Optimizing Cluster Trial Designs under Budget Constraints of Normal and Poisson Distributions

Wenyu Yang

Abstract

In this study, we focus on optimizing experimental designs for cluster trials under budgetary constraints, targeting the precise estimation of treatment effects. By simulating outcomes both under normal and Poisson distributions, our analysis examined the trade-offs between the number of clusters and the density of observations within each cluster. The study shows that configurations with a higher number of clusters, when balanced with the number of observations per cluster, enhance the precision of treatment effect estimates. Our results underscore the impact of the cost ratio between initial and subsequent samples within each cluster on achieving optimal designs.

The key findings highlight that while lower cost ratios facilitate data collection within clusters—thereby reducing the mean squared error (MSE)—higher cost ratios helps a distribution of available resources across fewer clusters to minimize compromise on data quality. Additionally, the extension of our study to include outcomes following a Poisson distribution provided deeper insights into the adaptability of our experimental designs under different statistical conditions, further emphasizing the importance of flexible design strategies in the face of varying underlying data distributions. This comprehensive analysis not only aligns with theoretical models but also offers actionable insights for designing efficient studies under real-world constraints, applicable across biomedical and social science fields.

Introduction

In experimental research, the design of efficient studies presents a significant challenge in cluster-randomized trials. Such trials are prevalent in public health, education, and biostatistics, where data naturally clusters within groups such as schools, communities, or patient cohorts. In these trials, each cluster is randomly assigned to either a treatment or a control

group, with observations recorded from multiple units within each cluster. The design complexities of these trials stem from the need to manage budget limitations and the correlation among observations within clusters, which can significantly affect the statistical power and cost-effectiveness of the study.

This study addresses these challenges by investigating optimal design strategies for cluster-randomized trials with the goal of minimizing the mean squared error (MSE) of treatment effect estimates while adhering to a predefined budget. The cost structure within cluster-randomized trials often differentiates between the expense of the first sample in a cluster (c1) and that of subsequent samples (c2), where c2 $<$ c1. Adding more clusters tends to reduce inter-cluster variability and potentially enhances the study's power, but it also leads to higher overall costs. Conversely, increasing the number of observations within a cluster is more cost-effective but may lead to diminishing returns due to increased correlation among observations. The goal of this research is to explore the balance between the number of clusters (n_clusters) and the number of observations per cluster (n_obs_per_cluster), aiming to identify configurations that optimize the trade-offs between cost and the precision of treatment effect estimates. Utilizing simulation studies, this project examines the impact of various configurations on MSE, taking into account both normally and Poisson-distributed outcome variables. The study employs a hierarchical modeling approach to account for cluster-level variability, simulating realistic scenarios where within-cluster measurements are influenced by shared environmental or biological factors.

Our study will focus on solving the three aims below:

AIM 1: Design a simulation study using the ADEMP framework to evaluate potential study designs.
This aspect of the study employs the ADEMP framework to methodically evaluate potential study designs. By simulating data across different combinations of cluster and observation numbers, this aim seeks to identify configurations that yield the lowest MSE, ensuring that the study design is systematic, reproducible, and reflective of real-world constraints.

AIM 2: Explore relationships between the underlying data generation parameters and the relative costs (c1/c2) and their impact on the optimal study design.
This aim focuses on the implications of varying cost ratios (c1/c2) on the study design. It assesses how different cost structures influence the balance between the number of clusters and the density of observations within clusters. This analysis is critical for understanding how budget allocations can be optimized to enhance the efficiency and effectiveness of cluster-randomized trials.

AIM 3: Extend the simulation study to the setting in which Y follows a Poisson distribution and explore how this impacts the results.
Extending the simulation study to scenarios where the outcome variable Y follows a Poisson distribution, this aim investigates how distributional assumptions about the data affect the optimal study designs and the precision of treatment effect estimates. Adapting the hierarchical model to reflect the log-linear relationship between the treatment effect and the mean

outcome, this extension evaluates the suitability of different configurations under the unique characteristics of count data.

