Mini-Project 2 Checkpoint 1

ECE/CS 498DS

Spring 2020

Abhinavi Madireddy (am49), Rishabh Gupta (rishabh7),

Yasaswi Boyapati (yasaswi2)

Task 1 - Question 0

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

The biologists need multiple samples to identify microbes with significantly altered abundance as increase in number of samples increases the accuracy of the results and increases the confidence of accuracy of results.

2. Number of samples analyzed: 764
3. Number of microbes identified: 149

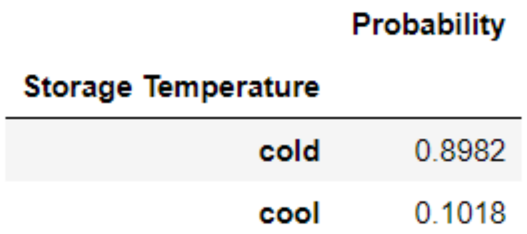
Task 1 – Question 1

- a. Factorization of joint probability distribution:

  S = storage Temp, M = collection method, C = contamination, T = lab time before processing, Q = quality
  $P(Q,C,S,M,T)=P(Q|C,T).P(C|S,M).P(S).P(M).P(T)$

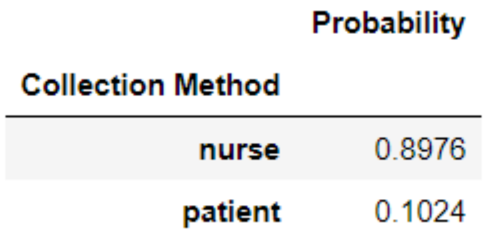
- b. Number of parameters needed to define conditional probability distribution: 11
- c. Conditional probability tables:

$P(S)$



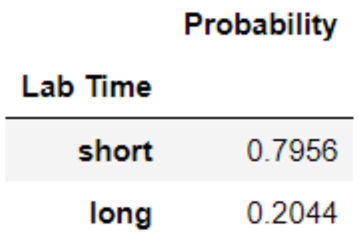
	Probability
Storage Temperature	
cold	0.8982
cool	0.1018

$P(M)$



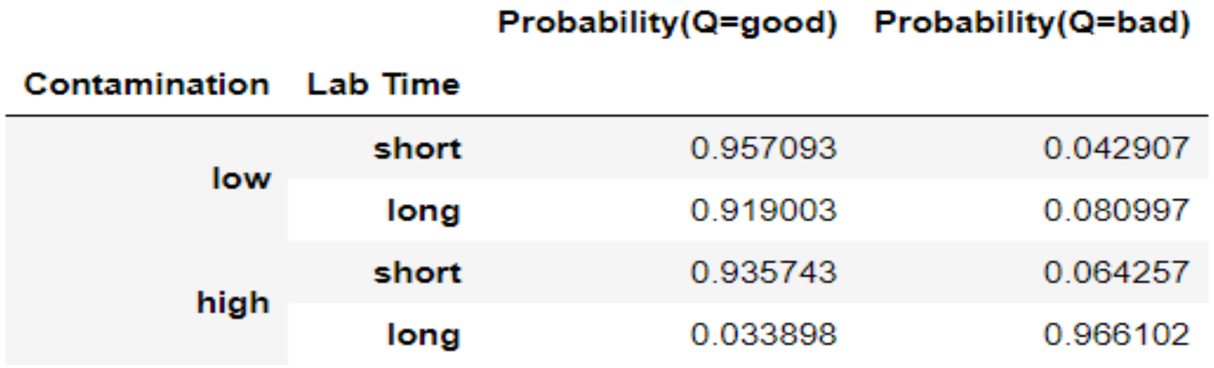
	Probability
Collection Method	
nurse	0.8976
patient	0.1024

