


eLIFE elifesciences.org

RESEARCH ARTICLE

Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren[1], Amy D Willis[2], Benjamin J Callahan[1,3]*

BIAS & DIVERSITY

- Sample species richness underestimates total richness
    - Rarefying *further* underestimates total richness

BIAS

- Bias = systematically wrong
- Data is not biased
- Some estimators are biased
- Studying bias requires a *model* and an *estimator*
- The "solution" depends on the problem!
  - We will continue to talk more about specific cases…

EVALUATING ESTIMATORS

- We want estimators to be
  - Accurate = correct on average = unbiased
  - Precise = usually close to its average = low variance
- Bias and variance are two criteria for evaluating estimators
  - e.g. Rarefying for diversity is low variance, high bias
  - e.g. Estimating diversity with breakaway and DivNet is higher variance, lower bias

VARIANCE

- Variance describes how much the estimates vary

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

- Variance actually isn't about the parameter
  - Some bad estimators have low variance

VARIANCE

- The variance reflects how far apart 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 0.2
  - 12, 12, 12, 13013, 12… => variance is 33 805 200
- A large change in the estimates equals a large variance
- Standard deviation = $\sqrt{\text{variance}}$

VARIANCE

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

- Every *random variable* has a variance
  - Estimators have variance
    - Estimators' variances are usually unknown - we need a $\hat{\text{Variance}}(\hat{\theta})$
    - We use $\sqrt{\hat{\text{Variance}}(\hat{\theta})}$ often in statistics — it's called the standard error
      - standard error = estimate of the standard deviation
- Observed outcomes have variance

SUMMARY SO FAR

- *Statistical thinking*
- Why we need statistical models
- Parameters of models
- Estimators of parameters
- Bias & variance as ways to evaluate estimators

THE PLAN

- Hypothesis testing

Now! ask us about multiple testing!

- Choosing a model

After!

ask us about confounders!
ask us about batch effects!

- Modeling with microbiome data
  - Abundance
    - 2 x lectures + 2x labs

This afternoon!

ask us about compositionality!
ask us about differential abundance!

  - Diversity:
    - Lecture + lab
- Experimental design

Tomorrow morning!

ask us about rarefaction!
ask us about diversity metrics!
ask us about ordination!
ask us about replicates!

- Questions — throughout!

INFERENCE

From Sarah

HYPOTHESIS TESTS

- You can now:
  - Describe your population and your sample
  - Define the parameter you are interested in
  - Construct and critique an estimator of your parameter of interest
- Let's put these ideas to answer your scientific questions!
  - Is the **true** value of your parameter of interest in your population of interest equal to a certain value? How much evidence do we have?

HYPOTHESIS TESTS

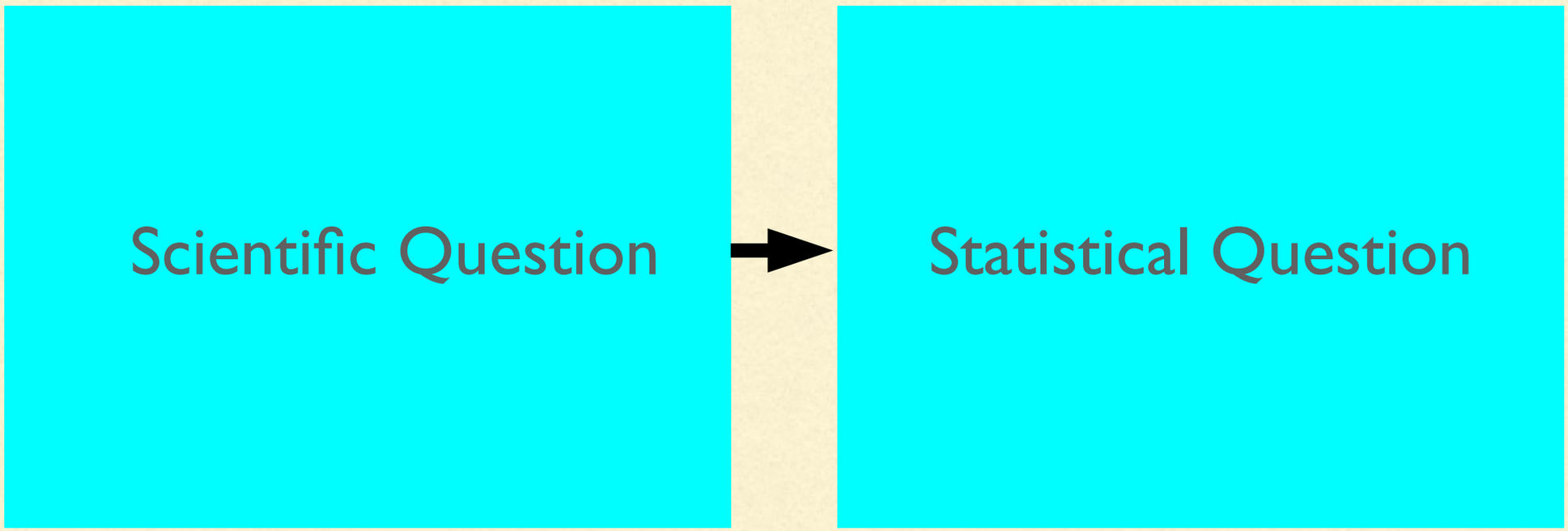
- What is a hypothesis test?

- Hypothesis: statement about a statistical parameter

- Test: way to ask "Do we have *enough* evidence to support our scientific claim?"

HYPOTHESIS

- First translate your scientific question into a statistical question



HYPOTHESIS

- First translate your scientific question into a statistical question

Does the presence of antibiotic resistance genes differ between chickens raised in farms versus chickens raised in households?

→

Statistical Question

HYPOTHESIS

- First translate your scientific question into a statistical question

Does the presence of antibiotic resistance genes differ between chickens raised in farms versus chickens raised in households?

→

Is the proportion of chickens that have genes conferring resistance against tetracycline different between chickens raised in farms versus households?

HYPOTHESIS

Ask yourself, do you have the data to estimate these parameters?

- First translate your scientific question into a statistical question

Does the presence of antibiotic resistance genes differ between chickens raised in farms versus chickens raised in households?

→

Is the proportion of chickens that have genes conferring resistance against tetracycline different between chickens raised in farms versus households?

HYPOTHESIS

- First translate your scientific question into a statistical question

Is *Cyanobacteria* differentially abundant between the Pacific and Atlantic oceans?

→

Your turn:
Statistical Question

HYPOTHESIS

- First translate your scientific question into a statistical question

Is *Cyanobacteria* differentially abundant between the Pacific and Atlantic oceans?

→

Is the mean cell concentration of *Cyanobacteria* in the Pacific different from the mean cell concentration of *Cyanobacteria* in the Atlantic?

HYPOTHESIS

Ask yourself, do you have the data to estimate these parameters?

- First translate your scientific question into a statistical question

Is *Cyanobacteria* differentially abundant between the Pacific and Atlantic oceans?

→

Is the mean cell concentration of *Cyanobacteria* in the Pacific different from the mean cell concentration of *Cyanobacteria* in the Atlantic?

HYPOTHESIS

- First translate your scientific question into a statistical question

Is *Cyanobacteria* differentially abundant between the Pacific and Atlantic oceans?

→

Is the relative abundance of *Cyanobacteria* in the Pacific different from the relative abundance of *Cyanobacteria* in the Atlantic?

