Optimal Experimental Design under Budget Constraints: Applications to Normal and Poisson Distributed Data

Tianna Chan

2024-12-05

Abstract

This project investigated optimal experimental design under budget constraints and varying parameters through a simulation study. Using a hierarchical model, we evaluate cluster randomized trials to estimate the treatment effect on Normal and Poisson simulated data. These methods are particularly relevant for applications like DNA sequencing, where clusters represent groups of samples processed together within a single sequencing run. Shared correlations within clusters, coupled with fixed and variable costs per sample, introduce unique statistical and budgetary challenges.

Our findings showed that more stable and accurate treatment effect estimates are consistently produced by designs with larger numbers of clusters and moderate within-cluster observations, compared to the smaller ones. Between-cluster and within-cluster variances significantly affect the estimates, especially in smaller clusters. Results remain largely consistent when extending to a Poisson distribution. Cost ratio analysis demonstrates slight differences between normal and poisson distributed data. For Normal data, lower or medium initial costs has a more precise result with lower mean squared error (MSE) and higher coverage or lower bias respectively. For Poisson data, lower initial costs consistently resulted in the lowest MSE, absolute bias, and higher coverage. These findings provide valuable insights into the trade-offs between cost, design complexity, and statistical performance, guiding researchers to optimize experimental designs under constrained resources.

Introduction

Cluster-randomized trials are essential for evaluating interventions when individual-level randomization is impractical or unethical. There are also applications in DNA sequencing, where clusters of DNA represents groups of samples processed together in a single sequencing run when they have shared correlations within clusters.

However, optimizing the design of such trials is challenging, especially under budget constraints. The cost per sample in DNA sequencing often include fixed and variable costs, which is parallel to our study design. Inefficient allocation of resources can lead to suboptimal designs that compromise the statistical power or precision of the estimated treatment effects.

In this project, we aim to design a simulation study to look at optimal experimental design under varying parameters, budget constraints, and cost structures. By leveraging the ADEMP framework (Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures), we simulate cluster-randomized trials under two common data distributions: Normal and Poisson. Our focus is on understanding how design elements—such as the number of clusters and observations per cluster—impact the estimation of the treatment effect (β) under different cost structures and parameters.

Through this work, we aim to provide practical guidance for researchers to optimize experimental designs, ensuring resource-efficient and statistically robust methodologies, especially in high-throughput fields like DNA sequencing.

Methods

We will design a simulation study using the ADEMP framework as follows.

Aims: To estimate the treatment effect β and determine the optimal number of clusters n and observations per cluster R under a budget and cost constraints.

Data-generating mechanisms:

(See R script `data_simulation.R` for data generation code)

X_i is the binary treatment variable that indicates whether a cluster is assigned to the treatment group $X_i = 1$ or control group $X_i = 0$. The hierarchical model for the normal distribution setting is given below.

- $\mu_{i0} = \alpha + \beta X_i$ (fixed effect, $\mu_{i0} = \alpha$ or $\mu_{i0} = \alpha + \beta$)
- $\mu_i \mid \varepsilon_i = \mu_{i0} + \varepsilon_i$ with $\varepsilon_i \sim N(0, \gamma^2)$, or in other words $\mu_i \sim N(\mu_{i0}, \gamma^2)$
- $Y_{ij} \mid \mu_i = \mu_i + e_{ij}$ with $e_{ij} \sim$ iid $N(0, \sigma^2)$, or in other words, $Y_{ij} \mid \mu_i \sim N(\mu_i, \sigma^2)$

For the hierarchical model for the poisson distribution setting,

- $log(\mu_i) \sim N(\alpha + \beta X_i, \gamma^2)$.
- $Y_{ij} \mid \mu_i \sim \text{Poisson}(\mu_i)$. Since the sum of iid Poisson random variables is still Poisson, we may simply have $Y_i := \sum_{j=1}^{R} Y_{i,j}$ and $Y_i \mid \mu_i \sim \text{Poisson}(R\mu_i)$ to reduce computational complexity.

Estimands: The estimand is β, the fixed treatment effect, representing the average difference between treatment and control groups.

Methods: The simulation study consists of three parts:

Part 1 (Varying Parameters - Normal Distribution): We will start with setting a total budget $B = \$2000$ in dollars, with sampling costs $c_1 = 20$ for the first observation in each cluster and $c_2 = 10$ for each additional observations in the same cluster, such that $c_2 < c_1$. The number of clusters and observations per cluster were varied from 5 to 15 in increments of 5 and 10 to 50 in increments of 2 respectively. These ranges represent scenarios with fewer, more resource-intensive clusters versus more clusters with potentially smaller sample sizes per clusters. With these constraints, we will find out the feasible number of clusters n and observations per cluster R designs that $Cost = n \cdot [c_1 + (R-1) \cdot c_2]$, and for each cluster size, we select the one with the maximum allowable cost for each cluster size.

We will then pair these designs with all combinations of design and parameter settings. The intercept α is fixed at 5 since changing it would only shift the result but not affecting the performance. The treatment effect (β) was set to represent small to large effects (1, 2, and 3), reflecting scenarios with varying levels of detectable differences. Between and within cluster variances (γ^2 and σ^2) were set to low and high levels (1 and 10) to capture conditions of low and high heterogeneity across and within clusters. Considering all possible cases with varying parameters, we repeat the simulations for 500 iterations to estimate the performance of each design using a linear mixed-effects model.

