Nicholas Ang
CS461

Project 2 Report

The project takes data from a dataset called GSE19804_series with data on people with lung cancer or normal lungs. From the data, we perform multiple calculations on it to create various histograms and graphs.

In question 3, we calculate the log fold change of the data which is the difference in the mean of the log base 2 of the cancer and normal groups. We also calculate the p-value using the t-test. The data is used to create a volcano plot where differentially expressed data is when the absolute log fold change is greater than one and the p-value is less than the significance level of 0.05.

Question 5 has the data permuted and t-scores are calculated using t-test on the permuted data. Permutations are good because they can be used to estimate how the population is. It is useful in seeing where things are changing compared to the population. New t-scores are calculated from the permuted data and compared to the observed data. When there is a large change or the t-score is more extreme, it is counted up. The empirical p-value at a gene is the count summed up in a row where the NULL values are more extreme than the OBSERVED values divided by 100. The p-values are then stored in a p.T vector.

Question 6 is similar but used euclidean distance. The OBSERVED values come from the calculated euclidean distance from the original data. The NULL values come from the calculated euclidean distance of permuted data. The empirical p-value is calculated similarly to question 5 where it is the count summed up in a row where the NULL values are more extreme than the OBSERVED values divided by 100. The p-values are then stored in a p.E vector.

Question 7 takes the empirical p-values from question 5 and question 6 and creates a histogram for each respective question. Both the p.T histogram and the p.E histogram form graphs that are nearly identical to each other. This makes sense as we are calculating the p-values of each gene and seeing how often a gene is differentially expressed. The differentially expressed genes in the dataset can be seen from using either euclidean distance or t-test. When performing the tests on permuted data which simulates the population, the occurrence of the differentially expressed genes can be compared to the original data and occurs roughly the same as in the original. Since the differentially expressed genes occur roughly the same, the p-values of each gene are nearly the same. The correlation between p.E and p.T is 0.990105 which supports how the p-values are nearly identical.