Honest statistics for microbial ecologists

Amy Willis

August 6 2016

OTU tables are huge and terrifying

At the end of our bioinformatics pipeline, we get our species abundance table

	A	B	C	D	E	F
1	amplicon	BAR	LAS	PIBa	PIBb	WIS
301281	79bd1a5bc510eded9f7abc97704816130ded2dce	0	1	0	0	0
301282	79ddb78bb1f9751699fe25fe3a61be7156a825b8	0	1	0	0	0
301283	7c3d613be383e8353818111da0f027498730fe2c	0	1	0	0	0
301284	7df0a84fd4e3e02f94229a26e7afe3acf5efb1ae	0	0	1	0	0
301285	7e9bb19bbec6d43c51290601b52eb3db24f2b68f	0	5	0	0	0
301286	7ec1e3d32273b4a2eb929c0ea656216210f6d8b2	1	0	0	0	0
301287	7f44fbc9119a37e3d6e0ede9edf432797915c5d1	0	0	0	0	0
301288	8075200e574b04416a374d70c21ea2b0a0a62081	0	1	0	0	0
301289	80ccc3a94c09fded098a558cab882f4c573d0f02	1	0	0	0	0
301290	80e8250b98b778a1c42933d3658d74c6f9ebf7ef	1098	729	92	30	52
301291	8122737296882b62ae4a079e9bb6be3b5f030c99	2	0	0	0	0
301292	82156752f11fc4b00c5f2b4814c90fef5f28354d	1	0	0	0	0
301293	83cea5dacc325d558a561027a9d4f8d637684732	1	0	0	0	0
301294	83f6cdd3292e490d92a1109739a20b2d9cacb926	0	0	0	0	0
301295	8476932bf4e4577ce7c4f8c23222194c7ea2726e	0	0	0	0	0

302,104 rows = 302,104 different microbes observed

52 columns = 52 lakes

OTU tables are huge and terrifying

Community-wide summaries help us digest information

- Richness
- Evenness

Our goals

1. Meaningful estimates of this information
2. A way to compare across communities

Our mutual misunderstandings

Parameters: True, unknown quantities associated with the environment/population under study

- μ, the true population mean (eg. shoe size of STAMPS 2016 participants)
- p, the true population proportion (eg. of microbes in the Phylum Cyanobacteria in the lake)
- β, a regression parameter (eg. number of microbes lost for each 1 unit increase in soil pH)

Our mutual misunderstandings

Estimates: The numbers/formulae we use to estimate the parameters

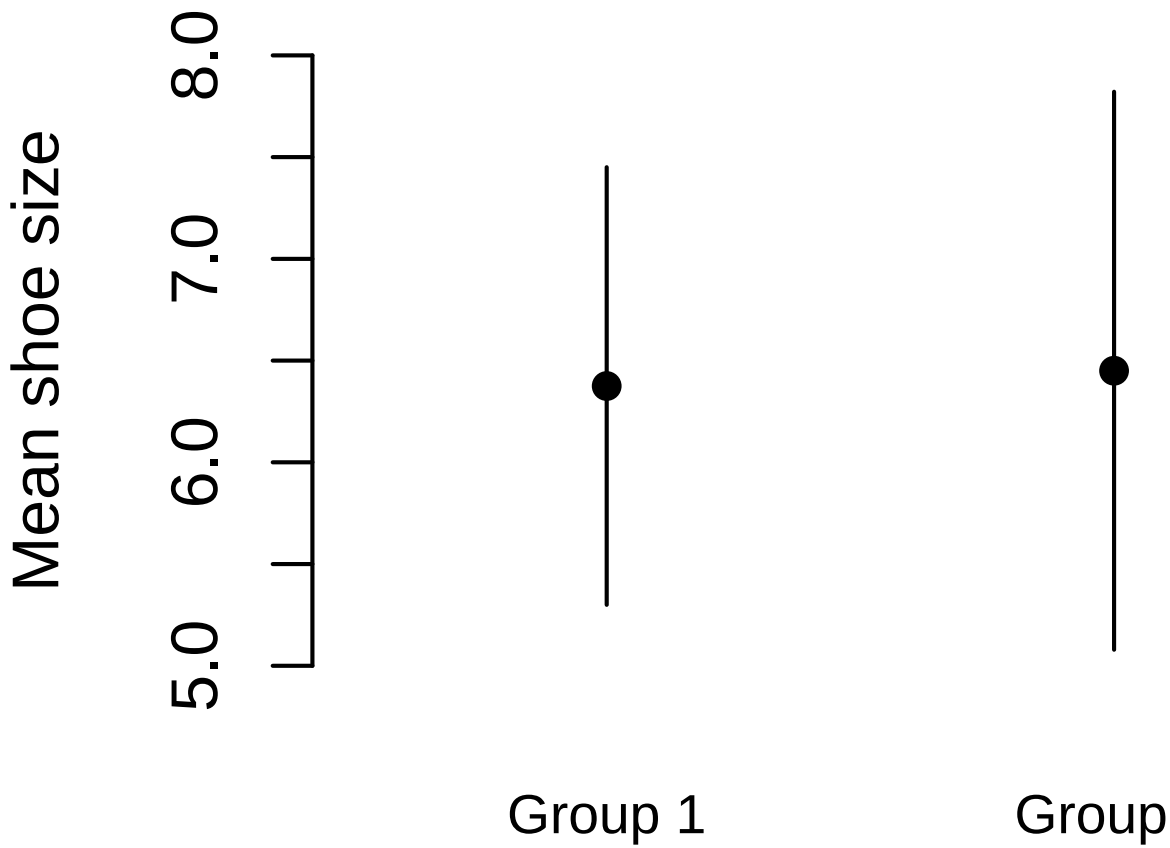
- $\overline{X}$, the sample mean
- $\hat{p}$, the sample proportion
- $\hat{\beta}$, your least-squares estimate of the regression parameter

