Mini Project II:
Unsupervised Stool Sample Analysis in Hepatic Encephalopathy

CS 498 DSU , Spring 2020

Cameron Ge(jg8)
Zinan Zhang(zinan2)
Anna Kalinowski(avk2)

Task 1:Data Cleaning and Visual Inspection

0. Getting started with Data

Q1:

Microbiomes vary from person to person, and multiple samples are needed to verify the specifics of each microbiome. Then, these samples are needed to figure out what the average abundance levels are and to identify any outliers. We can use the different sets of microbes to identify the causal relationship between these microbes and cirrhosis.

Q2:

There are 764 samples in total being analyzed

Q3:

In total 149 microbes are analyzed

1. Bayesian Network for quality control

A: Joint Probability Factorization

Factorization of the joint probability distribution:

$$P(storagetemp, collectionmethod, contamination, labtime, quality) =$$

$$P(quality|labtime, contamination)P(contamination|storagetemp, collectionmethod)P(labtime)P(storagetemp)P(collectionmethod)$$

B: Parameter Required:

The number of parameters required for defining the conditional probability:

$$P(quality|labtime, contamination)P(contamination|storagetemp, collectionmethod)P(labtime)P(storagetemp)P(collectionmethod)$$

=4+4+4=12

C : Constructing Probability Tables:

For P(storage temperature):

P(storage temp=cold)=0.8982

P(storage temp=cool)=0.1018

For p(collection method):

P(Collection Method = Nurse)=0.8976

P(Collection Method = Patient)=0.1024

For p(lab time):

P(Lab time = long)=0.2044

P(lab time = short)=0.7956

For P(contamination| storage temperature, collection method)

	strtmp	coll	cont = high	cont = low
0	cold	nurse	0.0439832	0.956017
1	cold	patient	0.0765766	0.923423
2	cool	nurse	0.0884354	0.911565
3	cool	patient	0.838235	0.161765

:

For P(Quality| Contamination, lab time):

	cont	labtime	qual = good	qual = bad
0	low	short	0.957093	0.0429069
1	low	long	0.919003	0.0809969
2	high	short	0.935743	0.064257
3	high	long	0.0338983	0.966102

D: Final Probability Tables:

	strtmp	coll	labtime	qual = good	qual = bad
0	cold	nurse	short	0.956154	0.043846
1	cold	nurse	long	0.880073	0.119927
2	cold	patient	short	0.955458	0.0445419
3	cold	patient	long	0.851225	0.148775
4	cool	nurse	short	0.955205	0.044795
5	cool	nurse	long	0.840729	0.159271
6	cool	patient	short	0.939197	0.0608033
7	cool	patient	long	0.177077	0.822923

E:

The criteria for bad quality data is that the probability that quality is good is under 0.5. By investigating our data set:

We sort the dataset and found out the only quality that fits P<=0.5 is:

	strtmp	coll	labtime	qual = good	qual = bad
7	cool	patient	long	0.177077	0.822923

Therefore we drop the events where lab time is long , storage temp is cool and collection method is by patient in the original data sets.

2. Data visualization

A: Verify Relative Abundance

```
b=newhe0.sum(axis=0)

x=[]
for j in range(1,b.size):
    if (b[j]>1+10e-10)|(b[j]<1-10e-10):
        x.append(j)
b.head()
print(x)
```

```
[]
```

```
b=newhe1.sum(axis=0)

y=[]
for j in range(1,b.size):
    if (b[j]>1+10e-10)|(b[j]<1-10e-10):
        y.append(j)
b.head()
print(y)
```

