Mini-Project 2 Checkpoint 1

ECE/CS 498DS

Spring 2020

Archit Patke (apatke), Haoran Qiu(haoranq4), Haoming Lu (hl36), Zhonghao Pan(zp3)

Task 1 - Question 0

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

In statistical analysis, sample size is an important consideration. Larger sample sizes provide more accurate mean values, identify outliers that could skew the data in a smaller sample and provide a smaller margin of error.

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

Task 1 – Question 1

- a. Factorization of joint probability distribution:

- P(Q, LT, C, CM, ST) = P(Q|LT, C) P(LT) P(C|S, CM) P(S) P(CM)

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

ST	P
Cold	0.8982
Cool	0.1018

CM	P
nurse	0.8976
patient	0.1024

LT	P
Long	0.2044
short	0.7956

P(Q|C,L)

Q	C	L	P
good	Low	Long	0.919
bad	Low	Long	0.081
good	low	Short	0.957
Bad	Low	Short	0.043
Good	High	Long	0.034
Bad	High	Long	0.966
Good	High	Short	0.936
Bad	High	Short	0.064

P(C|ST,CM)

C	ST	CM	P
low	cold	Nurse	0.956
high	cold	Nurse	0.044
low	cold	patient	0.923
high	cold	patient	0.077
low	cool	Nurse	0.912
high	cool	Nurse	0.088
low	cool	patient	0.162
high	cool	patient	0.838

Task 1 – Question 1 (contir

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

Q	T	CM	LT	P
Good	Cold	Nurse	long	0.88
bad	Cold	Nurse	long	0.12
Good	Cold	Nurse	Short	0.956
bad	Cold	Nurse	Short	0.044
Good	Cold	Patient	long	0.851
bad	Cold	Patient	long	0.149
Good	Cold	Patient	Short	0.956
bad	Cold	Patient	Short	0.044
Good	cool	Nurse	long	0.841
bad	cool	Nurse	long	0.159
Good	cool	Nurse	Short	0.955
bad	cool	Nurse	Short	0.045
Good	cool	Patient	long	0.177
bad	cool	Patient	long	0.823
Good	cool	Patient	Short	0.939
bad	cool	Patient	Short	0.061

- e. Total number of samples dropped:65

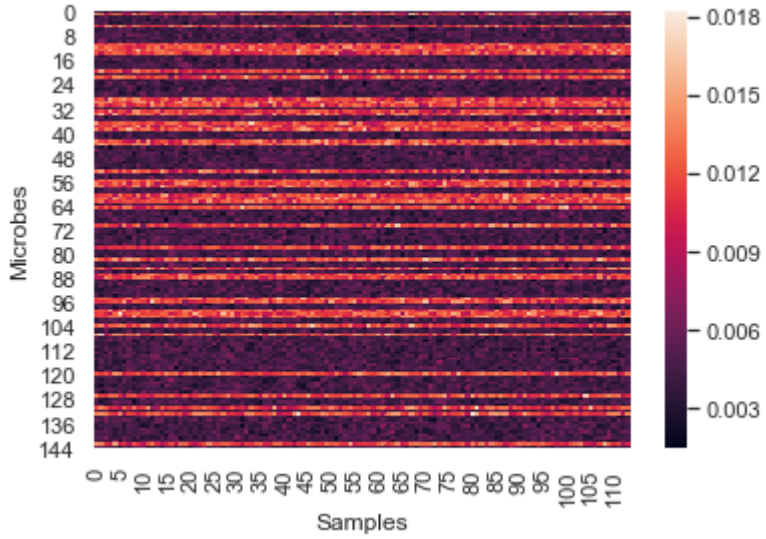
Task 1 – Question 2

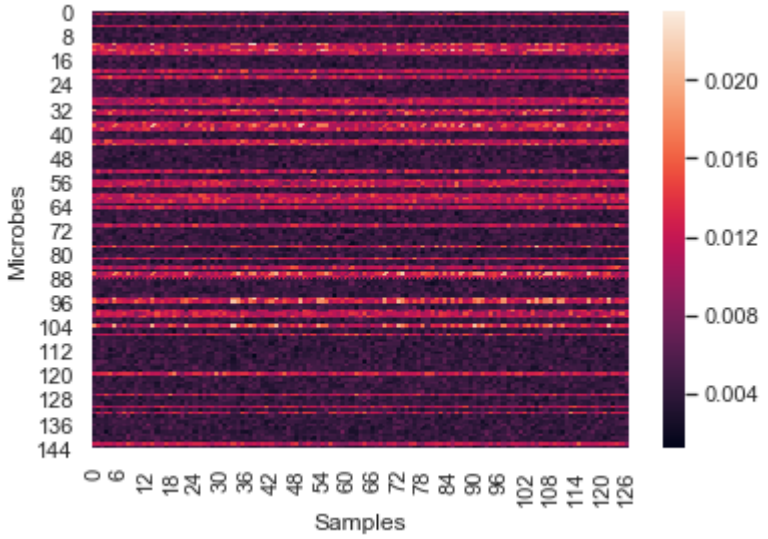
- 1. Number of samples removed:
  - HE0: 585
  - HE1: 572
- 2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

  Normalizing data allows input variables to have the same treatment in the model and the coefficients of a model are not scaled with respect to the scale of the inputs. Some machine learning algorithms require the input data to be normalized in order to converge to a good result. It also benefits seeking for relationship among two sets of data.However, the drawback is that we lost the original values (true data).

Task 1 – Question 3

- Heatmaps (HE0 on left HE1 on right):





- Summarize your observations

Samples within HE0 and within HE1 have similar microbe distributions. Compared to HE0, samples in HE1 have similar distribution patterns but darker (denser) amount of microbes. A type of microbe have similar relative abundance in HE0 and HE1.

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?

Heatmaps are graphical representations of data that utilize color-coded systems. The primary purpose of Heatmaps is to better visualize the volume of events within a dataset and assist in directing viewers towards areas on data visualizations that matter most. But they are not used for showing the real/true values. They are not used for showing the proportions of events (such as pie charts). They are not used for showing distribution of events.

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.

For 'Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter': H0 is that the distribution of 'Acidobacteria_Acidobacteria_Gp1_Telmatobacter_Telmatobacter' in samples of RelativeAbundance HE0 is the same as that of in samples of RelativeAbundance HE1

- 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	Number
0.1	32
0.05	25
0.01	15
0.005	12
0.001	9

Task 2 – Question 2

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

P-value is the probability of obtaining results as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.. Here, it means that the probability of the distribution of microbes in HE0 sample and that of HE1 sample are different at least as extreme as the results actually observed during the KS test, assuming that the two distribution are the same, is less than 0.05.

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

Uniform

- 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:

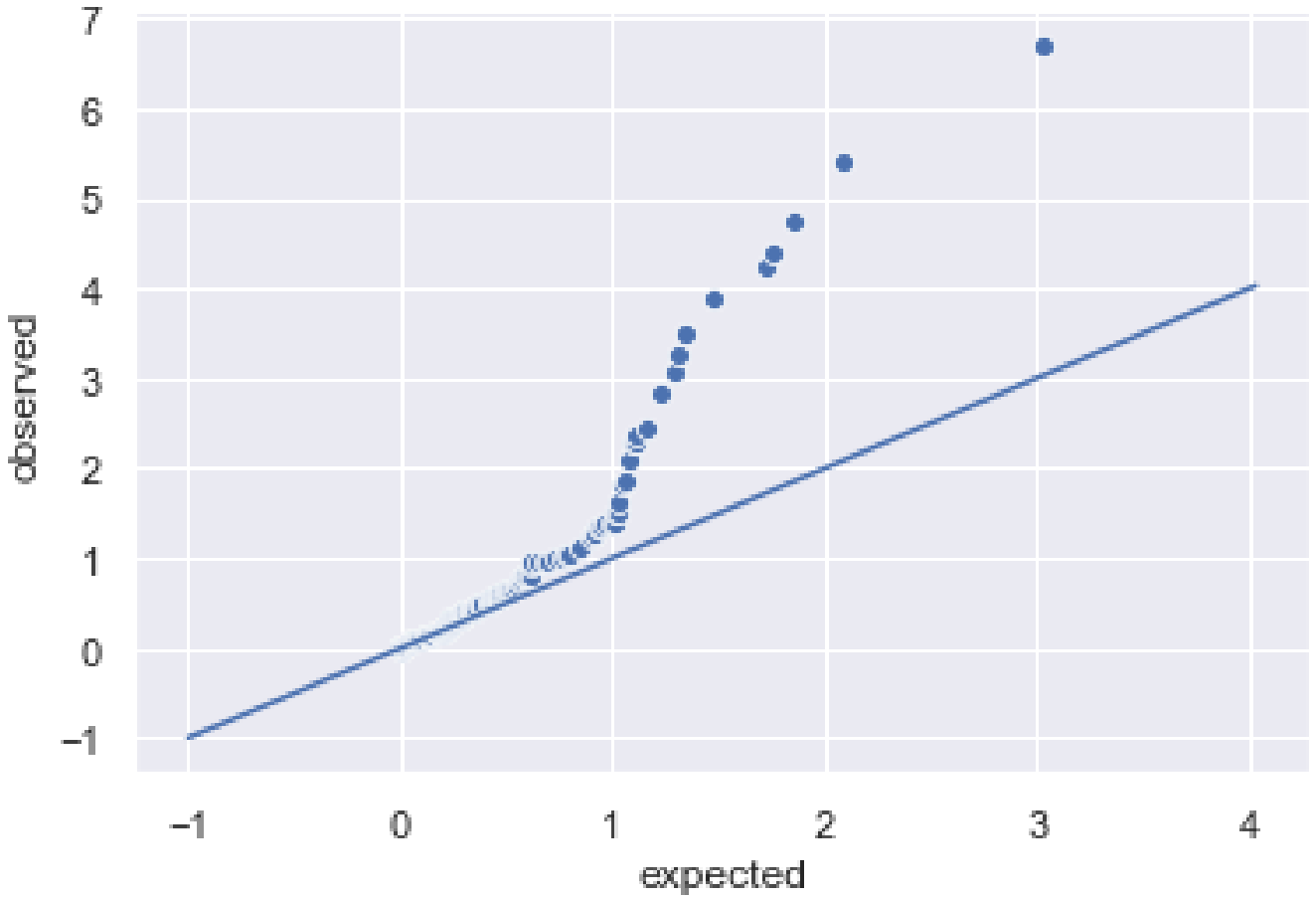
If no microbe's abundance was altered, the number would be

Smaller than 2.1c)

Alpha	Number	2.1C)
0.1	11.5	32
0.05	5.75	25
0.01	1.15	15
0.005	0.575	12
0.001	0.115	9

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?
- Since p values are small, taking log helps make smaller number larger.

- e.ii. What can you conclude from the Q-Q plot?
- Data relationship is not linear between expected and observed (it doesn't align with the x=y line), which means that expected and observed are not the same.