CSDS 313/413 — Homework 3

Pairwise Association

Camden Larson

November 11, 2025

Task 1: Associations for Binary Variables

Part (a): Two Variables in `p1a.csv`

Goal: Compute Mutual Information (MI), Jaccard Index (JI), and Pearson's χ^2. Assess statistical significance (permutation tests for MI & JI, analytical χ^2 test for Pearson).

Data and Setup

- Dataset: `Task1/data/p1a.csv`
- Number of samples: 198
- Chosen significance level: $\alpha = 0.05$
- Permutation count for MI & JI: $N = 10,000$

Computed Statistics

Statistic	**Observed**	**p-value**	**Test Type**	**Perms** N	α	**Decision**
Mutual Information (bits)	0.047527	0.00159984	Permutation (greater)	10,000	0.05	**Reject** H_0
Jaccard Index	0.000000	1.00000000	Permutation (greater)	10,000	0.05	Fail to reject H_0
Pearson's χ^2	7.936326	0.00484521	Parametric ($\chi^2_{df=1}$)	—	0.05	**Reject** H_0

Table 1: Association results for `Task1/data/p1a.csv`, columns (X) and (Y), N = 198.

	$Y = 0$	$Y = 1$	**Row sum**
$X = 0$	127	21	148
$X = 1$	50	0	50
Col sum	177	21	**198**

Table 2: 2×2 table for X vs. Y.

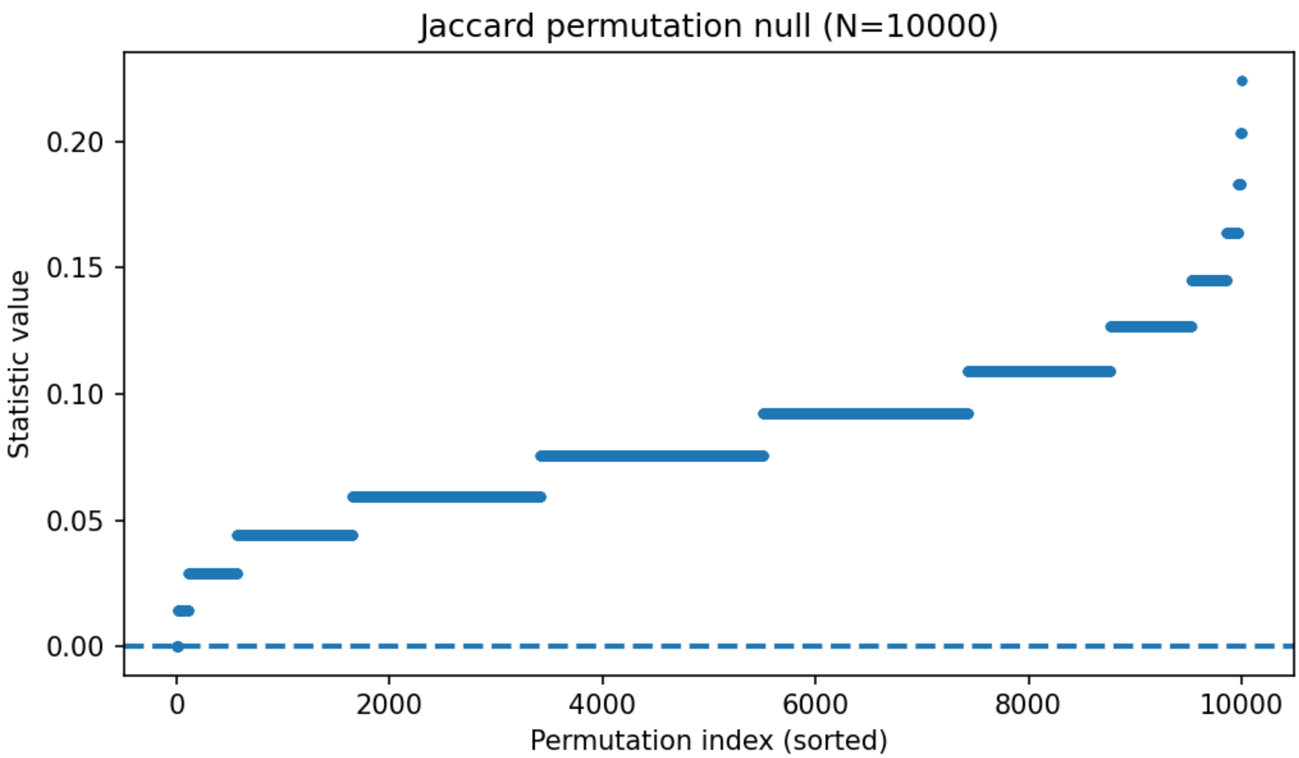
Permutation Tests (for MI & JI)

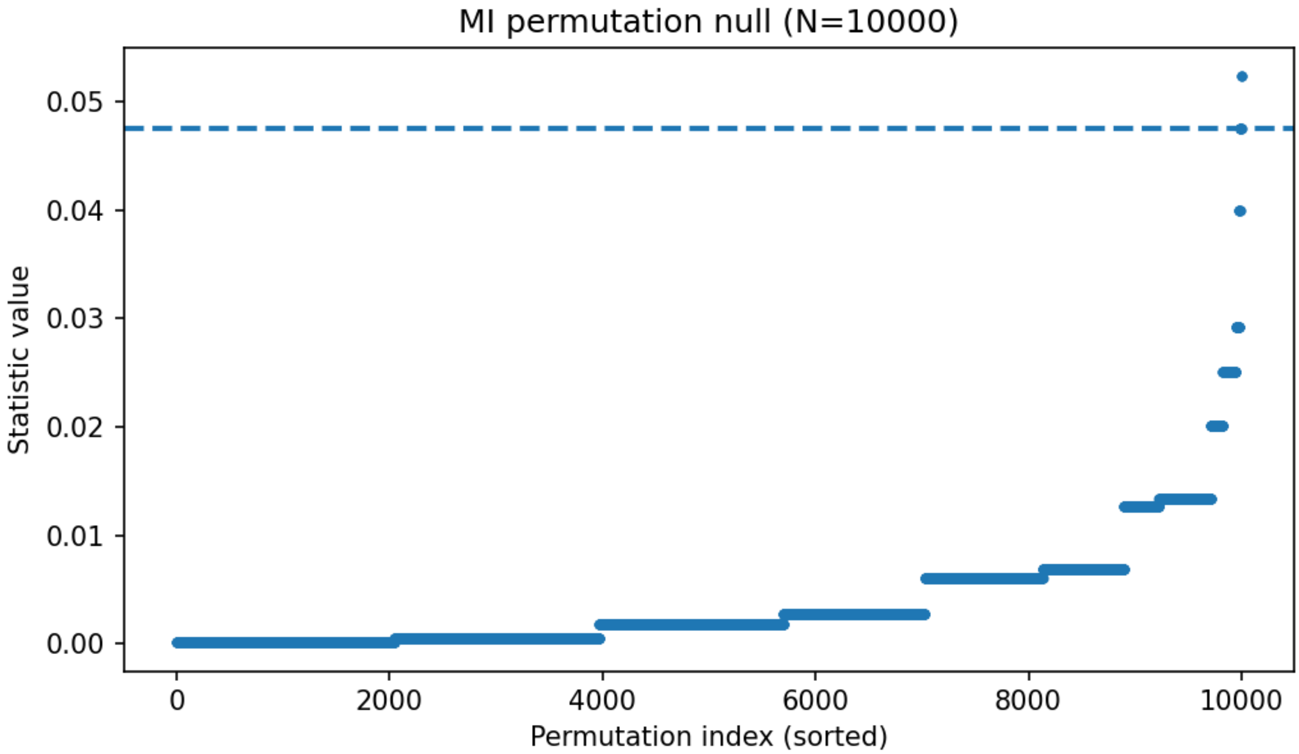
- MI: p-value (permutation) $= \dfrac{c+1}{N+1} = \dfrac{16}{10001} \approx 0.00159984$ (with $c = 15$, $N = 10000$)
- JI: p-value (permutation) $= \dfrac{c+1}{N+1} = \dfrac{10001}{10001} = 1.00000$ (with $c = 10000$, $N = 10000$)

Decision and Interpretation

Answer: Using $N = 198$, $\alpha = 0.05$, and $N_{\text{perm}} = 10{,}000$: we find MI = 0.047527 bits with permutation $p = 0.00159984$ and Pearson's $\chi^2 = 7.936326$ (df = 1), $p = 0.00484521$. Both are below α, so we **reject** H_0 and conclude there is a statistically significant association. The effect size from χ^2 is $\phi = \sqrt{\chi^2/N} \approx 0.200$, indicating a small-to-moderate association. In contrast, the Jaccard index is 0.000000 with permutation $p = 1.0$, so we fail to reject H_0 with Jaccard. This disagreement is expected: the Jaccard Index only captures co-occurrence of 1s and is blind to the observed mutual exclusivity (there were no cases where X = 1 and Y = 1, whereas MI and χ^2 use all table cells and detect negative dependence.

