Understanding Optimal Study Design in Causal Inference through Simulation Studies

Daniel Posmik

2024-12-01

Table of contents

Abstract

In this project, we explore optimal designs for cluster randomized trials. We begin by discussing the underlying data generating mechanisms and provide a detailed summary of our goals and procedures using the ADEMP framework. We simulate the results for different numbers of clusters and identify the point of minimum variance. We also provide a brief theoretical discussion on optimal design in cluster randomized trials, building off Raundenbush (1997). We then explore how the choices of different parameters impact the optimal study design. We vary the true fixed and treatment effects, the treatment assignment probability, the sampling costs, and the variances of the noise terms. Lastly, we extend our simulation study to the setting in which the outcome variable follows a Poisson distribution. All in all, we find that sampling more clusters tends to minimize the variance of the treatment effect estimator. This result holds across a reasonable range of varied parameters. That being said, this result is only sensitive to the cost of sampling individuals within a cluster. When individual sampling cost is low, we can expect to see the variance-minimizing number of clusters to be on the left side of the peak. These results remain consistent across the normal and Poisson cases.

Designing the Optimal Simulation Study

We shall begin our analysis with the discussion of the underlying data generating mechanisms (DGM) and the relationship amongst them. The below figure illustrates the structure of our data and statistical objects of interest:

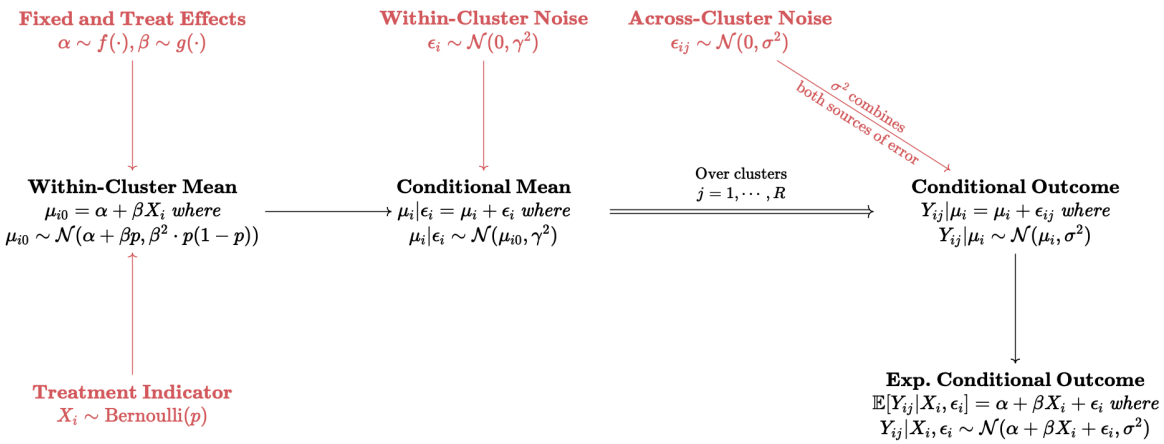

Figure 1: Data Generating Mechanisms and Data Structure

Here, the sources of variability are colored in light red. We let the treatment indicator X_i be a Bernoulli random variable with parameter p. The fixed effect α and treatment effect β underlie the DGMs $f(\cdot)$ and $g(\cdot)$, respectively. The noise terms ϵ_i and ν_j are normally distributed with expectation 0 and variances γ^2 and σ^2, respectively.

As a first step in this simulation project, we provide a detailed summary of our goals and procedures using the ADEMP framework, formalized by Morris et al. (2019). The ADEMP framework consists of five components.

