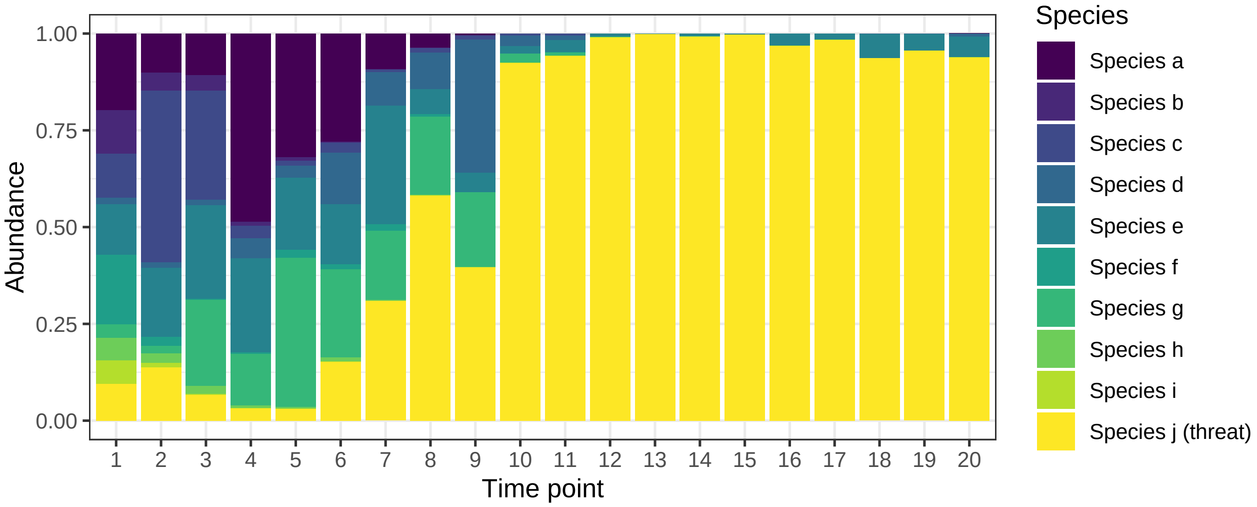
Announcements

Make sure to sign in on the google form.

Pset 7...

Detecting the unknown

During the Covid-19 pandemic, researchers put considerable effort 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?

$H_0 : \mu \leq 0$, $H_a : \mu > 0$. We will be using $\mu = 0$ as the null throughout because this is on the boundary and is therefore the hardest value of μ to reject. This null is composite because it involves a range.

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.

Because the day-to-day differences are Normal with unknown μ and σ, we will use the t-distribution. Therefore, under the null of $\mu = 0$,

$$T_{obs} = \frac{\bar{Y}}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

The p-value is obtained by looking for the probability of observing a more extreme t_{obs} statistic under the null than was actually observed: $p_1 = 1 - F_{t_{n-1}}(t_{obs})$. The rejection region is the values of $\vec{Y}$ that would cause us to reject the null: $R = \{\vec{y} : t_{obs} > Q_{t_{n-1}}(1 - \alpha)\}$ where we use α rather than $\alpha/2$ since the test is one-sided.

3. 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.

The power is

$$\begin{aligned}\beta_1(\mu, \sigma^2, \alpha) &= P(T_{obs} > Q_{t_{n-1}}(1-\alpha)) \\ &= P(F_{t_{n-1}}(T_{obs}) > 1-\alpha) \\ &= P\left(F_{t_{n-1}}\left(\frac{\bar{Y}}{\sqrt{\hat{\sigma}^2/n}}\right) > 1-\alpha\right) \\ &= E\left(I\left(F_{t_{n-1}}\left(\frac{\bar{Y}}{\sqrt{\hat{\sigma}^2/n}}\right) > 1-\alpha\right)\right)\end{aligned}$$

Since $\bar{Y}$ has the distribution $\mathcal{N}(\mu, \sigma^2/n)$ and $\hat{\sigma}^2$ has the distribution $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$ independently of $\bar{Y}$, we can write the expectation as a LOTUS integral using the Normal and Chi-squared densities:

$$\int_{-\infty}^{\infty}\int_0^{\infty} I\left(F_{t_{n-1}}\left(\frac{x}{\sqrt{\sigma^2 y/((n-1)n)}}\right) > 1-\alpha\right)\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2/n}\right)\frac{y^{n/2-1}e^{-y/2}}{2^{n/2}\Gamma(n/2)}dydx$$

4. 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 $\alpha = 0.05$?

Under the null, the population is equally likely to increase or decrease, so $P(I_i = 1) = 1/2$. Therefore, under the null, the test statistic $S = \sum_{i=1}^n I_i \sim \text{Bin}(n, 1/2)$. We can obtain an exact p-value (the probability of seeing data as extreme or more extreme under the null) by using the CDF of the binomial: $p_2 = 1 - F_{Bin(n,1/2)}(S-1)$. Since the test statistic takes on discrete values, the p-value will also take on discrete values, and we cannot ensure that the type I error rate is exactly α.