The hierarchical model employed in this study captures the complex structure of clustered data, where outcomes are influenced by both fixed effects. In the normal model, cluster-level variability is modeled using a Gaussian distribution, while in the Poisson model, the log-mean is modeled with a Gaussian random effect. These models ensure that the simulation study reflects realistic data-generating mechanisms encountered in practical applications. By addressing these aims, this study evaluates the optimal experimental designs for cluster-randomized trials under budget constraints. The findings are expected to inform researchers about efficient allocation of resources in the design of trials, enabling them to achieve robust and precise estimates of treatment effects. The results will be particularly valuable for fields where data collection is costly or time-intensive, such as biostatistics, social sciences, and epidemiology. This work not only contributes to the theoretical understanding of cluster-randomized trials but also offers practical recommendations for study design in resource-limited settings.

Methods

This study uses the ADEMP framework to rigorously evaluate experimental designs for cluster-randomized trials under stringent budget constraints. We focus primarily on optimizing designs that minimize the mean squared error (MSE) of treatment effect estimates while adhering to limited resources. By simulating both normally distributed and Poisson-distributed outcomes, we explore how various configurations affect the precision and accuracy of treatment effect estimates.

Our primary goal is to identify the optimal number of clusters n_clusters and observations per cluster that minimize MSE, given a fixed budget B. This exploration involves balancing the cost implications of initiating a cluster (c1) and adding subsequent samples within the same cluster (c2), where c2<c1, against the potential gains in reducing inter-cluster variability and enhancing statistical power.

The budget constraint can be expressed as: B = n_clusters c1 + n_obs c2 n_clusters

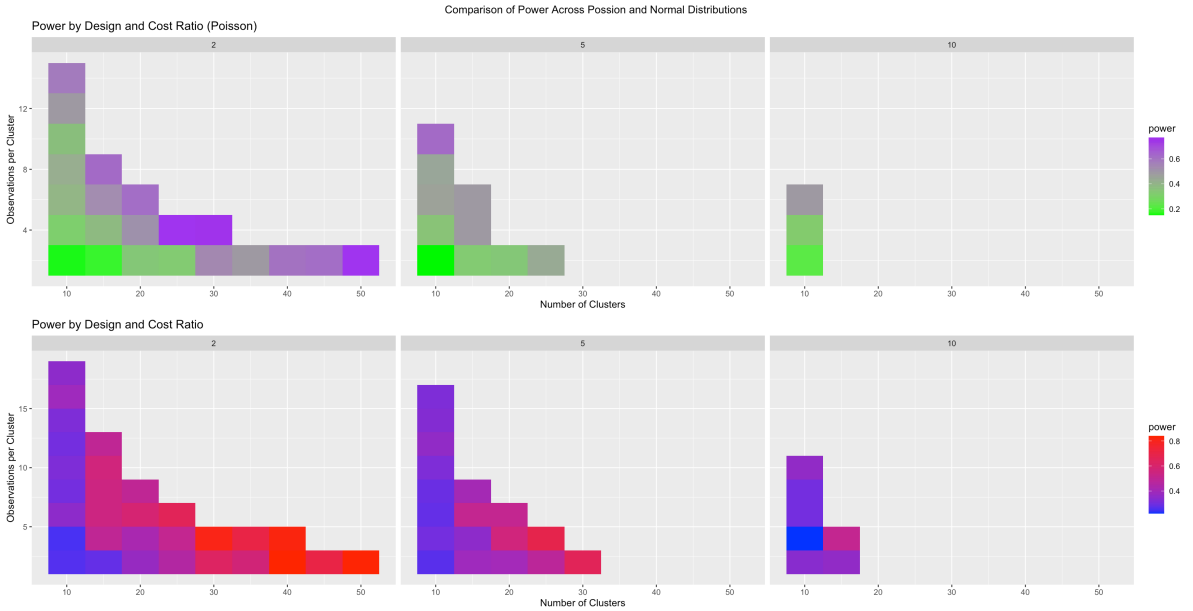
For normally distributed outcomes, our hierarchical data-generating mechanism posits that each cluster's mean outcome is influenced by both a fixed component, representing treatment assignment (i = + Xi + i), where is the intercept, represents the effect of the predictor (Xi), and i N(0, 2) is the random cluster-level effect. Individual observations within a cluster (Yij) are modeled as normally distributed around the cluster mean (i), with additional individual-level variance (2): Yij N(i, 2). This structure reflects the inherent correlation among observations within clusters, a typical characteristic in cluster-randomized trials. In scenarios where Y follows a Poisson distribution, the data-generating mechanism adapts to model the expected mean outcome for each cluster (i) using a log-linear relationship (log(i)

= + Xi), with observations within the cluster modeled as following a Poisson distribution based on the expected mean: Yij Poisson(i). This setup is tailored for count data, where outcomes exhibit heteroscedasticity, ensuring that within-cluster observations remain correlated while maintaining independence across clusters.

The treatment effect, defined as the average difference in outcomes between the treatment and control groups, is quantified by comparing the mean outcomes across these groups. The efficacy of different design configurations is assessed through simulations, constrained by the budget B, to reflect practical limits on resource availability. The cost of each experimental design is calculated by considering both c1 (cost of initiating a cluster) and c2 (cost of adding observations within a cluster), excluding designs that exceed the budget. For each feasible design, data are simulated under both normal and Poisson models, and the treatment effect is estimated. The MSE is calculated to determine the precision and accuracy of these estimates, identifying configurations that produce the lowest MSE as optimal. Furthermore, the study examines how variations in the cost ratio (c1/c2) affect design choices, analyzing how resource allocation preferences might shift the balance between the number of clusters and observations per cluster. This analysis provides crucial insights into how cost structures influence experimental design decisions in cluster-randomized trials.

Ultimately, the performance of each design is measured using three metrics: bias, variance, and MSE. Bias assesses the systematic differences between estimated and true treatment effects, variance captures the variability of these estimates across simulations, and MSE, combining both bias and variance, serves as the primary criterion for design evaluation. Designs with the lowest MSE are deemed optimal, striking the best balance between accuracy and precision. This methodical approach ensures that the study's findings are robust and widely applicable across various experimental settings.

