Assignment-3

Data Analytics

1. 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):
2. Use the above 2-way ANOVA framework to generate p-values for each row.
3. Draw the histogram of p-values.

Extra effort (will not be graded).

1. Understand the content of the slides titled "Correction for Multiple Testing", "Estimating n0", and "False Discovery Rate (FDR)". These may not covered in the lectures.
2. See if a better (than n) estimate for n0 is derivable from the histogram; justify your estimate.
3. Use an FDR cut-off of 0.05 to shortlist rows.
4. Create a shortlist of gene symbols from these rows.
5. Intersect with the following gene lists: Xenobiotic metabolism, Free Radical Response, DNA Repair, Natural Killer Cell Cytotoxicity.
6. 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.
7. Do the same for men.