Objectives	Explanation
Aims (A)	We want to determine an optimal study design in cluster randomized trials in terms of optimal number of clusters G^* and optimal number of sampled individuals R^*. For that purpose, we will use the combination of G^* and R^* that minimizes the empirical variance of the causal estimator $\widehat{\beta}$.
Data-Generating Mechanisms (D)	We will consider two cases of simulating the outcome variable Y. First, we will consider the normal case where $Y \sim \mathcal{N}(\mu_i, \sigma^2)$. In the last part of the project, we will go beyond the normal setting and consider a Poisson-distributed outcome variable, i.e. $Y \sim \text{Poisson}(R \cdot \mu_i)$. Across both of these cases, the estimates of β are extracted from a random effects model. Moreover, we will simulate our cluster-level treatment as a Bernoulli random variable with parameter p. Our within-cluster and across-cluster noise terms are 0 in expectation and have variances γ^2 and σ^2, respectively.
Estimand/ Target of Analysis (E)	Within the model, our estimand $\widehat{\beta}$ denotes the average treatment effect (ATE). Regarding our goal of optimizing our study design, we are interested in the optimal number of clusters G^* that minimizes the variance of $\widehat{\beta}$, conditional on design constraints like budget, cost, etc.

Objectives	Explanation
Methods (M)	In order to infer a robust result on the pair (G^*, R^*) that optimizes our study design, we simulate. Moreover, we will vary input parameters, i.e. the true fixed and treatment effect, the treatment assignment probability, and cost.
Performance Measures (P)	Our performance measure of choice is the minimum empirical variance across (G^*, R^*).

To begin, we will consider the case when the parameters are fixed. Our strategy will be set the parameters to reasonable values and observe the variance of the causal estimator $\hat{\beta}$ across different combinations of optimal number of individuals R and clusters G. Note that there is an inverse relationship between R and G.

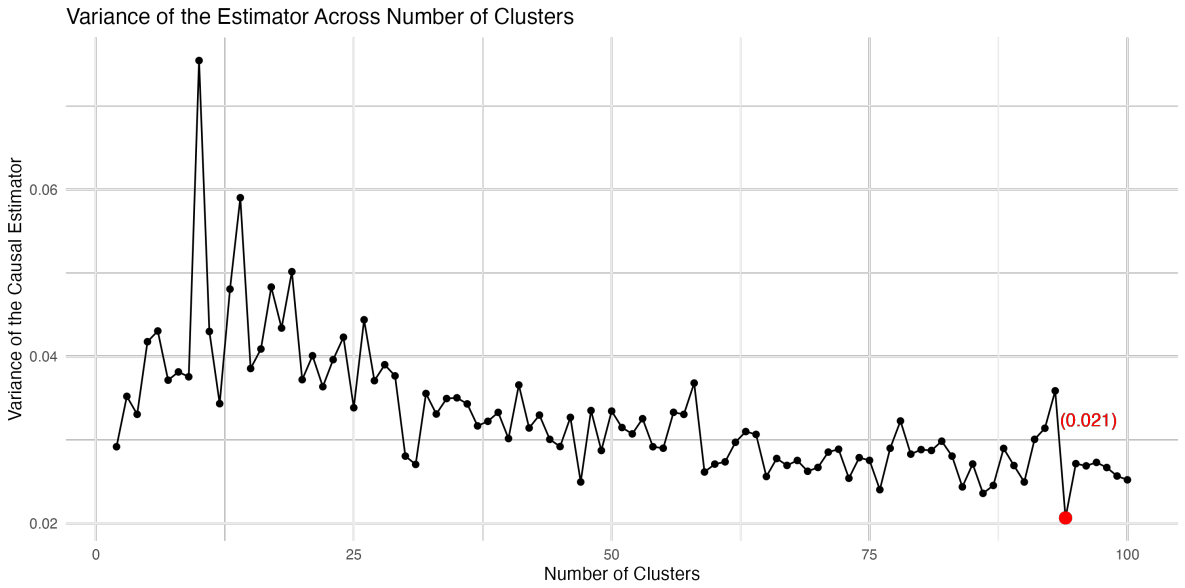

Figure 2: Variance of the Estimator Across Number of Clusters

We observe an interesting relationship between the choice of number of clusters and the empirical variance of the causal estimator. Although very small numbers of clusters lead to relatively low variance, the variance sharply increases as the number of clusters goes from very few to "a handful". After this spike, the variance of the causal estimator decreases and gradually approaches 0. We can locate the point of minimum variance at a large value of G, but not the largest. This initial spike is likely due to the cost c_2 we chose. If we chose a higher value for

c_2, we would likely less of a spike in the beginning since there is immediate benefit of sampling more clusters from the start.

