CSDS 313/413: Introduction to Data Analysis Fall 2025

Homework 3: Pairwise Association

©Vo Linh Chi Dao, Mehmet Koyutürk, Khoa Tran, Rui Yang, Crystal Zhu @ CWRU

There are two problems in this assignment.

1 Task 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.
- The file "p1b.csv"contains a binary matrix of size 199 x 15 which has 199 samples (individuals) and 15 variables (genomic variants). Each column represents a variable, and each entry (1 or 0) in the column indicates the presence of the corresponding variable in each of the 199 samples.

1.1 Part (a)

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

- Mutual Information
- Jaccard Index
- Pearson's chi-squared χ^2

For each of these statistics, assess the statistical significance by computing a p-value for the null hypothesis of "there is no association" (see *Note (1)* below). Select a significance level α that we can use to reject the null hypothesis and accept the alternate hypothesis "there is a non-zero association." if a p-value is less than α (see *Note (2)* below for the selection of α). For each of the three specified statistics, explain, in complete sentences, your findings: Is there a statististically significant association between the given variables? How strong does the association seem? Do the results of different statistics agree? If not, what does this mean (i.e., how do you explain this discrepancy)? Make sure to specify the values of the test statistics and the p-values as well as the selected α level and the number of permutations N you use for the permutation tests.

Note (1): The p-value in this context indicates the probability of observing a "more significant" result (for a statistic of interest like mutual information) than the observed result when there is no association between the genomic variants (which is our null hypothesis).

Note (2): You are welcome to select the commonly used 0.05 as your α threshold. However, you are encouraged to try different α values and observe how the conclusions of a researcher might change dramatically if they are not careful with the interpretations of the hypothesis test results.

Hint (1): For mutual information and Jaccard index, you need to do a Monte Carlo simulation (e.g., a permutation test) to assess statistical significance. For Pearson's χ^2, you can use the parametric distribution (i.e., Pearson's chi-squared test) instead of doing a Monte-Carlo simulation to generate the null distribution.

Hint (2): The idea of the permutation test is to randomly permute the entries in each column (variable) separately, in order to break any possible associations between the variables without modifying their distributions. This way, we can generate a null distribution for the test statistic (like mutual information) that represents our null hypothesis. Performing the permutation test is simple: First, generate N random permutations of the data and for each of these permuted data, compute the test statistic of interest. Then, count the number of times permuted data has more significant values (e.g., higher values for mutual information) than the actual data, let's denote this count c. Then, the p-value is simply equal to $\frac{c+1}{N+1}$. Notice that, a p-value obtained from a permutation test cannot be lower than $\frac{1}{N+1}$.

Hint (3): An alternative to a permutation test (or Monte-Carlo simulation in general) is to obtain a null distribution analytically by fitting a known parametric distribution. For example, it is known that Pearson's χ^2 follows a well-studied distribution called chi-squared distribution when the null hypothesis of "no association" is correct. Thus, it is possible to obtain a p-value analytically from the chi-squared distribution (indeed, this is what the Pearson's chi-squared test does).

Hint (4): To visualize the results of a permutation test, you can draw a sorted scatter plot while marking the observed value (as shown in 'example-figure1.jpg').

1.2 Part (b)

For each of the 105 (15×14/2) pairs of variables provided in "p1b.csv", repeat the steps in part (a) to compute the values of the test statistics as well as the p-values separately for: Mutual information, Jaccard Index and Pearson's chi-squared χ^2 statistics. Select an α level to reject the null hypothesis of no association. For each of the three test statistics, determine the pairs with significant association at α level while adjusting for false discovery rate (see *Note (3)* below). Draw scatter plots (or other appropriate visualizations) to compare the results of mutual information (MI), Jaccard Index (JI) and Pearson's chi-squared χ^2. How much do the results of these three test statistics agree to one another? Which two of these three statistics are most similar (MI-JI, MI-χ^2, or JI-χ^2)? If you were to use only one of these three statistics, which one would you prefer to use and why? If you were to consider only your preferred test statistic among the given three, would your conclusions (about the associations of the provided variable pairs) change a lot? Make sure to specify the number of significantly associated pairs and the number of overlap between the three test statistics. Also, specify the selected α level and the number of permutations N you use for the permutation tests.

Note (3): Here, differently from part (a), we are testing for multiple hypotheses (one for each variable pair). Therefore, if we were to reject a null hypothesis as suggested in part (a) (when

p-value is less than α), we would have a considerably higher false discovery rate (FDR) than α (Think: If each test has a α chance of failure, what is the probability that there is at least one test that fails? For $\alpha = 0.05$ and $n = 105$ tests, that would be 99.54% chance of failure!). Thus, we need to use an appropriate procedure that take into account the number of hypotheses to make sure the FDR is indeed less than α. For this purpose, you can apply one of the following two approaches: (1) Benjamini-Hochberg procedure (which is, in many fields, the go-to way of adjusting for multiple hypotheses) to determine the statistically significant pairs, (2) A Monte Carlo simulation specifically designed to control the FDR based on the distributions generated by the permutations (e.g., to compute the p-value of the top-scoring pair, we can use the distribution of the test statistics of the top-scoring pairs across all permutations, which would be a "multiple-hypthesis-corrected" null distribution).

Note (4): Using the Benjamini-Hochberg procedure, essentially a stricter threshold is used to reject a null hypothesis (e.g., the threshold of $\frac{\alpha}{n}$ for rejecting the null hypothesis of the most significant pair). Therefore, for estimating the p-values using permutation test, you need to use more permutations compared to part (a) to be able find a significant association. Thus, make sure that your number of permutations N is high enough to be able to find at least one significant pair.

2 Task 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:

- The file "p2a.csv"contains a matrix of size 2400 x 2 which has 2400 samples and 2 variables.
- The file "p2b.csv"contains a matrix of size 110 x 2 which has 110 samples and 2 variables.
- The file "p2c.csv"contains a matrix of size 2100 x 2 which has 2100 samples and 2 variables.

2.1 Part (a)

For the two variables provided in "p2a.csv", assess the association between them by computing Pearson correlation r_a and computing a p-value p_a 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 variables? What is the magnitude and the direction of the association?

2.2 Part (b)

Repeat part (a) for the variable pair provided in "p2b.csv"and compute Pearson correlation r_b and p-value p_b. Compare the Pearson correlations r_a and r_b as well as the p-values p_a and p_b. Explain your findings: Which variable pair (in part a or b) 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? Next, draw scatter plots (variable 1 vs. variable 2) to visualize the data for both part (a) and part (b). Which variable pair (in part a or b) has a stronger association do you think

according to the scatter plots? Does your conclusion agree with the comparisons of the p-values and correlation coefficients? If not, explain why.

2.3 Part (c)

Repeat part (b). But, this time compare the associations in the datasets "p2a.csv"and "p2c.csv". Make sure to answer the questions in part (b), this time for comparing the variable pairs in parts (a) and (c).