STATISTICS BOOTCAMP

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Assistant Professor

David Clausen — @davidandacat — PhD Candidate

Sarah Teichman — @sarah_teichman — PhD Candidate

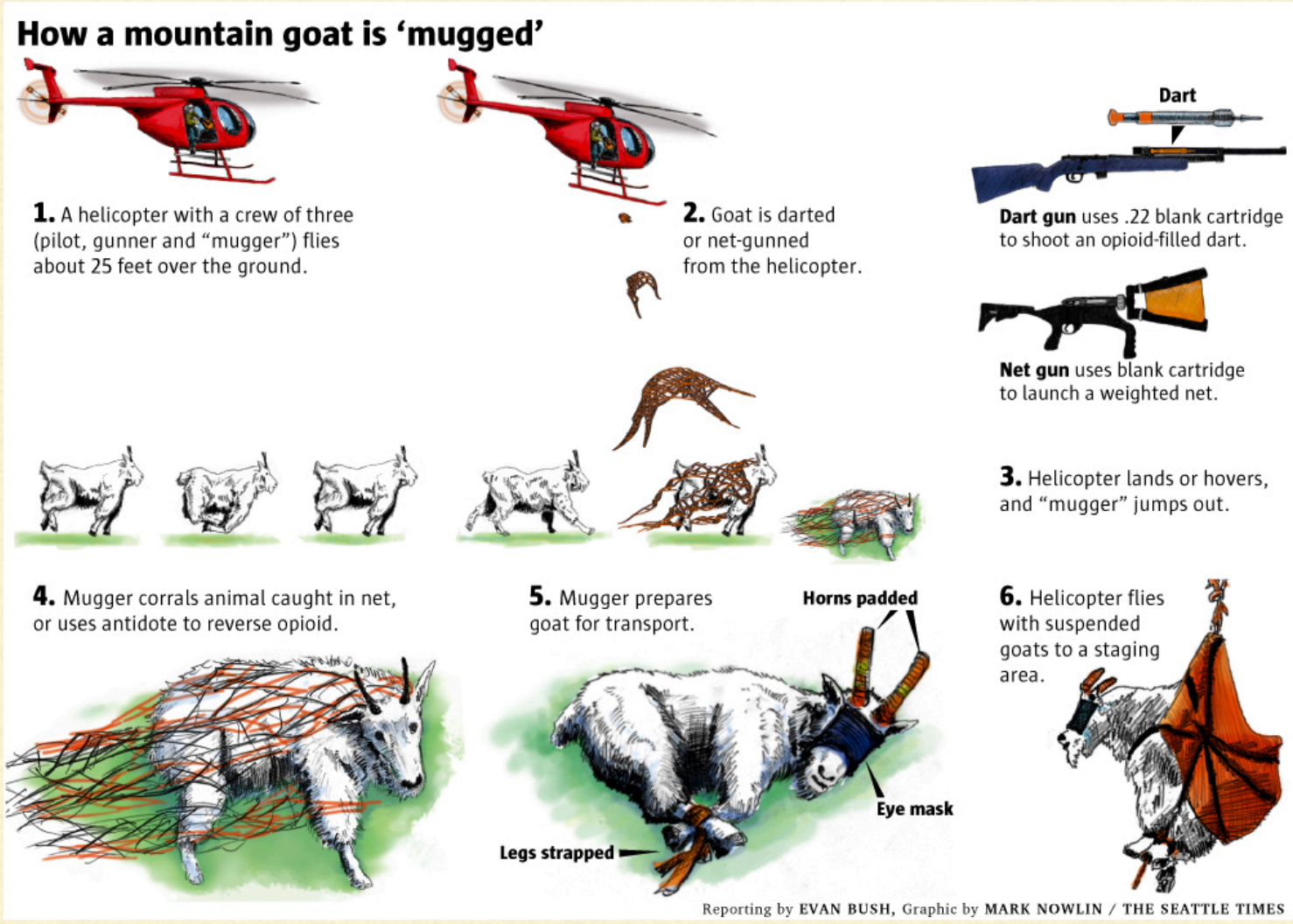
STATISTICS

- Who wants to know more statistics?
- Why?

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. Discuss model misspecification and its implications for hypothesis testing
4. Learn about statistical transparency and ethics
5. Draw connections with your previous statistics exposure (if any)

LEARNING PROCESS



STATISTICS

- Two different types of statistics
  - Inferential statistics
  - Exploratory statistics

EXPLORATORY STATISTICS

- What does your data say?
- How do we show what it says?
- How to visualise it?
- *Descriptive statistics*
- Not the focus of this talk (many other talks will discuss these ideas!)

INFERENTIAL STATISTICS

- "Use your data to say something about a population"

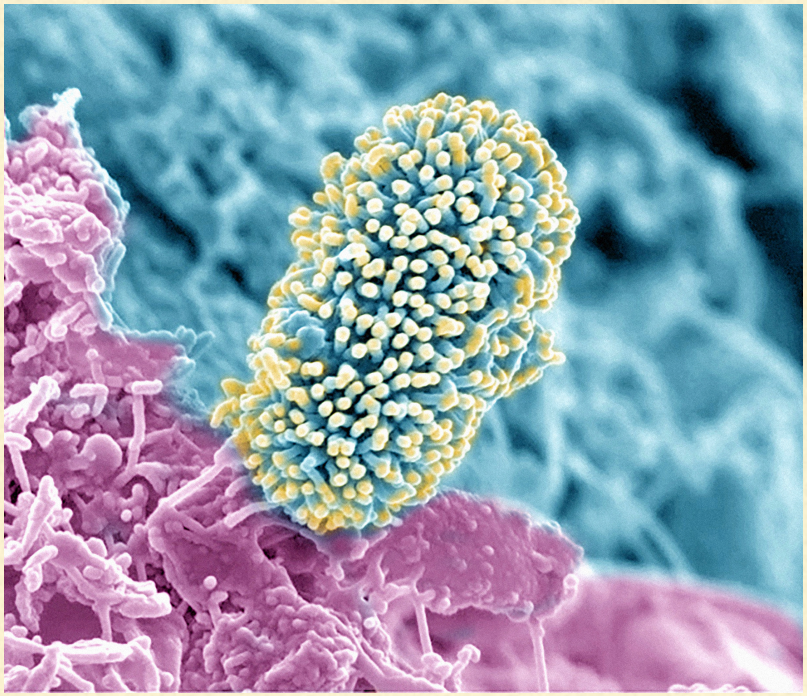
POPULATION

- 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



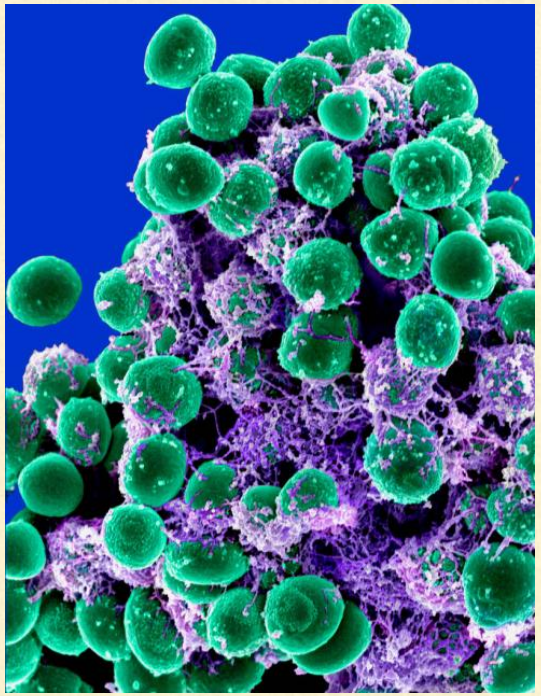


Photo credits: Pacific Northwest National Laboratory; NIAID

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 White women 25-45 y.o. with a clinical IBD diagnosis?
  - **Where?** … who also live in the city that your study was conducted in?
  - **When?** … between January 2022-March 202?
- 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

"SOMETHING ABOUT THE POPULATION"

- Statisticians have a formal concept of this
- "Parameter": a numerical characteristic of a model

PARAMETERS

- 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 Streptococcus in your saliva right now

PARAMETERS

- Example of a model:
  - Every #STAMPS2022 attendee carries methicillin-resistance S. aureus in their oral cavity, or they don't
- Example of a parameter of this model
  - The fraction of #STAMPS2022 attendees carrying S. aureus that have methicillin-resistance 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

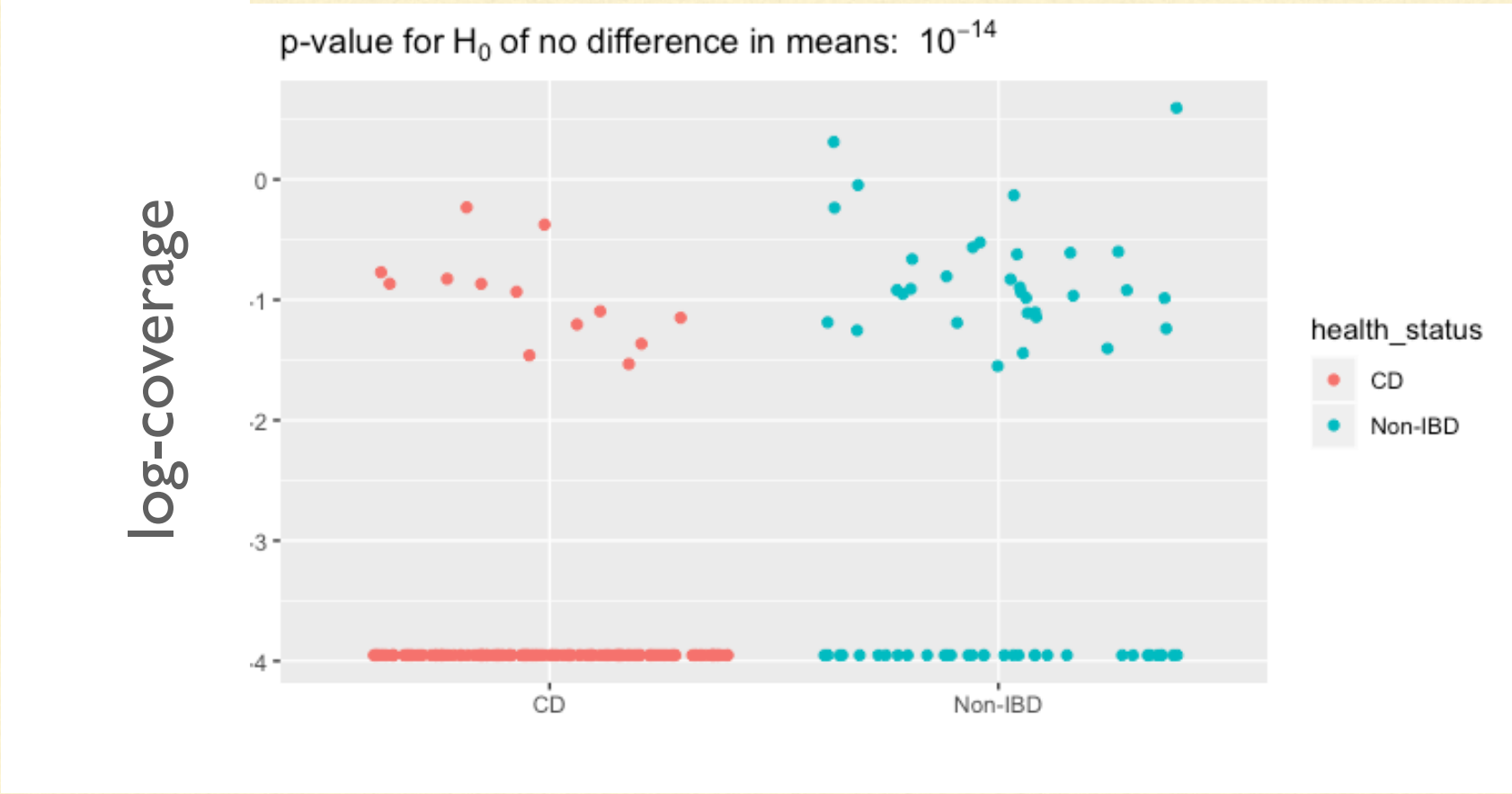
- Statisticians often estimate *means*
  - Never forget that this is a broad-range summary
  - Results about *means* say nothing about *individuals*!
  - Trade off of inferential statistics: need a population

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…

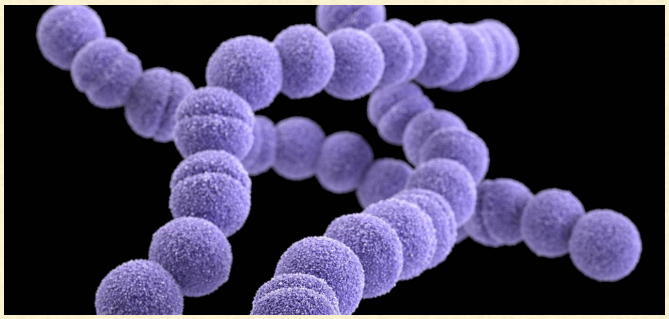
POPULATIONS VERSUS SAMPLES

- 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?