Before we summarize results, we want to interject a brief discussion on the theoretical work in the field of optimal design in cluster randomized trials. Rutterford et al. (2015) provide a broad survey and point to the work of Raudenbush (1997) and Moerbeck et al. (2000) as seminal contributions to the field. Raudenbush (1997) develops the estimators for the optimal number of clusters and individuals to sample from. Moerbeck et al. (2000) provide helpful examples around the former's results.

We can adjust the formulas of Raundenbush (1997) to our case. Subject to our cost function

$$T = c_1 + c_2 \cdot (R - 1)$$

we can now determine the optimal number of clusters and the optimal number of individuals to sample from. Adapting the results from Raudenbush (1997), we have the following:

$$\begin{aligned} R^* - 1 &= \frac{\sigma}{\gamma}\sqrt{\frac{c_1}{c_2}} \implies R^* = \lfloor \frac{\sigma}{\gamma}\sqrt{\frac{c_1}{c_2}} + 1 \rfloor \\ G^* &= \frac{T}{\frac{\sigma}{\gamma} \cdot \sqrt{c_1 \cdot c_2} + c_1} = \lfloor \frac{T}{(R^* - 1)c_2 + c_1} \rfloor \\ \mathrm{Var}(\hat{\beta}) &= \frac{4(\gamma^2 + \sigma^2/R^*)}{R^* \cdot G^*} \end{aligned}$$

where the $\lfloor \cdot \rfloor$ operator denotes the floor function, i.e. we must round down to the nearest integer. γ and σ denote the withon-cluster and across-cluster standard deviations, respectively. G^* and R^* are the optimal number of clusters and the optimal number of individuals to sample from, respectively. The equation for R^* tells us that larger R is best when the variability within a cluster is larger and when the cost of sampling additional individuals within a chosen cluster is larger, as compared with the cost of sampling more clusters. Overall, our goal is to select the combination of R and G that minimizes the variance of our treatment effect estimator $\hat{\beta}$.

Moreover, we note that that the intracluster correlation is denoted by $\rho = \frac{\gamma^2}{\gamma^2 + \sigma^2}$. As in Raudenbush (1997), we impose the constraint $\gamma^2 + \sigma^2 = 1$. Then, we have $\rho = \gamma^2$ and $\sigma^2 = 1 - \rho$ and having introduced the parameter ρ, we can rewrite the above expression as follows:

$$R^* = \lfloor \sqrt{\frac{1-\rho}{\rho}}\sqrt{\frac{c_1}{c_2}} + 1 \rfloor; \quad G^* = \lfloor \frac{T}{(R^* - 1)c_2 + c_1} \rfloor; \quad \mathrm{Var}(\hat{\beta}) \quad = \frac{4(\rho + (1-\rho)/R^*)}{R^* \cdot G^*}$$

yields the results summarized in Table 2.

Table 2: Optimal Sampling Strategy for Different Intracluster Correlations and Sampling Costs

Intracluster Correlation (rho)	**Cost of Sampling Individual (c2)**	**Optimal Number of Individuals (R)**	**Optimal Number of Clusters (G)**	**Variance of the Causal Estimator**
0.01	2	50	3	0.0007947
0.01	10	23	1	0.0092250
0.01	40	12	1	0.0308333
0.05	2	22	5	0.0033884
0.05	10	10	3	0.0193333
0.05	40	5	2	0.0960000
0.10	2	16	6	0.0065104
0.10	10	7	4	0.0326531
0.10	40	4	2	0.1625000
0.20	2	11	7	0.0141677
0.20	10	5	5	0.0576000
0.20	40	3	3	0.2074074
0.50	2	6	8	0.0486111
0.50	10	3	7	0.1269841
0.50	40	2	5	0.3000000

In Table 2, we see the results of our theoretical calculations. We observe that variance increases with intracluster correlation ρ and individual sampling cost c_2. This is a nice addition to the simulation results in cases where we may have access to important parameter information, such as the intracluster correlation, and we would like to derive G^* when we know our total number of individuals across clusters. Overall, we find that the value these theoretical results provide could provide needed context for researchers seeking to further test optimal study designs in cluster randomized trials.