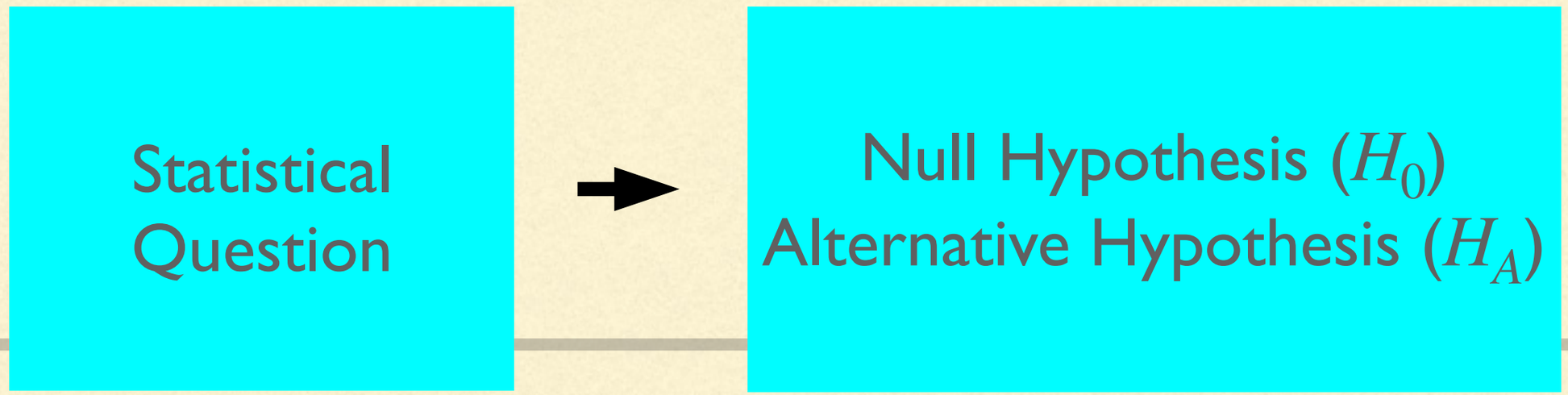
HYPOTHESIS

- Take two minutes to try to translate one of your scientific questions to a statistical question

Scientific Question → Statistical Question

HYPOTHESIS

- Your statistical question will inform your hypotheses
  - Null hypothesis (H_0): commonly accepted statement about parameter
  - Alternative hypothesis (H_A): scientifically interesting statement about parameter (usually), the opposite of the null hypothesis



HYPOTHESIS

- Your statistical question will inform your hypotheses
  - Null hypothesis (H_0): commonly accepted statement about parameter
  - Alternative hypothesis (H_A): scientifically interesting statement about parameter (usually), the opposite of the null hypothesis

Is the mean cell concentration of *Cyanobacteria* in the Pacific different from the mean cell concentration of *Cyanobacteria* in the Atlantic?

→

H_0: the mean cell concentration is the same for both groups
H_A: the mean cell concentration is different between the two groups

HYPOTHESIS

- Two possible conclusions from a hypothesis test:

1. Reject the null hypothesis.
   - your data is so extreme under the null hypothesis that it seems unlikely that it is true
2. Fail to reject the null hypothesis.
   - maybe the null hypothesis is true
   - maybe you just don't have sufficient evidence to reject it

HYPOTHESIS

- Two possible conclusions from a hypothesis test:

1. Reject the null hypothesis.

We have enough evidence to reject the hypothesis that mean cell concentration is the same in the Pacific and Atlantic oceans

2. Fail to reject the null hypothesis.

We do not have enough evidence to reject the hypothesis that mean cell concentration is the same in the Pacific and Atlantic oceans.

QUESTIONS

- Now we know what null and alternative hypotheses are, we'll soon move on to *testing*
- We're about to take a break
- Questions before we break?

BREAK

TESTING

- How do you know if you have enough evidence to reject the null hypothesis?

1. Calculate a test statistic from your sample

2. Ask how likely it would be to observe this test statistic if the null hypothesis were true in your population

TESTING

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

TESTING

- *Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

*(Wald) test statistic

TESTING

D = Mean concentration Pacific - Mean concentration Atlantic

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

0

Var(D)

TESTING

Larger difference: more evidence against H_0

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

TESTING

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

Larger difference: more evidence against H_0

Smaller value: more certainty about estimate

TESTING

- Test statistic often takes this form:

Larger difference: more evidence against H_0

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

Smaller value: more certainty about estimate

So, large t means large difference between estimate and null hypothesized value and/or high certainty in estimate

TESTING

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

Larger difference: more evidence against H_0

Smaller value: more certainty about estimate

And small t means small difference between estimate and hypothesized value and/or low certainty about estimate

TESTING

Larger difference: more evidence against H_0

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

Smaller value: more certainty about estimate

Suppose your standard error is half what it should be.
What happens to your test statistic?

TESTING

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

Larger difference: more evidence against H_0

Smaller value: more certainty about estimate

Your test statistic is twice what it should be!

TESTING

- Test statistic often takes this form:

$$t = \frac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$$

Larger difference: more evidence against H_0

Smaller value: more certainty about estimate

You've mistakenly doubled your evidence against H_0!

TESTING

- How do you know if you have enough evidence to reject the null hypothesis?

1. Calculate a test statistic from your sample ✅

2. Ask how likely it would be to observe this test statistic if the null hypothesis were true in your population

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

Probability,

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

Probability, when the null hypothesis is true,

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|\overset{\bigstar}{T}| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

Probability, when the null hypothesis is true, that we would calculate a test statistic from our data

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \overset{\bigstar}{\geq} |t| \,\middle|\, H_0 \text{ true} \right)$$

Probability, when the null hypothesis is true, that we would calculate a test statistic from our data as or more extreme

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \geq |\bigstar t| \,\middle|\, H_0 \text{ true} \right)$$

Probability, when the null hypothesis is true, that we would calculate a test statistic from our data as or more extreme as the one in this sample.

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

To calculate this probability, we need to specify a distribution for T under the null hypothesis

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

Often, we get to say that $T \sim N(0,1)$

TESTING

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true
- Formally,

$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

Often, we get to say that $T \sim N(0,1)$

Why? The Central Limit Theorem!

TESTING

A p-value!

- Ask how likely it would be to observe this test statistic (or a test statistic more extreme) if the null hypothesis were true

- Formally,

$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

Often, we get to say that $T \sim N(0,1)$

Why? The Central Limit Theorem!

P-VALUE

- A p-value tells us how extreme our results are in a world in which our null hypothesis is true

P-VALUE

p-value = 0.02

The probability that we would observe a difference in mean cell concentration of *Cyanobacteria* between the Pacific and the Atlantic ocean as or more extreme than the difference in our sample, if there is truly no difference in our population, is 2%.

Therefore, we reject our null hypothesis.

ALPHA LEVEL

- When can we reject the null hypothesis?
  - The alpha level (α) of a test is our threshold
  - We reject the null hypothesis when the p-value is less than our alpha level

ALPHA LEVEL

- How do we choose a good alpha level?
  - It depends!
  - Recall, a p-value tells us how unlikely our results are in a world in which our null hypothesis is true
  - 0.01? 0.05? 0.20?

ALPHA LEVEL

- How do we choose a good alpha level?
  - It depends!
  - Recall, a p-value tells us how unlikely our results are in a world in which our null hypothesis is true
  - 0.01? 0.05? 0.20?

A usual choice for the alpha level is 0.05

VALID HYPOTHESIS TEST

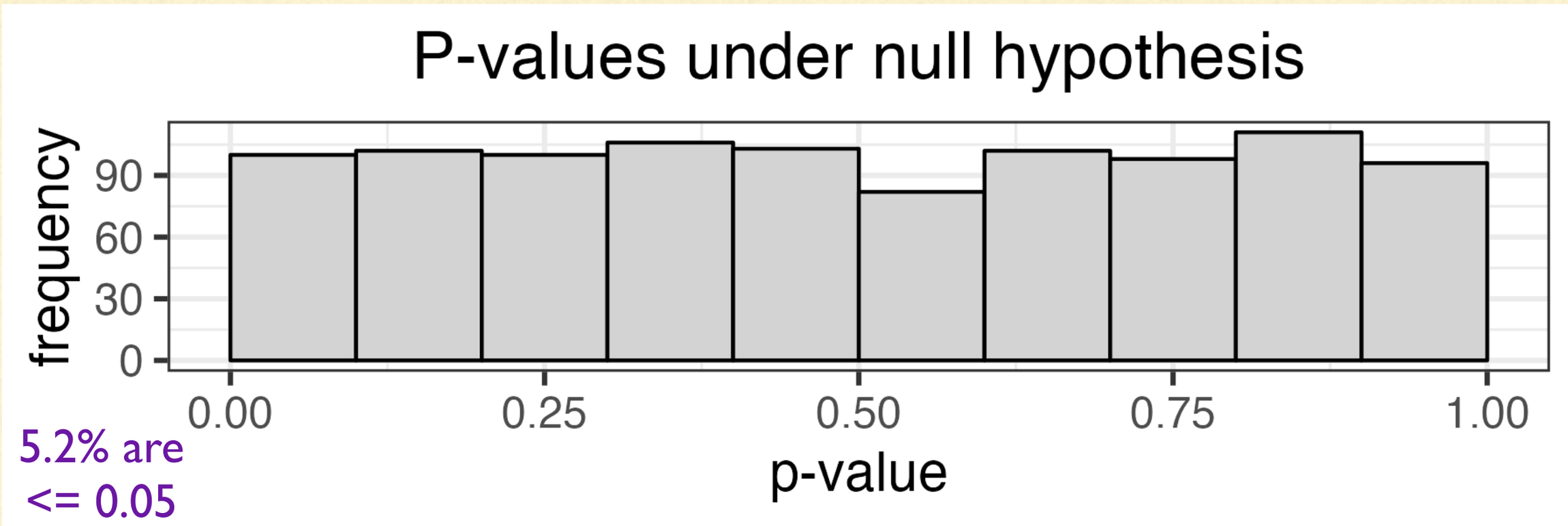
- A valid hypothesis test will reject the null hypothesis exactly $\alpha \times 100\,\%$ of the time **when it is true**
  - Less often = understating your evidence against H_0
    - your test is conservative
  - More often = overstating your evidence against H_0
    - your test is anticonservative

VALID HYPOTHESIS TEST

- Let's say that we sample data from a population for which **the null hypothesis is true** 1000 times
- For each sample, we calculate a test statistic and a p-value
- What should the distribution of our p-values look like?

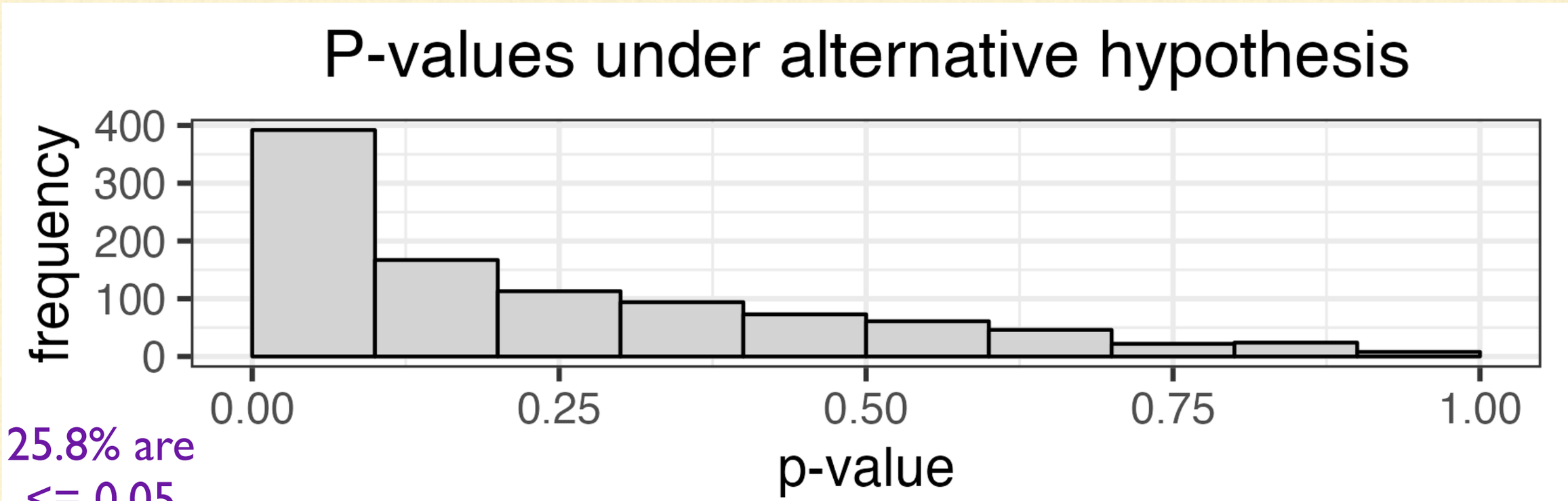
VALID HYPOTHESIS TEST

- Let's say that we sample data from a population for which **the null hypothesis is true** 1000 times
- For each sample, we calculate a test statistic and a p-value



VALID HYPOTHESIS TEST

- Let's say that we sample data from a population for which **the null hypothesis is false** 1000 times
- For each sample, we calculate a test statistic and a p-value



VALID HYPOTHESIS TEST

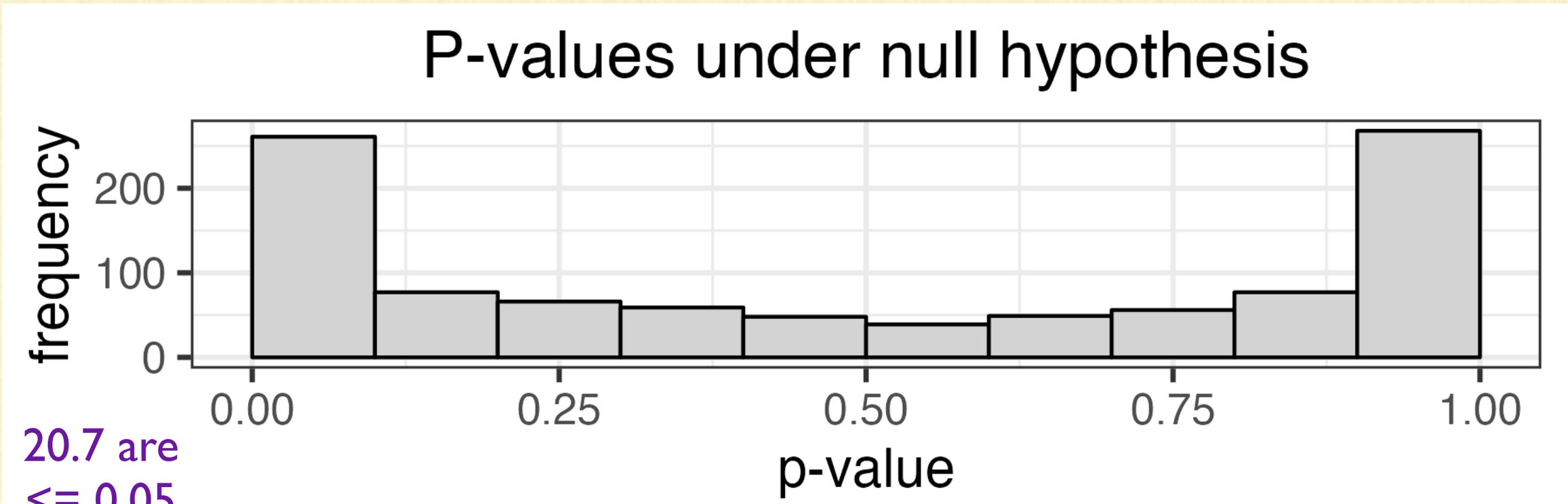
- Why might a hypothesis test be invalid?
  - Hint: recall our discussion of standard errors!

MODEL MISSPECIFICATION

- p-values require estimates of the variance (standard error)
- estimates of the variance require models
- wrong model → wrong variance → wrong p-value!