Figure: Permutation Nulls





Summary for Part (b): All 105 Pairs in `p1b.csv`

Setup. Pairs = 105, samples $N = 198$, FDR level $\alpha = 0.05$, permutations for MI & JI $N_{\text{perm}} =$ 50,000.

Method	Significant Pairs (BH-FDR @ α)	Total Pairs	Notes
Mutual Information (MI)	93	105	permutation p-values
Jaccard Index (JI)	57	105	permutation p-values
Pearson's χ^2	93	105	parametric p-values

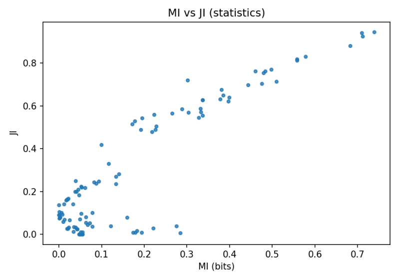
Table 3: Number of significantly associated pairs after Benjamini–Hochberg FDR control at $\alpha =$ 0.05.

Overlap Set	Count
$\text{MI} \cap \text{JI}$	57
$\text{MI} \cap \chi^2$	93
$\text{JI} \cap \chi^2$	57
$\text{MI} \cap \text{JI} \cap \chi^2$	57

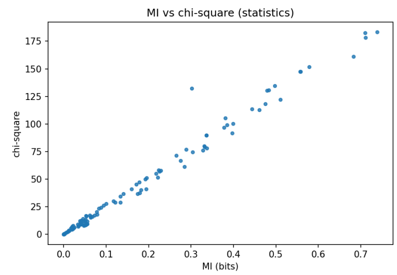
Table 4: Overlaps among the sets of significant pairs (BH-FDR @ $\alpha = 0.05$).

Answer: At FDR $\alpha = 0.05$, MI and χ^2 each flagged 93 out of 105 pairs as significant, while JI flagged 57 as significant. The overlaps show that all JI significant pairs are also significant by

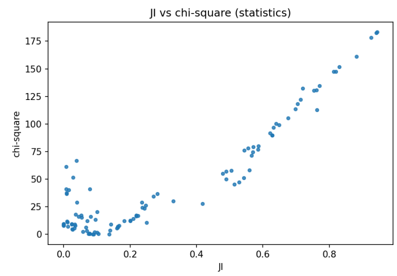
MI and χ^2 (the overlap of all three was 57), and MI and χ^2 agree on 93 pairs. Hence, MI and χ^2 are the most similar. JI is more conservative because it captures only co-occurrence of 1s and is insensitive to negative association/mutual exclusivity like I mentioned above. If restricted to a single metric, MI or χ^2 would preserve nearly the same conclusions for this dataset.



(a) MI vs JI (statistics)

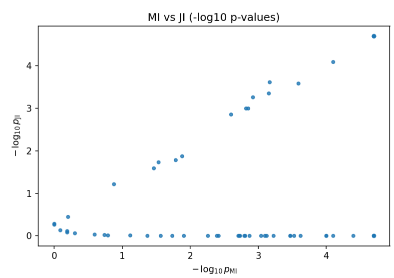


(b) MI vs χ^2 (statistics)

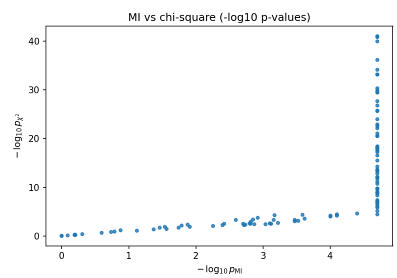


(c) JI vs χ^2 (statistics)

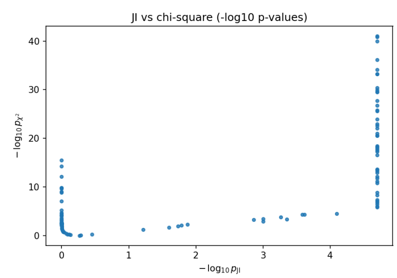
Figure 1: Pairwise comparisons of the test *statistics* across all 105 pairs.