Result

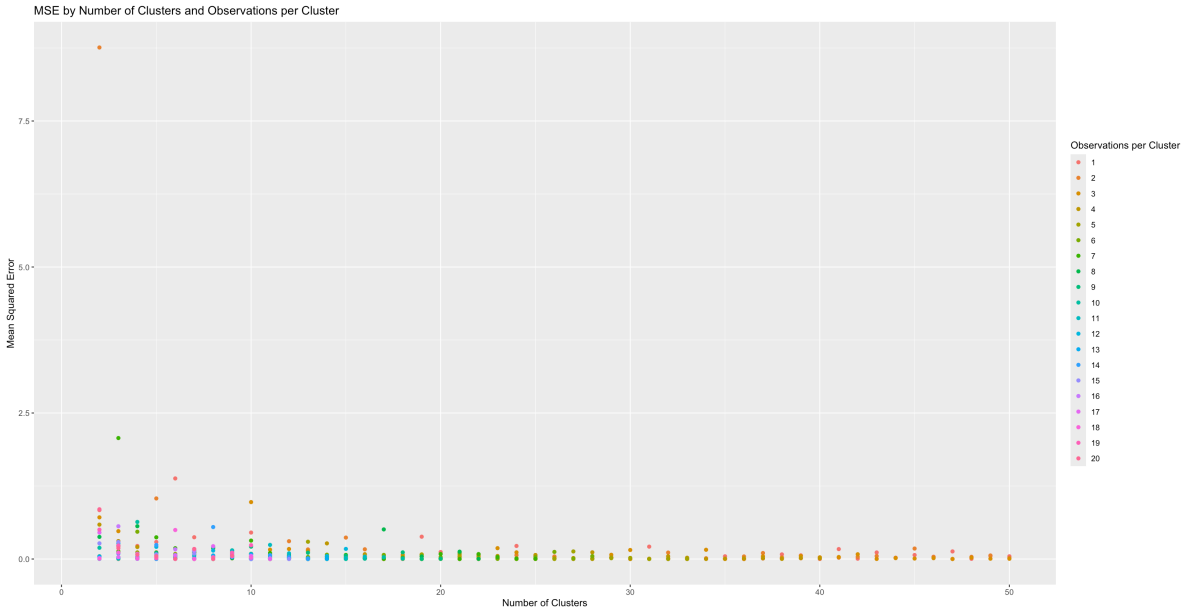


For AIM2, Our heatmap provides clear insights into the relationship between the underlying data generation mechanisms, relative costs (c1/c2), and their impact on optimal study design. For the Poisson distribution, power is highly sensitive to the allocation of clusters and observations under varying cost ratios. At a low cost ratio of 2, designs with a larger number of clusters achieve higher power because the reduced marginal cost of additional observations (c2) allows for more flexibility in spreading resources across clusters. This configuration helps Poisson outcomes, which benefit from increased between-cluster variability due to their variance-mean dependence. As the cost ratio increases to 5, the optimal design shifts to fewer clusters, and power is constrained to a narrower range. This reflects the growing expense of spreading resources widely, limiting the ability to increase cluster numbers and forcing a balance between cluster count and within-cluster density. At a high cost ratio of 10, power becomes concentrated in designs with very few clusters (around 10) and minimal within-cluster observations. This dramatic reduction in flexibility highlights the disproportionate impact of increasing c1/c2 on Poisson outcomes, where high within-cluster correlation and the variance-mean relationship make optimal designs particularly sensitive to cost constraints.

For the Normal distribution, the relationship between cost ratios and study design is comparatively more stable due to the homoscedastic nature of the data generation mechanism. At a cost ratio of 2, power gradients show a preference for designs with a larger number of clusters (10 to 25), indicating that spreading resources across more clusters remains effective in achieving high power. However, the Normal model exhibits greater robustness to reductions in cluster count compared to the Poisson model, as its variance structure is less dependent on the balance between clusters and within-cluster observations. At a cost ratio of 5, the design shifts slightly toward fewer clusters (10 to 20) with more observations per cluster, as the increasing

cost ratio limits the feasibility of maximizing cluster count. Even so, power remains relatively high across a broader range of designs compared to the Poisson model. At a cost ratio of 10, the Normal model demonstrates resilience in maintaining power, with optimal designs concentrated in fewer clusters (around 10) and denser within-cluster observations. Unlike the Poisson model, the Normal distribution's lack of variance-mean dependence allows it to sustain power under more constrained budgets, making it less sensitive to increasing c1/c2 ratios.

