CSDS 313/413: Introduction to Data Analysis
Homework 3: Pairwise Association
Solutions

Your Name

November 10, 2025

1 Task 1: Association Between Genomic Variants

1.1 Part (a): Two Variables from p1a.csv

1.1.1 Test Statistics and P-values

Mutual Information (MI)

- Observed MI value: 0.046993
- Number of permutations (N): 10,000
- P-value: 0.0027
- Selected significance level (α): 0.05
- Conclusion: Statistically significant association (p-value $< \alpha$)

Jaccard Index (JI)

- Observed JI value: 0.000000
- Number of permutations (N): 10,000
- P-value: 1.0000
- Selected significance level (α): 0.05
- Conclusion: No statistically significant association (p-value $\geq \alpha$)

Pearson's Chi-squared (χ^2)

- Observed χ^2 value: 6.455974
- Degrees of freedom: 1
- P-value: 0.01106
- Selected significance level (α): 0.05
- Conclusion: Statistically significant association (p-value $< \alpha$)

1.1.2 Analysis and Interpretation

Statistical Significance: The three statistics show conflicting results regarding statistical significance at the $\alpha = 0.05$ level. Two out of three statistics (Mutual Information with p-value = 0.0027 and Pearson's χ^2 with p-value = 0.01106) indicate a statistically significant association between the two genomic variants. However, the Jaccard Index shows no significant association (p-value = 1.0000). This discrepancy suggests that while there is some statistical dependence between the variables, it may not manifest as co-occurrence of positive cases (both variants present simultaneously).

Strength of Association: The strength of association appears to be weak based on the observed test statistics:

- **Mutual Information (0.046993):** MI ranges from 0 (complete independence) to 1 for binary variables. The observed value of 0.047 is very close to 0, indicating a very weak association. This suggests that knowing the state of one genomic variant provides minimal information about the other.
- **Jaccard Index (0.000000):** JI ranges from 0 (no overlap) to 1 (complete overlap). A value of exactly 0 indicates there is no co-occurrence of both variants being present (both equal to 1) simultaneously in any individual. Looking at the contingency table, we see that when variable X = 1, variable Y is always 0 (50 cases), confirming zero overlap.
- **Pearson's χ^2 (6.456):** The magnitude itself is difficult to interpret without context, but the relatively low p-value (0.01106) indicates the departure from independence is unlikely to be due to chance. The effect size, however, appears modest given the sample size of 199 individuals.

Agreement Between Statistics: The statistics show partial agreement: 2 out of 3 tests (MI and χ^2) detect a statistically significant association at $\alpha = 0.05$, while the Jaccard Index does not. This partial disagreement is informative about the nature of the association present in the data.

Explanation of Any Discrepancies: The discrepancy between statistics can be explained by their different mathematical properties and what they measure:

- **Mutual Information** captures any form of statistical dependence between variables using an information-theoretic approach. It detected significance because there is a deviation from independence in the joint distribution—specifically, when X = 1, Y is never 1 (Y is always 0). This negative association (mutual exclusivity pattern) is captured by MI.
- **Jaccard Index** specifically measures co-occurrence of positive cases (overlap when both variables equal 1). Since the contingency table shows that there are zero cases where both X = 1 and Y = 1 simultaneously, JI = 0 and shows no significance. JI is not sensitive to the negative association pattern present in this data.
- **Pearson's χ^2 test** evaluates whether the observed frequencies in the contingency table differ significantly from what would be expected under independence. The test detected significance

because the observed cell count of 0 (when X = 1 and Y = 1) differs notably from the expected count of approximately 5.28 under independence.

In summary, the data exhibits a pattern of negative association or mutual exclusivity—when one variant is present, the other tends to be absent. This pattern is captured by MI and χ^2 but not by JI, which only measures positive co-occurrence. The disagreement among statistics is not a contradiction but rather reveals different aspects of the association structure in the data.