(a) MI vs JI ($-\log_{10} p$)



(b) MI vs χ^2 ($-\log_{10} p$)



(c) JI vs χ^2 ($-\log_{10} p$)

Figure 2: Pairwise comparisons of the *significance* ($-\log_{10} p$) across all 105 pairs.

Task 2: Associations for Continuous Variables

Part (a): `p2a.csv`

Goal: Compute Pearson correlation r_a and two-sided p-value p_a; decide at α.

Data and Setup

- Dataset: `Task2/data/p2a.csv`
- Samples: $N_a = 2400$
- Significance level: $\alpha = 0.05$

Computed Statistics

- Pearson correlation: $r_a = 0.380875$
- Two-sided p-value: $p_a = 1.04095 \times 10^{-83}$
- 95% CI for r_a (Fisher z): $[0.346,\ 0.415]$

Decision and Interpretation

Answer: At $\alpha = 0.05$, we reject H_0 (no linear association) since $p_a \ll \alpha$. The association is positive with moderate magnitude ($r_a \approx 0.381$; 95% CI $[0.346, 0.415]$), indicating that larger values of X tend to be associated with larger values of Y.

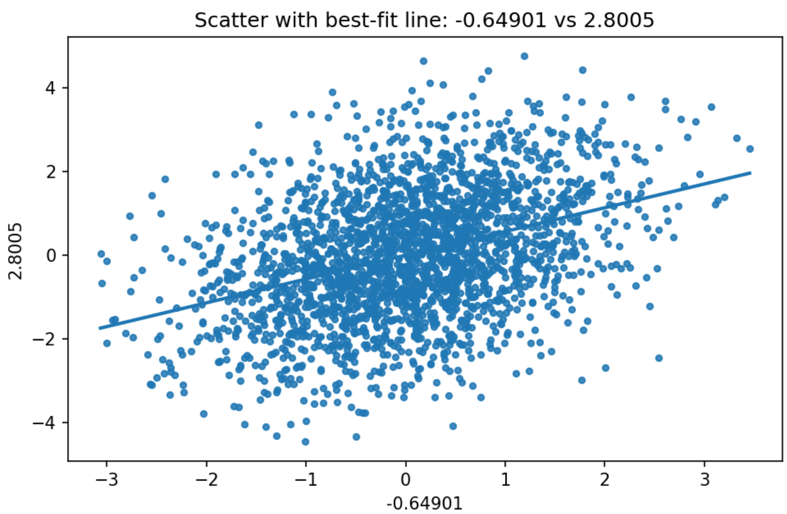


Figure 3: Scatter plot with best-fit line for `p2a.csv`.

Part (b): `p2b.csv`

Goal: Compute Pearson correlation r_b and two-sided p-value p_b; compare with Part (a).

Data and Setup

- Dataset: `Task2/data/p2b.csv`
- Samples: $N_b = 109$
- Significance level: $\alpha = 0.05$

Computed Statistics

- Pearson correlation: $r_b = 0.932898$
- Two-sided p-value: $p_b = 2.87194 \times 10^{-49}$
- 95% CI for r_b (Fisher z): $[0.903,\ 0.954]$

Decision and Interpretation

Answer: At $\alpha = 0.05$, we reject H_0 because $p_b \ll \alpha$. The association is positive with large magnitude ($r_b \approx 0.933$; 95% CI $[0.903, 0.954]$).

Comparison to Part (a)

- **By correlation magnitude:** $|r_a| = 0.380875$ (Part a, $N_a = 2400$) vs. $|r_b| = 0.932898$ (Part b, $N_b = 109$) $\Rightarrow$ Part (b) is stronger by $|r|$.

- **By p-value:** $p_a = 1.04095 \times 10^{-83}$ vs. $p_b = 2.87194 \times 10^{-49}$ $\Rightarrow$ Part (a) is more significant by p (much smaller p), largely due to its much larger sample size.
- **Do $|r|$ and p agree?** No. The discrepancy arises because p-values reflect both effect size and sample size; with very large N, even a moderate r yields an extremely small p.
- **Visual comparison:** The scatter for Part (b) should appear more tightly clustered around a line (consistent with *large* r_b), whereas Part (a) shows a broader cloud (consistent with *moderate* r_a). Thus, the visual impression agrees with the $|r|$ comparison (Part b looks stronger) but not with the p-value ranking (Part a more "significant" due to N).