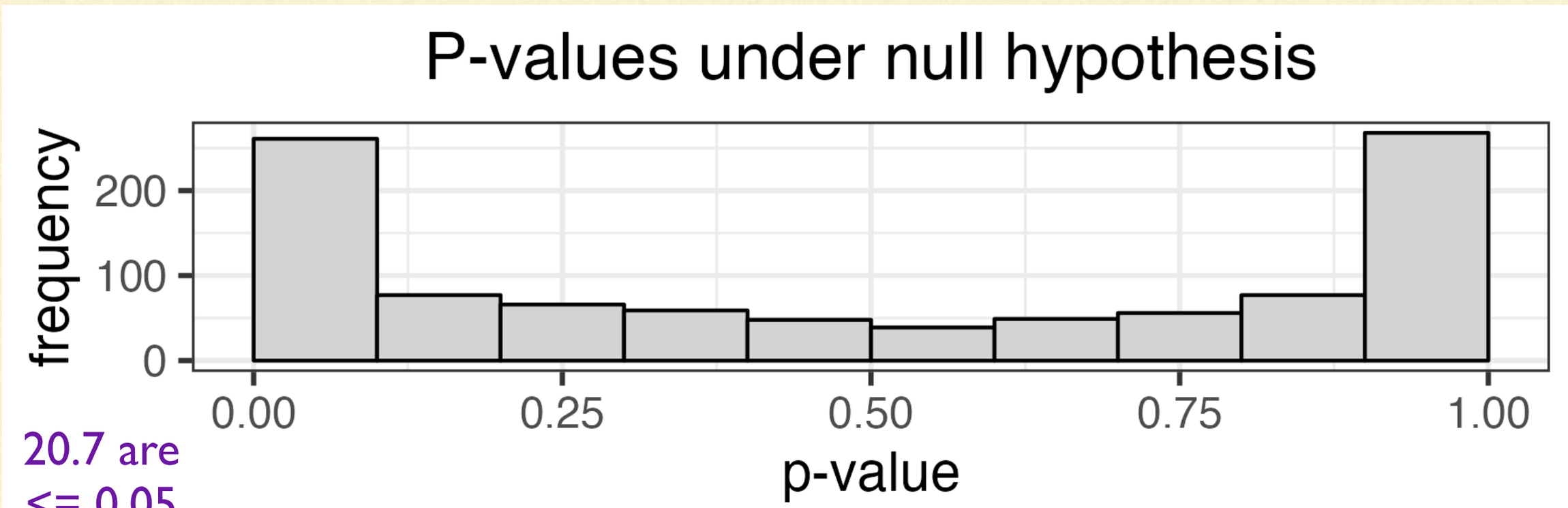
MODEL MISSPECIFICATION

- What if we have the wrong model, causing us to mistakenly estimate a standard error that is 1/2 the correct standard error?



MODEL MISSPECIFICATION

- This test is no longer valid! It will reject the null hypothesis more than $\alpha \times 100\,\%$ of the time.



TYPE I ERROR

- Type I error = rejecting the null, **when it is true**
  - "False positive"
- For a valid test, $Pr(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$

MULTIPLE TESTING

- **Setting:** Your colleague is conducting a microbiome-wide association study (MWAS) to understand the microbiome's relationship with colorectal cancer. They run 1000 tests to look for differentially abundant taxa. They find that at an alpha level of 0.05, 50 taxa are associated with cancer. They publish the following:

"50 new taxa confirmed to be associated with colorectal cancer!"

- What's wrong with this headline? How would you report these results?

MULTIPLE TESTING

- 2 independent tests:
  - Probability you don't reject H_0 for Test 1 $= .95$
  - Probability you don't reject H_0 for Test 2 $= .95$
  - Probability you don't reject H_0 for both tests $= .95 \times .95 = .9025$
- Probability you make at least one type 1 error: ~10%

MULTIPLE TESTING

- 3 independent tests:
  - Don't reject H_0 for all tests $= .95 \times .95 \times .95 = .8574$
- Probability you make at least one type I error: ~14%

MULTIPLE TESTING

m	Probability of at least one Type I error
10	0.40
100	0.994
1,000	~1

MULTIPLE TESTING

- So, what can we do when we need multiple tests?
- Instead of controlling Type I error rate separately for each test, consider:
  - **Family-wise Error Rate (FWER):** probability of at least one type I error
  - **False Discovery Rate (FDR):** the expected proportion of type I errors among the rejected hypotheses

MULTIPLE TESTING

- Instead of controlling Type I error rate separately for each test, consider:
  - **Family-wise Error Rate (FWER):** probability of at least one type I error
    - Use Bonferroni correction, divide α by number of tests, use this as threshold for rejecting null hypothesis
  - **False Discovery Rate (FDR):** the expected proportion of type I errors among the rejected hypotheses
    - Can use q-values instead of p-values

MULTIPLE TESTING

- q-values
  - Adjusted p-values to control FDR instead of Type I error rate
- In their MWAS study, your colleague found one species with a p-value of 0.00005 and a q-value of 0.03
  - p-value: the probability they would see a test statistic as extreme as the one observed for a non-differentially abundant species is 0.00005
  - q-value: 3% of the species that were tested and had test statistics even more extreme than the one observed would be false positives

MULTIPLE TESTING

- There are a number of other methods to avoid issues with multiple testing

- **BUT your best bet is limiting formal testing to primary hypotheses**

TYPE 2 ERROR & POWER

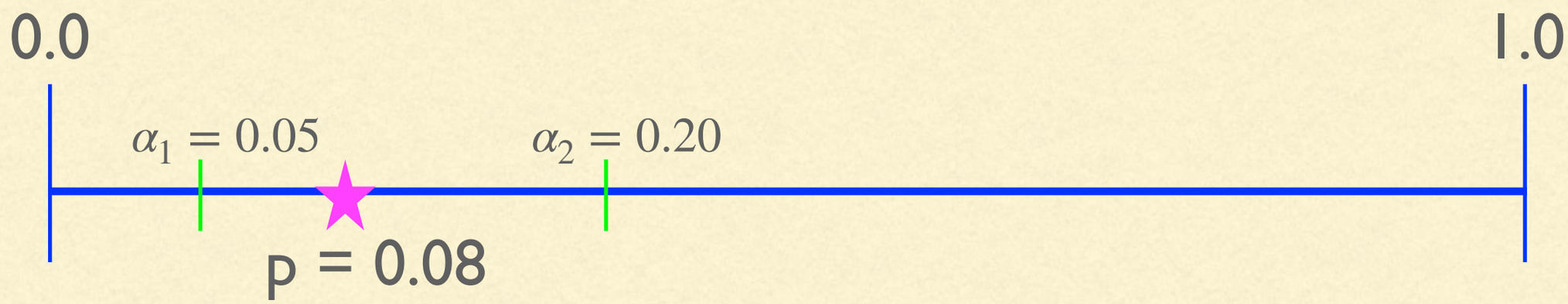
- Our alpha level specifies the probability of a Type 1 error
- We can also commit Type 2 errors: "false negatives"
- Power: probability of correctly rejecting the null hypothesis, **when it is false**
  - $Pr(\text{reject } H_0 \mid H_0 \text{ is false})$
  - 1 - probability of type 2 error

TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level (increasing Type I error rate)
- **Exercise**: Why?

TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level (increasing Type I error rate)
- **Exercise**: Why?

