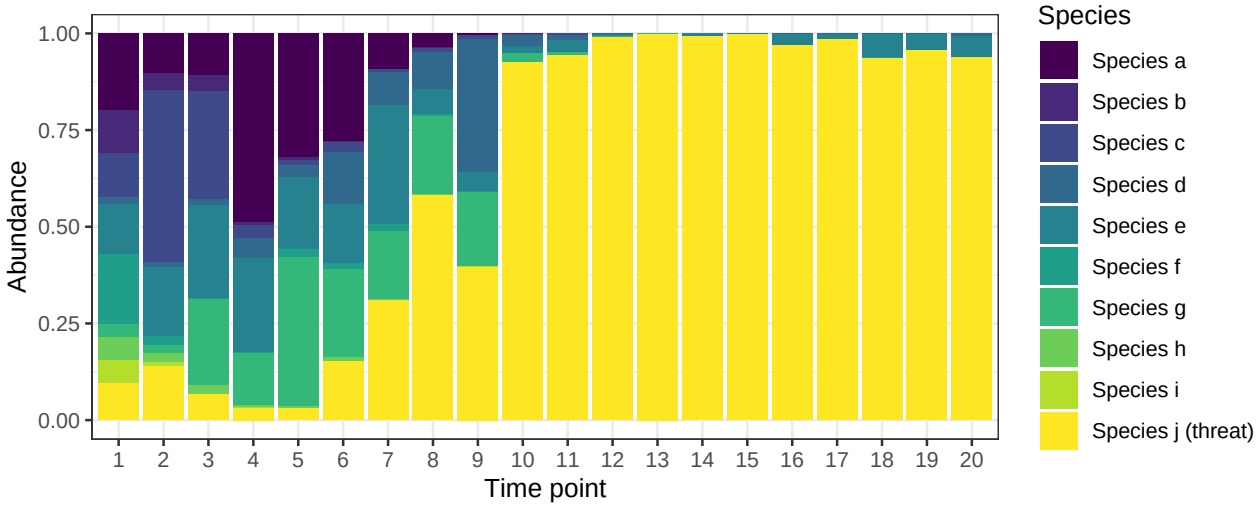
Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 7 due Friday 3/31

Detecting the unknown

During the Covid-19 pandemic, considerable effort was put into estimating Covid-19 community case counts based on wastewater testing. With an eye towards detecting future pandemics early, the Nucleic Acid Observatory was founded in 2021 with the intention of performing daily metagenomic sequencing on wastewater and major waterways for early biological threat detection. Because a pandemic is likely to involve a previously uncharacterized microbe, the detection will rely on the assumption that a novel threat will show exponential growth compared to the background fluctuations of other microbes.



In the plot above, we want to be able to detect threats like species j before they become dominant. To make these detections, we will assume the daily change in log abundance for a species is independent $Y_t \sim \mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 unknown (raw species abundances in microbiome data are often assumed to have log-normal distributions). If $\mu > 0$, on the original scale, the species will grow exponentially over time, indicating a threat.

1. Write a one-sided null and alternative hypothesis. Is this null simple or composite?

2. Suppose we have observed day-to-day differences $Y_1, ..., Y_n$ for a particular species. Construct an exact test statistic and give its distribution under the null. Show how you would find a p-value p_1 for the observed test statistic t_{obs}. State the rejection region for a significance level α.

