# **Report 1**

**Name: William Liu**

**E-mail:** [**dishenliu@gmail.com**](mailto:dishenliu@gmail.com)

**Last four digits of student ID: 1174**

1a)

Taking log of the data can restore symmetry to the data, thus gives normal or near normal distribution. This is shown in hw1item1a\_1.jpeg. Before, it was very skewed, most mass is on the left.

1b)

After standardization, it will be the Z distribution. Even though the normalized histography look the same as standardized one, however, standardized means: subtracting the mean and divided by the sd.

1c)

I will be taking 3 standard deviation(or 4 sd variables) of numbers because in normal distribution, which I have already done in this case, would have the coverage of 99.7% of the set. In addition, the P-value given beyond the 3rd S.D. is also insignificant, which would shows that I can roll the other S.D. out.

1d)

hw1item1c\_1.pdf and hw1item1c\_2.pdf

\_1 shows the raw data while \_2 shows the normalized and scaled version of the data.

1e)20 variables total: L1-L5, oL1-oL5, lag1-lag5, olag1-olag5

2a) hw1item2a.R

hw1item2a.pdf

2b) hw1item2b.R

hw1item2b.pdf

Question2c:

filename: hw1item2c.R

graph: hw1item2c.pdf

Questions2d:

filename: hw1item2d.R

graph: hw1item2d.pdf

Question2e:

filename: hw1item2e.R

graph: hw1item2e.pdf

Question2f:

filename: hw1item2f.R

The peaks is much smoother and able to max values. Threshold did not change. Classification still remain separated.

Question3:

filename: hw1item3.R

criteria:

I used Gaussian kernel for density estimation, and used default bandwidth for each density estimation,

If I got more than one intersections, I choose the one that is closest to median value of the parameter dataset and define that as the threshold.

I would choose ipdRatio and score for the subset of parameters used for classification, since they should have high accuracy

Question4:

filename: hw1item4.R

Using KNN classification, the criteria is if total of unmethylated is higher than unmethylated. Then test whether data is unmethylated or methylated. The model suppose to give accuracy for each parameter, as well as when k is different each time.

I believe best k would be k=5 because it should give the highest accuracy and have the most parameter that gives high accuracies. ipdRatio should have the highest accuracy for both models.

False negative:

False positive: