Statistics foundations

Statistical Diversity Lab @ University of Washington
Amy Willis — @AmyDWillis — Associate Professor
Shirley Mathur — PhD Candidate
Sarah Teichman — PhD Candidate
María Valdez — PhD Candidate

Pep talk

Microbial universe

- Y_{ij} = true number of unit j in sample i
  - e.g., there are 7,455,469 16S DNA copies/ml of *S epidermidis* on my finger

😻 Y_{ij} 💸	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

Question of the day

- What can (and can't) we learn about Y_{ij}'s from HTS?

Review: Parameters

- **Parameters** are
  - numerical summaries of a population
  - here, *functions* of Y_{ij}'s

😻 Y_{ij} 💸	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

Review: Parameters

- Examples of microbial parameters
  - ...

Activity: Parameters

Examples of microbial parameters	Non-comparative	Comparative
Summarise one sample		
Summarise multiple samples		

Don’t take this *too* seriously. I made it up, but I think it’s helpful.

Review: Data

- W_{ij} = number of times unit j observed in sample i

🌧️ W_{ij} 😿	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

🦉*The* question🦚: How do we connect the W_{ij}'s to the $Y'_{ij}s$?

Review: Estimators

- Parameters are *unknown*
- We *estimate* parameters using our data
- We call these functions of our data *estimators*

Example:
Shannon diversity

- Shannon diversity is a *parameter*

$$\alpha_i := -\sum_{j=1}^{J} p_{ij} \log p_{ij} \qquad \text{for } p_{ij} := \frac{Y_{ij}}{\sum_{j=1}^{J} Y_{ij}}$$

- A *function* of the true, unknown Y_{ij}'s… thus, a parameter!

Example:
Shannon diversity

- We can *estimate* Shannon diversity

- The most common estimator is the “plug-in” estimator

$$\hat{\alpha}_i := - \sum_{j=1}^{J} \hat{p}_{ij} \log \hat{p}_{ij} \text{ for } \hat{p}_{ij} := \frac{W_{ij}}{\sum_{j=1} W_{ij}}$$

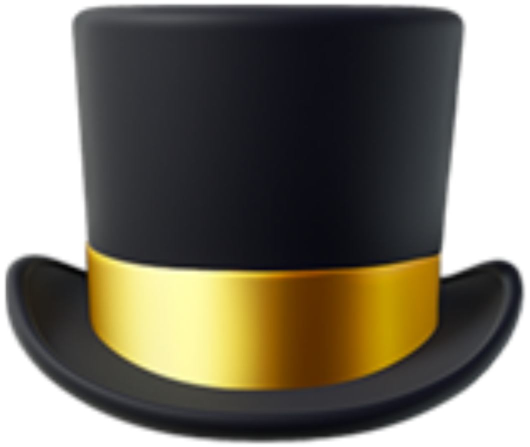
- A *function* of the observed W_{ij}’s… thus, an estimator!

Estimators: notation

- The parameter *Amy*:

Estimators: notation

- An estimator of the parameter *Amy*:



Example: differences in log-ratios

- Here's a different parameter: $\beta_{j,j'}$

average of **treatment** samples' $\log\left(\dfrac{Y_{ij}}{Y_{ij'}}\right)$

minus

average of **control** samples' $\log\left(\dfrac{Y_{ij}}{Y_{ij'}}\right)$

Example: differences in log-ratios

- Having defined the parameter $\beta_{j,j'}$ as

average of **treatment** samples' $\log\left(\dfrac{Y_{ij}}{Y_{ij'}}\right)$

minus

average of **control** samples' $\log\left(\dfrac{Y_{ij}}{Y_{ij'}}\right)$

…come up with an estimator of $\beta_{j,j'}$

Bonus points: *any* justification for your estimator

Example:
differences in log-ratios

Example: differences in log-ratios

- Congratulations! You just came up with a great estimator!
- Under broad assumptions, this estimator is
  - consistent guaranteed to get closer and closer to the "right" answer as you collect more data
  - efficient the lowest variance out of all possible consistent estimators
  - easy to study

Uncertainty

- Even the best estimators are typically *wrong*: $\hat{\beta}_{j,j'} \neq \beta_{j,j'}$
- All estimators have *uncertainty*

Uncertainty

- Our data is *random*
    - Communities vary spatially and temporally
    - Inexhaustive counting processes
    - …
- Our estimators are functions of our data… so our estimators are *random*

Evaluating estimators

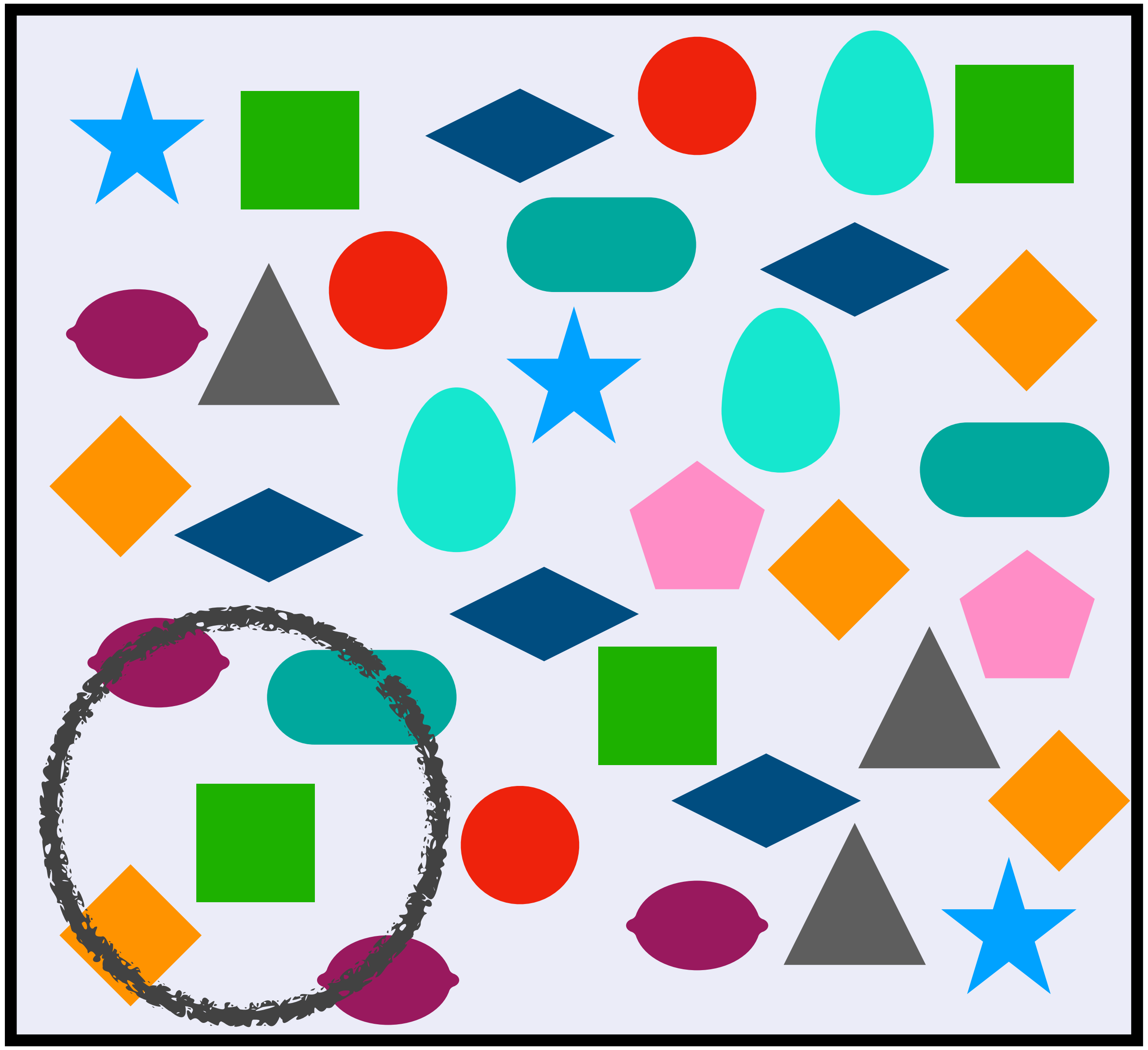
- We want estimators to be
  - Accurate = correct on average = unbiased
  - Precise = usually close to their average = low variance

Bias

- Bias = average value of estimator − true value of parameter
  - e.g., average $\hat{\beta}_{j,j'} - \beta_{j,j'}$

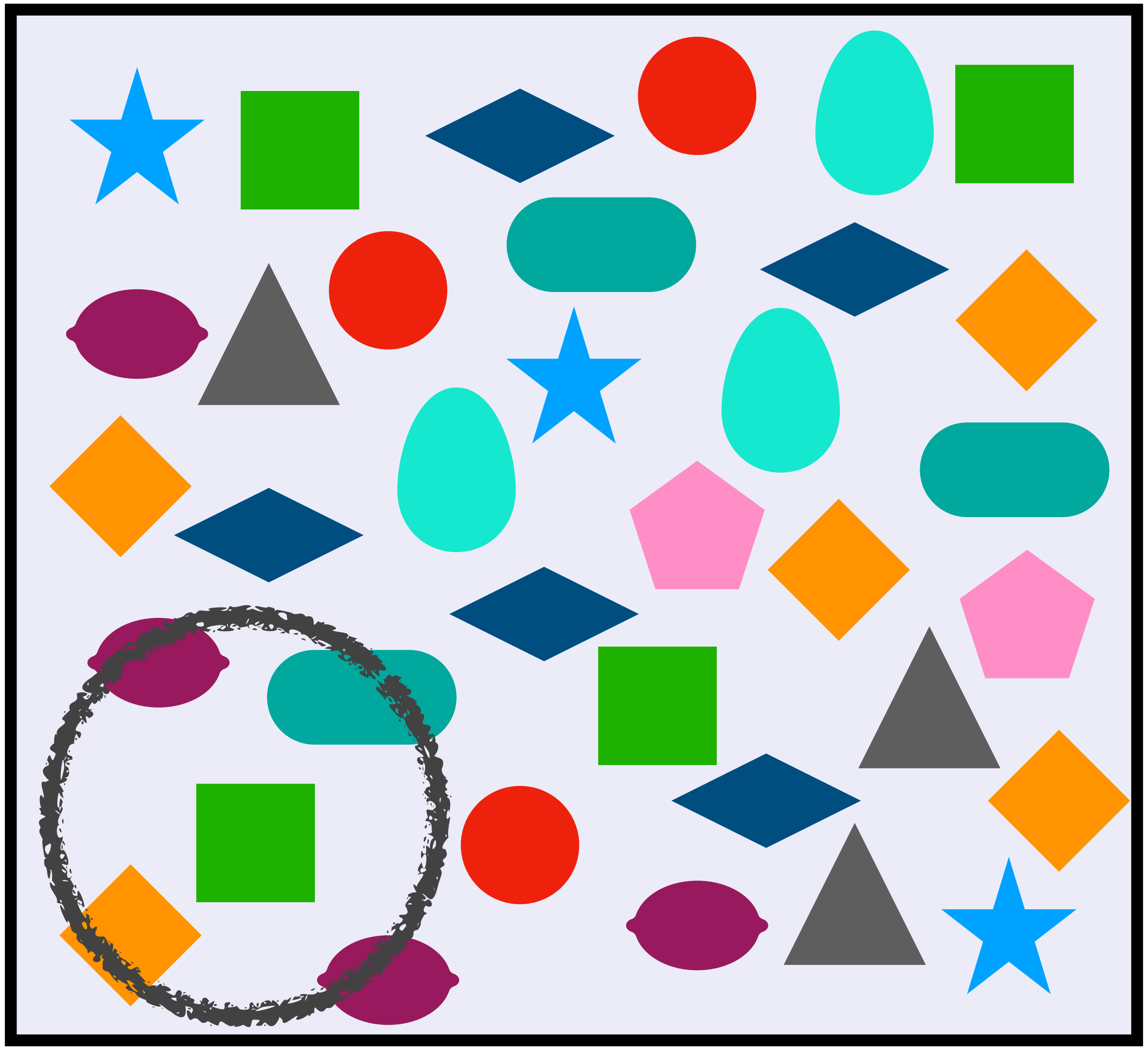
Bias: species richness

- Parameter: total species richness
    - $C = 10$
- Estimator: observed species richness