Part 2 (Varying Cost Ratio - Normal Distribution): We will explore relationships between the underlying data generation mechanism parameters and the relative costs c_1/c_2 and how these impact the optimal study design. We repeat the analysis by fixing the parameters and vary the relative costs. The treatment effect (β), within and between cluster variances (γ^2 and σ^2) will be fixed at 1.

Part 3 (Extension to Poisson Model): We will repeat part 1 and 2 but changing to the setting in which Y follows a Poisson distribution. We aim to determine whether a Poisson model leads to different optimal designs under similar budget constraints and parameter settings, which we will compare the performance measures (absolute bias, variance, MSE, and coverage) between the normal and Poisson settings to assess how the data distribution affects design efficiency.

Performance measures: Since our target is the treatment effect, we evaluate the absolute bias, variance, mean square error (MSE), and coverage. Absolute bias measures systematic error, variance captures variability, MSE combines these two components, and coverage evaluates interval accuracy.

The formula for these performance measures is as follows: Absolute Bias: $|E[\hat{\beta}_\alpha] - \beta_1|$; Variance: $Var[\hat{\beta}_\alpha]$; MSE: $Abs.Bias^2 + Variance$; Coverage: The proportion of simulations where the true β_1 lies within the confidence interval for $\hat{\beta}_\alpha$.

Results

Table 1 shows the head of the simulated data. All simulated data for the varying parameters parts of the project follows the same structure. Additional column for the cost ratio information is included in the varying cost ratio data. With 500 iterations each combination, there are 18,000 and 9,000 observations simulated for the varying parameters data in Normal and Poisson distribution respectively. 3500 observations were simulated for the cost ratio sections.

Table 1: Head of Generated Data

beta	gamma2	sigma2	n_clusters	R_per_cluster	replication	estimated_beta	lower_confint	upper_confint
1	1	1	15	12	1	0.760	-0.128	1.649
1	1	1	15	12	2	0.898	-0.415	2.211
1	1	1	15	12	3	0.247	-0.688	1.182

Part 1 - Varying Parameters - Normal Distribution

Tables below show a subset of the results of varying only one of the parameters, and the rest are fixed at 1, with number of clusters as 15 and 12 observations per cluster. The beta estimate mean is very close to each true beta, and it had a high coverage for all the cases here **(Table 2)**. When we increase γ^2, the variance of the beta estimate and MSE significantly increase **(Table 3)**. When we increase σ^2, the variance of the beta estimate slightly increased, along with the MSE, but the absolute bias is close to 0 with a higher coverage **(Table 4)**.

Table 2: Varying Beta (gamma=1, sigma=2, n=15, R=12)

True Beta	Beta Estimate Mean	Variance	Coverage	Absolute Bias	MSE
1	0.976	0.322	0.922	0.024	0.323
2	2.019	0.328	0.926	0.019	0.328
3	3.003	0.315	0.928	0.003	0.315

Table 3: Varying gamma (Beta=1, sigma=1, n=15, R=12)

Gamma^2	Beta Estimate Mean	Variance	Coverage	Absolute Bias	MSE
1	0.976	0.322	0.922	0.024	0.323
10	0.957	2.848	0.930	0.043	2.850

Table 4: Varying sigma (Beta=1, gamma=1, n=15, R=12)

Sigma^2	Beta Estimate Mean	Variance	Coverage	Absolute Bias	MSE
1	0.976	0.322	0.922	0.024	0.323
10	1.000	0.457	0.950	0.000	0.457

The plots in the figure below shows the distribution of mean and variance of beta as we vary β, γ^2, and σ^2, faceted by the cluster design, and they represent the results from all possible combinations of β, γ^2, and σ^2 (**Figure 1**). From the first subplot, higher true beta generally lead to higher beta estimates, and the variability increases with smaller clusters (n = 5, R = 38). The variability is smallest for cluster design n = 15, R = 12, which is the largest cluster number and moderate level of observations per cluster. We can have a better view of this in the right, where we see the cluster design n = 5, R = 38 has the highest beta estimate variance. For the between-cluster variance γ^2, increasing γ^2 does not seem to affect beta mean significantly, but we see a significant difference for the variance. Higher γ^2 results in higher beta variance and smaller clusters (n = 5, R = 38) are more affected. Increasing σ^2 does not affect beta mean significantly, but there is higher variance in smaller clusters (n = 5, R = 38).

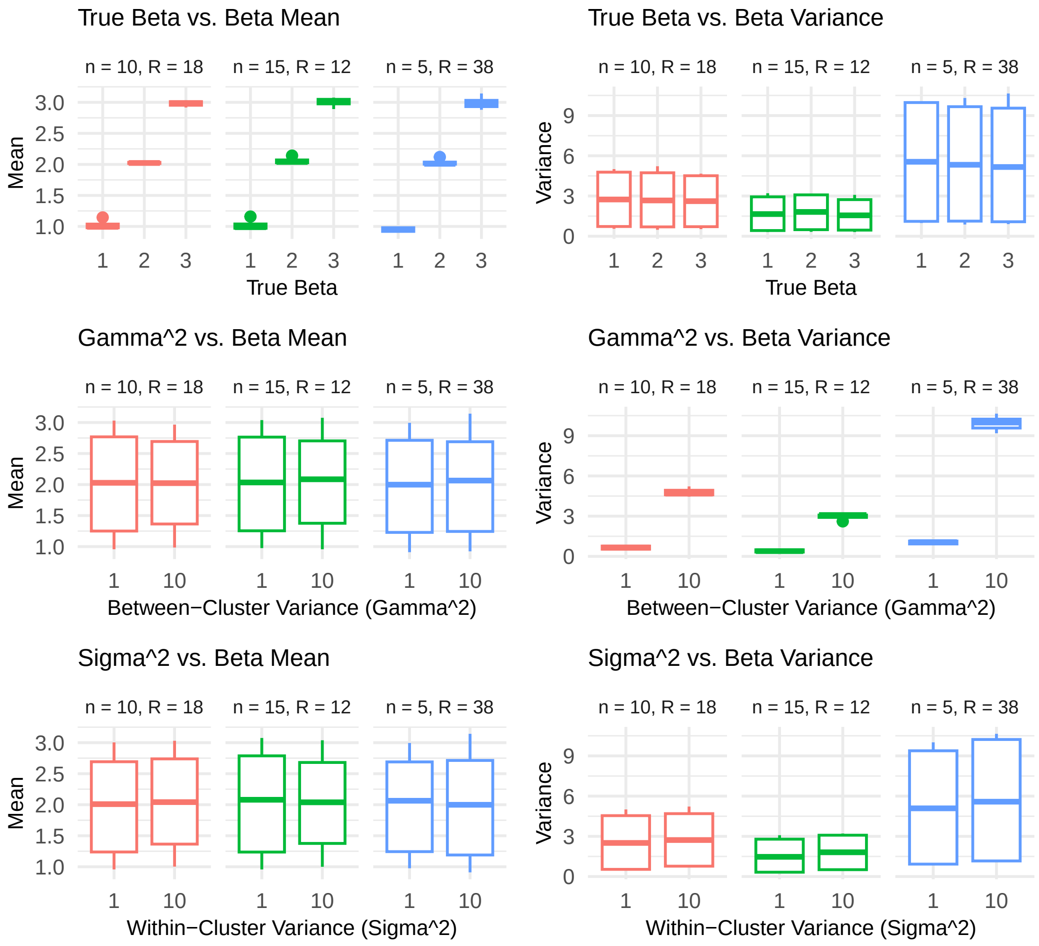

Figure 1: Results from Normal Distrbution - Varying Parameters

Figure 2 shows the boxplots of the three key performance metrics: Mean Squared Error

(MSE), Bias, and Coverage by cluster design. It is observable that smaller clusters with large observations (n = 5, R = 38) have the highest variability among all performance metrics, which indicates unstable and lower performance. Meanwhile, the larger clusters with moderate number of observations (n = 15, R = 12) have the most stable performance, with the lowest mean MSE and highest mean coverage. These findings suggest prioritizing larger clusters with moderate repetition to achieve stable and accurate parameter estimates.

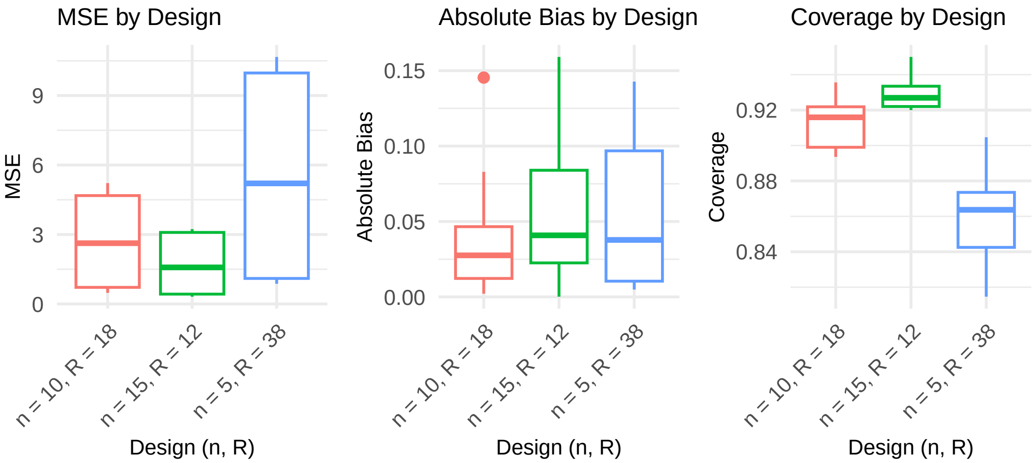

Figure 2: Normal - Performance Metrics by Cluster Design

Part 2 - Varying Cost Ratio - Normal Distribution

Table 5 shows all the cluster designs chosen by maximizing cost and varying the cost ratio.

Table 5: Cluster Designs

n_clusters	R_per_cluster	cost	ratio
15	12	1950	2:1
10	18	1900	2:1
5	38	1950	2:1
10	16	2000	50:1
5	36	2000	50:1
10	10	1900	10:1
5	30	1950	10:1

The plots below **(Figure 3)** illustrate how the cost ratio (c1: c2) affects the three key performance metrics: Mean Squared Error (MSE), Bias, and Coverage. As the cost ratio increases from 2:1 to 50:1, we observe distinct trends. In the MSE by Cost Ratio (left panel), the 2:1 cost ratio exhibits the highest variability but lowest mean in MSE, indicating that this cost ratio combination produces more varied results but still precise on average. The 10:1 ratio has the lowest MSE variability, suggesting that designs with a slightly higher initial cost for the first sample yield more stable estimates. For the Absolute Bias by Cost Ratio (middle panel), we have the lowest absolute bias for ratio 10:1. For Coverage by Cost Ratio (right panel), the 10:1 ratio has the highest mean, but the one for 2:1 is just marginally below, and the full range goes higher. The 50:1 ratio shows much lower and variable coverage, likely due to the trade-off between cluster size and the number of clusters. Thus, one should choose 2:1 to minimize MSE and maximize coverage, 10:1 to minimize absolute bias. It is important to consider which performance metric we want to use in this case.

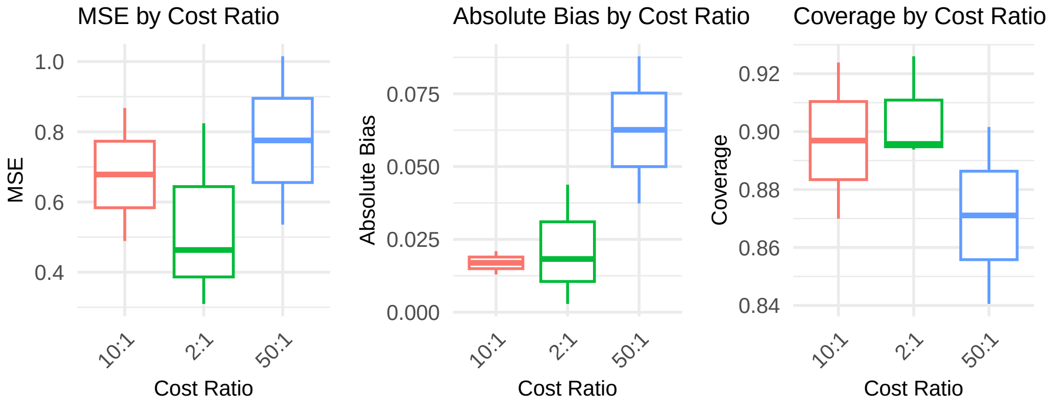

Figure 3: Normal - Performance Metrics by Varying Cost Ratio

Part 3 - Extension to a Poisson Model

We altered the outcome and re-generated simulations with a poisson model. Similar to that in part 1, the tables below are subset of the results of varying only one parameter, where the rest are fixed at 1 and we set the number of clusters as 15, with 12 observations per cluster. The beta estimate mean is also close to each true beta **(Table 6)**. When we increase γ, the variance of the beta estimate and MSE also significantly increase **(Table 7)**. The data generation process for poisson no longer include σ, so we do not have a table for that.

Table 6: Varying Beta (gamma=1, n=15, R=12)

True Beta	Beta Estimate Mean	Variance	Coverage	Absolute Bias	MSE
1	0.291	0.910	0.004	0.291	n = 15, R = 12
2	0.284	0.916	0.000	0.284	n = 15, R = 12
3	0.284	0.906	0.034	0.285	n = 15, R = 12

Table 7: Varying gamma (Beta=1, n=15, R=12)

Gamma^2	Beta Estimate Mean	Variance	Coverage	Absolute Bias	MSE
1	0.291	0.91	0.004	0.291	n = 15, R = 12
10	2.797	0.92	0.066	2.802	n = 15, R = 12

The plots in the figure below again shows the distribution of mean and variance of beta as we vary β, γ^2, and σ^2, faceted by the cluster design, and represents the results from all possible combinations of β, γ^2, and σ^2 but in a poisson distribution **(Figure 4)**. The results were almost identical to the normal case. We still see that higher true beta is leading to higher beta estimates, and the variability increases in smaller clusters (n = 5, R = 38). The variability is still smallest for cluster design n = 15, R = 12, which is the largest cluster number and moderate level of observations per cluster. The subplot on the right shows more detail on the variance distribution. For the between-cluster variance γ^2, increasing γ^2 does not have a consistent effect on beta mean between designs, but we see a significant difference for the variance. Higher γ^2 results in higher beta variance, and this is particularly observed in smaller clusters.