$P(T)$



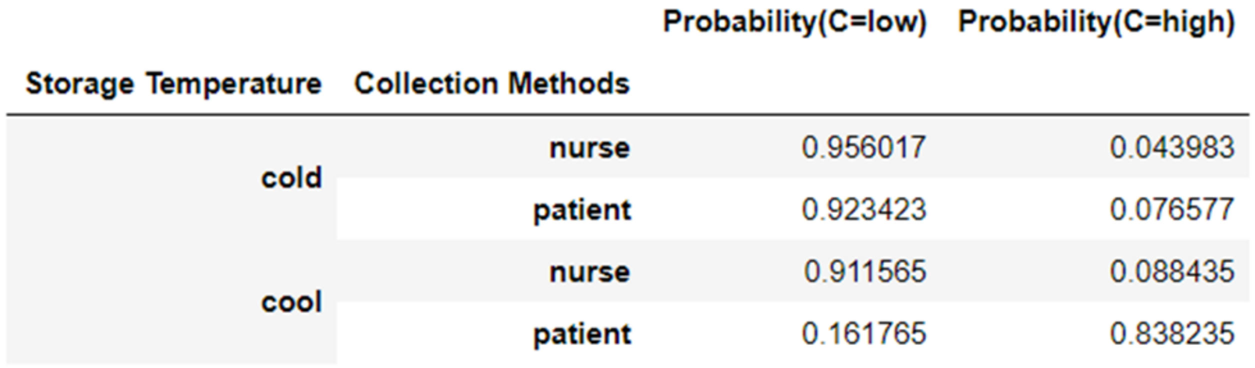
	Probability
Lab Time	
short	0.7956
long	0.2044

$P(Q|C,T)$



		Probability(Q=good)	Probability(Q=bad)
Contamination	**Lab Time**		
low	short	0.957093	0.042907
	long	0.919003	0.080997
high	short	0.935743	0.064257
	long	0.033898	0.966102

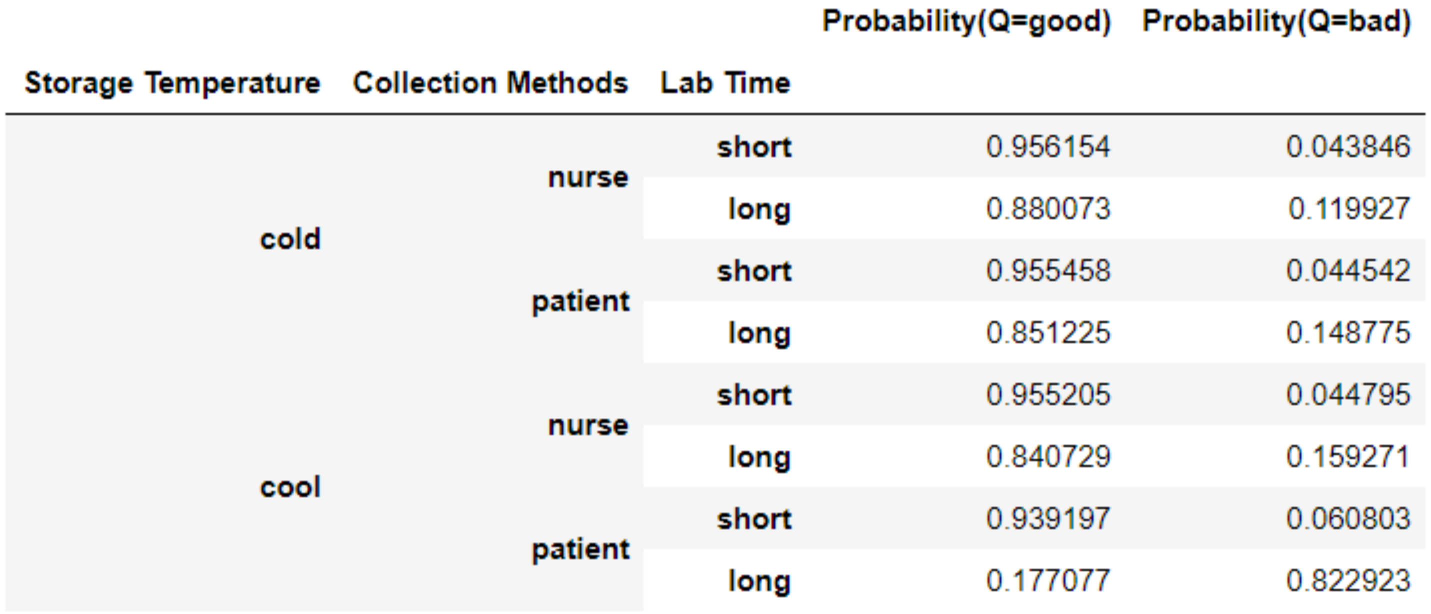
$P(C|S,M)$



		Probability(C=low)	Probability(C=high)
Storage Temperature	**Collection Methods**		
cold	nurse	0.956017	0.043983
	patient	0.923423	0.076577
cool	nurse	0.911565	0.088435
	patient	0.161765	0.838235