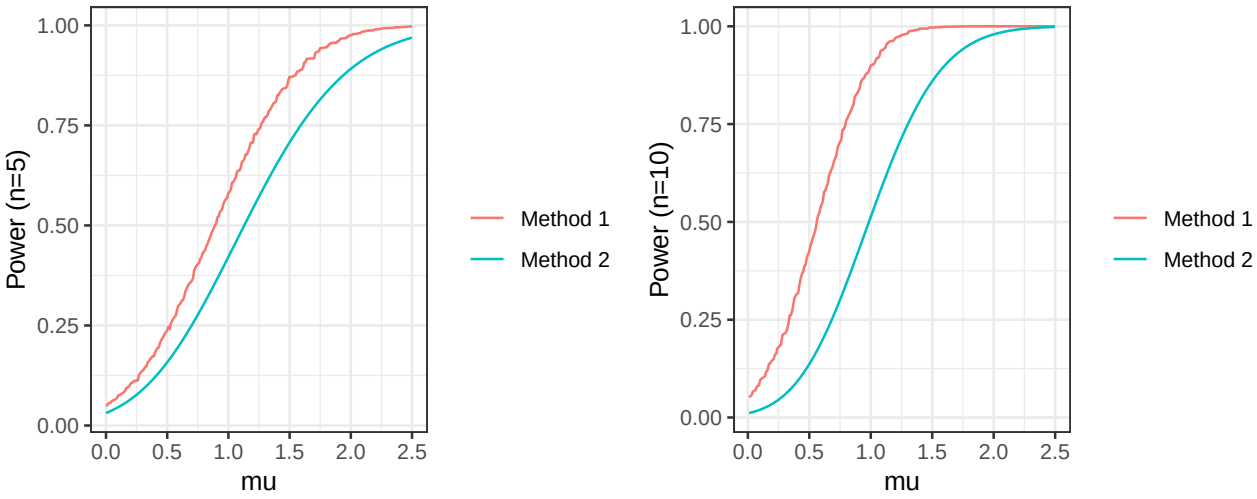
3. Show that a confidence interval with width $2\bar{Y}$ covers μ with probability $1 - 2p_1$.

4. Find the power for the test $\beta(\mu, \sigma^2)$ at significance level α (the probability of rejecting the null given the true parameters μ and σ^2). Leave the answer as an expectation that could be calculated with LOTUS and explain how you would calculate it numerically.

5. Another way to test our hypotheses is to use the proportion of days the abundance of a species increased. Let I_i be the indicator that $Y_i > 0$. Construct a test statistic based on I_i and give its distribution under the null. Show how you would find a p-value for the observed test statistic. Can this test be constructed to give an exact type I error rate of $\alpha = 0.05$?

6. Find the power for the test $\beta(\mu, \sigma^2, \alpha)$. How does this compare to before?

7. Fixing $\sigma^2 = 1$ and $\alpha = 0.05$, the following plot shows the power of each method as a function of μ from $\mu = 0$ to $\mu = 2.5$ on the log scale for $n = 10$. Which method performs better and why? Why is the first method so jumpy? Why is the second method not 0.05 at $\mu = 0$?



8. A receiver operator characteristic (ROC) curve plots the true positive rate against the false positive rate to show the accuracy of a binary predictor. The curve can be considered the result of evaluating many thresholds and plotting the true positive and false positive rate at each. A curve that goes from (0,0) to (0,1) to (1,1) is a perfect classifier, and a curve that follows the $y = x$ line shows no predictive value. Give two pairs of parametric equations that would give a proper ROC curve for each method.

9. The vast majority of tested microbes will not be pathogenic. In particular, assume that 1 out of every k microbes is pathogenic for some large k. The false discovery rate is the proportion of tests called as significant in which the null is actually true. What is the false discovery rate for each test as a function of $k, n, \mu, \sigma^2, \alpha$? What are these for large values of k? What does this indicate?

10. Perform a Wald test based on the second test statistic. What is the p-value if the microbe increased in abundance on 8 of the 10 observed days? Recall that for $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} I_i$, the MLE for the true proportion of times the microbe's abundance increases, $\hat{p} \xrightarrow{d} \mathcal{N}(p, \mathcal{I}_{\vec{Y}}^{-1}(p))$ with $\mathcal{I}_{\vec{Y}}(p) = \frac{n}{p(1-p)}$.

11. Perform a likelihood ratio test based on the second test statistic. What is the p-value if the microbe increased in abundance on 8 of the 10 observed days? Recall that the likelihood test statistic $\Lambda(\vec{Y}) = 2\log\left(\frac{L(\hat{p};\vec{Y})}{L(p_0;\vec{Y})}\right) \xrightarrow{d} \chi_1^2$ under the null.

12. How do these compare to the exact p-value?