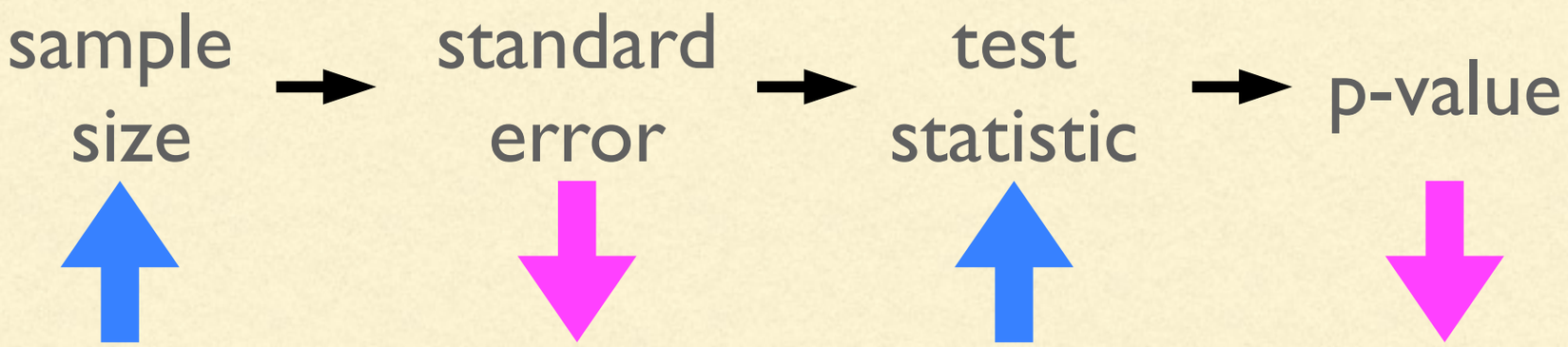


TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level (increasing Type I error rate)
  - **Exercise**: Why?
- We can increase power *without sacrificing Type I error* by increasing our sample size (if we have the $$$)

TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level (increasing Type I error rate)
  - **Exercise**: Why?
- We can increase power *without sacrificing Type I error* by increasing our sample size (if we have the $ $$)



QUESTIONS?

- Questions before we move on to modeling?

REGRESSION MODELS

- **Inferential statistics:** What can you say about the population your data represents?
  - Parameters are summaries of populations
    - Common parameters arise in *regression models*

REGRESSION MODELS

- Regression models take the form

  functional of outcome variable = function of predictor variables

- e.g.,

- expected diversity$_i = \beta_0 + \beta_1 \times \mathbf{1}_i$ is from lakewater (not seawater)}

  - $\hat{\beta}_0$ is an estimate of the expected (mean) diversity in seawater environments

  - $\hat{\beta}_1$ is an estimate of the difference in expected diversity in lake vs seawater environments

REGRESSION MODELS

- Regression models take the form

  functional of outcome variable = function of predictor variables

- e.g.,

  - $\text{expected}\ \text{diversity}_i = \beta_0 + \beta_1 \times \mathbf{1}_i$ is from lakewater (not seawater)

    - $\hat{\beta}_0$ is an estimate of the expected (mean) diversity in seawater environments

    - $\hat{\beta}_1$ is an estimate of the difference in expected diversity in lake vs seawater environments

REGRESSION MODELS

- Example of regression model: 🌽 corncob 🌽

$$\text{expected counts}_{ij} = M_i p_{ij}$$

$$\text{logit}\left(p_{ij}\right) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{i1} + \ldots + \beta_{pj}X_{ip}$$

- $\hat{\beta}_{kj}$ is an estimate of the difference in the logit-transformed relative abundance of taxon j between environments that differ by 1 unit in $X_{\cdot k}$ but are alike in $X_{\cdot 1}, \ldots, X_{\cdot k-1}, X_{\cdot k+1}, \ldots, X_{\cdot p}$

REGRESSION MODELS

- Another example of a regression model: DESeq2

$$\text{expected counts}_{ij} = s_i p_{ij}$$

$$\log_2\left(p_{ij}\right) = \beta_{0j} + \beta_{1j}X_{i1} + \ldots + \beta_{pj}X_{ip}$$

- $2^{\hat{\beta}_{kj}}$ is an estimate of the multiplicative difference in the relative abundance of taxon j between environments that differ by 1 unit in $X_{\cdot k}$ but are alike in $X_{\cdot 1}, \ldots, X_{\cdot k-1}, X_{\cdot k+1}, \ldots, X_{\cdot p}$

REGRESSION MODELS

- Another example: PERMANOVA

$$\text{Centroid for sample } i \text{ using distance } d$$

$$= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

- Difficult to interpret the β's
- Very commonly used despite limited possible insights

TYPES OF VARIABLES

functional of outcome variable = function of predictor variables

- Outcome variable
  - Choose something you actually care about
  - Ok to defy conventions
- Functional
  - means, rates, true underlying proportions…
  - Stick to conventions

TYPES OF VARIABLES

functional of outcome variable = function of predictor variables

- There are different types of predictor variables

1. Predictor of interest

   - The main thing you set out to study

   - Always include

TYPES OF VARIABLES

2. Confounders

- Associated with predictor of interest
- Causally associated with outcome
- Not in causal pathway of interest

- e.g., In estimating the causal effect of smoking on lung function in teenagers, age is a confounder

- Need a causal model (& ideally some more training/collaboration)

TYPES OF VARIABLES

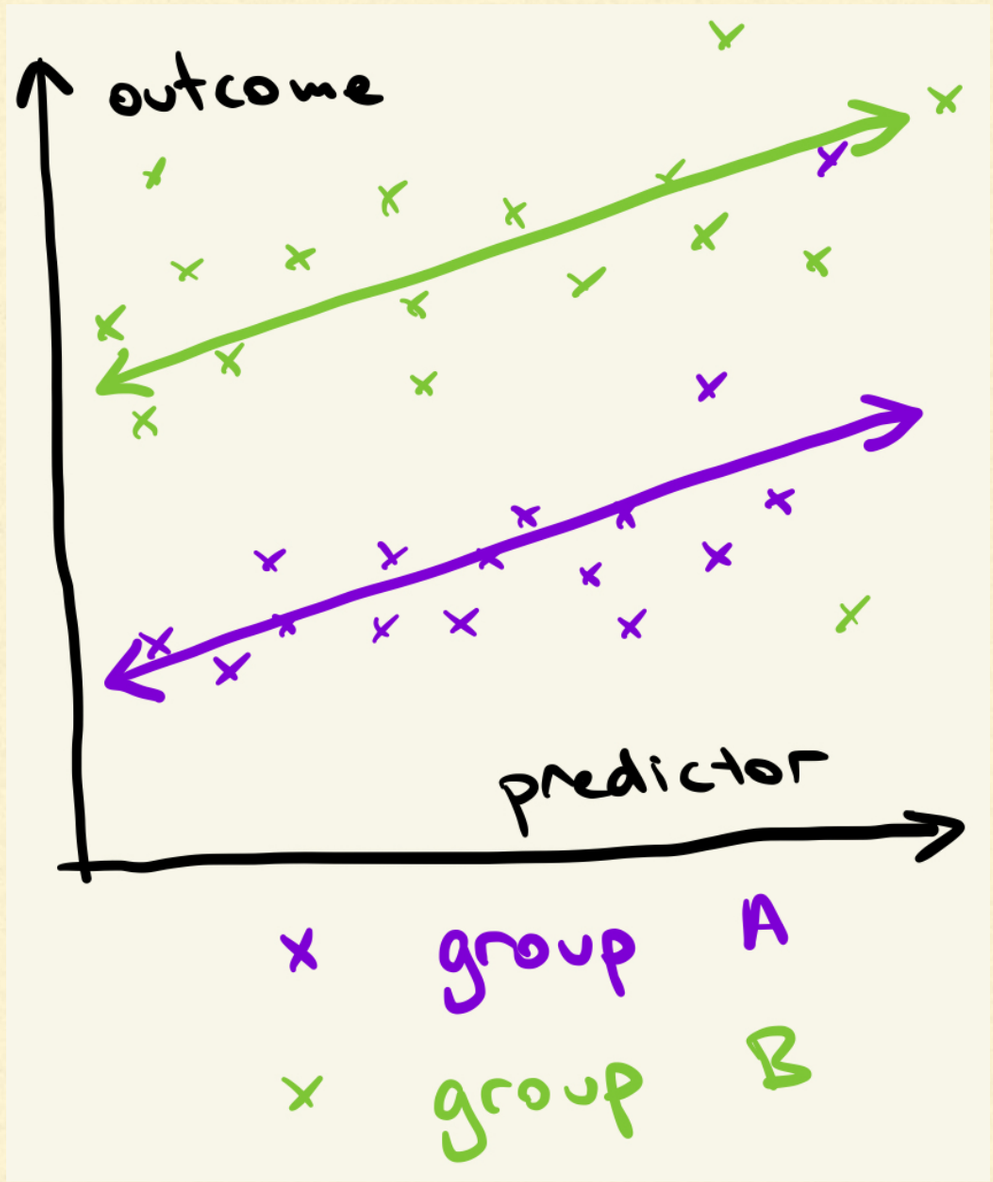
2. Confounders

- None of the following are confounders
  - Batch
  - Sequencing technology
  - Any measurement variables
- Variables associated with the measurement process *cannot* be causally associated with outcome
- "Confounders" is more often misused than correctly used