Estimates are random, and have errors. Parameters are fixed.

A somewhat familiar example

We want to estimate the population mean μ by the sample mean $\overline{X}$. The central limit theorem tells us that

- The estimate is normally distributed around the parameter
- A good estimate of the standard deviation of the estimate is $s/\sqrt{(n)}$



Warnings

- Both of these guarantees only necessarily apply when estimating a mean
  - A different procedure should apply when estimating *not-mean* parameters
- Dividing by n: the standard error in the estimate of the mean versus the standard deviation of the data

What changes?

Sometimes we are interested in targets other than means.
Considerations to make

- What estimate do you use?
- What estimate of the standard deviation of the estimate ("standard error") will you use?
- Do you know the distribution of the estimate?

The thinking is the same, but the mechanics may be different!

An unfamiliar example: Shannon diversity & evenness

Consider a community of 1000 taxa:

- Equal proportions of each: 9.966
- 1 taxon comprises 20%, and the rest split the rest: 8.693
- 1 taxon comprises 80% and the rest split the rest: 2.715
- 1 taxon comprises 50%, another 25%, another 12.5%. . . : 2

Shannon index increases as community gets more even

How many myths arose

True (population) Shannon diversity:

$$-\sum_{taxa \in pop' n} p_{taxa} \log_2 p_{taxa}$$

Estimate of Shannon diversity:

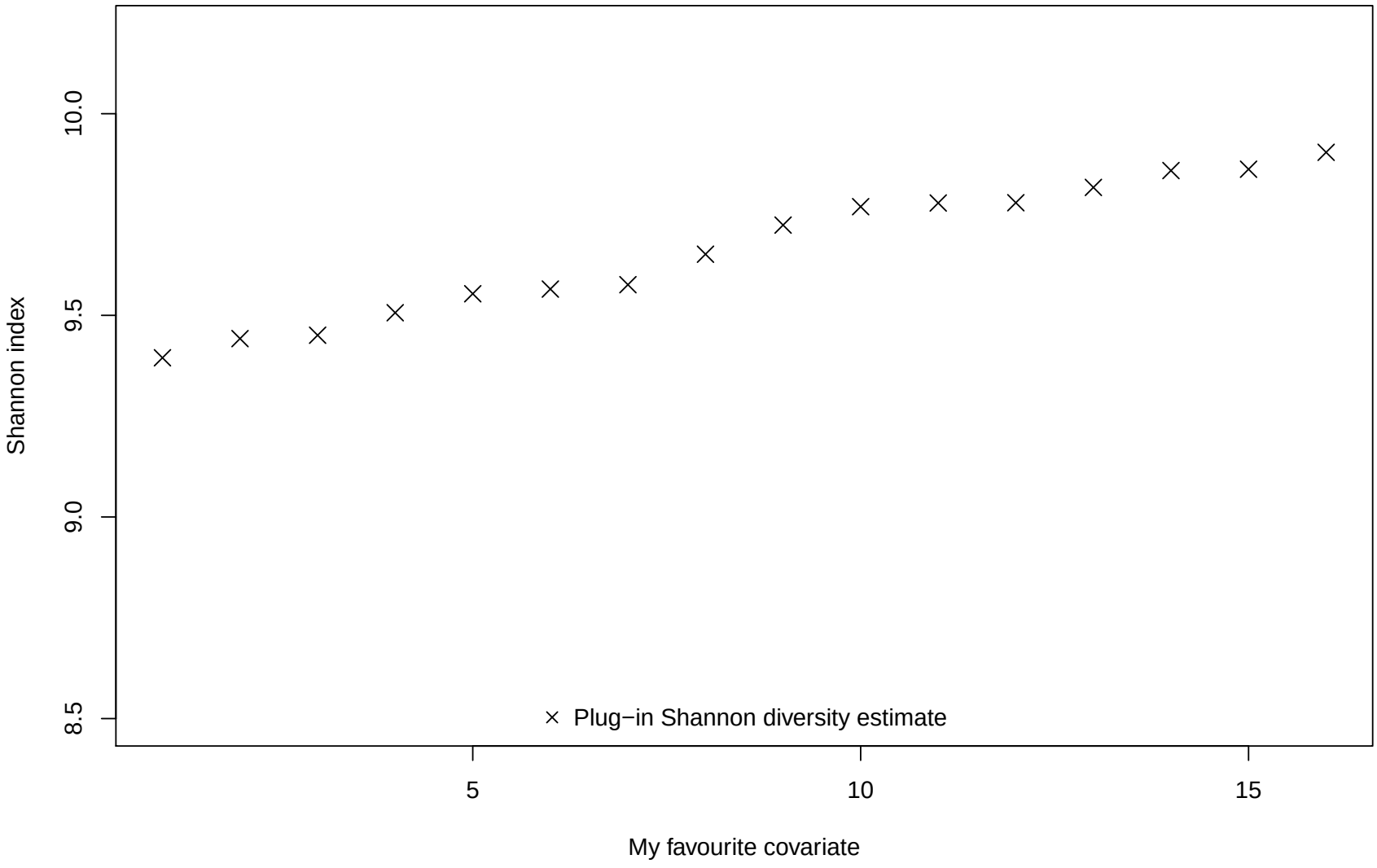
$$-\sum_{taxa \in sample} \hat{p}_{taxa} \log_2 \hat{p}_{taxa}$$

