MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 1, 3–4

On Zero-Count Correction Strategies in Tetrachoric Correlation Estimation

Jeongwon Choi and Hao Wu

Department of Psychology and Human Development, Vanderbilt University, Nashville, TN, USA

A zero-frequency cell in a 2 × 2 contingency table often results in a tetrachoric correlation estimate close to 1 or −1. Although many ways exist to correct such cells, they have not been thoroughly investigated, with most studies focusing on adding 0.5 to the zeros (e.g. Savalei, 2011; Yang & Weng, 2024). This study explores various correction strategies, including adding values of 0.1, 0.25, and 0.5, and different ways to add, such as adding only to zero cells, keeping margins, and adding to all cells when zero cells exist or regardless of their presence.

First, a simulation with varying sample sizes (50, 100, 200), correlations (0.3, 0.5, 0.7, 0.9), and thresholds (−1.5, −1.0, −0.8, 0.8, 1.0, 1.5) was conducted to evaluate different estimates of a single tetrachoric correlation. We compared using unadjusted versus adjusted thresholds during the second stage of the two-stage procedure. Correlation estimates were evaluated using root mean squared error (RMSE), mean absolute error (MAE) (Figure 1), MAE of Fisher's z-transformed correlations, and the noncoverage rates of the 95% Wald CI. Our results indicate that the choice between adjusted and unadjusted thresholds has minimal impact, and smaller added values are more effective as the correlation increases and the thresholds are farther apart.

Second, we evaluated the correction strategies for estimating a correlation matrix. Data were generated from a 6-variate normal distribution with a sample size of 50 and then discretized using thresholds. Simulation conditions included three sets of population correlations (0.4, 0.8, or a mix) and thresholds of 1.5, 1.0, and 0.8, with positive or mixed signs. Results were evaluated by the number of positive definite correlation matrices and the average weighted mean square error (AWMSE). The results indicate that adding larger values increases the counts of positive-definite matrices, with adding 0.5 and keeping margins being the most effective. The AWMSE decreases with larger added values for positive thresholds, while for mixed-signed thresholds, the best result was achieved with 0.25 at correlations of 0.4, no correction at correlations of 0.8, and 0.5 for mixed correlations.

Finally, we evaluated the correction strategies in estimating a confirmatory factor analysis (CFA) model for binary data, using a one-factor model with four variables. With a sample size of 50, data were generated from a 4-variate normal distribution and discretized using thresholds. Population loadings were 0.4 or 0.7, and thresholds were positive or of mixed signs, with absolute values of 1.5, 1.0, and 0.8. The CFA model was estimated using diagonally weighted least squares. The results indicate that bias and RMSE are largely influenced by the added value. For extreme thresholds, smaller added values minimize bias; otherwise, medium and larger added values produce less biased estimates. In terms of RMSE, for positive thresholds, larger added values are better; for mixed signed thresholds, smaller ones are better.

Overall, our simulations demonstrate that different correction methods perform differently for different combinations of the correlations and thresholds, and no single approach works best in every situation. A better way to resolve this issue is to identify a theoretically motivated (rather than empirical) correction. This will be approached in future research.

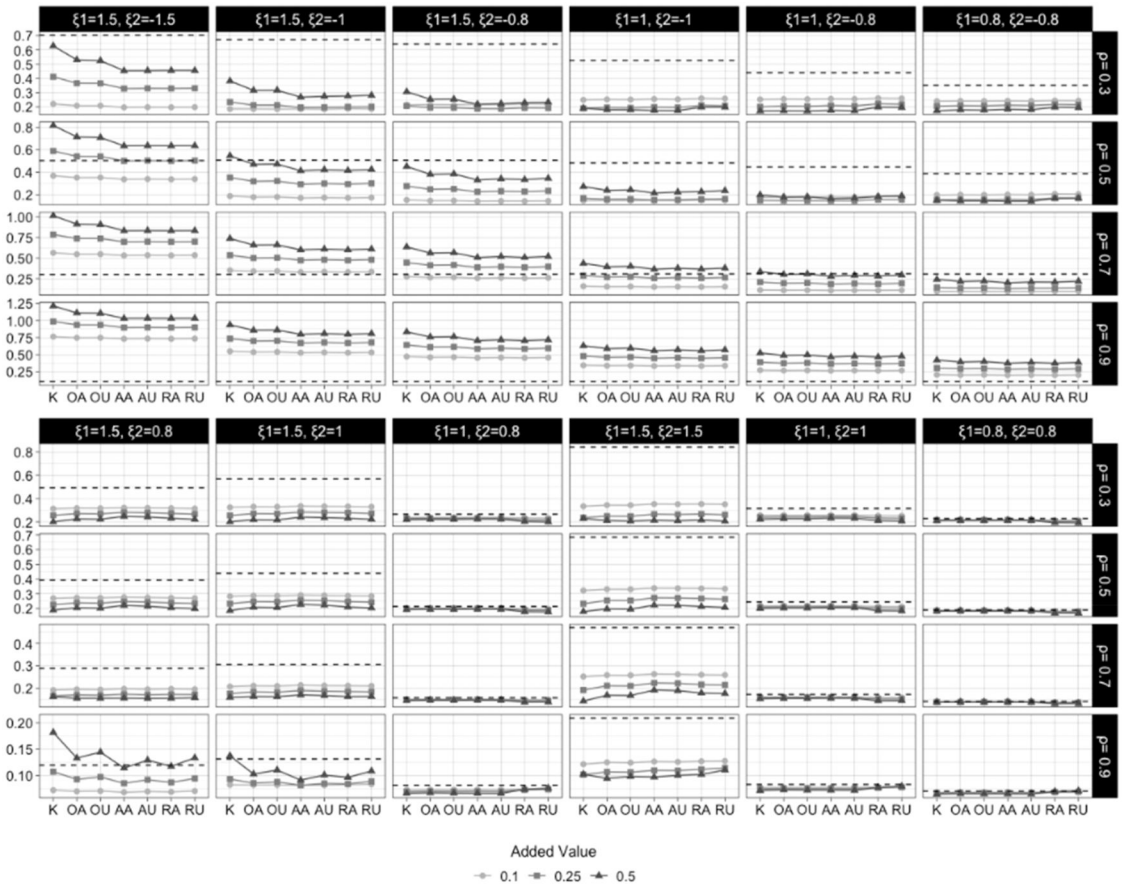

Figure 1. Mean Absolute Error (MAE) for Point Estimates of the Correlation. K: Keep the marginal, for which the thresholds stay the same; OA/OU: Only add to the zero cell and use adjusted/unadjusted thresholds in the second stage of estimation; AA/AU: Add to all cells when zero is present in the table, and use adjusted/unadjusted thresholds in the second stage; RA/RU: Add to all cells regardless of the presence of zero, and use adjusted/unadjusted thresholds in the second stage. The dotted horizontal bar in each panel represents no correction.

References

Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations. *Structural Equation Modeling*, *18*(2), 253–273. https://doi.org/10.1080/10705511.2011.557339

Yang, T.-R., & Weng, L.-J. (2024). Revisiting Savalei's (2011) research on remediating zero-frequency cells in estimating polychoric correlations: A data distribution perspective. *Structural Equation Modeling*, *31*(1), 81–96. https://doi.org/10.1080/10705511.2023.2220919