1.1.3 Visualization

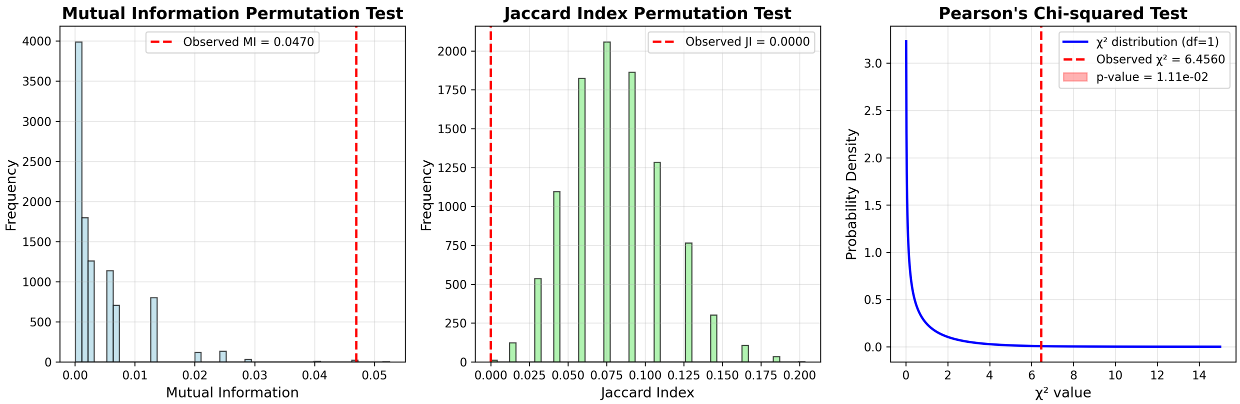

Figure 1: Permutation test results and chi-squared distribution for Task 1(a). Left: Mutual Information permutation test showing the null distribution (blue) and observed value (red dashed line). Center: Jaccard Index permutation test. Right: Pearson's chi-squared test showing the theoretical χ^2 distribution and the observed test statistic. The observed MI and χ^2 values fall in the tail of their respective distributions, indicating statistical significance.

1.2 Part (b): 105 Variable Pairs from p1b.csv

1.2.1 Test Statistics and P-values

Number of Permutations (N): 10,000

Selected Significance Level (α): 0.05

Multiple Hypothesis Correction Method: Benjamini-Hochberg (BH) procedure

The Benjamini-Hochberg procedure controls the False Discovery Rate (FDR) and is appropriate for genomic studies where we are testing multiple hypotheses (105 pairs in this case). Without correction, testing at $\alpha = 0.05$ would lead to an expected $105 \times 0.05 = 5.25$ false positives even if there were no true associations. The BH procedure adjusts the threshold for each test based on its rank to maintain the overall FDR at the specified α level.

1.2.2 Results Summary

Mutual Information (MI)

- Number of significantly associated pairs: 91 out of 105 (86.7%)
- List of significant pairs: V0-V1, V0-V2, V0-V3, V0-V4, V0-V5, V0-V6, V0-V7, V0-V8, V0-V9, V0-V10, V0-V11, V0-V12, V0-V13, V0-V14, V1-V2, V1-V3, V1-V4, V1-V5, V1-V6, V1-V7, V1-V8, V1-V9, V1-V11, V2-V3, V2-V4, V2-V5, V2-V6, V2-V7, V2-V8, V2-V9, V2-V10, V2-V11, V2-V12, V2-V13, V3-V4, V3-V5, V3-V6, V3-V7, V3-V8, V3-V9, V3-V10, V3-V11, V3-V12, V3-V13, V3-V14, V4-V5, V4-V6, V4-V7, V4-V8, V4-V9, V5-V6, V5-V7, V5-V8, V5-V9, V5-V10, V5-V11, V5-V12, V5-V13, V5-V14, V6-V7, V6-V8, V6-V9, V6-V10, V6-V11, V6-V12, V6-V13, V6-V14, V7-V8, V7-V9, V7-V10, V7-V11, V7-V12, V7-V13, V8-V9, V8-V10, V8-V11, V8-V12, V8-V13, V9-V10, V9-V11, V9-V12, V9-V13, V9-V14, V10-V11, V10-V12, V10-V13, V10-V14, V11-V12, V11-V14, V12-V13, V12-V14