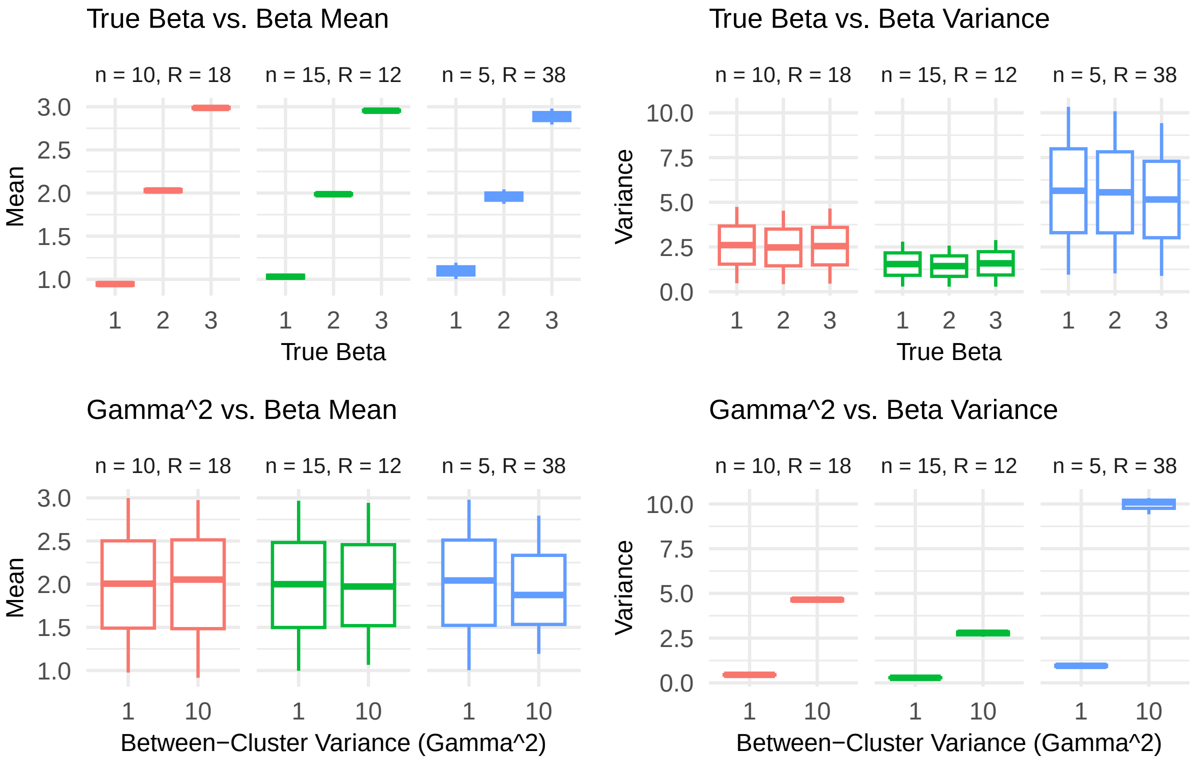

Figure 4: Results from Poisson Distrbution - Varying Parameters

Figure 5 shows the boxplots of the three key performance metrics: Mean Squared Error (MSE), Bias, and Coverage by cluster design. Similarly, smaller clusters with large observations (n = 5, R = 38) have the highest variability in all performance metrics, which means unstable and lower performance. Meanwhile, the larger clusters with moderate number of observations (n = 15, R = 12) have the most stable performance (least variability), with the lowest mean MSE and highest coverage. These findings also suggest prioritizing larger clusters with moderate repetition to achieve stable and accurate parameter estimates in a poisson distributed data.

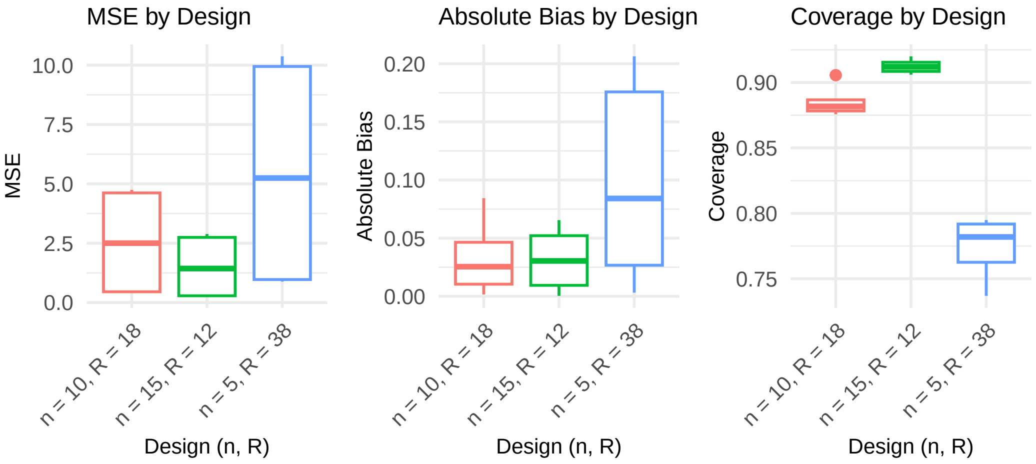

Figure 5: Poisson - Performance Metrics by Cluster Design

The plots below illustrate how the cost ratio (c1: c2) affects the three key performance metrics: Mean Squared Error (MSE), Bias, and Coverage **(Figure 6)**. There are a number of differences compared to the normal distributed data. As the cost ratio increases from 2:1 to 50:1, we again observe distinct trends. The lowest mean MSE is still 2:1. For the absolute biases, the mean bias is the lowest for 2:1, which is significantly different from the normal distribution case. Additionally, the absolute bias variability is the highest for 50:1. We also see that the 2:1 ratio achieves the highest mean coverage, suggesting that designs with lower initial cost seem to perform better in ensuring coverage. Thus, we conclude that a lower cost ratio (2:1) is an optimal cost ratio for poisson distributed data since it has the lowest MSE/bias and great coverage.