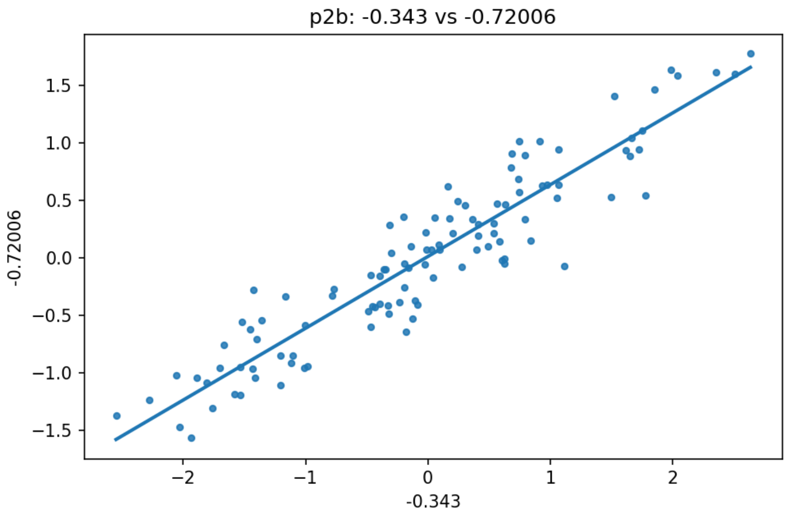


Figure 4: Scatter plot with best-fit line for `p2b.csv`.

Part (c): `p2c.csv`

Goal: Compute Pearson correlation r_c and two-sided p-value p_c; decide at α.

Data and Setup

- Dataset: `Task2/data/p2c.csv`
- Samples: $N_c = 2099$
- Significance level: $\alpha = 0.05$

Computed Statistics

- Pearson correlation: $r_c = 0.041055$
- Two-sided p-value: $p_c = 0.0600291$
- 95% CI for r_c (Fisher z): $[-0.002,\ 0.084]$

Decision and Interpretation

Answer: At $\alpha = 0.05$, we fail to reject H_0 (no linear association) since $p_c \approx 0.060 > \alpha$. The association is positive but of negligible (not statistically significant) magnitude ($r_c \approx 0.041$; 95% CI $[-0.002, 0.084]$), indicating little to no linear relationship in this sample.

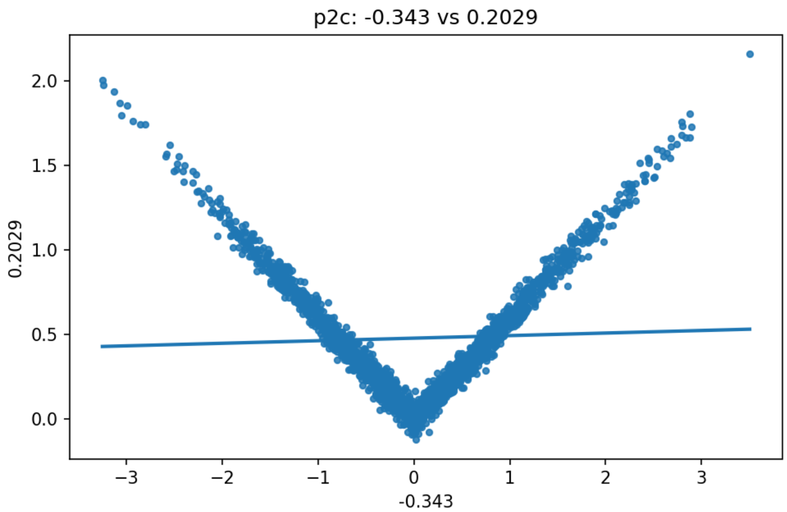


Figure 5: Scatter plot with best-fit line for `p2c.csv`.

Comparison to Part (a)

Answer:

- **By correlation magnitude:** $|r_a| = 0.380875$ (Part a, $N_a = 2400$) vs. $|r_c| = 0.041055$ (Part c, $N_c = 2099$) $\Rightarrow$ Part (a) is much stronger by $|r|$ (moderate vs. negligible).
- **By p-value:** $p_a = 1.04095 \times 10^{-83}$ vs. $p_c = 0.0600291 \Rightarrow$ Part (a) is far more significant; Part (c) is not significant at $\alpha = 0.05$.
- **Do $|r|$ and p agree?** Yes. Both metrics indicate a stronger association in Part (a).
- **Rationale:** Part (a) shows a clear, positive, moderate linear relationship (95% CI [0.346, 0.415]), while Part (c) shows a negligible association with a CI spanning zero ($[-0.002, 0.084]$), consistent with no meaningful linear effect.