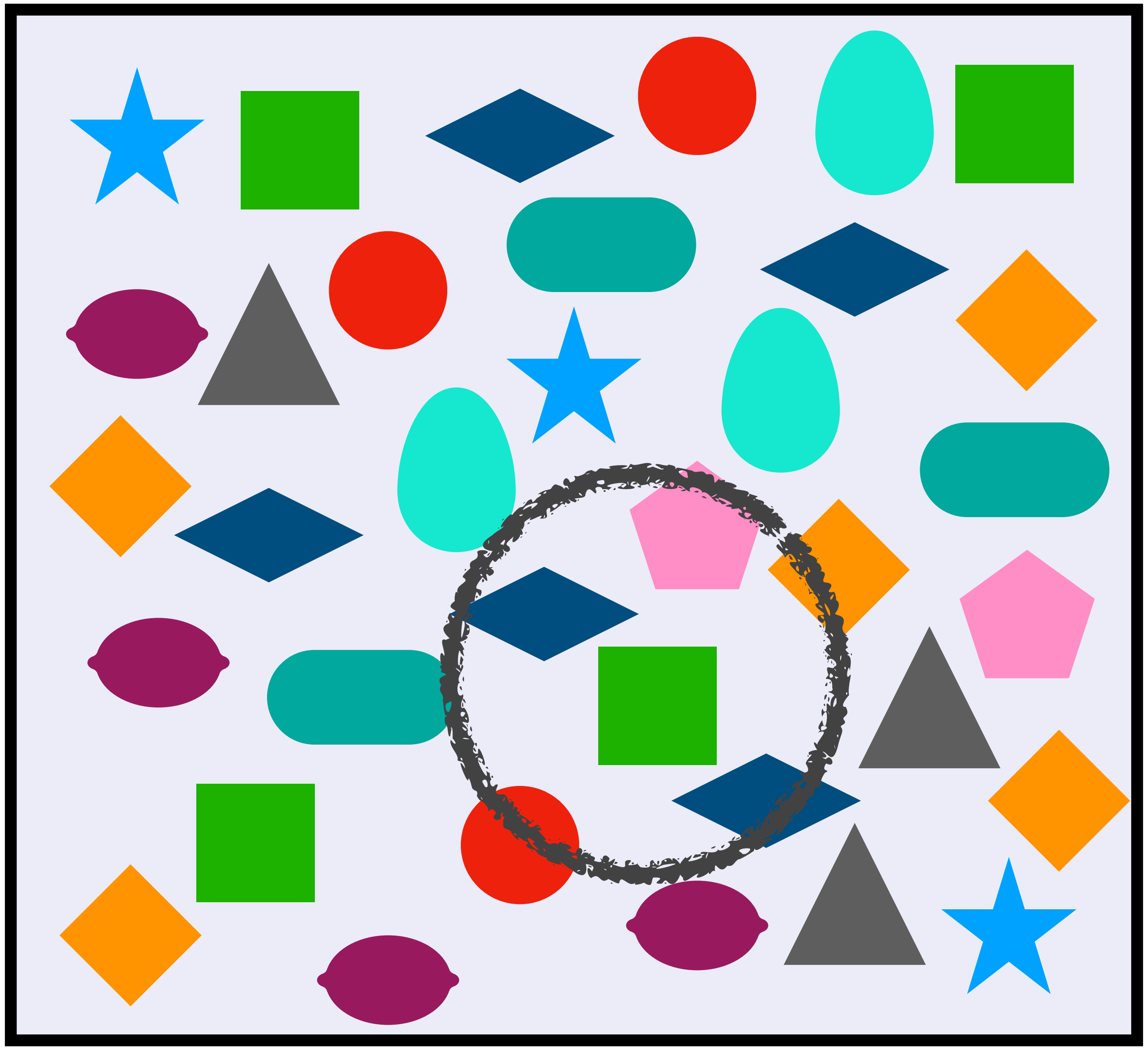
Bias:
species richness

- Parameter: total species richness
  - $C = 10$
- Estimator: observed species richness
  - $\hat{C} = 4$



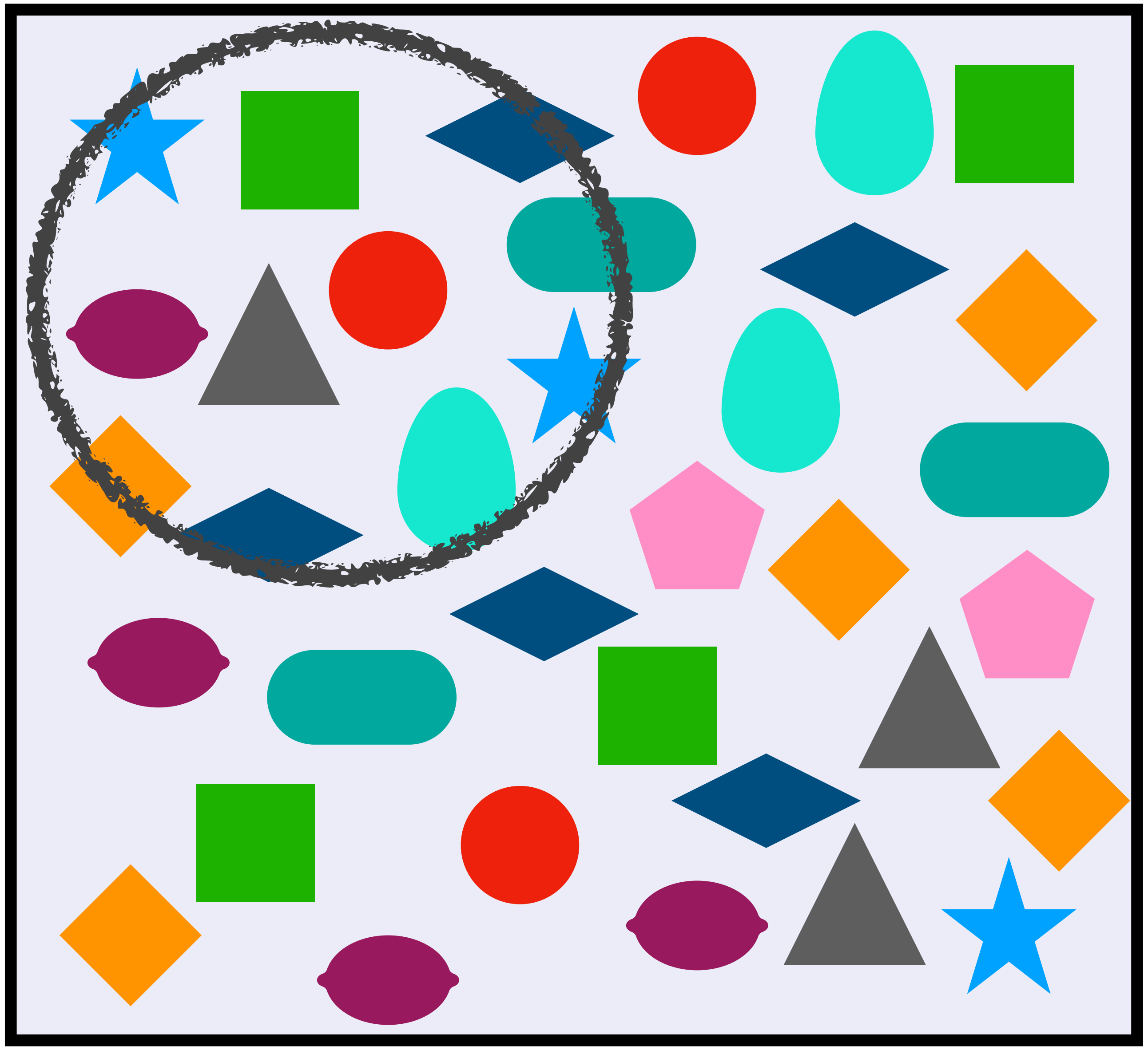
Bias: species richness

- Parameter: total species richness
    - $C = 10$
- Estimator: observed species richness
    - $\hat{C} = 5$



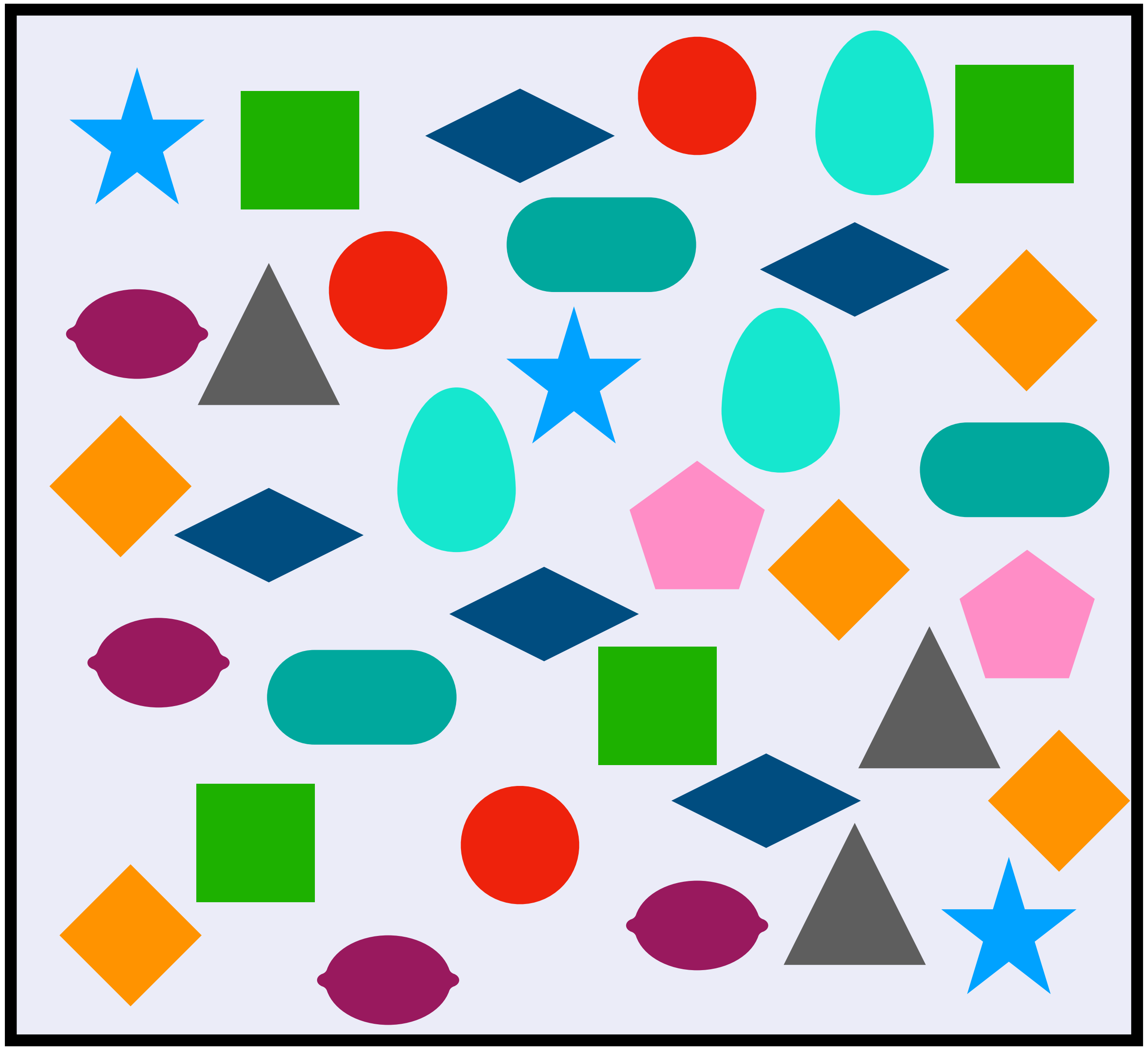
Bias:
species richness

- Parameter: total species richness
  - $C = 10$
- Estimator: observed species richness
  - $\hat{C} = 9$