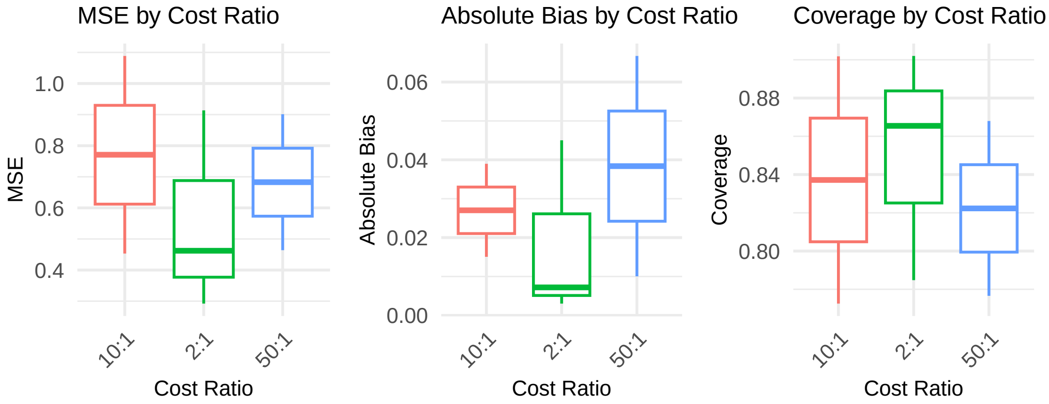

Figure 6: Poisson - Performance Metrics by Varying Cost Ratio

Discussion

The results provides a comprehensive evaluation of optimal experimental designs for cluster-randomized trials under different parameters and budget constraints, with actionable insights for balancing cost and statistical performance. Incorporating both Normal and Poisson data distributions also make the findings broadly applicable to numerous fields. The utilization of the ADEMP framework ensures a structured approach to study design with clear aims, data-generation mechanisms and rigorous evaluation metrics. Reproducibility is also ensured – the data simulation R script included formulas that one could easily simulate data by specifying parameters, budget constraints, and data distributions.

Despite the strengths of this project, several limitations should be addressed. While the simulation is thorough, the findings have not been validated using real-world data. Applying these designs to actual real-world data would strengthen their generalizability. Additionally, the use of only 500 iterations to each combination might also be too few for a robust statistical analysis. Increasing the number of iterations in future studies to 1,000 or more would improve the stability of performance metrics, such as mean squared error, bias, and coverage, leading to more robust conclusions. The cost model used in this study is another limitation. Although it effectively illustrates the trade-offs between fixed and variable costs, it is simplified and may not fully capture all the complexities of real-world budget designs.

Conclusions

This study demonstrates that optimal experimental design in cluster-randomized trials depends critically on the interplay between budget constraints, cost structures, and data-generating mechanisms.

Our findings highlight that designs with larger numbers of clusters and moderate within-cluster observations (n = 15, R = 12) consistently yield more stable and accurate estimates of the treatment effect, compared to the smaller ones. Variations in between-cluster (γ^2) and within-cluster (σ^2) variances significantly affect the estimates, with smaller clusters being more sensitive to increased variance. When extending to a Poisson distribution, results remain largely consistent, although overall lower absolute bias and variability are observed.

Cost ratio analysis demonstrates differences between normal and poisson distributed data. For normal, small initial costs (2:1) have a more precise result with lower MSE and higher coverage, while medium initial costs (10:1) has the lowest absolute bias. However, for poisson, lower initial costs (2:1) consistently have the lowest MSE, absolute bias, and higher coverage. These findings provide valuable insights into the trade-offs between cost, design complexity, and statistical performance.

Overall, this work offers a practical framework for optimizing experimental designs under constrained resources, with broad applicability across fields, including clinical trials and DNA sequencing. Future studies can build on these findings by incorporating real-world data validation, more complex cost models, and additional data-generating mechanisms to further refine design recommendations.