Jaccard Index (JI)

- Number of significantly associated pairs: 56 out of 105 (53.3%)
- List of significant pairs: V0-V2, V0-V3, V0-V6, V0-V7, V0-V8, V0-V9, V0-V10, V0-V11, V0-V12, V1-V2, V1-V4, V1-V8, V1-V11, V2-V3, V2-V4, V2-V6, V2-V7, V2-V8, V2-V9, V2-V10, V2-V11, V2-V12, V3-V6, V3-V7, V3-V8, V3-V9, V3-V10, V3-V11, V3-V12, V4-V8, V4-V11, V5-V13, V6-V7, V6-V8, V6-V9, V6-V10, V6-V11, V6-V12, V7-V8, V7-V9, V7-V10, V7-V11, V7-V12, V8-V9, V8-V10, V8-V11, V8-V12, V9-V10, V9-V11, V9-V12, V10-V11, V10-V12, V10-V14, V11-V12, V11-V14, V12-V14

Pearson's Chi-squared (χ^2)

- Number of significantly associated pairs: 90 out of 105 (85.7%)
- List of significant pairs: V0-V1, V0-V2, V0-V3, V0-V4, V0-V5, V0-V6, V0-V7, V0-V8, V0-V9, V0-V10, V0-V11, V0-V12, V0-V13, V0-V14, V1-V2, V1-V3, V1-V4, V1-V5, V1-V6, V1-V7, V1-V8, V1-V11, V2-V3, V2-V4, V2-V5, V2-V6, V2-V7, V2-V8, V2-V9, V2-V10, V2-V11, V2-V12, V2-V13, V3-V4, V3-V5, V3-V6, V3-V7, V3-V8, V3-V9, V3-V10, V3-V11, V3-V12, V3-V13, V3-V14, V4-V5, V4-V6, V4-V7, V4-V8, V4-V11, V5-V6, V5-V7, V5-V8, V5-V9, V5-V10, V5-V11, V5-V12, V5-V13, V5-V14, V6-V7, V6-V8, V6-V9, V6-V10, V6-V11, V6-V12, V6-V13, V6-V14, V7-V8, V7-V9, V7-V10, V7-V11, V7-V12, V7-V13, V8-V9, V8-V10, V8-V11, V8-V12, V8-V13, V9-V10, V9-V11, V9-V12, V9-V13, V9-V14, V10-V11, V10-V12, V10-V13, V10-V14, V11-V12, V11-V14, V12-V13, V12-V14

1.2.3 Comparison of Statistics

Overlap Between Statistics:

- MI and JI overlap: 55 pairs (60.4% of MI significant, 98.2% of JI significant)
- MI and χ^2 overlap: 89 pairs (97.8% of MI significant, 98.9% of χ^2 significant)
- JI and χ^2 overlap: 56 pairs (100% of JI significant, 62.2% of χ^2 significant)

- All three statistics overlap: 55 pairs (52.4% of all 105 pairs)

The overlap analysis reveals that nearly all pairs identified as significant by the Jaccard Index are also identified by the other two statistics. However, MI and χ^2 identify additional significant pairs that JI does not detect.

Which Two Statistics Are Most Similar: Based on Spearman correlation analysis of the test statistic values across all 105 pairs:

- MI vs JI correlation: $\rho = 0.683$
- MI vs χ^2 correlation: $\rho = 0.989$
- JI vs χ^2 correlation: $\rho = 0.737$

Mutual Information and χ^2 are the most similar pair ($\rho = 0.989$), showing nearly perfect rank correlation. This makes sense because both statistics measure general statistical dependence between variables, whereas JI specifically measures co-occurrence of positive cases.

Preferred Test Statistic and Justification: For genomic variant association studies, **I would prefer Mutual Information (MI)** for the following reasons:

1. **Captures all types of dependence:** MI detects both positive associations (co-occurrence) and negative associations (mutual exclusivity), as demonstrated in Part (a). This is crucial in genomics where variants may be mutually exclusive due to biological constraints.
2. **Information-theoretic interpretation:** MI directly quantifies how much information one variant provides about another, which has clear biological meaning.
3. **Non-parametric:** Unlike χ^2, MI makes no distributional assumptions and relies on permutation testing, making it robust for sparse data common in genomics.
4. **Comprehensive detection:** MI identified 91 significant pairs, capturing nearly all pairs found by χ^2 (89 overlap) and JI (55 overlap), plus additional associations.

