Effects Of Smoking Report

Aditya Manjunatha

Assignment 3

1 Setting the hypothesis

H_0 :- There is no interaction between smoking status and gender on gene expression.
The effect of smoking on gene expression is independent of gender. In other words, there is no interaction between smoking status and gender. The change in gene expression due to smoking is the same for both men and women.
H_1 :- There is an interaction between smoking status and gender on gene expression.
The effect of smoking on gene expression depends on gender. In other words, there is a statistically significant interaction between smoking status and gender. The change in gene expression due to smoking is different between men and women.

2 Functions used in assignment

2.1 load_and_preprocess_gene_data(filepath)

This is where are preprocessing our dataset.
The following are the modifications I made to the input file.

- Removed the 'Go' column as told in slides
- Removed Rows which had Nan as their GeneSymbol and GeneEntrezID
- Since Original data was in log base 2 scale. We exponentiate(ie raise to the power of 2) all the numeric-columns.

I am also keeping track of GeneIdentifiers (ie the GeneSymbol) as I will be using it later.

2.2 creating_matrices

Here we are constructing the A and B design matrices.
There are 4 groups (Male Non-Smoker, Female Non-Smoker, Male Smoker, Female Smoker) and 12 samples per group, making 48 total samples.

The matrices are filled with specific patterns representing the interaction model (matrix A) and the additive (null) model (matrix B).
Matrix A (Interaction Model):
Each row in matrix A represents a sample's group (gender and smoking status). This matrix explicitly models the interaction between smoking status and gender.
First column: Corresponds to the "Male Non-Smoker" group.
Second column: Corresponds to the "Female Non-Smoker" group.
Third column: Corresponds to the "Male Smoker" group.
Fourth column: Corresponds to the "Female Smoker" group.
Matrix B (Null Model):
Each row in matrix B represents the additive effects of gender and smoking status without interaction.
First two columns: Represent the gender effect (Male and Female).
Last two columns: Represent the smoking status effect (Non-Smoker and Smoker).
The two matrices are crucial for the 2-way ANOVA framework, as we compare these models to detect whether there's a significant interaction between smoking status and gender.

2.3 calculate_f_statistic

This function computes the F-statistic for a single gene (probe) based on its expression values.

Inputs:
Probe: A vector containing gene expression values for one probe across all samples.
A_mat: The design matrix A for the null model (no interaction).
B_mat: The design matrix B for the interaction model.
df_numerator and df_denominator: The degrees of freedom for the F-test.
F-statistic Calculation:
num = probe.T @ (A_mat - B_mat) @ probe:
This computes the numerator of the F-statistic, which measures how much the interaction model improves the fit compared to the null model.
den = probe.T @ (identity_matrix - A_mat) @ probe:
This computes the denominator, which represents the variance explained by the null model.
f_statistic = (num * df_numerator) / (den * df_denominator):
The F-statistic is the ratio of the improvement due to the interaction model over the residual variance of the null model. If the denominator is zero (indicating no variance), the function returns None.

2.4 calculate_p_value

This is a straighforward function which given the F-value for a row, gives the corresponding P-value.

2.5 compute_pvalues

Input:
data: The numeric matrix of gene expression values (one row per gene, one column per sample).
A and B: Design matrices representing the null and interaction models.
Matrix Operations:
A_mat = A @ pinv(A.T @ A) @ A.T:
This creates a projection matrix for the null model.
B_mat = B @ pinv(B.T @ B) @ B.T:
This creates a projection matrix for the interaction model.
Degrees of Freedom:
The function calculates df_num (degrees of freedom for the numerator) and df_den (for the denominator) based on the rank of A and B.
Loop over genes: For each probe (gene):
F-statistic: The function computes the F-statistic using the calculate_f_statistic function.
P-value: It calculates the corresponding p-value using the calculate_p_value function.
Store P-values: It stores the computed p-values for all genes.

2.6 Plotting the P values

This is the plot we get for 100 bins From here I got that there are 432 genes

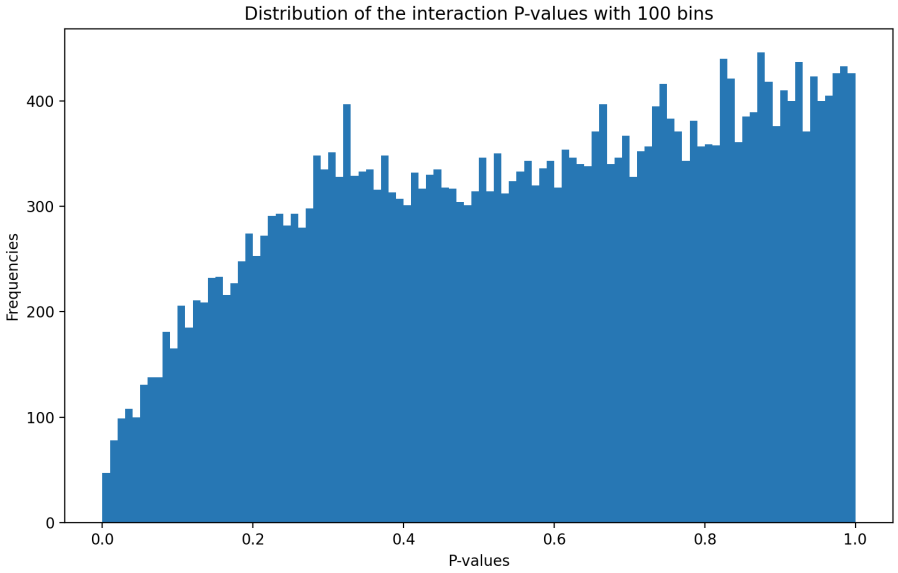

Figure 1: P value Histogram

whose P values are less than the significance level of 0.05. It means that for those genes, the interaction between smoking status and gender has a statistically significant effect on gene expression.
This means that the effect of smoking on gene expression differs between men and women for these genes.
The significant interaction implies that the change in gene expression due to smoking is not the same in men and women. For some genes, smoking might lead to an increase in expression in one gender and a decrease (or no change) in the other.
So we reject the Null Hypothesis for 432 genes.

2.7 Genes for which Pvalues are less than α

This is the link for the csv file of significant genes :-
It is stored in my one-drive. You must be able to access it. If not, please contact me :-
Click here for .csv file