Bias:
species richness

Observed species richness is *negatively* biased = too small on average

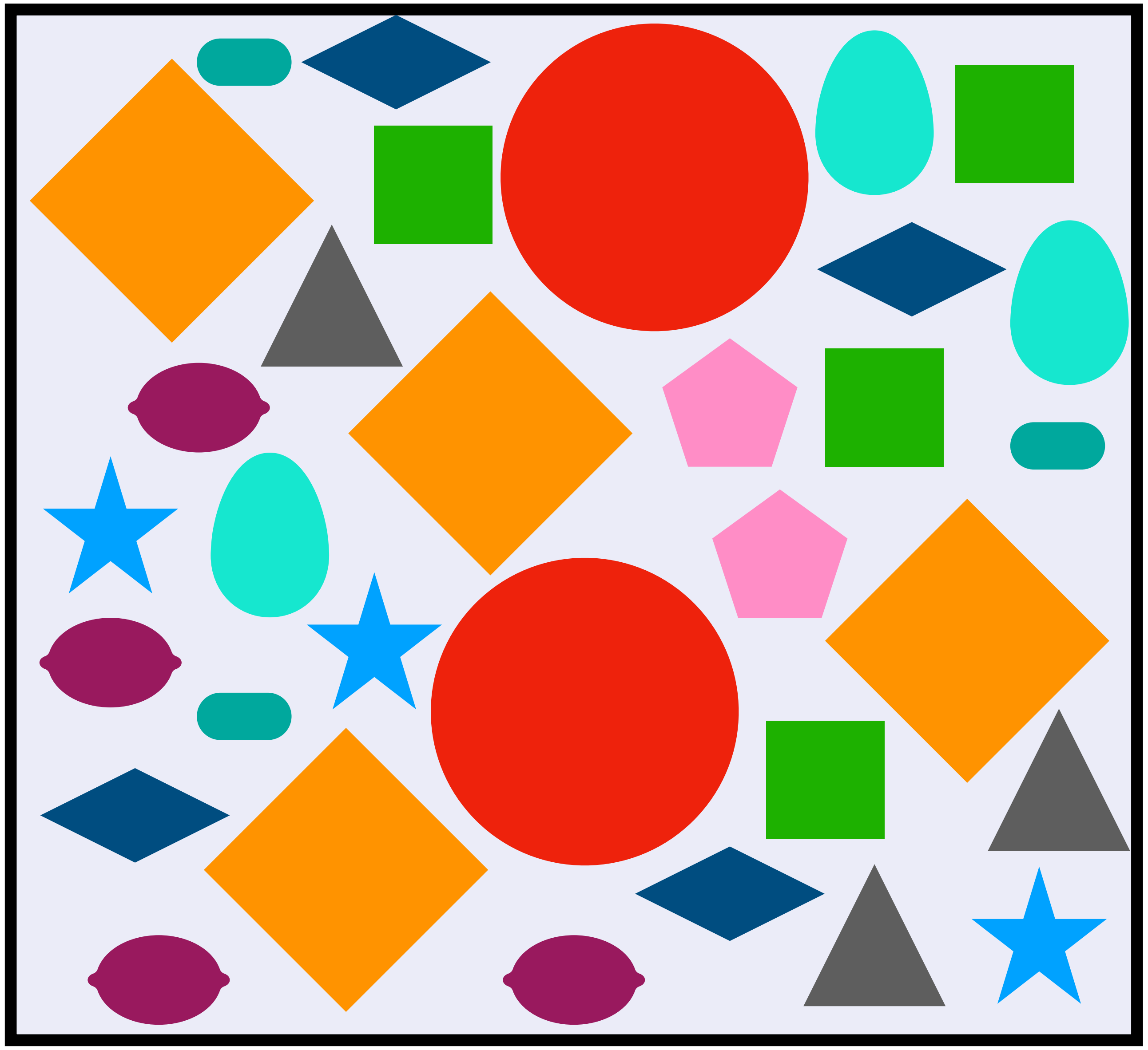


Bias:
relative abundance

- Parameter: true relative abundance of ◆

- Estimator: observed relative abundance of ◆

Activity: Estimate the bias



Hint: there are 32 members of this community

Variance

$$\text{variance} = \text{average of } (\text{estimator} - \text{average value of estimator})^2$$

- If 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 ~27 000 000

Variance

$$\text{variance} = \text{average of } (\text{estimator} - \text{average value of estimator})^2$$

- Hard to compare size of variance relative to estimate itself

Standard deviation = $\sqrt{\text{variance}}$

Variance

$$\text{variance} = \text{average of } (\text{estimator} - \text{average value of estimator})^2$$

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

- Variances are *unknown*
- *True* variance vs *estimate* of variance

Standard error = estimate of the standard deviation

Evaluating estimators

- Bias and variance are two criteria for evaluating estimators
- We can evaluate estimators' bias and variance under different *assumptions*

Evaluating estimators

- To rationally compare methods, we need to be clear about
  - The parameter
  - The estimator
  - The assumptions we want to make

We'll come back to this in specific cases:
diversity, differential abundance…

Examples of assumptions

- We saw all the species that were present
- All species are equally easy to detect
- Amplicon counts follow a zero-inflated Negative Binomial distribution
- …

Examples of assumptions

- Taxa are consistently over/underdetected within a sequencing batch

- Measurements taken from different participants are independent

- The more deeply I sequence, the more likely I am to see something that's present

Identifiability

- Assumptions aren’t bad… they’re *necessary*
- You need to make assumptions to make a parameter *identifiable*

Identifiability

- Why can't we estimate Y_{ij} from W_{ij}? W_{ij} from HTS
- Because the assumptions needed to estimate Y_{ij} from W_{ij} aren't plausible…
  - They don't allow us to *identify* Y_{ij}
- We'll talk about this more this afternoon!

Y = true abundances, W = observed data

The life of a statistician

- Statisticians do the following
  - Choose assumptions
  - Show that the parameter is identifiable using the data + assumptions
  - Derive an estimator
  - Write software & make it useful for others
- These steps allow us to *learn about the universe*, while understanding the *limitations of our methods*

The life of a microbial ecologist

- Choose a parameter meaningful & identifiable
- Choose a sensible estimator reasonable assumptions
- Communicate the estimate of the parameter, and a measure of its uncertainty
- (If appropriate) Perform a valid test about the value of the parameter

Recap

- Everyone here cares about different things…
  - Presence of ARGs
  - Abundance of *Fusobacteria*
  - The diversity of protists
  - …
- These are different *parameters*

Recap

- The difference between parameters and estimators is not widely discussed
- Microbial ecologists may take for granted that there is only one way to estimate parameters…
  - Plug-in estimates
  - Black box estimates

…and are often left out of the conversation about what assumptions are needed and reasonable

The plan

- We've used a couple of different running examples to illustrate specific concepts, but I haven't recommended specific estimators…
  - I will!

The plan

- Now
  - Hypothesis testing
  - Regression models
- This afternoon
  - Abundance
- Tomorrow morning
  - Diversity
  - Experimental design
  - Descriptive stats
  - Miscellanea

Inference

Hypothesis tests

- From lecture so far, you can:
    - Define the parameter you are interested in
    - Construct and critique an estimator of your parameter of interest
- Let’s use these ideas to answer your scientific questions!

Hypothesis tests

- From lecture so far, you can:
  - Define the parameter you are interested in
  - Construct and critique an estimator of your parameter of interest
- Let’s use these ideas to answer your scientific questions!
  - What is the goal of a hypothesis test?

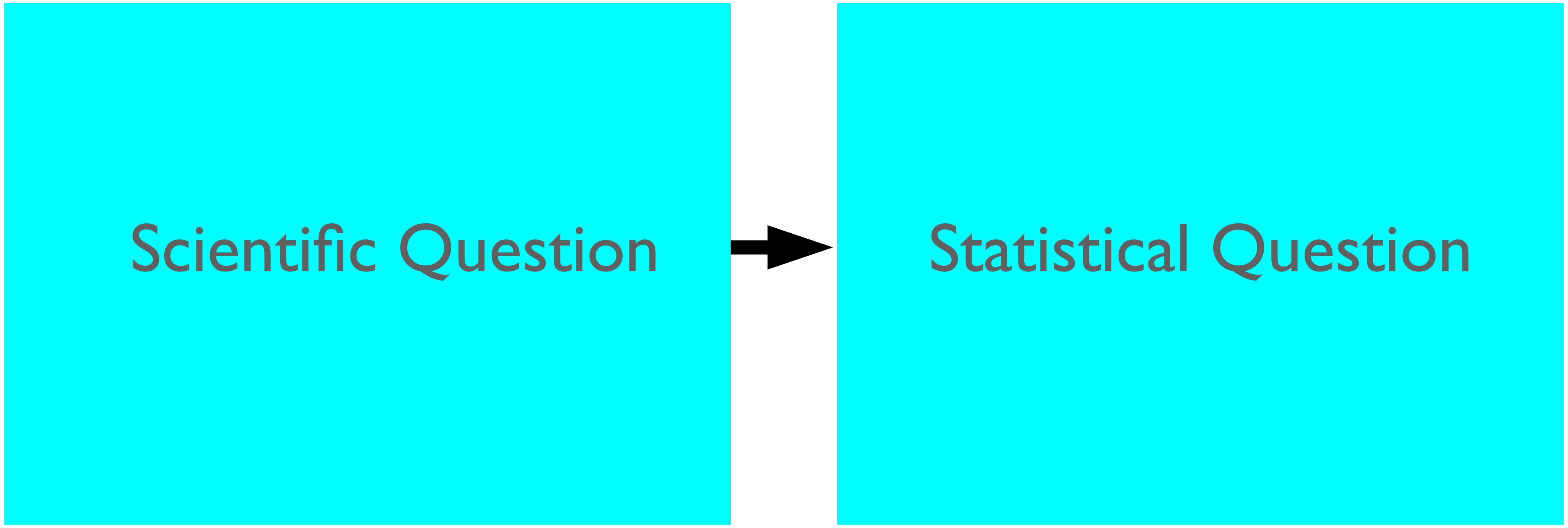
Think, pair, share!

Hypothesis tests

- What is a hypothesis test?
    - Hypothesis: statement about a statistical parameter reflecting a scientific claim
    - Test: way to ask “How much evidence do we have to support our scientific claim?”

Hypothesis

- First translate your scientific question into a statistical question

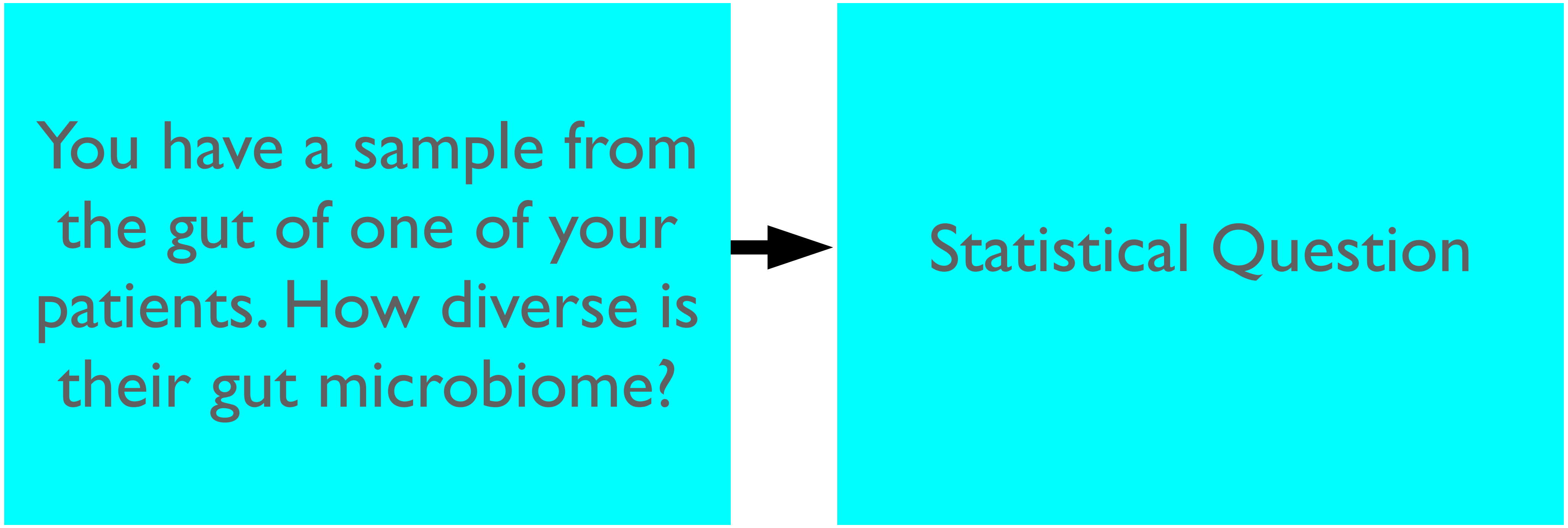


Scientific Question

Statistical Question

Hypothesis

- First translate your scientific question into a statistical question

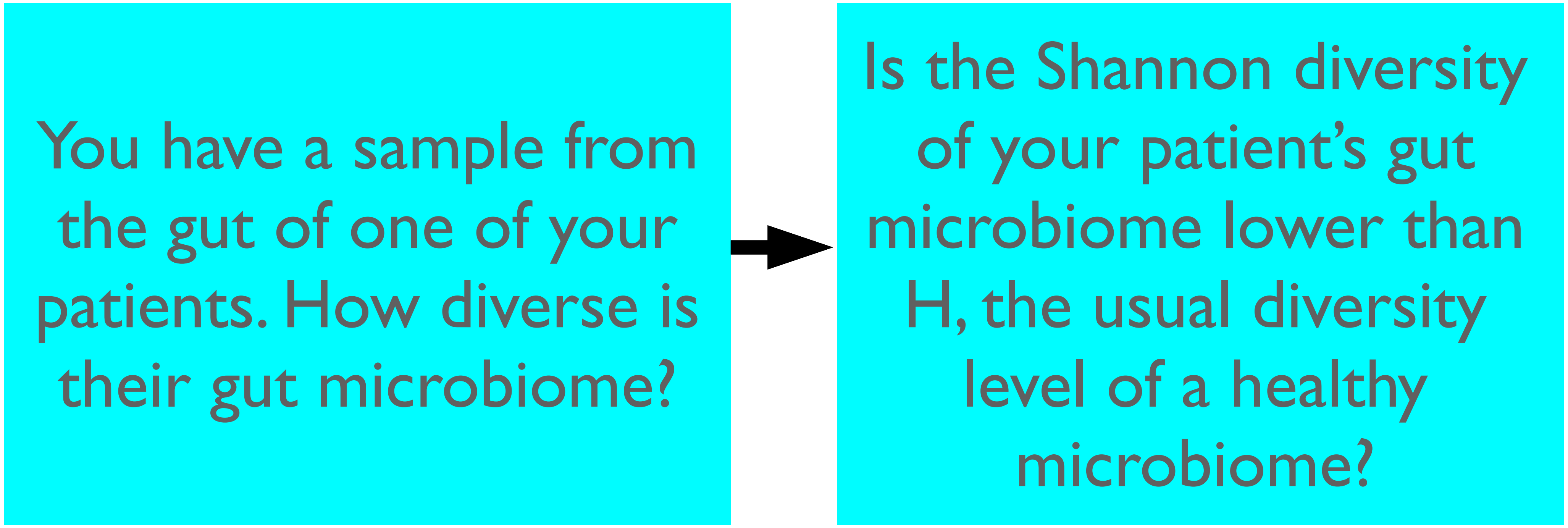


You have a sample from the gut of one of your patients. How diverse is their gut microbiome?

Statistical Question

Hypothesis

- First translate your scientific question into a statistical question

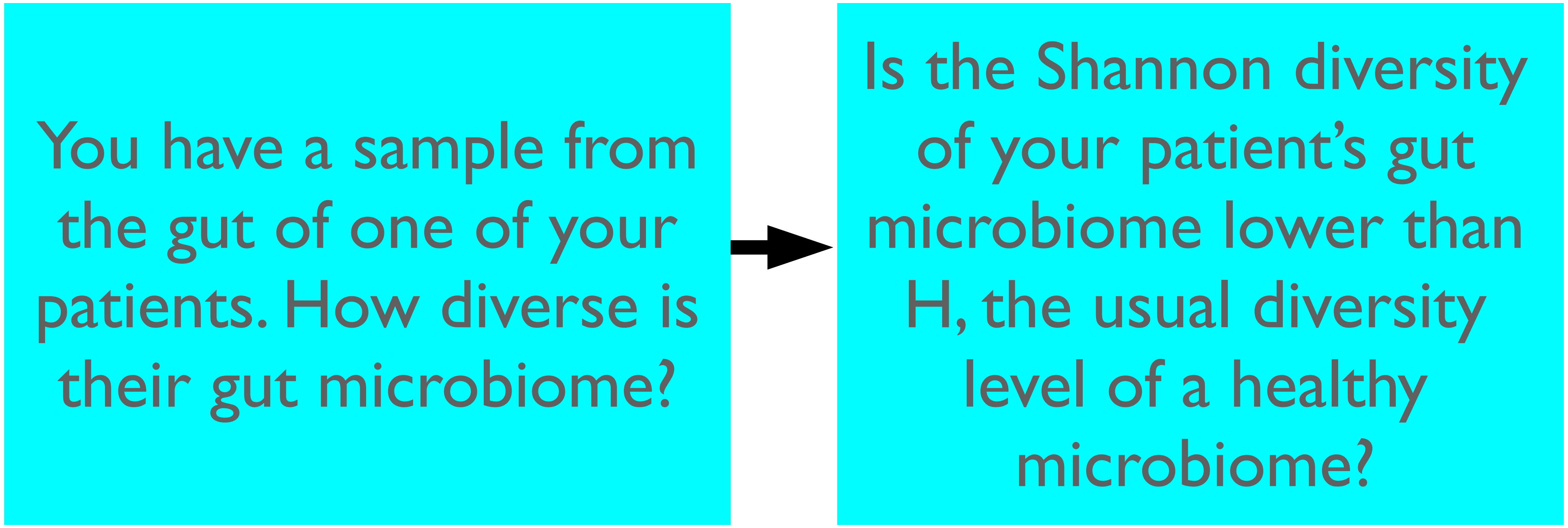


You have a sample from the gut of one of your patients. How diverse is their gut microbiome?

Is the Shannon diversity of your patient's gut microbiome lower than H, the usual diversity level of a healthy microbiome?

Hypothesis

- First translate your scientific question into a statistical question



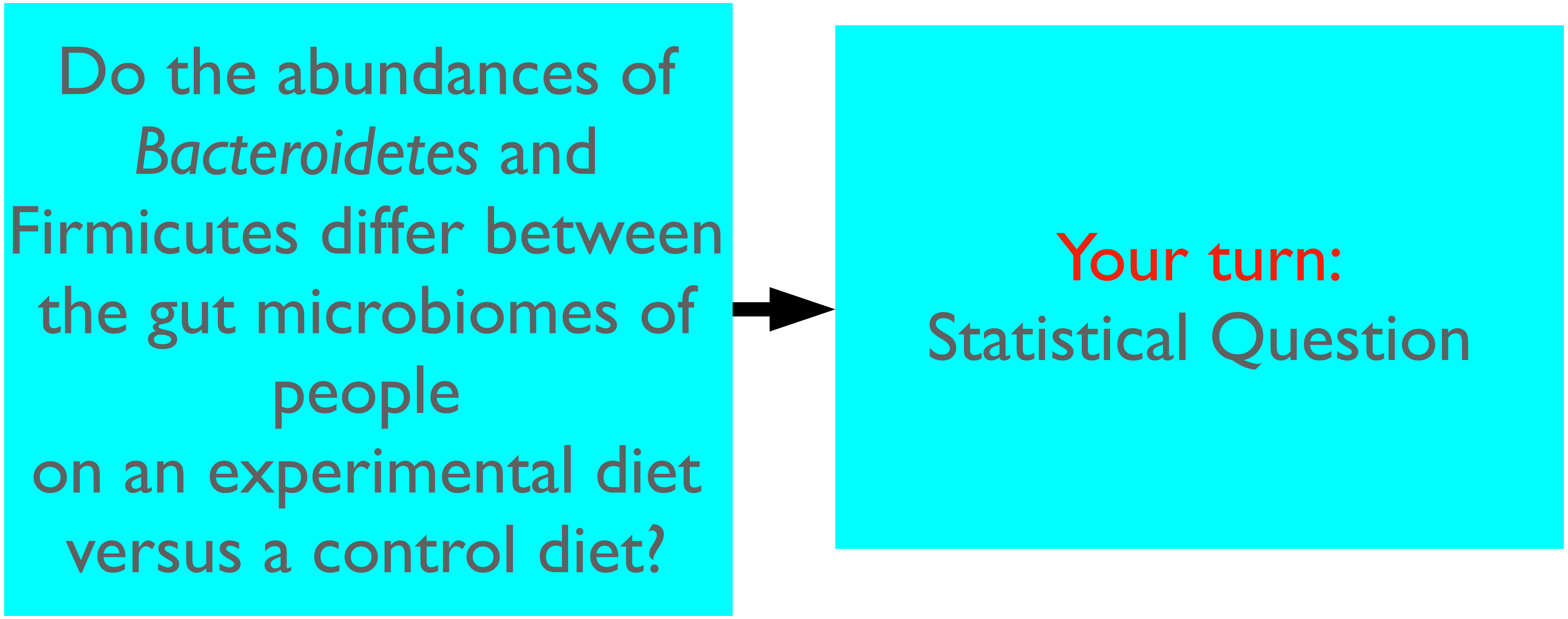
You have a sample from the gut of one of your patients. How diverse is their gut microbiome?

Is the Shannon diversity of your patient's gut microbiome lower than H, the usual diversity level of a healthy microbiome?

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

Hypothesis

- First translate your scientific question into a statistical question

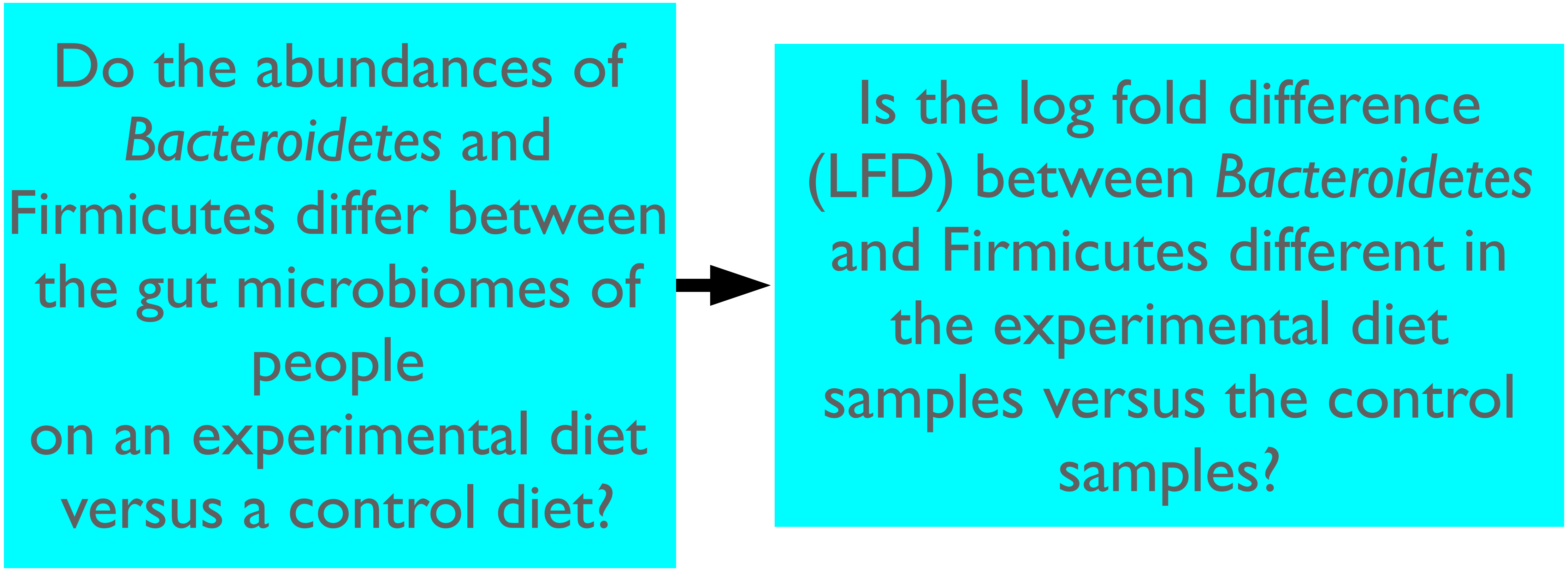


Do the abundances of *Bacteroidetes* and Firmicutes differ between the gut microbiomes of people on an experimental diet versus a control diet?

Your turn:
Statistical Question

Hypothesis

- First translate your scientific question into a statistical question

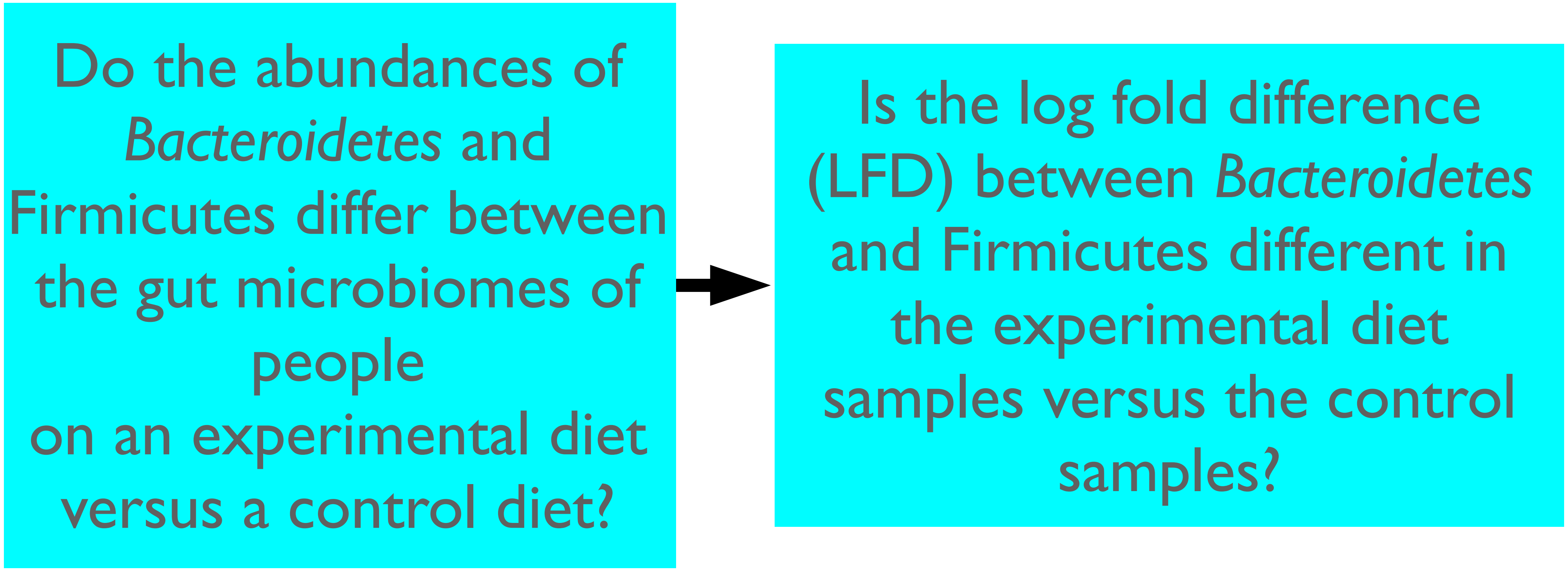


Do the abundances of *Bacteroidetes* and Firmicutes differ between the gut microbiomes of people on an experimental diet versus a control diet?