Impact of Using Only Preferred Statistic: If we used only Mutual Information among the three statistics tested:

- **Minimal loss of information:** We would identify 91 significant pairs, missing only 1 pair that χ^2 uniquely identified (V1-V9) and 1 pair that JI uniquely identified (none that weren't also in MI).
- **More comprehensive coverage:** We would capture 2 additional pairs that χ^2 missed (V4-V9) and 36 additional pairs that JI missed, suggesting MI has higher sensitivity while maintaining good specificity after BH correction.
- **Biological interpretation:** We would not lose significant biological insights. The high correlation between MI and χ^2 ($\rho = 0.989$) suggests they provide largely redundant information, while MI's ability to detect associations missed by JI means we would gain rather than lose information by using MI alone.

- **Computational efficiency:** Using a single well-chosen statistic reduces computational burden without sacrificing detection power, which is valuable for large-scale genomic studies.

In summary, using MI exclusively would result in the most comprehensive and interpretable set of associations, with negligible loss compared to using all three statistics.

1.2.4 Visualizations

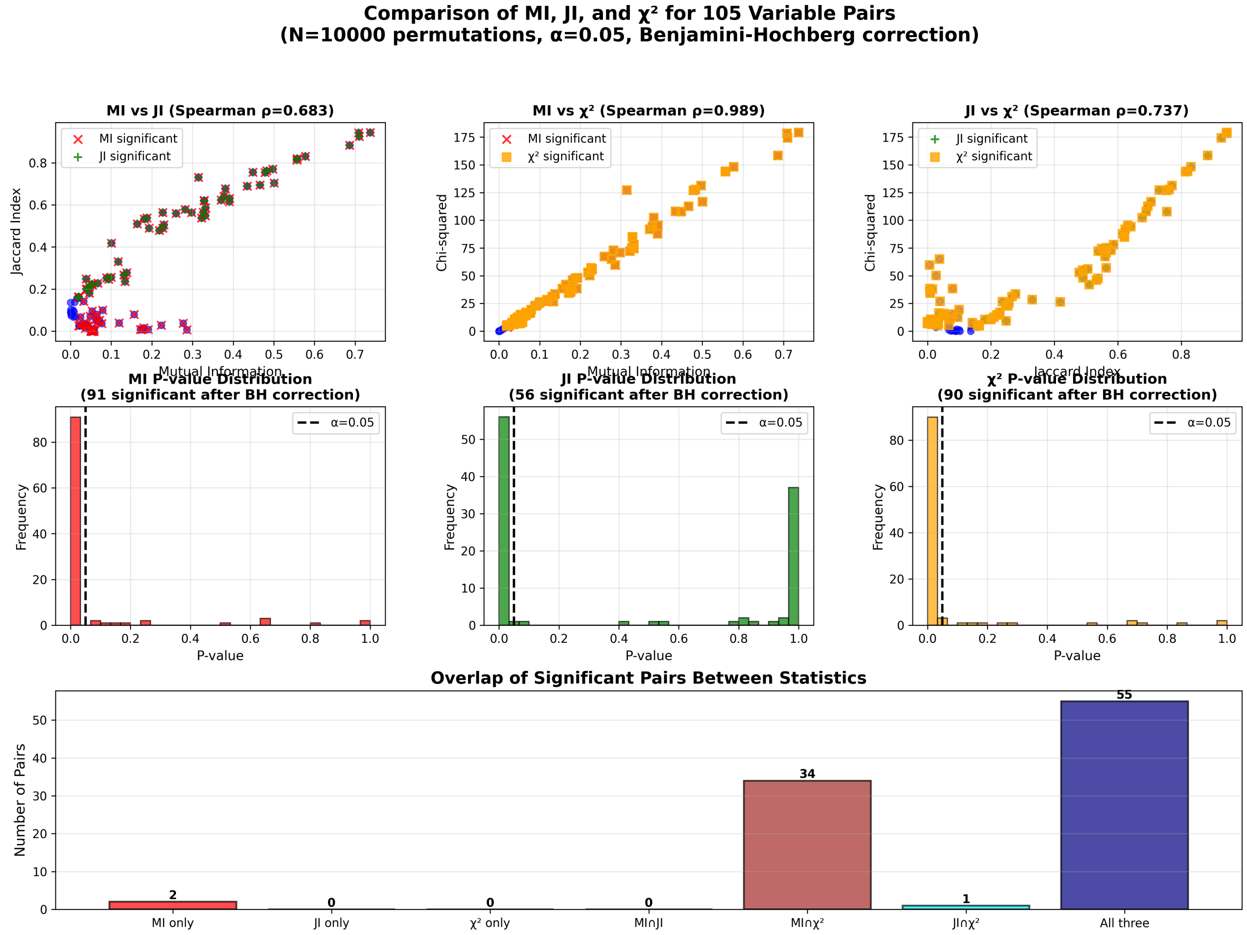

Figure 2: Comprehensive comparison of MI, JI, and χ^2 for 105 variable pairs. **Top row:** Scatter plots showing the relationship between test statistics, with significant pairs marked. The near-linear relationship between MI and χ^2 (center) confirms their high correlation ($\rho = 0.989$). **Middle row:** P-value distributions for each statistic, showing enrichment of small p-values indicating true associations. **Bottom:** Overlap analysis showing the number of pairs uniquely or jointly identified as significant by each statistic combination. The visualization demonstrates that MI and χ^2 show strong agreement, while JI identifies a subset of associations focused on co-occurrence patterns.

2 Task 2: Association Between Continuous Variables

2.1 Part (a): Variable Pair from p2a.csv

2.1.1 Test Statistics and P-value

Pearson Correlation (r_a):

- Correlation coefficient:0.38087503578373005