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

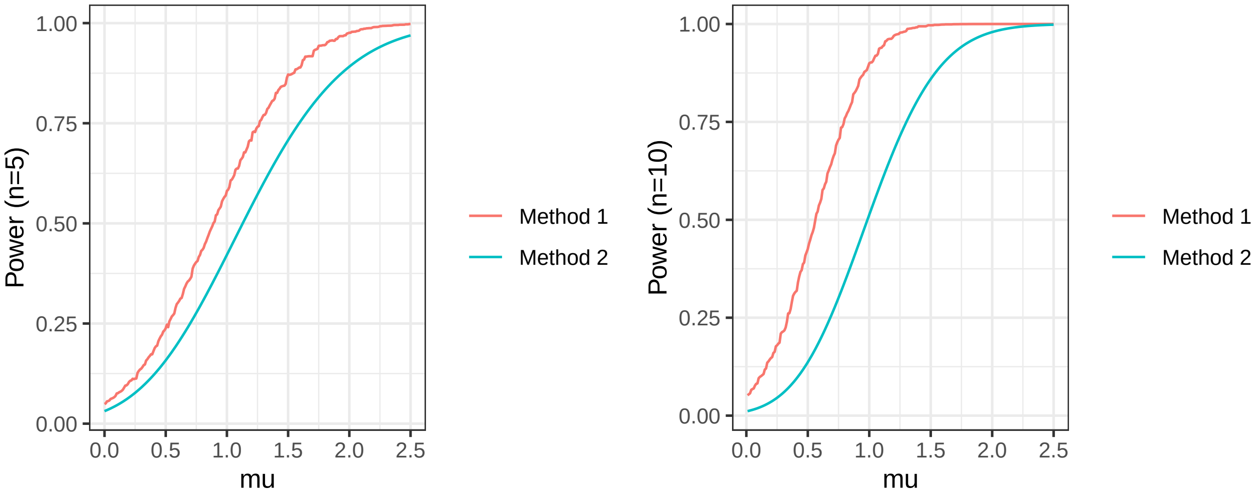
In general, the probability the microbe will increase in abundance is

$$P(I_i = 1) = P(Y_i > 0) = P\left(\frac{Y_i - \mu}{\sigma} > -\mu/\sigma\right) = 1 - \Phi(-\mu/\sigma) = \Phi(\mu/\sigma)$$

Therefore, $S \sim \text{Bin}(n, \Phi(\mu/\sigma))$. Then, the power is

$$\begin{aligned}\beta_2(\mu, \sigma^2, \alpha) &= P(S \geq Q_{Bin(n,1/2)}(1-\alpha) + 1) \\ &= 1 - \sum_{k=0}^{Q_{Bin(n,1/2)}(1-\alpha)} \binom{n}{k} (\Phi(\mu/\sigma))^k (1 - \Phi(\mu/\sigma))^{n-k}\end{aligned}$$

6. 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 $\mu = 0$?

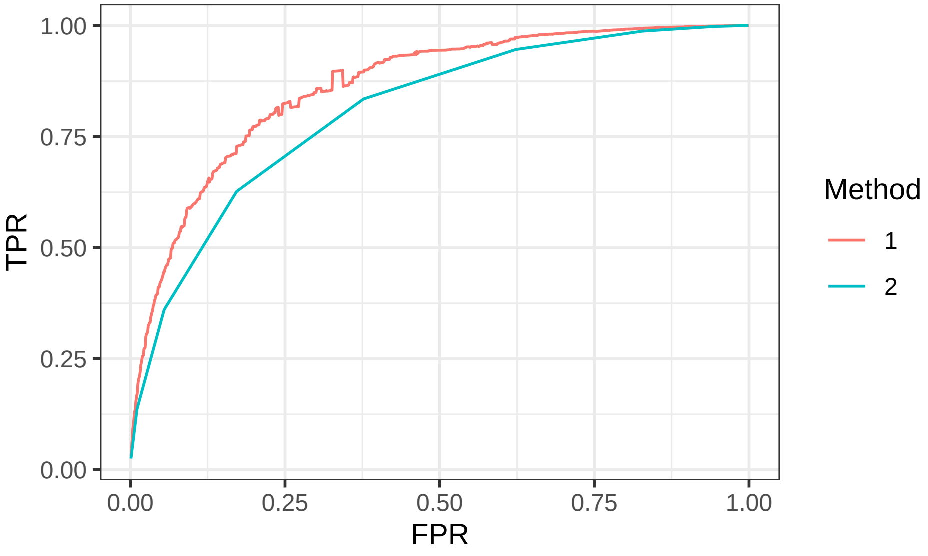


The first method is more powerful because it places more restrictive assumptions on the distribution; method 2 lacks power by up to 0.3 relative to method 1 depending on μ. Both are more powerful with more days of observations. Method 1's power is jumpy because it is calculated with numeric integration. The second method is not 0.05 at $\mu = 0$ because it is based on a discrete distribution, so it will have a type 1 error rate of at most, but not necessarily, $\alpha = 0.05$.

