Mini-Project 2 Checkpoint 1

ECE/CS 498DS

Spring 2020

Akhilesh Somani (somani4)
Gowtham Kuntumalla (gowtham4)
Manan Mehta (mananm2)

Task 1 - Question 0

1. Why do biologists need multiple samples to identify microbes with significantly altered abundance?

Biologists need multiple samples to be sure that the data is statistically significant. Hypothesis needs to be backed by data. This helps them to conclude, with a greater confidence, which microbes are present in more numbers than usual.

2. Number of samples analyzed (in context of HE0): 764 samples
3. Number of microbes identified: 149

Task 1 – Question 1

- a. Factorization of joint probability distribution:

T = storage temp, M = collection method, C = contamination, L = Lab time, Q = quality

$$P(joint) = P(Q, C, L, M, T) = P(Q|C, L, M, T) * P(C|M, T, L) * P(M|T, L) * P(T|L) * P(L)$$
$$P(Q, C, L, M, T) = P(Q|C, L) * P(C|M, T) * P(M) * P(T) * P(L)$$

- b. Number of parameters needed to define conditional probability distribution:

Each feature takes 2 posisble values. For the CPTs:

P(Quality | Contamination, Lab Time): 4 parameters

P(Contamination | Storage Temp, Collection Method): 4 parameters

P(Storage Temp): 1 parameters

P(Collection Method): 1 parameters

P(Lab Time): 1 parameters

Thus, we need (4 + 4 + 1 + 1 + 1) = **11 parameters**

Task 1 – Question 1 (continued)

- C. Conditional probability tables:

P(Contamination | Storage Temp, Collection Method)

	strtmp	coll	cont = low	cont = high
0	cold	nurse	0.956017	0.043983
1	cold	patient	0.923423	0.076577
2	cool	nurse	0.911565	0.088435
3	cool	patient	0.161765	0.838235

P(Quality | Contamination, Lab Time)

	cont	labtime	qual = good	qual = bad
0	low	short	0.957093	0.042907
1	low	long	0.919003	0.080997
2	high	short	0.935743	0.064257
3	high	long	0.033898	0.966102

P(Storage Temp)	{'cold': 0.8982, 'cool': 0.1018}
P(Collection Method)	{'nurse': 0.8976, 'patient': 0.1024}
P(Lab Time)	{'short': 0.7956, 'long': 0.2044}

Task 1 – Question 1 (continued)

- d. Table of P(Quality|Storage Temp, Collection Method, Lab Time)

	strtmp	coll	labtime	qual = good	qual = bad
0	cold	nurse	short	0.955112	0.044888
1	cold	nurse	long	0.887962	0.112038
2	cold	patient	short	0.943978	0.056022
3	cold	patient	long	0.862069	0.137931
4	cool	nurse	short	0.972376	0.027624
5	cool	nurse	long	0.822785	0.177215
6	cool	patient	short	0.960784	0.039216
7	cool	patient	long	0.117647	0.882353

