CSE 5520 Fall 2021
Homework 5 (Due 11:59 pm, Sunday, October 17, 2021 at HuskyCT)

Visualization of Hypothesis Testing and Network

This homework is to help you practice with some data visualization fundamentals. You are expected to use these visualization techniques and others in your final project. You are required to do this exercise in Python. All plots/graphs must have titles and x-y coordinate tick labels.

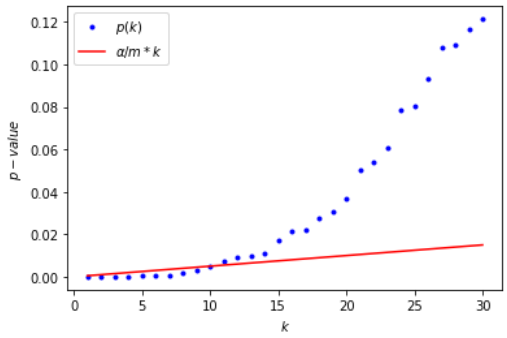
Part 1: P value in hypothesis testing

Consider again the gene expression data sets introduced in Homework 3/4. Your analysis is limited to comparing the female/male Dementia group and the female/male Non-Dementia group, using 'GSE44768_CR_40_54_Combined.csv'. Your study goal is to calculate p-values from multiple t-tests and correct p-values using Benjamini-Hochberg (BH) correction for 2 genes, LAMP2 and BSN.

Step 1. Focus on LAMP2. Your plan is to perform six different hypothesis testings (*t*-test) for this gene, **a)** Non-Dementia Female vs. Alzheimer's Female, **b)** Non-Dementia Female vs. Non-Dementia Male, **c)** Non-Dementia Female vs. Alzheimer's Male**, d)** Alzheimer's Female vs Non-Dementia Male, **e)** Alzheimer's Female vs Alzheimer's Male, **f)** Non-Dementia Male vs Alzheimer's Male. Calculate *p*-values based on *t*-test for each case. Discuss what you can say based on the p-values you got, assuming the significance level $\alpha = 0.05$.

Step 2. Produce 2x3 plot of 6 Gardner-Altman estimation plots from Step 1 a) ~f). The first row should include Gardner-Altman estimation plots of a), b) and c), and the second row Gardner-Altman estimation plots of d), e) and f),

Step 3. You like to perform Benjamini-Hochberg(BH) correction for the p-values found in Step1.
a) Plot sorted p-values (ascending order) in blue dots.



b) Draw a red line with slope $\frac{\alpha}{m}$ on top of the plot a), where α is the desired FDR (0.05) and *m* is total number of comparisons (6). Your plot should look like the plot give below, except in your case the range of x axis will be 1 ~ 6 (Note that we have only 6 *p*-values).

c) Show the largest *k* that is below the red line.

d) Show the corresponding *p*-values up to the largest k found in c). For example, the largest *k* that is smaller than the red line is 4, 4 smallest *p*-values should be recognized.

e) Plot the found *p*-values with green dots on top of the plot b).

Step 4. Add a Jupyter lab markdown cell to compare the Hierarchical clustering results you did for HW4 with what you have produced in Steps 1, 2, and 3. Do Hierarchical clustering and Gardner-Altman estimation plots show the similar analysis results? Do Hierarchical clustering and the raw p-values indicate the similar analysis results? How about the corrected p-values? Is there any difference between the p-values and corrected p-values when you contrast these values with the outcome from Hierarchical clustering? Please note that your goal is trying to build a

case by showing/contrasting both numerical values and plots (This course tile is “Data Visualization and Communication”!).

Step 5. Repeat Steps1 ~ 4 for the gene BSN.

Step 6. Extra Credit: If you think there is a better gene than LAMP2 or BSN to contrast various plots in the above steps, you are welcome to repeat the steps with the gene of your choice. Hint – You could use your heatmap you generated in HW4 to select what could provide an interesting case contrasting different plots (i.e., genes that can differentiate different patient groups).

Part 2: Network visualization

Consider the Alzheimer data set GSE44768_CR_alz_female_reduced.csv which is available at HuskyCT’s Data folder.

Step 1. Create and show a 50 x 50 correlation matrix for pairs of genes for the female Alzheimer’s patients for CR by appropriately thresholding their Pearson correlation coefficients. That is, you include top 50 pairs having higher “absolute” correlation coefficient values | r |.

Step 2. Use Networkx to turn the correlation matrix of Step 1 into a network of undirected edges. That is, you try to show how top 50 correlated genes could be interacting with each other. If r is positive, the edge should be red color and if negative, green color, as illustrated in lecture slide.

Step 3. This time, use Pyvis to repeat what you have done for Step 2.

You upload your Jupyter notebook in HuskyCT. The file name should be of the following format: HWn_Doe where n is the homework number and Doe denotes the last name.

HWs and Projects, 5% penalty for one day late submission. No acceptance after 5 days late. Extension is allowed only with the supporting medical record.