Is the log fold difference (LFD) between *Bacteroidetes* and Firmicutes different in the experimental diet samples versus the control samples?

Hypothesis

- First translate your scientific question into a statistical question



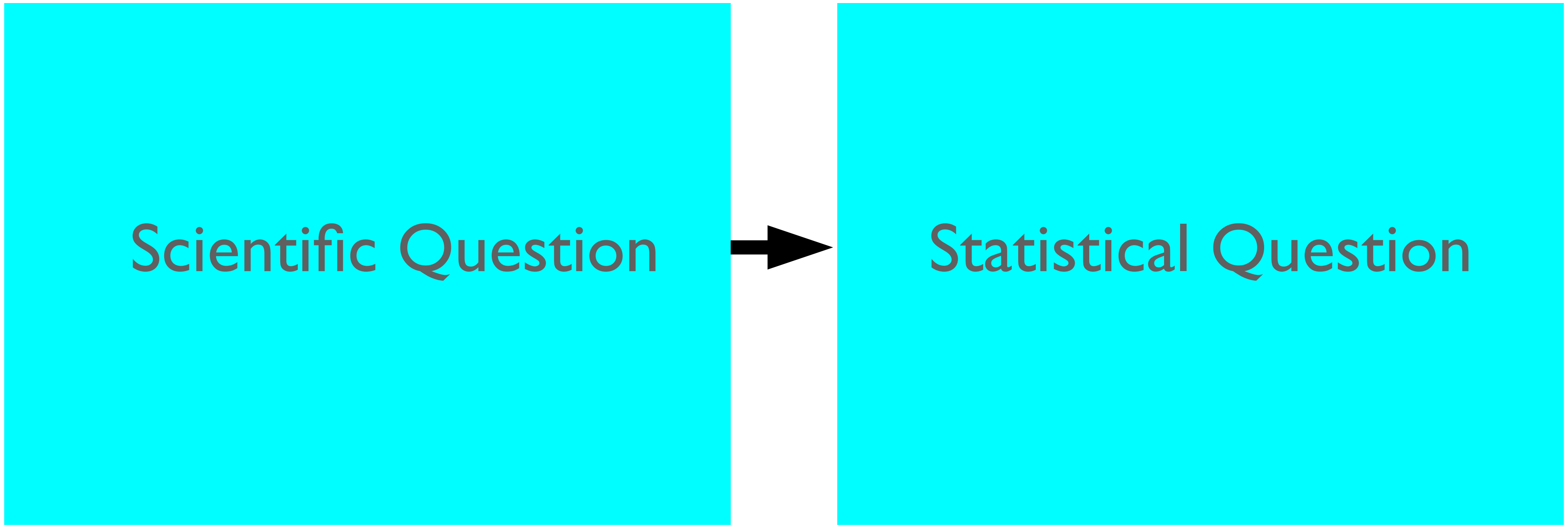
Do the abundances of *Bacteroidetes* and Firmicutes differ between the gut microbiomes of people on an experimental diet versus a control diet?

Is the log fold difference (LFD) between *Bacteroidetes* and Firmicutes different in the experimental diet samples versus the control samples?

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

Hypothesis

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

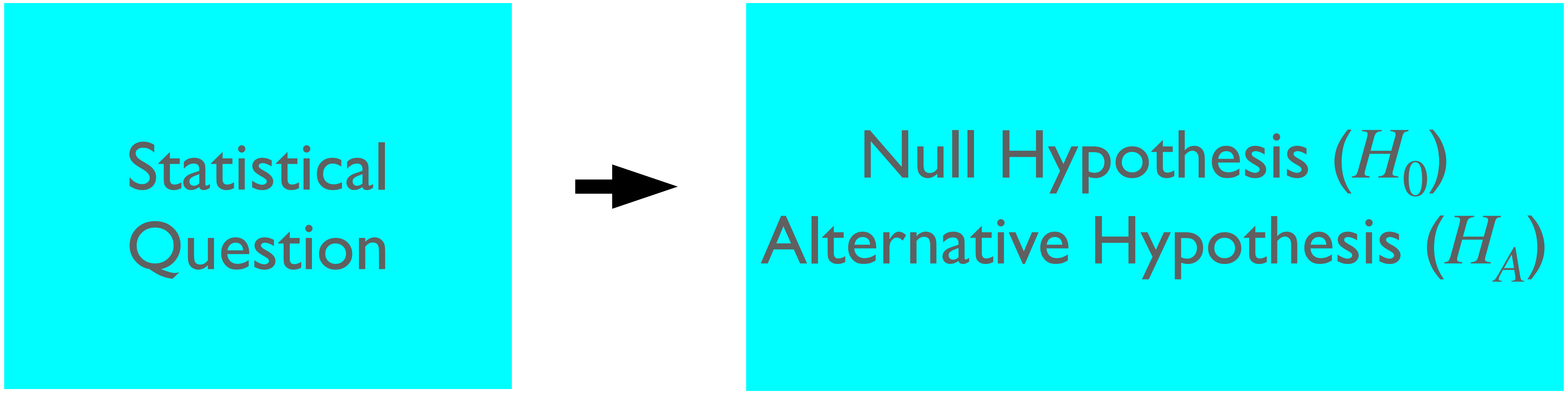


Scientific Question

Statistical Question

Hypothesis

- Your statistical question will inform your hypotheses
  - Null hypothesis (H_0): neutral 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): neutral statement about parameter

  - Alternative hypothesis (H_A): scientifically interesting statement about parameter (usually), the opposite of the null hypothesis

Is the LFD between *Bacteroidetes* and Firmicutes different in the experimental diet samples versus the control diet samples?

→

H_0: the LFD is the same for both groups
H_A: the LFD 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 the LFDs are the same for the experimental and control diets.

2. Fail to reject the null hypothesis.

We do not have enough evidence to reject the hypothesis that the LFDs are the same for the experimental and control diets.

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

$$D = \text{Mean LFD}_{\text{Experimental}} - \text{Mean LFD}_{\text{Control}}$$

- Test statistic often takes this form:

0

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

$$\sqrt{\text{Var(D)}}$$

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

Large difference between estimate and null hypothesized value and/or high certainty in estimate cause larger t

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

Small difference between estimate and hypothesized value and/or low certainty about estimate cause smaller t

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

Suppose your estimated standard error is half of the true standard error. What happens to your test statistic?

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

You’ve mistakenly doubled your test statistic!

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

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

- Your friendly local statistician has come up with an exciting new procedure to test any hypothesis you want. It looks like this:

```
my_test <- function(data) {
      return(p = 0.001)
}
```

- Is this a p-value? Why or why not?
  - Think, pair, share!

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, 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 LFD of Bacteroidetes to Firmicutes between the experimental diet and the control diet as or more extreme than the difference in our sample, if there is no difference in the LFD across our control and experimental unit populations, is 2%.

Therefore, we reject our null hypothesis.

Testing

- Coming back to earlier scenario!
- Your friendly local statistician has come up with an exciting new procedure to test any hypothesis you want. It looks like this:

```
my_test <- function(data) {
        return(p = 0.001)
}
```

- Is this a p-value? Why or why not? Has your reasoning changed?
    - Think, pair, share!

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

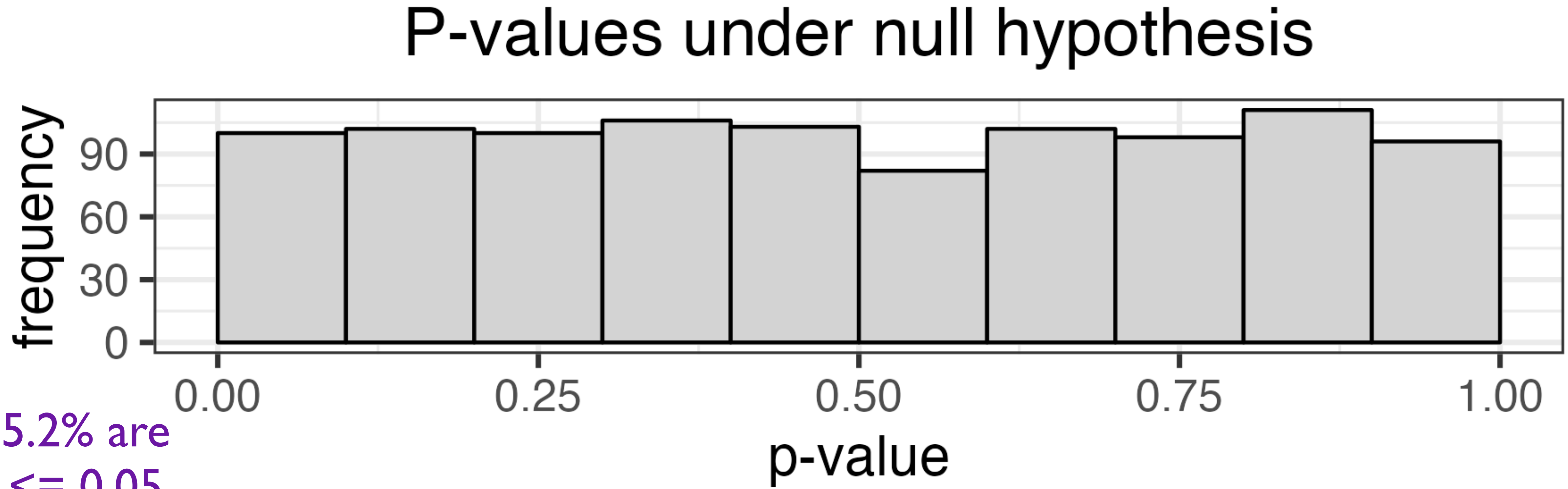
- A valid hypothesis test will reject the null hypothesis exactly $\alpha \times 100\,\%$ of the time **when it is true**
  - Less often = understating your evidence against H_0
    - your test is conservative
  - More often = overstating your evidence against H_0
    - your test is anticonservative