All in all, find that sampling more clusters tends to minimize the variance of the treatment effect estimator. Specifically, in our simulation above, choosing $G^* = 94$ minimizes the variance of the causal estimator at $\text{Var}(\widehat{\beta}) = 0.021$.

The Impact of Parameter Choices on Optimal Study Design

As a logical next step, we hope to explore how the choices of different parameters impact the optimal study design. We are most interested in varying:

- α, β: The true fiexed and true treatment effect, respectively
- p: The treatment assignment probability
- c_1, c_2: The sampling costs
- γ^2: The variances of the noise terms (holding σ^2 constant)

This step is distinct from the theoretical results in the previous section in that we are returning to the simulation framework. To simulate the costs of sampling, we will hold constant c_1, effectively varying the ratio of initial cluster cost and cost per individual. For the treatment assignment parameter p, we will choose one low and one high value, reflecting scenarios where the treatment may be significantly more or less likely to be assigned. For the treatment effect β, we will choose a "low" and "high" value to explore the impact of the treatment effect on the optimal study design. We will proceed similarly for α. We will keep other parameters constant. Namely, for simplicity, we will keep the budget fixed. For simplicity, we will keep σ^2 fixed and vary γ^2.

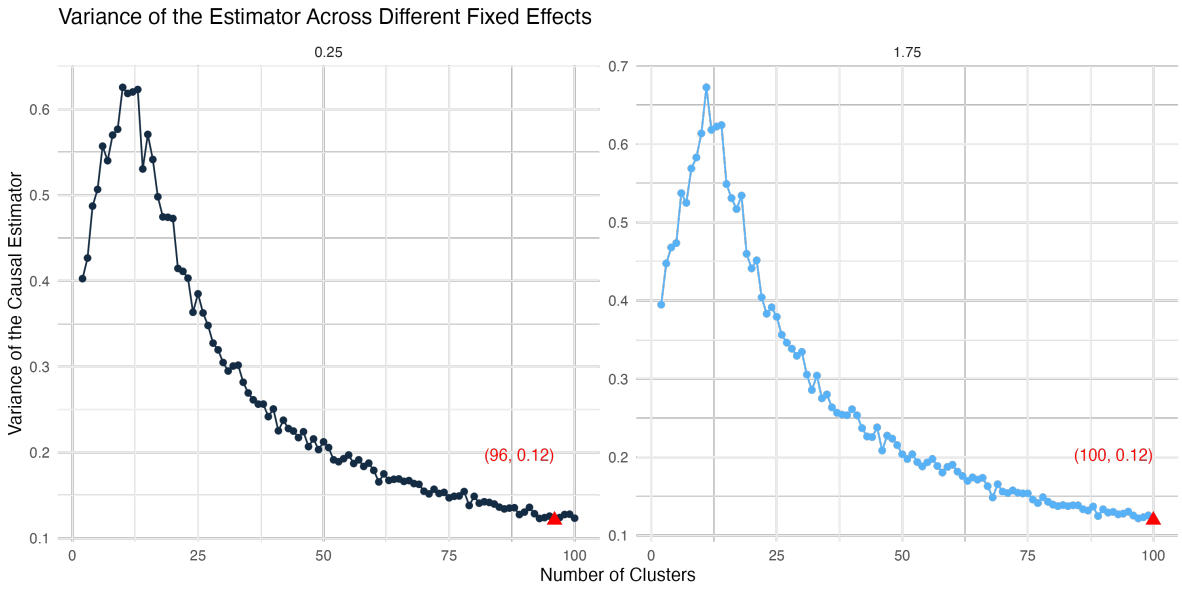

Figure 3: Variance of the Estimator Across Different Fixed Effects

Varying across α: Looking at the different values of α, we observe that the variance of the treatment effect estimator is relatively stable across different fixed effects. This is intuitive, since the fixed effect is shared across treatment groups. We also observe that the number of clusters that minimizes the variance of the treatment effect estimator is unformly large.