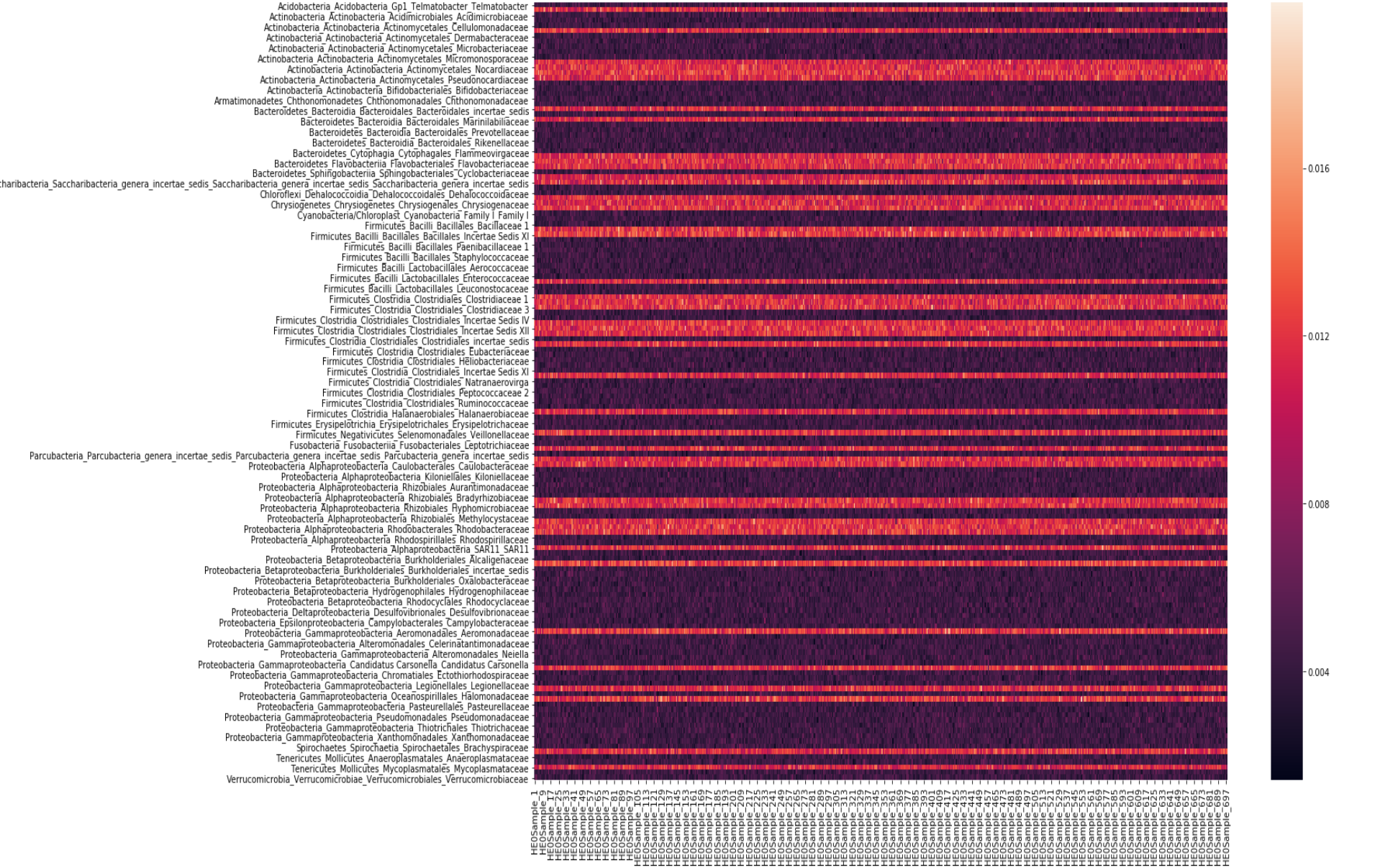
```
[]
```

By inspection no data trace exceeds the minimum error therefore we are provided with relative abundance data

