Data Analytics:Assignment-4

Kishalay Das (SR Number-15938)

October 30, 2019

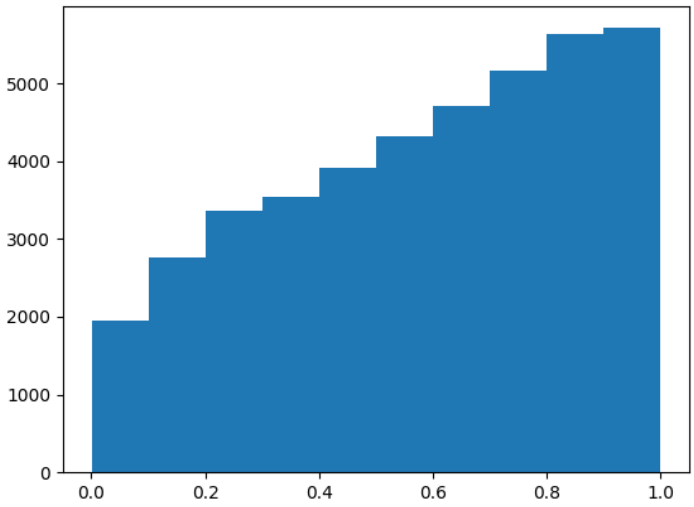
1.Analysis Effect of Smoking :

Part-1:Use the above 2-way ANOVA framework to generate p-values for each row

Approach Used :

1. Read the Raw-Data-GeneSpring.txt file and store 41093 probe values for different genes.
2. Generate A and A' matrix using the way described in the class.
3. Find ***F-Statistics*** for each probe given the formula in the slide.
4. Find ***P-Value*** using **scipy.stats.cdf** library function.

Part-2: Draw the Histogram of p-values



Part-3: Estimate n_0 for is derivable from this histogram

As we see from the hostogram the bias is closed to 1(Greater Density near 1 than 0), we take n as a conservative estimate of n_0.Hence estimate for n_0 is n i.e 41093

Part-4: Use an FDR cut-off of 0.05 to shortlist rows

As estimate for n_0 is n,so it is not possible to shortlist based on FDR.

Part-5:Create a shortlist of gene symbols from these rows

Refer Code.

Part-6: Intersect with the following gene lists: Xenobiotic metabolism, Free Radical Response, DNA Repair, Natural Killer Cell Cytotoxicity

Xenobiotic metabolism Genes:

GENE-ID	P-VALUE
SULT1A1	0.01642625207622117
AOC2	0.017717851223398196
CYP2S1	0.010012074301137153
AADAC	0.04924606948941401
HNF4A	0.03667822789527997
AS3MT	0.010454010328647234

Free Radical Response Genes:

No gene Found in the intersection.

DNA Repair Genes:

GENE-ID	P-VALUE
PNKP	0.049025748670447955

Natural Killer Cell Cytotoxicity Genes:

GENE-ID	P-VALUE
IFNG	0.042358049319651925
KLRC2	0.01869479068125468
PTPN6	0.008651502201786454
HLA-C	0.024270295367574524
PRF1	0.047549067003757495
HLA-E	0.03907015136369796
HLA-G	0.020670652893200914

Part-7:Report intersection counts for each list, split into four groups; going down in women smokers vs non-smokers/going up in women smokers vs non-smokers x ditto for men

Approach 1:Taken mean of all the values for individual probe for Male Smoker,Male NonSmoker,Female Smoker,Female NonSmoker

Women Smokers vs non-smokers Up Genes	'SULT1A1', 'AOC2', 'CYP2S1', 'HNF4A', 'PNKP', 'PTPN6', 'HLA-C', 'HLA-E', 'HLA-G'
Women Smokers vs non-smokers Down Genes	'AADAC', 'AS3MT', 'IFNG', 'KLRC2', 'PRF1'
Men Smokers vs non-smokers Up Genes	'AADAC', 'HNF4A', 'AS3MT', 'IFNG', 'KLRC2', 'PRF1', 'HLA-E', 'HLA-G'
Men Smokersvs non-smokers Down Genes	'SULT1A1', 'AOC2', 'CYP2S1', 'HNF4A', 'PNKP', 'PTPN6', 'HLA-C', 'HLA-E', 'HLA-G'

Approach 1:Taken Median of all the values for individual probe for Male Smoker,Male NonSmoker,Female Smoker,Female NonSmoker

Women Smokers vs non-smokers Up Genes	'SULT1A1', 'AOC2', 'CYP2S1', 'AADAC', 'HNF4A', 'KLRC2', 'PTPN6', 'HLA-C', 'HLA-E', 'HLA-G'
Women Smokers vs non-smokers Down Genes	'AS3MT', 'PNKP', 'IFNG', 'HLA-C', 'PRF1', 'HLA-E', 'HLA-G'
Men Smokers vs non-smokers Up Genes	'AOC2', 'AADAC', 'HNF4A', 'AS3MT', 'IFNG', 'KLRC2', 'PRF1', 'HLA-E', 'HLA-G'
Men Smokersvs non-smokers Down Genes	'SULT1A1', 'CYP2S1', 'PNKP', 'PTPN6', 'HLA-C', 'HLA-E'