TYPES OF VARIABLES

3. Precision variables

- Associated with response
  - Not associated with predictor of interest
- Helps to improve precision
- e.g., batch effects, tank effects
- e.g. in human microbiome: age, sex…
- Often capture "technical variation"



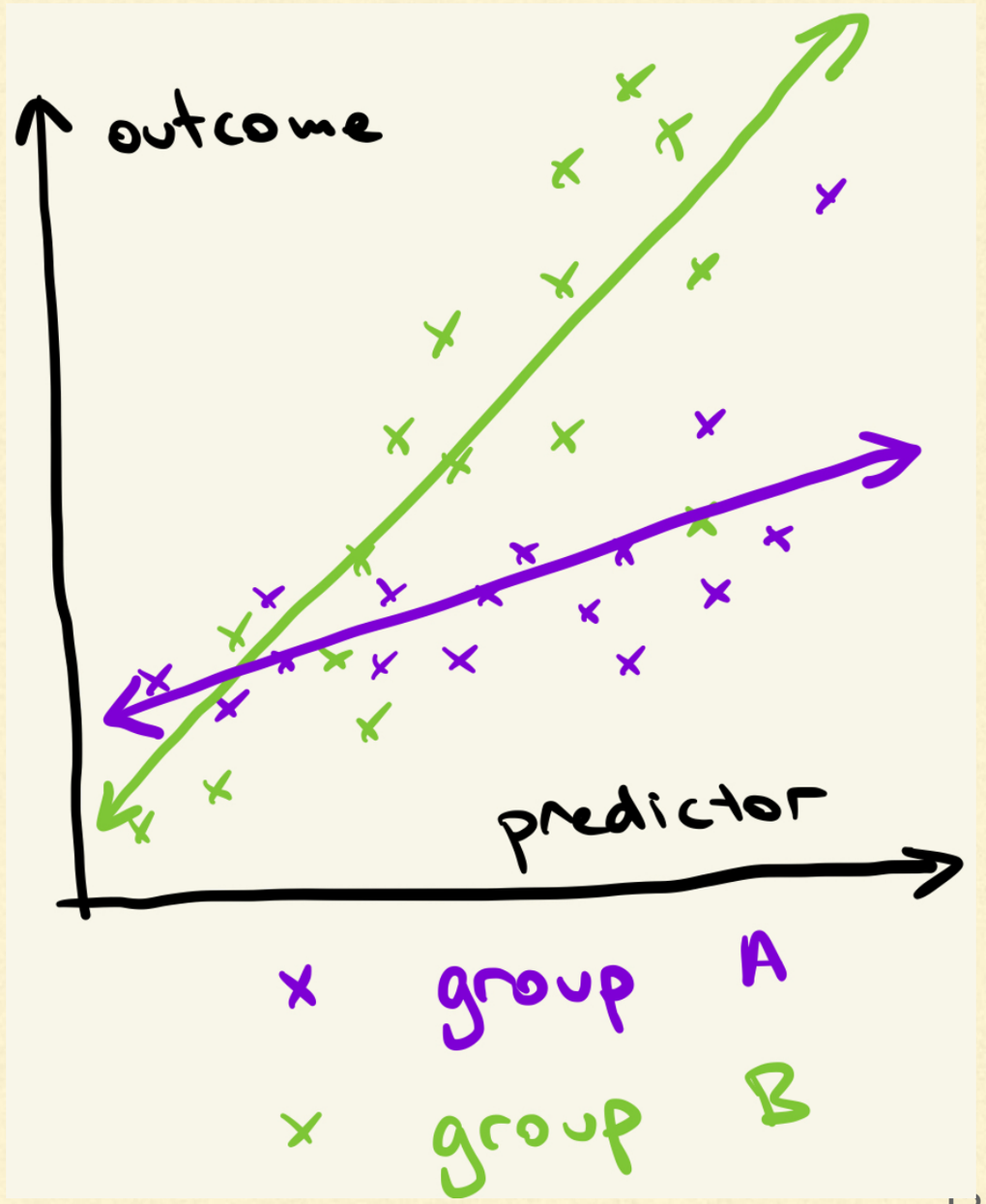
TYPES OF VARIABLES

4. Effect modifiers

- Association b/w response & predictor of interest differs for different values of an effect modifier

- "interaction" between variables

- Sometimes, effect modification is the predictor of interest



TYPES OF VARIABLES

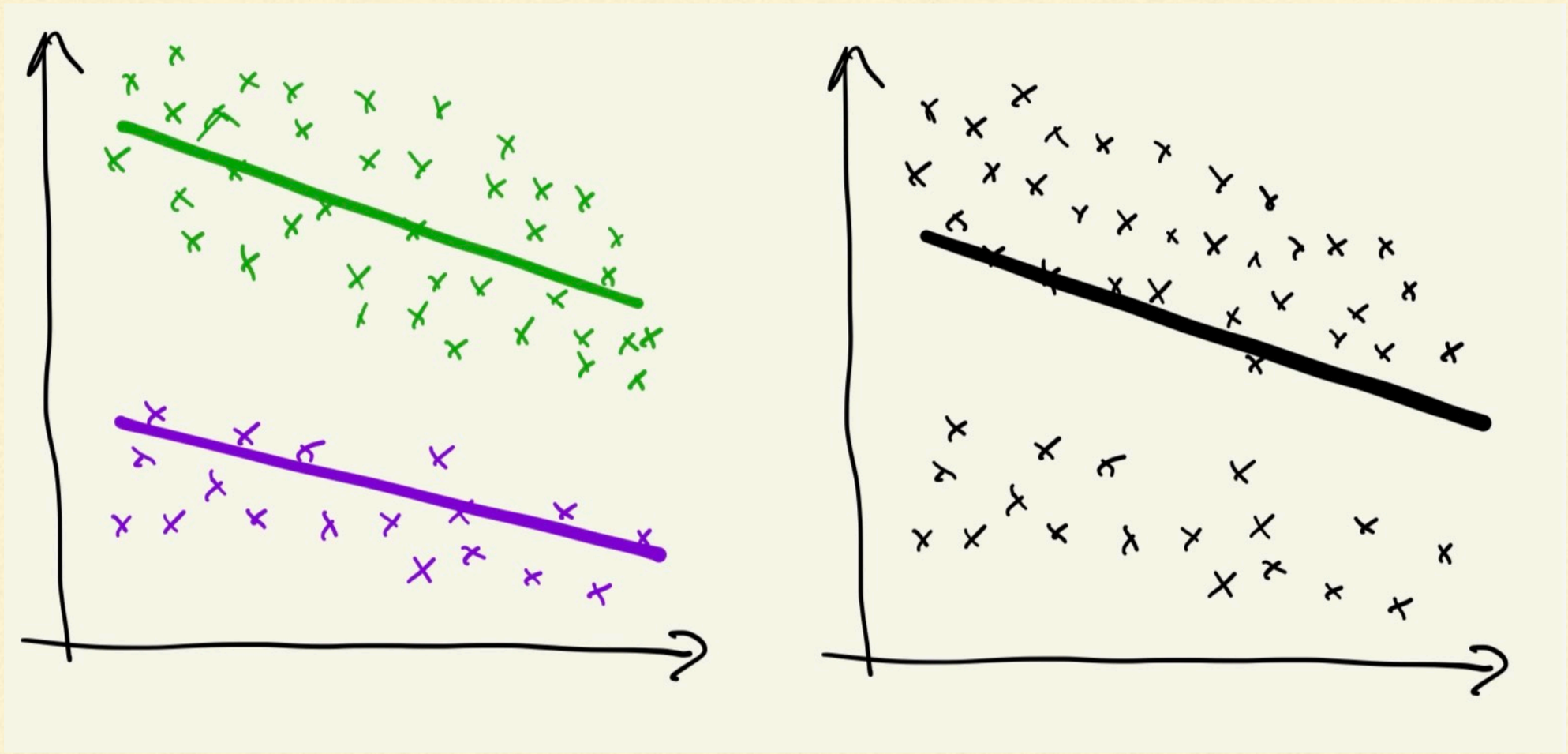
- Precision variables and effect modifiers
  - There almost always will be many unmeasured or unmeasurable precision variables
  - There almost always will be many unmeasured or unmeasurable effect modifiers
  - This is *fine!* You don't need to include all PVs and EMs in your model!