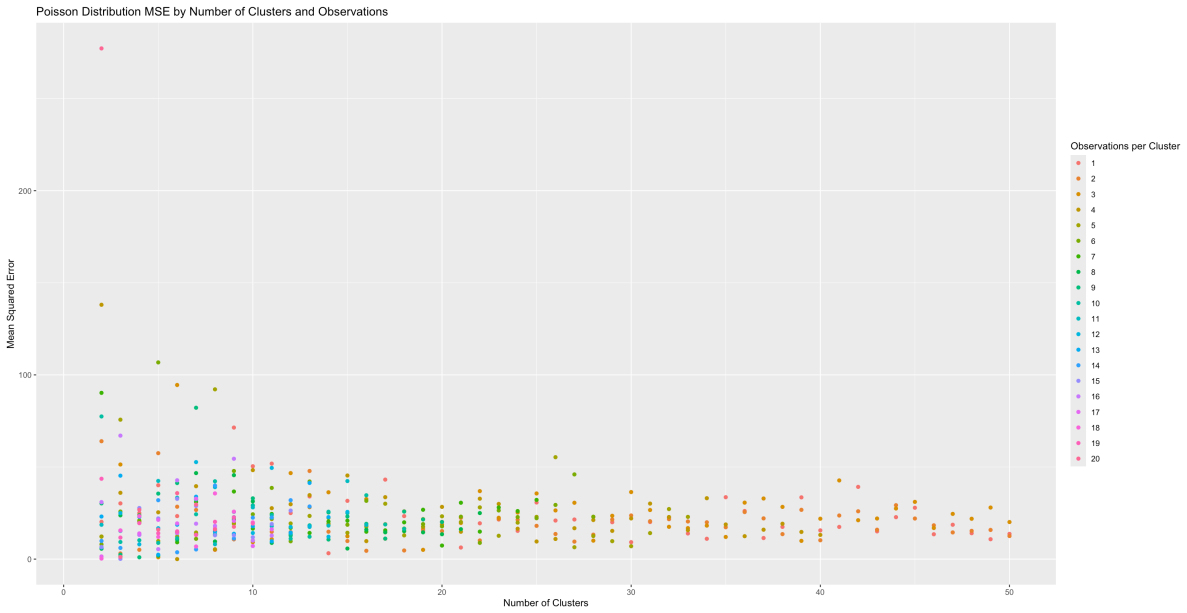
Overall, the underlying data generation mechanisms significantly influence the impact of cost ratios on optimal study design. For Poisson outcomes, the variance-mean relationship amplifies sensitivity to cost constraints, necessitating careful balance between cluster count and within-cluster density. High cost ratios disproportionately reduce power in Poisson models, emphasizing the importance of spreading resources across clusters when costs allow. In contrast, the Normal model's homoscedastic variance structure supports a broader range of feasible designs, enabling higher power even with fewer clusters and more observations per cluster under constrained budgets. These findings underscore the critical role of understanding data generation mechanisms and cost structures when designing cluster randomized trials, as they directly inform resource allocation strategies for maximizing power within budgetary constraints.



For Aim1, The comprehensive analysis of Mean Squared Error (MSE) relative to the number of clusters and observations per cluster provides a nuanced understanding of its impact on the precision of treatment effect estimates in cluster-randomized trials. The data presented in your chart clearly indicate that MSE generally declines as the number of clusters increases, particularly when the clusters include a moderate number of observations. This trend highlights a critical aspect of experimental design: leveraging a higher number of clusters to significantly enhance the precision of experimental outcomes, thereby providing a robust foundation for credible scientific conclusions. Delving deeper into the data, it is evident that configurations

with a greater number of clusters—especially those exceeding 40 clusters—achieve particularly low MSE values, irrespective of the number of observations per cluster. This suggests that the benefit of adding more observations within a cluster plateaus beyond a certain point, thus supporting a strategy focused on increasing the number of clusters rather than merely augmenting the size of each cluster. The strategic implication here is profound, as it suggests that resources could be better allocated towards expanding the number of clusters rather than increasing cluster size, which may yield diminishing returns in terms of reducing MSE.

Analyzing the configuration details, the lowest MSE is observed across 50 clusters. This configuration, regardless of the number of observations per cluster, consistently shows lower MSE values, thereby underscoring the advantage of a broad cluster distribution. This finding demonstrates how effectively diluting within-cluster correlation, which can skew results if one or few clusters have atypical responses, enhances the reliability of treatment effect estimates. The broader distribution of clusters helps mitigate the influence of any single cluster's peculiar characteristics. This is particularly important in settings where the clusters might have inherent variability that could influence the treatment effect. By increasing the number of clusters, the influence of outliers is minimized, and the treatment effect can be estimated with greater accuracy and less bias. This also increases the robustness and generalizability of the results, as the estimates are less likely to be influenced by the idiosyncrasies of a smaller number of clusters.



In the third aim of our simulation study, we explore the implications of using a Poisson distribution for the outcome variable Y, which is notably different from a normal distribution due to its variance being inherently tied to the mean. This relationship provides a distinct context for evaluating treatment effect estimation, highlighting how variance characteristics evolve with the response's intensity.The Mean Squared Error (MSE) graph for the Poisson

distribution, detailing the impact of the number of clusters and observations per cluster, reveals several critical insights. Notably, the variation in MSE across different configurations is more pronounced than what one might expect in settings following a normal distribution. This increased variability is attributable to the Poisson distribution's sensitivity to changes in the mean, where even small fluctuations can significantly affect the variance.