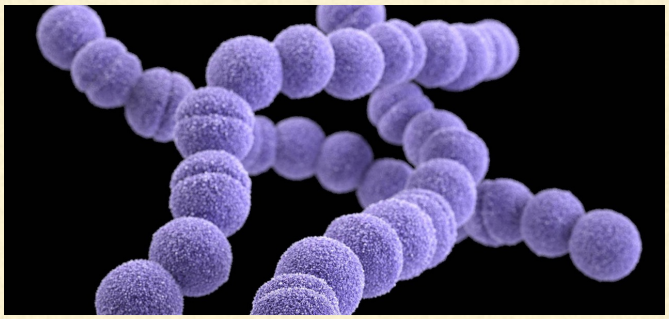
RELATIVE ABUNDANCE

- 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?
    - We will return to this question in a few slides…

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*:

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}
- *Which* do you prefer and *why*?

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

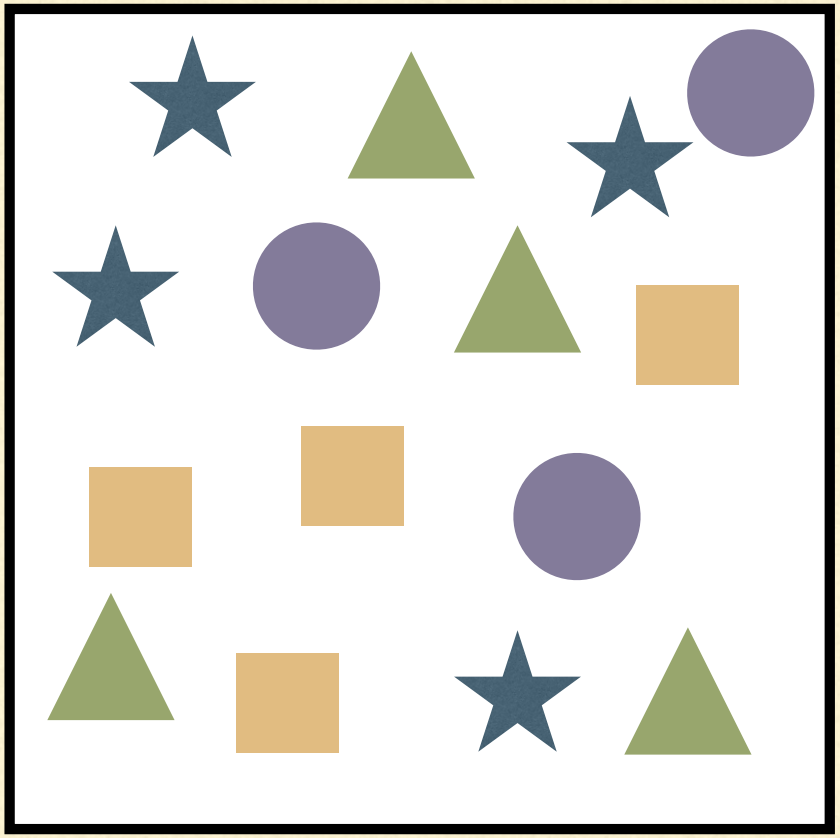
- An estimator of a parameter
  - is unbiased if
    - its bias is zero under the model
- The distribution of the estimator depends on the distribution of the data...
  - i.e., your model

BIAS

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

EXAMPLE

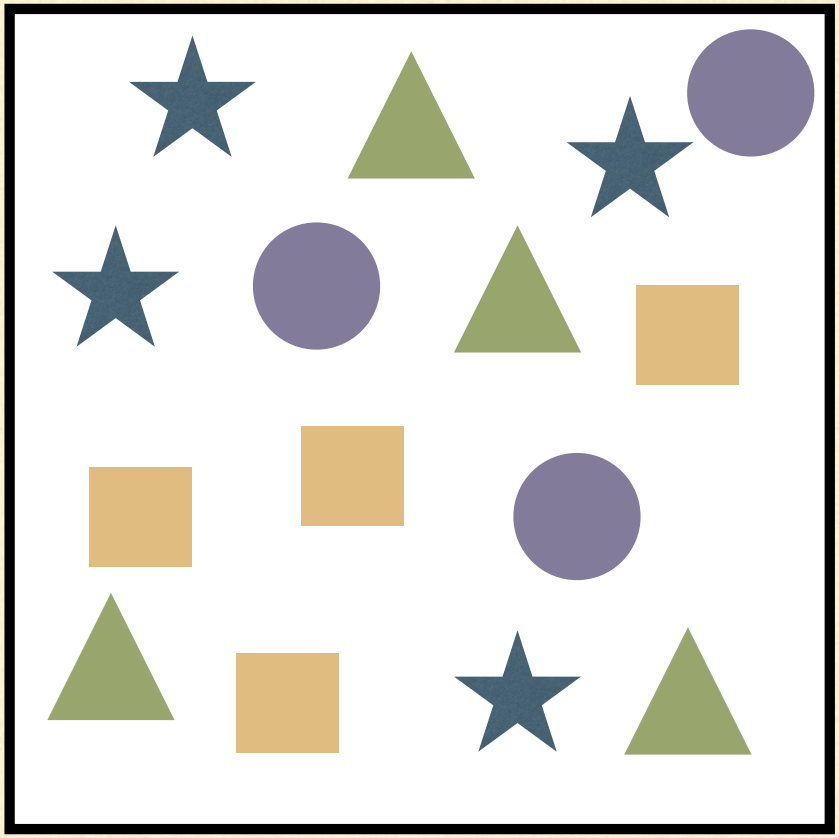
- What are the true relative abundances in the community?
- What's the (true) richness?
- Shannon diversity?



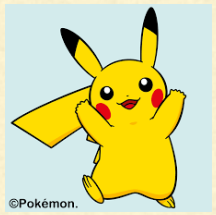
EXAMPLE

True abundances:

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

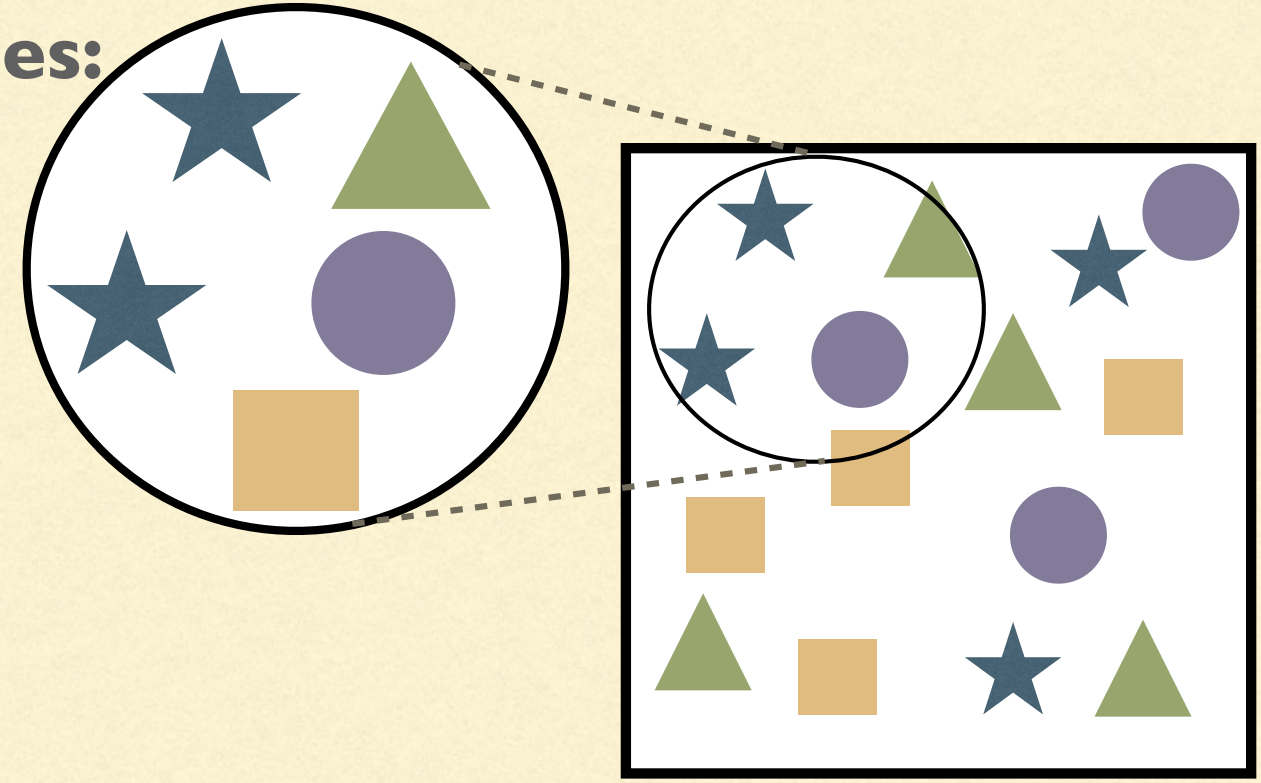


EXAMPLE



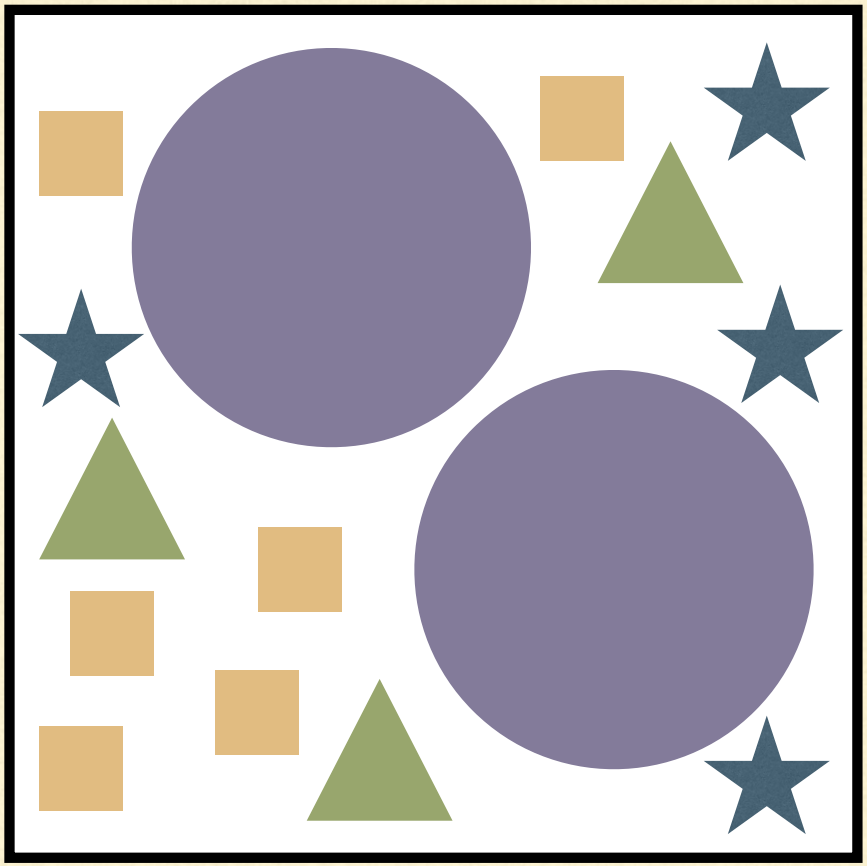
Observed abundances:

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



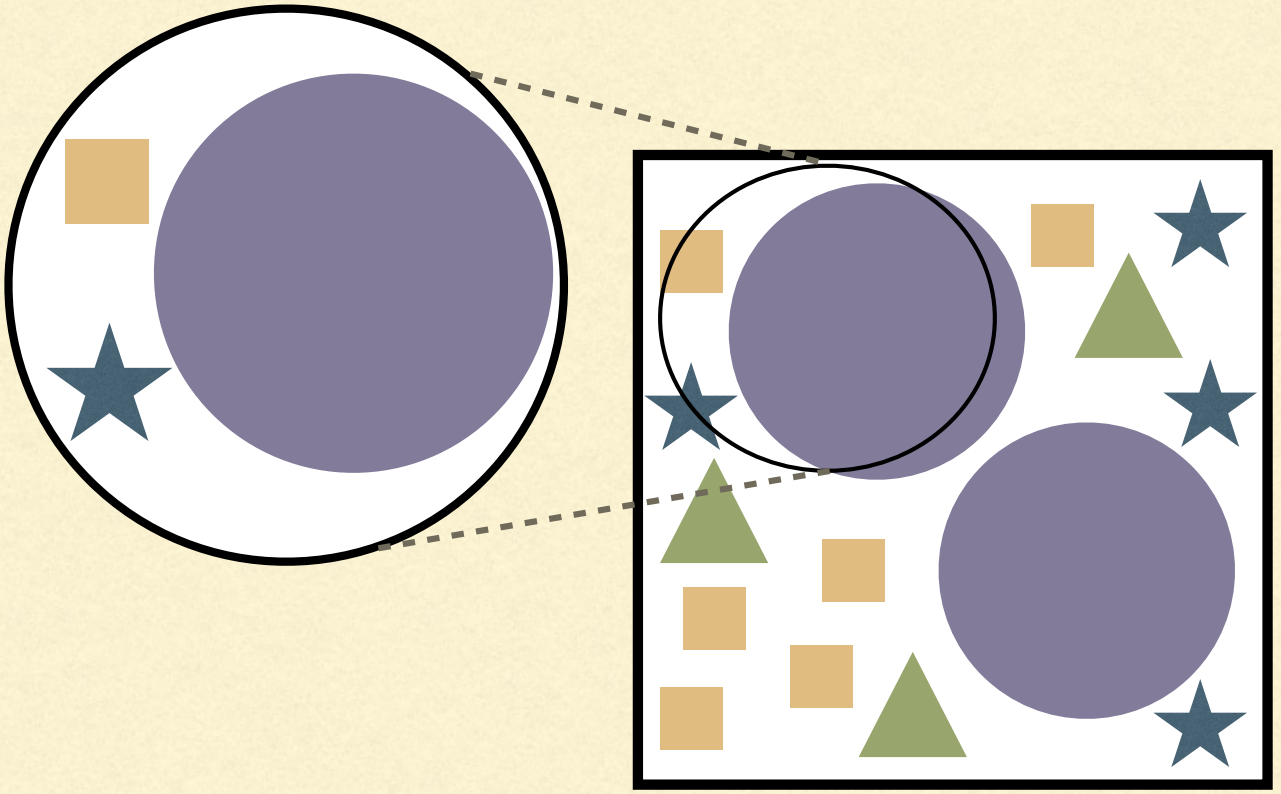
BIAS

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



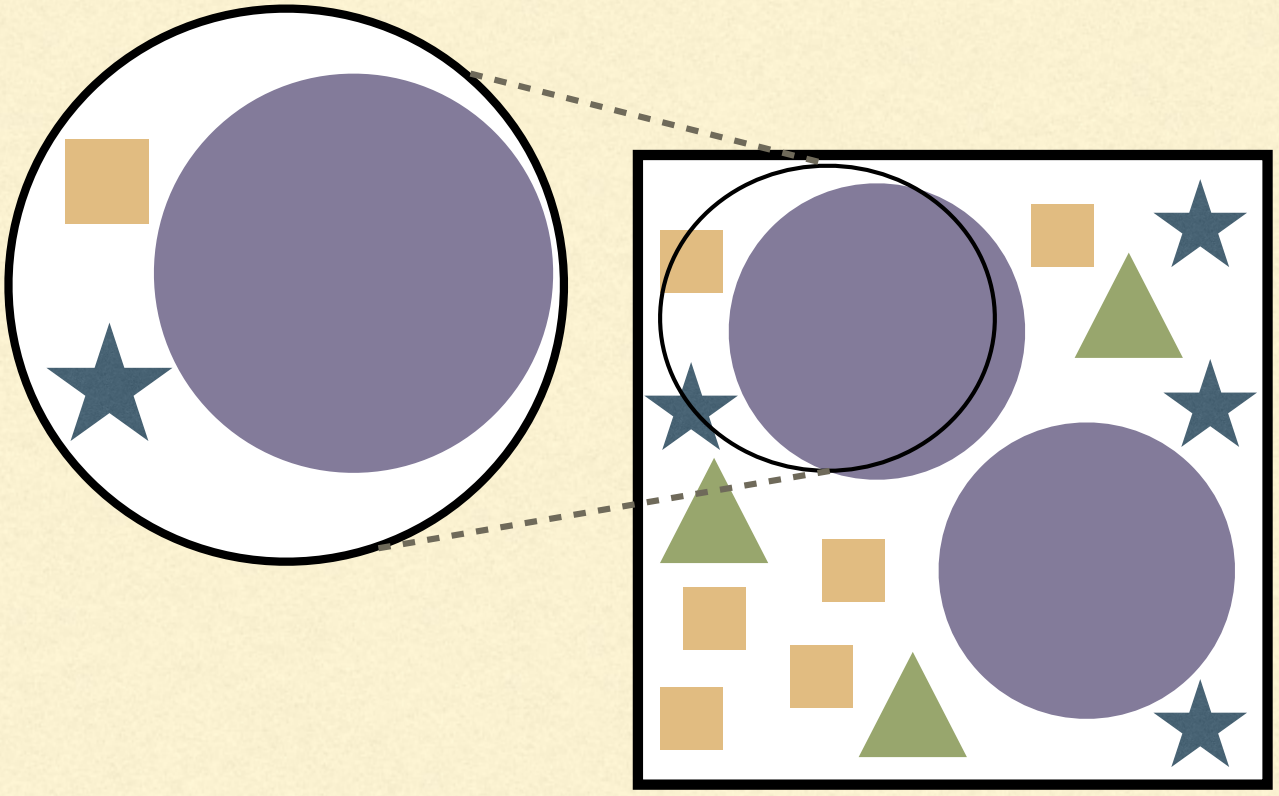
BIAS

-  = 1/3
-  = 1/3
-  = 1/3



BIAS

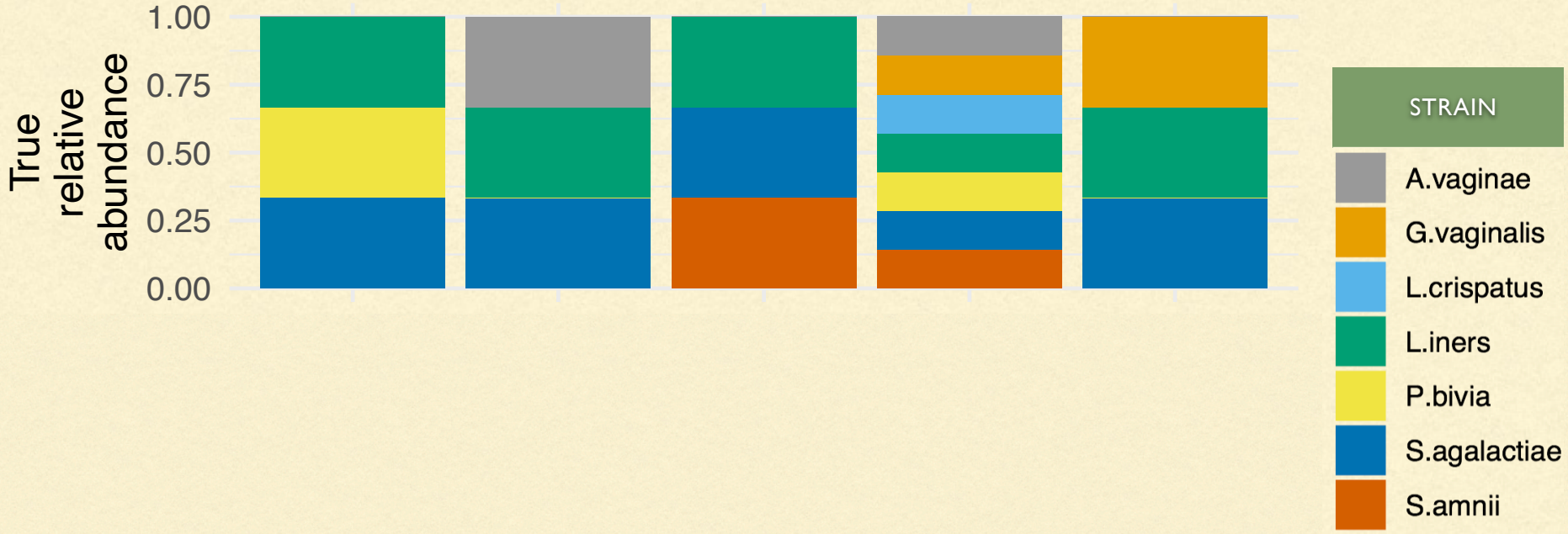
- ★ = 5/15
- ● = 5/15
- ■ = 5/15



Truth:

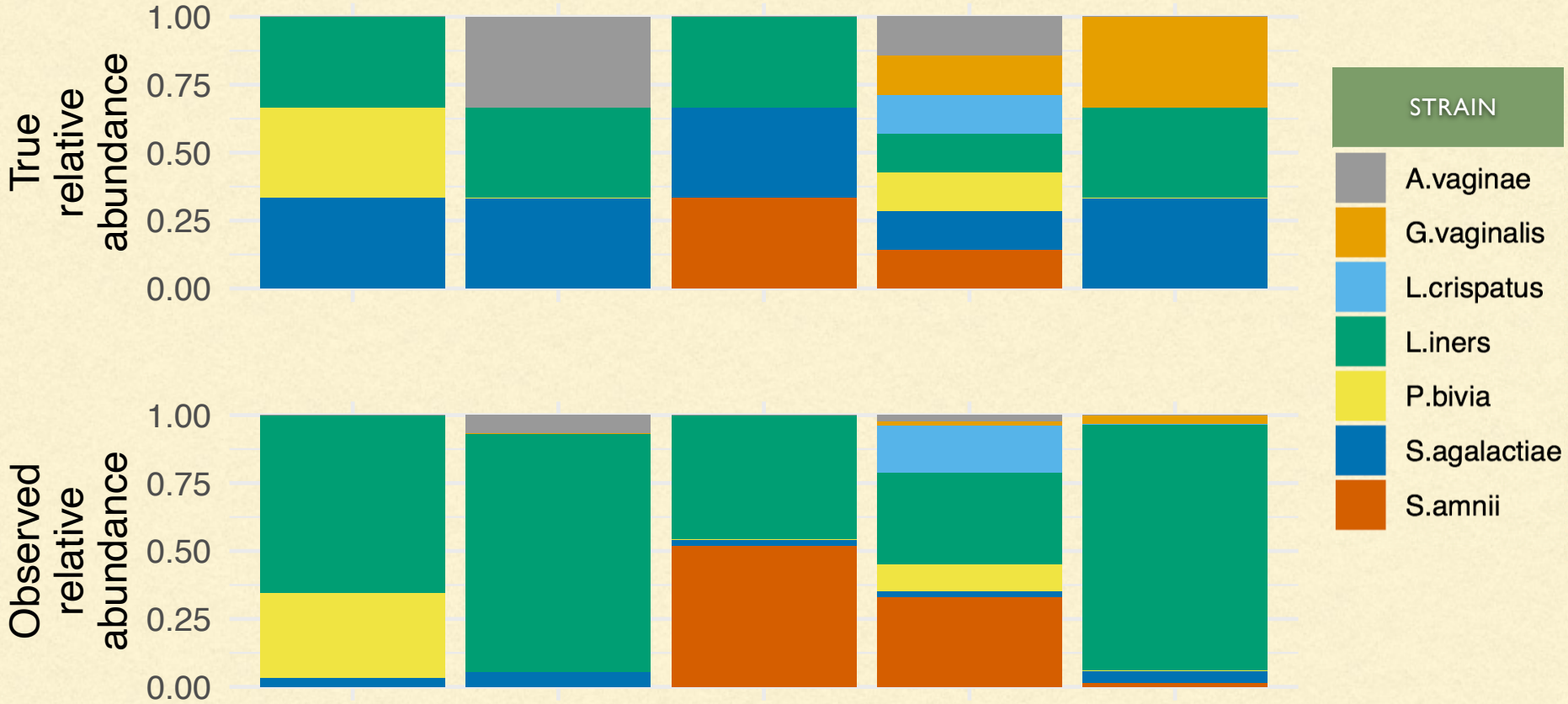
★ = 4/15 ● = 2/15 ▲ = 3/15 ■ = 6/15

BIAS AND RELATIVE ABUNDANCE



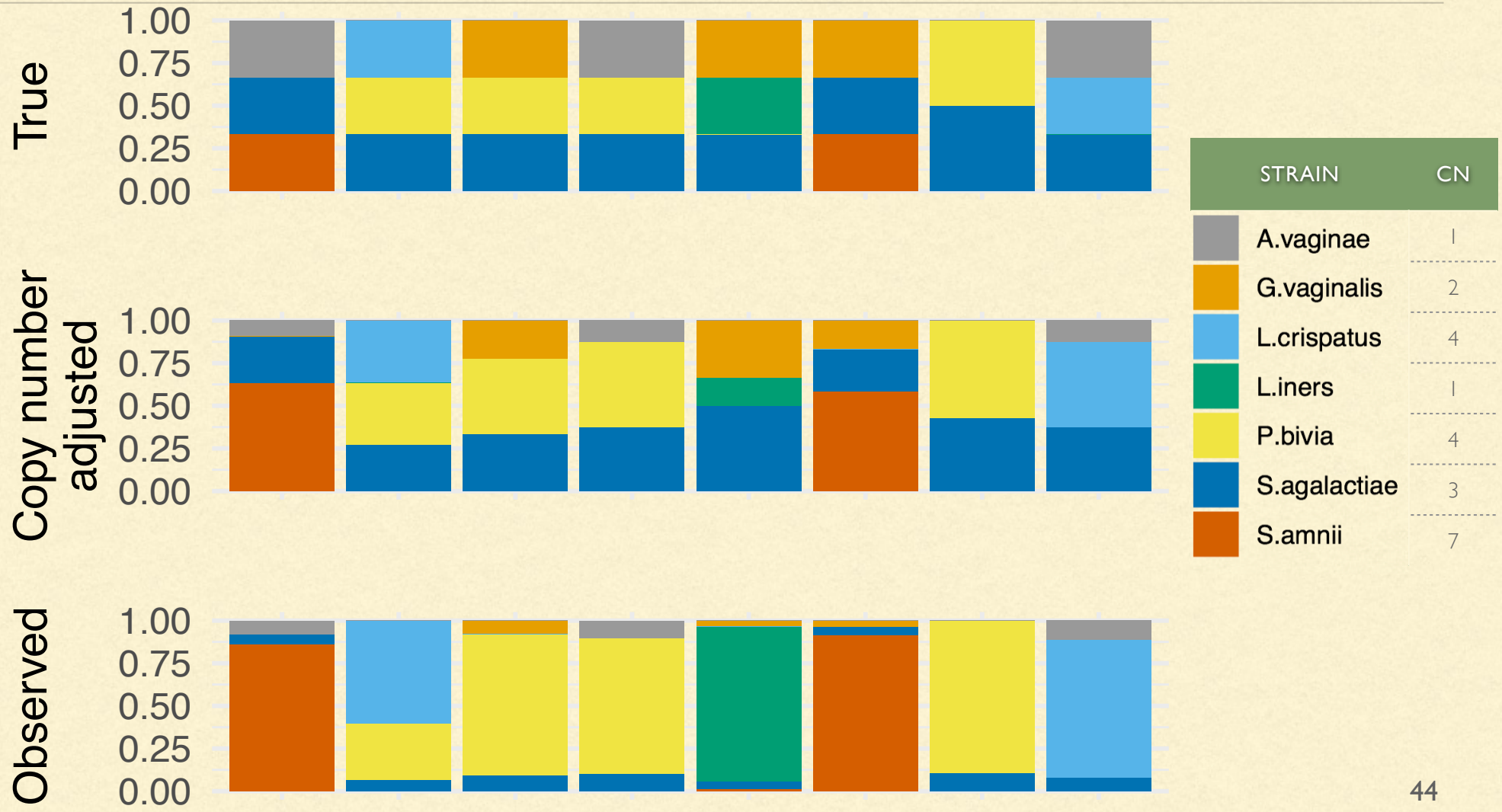
Data from Brooks *et al* (2015); figure adapted from McLaren, Willis & Callahan (2019)

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

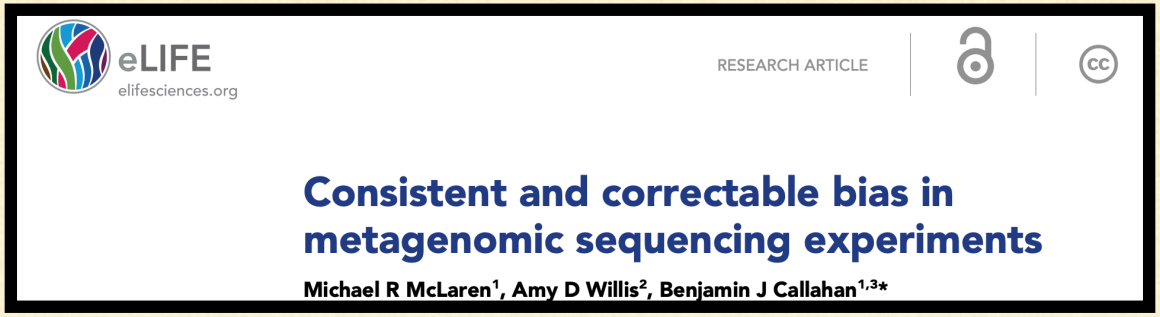
Taxon specific detection efficiencies

make

sample relative abundance

biased for

actual relative abundance



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!
  - More discussion this afternoon when we talk about estimation

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 33805200
- A large change in the estimates equals a large variance
- Standard deviation = $\sqrt{\text{variance}}$

VARIANCE

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

- Any *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
- If we model a *random variable* with a linear regression model, we can *partition* variance

SUMMARY SO FAR

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

QUESTIONS

- We're about to take a break
- Any questions before we break?

BREAK

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.,
  - $\text{expected diversity}_i = \beta_0 + \beta_1 \times \mathbf{1}_i \text{ is from lakewater (not seawater)}$
    - $\hat{\beta}_0$ 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

Centroid for sample i 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 causal effect of smoking on lung function in teenagers, age is a confounder
- Need a causal model (and serious training/collaboration)

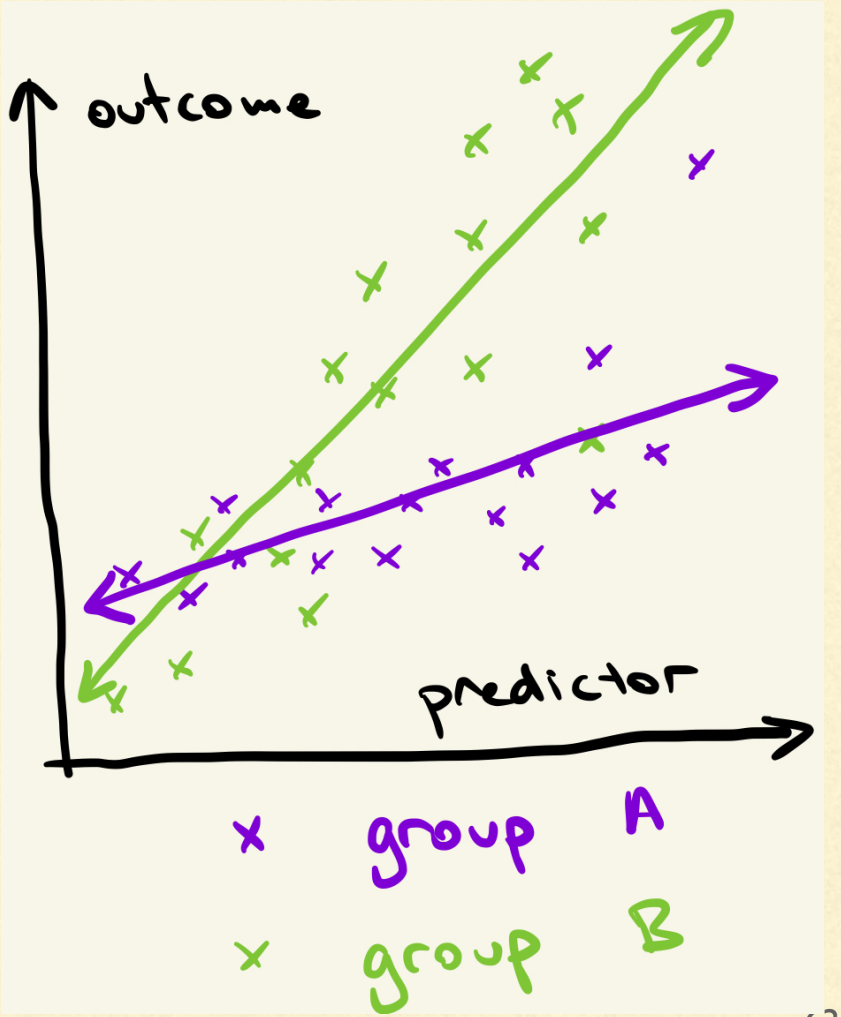
TYPES OF VARIABLES

3. Effect modifiers

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

- "interaction" between variables

- Don't always need to include in model!

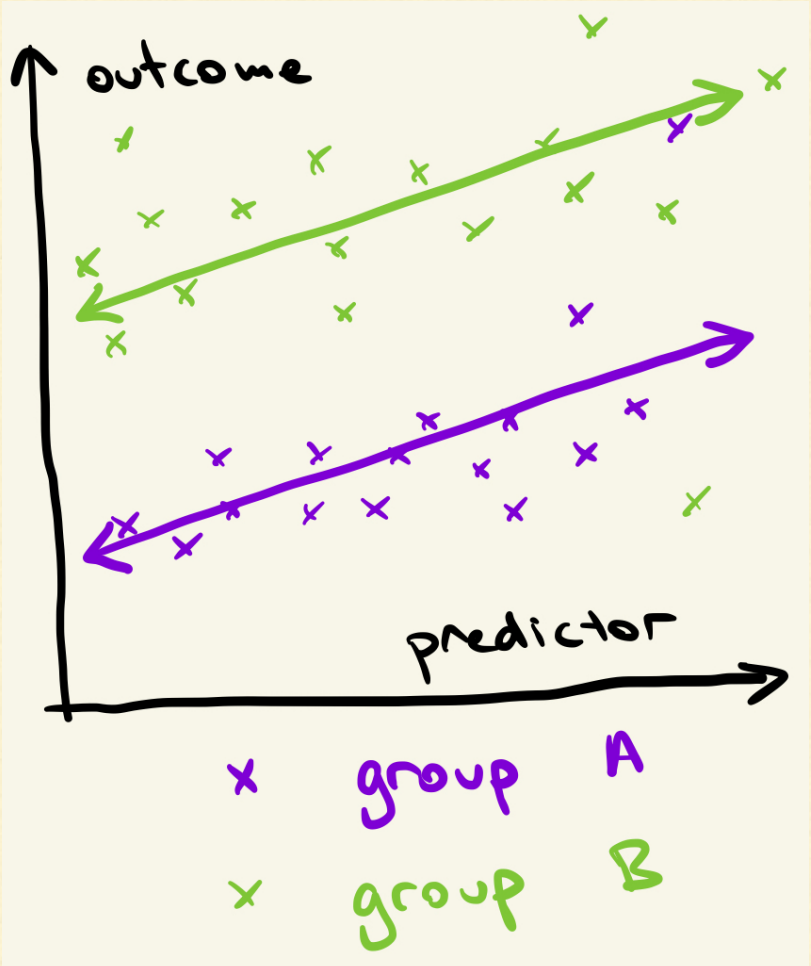


TYPES OF VARIABLES

4. 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"



REFLECTING ON VARIATION

- Group exercise (5 minutes)
  - *Why might two studies find different results?*
    - Researchers in Boston have published that the Firmicutes-to-Bacteroides ratio is lower in folx w IBD compared to "healthy" controls
    - Researchers in Seattle have published that the F-to-B ratio is greater in folx w IBD compared to "healthy" controls
    - Both used 16S sequencing
  - List some reasons why this could have happened

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!

INFERENCE

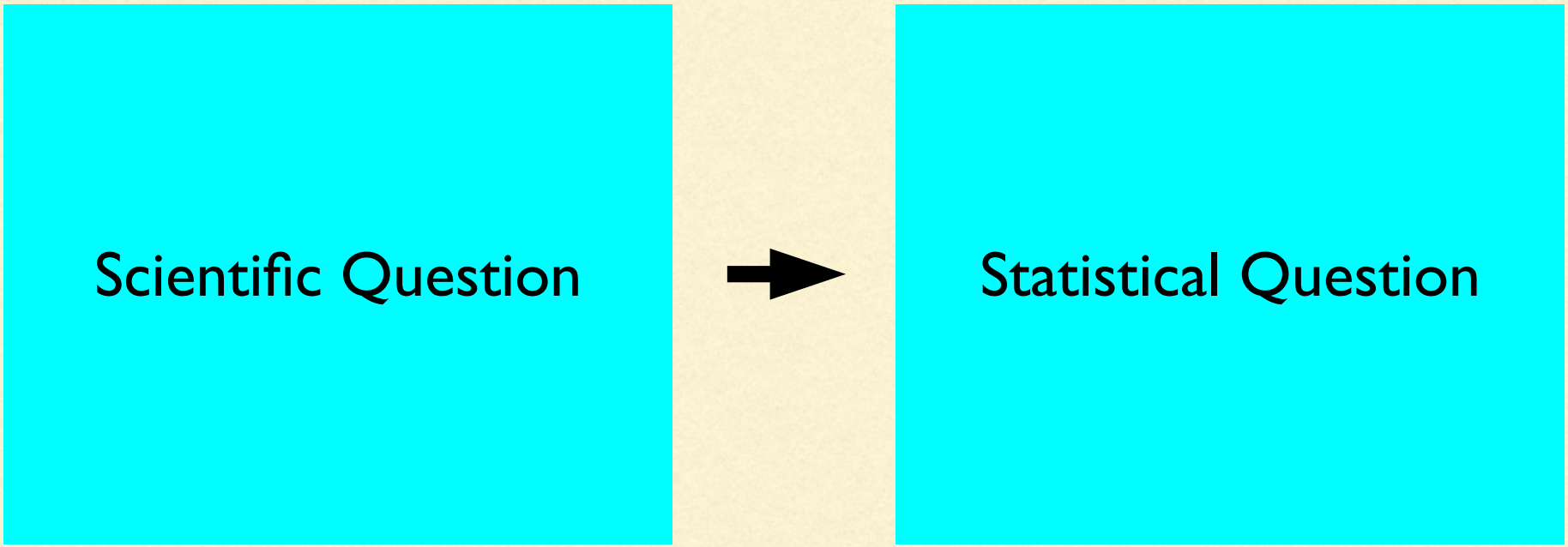
From David & Sarah

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

Do mountain goats that have been relocated have a different life expectancy than non-relocated goats?

Is the mean life span of mountain goats that have been relocated different from the mean life span of mountain goats that have not been relocated?

HYPOTHESIS

Ask yourself, do you have the data for these parameters?

- First translate your scientific question into a statistical question

Do mountain goats that have been relocated have a different life expectancy than non-relocated goats?

Is the mean life span of mountain goats that have been relocated different from the mean life span of mountain goats that have not been relocated?

HYPOTHESIS

- First translate your scientific question into a statistical question

Which taxa are differentially abundant across two groups?

Your turn! Can you turn this into a statistical question?

HYPOTHESIS

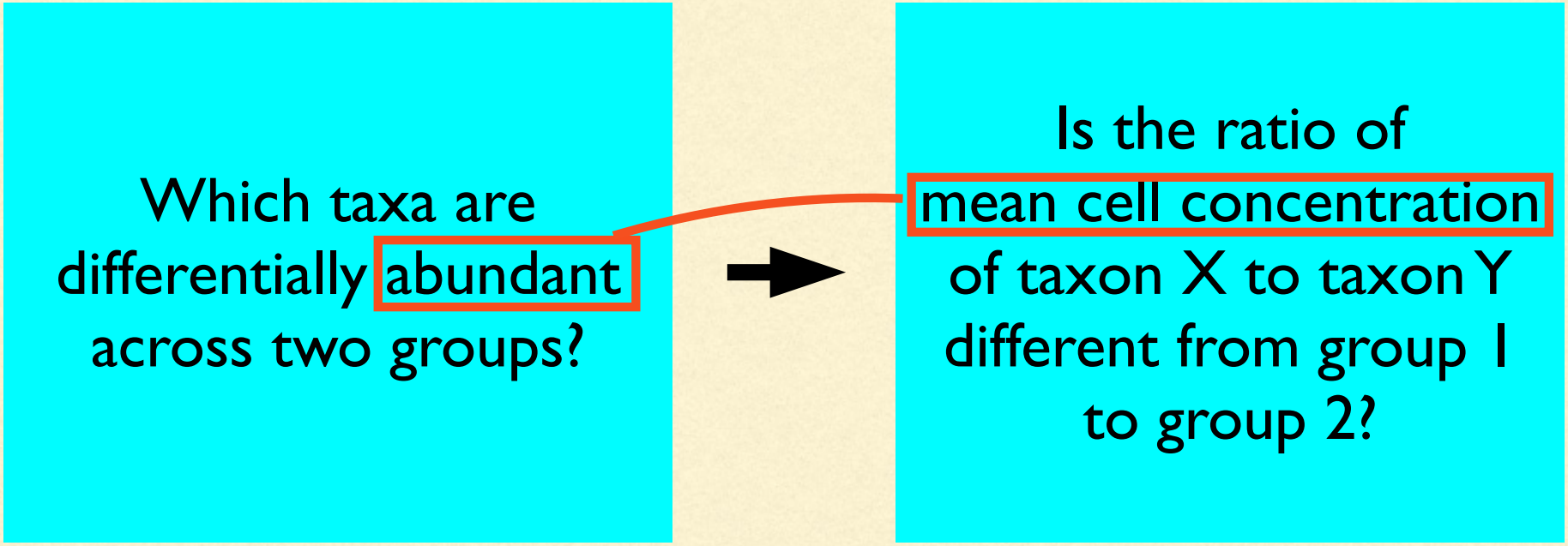
- First translate your scientific question into a statistical question

Which taxa are differentially abundant across two groups?

Is the ratio of mean cell concentration of taxon X to taxon Y different from group 1 to group 2?

