Assignment-3

Y Rithvik

September 20, 2023

1 Implementation Summary

1.1 Importing Data

Implemented the following in *import_data* function

- The function takes a path to a text file as input.
- The function reads the data from each row of the text file, converts the entries of columns 1 to 49 into a list of floats, and appends them to the data list.
- The data list is then converted into a numpy array and returned.

1.2 Constructing N and D matrices

- Constructed the binary encoded N matrix relative to the null hypothesis which states that there exist 4 real numbers m, f, s, ns such that the means of the four underlying distributions are m+s, m+ns, f+s, f+ns respectively.
- Constructed the binary encoded D matrix relative to the alternate hypothesis which states that there exist 4 real numbers m_s, m_ns, f_s and f_ns such that the 4 underlying distributions could have arbitrary means m_s, m_ns, f_s and f_ns respectively.

1.3 Computing F-statistic and p-value

- Computed the value of the F-statistic of each row according to the following formula

$$\frac{1/(\text{rank}(D) - \text{rank}(N))}{1/(n - \text{rank}(D))} \times \left(\frac{X^T \left(I - N \left(N^T N \right)^{\dagger} N^T \right) X}{X^T \left(I - D \left(D^T D \right)^{\dagger} D^T \right) X} - 1 \right) \tag{1}$$

- Computed the probability mass for the $F(\text{rank}(D) - \text{rank}(N), n - \text{rank}(D))$ distribution that lies to the right of f-statistic value; which is the p-value by using *scipy.stats.f.cdf* function
- Plotted histogram of p-values ($bins = 25$) of all the rows using matplotlib.

2 Results

2.1 Plots

Histogram plot of p-values with 25 bins.

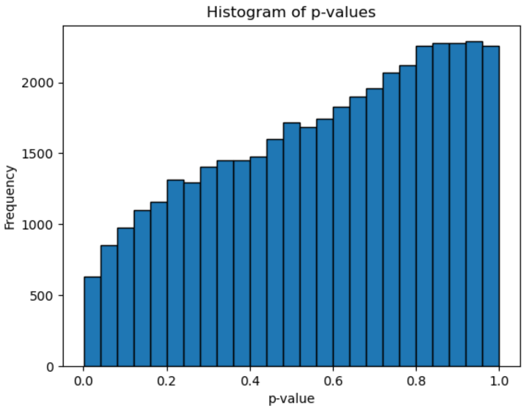

Figure 1: Histogram of p-values

2.2 Observations

- It is observed from the histogram that there are more rows with p-values closer to 1, this implies that there are more rows that reject the null hypothesis which further implies that there are more rows of interest.
- The Above observations might be because the same gene can be attributed to different probes and the histogram is for probes and not genes, and there might be more probes that detect the gene of interest.