7. 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.

For method 1, we have $\text{FPR}(\alpha) = \alpha$ and $\text{TPR}(\alpha) = \beta_1(\mu, \sigma^2, \alpha)$. For method 2, we have $\text{FPR}(\alpha) = 1 - F_{Bin(n,1/2)}(Q_{Bin(n,1/2)}(1-\alpha))$ and $\text{TPR}(\alpha) = \beta_2(\mu, \sigma^2, \alpha)$.

For $\mu = 0.5$, $\sigma^2 = 1$, and $n = 10$, these give the curves:



Both show moderate but not excellent discriminatory power.

8. 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?

Using Bayes' rule, for method 1 we have

$$\text{FDR}(k, n, \mu, \sigma^2, \alpha) = \frac{\frac{k-1}{k}\alpha}{\frac{k-1}{k}\alpha + \frac{1}{k}\beta_1(\mu, \sigma^2, \alpha)} = \frac{(k-1)\alpha}{(k-1)\alpha + \beta_1(\mu, \sigma^2, \alpha)}$$

For method 2 we have

$$\begin{aligned}\text{FDR}(k, n, \mu, \sigma^2, \alpha) &= \frac{\frac{k-1}{k}(1 - F_{Bin(n,1/2)}(Q_{Bin(n,1/2)}(1-\alpha)))}{\frac{k-1}{k}(1 - F_{Bin(n,1/2)}(Q_{Bin(n,1/2)}(1-\alpha))) + \frac{1}{k}\beta_2(\mu, \sigma^2, \alpha)} \\ &= \frac{(k-1)(1 - F_{Bin(n,1/2)}(Q_{Bin(n,1/2)}(1-\alpha)))}{(k-1)(1 - F_{Bin(n,1/2)}(Q_{Bin(n,1/2)}(1-\alpha))) + \beta_2(\mu, \sigma^2, \alpha)}\end{aligned}$$

For a fixed α, since β_1 and β_2 are between 0 and 1, when k is large, these will be approximately 1: almost all microbes flagged as pathogenic exponential replicators will not actually be threats.

9. 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 $\mathcal{I}_{\vec{Y}}(p) = \frac{n}{p(1-p)}$.

Under the null, $\hat{p} \xrightarrow{d} \mathcal{N}(1/2, \mathcal{I}_{\vec{Y}}^{-1}(1/2))$, so for large n

$$\frac{\sqrt{n}\,(\hat{p} - 1/2)}{\sqrt{1/2(1-1/2)}} = 2\sqrt{n}\,(\hat{p} - 1/2) \sim \mathcal{N}(0, 1)$$

approximately, so we reject H_0 if $2\sqrt{n}\,(\hat{p} - 1/2) > Q_{\mathcal{N}(0,1)}(1-\alpha)$. For $\hat{p} = 0.8$ and $n = 10$, $2\sqrt{n}\,(\hat{p} - 1/2) \approx 1.90$, which gives a p-value of 0.029.

```
n <- 10
phat <- 0.8
pnorm(2 * sqrt(n) * (phat-0.5), 0, 1, lower.tail = F)
```

```
## [1] 0.02888979
```

10. 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.

Normally, we will find $\Lambda(\vec{Y})$ and reject the null if $\Lambda(\vec{Y}) > Q_{\chi_1^2}(1-\alpha)$. However, because we have a one-sided null, the possible parameter space is $[1/2, 1]$ with a null on the boundary. For the same reason as biohazard 8.6.3, the distribution of the test statistic under the null is now a point mass of 1/2 at 0 plus the χ^2 density scaled by a factor of 1/2. Thus, we need to correct the p-value as follows: if $\hat{p}$ is less than or equal to 0.5, the p-value is 1. Otherwise, it is $1 - F_{\chi_1^2}(\Lambda(\vec{Y}))/2$. We get the test statistic as follows:

$$L(p; \vec{Y}) = p^{n\hat{p}}(1-p)^{n-n\hat{p}}$$

$$\implies \Lambda(\vec{Y}) = 2\log\left(\frac{L(\hat{p};\vec{Y})}{L(1/2;\vec{Y})}\right) = 2\log\left(\frac{\hat{p}^{n\hat{p}}(1-\hat{p})^{n-n\hat{p}}}{1/2^n}\right) = 2\left(n\log(2) + n\hat{p}\log(\hat{p}) + (n - n\hat{p})\log(1-\hat{p})\right)$$

```
pchisq(2 * (n * log(2) + n * phat * log(phat) + (n - n * phat) * log(1 - phat)),
       df = 1, lower.tail = F) / 2
```

```
## [1] 0.02480051
```

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

Since $n\hat{p} \sim \text{Bin}(n, p)$, the exact p-value is given as $1 - F_{\text{Bin}(n,1/2)}(n\hat{p} - 1)$, the probability of observing data as or more extreme than was actually observed.

```
1 - pbinom(8-1, 10, 1/2)
```

```
## [1] 0.0546875
```