Task 1 – Question 1 (continued)

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



			Probability(Q=good)	Probability(Q=bad)
Storage Temperature	Collection Methods	Lab Time		
cold	nurse	short	0.956154	0.043846
		long	0.880073	0.119927
	patient	short	0.955458	0.044542
		long	0.851225	0.148775
cool	nurse	short	0.955205	0.044795
		long	0.840729	0.159271
	patient	short	0.939197	0.060803
		long	0.177077	0.822923

- e. Total number of samples dropped: 65

Task 1 – Question 2

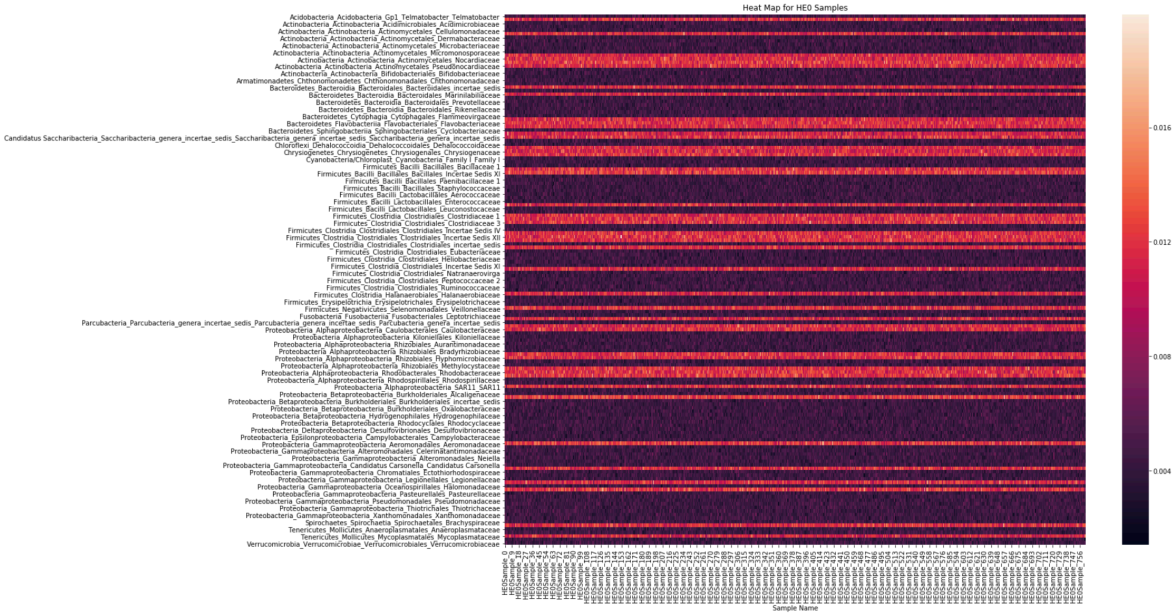
- 1. Number of samples removed: 0
- 2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