Code Appendix

```
# Load libraries
library(readr)
library(tidyverse)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(gridExtra) # For grid arrange
library(ggpubr)
library(lme4) # For lmer

th <- theme(plot.title = element_text(size = 8.8),
            axis.title.x = element_text(size = 8),
            axis.title.y = element_text(size = 8),
            axis.text.x = element_text(size = 8),
            axis.text.y = element_text(size = 8),
            legend.position = "bottom",
            legend.justification = "center",
            legend.box = "horizontal",
            legend.title = element_text(size = 6),
            legend.text = element_text(size = 6),
            legend.key.size = unit(0.3, 'cm'),
            strip.text.x = element_text(size = 7)
            )
simulation_results <- read_csv("outputs/normal_varied_parameter.csv")

head(simulation_results, 3) %>%
  kable(booktabs = TRUE, digits = 3, caption = "Head of Generated Data") %>%
  kableExtra::kable_styling(font_size = 7,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )
results <- simulation_results %>%
  group_by(beta, gamma2, sigma2, n_clusters, R_per_cluster) %>%
  summarize(
    mean_beta = mean(estimated_beta, na.rm = TRUE),
    Variance = var(estimated_beta, na.rm = TRUE),
    Coverage = mean(beta >= lower_confint & beta <= upper_confint, , na.rm =
     ↪  TRUE)
  ) %>%
  mutate(
    Abs.Bias = abs(mean_beta - beta),
    MSE = Abs.Bias^2 + Variance,
  ) %>%
  mutate(
    group = case_when(
      n_clusters == 15 & R_per_cluster == 12 ~ "n = 15, R = 12",
      n_clusters == 10 & R_per_cluster == 18 ~ "n = 10, R = 18",
      n_clusters == 5 & R_per_cluster == 38 ~ "n = 5, R = 38",
    )

  )


# Vary beta while fixing other parameters
vary_beta <- results %>%
  filter(gamma2 == 1, sigma2 == 1, n_clusters == 15, R_per_cluster == 12)

vary_beta[,c(1,6:10)] %>%
  kable(booktabs = TRUE, digits = 3, caption = "Varying Beta (gamma=1,
   ↪  sigma=2, n=15, R=12)",
        col.names = c("True Beta", "Beta Estimate Mean", "Variance",
                      "Coverage", "Absolute Bias", "MSE")
          ) %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )

# Vary gamma2 while fixing other parameters
vary_gamma2 <- results %>%
  filter(beta == 1, sigma2 == 1, n_clusters == 15, R_per_cluster == 12)

vary_gamma2[c(2,6:10)] %>%
  kable(booktabs = TRUE, digits = 3, caption = "Varying gamma (Beta=1,
   ↪  sigma=1, n=15, R=12)",
        col.names = c("Gamma^2", "Beta Estimate Mean", "Variance",
                      "Coverage", "Absolute Bias", "MSE")
          ) %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )

# Vary sigma2 while fixing other parameters
vary_sigma2 <- results %>%
  filter(beta == 1, gamma2 == 1, n_clusters == 15, R_per_cluster == 12)

vary_sigma2[c(3,6:10)] %>%
  kable(booktabs = TRUE, digits = 3, caption = "Varying sigma (Beta=1,
   ↪  gamma=1, n=15, R=12)",
        col.names = c("Sigma^2", "Beta Estimate Mean", "Variance",
                      "Coverage", "Absolute Bias", "MSE")
          ) %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )

# Beta vs. Beta estimate mean
a <- ggplot(results, aes(x = factor(beta), y = mean_beta, color = group)) +
  geom_boxplot() +
  facet_wrap(~ group , ncol = 3) +
  labs(
    title = "True Beta vs. Beta Mean",
    x = "True Beta",
    y = "Mean",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Beta vs. Beta estimate variance
b <- ggplot(results, aes(x = factor(beta), y = Variance, color = group)) +
  geom_boxplot() +
  facet_wrap(~ group , ncol = 3) +
  labs(
    title = "True Beta vs. Beta Variance",
    x = "True Beta",
    y = "Variance",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Gamma vs. Beta estimate mean
```

```r
c <- ggplot(results, aes(x = factor(gamma2), y = mean_beta, color = group)) +
  geom_boxplot() +
  facet_wrap(~ group, ncol = 3) +
  labs(
    title = "Gamma^2 vs. Beta Mean",
    x = "Between-Cluster Variance (Gamma^2)",
    y = "Mean",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Gamma vs. Beta estimate variance
d <- ggplot(results, aes(x = factor(gamma2), y = Variance, color = group)) +
  geom_boxplot() +
  facet_wrap(~ group , ncol = 3) +
  labs(
    title = "Gamma^2 vs. Beta Variance",
    x = "Between-Cluster Variance (Gamma^2)",
    y = "Variance",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Sigma vs. Beta estimate mean
e <- ggplot(results, aes(x = factor(sigma2), y = mean_beta, color = group)) +
  geom_boxplot() +
  facet_wrap(~ group, ncol = 3) +
  labs(
    title = "Sigma^2 vs. Beta Mean",
    x = "Within-Cluster Variance (Sigma^2)",
    y = "Mean",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Sigma vs. Beta estimate variance
f <- ggplot(results, aes(x = factor(sigma2), y = Variance, color = group)) +
  geom_boxplot() +
  facet_wrap(~ group , ncol = 3) +
  labs(
    title = "Sigma^2 vs. Beta Variance",
    x = "Within-Cluster Variance (Sigma^2)",
    y = "Variance",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

ggarrange(a, b, c, d, e, f, ncol = 2, nrow = 3, legend = "none")
a <- ggplot(results, aes(x = group, y = MSE, color = group)) +
    geom_boxplot()+
    labs(
        title = "MSE by Design",
        x = "Design (n, R)",
        y = "MSE"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

b <- ggplot(results, aes(x = group, y = Abs.Bias, color = group)) +
    geom_boxplot()+
    labs(
        title = "Absolute Bias by Design",
        x = "Design (n, R)",
        y = "Absolute Bias"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

c <- ggplot(results, aes(x = group, y = Coverage, color = group)) +
    geom_boxplot()+
    labs(
        title = "Coverage by Design",
        x = "Design (n, R)",
        y = "Coverage"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggarrange(a, b, c, ncol = 3, legend = "none")

cluster_designs <- read_csv("outputs/all_designs.csv")

cluster_designs %>%
  kable(booktabs = TRUE, digits = 3, caption = "Cluster Designs") %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )
simulation_results2 <- read_csv("outputs/normal_varied_cost.csv")

results2 <- simulation_results2 %>%
  group_by(beta, gamma2, sigma2, n_clusters, R_per_cluster, ratio) %>%