The above estimate, called the plug-in estimate, is only one way to estimate the true Shannon diversity. And we know basically nothing about it!

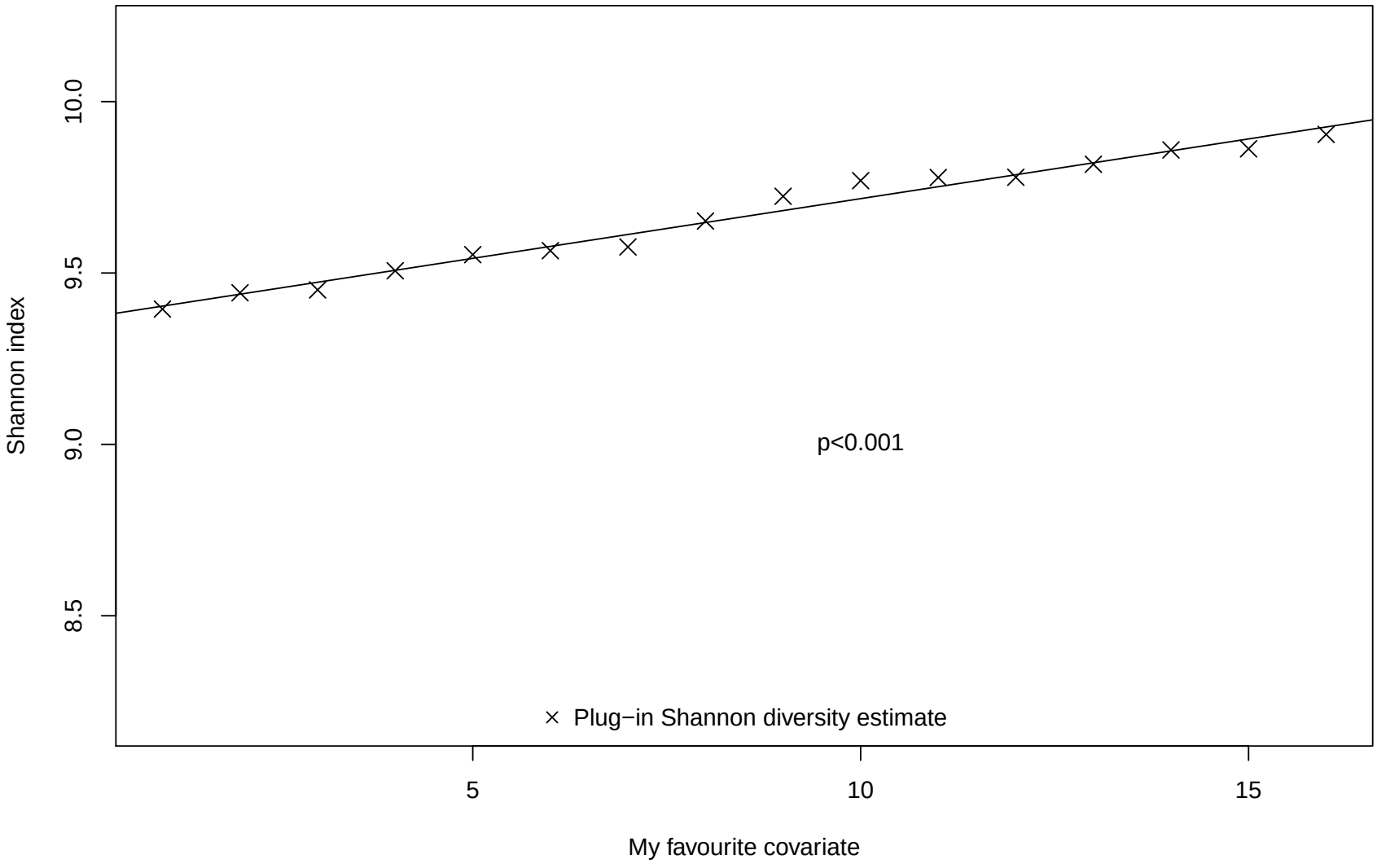
Two major issues

- Our plug-in estimate may not be the best estimate of the Shannon diversity
- It is almost never quoted with a standard error: no attention is paid to its robustness