WHAT HAPPENS WHEN WE OMIT VARIABLES?

- Unmodeled precision variables and effect modifiers get "averaged over"
  - This is model misspecification?

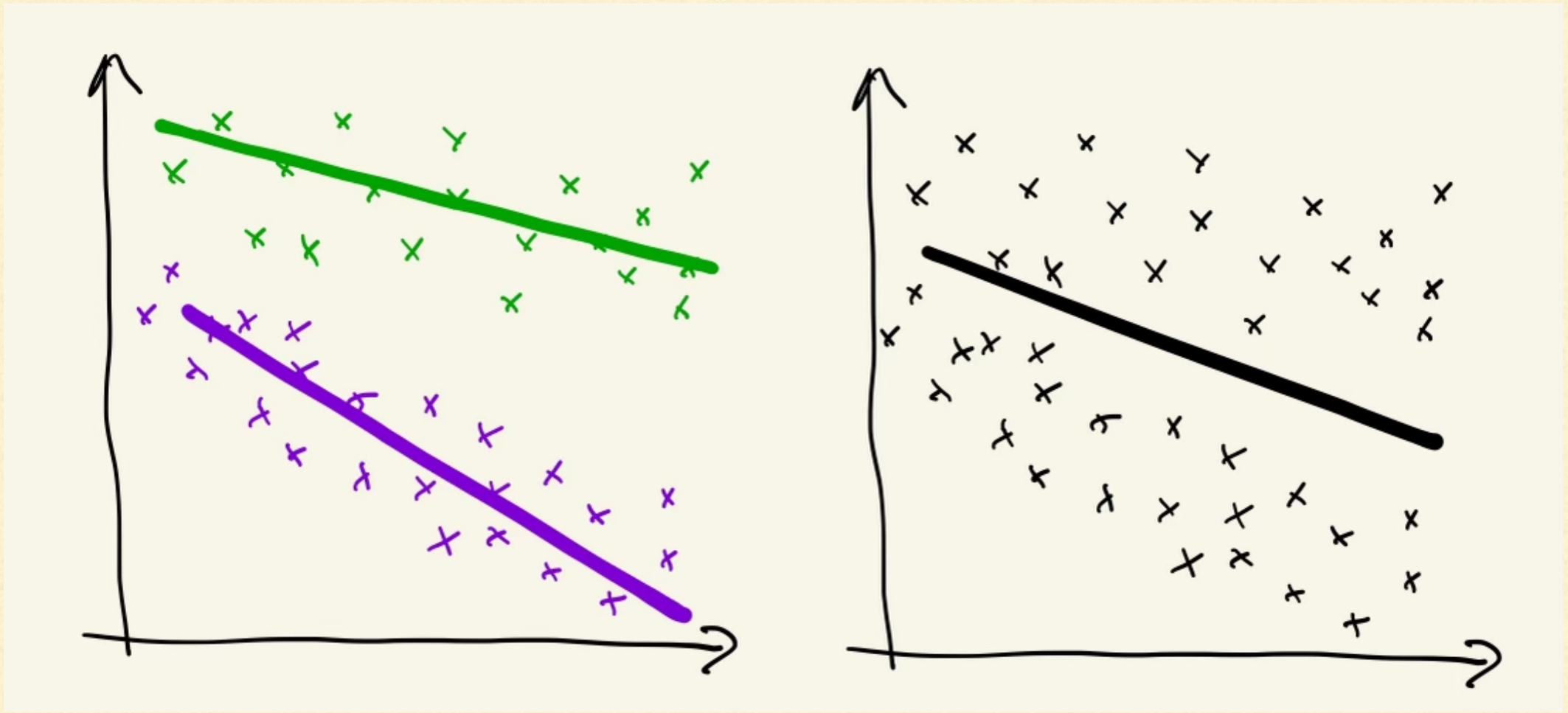
WHAT HAPPENS WHEN WE OMIT VARIABLES?

- Unmodeled precision variables get "averaged over"



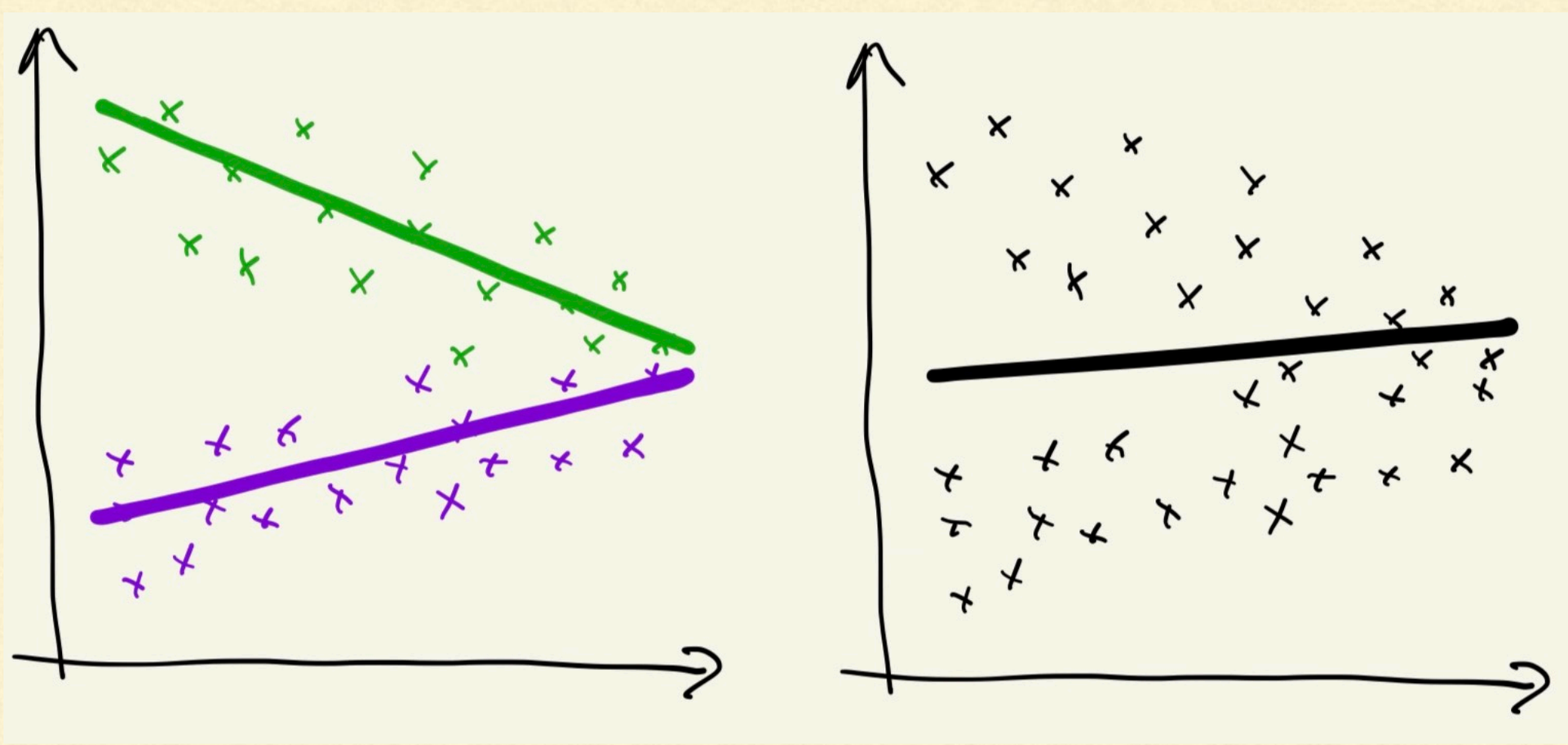
WHAT HAPPENS WHEN WE OMIT VARIABLES?