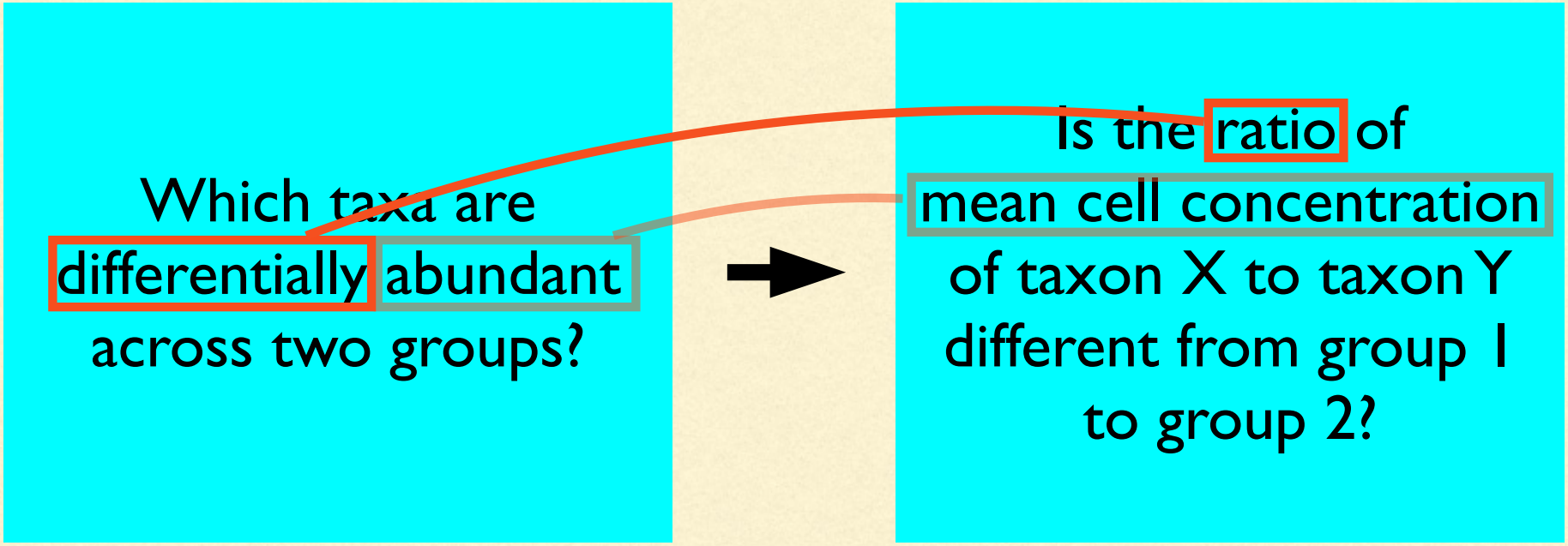
HYPOTHESIS

- First translate your scientific question into a statistical question



HYPOTHESIS

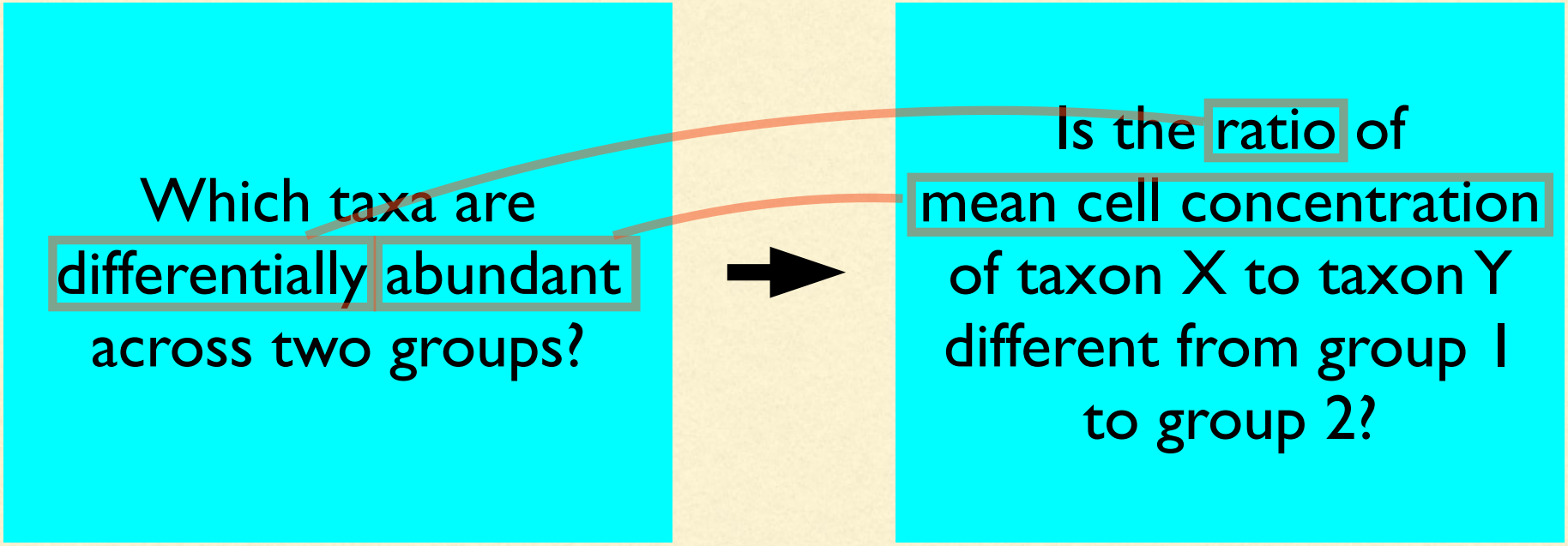
- First translate your scientific question into a statistical question



HYPOTHESIS

Ask yourself, do you have the data for these parameters?

- First translate your scientific question into a 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

Statistical Question → Null Hypothesis(H_0) Alternative Hypothesis(H_A)

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 ratio of mean cell concentration of taxon X to taxon Y different from group 1 to group 2?

H_0: ratios are the same for both groups
H_A: ratios are different between the two groups

HYPOTHESIS

- Two possible conclusions from a hypothesis test:

  1. Reject the null hypothesis. This provides support in favor of the alternative hypothesis.

  2. Fail to reject the null hypothesis. This means you don't have enough information to support the alternative hypothesis.

HYPOTHESIS

If you've ever written a proof by contradiction, this way of thinking might be familiar!

- Two possible conclusions from a hypothesis test:

1. Reject the null hypothesis. This provides support in favor of the alternative hypothesis.

2. Fail to reject the null hypothesis. This means you don't have enough information to support the alternative hypothesis.

TESTING

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

    1. Calculate a test statistic

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

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:

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

Larger difference: more evidence against H_0

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:

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

Larger difference: more evidence against H_0

Smaller value: more certainty about estimate

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

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 or low certainty about 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

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

Suppose your standard error is half what it should be.
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

Suppose your standard error is half what it should be. 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~~

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

TESTING

- Ask how likely it would be to observe this test statistic 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 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 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 if the null hypothesis were true
  - Formally,

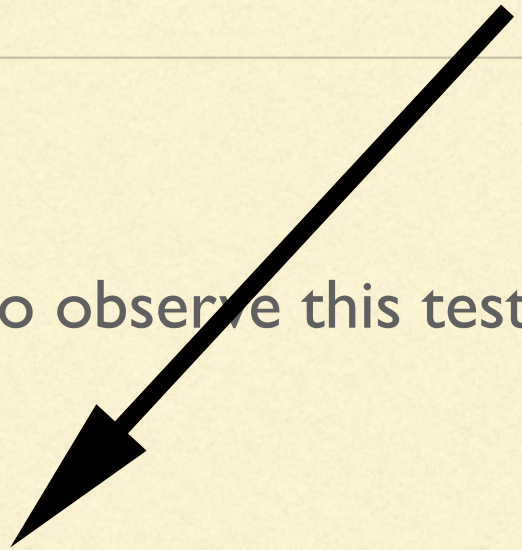
$$Pr\left(|T| \geq |t| \,\middle|\, H_0 \text{ true} \right)$$

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 if the null hypothesis were true
  - Formally,

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

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

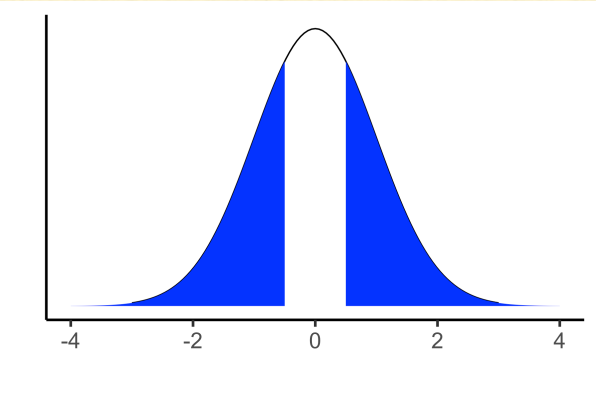
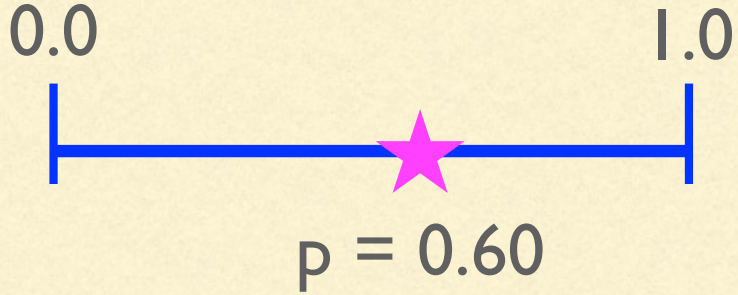
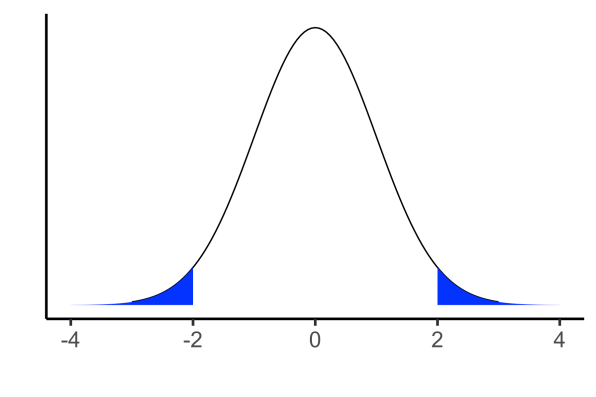
Why? The Central Limit Theorem!

P-VALUE

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

P-VALUE

Recall, $t = \dfrac{\text{estimate} - H_0 \text{ value}}{\text{standard error}}$

	test statistic	p-value	conclusion
t = 0.5			Less evidence against H_0
t = 2.0			More evidence against H_0

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 = you are lying to your readers
    - Understating your evidence against H_0
  - More often = you are lying to your readers
    - Overstating your evidence against H_0

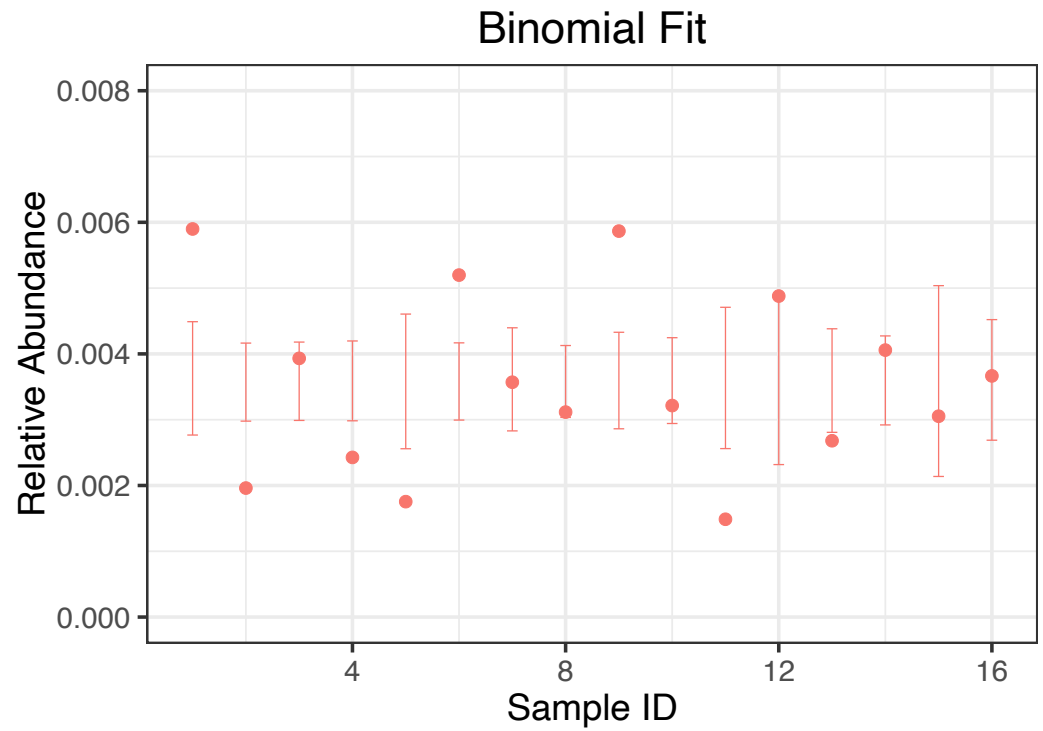
VALID HYPOTHESIS TEST

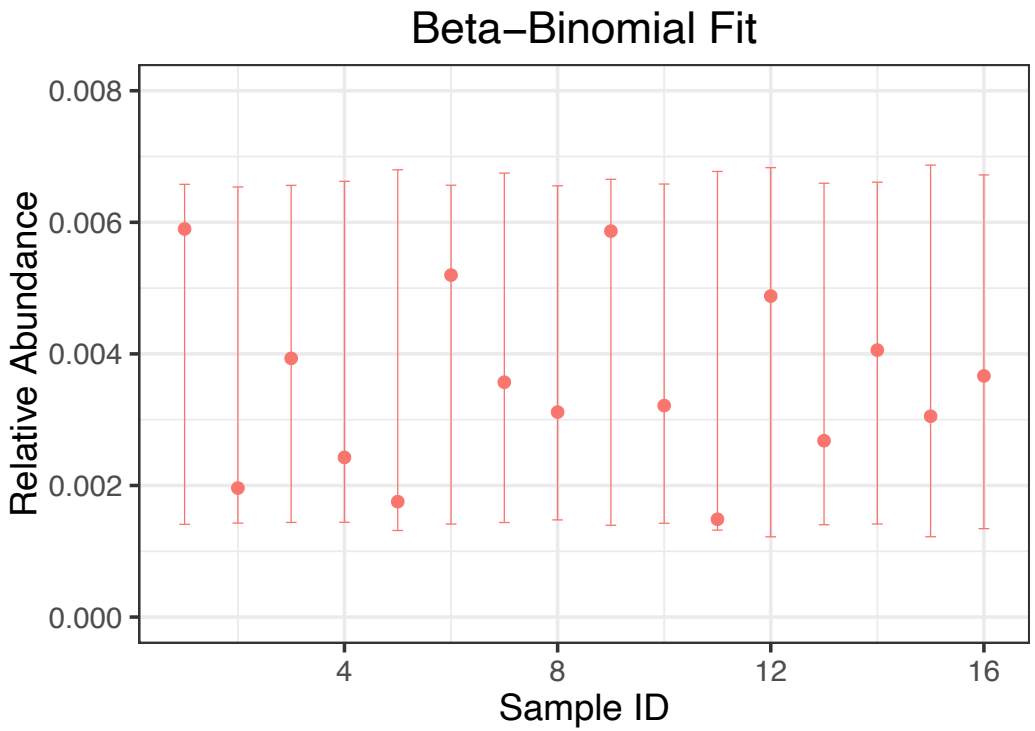
- Why might a hypothesis test be invalid?
  - Hint: recall our discussion of standard errors!