A striking observation from the data is that configurations with a larger number of observations per cluster generally exhibit lower MSEs. This suggests that enhancing granularity within clusters can profoundly improve the precision of treatment effect estimates under the Poisson model. The chart shows that the MSE values tend to decline as the number of clusters increases, especially when each cluster consists of more than five observations. This trend is vital as it indicates that distributing observations across more clusters can mitigate the impact of over-dispersion—a characteristic typical of Poisson-distributed data—which is essential for reducing estimation errors in treatment effects.Configurations that include around ten clusters, with each cluster having between six to twelve observations, demonstrate markedly lower MSE compared to configurations with either fewer clusters or fewer observations per cluster. This pattern accentuates the importance of having a sufficient sample size within each cluster to address the unique challenges posed by the Poisson model. In contrast to normal distribution models where variance remains constant, the variance in a Poisson model scales with the mean. It becomes crucial, therefore, to ensure that there are enough observations within clusters to stabilize the variance estimates.Furthermore, an increase in MSE with fewer observations per cluster likely reflects a greater relative impact of random variation, which can disproportionately influence the mean response's accuracy within each cluster. Optimal configurations thus require not only a strategic choice of the number of clusters but also a careful balance in the number of observations within each cluster. This balance is imperative to effectively manage the inherent variability of the Poisson distribution, ensuring the design is robust enough to handle over-dispersion and accurately capture variance related to the treatment effects.

Above all, we found a nuanced view of how distribution assumptions and the number of observations per cluster influence statistical power. We observed that in scenarios with higher cost ratios, there was a discernible drop in power. This indicates the negative impact of higher initial costs per cluster on the experimental power, highlighting the challenge of balancing cost and efficiency in experimental design, particularly under budget constraints. Furthermore, our comparison of power across Poisson and normal distributions illuminated the sensitivity of experimental outcomes to the assumed underlying distribution. The normal distribution typically assumes homogeneity in variance across observations, in contrast to the Poisson distribution where variance increases with the mean. This variance structure leads to different power outcomes, where Poisson-distributed data may require more careful planning regarding the number of clusters and observations per cluster to maintain adequate statistical power.

Regarding Mean Squared Error (MSE) trends and implications, our examination of the bottom charts further delineated the relationship between MSE, the number of clusters, and the observations per cluster. We found pronounced variability in MSE under the Poisson distribution, as opposed to the normal distribution, reflecting the impact of the distribution's inherent

properties on error rates. Notably, configurations with a greater number of observations per cluster generally exhibited lower MSE. This underscores the effectiveness of increasing observation granularity within clusters in reducing error rates and enhancing the precision of treatment effect estimates under the Poisson model. This aspect of our findings is crucial for researchers aiming to design experiments that minimize error while maximizing the reliability of their results in studies where the Poisson distribution is appropriate.

Conclusion

In our simulation study focused on cluster-randomized trials, we systematically explored the impacts of different configurations and cost structures across several scenarios, guided by the ADEMP framework. Our analysis demonstrated that increasing the number of clusters generally leads to reduced Mean Squared Error (MSE), particularly when these clusters contain a moderate number of observations. This configuration maximized the precision of treatment effect estimates by effectively minimizing errors due to intra-cluster correlation. We observed the lowest MSE values in configurations featuring around 50 clusters, suggesting that a larger number of smaller clusters is most advantageous for obtaining robust and credible scientific results. In examining the relationship between data generation parameters and relative costs (c1/c2), we found that cost structures significantly impact the optimal design of cluster-randomized trials. At lower cost ratios, where the cost of adding additional observations (c2) is much less than the initial cluster cost (c1), designs incorporating more clusters proved beneficial. This arrangement allowed for greater flexibility in distributing resources across clusters, enhancing the trial's power. However, as the cost ratio increased, the feasibility of using numerous clusters diminished, necessitating a design shift toward fewer clusters with potentially more observations per cluster. This strategic balance is crucial for managing rising costs while maintaining statistical power and the precision of treatment effect estimations.