- Unmodeled effect modifiers get "averaged over"



WHAT HAPPENS WHEN WE OMIT VARIABLES?

- Unmodeled effect modifiers get "averaged over"



MODELS

- Think about your
  - Scientific question
  - Model, including relevant variables
  - Experimental design

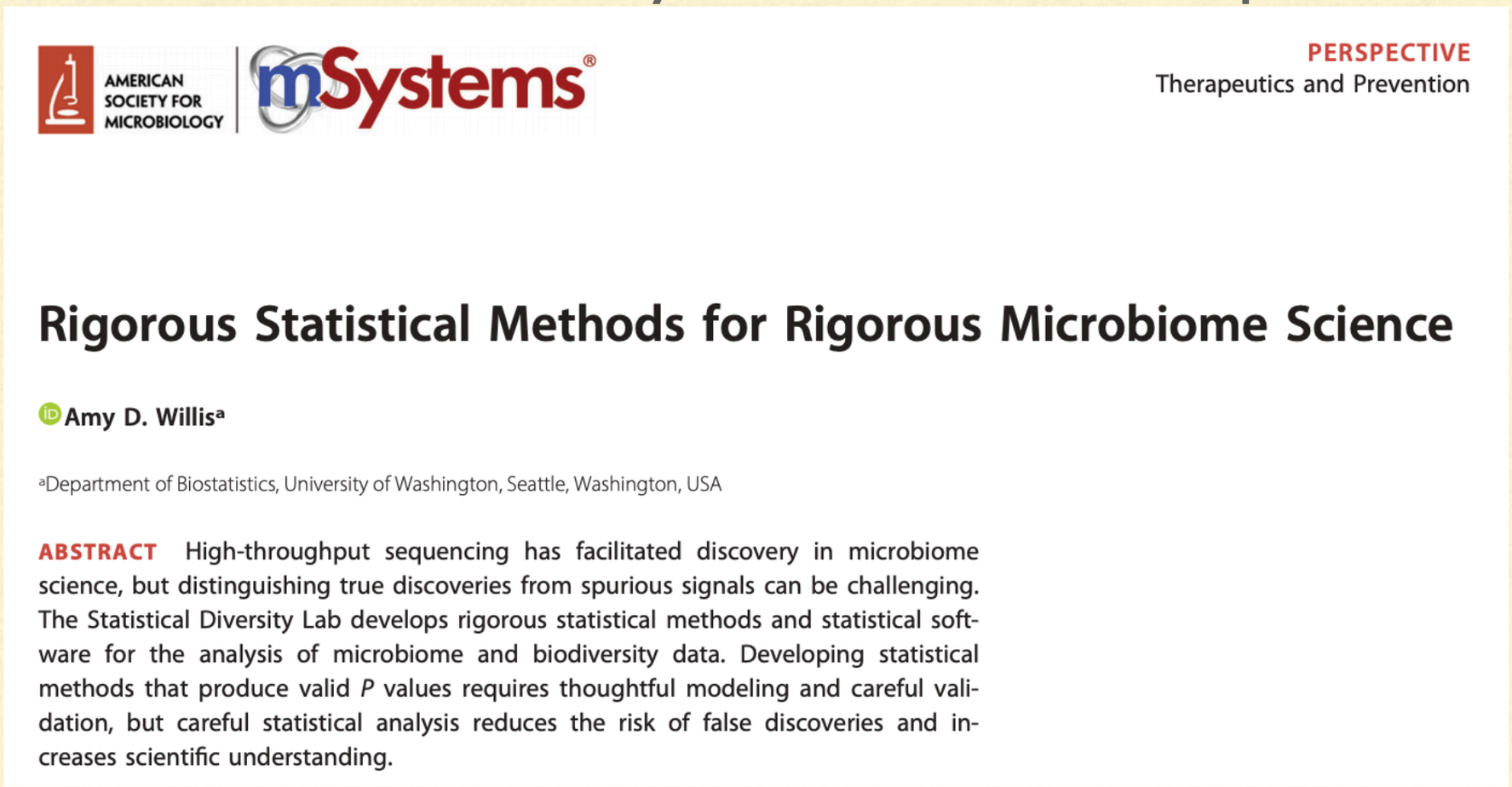
  before collecting your expensive, precious data!

- You may realize that you can't answer your question with the data you have…
  - …or that something else is even more interesting to you!

WHO, AGAIN?

statistical diversity lab

- We are the statistical diversity lab, and we develop…



AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
Therapeutics and Prevention

Rigorous Statistical Methods for Rigorous Microbiome Science

Amy D. Willis[a]

[a]Department of Biostatistics, University of Washington, Seattle, Washington, USA

ABSTRACT High-throughput sequencing has facilitated discovery in microbiome science, but distinguishing true discoveries from spurious signals can be challenging. The Statistical Diversity Lab develops rigorous statistical methods and statistical software for the analysis of microbiome and biodiversity data. Developing statistical methods that produce valid *P* values requires thoughtful modeling and careful validation, but careful statistical analysis reduces the risk of false discoveries and increases scientific understanding.

statisticaldiversitylab.com

@AmyDWillis adwillis@uw.edu

WHO, AGAIN?

statistical diversity lab

We work on what we believe to be the most critical methodological needs in microbial science and the most series shortcomings of existing analysis methods. Along with our research, we see outreach, education, and collaboration as a core part of this mission.

statisticaldiversitylab.com

@AmyDWillis adwillis@uw.edu

MANY THANKS

- Sarah Teichman $\times 10^6$
- The instructional team
  - For making this amazing workshop #STAMPS2023 happen
- YOU!
  - For engaging in reproducible and ethical science

RECAP

RECAP

STATISTICS BOOTCAMP

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Associate Professor

Sarah Teichman — @sarah_teichman — PhD Candidate

WHAT'S COMING NEXT?

- This session focused on *principles* of applied stats for microbiome data analysis

- We'll speak about specific models, estimands, analyses, etc., after lunch

- That was 11 years of work on my mental model distilled into ~2 hours… errr… questions?

WHAT'S COMING NEXT?

- Time permitting…
  - Case study!
- If not…
  - Lunch!

CASE STUDY 🤠🐄💩🐮💉

Who: A cohort of 30 existing dairy workers, 30 new dairy workers, and 30 community controls

What: Fecal, nasal, blood samples collected at baseline enrollment, 3 month, 6 month, 12 month, and 24 month follow-ups. Survey data on demographics, antibiotic use, & food frequency questionnaire.

When: September 2017 - December 2020.

Where: A conventional dairy farm in Yakima Valley, WA

🤠🐄💩🐮💉 CONT…

Hypothesis: Working on a dairy farm places people at a higher risk of acquiring antibiotic resistance genes.

Problem: You are interested in producing a pilot investigation into this hypothesis. However, you have a limited budget where you can only sequence ~300 million reads. You have enough money for library prep of up to ~30 samples.

Discuss: How might you design your study? What model will you investigate? How many and which samples do you choose?

DISCUSSION 🤠🐄💩🐮💉

Who/What/When/Where: 30 existing dairy workers, 30 new dairy workers, and 30 community controls working in a conventional dairy farm in Yakima Valley, WA. Baseline, 3 mo, 6 mo, 12 mo, 24 mo: Fecal & nasal samples. Data on demographics, antibiotic use, diet, blood samples. Data collected between September 2017- December 2020.

Hypothesis: Working on a dairy farm places people at a higher risk of acquiring antibiotic resistance genes.

Constraints: Shotgun sequencing up to 300 million reads (💰💰); library prep for up to 30 samples.

Discuss: How might you design your study? What model will you investigate? How many and which samples do you choose? How deeply do you sequence? What controls do you take?