Introduction to Data Science (CSCI 4341)
Homework: Association and Statistical Inference
Due Date: November 7, 2021

1 Problem 1.

In this exercise, our aim is to quantify the association between genomic variants in the plant Arabidopsis Thaliana while assessing the statistical significance of their association. In this context, the variables are genomic variants and the samples are individual plants. We will use the two datasets that are provided with the assignment:

• The file “p1a.csv” contains a binary matrix of size 199 x 2 which indicates the presence of two genomic variants for 199 individuals. A value of 1 indicates that the genomic variant is present for the corresponding individual, and 0 indicates it is not.

For the two variables provided in “p1a.csv”, assess the association between them by computing each of the following statistics:

• Mutual Information

• Jaccard Index

• Chi-squared $\chi 2$

2 Problem 2

In this exercise, our aim is to quantify the associations between continuous variables and assess the statistical significance of these associations. For this purpose, we will use the three datasets that are provided with the assignment. Each column in these datasets represent a gene, and the value represents the expression of that gene in different samples. We would like to identify how these genes are associated.

• The file “p2a.csv”contains a matrix of size 2400 x 2 which has 2400 samples and 2 genes.

• The file “p2b.csv”contains a matrix of size 110 x 2 which has 110 samples and 2 genes.

• The file “p2c.csv”contains a matrix of size 2100 x 2 which has 2100 samples and 2 genes.

2.1 Part (a)

For the two variables provided in “p2a.csv”, assess the association between them by computing Pearson correlation ra and computing a p-value pa for the null hypothesis of no association. Select a significance level α and reject the null-hypothesis if the p-value is less than α. Explain, in complete sentences, your findings: Is there a statistically significant association (at α level) between the provided genes? What is the magnitude and the direction of the association?

Introduction to Data Science (CSCI 4341)
Homework: Association and Statistical Inference
Due Date: November 7, 2021

2.2 Part (b)

Repeat part (a) for the variable pair provided in “p2b.csv”and compute Pearson correlation rb and p-value pb. Select a significance level α and reject the null-hypothesis if the p-value is less than α. Explain, in complete sentences, your findings: Is there a statistically significant association (at α level) between the provided genes? What is the magnitude and the direction of the association?

2.3 Part (c)

Repeat part (a) for the variable pair provided in “p2c.csv”and compute Pearson correlation rc and p-value pc. Select a significance level α and reject the null-hypothesis if the p-value is less than α. Explain, in complete sentences, your findings: Is there a statistically significant association (at α level) between the provided genes? What is the magnitude and the direction of the association?

2.4 Part (d)

Draw scatter plots (variable 1 vs. variable 2) to visualize the data for all parts (a, b, c).

2.5 Part (e)

Explain your findings: Which variable pair (in part a or b or c) has a stronger association according to the comparison of the correlations? Which variable pair has a stronger association according to the comparison of the p-values? Do the comparisons according to correlation coefficients and p-values agree on which variable pair indicate the same stronger association? If not, why is there such a discrepancy? Which gene pair (in part a or b or c) has a stronger association do you think according to the scatter plots?