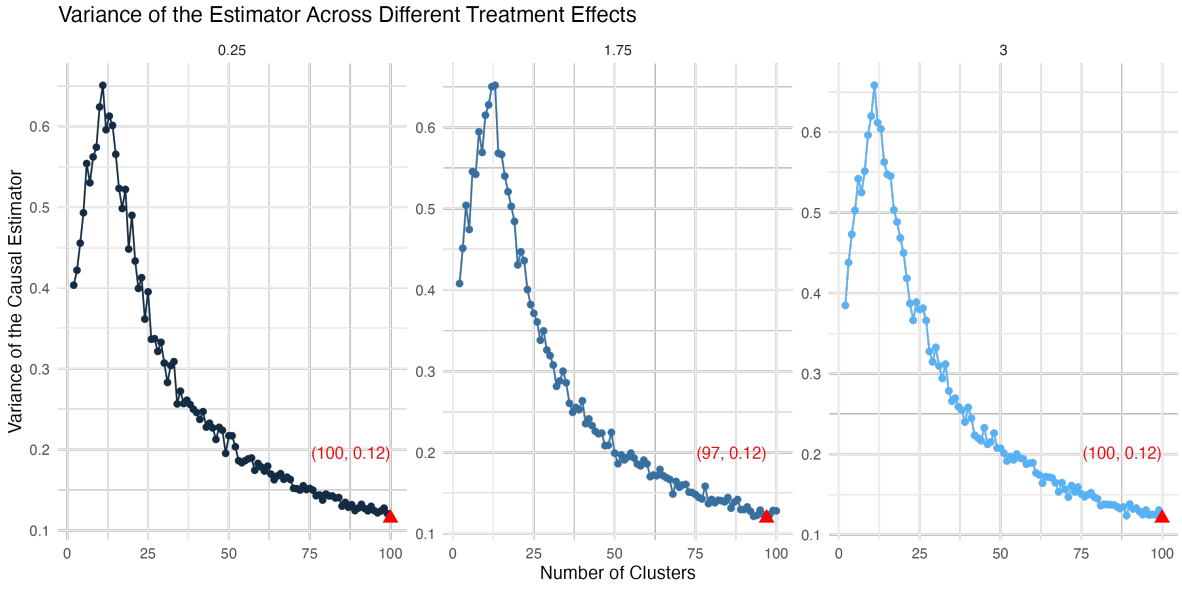

Figure 4: Variance of the Estimator Across Different Treatment Effects

Varying across β: The treatment effect β does not seem to have a large impact on its variance, either. This is somewhat surprising since we would have expected a slightly higher variance-minimizing number of clusters when the treatment effect is lower. We would have expected this to be due to the fact that treatment is randomized at the cluster-level, and when the treatment effect is lower, we need more clusters to detect the effect.

