The Effects of Smoking on Gene Expression

Ashish Ranjan

September 20, 2023

1 Introduction

This report presents an analysis of gene expression data to investigate the effects of smoking on gene expression patterns. The data was collected from white blood cells of 48 individuals, and we aim to identify which genes show significant differences in expression between different groups, including male non-smokers, male smokers, female non-smokers, and female smokers.

2 Data Description

The gene expression data consists of 41,094 probes, with each probe corresponding to a specific gene. The data is organized into four groups:

1. 12 Male Non-smokers (Samples 106-117)
2. 12 Male Smokers (Samples 118-129)
3. 12 Female Non-smokers (Samples 130-141)
4. 12 Female Smokers (Samples 142-153)

The values in the data are logarithmically transformed and have some zero values due to thresholding.

3 Data Processing and Implementation

3.1 Data Preprocessing

We began the analysis by loading the gene expression data and then removing the last three columns: Probe name, Gene Symbol, and Entrez Gene Id.

3.2 Hypothesis Matrices

To perform the two-way analysis of variance (ANOVA), we constructed two matrices, N (for the null hypothesis) and D (for the alternative hypothesis). The degrees of freedom for matrix N are 3, while matrix D has 4 degrees of freedom.

3.3 F-Statistics Calculation

The F-statistic for each gene was calculated using matrix operations involving N and D. These calculations are essential in testing the significance of gene expression differences across various groups.

3.4 P-Value Calculation

We used the scipy.stats library to calculate p-values based on the computed F-statistics. These p-values provide insights into the significance of the observed differences in gene expression.

3.5 Histogram of P-Values

To visualize the distribution of p-values, we created a histogram. The histogram was divided into 10 bins to provide a clear representation of the data.

4 Results

4.1 Histogram of P-Values

The histogram of p-values is presented below. It illustrates the distribution of p-values obtained from the analysis. The choice of 10 bins helps in visualizing the significance of the results.

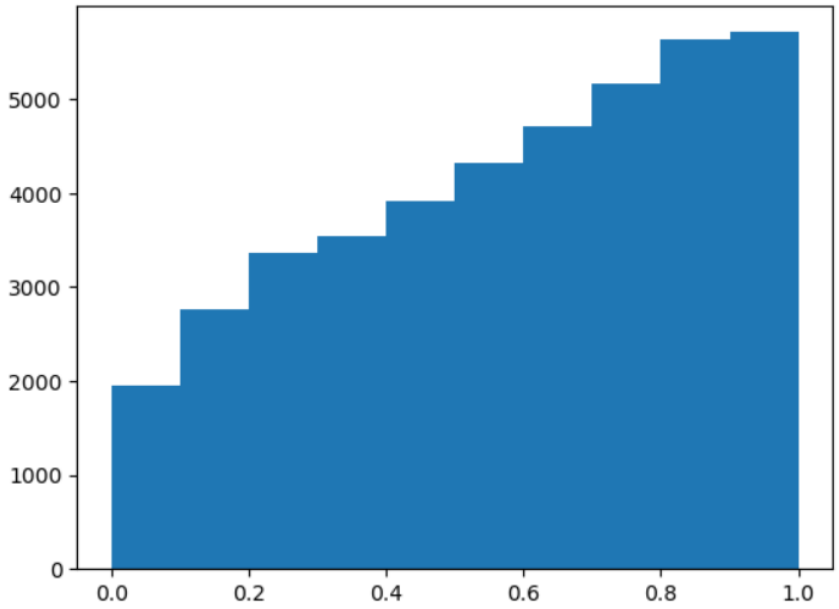

Figure 1: Histogram of P-Values

The histogram provides valuable information about the overall significance of the gene expression differences observed in the analysis.

5 Conclusion

In conclusion, our analysis of gene expression data suggests that certain genes exhibit significant differences in expression levels across different groups, particularly in relation to smoking status and gender. These findings provide valuable insights into the molecular mechanisms associated with smoking.