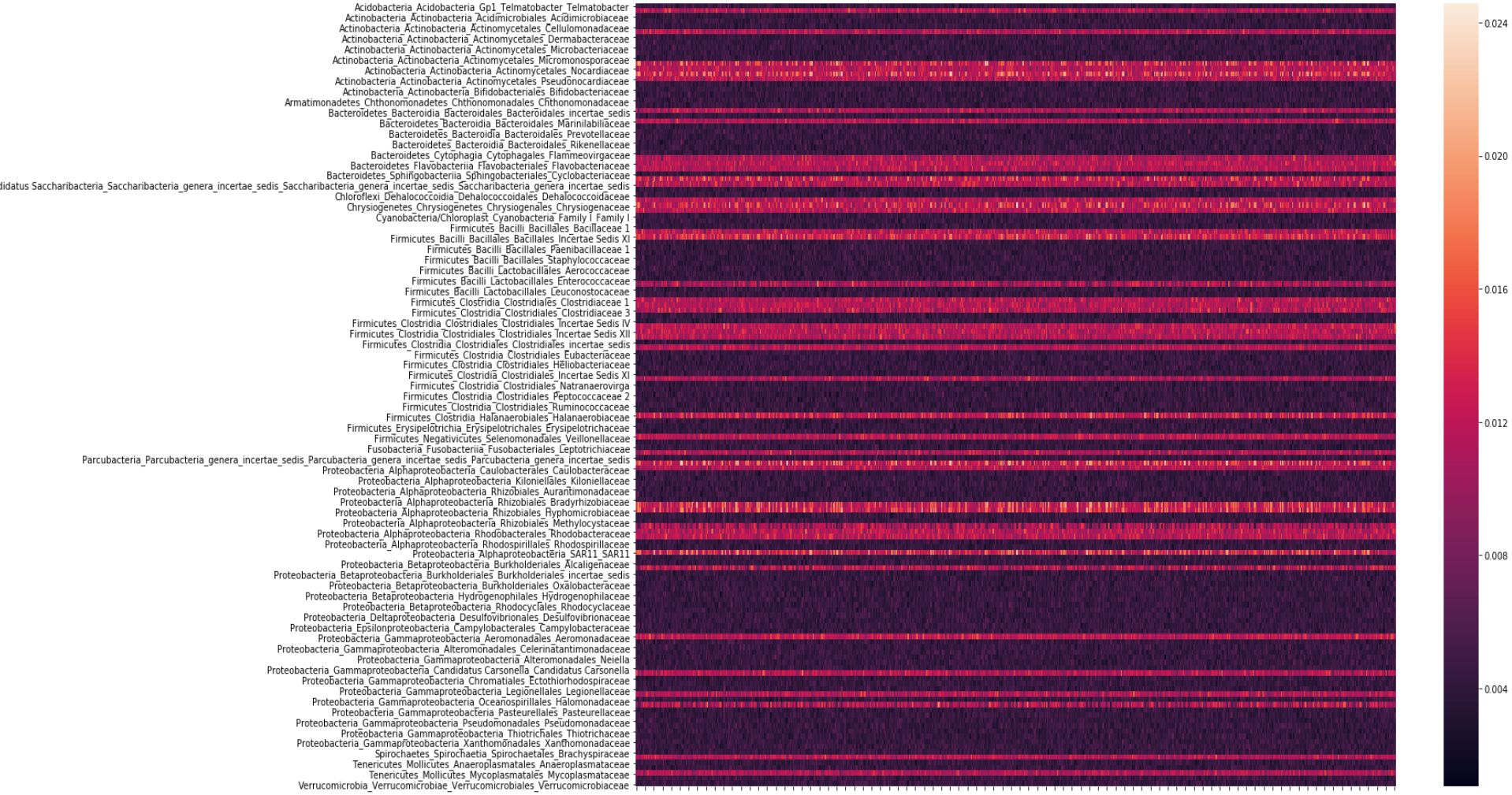
B: Benefits and drawbacks:

The benefits of the Relative Abundance Data is to get an intuitive impression conform to specific patterns that are among the best-known and most-studied patterns, which might be helpful to the study; however the Data might omit some of the components that are not considered within all the species, therefore these information would be lost if the normalization is performed.

Heatmap for HE0



Heatmap for HE1



Task 2:Statistical Analysis

1. Kolmogorov-Smirnov (KS) Test:

a. Calculate p-value.

