Exercise

- Is a t-test on alpha diversity inferential statistics or exploratory statistics?

This is a very, very advanced exercise!!!

Also, a useful learning tool!

Audience responses

- Question: Is a t-test on alpha diversity inferential statistics or exploratory statistics?
- “It’s exploratory statistics” 🦌🦌🦌🦌🦌🦌🦌🦌🦌🦌🦌🦌
    - “You are transforming your data to calculate alpha diversity, so it’s exploratory”
    - “You have to look at your data to calculate alpha diversity, so it’s exploratory”
        - “There’s no parameter here, so it’s exploratory”
- “It’s inferential statistics!” 🐰🐰🐰🐰🐰
    - “You said all hypothesis tests are inferential”
    - “You said that testing only happens in the inferential paradigm”
- “It’s both!”
- “Isn’t calculating Shannon diversity is bioinformatics, not statistics?”

Amy’s response

- Amy: “It’s exclusively, 100% inferential!”

Amy’s response

- If it’s inferential, we need
  - a parameter
  - a model
  - an estimator, and
  - a hypothesis
- What are they?
- (This is an advanced activity! I did not give sufficient background for you to work through this by yourselves!)

The procedure

- Let’s start by writing out the procedure, then study what underpins it

The procedure

1. Estimate each sample's true Shannon diversity by its "plug in" sample Shannon diversity

   $\hat{\alpha}_i = -\sum_j \hat{p}_{ij} \log \hat{p}_{ij}$ where $\hat{p}_{ij} := \frac{W_{ij}}{\sum_j W_{ij}}$

2. "Do a t-test comparing $\hat{\alpha}_1^{ctrl}, \ldots \hat{\alpha}_{n_2}^{ctrl}$ and $\hat{\alpha}_1^{tmt}, \ldots \hat{\alpha}_{n_2}^{tmt}$"

   - Find the sample average Shannon diversity of each group, call them $\hat{\bar{\alpha}}_i^{tmt}$ and $\hat{\bar{\alpha}}_i^{ctrl}$
   - Calculate a test statistic $T = \frac{\hat{\bar{\alpha}}_i^{tmt} - \hat{\bar{\alpha}}_i^{ctrl}}{\text{std error in } \hat{\bar{\alpha}}_i^{tmt} - \hat{\bar{\alpha}}_i^{ctrl}}$
     - There are a number of ways to calculate the denominator, and it's not important for this exercise
   - Find Pr(|calculated value of test statistics| > 1.96)

3. This gives a p-value for the null hypothesis of equality of means of Shannon diversity across treatment and control samples

Parameters

- Sample-level: Every sample has its own true Shannon diversity

- The Shannon diversity of sample i is defined to be

$$\alpha_i := -\sum_j p_{ij} \log p_{ij} \text{ where } p_{ij} := \frac{Y_{ij}}{\sum_j Y_{ij}}$$

- α_i depends on the Y_{ij}, which are also unknown parameters, so it is also an unknown parameter!

Parameters (ctd.)

- "Metadata"/covariate-level: Every treatment/control group has its own true average Shannon diversity

  - Let's call these averages $\bar{\alpha}_i^{tmt}$ and $\bar{\alpha}_i^{ctrl}$

    - i.e., If we knew the true Shannon diversity of the gut microbiome of every single person in the world who was treated, the average of those Shannon diversities would be $\bar{\alpha}_i^{tmt}$; similarly for control

  - Here, average = mean (not median, eg)

- Finally, we are interested in the difference between the true averages of the treatment and control groups

  - Our final, target parameter is $\bar{\alpha}_i^{tmt} - \bar{\alpha}_i^{ctrl}$

Model / assumptions

- What assumptions underpin this model?*
- We're estimating true Shannon diversity using $\hat{\alpha}_i = -\sum_j \hat{p}_{ij} \log \hat{p}_{ij}$ where $\hat{p}_{ij} := \frac{W_{ij}}{\sum_j W_{ij}}$
    - By replacing our true abundances with observed abundances, we're assuming that <u>empirical relative abundances $\hat{p}_{ij}$</u> are good estimates of <u>true relative abundances p_{ij}</u>
    - i.e., we're assuming that microbes are sampled uniformly-at-random (Model 2, Slide 17!)
- Relatedly, we're assuming that all units that were present were observed (i.e., no missing diversity)

*You could think of "models" as "assumptions", if you find it helpful

I wish I said this earlier

Model / assumptions

- What other assumptions underpin this model?
- We’re assuming that all our Shannon diversity estimates are *unbiased*
    - Neither too big nor too small on average
    - Not consistently over/underestimating
- We’re assuming that all our Shannon diversity estimates have the same uncertainty

Hypothesis

- A “t-test” tests that a difference is means is zero

- Here, our null hypothesis is that $\bar{\alpha}_i^{tmt} = \bar{\alpha}_i^{ctrl}$

Some perspective

- The proposed approach is only <u>one</u> way to test the hypothesis that $\bar{\alpha}_i^{tmt} = \bar{\alpha}_i^{ctrl}$

- It has a number of challenges!

Challenges

- A ranked list of challenges
  - **Biological units are typically not well-detected, so replacing p_{ij} by $\hat{p}_{ij}$ doesn't work well**
    - Model 2 is not well supported by control data
    - We often fail to detect taxa, especially low abundance taxa
      - $\hat{p}_{ij} = 0$ when $p_{ij} > 0$
      - This means we often *underestimate* Shannon diversity using plug-in estimates
    - Our uncertainties in the $\hat{\alpha}_i$'s (as estimates of α_i's) are not equal
      - Not all samples equally deeply sequenced
      - More uncertainty in estimating diversity in communities with many rare members

Perspective

- Do these challenges make a t-test on Shannon diversity useless?
    - Probably not!
    - There are situations where the true effect may swamp the impact of the challenges
- Are there better ways to test equality of Shannon diversity?
    - ~~Yes?~~ Probably!
    - But, talk to the StatDivLab, too!
    - Amy: I doubt you *really* care about Shannon diversity — I think you've seen other people use it, and therefore thought it must be reasonable — so, I'll probably ask *why* you care about this parameter