MODEL MISSPECIFICATION

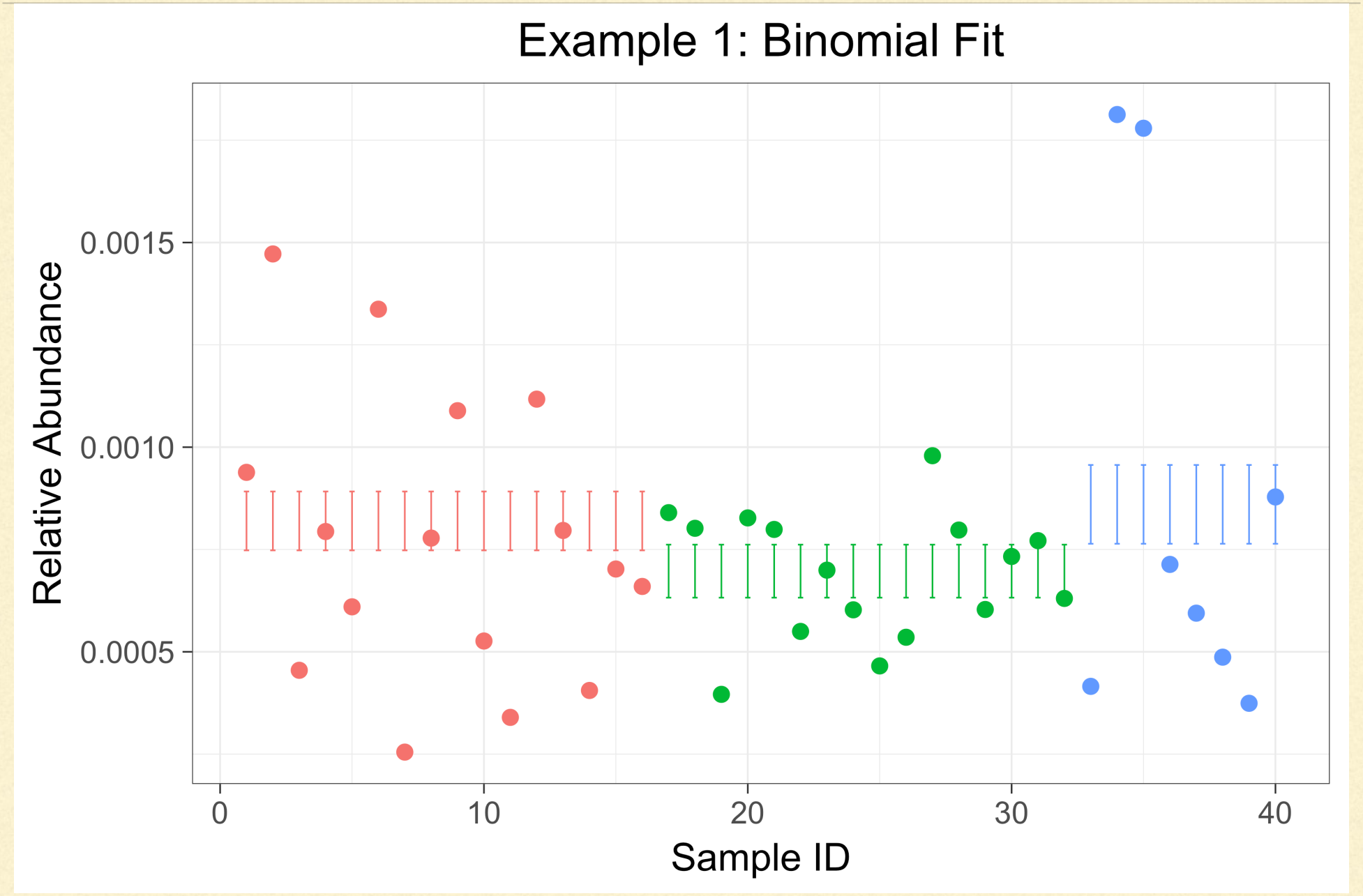
- p-values require estimates of the variance
- estimates of the variance require models
- wrong model → wrong variance → wrong p-value!

MODEL MISSPECIFICATION

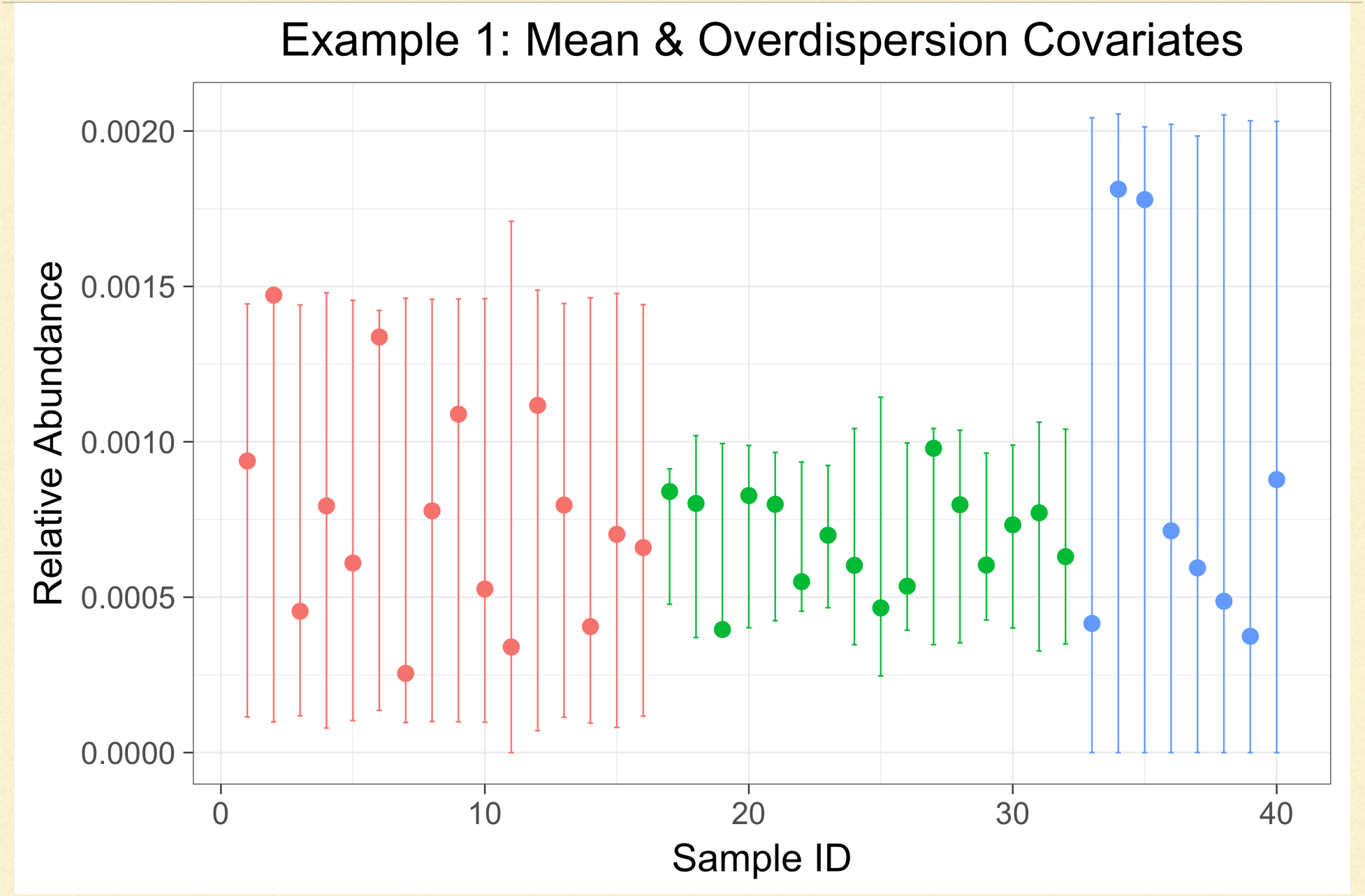




🌽🌽 Martin, Witten, & Willis (2019) 🌽🌽



p-value: 0.0058 😿😿



p-value: 0.3782 🌽🌽

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 species. They find that at an alpha level of 0.05, 50 species are associated with cancer. They publish their findings, reporting their statistically significant p-values.

- Your local statistician says "wait a minute, I'm concerned about your Type I errors." Why?

MULTIPLE TESTING

- 2 independent tests:
  - Don't reject H_0 for Test 1 $= .95$
  - Don't reject H_0 for Test 2 $= .95$
  - Don't reject H_0 for both tests $= .95 \times .95 = .9025$
- Don't make any type 1 errors: ~90%

MULTIPLE TESTING

- 3 independent tests:
  - Don't reject H_0 for all tests $= .95 \times .95 \times .95 = .8574$
- Don't make any type I error: 86%

MULTIPLE TESTING

m (independent) tests

$$(1-\alpha)^m = P(\text{don't reject any } H_0 \,|\, \text{all } H_0)$$

$$1-(1-\alpha)^m = P(\text{reject any } H_0 \,|\, \text{all } H_0)$$

MULTIPLE TESTING

m (independent) tests

$$(.95)^m = P(\text{don't reject any } H_0 \,|\, \text{all } H_0)$$

$$1 - (.95)^m = P(\text{reject any } H_0 \,|\, \text{all } H_0)$$

MULTIPLE TESTING

m	$1 - (.95)^m$
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, 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, consider:
  - **Family-wise Error Rate (FWER):** probability of at least one type I error
    - Use Bonferroni correction, divide α by number of tests
  - **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

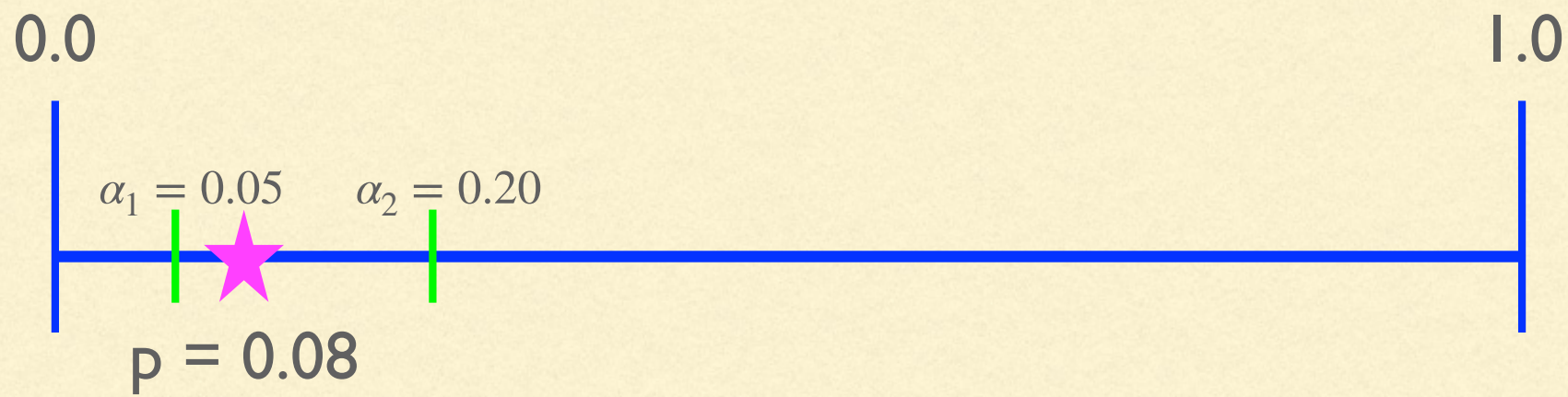
- Our alpha level specifies the probability of a Type 1 error
- We can also commit Type 2 errors: "false negatives"
- Power: probability of correctly rejecting the null hypothesis, **when it is false**
  - $Pr(\text{reject } H_0 \mid H_0 \text{ is false})$
  - 1 - probability of type 2 error

TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level
  - **Exercise**: Why?

TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level
  - **Exercise**: Why?

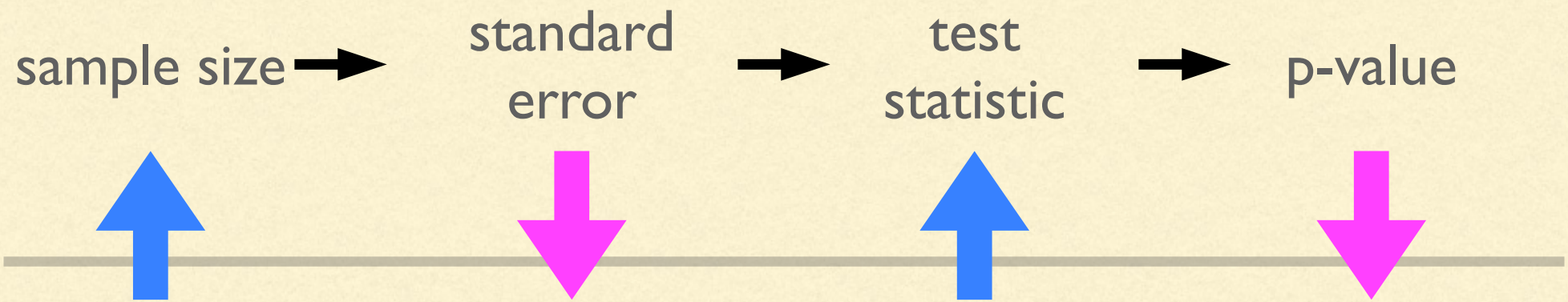


TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level
  - **Exercise**: Why?
- We can increase power *without sacrificing Type 1 error* by increasing our sample size

TYPE 2 ERROR & POWER

- We can increase power by increasing our alpha level
  - **Exercise**: Why?
- We can increase power *without sacrificing Type I error* by increasing our sample size



AMY'S RECAP ON P-VALUES

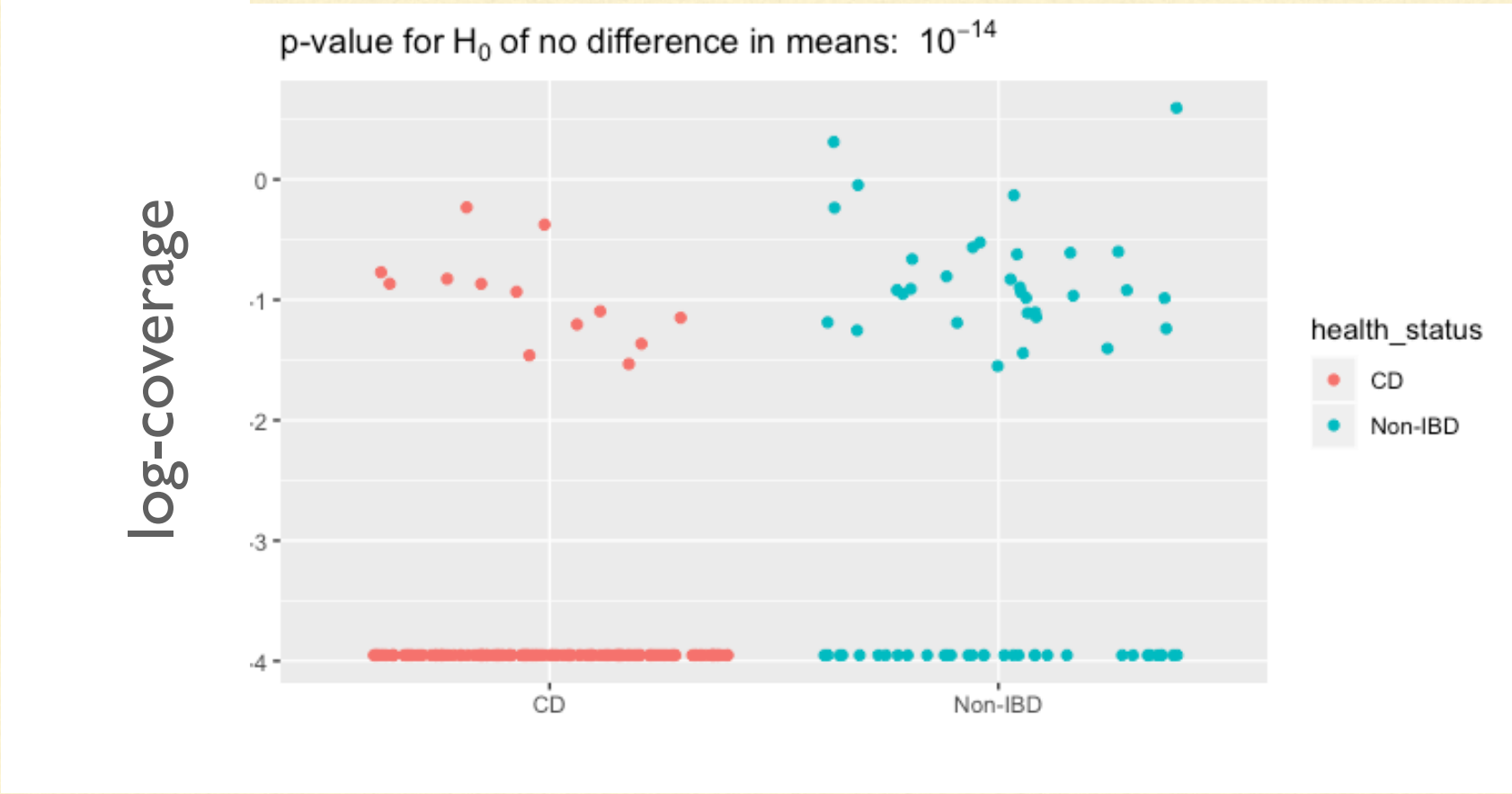
- Current conventions for statistical analysis in science are inherently unscientific
  - Yes, this is unfair
- You have some obligations
  - Know what a hypothesis test is doing: what is the parameter, model, null hypothesis…
  - Do not consider "p < [favourite number]" a benchmark for publication!

EVERY P-VALUE NEEDS A PLOT

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

EVERY P-VALUE NEEDS A PLOT

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



WHAT ELSE CAN WE DO?

- Be scientific
  - "What else could have driven this result?"
  - Corroborate your findings using multiple approaches
    - e.g. multiple HTS technologies, qPCR, explore literature with consideration of their methods, *in silico* modelling…

WHAT ELSE CAN WE DO?

- Understand your methods
  - 🐤 test
  - Plot your data & publish your figures, not (just) p-values
  - Ask the developers!

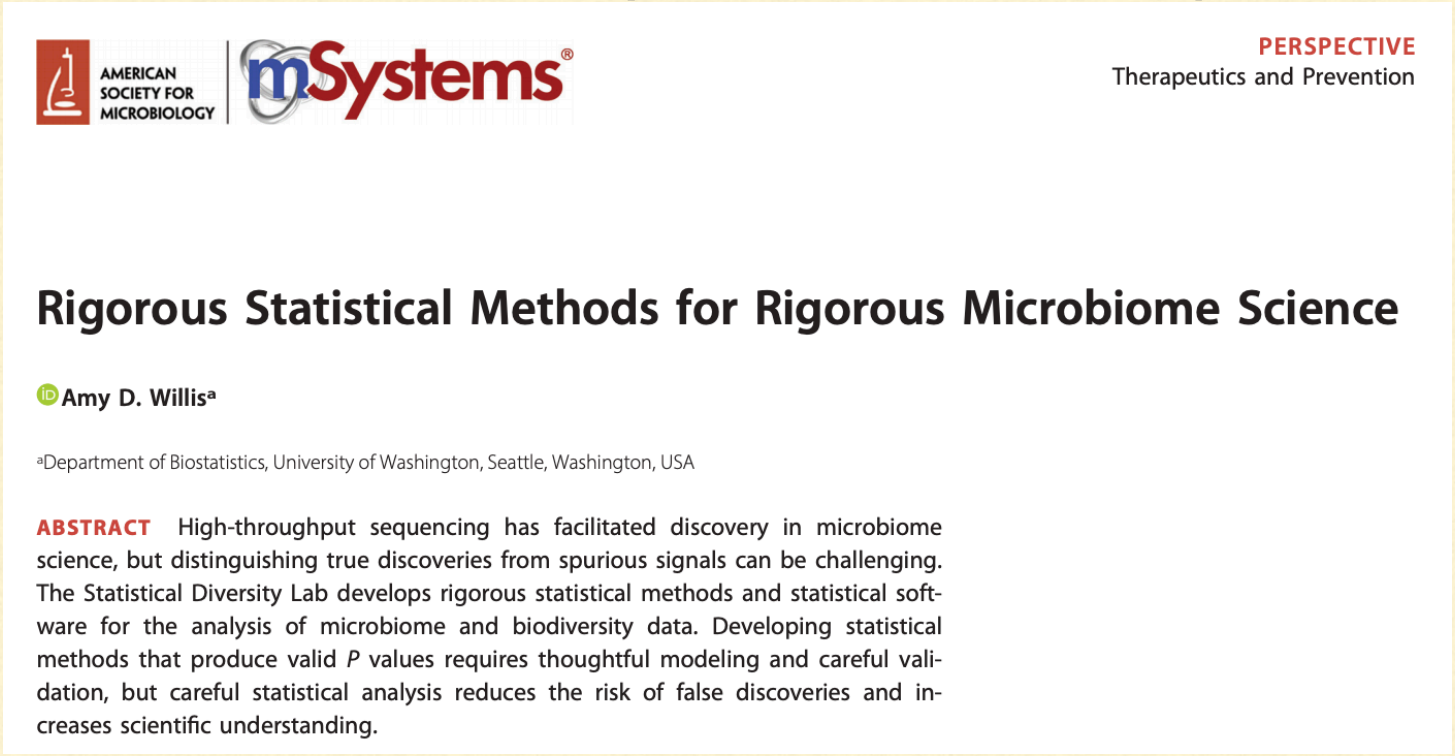
WHAT ELSE CAN WE DO?

- 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

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?

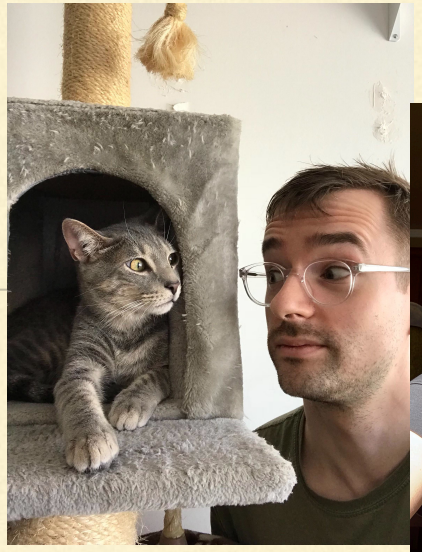
statistical diversity lab

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 & David Clausen
- Tracy & Mihai & Mike & rest of team
  - For making this amazing workshop #STAMPS2022 happen
- YOU!
  - For engaging in reproducible and ethical science

RECAP

RECAP

STATISTICS BOOTCAMP

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Assistant Professor

David Clausen — @davidandacat — PhD Candidate

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 >150 hours of graduate training in applied statistics in ~2 hours
  - Errr… questions?
- Case study

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?