Below is the head of first 10 trails of the p-value list for 149 microbes:

```
p_values = pd.Series(p_values)
p_values
```

```
: 0      1.740540e-01
  1      2.952649e-03
  2      1.056760e-01
  3      5.306311e-01
  4      9.335294e-01
  5      2.913301e-04
  6      3.035425e-01
  7      1.202459e-01
  8      7.545833e-01
  9      8.736905e-01
  10     6.650881e-01
```

b.

The Null hypothesis states that the microbe's abundance is not altered since the difference between the abundance in the two data sets for the very same microbe is not statistically significant.

c.

The number of microbes, with significantly altered alpha values, are shown in the right:

	alpha	count
0	0.100	50.0
1	0.050	37.0
2	0.010	27.0
3	0.005	26.0
4	0.001	21.0

2. Multiple Testing

a.

The p-value in this case is the threshold of whether a certain microbe's abundance is altered or not. When the p-value is smaller than 0.05 it means the null hypothesis should be rejected and teh microbe's abundance is altered, because the difference observed in the two samples are statistically significant. If the p-value is larger than 0.05 than the null hypothesis is not too be rejected thus the abundance is not altered.

b.

The p-value follows uniform distribution if null hypothesis is true.

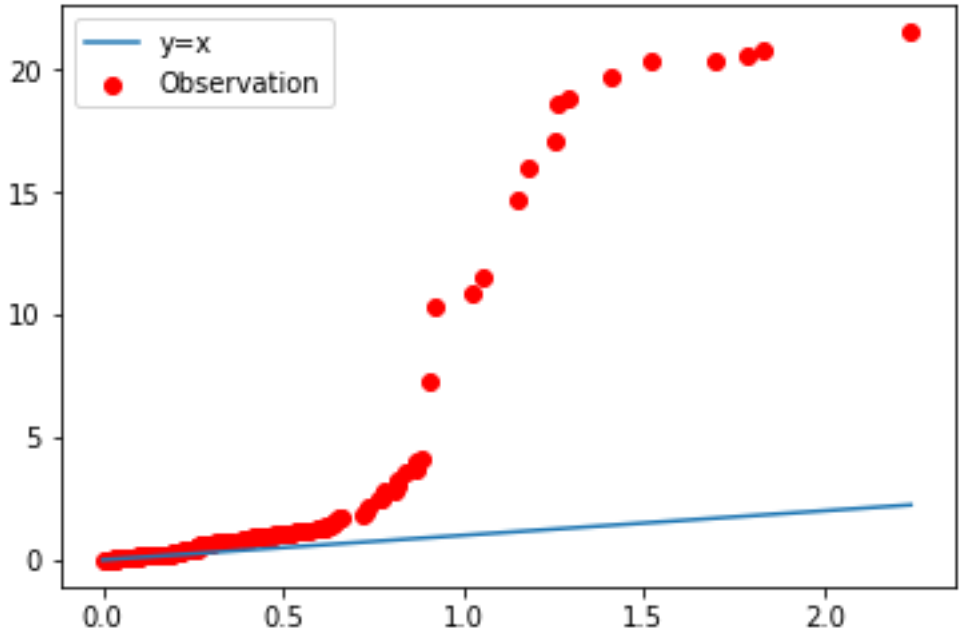
c.

Below is Expected Number of Significant p-values , in comparison to the actual observed number of significant altered microbes

	alpha	Expected	Actual
0	0.100	14.900	50.0
1	0.050	7.450	37.0
2	0.010	1.490	27.0
3	0.005	0.745	26.0
4	0.001	0.149	21.0

d.

A q-q plot is made on the right:



e.

i.

Taking the -log10 value of the p-values would make the graph more intuitive and visible(-log10(0,1)=0.1, 0.01 becomes 2, 0.001 becomes 3, etc.)

ii.

The plot does not aligh with x=y therefore the null hypothesis is not true