  summarize(
    mean_beta = mean(estimated_beta, na.rm = TRUE),
    Variance = var(estimated_beta, na.rm = TRUE),
    Coverage = mean(beta >= lower_confint & beta <= upper_confint, , na.rm =
     ↪  TRUE)
  ) %>%
  mutate(
    Abs.Bias = abs(mean_beta - beta),
    MSE = Abs.Bias^2 + Variance,
  )

a <- ggplot(results2, aes(x = ratio, y = MSE, color = ratio)) +
    geom_boxplot()+
    labs(
        title = "MSE by Cost Ratio",
        x = "Cost Ratio",
        y = "MSE"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

b <- ggplot(results2, aes(x = ratio, y = Abs.Bias, color = ratio)) +
    geom_boxplot()+
    labs(
        title = "Absolute Bias by Cost Ratio",
        x = "Cost Ratio",
        y = "Absolute Bias"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

c <- ggplot(results2, aes(x = ratio, y = Coverage, color = ratio)) +
    geom_boxplot()+
    labs(
        title = "Coverage by Cost Ratio",
        x = "Cost Ratio",
        y = "Coverage"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggarrange(a, b, c, ncol = 3, legend = "none")

simulation_results3.1 <- read_csv("outputs/poisson_varied_parameter.csv")

results3.1 <- simulation_results3.1 %>%
  group_by(beta, gamma2, n_clusters, R_per_cluster) %>%
  summarize(
    mean_beta = mean(estimated_beta, na.rm = TRUE),
    Variance = var(estimated_beta, na.rm = TRUE),
    Coverage = mean(beta >= lower_confint & beta <= upper_confint, , na.rm =
     ↪  TRUE)
  ) %>%
  mutate(
    Abs.Bias = abs(mean_beta - beta),
    MSE = Abs.Bias^2 + Variance,
  ) %>%
  mutate(
    group = case_when(
      n_clusters == 15 & R_per_cluster == 12 ~ "n = 15, R = 12",
      n_clusters == 10 & R_per_cluster == 18 ~ "n = 10, R = 18",
      n_clusters == 5 & R_per_cluster == 38 ~ "n = 5, R = 38",
    )

  )

# Vary beta while fixing other parameters
vary_beta <- results3.1 %>%
  filter(gamma2 == 1, n_clusters == 15, R_per_cluster == 12)

vary_beta[,c(1,6:10)] %>%
  kable(booktabs = TRUE, digits = 3, caption = "Varying Beta (gamma=1, n=15,
   ↪  R=12)",
        col.names = c("True Beta", "Beta Estimate Mean", "Variance",
                      "Coverage", "Absolute Bias", "MSE")
          ) %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )

# Vary gamma2 while fixing other parameters
vary_gamma2 <- results3.1 %>%
  filter(beta == 1, n_clusters == 15, R_per_cluster == 12)

vary_gamma2[c(2,6:10)] %>%
  kable(booktabs = TRUE, digits = 3, caption = "Varying gamma (Beta=1, n=15,
   ↪  R=12)",
        col.names = c("Gamma^2", "Beta Estimate Mean", "Variance",
                      "Coverage", "Absolute Bias", "MSE")
          ) %>%
  kableExtra::kable_styling(font_size = 9,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position")
                            )
# Beta vs. Beta estimate mean
a <- ggplot(results3.1, aes(x = factor(beta), y = mean_beta, color = group))
 ↪  +
  geom_boxplot() +
  facet_wrap(~ group , ncol = 3) +
  labs(
    title = "True Beta vs. Beta Mean",
    x = "True Beta",
    y = "Mean",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Beta vs. Beta estimate variance
b <- ggplot(results3.1, aes(x = factor(beta), y = Variance, color = group)) +
  geom_boxplot() +
  facet_wrap(~ group , ncol = 3) +
  labs(
    title = "True Beta vs. Beta Variance",
    x = "True Beta",
    y = "Variance",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Gamma vs. Beta estimate mean
c <- ggplot(results3.1, aes(x = factor(gamma2), y = mean_beta, color =
↪  group)) +
  geom_boxplot() +
  facet_wrap(~ group, ncol = 3) +
  labs(
    title = "Gamma^2 vs. Beta Mean",
    x = "Between-Cluster Variance (Gamma^2)",
    y = "Mean",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

# Gamma vs. Beta estimate variance
d <- ggplot(results3.1, aes(x = factor(gamma2), y = Variance, color = group))
↪  +
  geom_boxplot() +
  facet_wrap(~ group , ncol = 3) +
  labs(
    title = "Gamma^2 vs. Beta Variance",
    x = "Between-Cluster Variance (Gamma^2)",
    y = "Variance",
    color = "Cluster Design"
  ) +
  theme_minimal() + th

ggarrange(a, b, c, d, ncol = 2, nrow = 3, legend = "none")
a <- ggplot(results3.1, aes(x = group, y = MSE, color = group)) +
    geom_boxplot()+
    labs(
        title = "MSE by Design",
        x = "Design (n, R)",
        y = "MSE"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

b <- ggplot(results3.1, aes(x = group, y = Abs.Bias, color = group)) +
    geom_boxplot()+
    labs(
        title = "Absolute Bias by Design",
        x = "Design (n, R)",
        y = "Absolute Bias"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

c <- ggplot(results3.1, aes(x = group, y = Coverage, color = group)) +
    geom_boxplot()+
    labs(
        title = "Coverage by Design",
        x = "Design (n, R)",
        y = "Coverage"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggarrange(a, b, c, ncol = 3, legend = "none")

simulation_results3.2 <- read_csv("outputs/poisson_varied_cost.csv")

results3.2 <- simulation_results3.2 %>%
  group_by(beta, gamma2, n_clusters, R_per_cluster, ratio) %>%
  summarize(
    mean_beta = mean(estimated_beta, na.rm = TRUE),
    Variance = var(estimated_beta, na.rm = TRUE),
    Coverage = mean(beta >= lower_confint & beta <= upper_confint, , na.rm =
     ↪  TRUE)
  ) %>%
  mutate(
    Abs.Bias = abs(mean_beta - beta),
    MSE = Abs.Bias^2 + Variance,
  )
a <- ggplot(results3.2, aes(x = ratio, y = MSE, color = ratio)) +
    geom_boxplot()+
    labs(
        title = "MSE by Cost Ratio",
        x = "Cost Ratio",
        y = "MSE"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
b <- ggplot(results3.2, aes(x = ratio, y = Abs.Bias, color = ratio)) +
    geom_boxplot()+
    labs(
        title = "Absolute Bias by Cost Ratio",
        x = "Cost Ratio",
        y = "Absolute Bias"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

c <- ggplot(results3.2, aes(x = ratio, y = Coverage, color = ratio)) +
    geom_boxplot()+
    labs(
        title = "Coverage by Cost Ratio",
        x = "Cost Ratio",
        y = "Coverage"
    ) +
    theme_minimal() + th +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggarrange(a, b, c, ncol = 3, legend = "none")
```