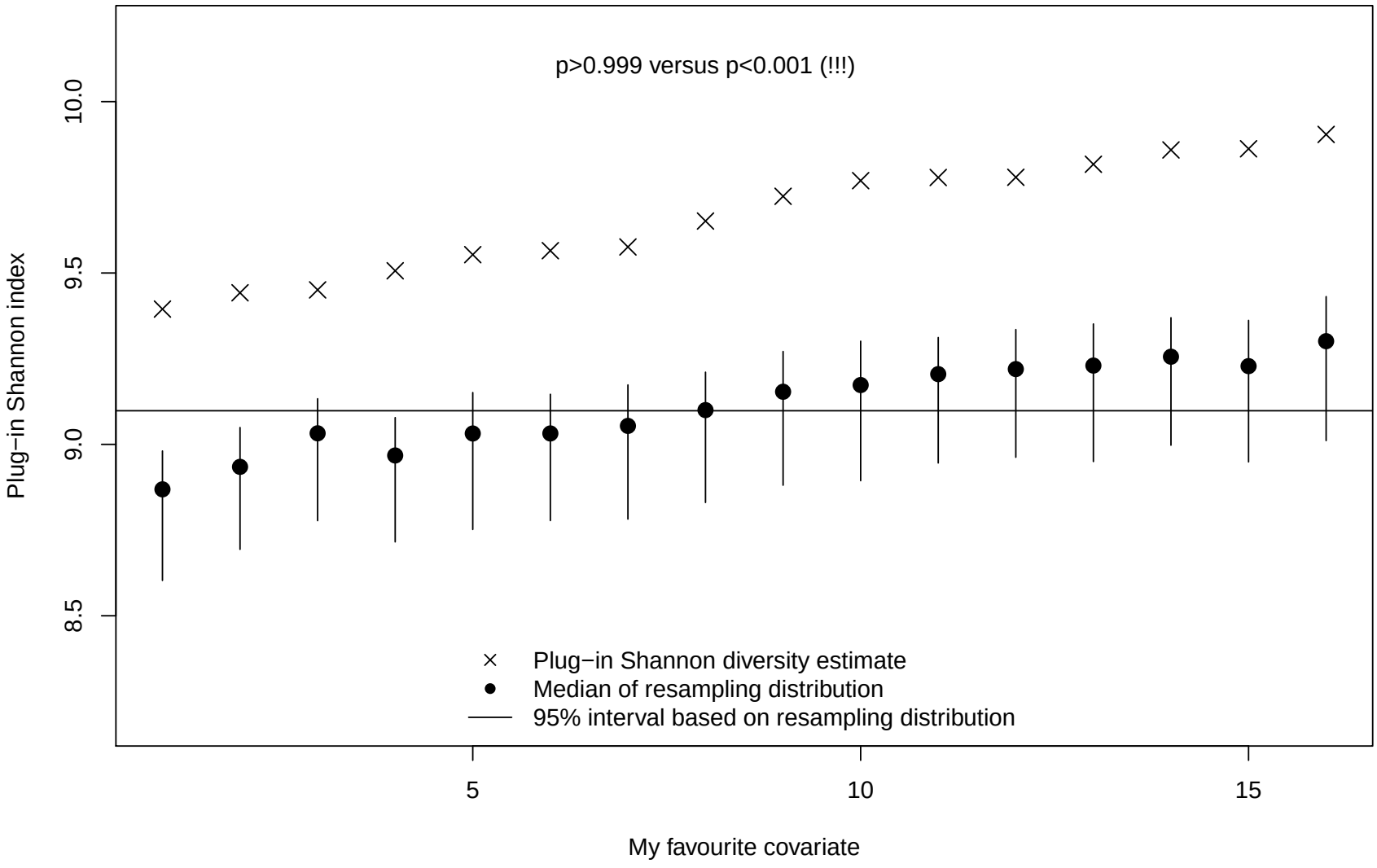
How this can be misleading



How this can be misleading



Correcting the picture

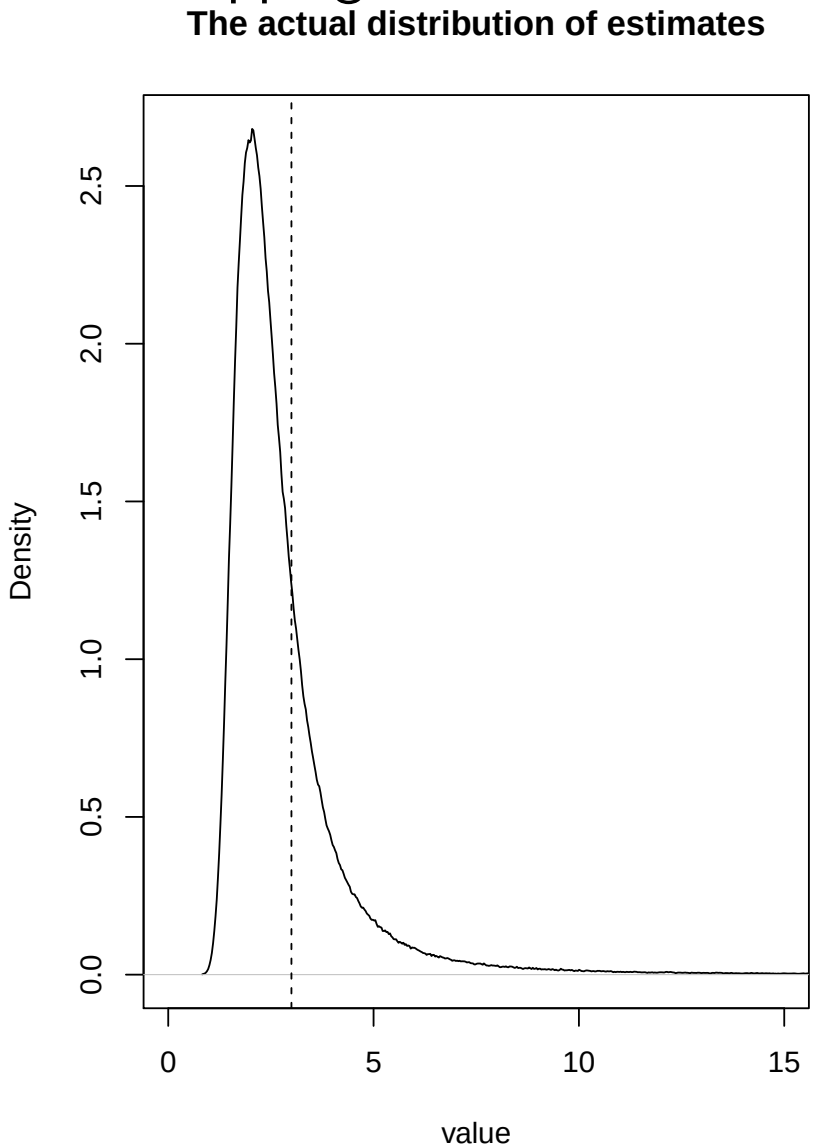


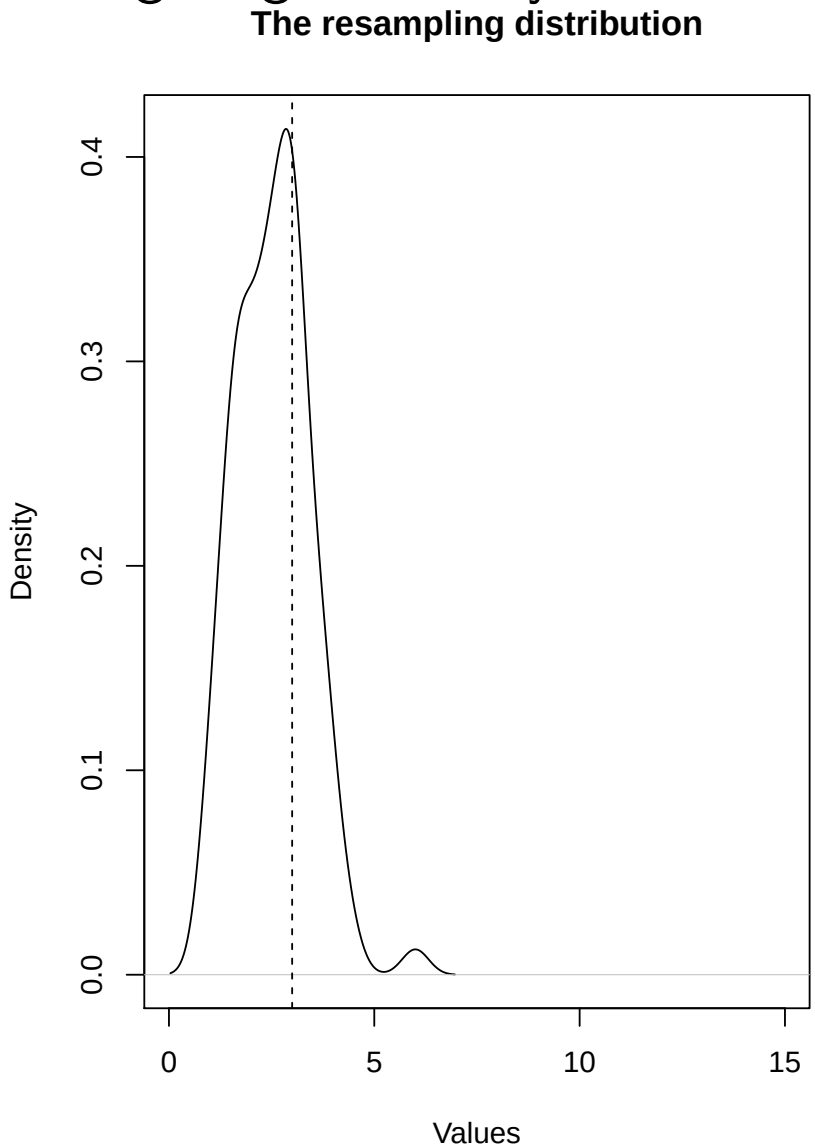
The core problem: no theory

Unlike for means, we don't know how diversity index estimates are distributed (no central limit theorems)

The best we can do for now: Bootstrapping

Bootstrapping is a useful tool for investigating sensitivity





Bootstrapping

Bootstrapping is a useful tool for approximating standard errors of understudied estimates, eg. plug-in diversity indices!

- Bootstrap standard errors generally understate true error

You cannot bootstrap yourself out of a bad sample. . .
But a reasonable sample may give you a more reasonable idea of the estimator variability

How to do this

The **R** package `breakaway` has a simple implementation called `resample_estimates`

```
set.seed(7)
my_resamples <- replicate(100,
    resample_estimate(otus, shannon))
head(my_resamples, 5)
```

```
## [1] 9.580679 9.566047 9.580123 9.595058 9.588148
```

