Research Ireland Centre for Research Training in Foundations of Data Science

Taighde Éireann Research Ireland

The Statistical Power of Alcohol Consumption Measures

Cormac Monaghan[1,3], Dr. Rafael de Andrade Moral[2], Dr. Joanna McHugh Power[3]

[1]Hamilton Institute, Maynooth University
[2]Department of Mathematics and Statistics, Maynooth University
[3]Department of Psychology, Maynooth University

Background

- Understanding alcohol use epidemiology is essential for public health surveillance, policy, and intervention design.
- However, while self-report surveys remain the primary tool for measuring consumption, their reliability depends heavily on methodological precision.
- Using simulated data, we systematically evaluate the statistical power of three widely used alcohol consumption measures to determine which best detects associations in epidemiological research.

Measure	Response
Quantity frequency	Count data
Graduated frequency	Ordinal data
Binary classification	Binary data

Methods

We conducted a simulation study to compare the statistical power of different alcohol measurement approaches:

- **Data Simulation**
  - Generated a large-scale dataset reflecting real-world alcohol consumption distributions.
  - Simulated **7000 random samples** across varying sample sizes $n \in \{100, 250, 500, 1000, 5000, 10000\}$
- **Modelling Approach**
  - Fitted four generalized linear models based on response variable type:
    - **Negative binomial (quantity-frequency)**

$$Y_i \sim \text{NB}(\mu_i, \theta)$$
$$\log(\mu_i) = \boldsymbol{\beta}^\top X_i$$

    - **Zero-inflated negative binomial (quantity-frequency)**

$$Y_i \sim \begin{cases} \pi + (1-\pi) \cdot \text{NB}(0 \mid \mu_i, k) & \text{if } y = 0 \\ (1-\pi) \cdot \text{NB}(y \mid \mu_i, k) & \text{if } y > 0 \end{cases}$$
$$\log(\mu_i) = \boldsymbol{\beta}^\top X_i$$
$$\log\left(\frac{\pi}{1-\pi}\right) = \gamma^\top z_i$$

    - **Ordinal logistic regression (graduated frequency)**

$$\log\left(\frac{P(Y_i \leq \ell)}{P(Y_i > \ell)}\right) = \zeta_\ell - \boldsymbol{\beta}^\top X_i$$

    - **Binary logistic regression (binary classification)**

$$Y_i \sim \text{Bernoulli}(\pi_i)$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\beta}^\top X_i$$

Results

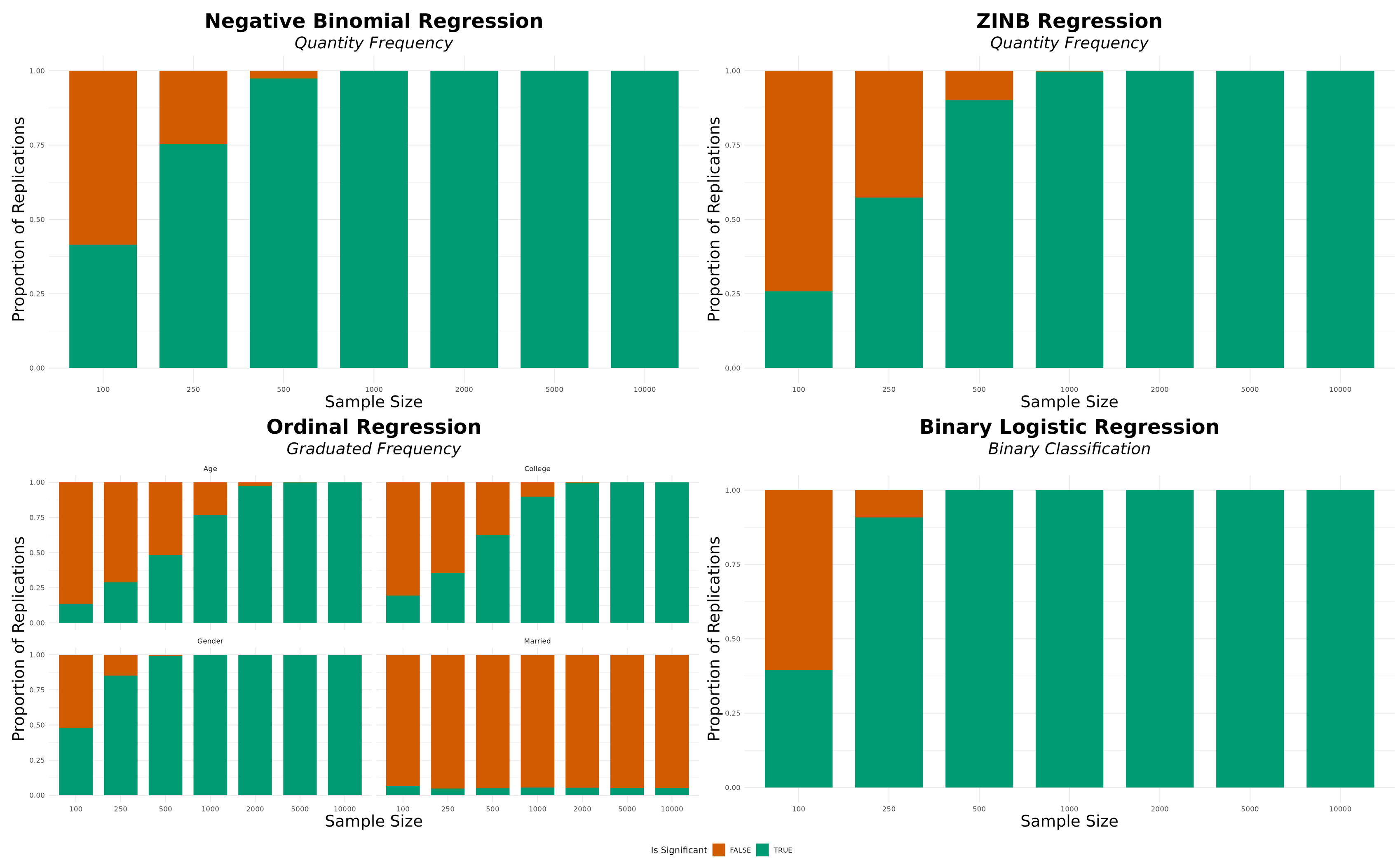

Figure 1: Proportion of Significant Wald Values

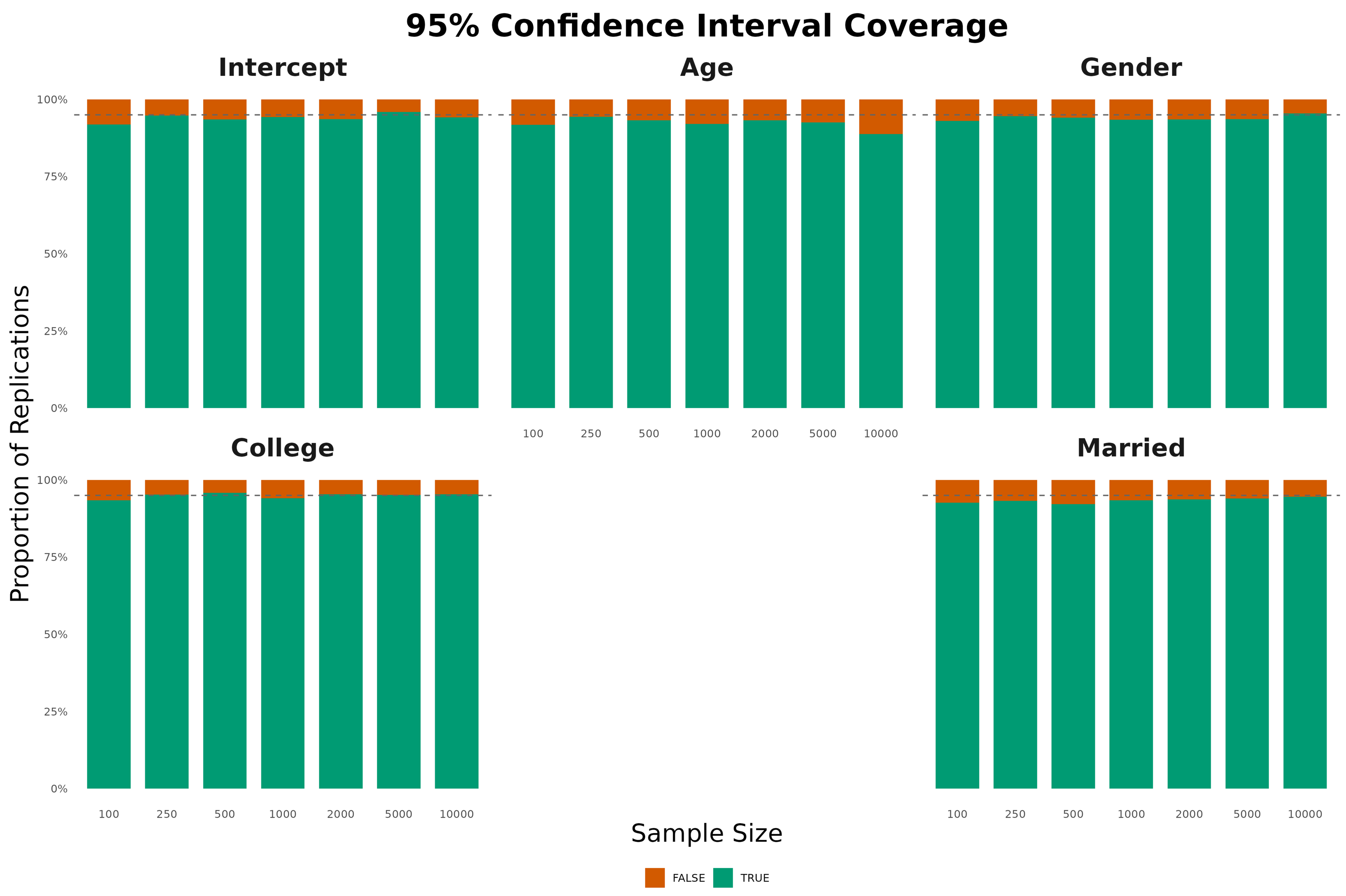

Figure 2: Proportion of 95% CI Coverage

Conclusion

- With small sample sizes, binary data is the most robust approach for measuring alcohol consumption. However, with larger sample sizes, count data performs equivalently.
- Given these findings, we recommend the use of binary classifications when working with small sample sizes due to their robust performance.
- For large sample sizes, count data can offer greater precision by capturing more detailed variations in consumption levels.
- These findings have important implications for public health research and policy, where accurate measurement is critical. Future studies should explore how these methods perform in broader populations, including high-risk groups and cross-cultural settings, to ensure generalizability.

Acknowledgements

This work has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 18/CRT/6049