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

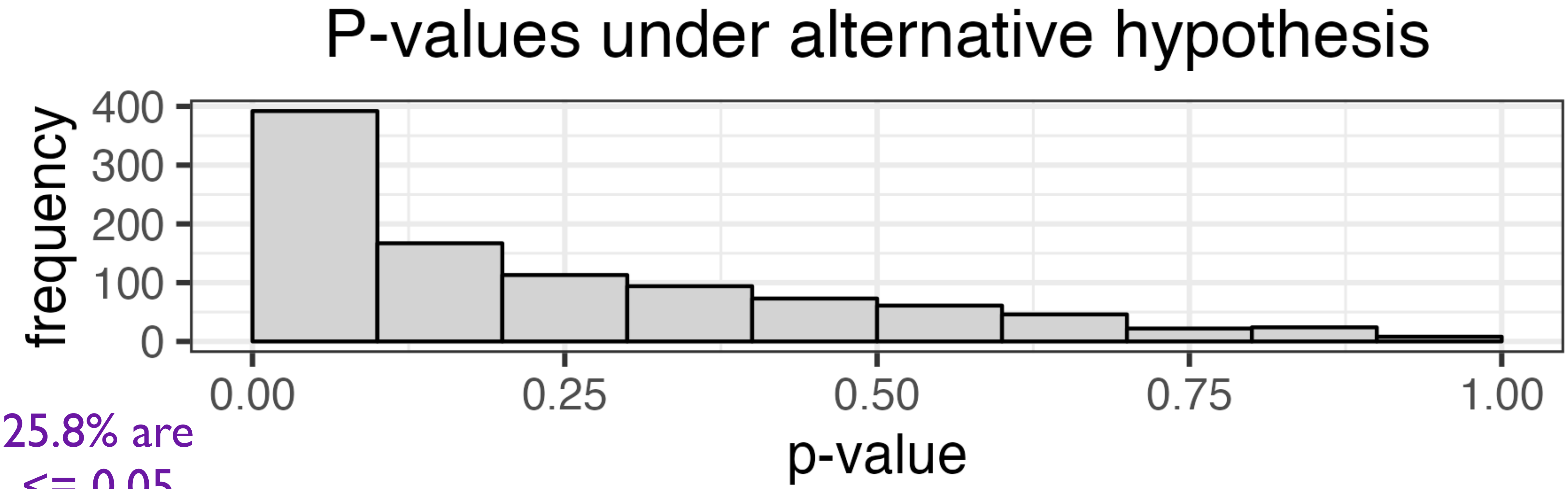


5.2% are <= 0.05

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



25.8% are <= 0.05

Valid hypothesis test

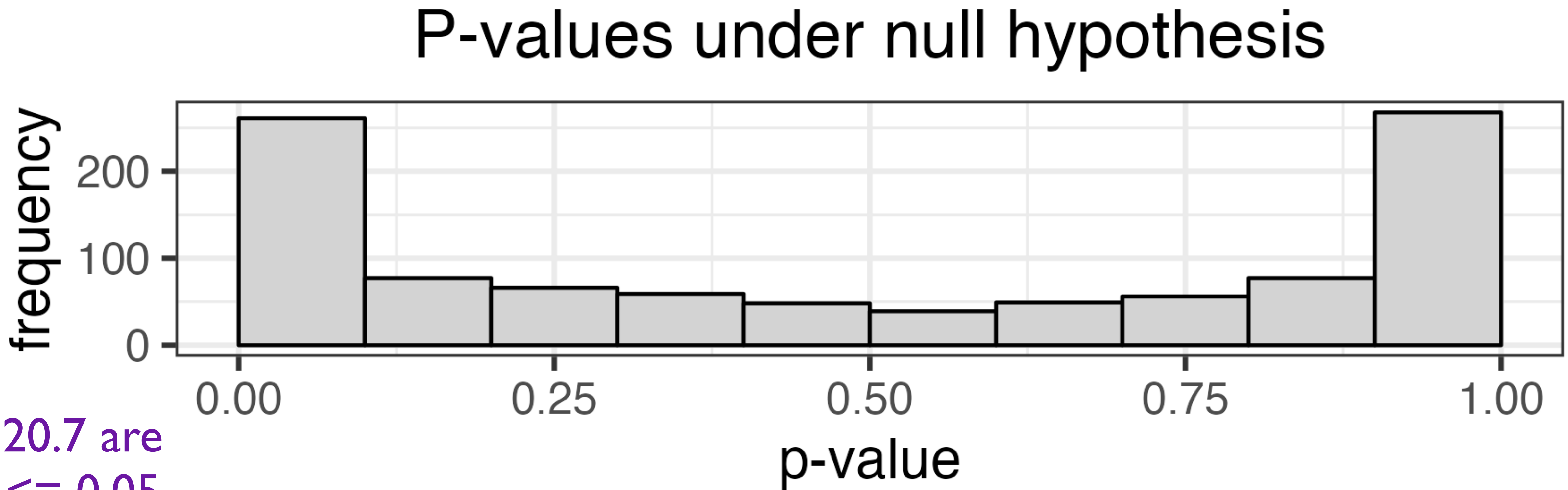
- Why might a hypothesis test be invalid?
  - Hint: recall our discussion of standard errors!

Model misspecification

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

Model misspecification

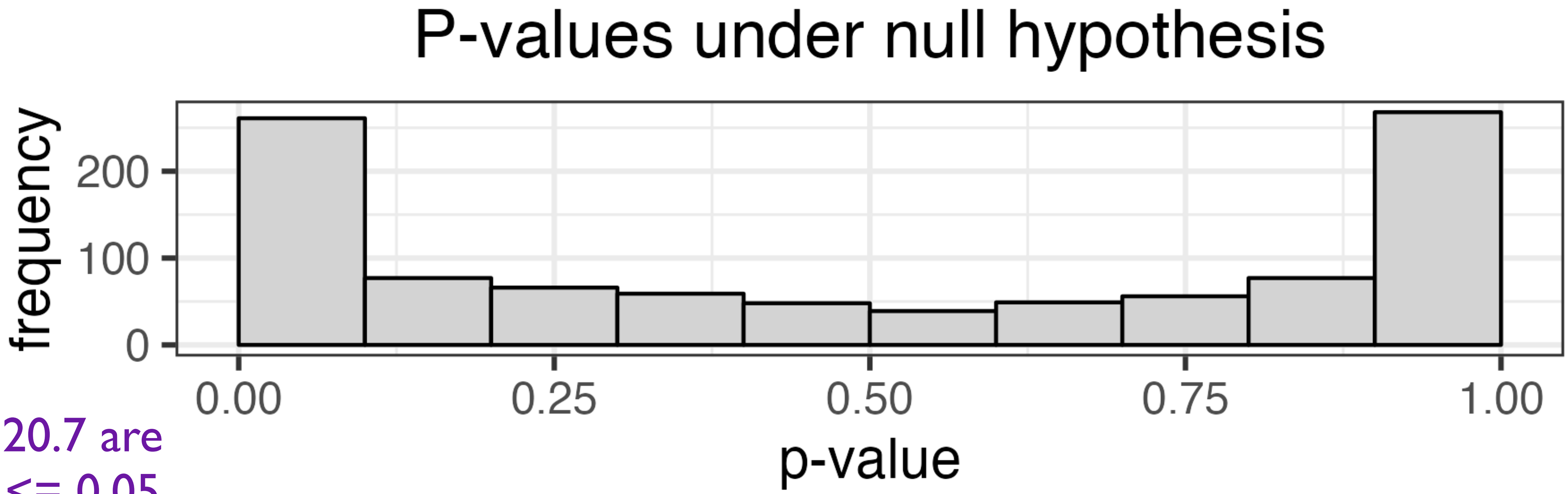
- What if we have a bad model, causing us to mistakenly estimate a standard error that is 1/2 the true standard error?



20.7 are <= 0.05

Model misspecification

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



20.7 are <= 0.05

Type I error

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

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 1 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 1 error

    - **False Discovery Rate (FDR):** the expected proportion of type 1 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 1 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 1 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

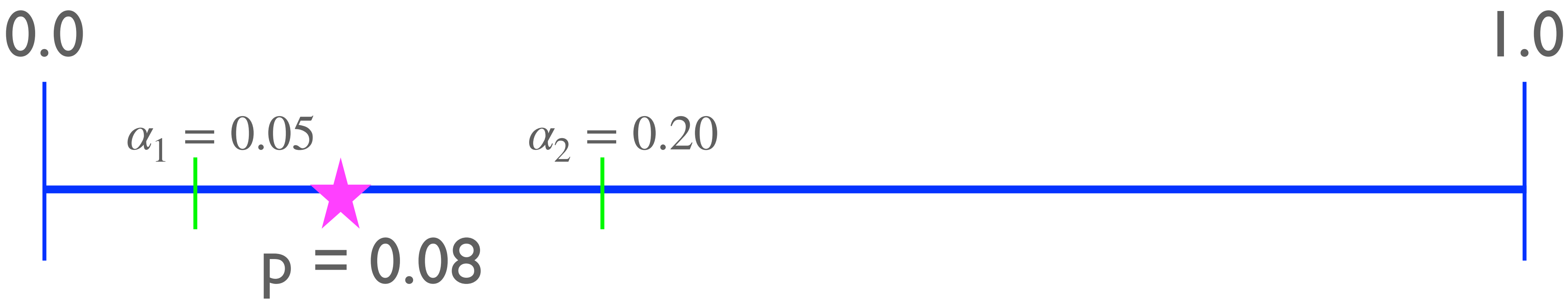
- Our alpha level specifies the probability of a Type 1 error
- We can also commit Type 2 errors: “false negatives”
- Power: probability of correctly rejecting the null hypothesis, **when it is false**
  - 1 - probability of type 2 error

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?

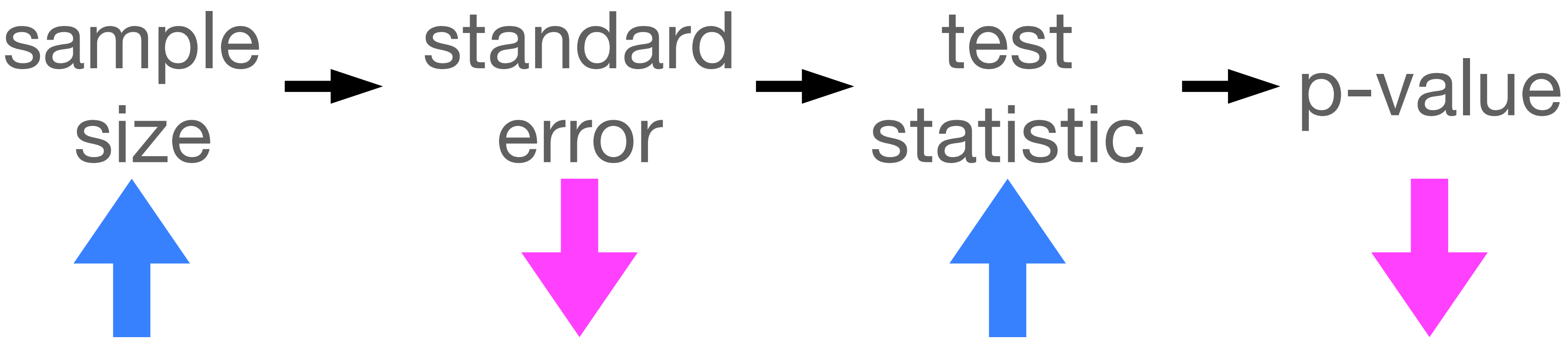


Type 2 Error & Power

- We can increase power by increasing our alpha level (increasing Type I error rate)
    - **Exercise:** Why?
- We can increase power *without sacrificing Type 1 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 1 error* by increasing our sample size (if we have the $$$)



sample size → standard error → test statistic → p-value

Questions?

Comparing populations of samples

Choosing sensible comparative parameters

Parameters

- You choose!
- Considerations
  - Identifiability
  - Meaningfulness = make sensible comparisons

Comparative parameters

- How do we compare populations of samples?
    - Regression models
- How do we choose what regression model?
    - Don’t! Choose parameters!
- What are sensible comparisons to make?
    - Consider adjustment sets

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

Regression models

- Regression models take the form

  functional of outcome variable = function of predictor variables

- e.g.,

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

    - $\hat{\beta}_0$ is an estimate of the average diversity in seawater environments

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

Regression models

- Example of regression model: 🦤 radEmu 🦤

$$\log \text{expected counts}_{ij} = s_i + e_j + \beta_{0j} + \beta_{1j}X_{i1} + \ldots + \beta_{pj}X_{ip}$$

- $\hat{\beta}_{kj}$ is an estimate of the log fold difference in the absolute abundance of taxon j between environments that differ by 1 unit in $X_{\cdot k}$ but are alike in $X_{\cdot 1}, \ldots, X_{\cdot k-1}, X_{\cdot k+1}, \ldots, X_{\cdot p}$, relative to the median of these log fold differences across all taxa

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 of a regression model: betta

$$\text{mean } C_i = \beta_{0j} + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

  - $\hat{\beta}_k$ is an estimate of the difference in the true species richness in environments that differ by 1 unit in $X_{\cdot k}$ but are alike in $X_{\cdot 1}, \ldots, X_{\cdot k-1}, X_{\cdot k+1}, \ldots, X_{\cdot p}$

https://github.com/adw96/breakaway

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
    - Summarise the distribution (random variable ➡️ summary)
    - 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

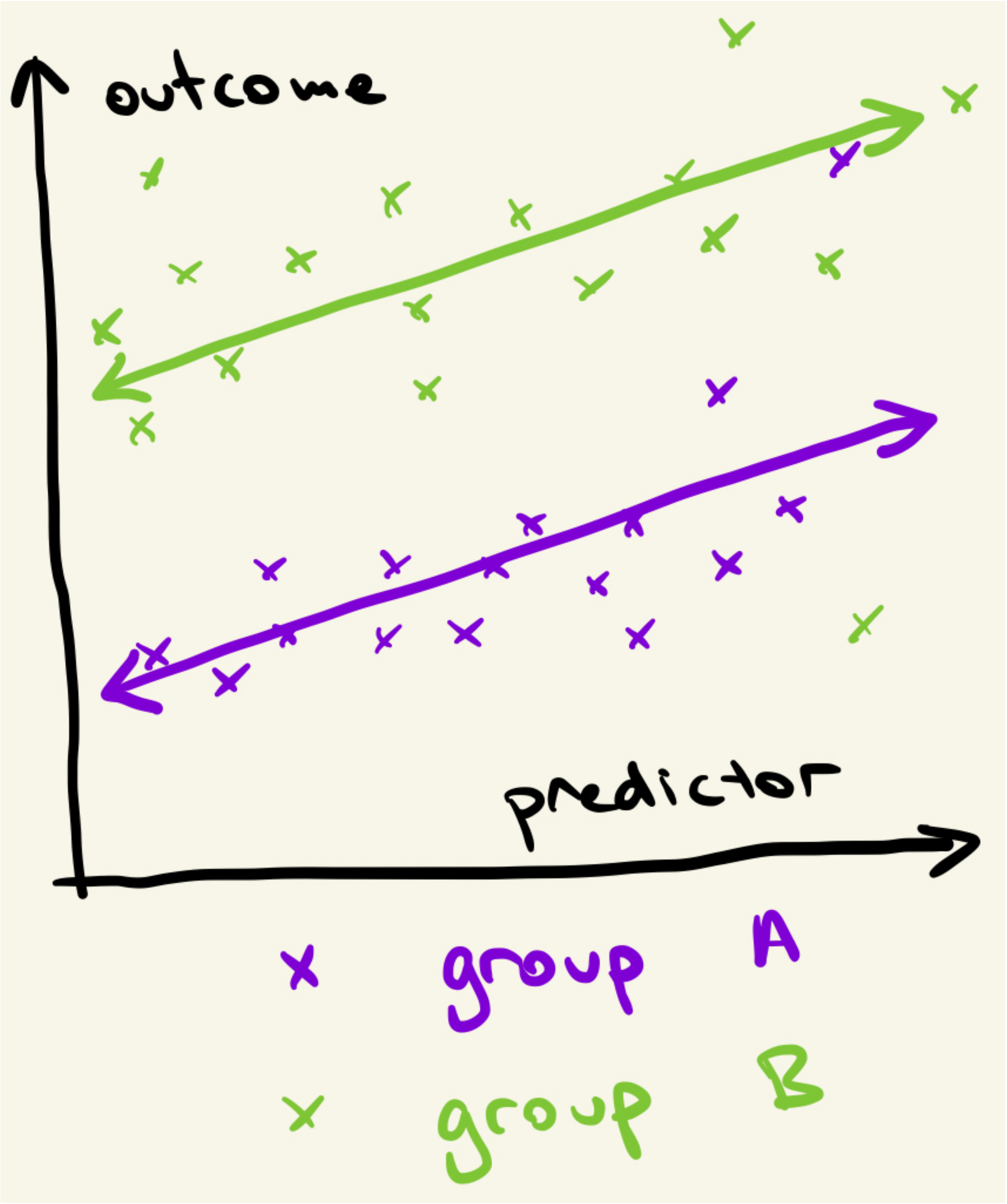
1. Predictor of interest
    - The main thing you set out to study
    - Always include

Types of variables

2. Precision variables

   - Associated with outcome
   - 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

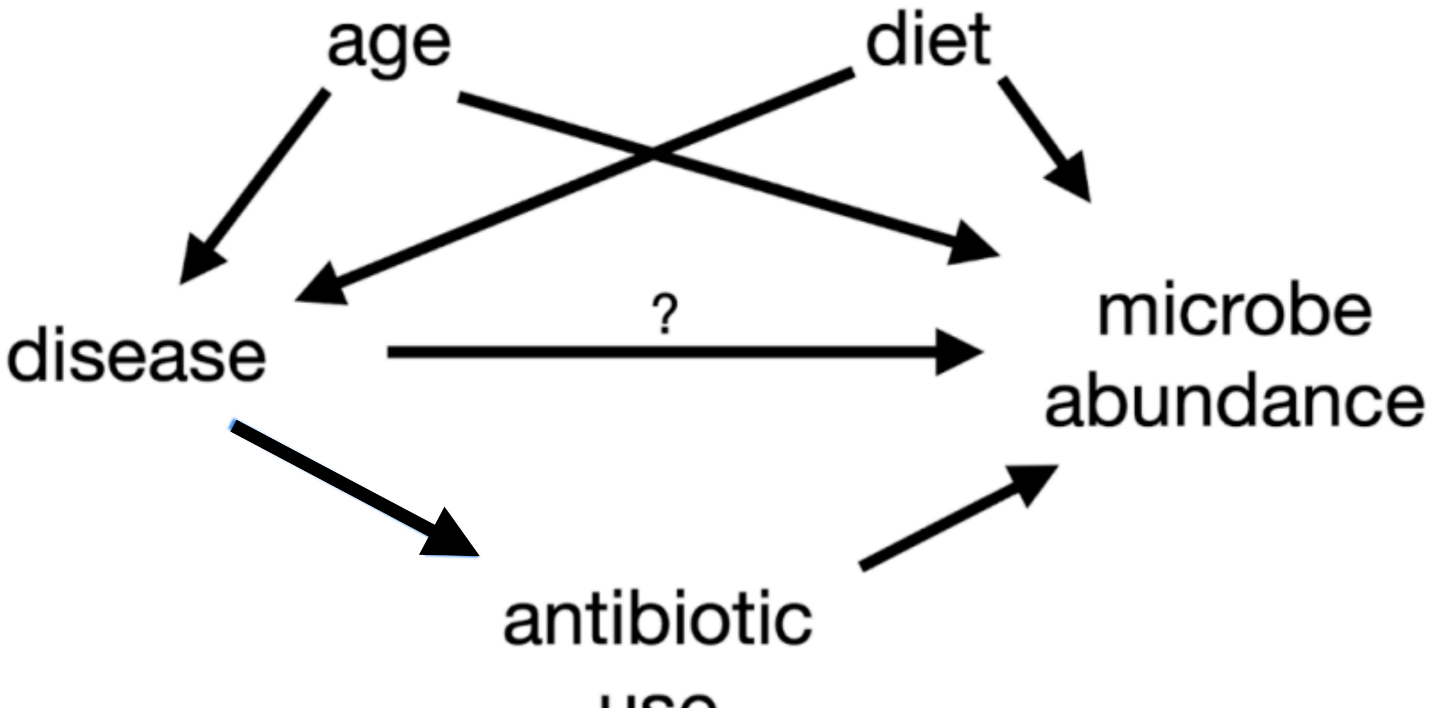
3. Confounders

- “Common causes” of **predictor of interest** and **outcome**
- None of the following are confounders for true microbial abundances Y_{ij}
  - Batch
  - Sequencing technology
  - Any measurement variables (depth…)
- Variables associated with the measurement process *cannot* be causally associated with outcome
- “Confounders” is more often misused than correctly used

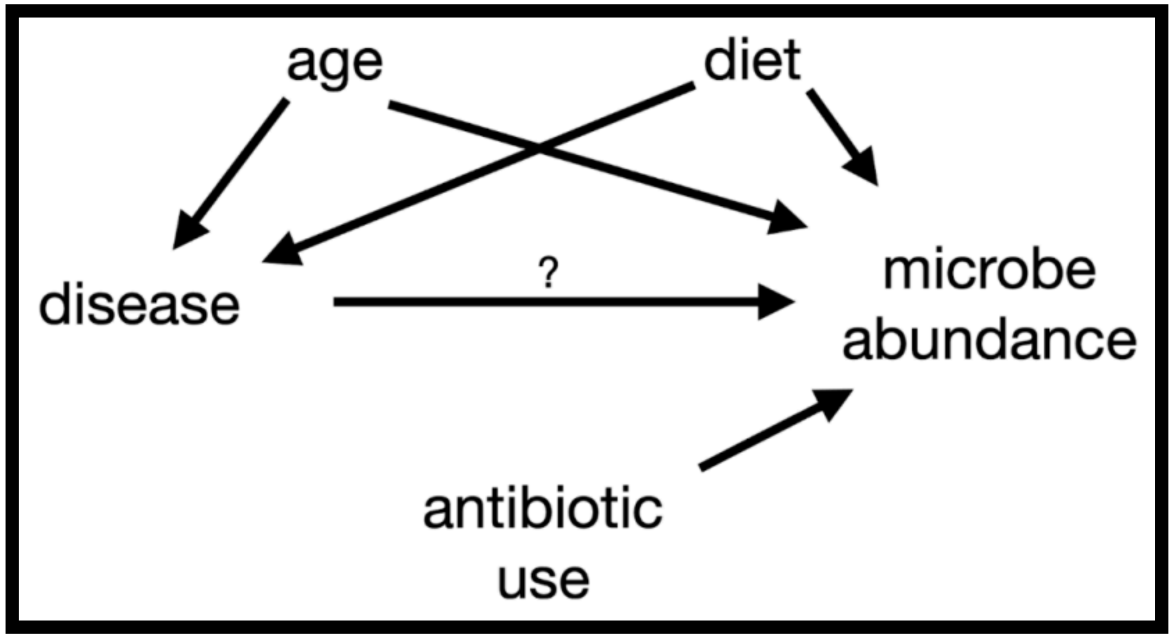
Types of variables

- “Common causes” requires you to write down causal assumptions

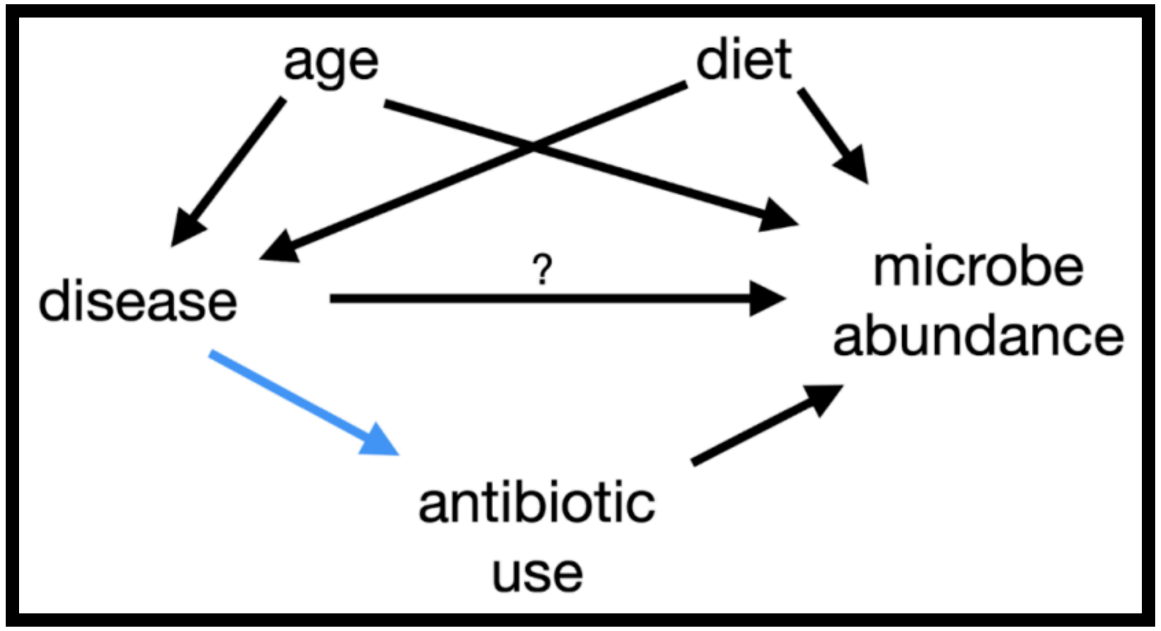
- Causal assumptions = a hypothesized list of causes and outcomes



Adjustment sets



Adjust for: age, diet, antibiotic use



Adjust for: age, diet

- Even if not attempting “causal inference,” write down causal assumptions to choose adjustment sets
  - https://www.r-causal.org/chapters/05-dags
  - dagitty::adjustmentSets()