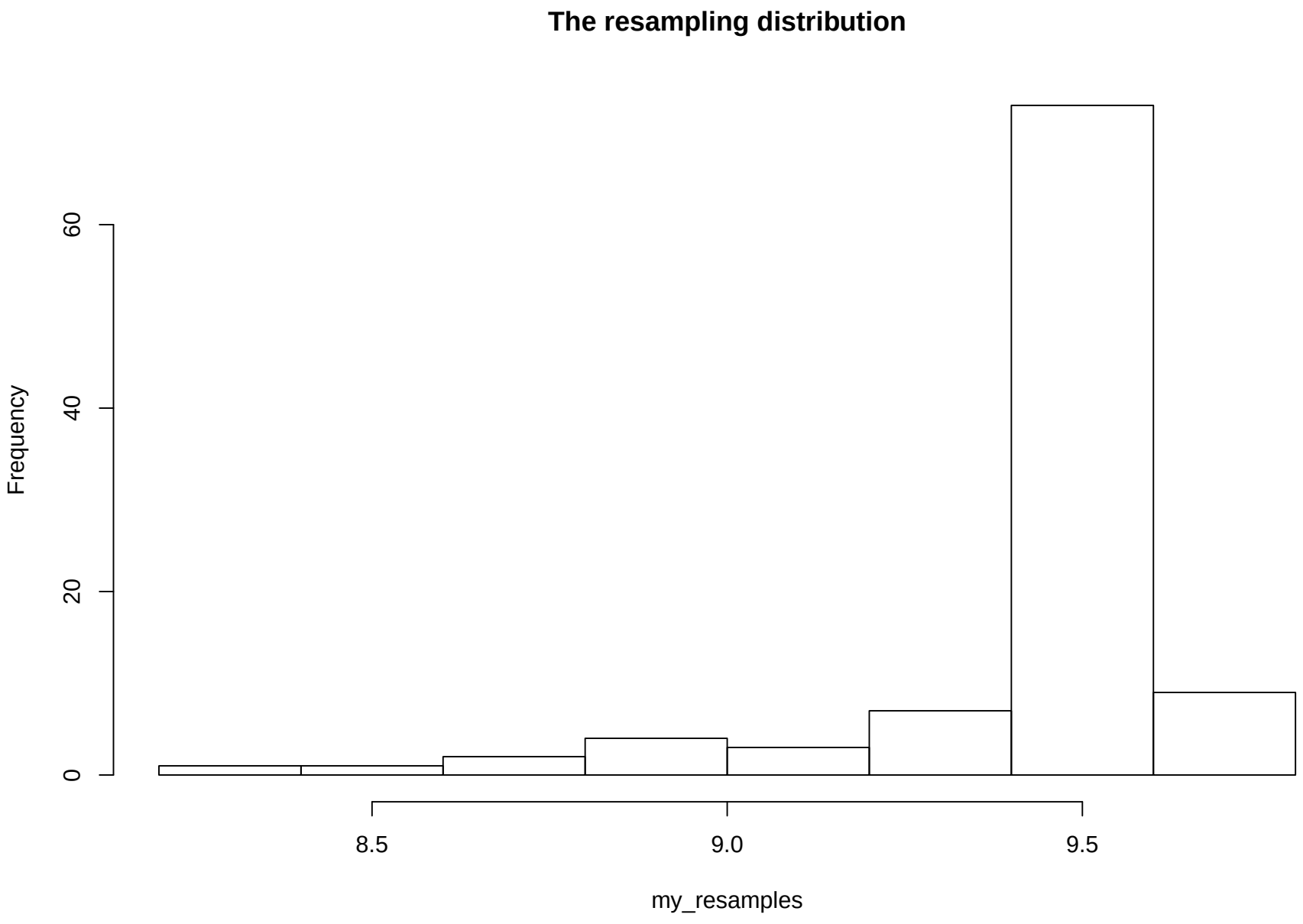
```
mean(my_resamples)
```

```
## [1] 9.455717
```

```
sd(my_resamples)
```

```
## [1] 0.245028
```