- e. Total number of samples dropped: 65 (for HE0) + 65 (for HE1) = 130 samples

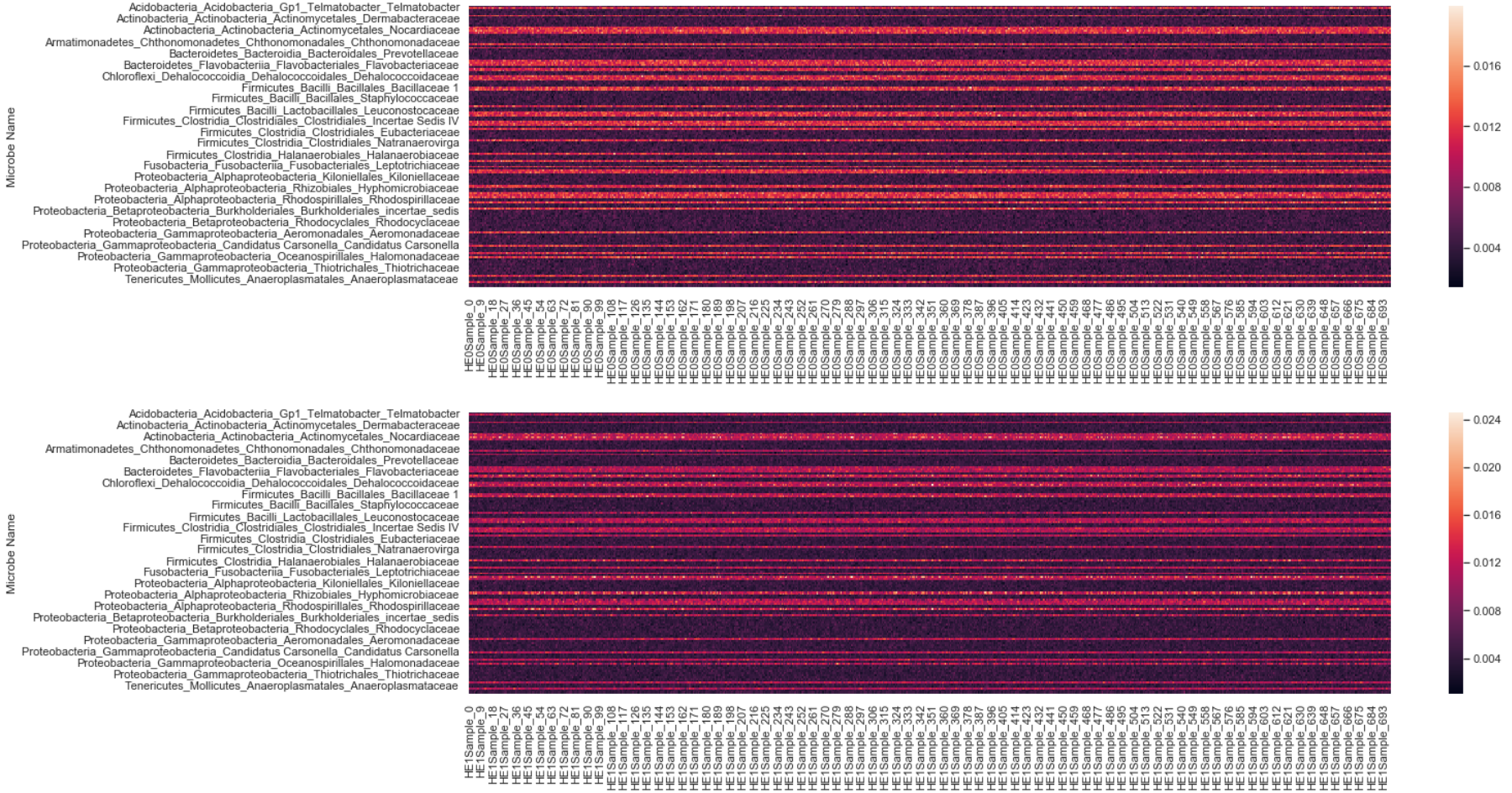
Task 1 – Question 2

- 1. Number of samples removed: 0
- 2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

While using relative abundance data, we have scaled the variance of the data and hence, we give equal emphasis to the variation for each bacteria. This normalization gives us a constrained snapshot of the relative distributions of microbes in a specific sample. There is a problem in doing this. We do not know the exact number of the bacteria present, which may be important to know rather than just the relative abundance. For e.g. - A relative abundance of 0.5:0.5 might mean 100:100 bacteria or 100k:100k bacteria. If there is a constraint on the number of bacteria to do some analysis, then this information is lost by scaling it.

Task 1 – Question 3

- Heatmaps (HE0 on top and HE1 on bottom):



Task 1 – Question 3 (continued)

- Summarize your observations

The heatmaps help in visualize at a glance the trend between the relative abundance of different bacteria in all the samples. The darker zones refer to low abundance and lighter zones correspond to higher abundance. A preliminary glance at the heatmaps tell us that the trend for relative abundance for all bacteria is same for both HE0 patients and HE1 patients. The heatmaps also show that the relative abundance of a particular bacteria among different samples is also same (which is expected because of data cleaning).

- Which aspects of the data are the heatmaps good at highlighting? What types of things are heatmaps less suitable for?

A heatmap is a graphical representation of data where values are depicted by color. Heatmaps make it easy to visualize complex data and understand it at a glance. The problem is that when we perceive shading, our brains tend to think in terms of relativities. That is, it notices sharp contrasts between adjacent bits of an image. However, we are poor at comparing shading in non-adjacent regions of a visualization.

Task 2 – Question 1

- b. What is the null hypothesis of the KS test in our context? Use one microbe as an example to explain your answer.

Ho for the KS Test is that the 2 samples tested are drawn from the same underlying distribution. In our context, it can be interpreted as no significantly altered expression of a particular microbe in the stool samples from HE0 and HE1 patients.

- c. Count the number of microbes with significantly altered expression at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Summarize your answers in a table below:

Alpha Level	Number of bacteria with altered expressions
0.1	50
0.05	37
0.01	27
0.005	26
0.001	21

Task 2 – Question 2

- a. What does a p-value of 0.05 represent in our context?

P-value, in general, is the probability of observing the test statistic or a more extreme value assuming $H0$ is true. In our context, a p-value of 0.05 represents a 5% probability of observing the KS test statistic (D-statistic), given that there is no significantly altered expression of the microbe in the HE0 and HE1 samples. In simple words, P-value of 0.05 represents 5% probability of rejecting H0 falsely. In our context, H0: for a microbe, both HE0 and HE1 sample follow same distribution.

- b. If the null hypothesis is true, what distribution will the p-values follow?

If the null hypothesis is true, the p-values will follow a unifom distribution. The reason is how we define $\alpha\alpha$ as the probability of erroneously rejecting $H0$H0. We reject $H0$H0 when p-value < $\alpha\alpha$ and the only way this holds for any value of $\alpha\alpha$ is when p is uniformly distributed.