Further extending our study to include outcomes where the variable Y follows a Poisson distribution, we identified unique impacts due to the Poisson model's variance-mean dependency. Configurations with a greater number of observations per cluster consistently showed lower MSEs, highlighting the importance of sufficient sample sizes within clusters to effectively accommodate the inherent variability of the Poisson distribution. Additionally, the decline in MSE with an increased number of clusters was more pronounced, emphasizing the need to distribute observations across a larger number of clusters to mitigate the effects of over-dispersion typical in Poisson-distributed data.These findings collectively enhance our understanding of how both the statistical properties of data distributions and economic considerations shape the design of cluster-randomized trials. Our study provides essential guidance for optimizing experimental designs, ensuring that trials are both cost-effective and statistically powerful, and tailored to the specific distributional characteristics of the outcome data. This approach not only addresses the logistical and financial constraints of trial design but also ensures the integrity and accuracy of treatment effect estimations.

Reference

1. For the "Stat Med" article from 2019: Morris, Tim P., Ian R. White, and Michael J. Crowther. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine*, vol. 38, no. 11, 16 Jan. 2019, pp. 2074–2102. Wiley Open Access Collection, doi:10.1002/sim.8086. PMCID: PMC6492164, PMID: 30652356.
2. For the "Methods in Ecology and Evolution" article from 2024: Williams, Coralie, et al. "Transparent Reporting Items for Simulation Studies Evaluating Statistical Methods: Foundations for Reproducibility and Reliability." *Methods in Ecology and Evolution*, First published 30 Sept. 2024, https://doi.org/10.1111/2041-210X.14415. Handling Editor: Aaron Ellison.
3. For the "Computers & Industrial Engineering" article from December 2024: Badakhshan, Ehsan, Navonil Mustafee, and Ramin Bahadori. "Application of Simulation and Machine Learning in Supply Chain Management: A Synthesis of the Literature Using the Sim-ML Literature Classification Framework." *Computers & Industrial Engineering*, vol. 198, Dec. 2024, Article 110649. Elsevier, doi:10.1016/j.cie.2024.110649.
4. For the "Research Policy" article from October 2024: Belik, Ivan, Prasanta Bhattacharya, and Eirik Sjåholm Knudsen. "A Case for Simulated Data and Simulation-Based Models in Organizational Network Research." *Research Policy*, vol. 53, no. 8, Oct. 2024, Article 105058. Elsevier, doi:10.1016/j.respol.2024.105058.
5. For the article in "Psychological Methods" from 2024: Lang, Jonas W. B., and Paul D. Bliese. "The Plausibility of Alternative Data-Generating Mechanisms: Comment on and Attempt at Replication of Dishop (2022)." *Psychological Methods*, 2024. Advance online publication. doi:10.1037/met0000650.
6. For the "BMC Genomics" article from April 2020: Assefa, Alemu Takele, Jo Vandesompele, and Olivier Thas. "On the Utility of RNA Sample Pooling to Optimize Cost and Statistical Power in RNA Sequencing Experiments." *BMC Genomics*, vol. 21, 19 Apr. 2020, Article 312, doi:10.1186/s12864-020-6721-y, PMCID: PMC7168886, PMID: 32306892.
7. For the "Econometrics Journal" article from January 2020: Carneiro, Pedro, Sokbae Lee, and Daniel Wilhelm. "Optimal Data Collection for Randomized Control Trials." *The Econometrics Journal*, vol. 23, no. 1, Jan. 2020, pp. 1–31, doi:10.1093/ectj/utz020.
8. For the "Journal of Thoracic Disease" letter from March 2015: Viti, Andrea, Alberto Terzi, and Luca Bertolaccini. "A Practical Overview on Probability Distributions." *Journal of Thoracic Disease*, vol. 7, no. 3, Mar. 2015, pp. E7-E10, doi:10.3978/j.issn.2072-1439.2015.01.37, PMCID: PMC4387424, PMID: 25922757.
9. "Project 3: Simulation Studies." PHP 2550: [**Practical Data Analysis**], Brown University, Fall 2024. Canvas

GITHUB REPO: https://github.com/nice-jessica/project3-2550