DA Assignment 3

Shreeman Agrawal

September 2024

Assignment Problem Statement

Given: Data is generated from white blood cells from 48 individuals reference. A single file with 48 columns of data, plus some auxiliary columns

Problem: Identify the genes that respond different to smoke in men vs women (Smoking Status x Gender vs the Smoking Status + Gender null):

- Use the above 2-way ANOVA framework to generate p-values for each row.
- Draw the histogram of p-values.

2-Way ANOVA Framework : F-Statistic

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

- Null Hypothesis (numerator): The SmokingxGender interaction is purely additive, i.e., there exist numbers m,f,s,ns, such that the means of the four underlying distributions are m+s, m+ns, f+s, f+ns respectively (Fig. 1)
- The Alternative hypothesis (denominator): The SmokingxGender interaction is arbitrary, the 4 underlying distributions could have arbitrary means m_s, m_ns, f_s, f_ns respectively (Fig. 2)

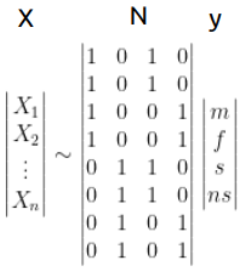

Figure 1: Null Hypothesis

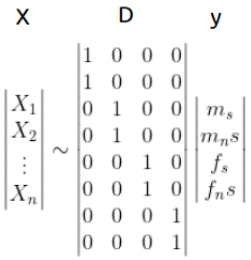

Figure 2: Alternate Hypothesis

Results

In the analysis, we calculate the F-statistic for each gene in the dataset to evaluate the significance of differences in gene expression between two groups, adjusted by the specific scaling factor. Then, p-values are derived from the F-distribution to assess the statistical significance of these differences.

The resulting histogram is plotted below (Fig. 3).

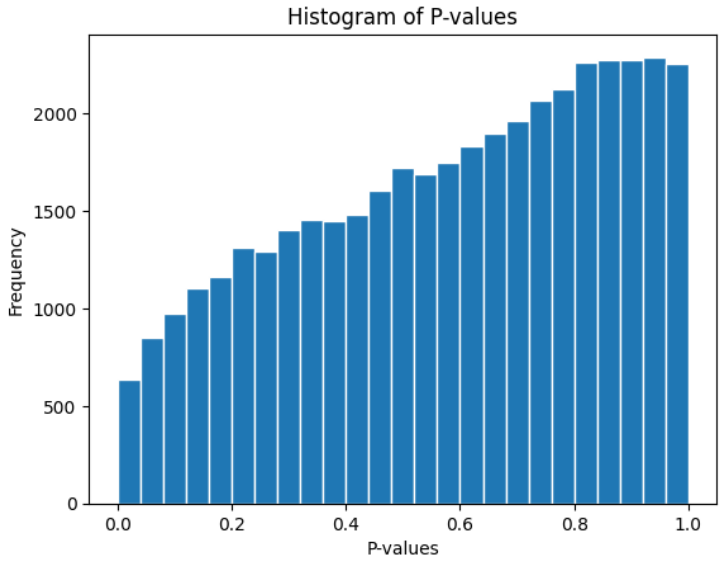

Figure 3: Histogram of p-values