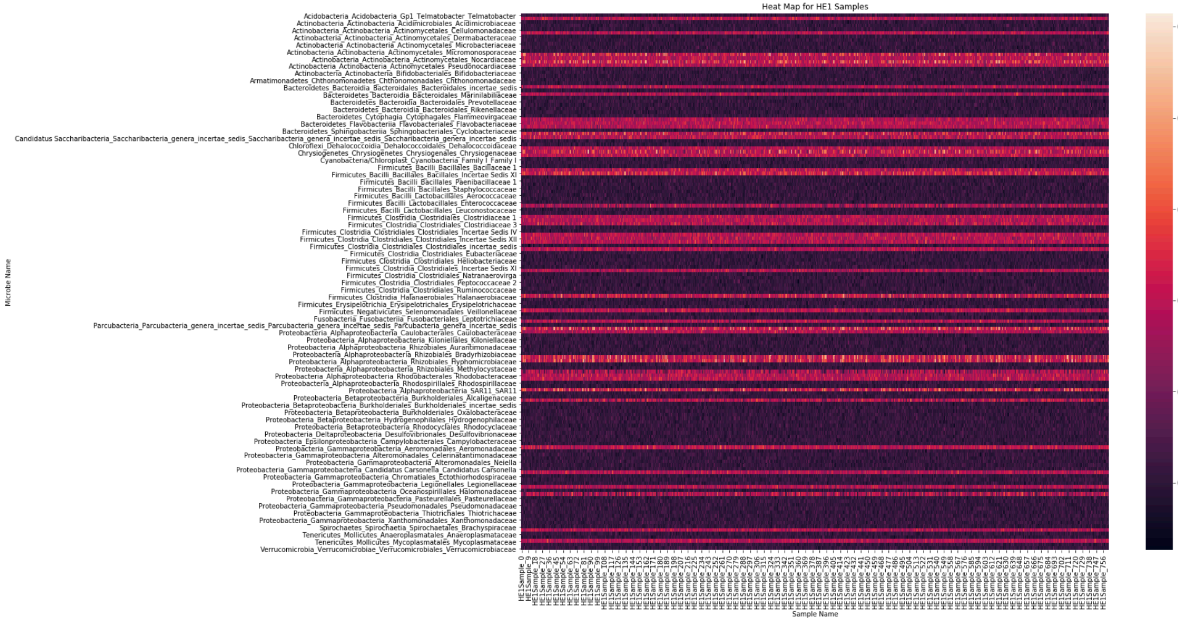
The benefits of relative abundance would be that it helps to easily realize which microbe is in highest and lowest quantity relative to others. But the drawback is we can't know the actual values of the samples as we only have relative abundance.

We may lose some information in rare cases, like in presence of extreme outliers.

Task 1 – Question 3

- Heatmaps (HE0 on left HE1 on right):





- Summarize your observations

From the heat map, we can see that each microbe is almost at same percentage in all samples. For example, in HE0 Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter is at approximately 0.004 in all samples and likewise for other microbes.

Task 1 – Question 3 (continued)

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

The heatmaps are good at highlighting the data that matter the most. But with data that does not contain variation, the heatmap do not provide any valuable information. In our case, the relative abundance looks almost consistent and heatmap is of no particular use.

Task 2 – Question 1

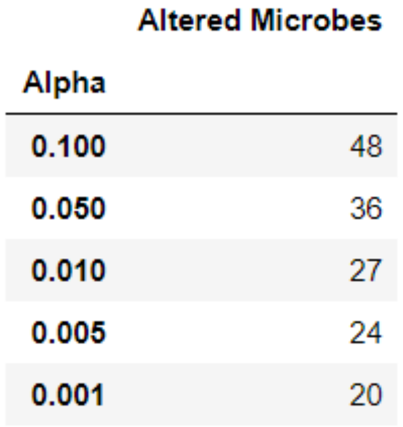
- b. What is the null hypothesis of the KS test in our context? Use one microbe as an example to explain your answer.

Considering the microbe- Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter for the KS test.

Null Hypothesis-

The distributions of Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter in HE0 and HE1 samples are same.

- 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:



	Altered Microbes
Alpha	
0.100	48
0.050	36
0.010	27
0.005	24
0.001	20

Task 2 – Question 2

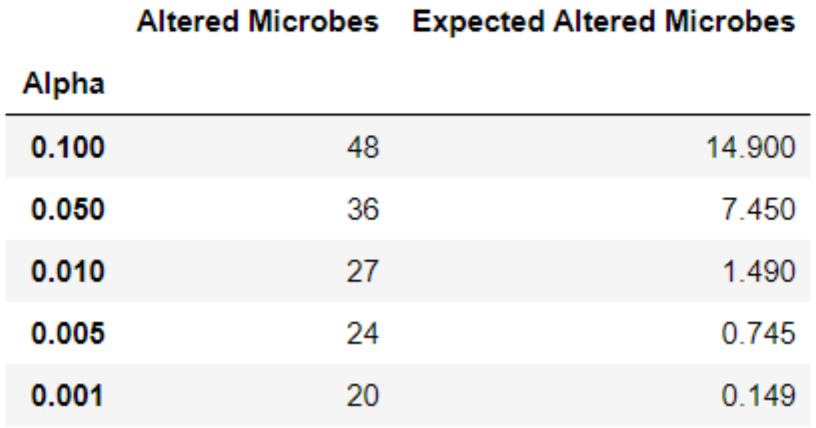
- a. What does a p-value of 0.05 represent in our context?

when p-value is less than 0.05, it means the samples do not come from same distribution, otherwise, they come from same distribution.

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

When the null hypothesis is true and the underlying random variable is continuous, then the probability distribution of the p-value is uniform on the interval [0,1]. However, if the alternative hypothesis is true, the distribution is dependent on sample size and the true value of the parameter being studied.

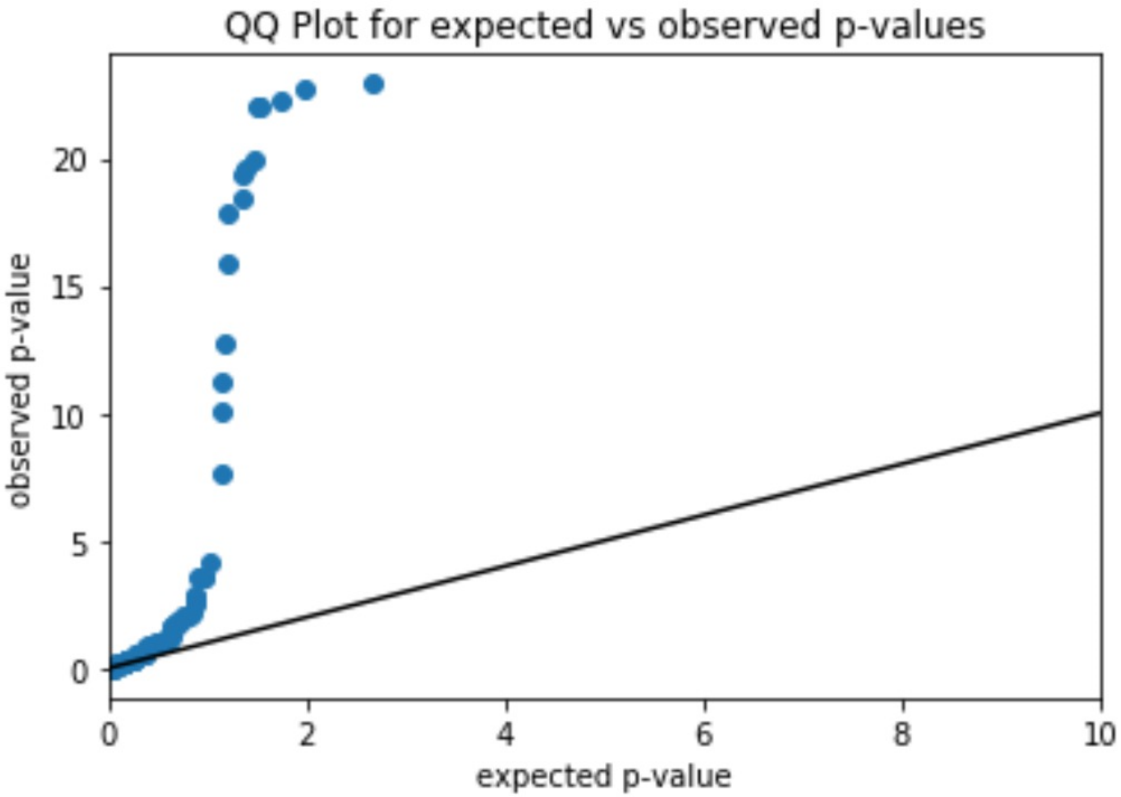
- c. If no microbe’s abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.c. Show the comparison in a table below:



	Altered Microbes	Expected Altered Microbes
Alpha		
0.100	48	14.900
0.050	36	7.450
0.010	27	1.490
0.005	24	0.745
0.001	20	0.149

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?

Taking –log10() of the p-values helps in normalizing the p –values, the scaled p-values distributions visualizations are clearer.

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

Since the values plotted do not fall over the x=y line, it can be concluded that the expected p-value and observed p-value do not belong to the same distribution.