E0-259-O Assignment 2 - Data Analytics

Authored by Deepank Dixit : SR - 20269, M.Tech AI (Online)

June 5, 2022

1 Assignment Problem Statement

As a first step to 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 Smoking Status x Gender(Alternate Model)

$$\begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{48} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \vdots & & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} male \\ female \\ smoker \\ non-smoker \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{48} \end{bmatrix}$$

3 Smoking Status + Gender (null model)

$$\begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{48} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & & & \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} male_{non-smoker} \\ male_{smoker} \\ female_{non-smoker} \\ female_{smoker} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{48} \end{bmatrix}$$

4 F-Statistic calculation

We iterate over all the 41093 rows of the dataframe and retrieve the gene probe measurements for all 48 individuals of each row (h). Then computed the f-statistic value using the following formula. At the end of for loop, we'll have 41093 f-statistic values.

$$\hat{f}_{Status,Gender} = \frac{\vec{\hat{h}}^T (A(A^T A)^\dagger A^T - A'(A'^T A')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I - (A(A^T A)^\dagger A^T)\vec{\hat{h}}} * \frac{n - rank(A)}{rank(A) - rank(A')}$$

Figure 1: f-stat formula

5 p-value calculation and Final output

Then using scipy.stats.f.cdf() function, we plot the F_statistics distribution using. We calculate the p-value using this distribution. The p-values generated for the whole 41093 rows of data is represented using a histogram as shown below.

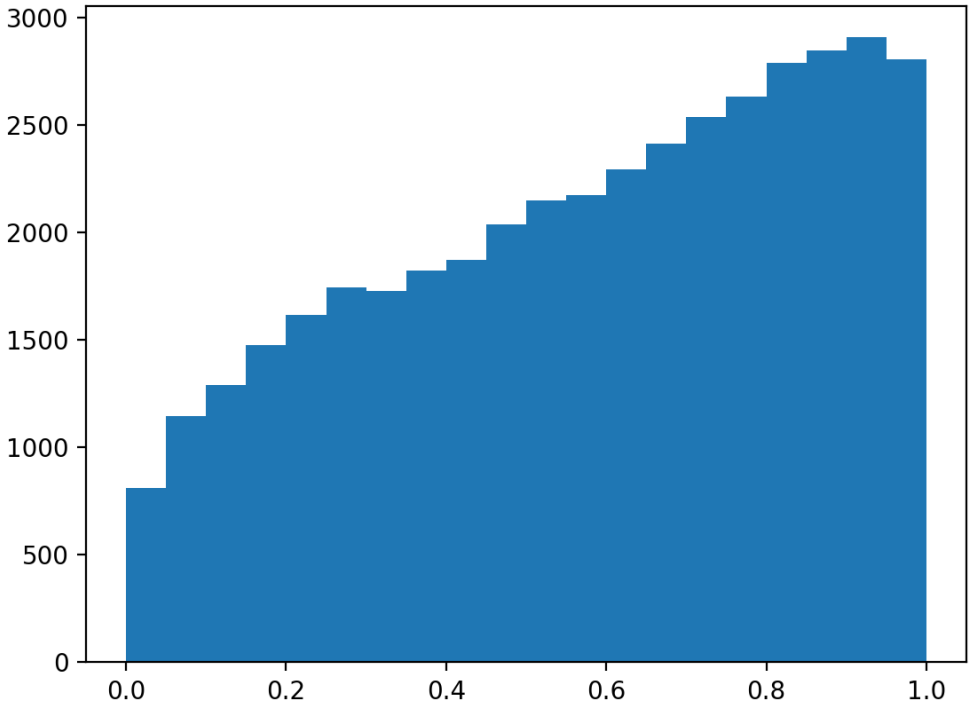

Figure 2: Histogram of p-values Generated