Distributions

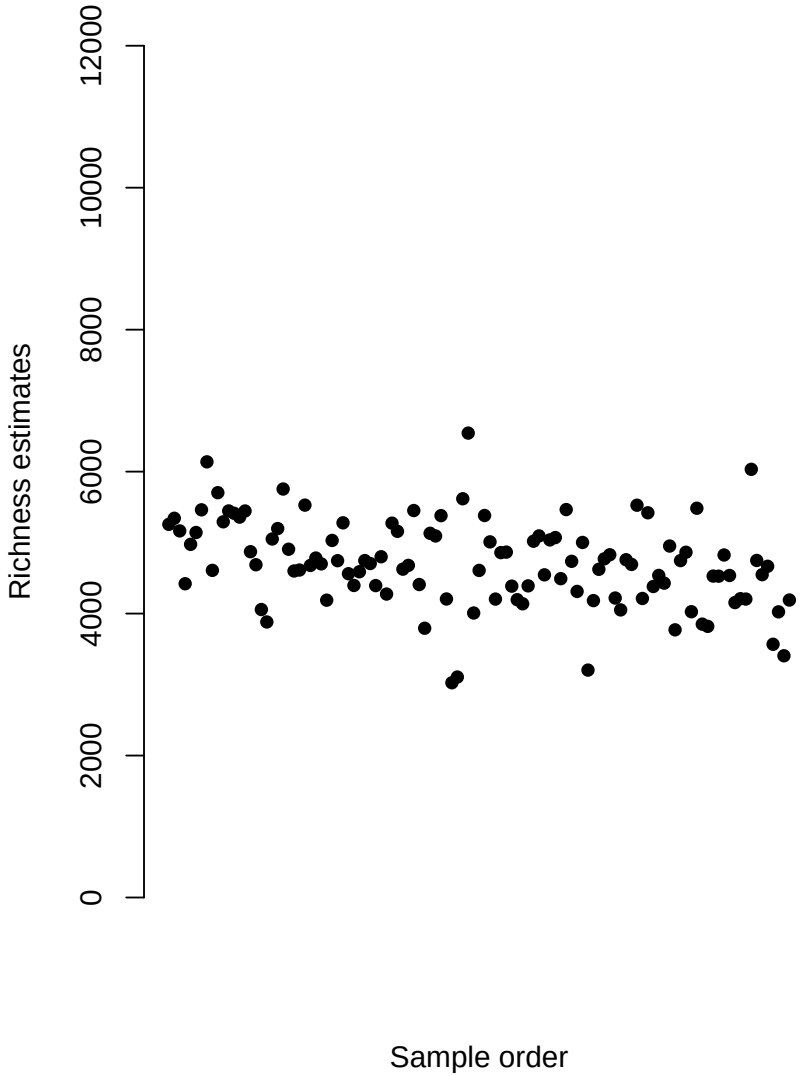


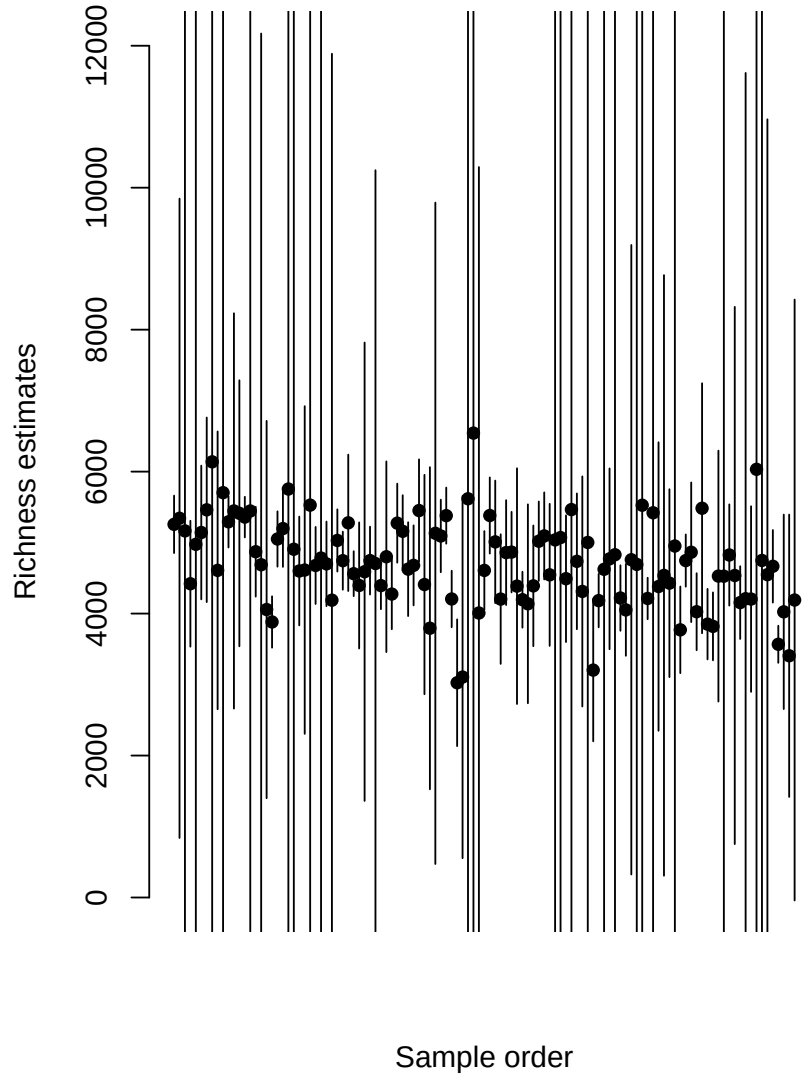
What next?

Great! Now we have some idea how variable our estimates are!
How do we use them?

Modelling and inference

The way we formally test for changes with covariates *changes* when we have error





betta: A better way to do testing in microbial cases

```
betta(ests, ses, my_X)$table
```

```
##              Estimates Standard Errors p-values
## No Amdmt    4680.4121         63.51318    0.000
## Biochar        0.0000        100.79685    1.000
## Biomass     -497.2997        158.99322    0.002
```

"We reject the null hypothesis that fresh biomass additions have no effect on species richness (p=0.002) and conclude an average loss of 497 taxa compared to non-fertilized controls."

The methodology underneath

betta works by letting each "observation"" be blurry: we have ambiguity about where the observation falls (sampling variability) in addition to true "noise" in the pattern

total variability = estimation error + noise

NEVER

NEVER use rarefaction-based estimates.

You can always do better without rarefying.

Closing remarks: To be continued

The proposal from today is a very rough solution to a very big problem

I believe it points us in the right direction of better variability accounting

It is a temporary solution while statisticians' models and testing frameworks continue to be developed

The Point

A basic procedure that accounts for variability is better than nothing

The Plan

I'm now going to do a 15 minute lecture on species richness, then a 30 minute lab to demonstrate this stuff

Questions on the content of this talk?