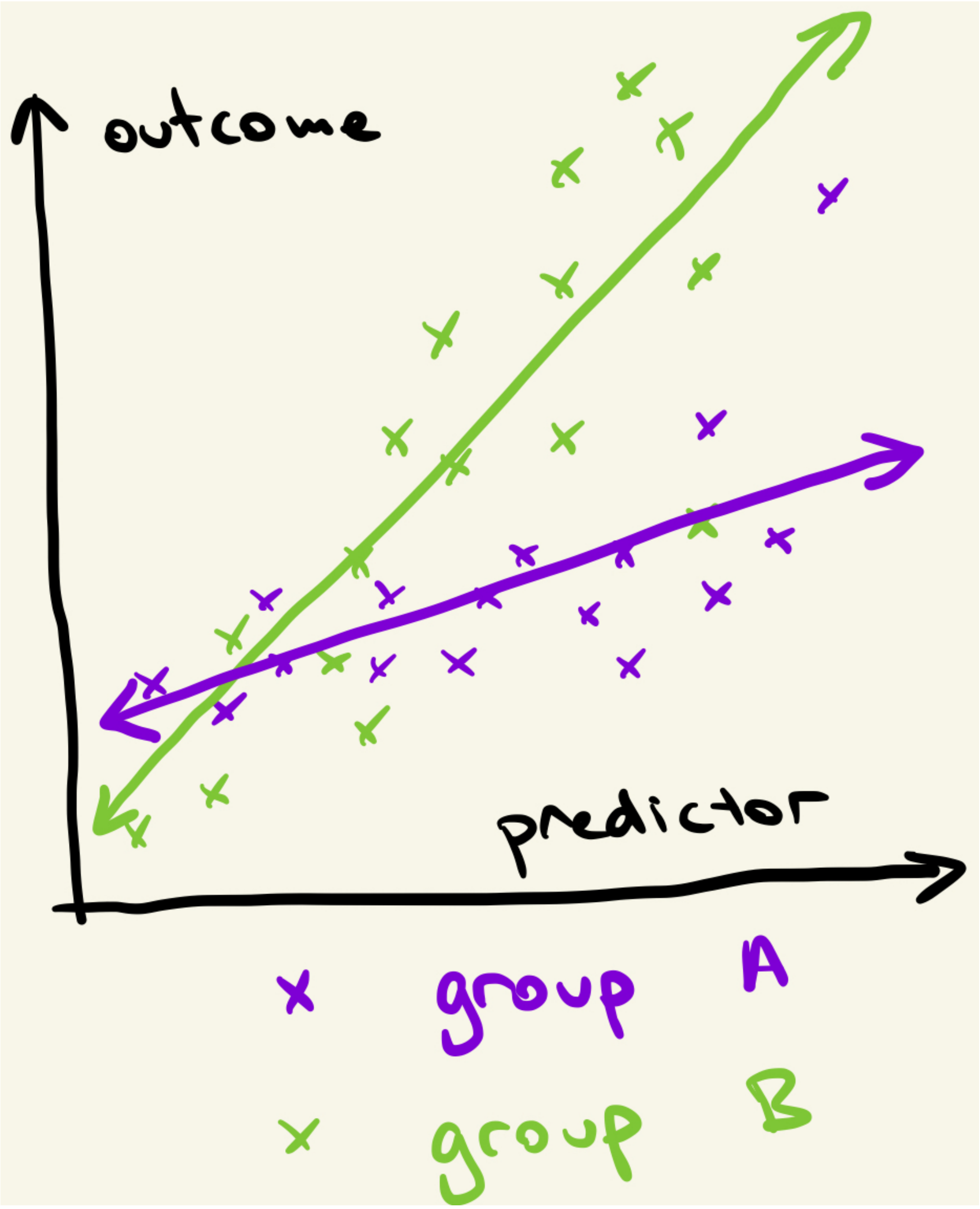
Types of variables

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

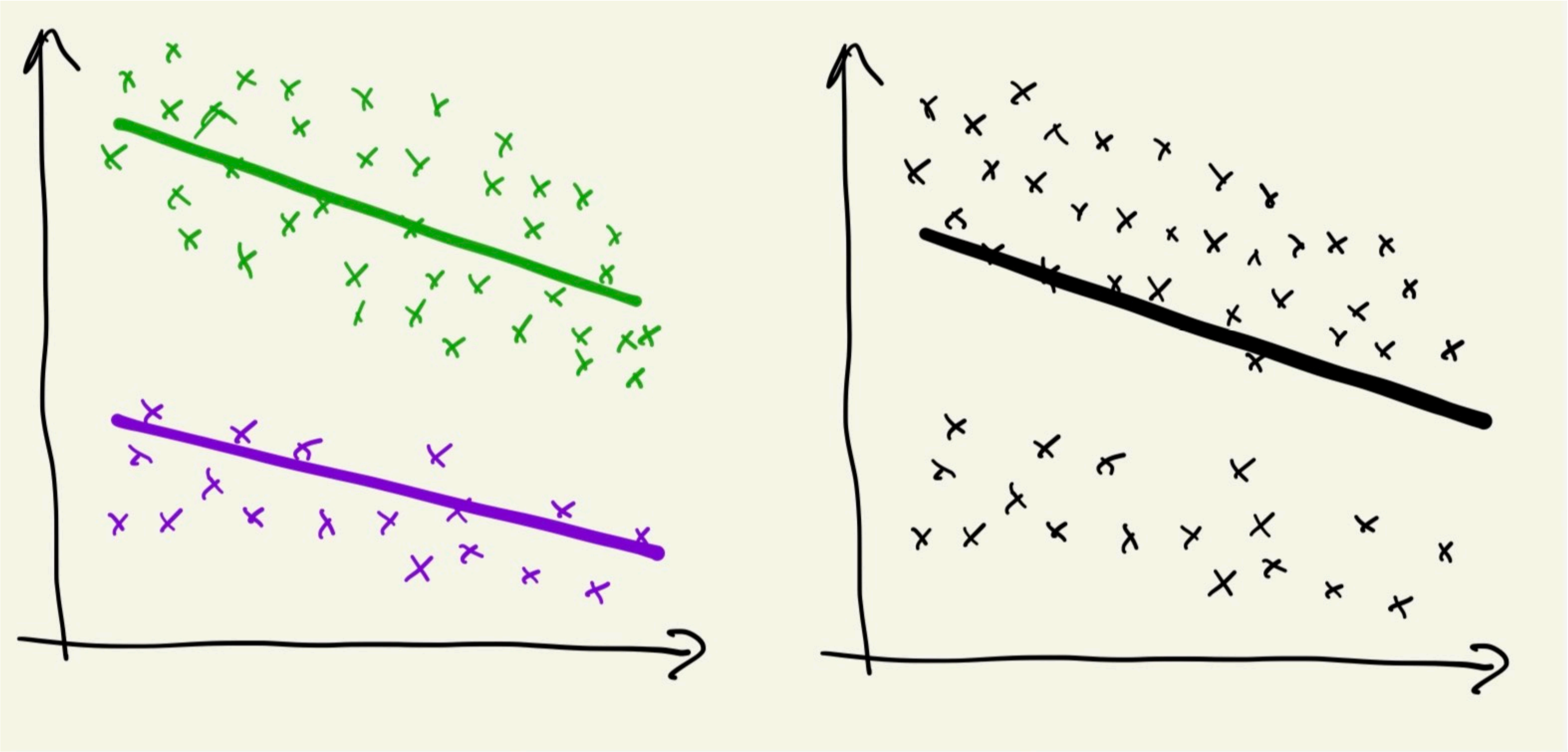
- Precision variables and effect modifiers
  - There almost always will be many unmeasured or unmeasurable precision variables
  - There almost always will be many unmeasured or unmeasurable effect modifiers
  - This is *fine!* You don't need to include all PVs and EMs in your model!

What happens when we omit variables?

- Unmodeled precision variables and effect modifiers get “averaged over”
    - Not necessarily a problem

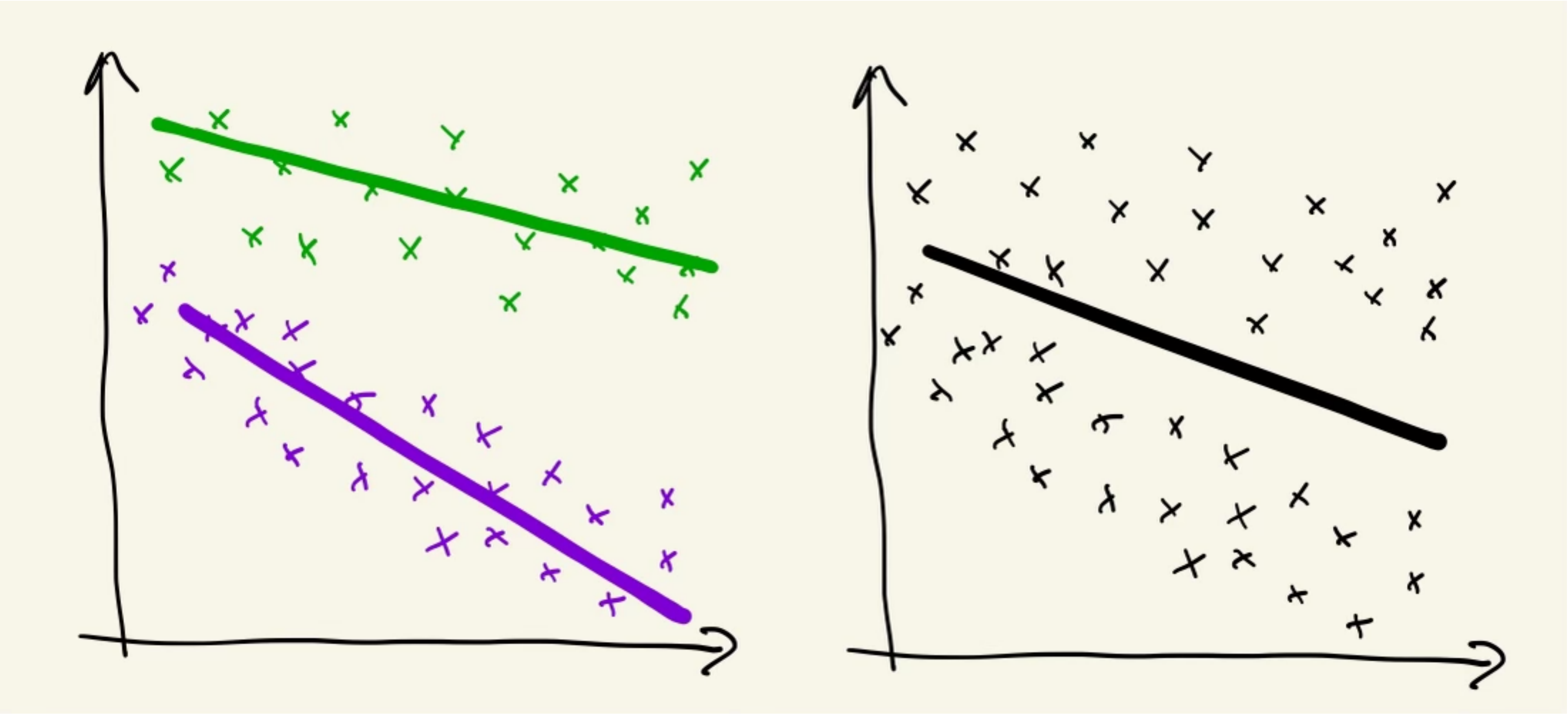
What happens when we omit variables?

- Unmodeled precision variables get “averaged over”



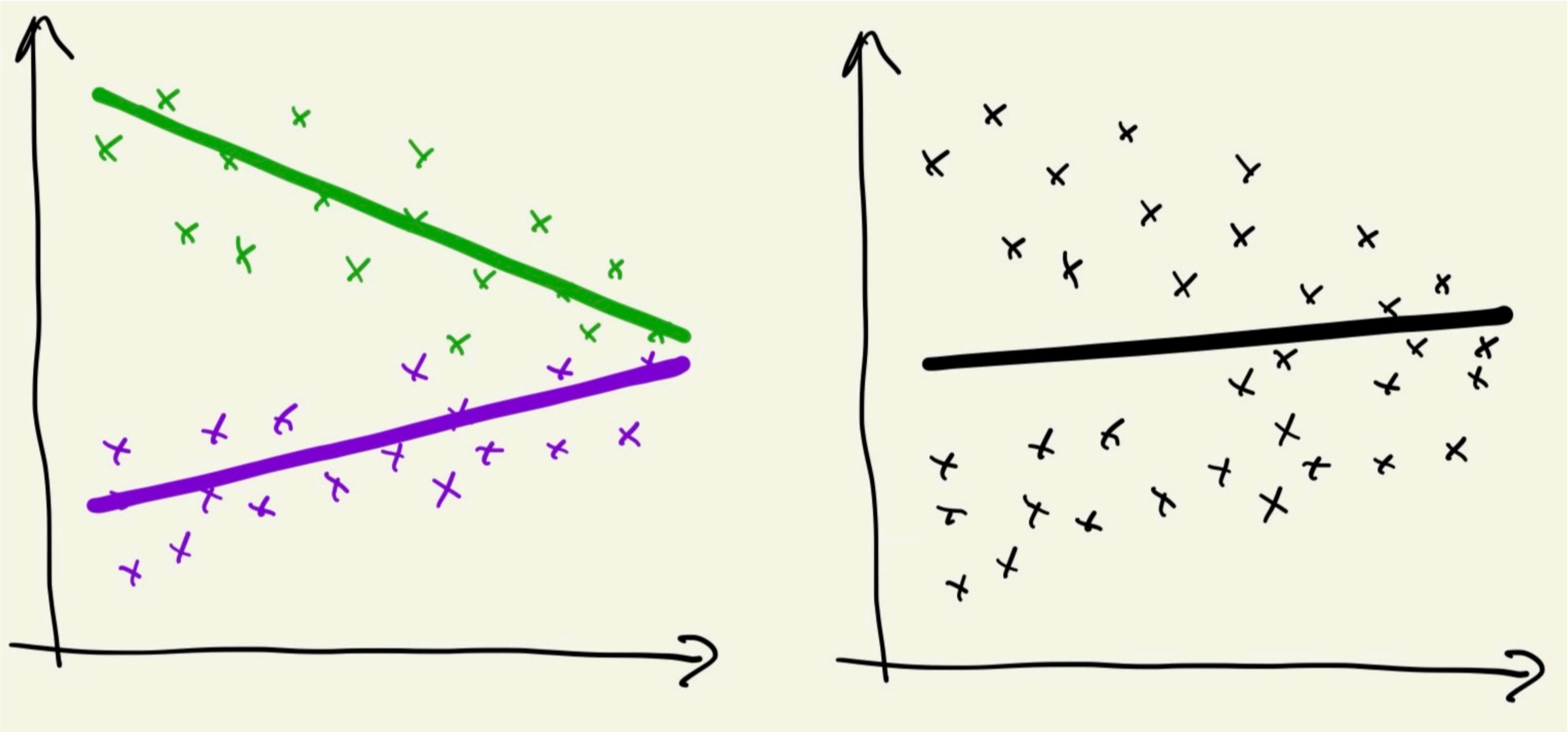
What happens when we omit variables?

- 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!

Amy’s wish list

- You choose a meaningful parameter to estimate
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

Key points

- ~~Dream big~~ Parameters are what you want
- ~~Life’s tough~~ Estimators are what you get
- ~~Dust yourself off~~ You need to make assumptions to estimate parameters
- ~~Believe~~ Data isn’t biased… but estimators can be
- ~~Resist peer pressure~~ Choose meaningful parameters

- This afternoon
  - abundance
- Tonight
  - (Optional) Parameters lab
- Tomorrow morning
  - Diversity
  - Experimental design
  - Descriptive stats
  - Miscellanea

Questions
&
Discussion

Statistics foundations

Statistical Diversity Lab @ University of Washington
Amy Willis — @AmyDWillis — Associate Professor
Shirley Mathur — PhD Candidate
Sarah Teichman — PhD Candidate
María Valdez — PhD Candidate