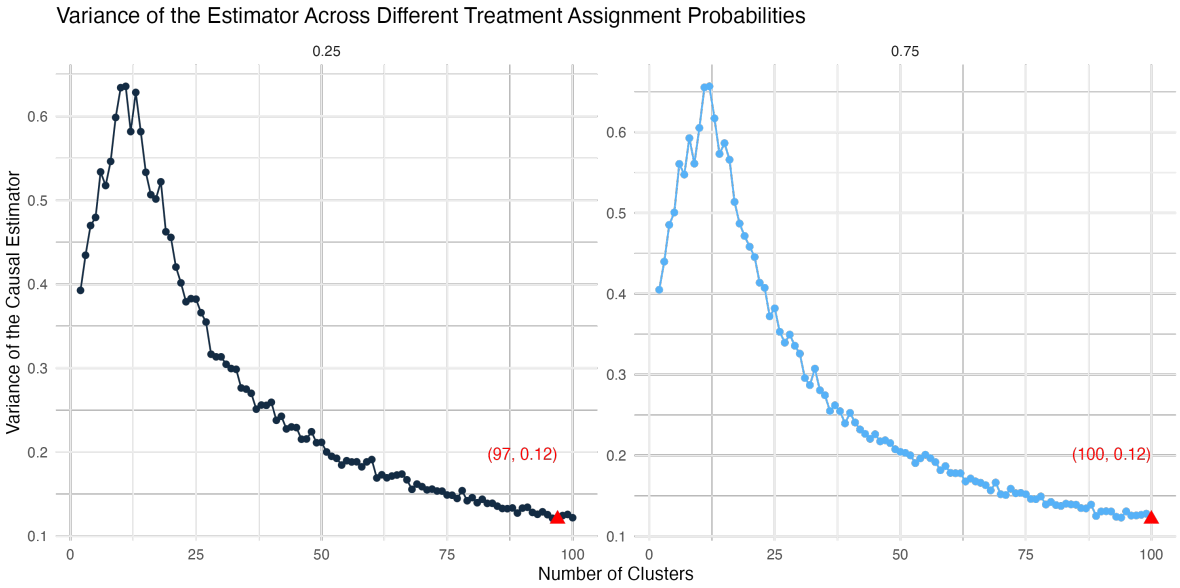

Figure 5: Variance of the Estimator Across Different Treatment Assignment Probabilities

Varying across p: Interestingly, when we vary the probability of receiving treatment p, we observe the result we expected for varying β slightly more clearly. Similarly, when p is lower, we likely need more clusters to optimize our estimation procedure. Again, this is likely due to the need to detect a meaningful treatment effect across more clusters, since if $X_i = 0$, the β term is zero, as well.

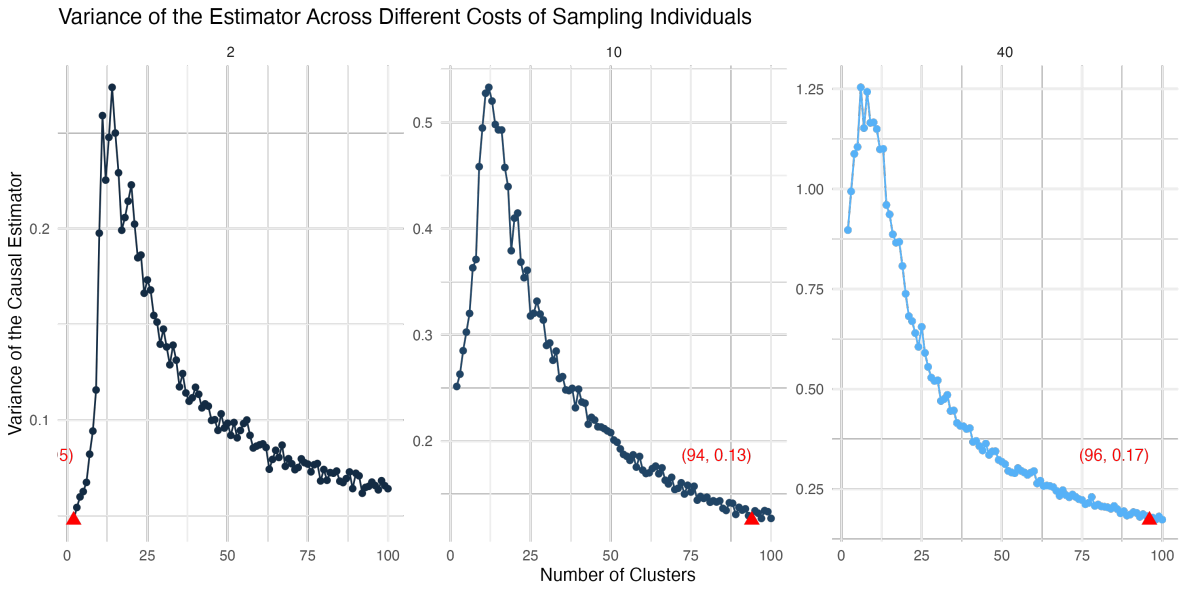

Figure 6: Variance of the Estimator Across Different Costs of Sampling Individuals

Varying across c_2: When collapsing by the individual sample cost c_2, we observe find that the number of clusters that minimizes the variance of the treatment effect estimator increases with c_2. For the case when c_2 is very small (i.e. $c_2 = 2$), we even find that the the variance-minimizing number of clusters is on the left side of the peak. This can be easily explained: If the cost of sampling individuals is low, we can greedily sample individuals with few clusters. Intuitively, a smaller amount of clusters "gets the job done" when the cost of sampling individuals is low.