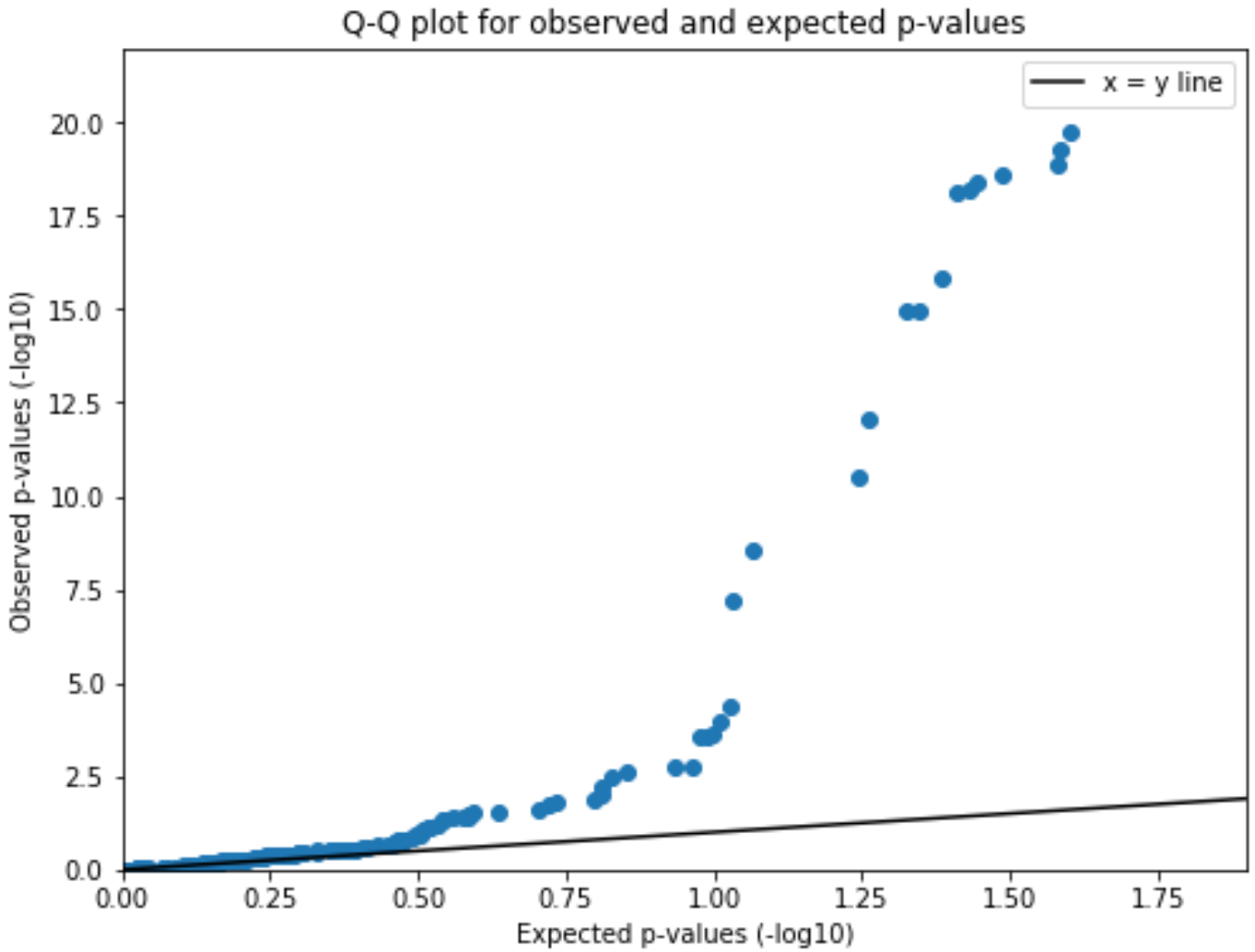
- c. If no microbe's abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.c. Show the comparison in a table below:

If no microbe's abundance was altered, which is to say that $H0$ is true, the significant p-values will be uniformly distributed. Thus, for an α value of 0.1, we expect to see 10% of the total number of samples (and so on). (We round the number of microbes to 150 instead of 149 here)

Alpha Level	Number of significant p-values if H0 true	Number observed from Task 2.1.c
0.1	15	50
0.05	8	37
0.01	2	27
0.005	1	26
0.001	0	21

Task 2 – Question 2 (continued)

- d. Q-Q plot:



Task 2 – Question 2 (continued)

- e.i. How does taking the -log10() of the p-values help you visualize the p-value distribution?

Function -log10 blows up the p-values closer to 0. For example -log(0.001) = 3 and -log(0.01) = 2. Data above 0.1 is less emphasized. This helps us focus more on the lower numerical values of p_value which are critical when making decision on elimination of H0

- e.ii. What can you conclude from the Q-Q plot?

Q-Q doesn't align with the x=y line hence the distributions are quite different, we can say expected and observed p-values follow different distributions. Assumption “H0 = True” is probably false. There is a difference between HE0 and HE1 samples and this difference is explained