Bounding Pearson's Correlation for Ordinal Data: Theory, Symmetry, and Empirical Consequences

Author names omitted for review

Abstract

Pearson's correlation coefficient r is commonly used with ordinal variables, despite being designed for interval-scale data. This paper develops a theoretical framework for bounding r given fixed marginal distributions. Using rearrangement inequalities and connections to Fréchet bounds, we derive conditions for achieving the minimum and maximum correlations, show when symmetry fails ($r_{\min} \neq -r_{\max}$), and prove that the expected value of r under randomization is zero. Simulations illustrate how marginal skew and asymmetry influence the attainable range of r, with implications for factor analysis and structural modeling.

Contents

1 Introduction

- Pearson's r is widely used even when the variables are ordinal.
- Ordinal variables violate the assumptions of interval measurement, yet researchers apply r due to its familiarity and convenience.
- But r is constrained by the marginals of the two variables — its range is not always $[-1, 1]$.
- This paper develops the theoretical and empirical framework to:
  (i) Derive the min and max of r for fixed marginals
  (ii) Relate these to classical probabilistic bounds (Fréchet–Hoeffding)
  (iii) Identify when bounds are symmetric and when they are not
  (iv) Prove that $\mathbb{E}[r] = 0$ under randomization
  (v) Illustrate implications for interpreting correlation-based analyses

2 Setup and Notation

- Define ordinal variables X and Y with K_X, K_Y categories.
- Let $p_i = P(X = x_i), q_j = P(Y = y_j)$; fixed marginals.
- Pearson's correlation:
$$r = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$
- Our goal: Determine the extreme values of r over all joint distributions consistent with $\{p_i\}, \{q_j\}$.

3 Theory: Bounding Pearson's r with Fixed Marginals

3.1 (1) Finding the Extremal Joint Distributions

- Define the optimization problem: maximize/minimize $\mathbb{E}[XY]$ over all couplings.
- Use rearrangement inequalities and greedy coupling to construct $F_{\max}(x, y)$.
- Show examples with 4×5 and 5×5 marginals.

3.2 (1a) Closed-Form Expressions

- Present analytic expressions for the min and max of $\mathbb{E}[XY]$, where derivable.
- Discuss bounds when exact forms are not attainable due to discreteness.

3.3 (2) Relationship to Fréchet–Hoeffding Bounds

- Show that min/max r corresponds to extremal cases of joint distribution under Fréchet bounds.
- Link this to the feasible region defined by Boole–Fréchet inequalities.

3.4 (3) Symmetry, Asymmetry, and the Role of Ties

- Theoretical result: symmetry holds under symmetric marginals.
- When marginals are asymmetric or tied, $r_{\min} \neq -r_{\max}$.
- Explore how entropy of marginals predicts symmetry breaking.

3.5 (4) When Can $r_{\min} > 0$?

- Conditions under which the lower bound on r is strictly positive.
- Example simulations to show how skewed marginals can eliminate negative correlation.

3.6 (5) Center of the Randomization Distribution

- Proof: under uniform random permutation (null of independence), $\mathbb{E}[r] = 0$, regardless of marginals.
- Use this to justify randomization tests even when marginals are highly asymmetric.

4 Empirical Illustrations: Simulating the Randomization Distribution

4.1 Simulation Design

- Fixed-marginal randomization using permutation methods (existing R code).

- Tables: 4×4, 4×5, 3×7, 5×5, 7×10.
- Varying marginal shape: symmetric, skewed, uniform, bimodal.

4.2 Observed Range of r under the Null

- Show full distribution of r under permutation.
- Overlay theoretical min/max — often quite narrow.
- Demonstrate how symmetry or skew influences spread and center.

4.3 When Is r Meaningful?

- Compare observed correlations from real data to attainable range under null.
- Situations where a modest r (e.g., .25) is near-maximal vs. those where it's central.
- Implications for interpreting correlation-based results in practice.

5 Implications for Applied Work

5.1 Consequences for Statistical Modeling

- SEM, factor analysis: attenuated r leads to over-factoring or underestimation.
- Justification for polychoric correlation vs. bounded Pearson's r.

5.2 Recommendations

- Check whether observed r is near the bounds given marginals.
- Report bounds alongside point estimates.
- Use permutation testing or bounds-aware diagnostics for correlation interpretation.

6 Conclusion

- Pearson's correlation has hidden constraints when applied to ordinal data.
- We derive sharp theoretical bounds and show when and why symmetry breaks.

- Our results support more nuanced interpretation of correlation in ordinal contexts.
- A lightweight R package accompanies this paper, allowing users to compute bounds and simulate randomization distributions.

Technical Appendix (Optional)

- Formal derivations of min/max $E[XY]$.
- Symmetry-breaking conditions.
- Greedy algorithm for coupling construction (pseudocode).