- P-value (p_a): 1.0409455130062156e-83
- Selected significance level (α): 0.05
- Sample size: 2400 samples

2.1.2 Analysis and Interpretation

Statistical Significance: The p-value is an extremely small value, 1.04 e-83. This is far below the chosen significance level, 0.05. Therefore, we reject the null hypothesis. There is no linear relationship between the two variables in p2a. The correlation is statistically significant. The correlation likely did not happen by random chance.

Magnitude of Association: The correlation coefficient, 0.38, indicates a moderate association between the two variables. There is a relationhsip present but it not a strong one. The variation is variable 1 is not wholly explained by the other variable in the dataset.

Direction of Association: The correction is positive. Given that the correlation coefficient has a positive sign, it indicates that as one variable increases, the other increases. The association indicates an upward trend in the sample size.

2.1.3 Visualization

2.2 Part (b): Comparison of p2a.csv and p2b.csv

2.2.1 Test Statistics and P-value for p2b.csv

Pearson Correlation (r_b):

- Correlation coefficient:
- P-value (p_b):
- Selected significance level (α):
- Sample size:

2.2.2 Comparison Analysis

Comparison of Correlations (r_a vs r_b):

Stronger Association Based on Correlations:

Comparison of P-values (p_a vs p_b):

Stronger Association Based on P-values:

Agreement Between Correlations and P-values:

Explanation of Any Discrepancy:

Visual Assessment from Scatter Plots:

Agreement with Statistical Measures:

2.2.3 Visualizations

2.3 Part (c): Comparison of p2a.csv and p2c.csv

2.3.1 Test Statistics and P-value for p2c.csv

Pearson Correlation (r_c):

- Correlation coefficient:
- P-value (p_c):
- Selected significance level (α):
- Sample size:

2.3.2 Comparison Analysis

Comparison of Correlations (r_a vs r_c):

Stronger Association Based on Correlations:

Comparison of P-values (p_a vs p_c):

Stronger Association Based on P-values:

Agreement Between Correlations and P-values:

Explanation of Any Discrepancy:

Visual Assessment from Scatter Plots:

Agreement with Statistical Measures:

2.3.3 Visualizations

3 Code Appendix