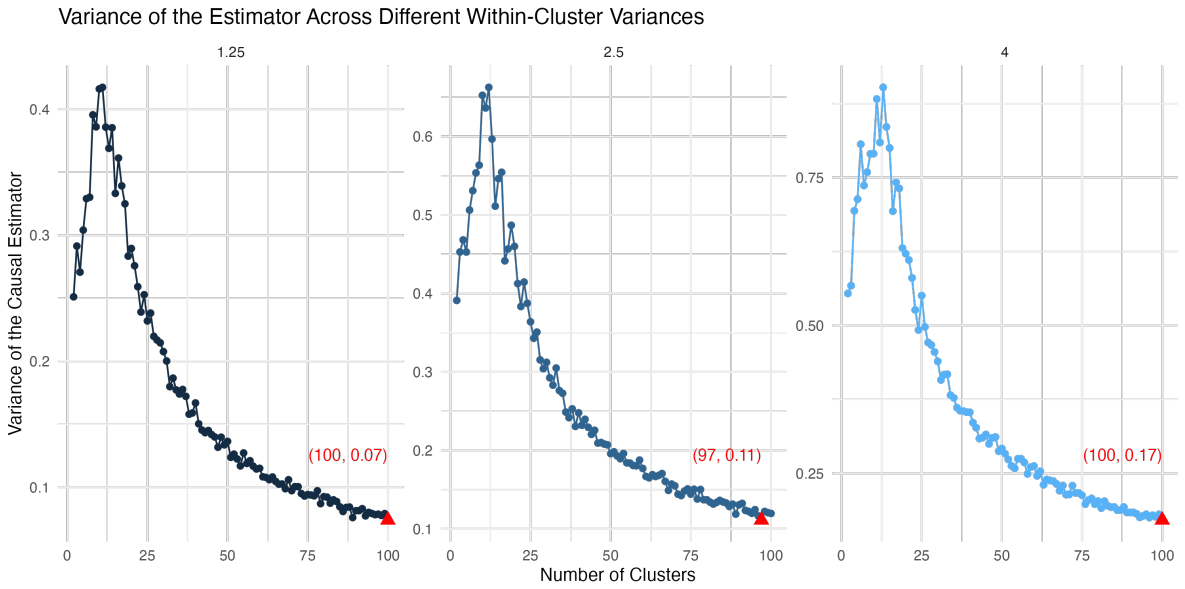

Figure 7: Variance of the Estimator Across Different Within-Cluster Variances

Varying across γ^2: We see that varying γ^2 does not have a large impact on the number of clusters that minimizes the variance of the treatment effect estimator. This is somewhat surprising, since when lowering/raising the within-cluster variance γ^2, holding constant σ^2, we effectively lower/ raise the intracluster correlation ρ. This means that if γ^2 were lower, we would expect the individuals to be more similar across clusters, and thus need fewer clusters. We fail to observe this in our simulation but this may well be because our parameter choices are not extreme enough.

All in all, we find that the number of clusters that minimizes the variance of the treatment effect estimator is relatively stable across different choices of parameters. The notable exception here is the cost of sampling individuals within a cluster. When that is low, we can expect to see the variance-minimizing number of clusters to be on the left side of the peak, i.e. low.

Beyond Normal: A Look at Count Data

Lastly, we will consider the case when our data are not normally distributed. Specifically, suppose we have

$$Y_i|\mu_i \sim \text{Poisson}(R\mu_i) \text{ where } Y_i := \sum_{j=1}^{R} Y_{i,j}; \quad Y_{ij}|\mu_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) \sim \mathcal{N}(\alpha + \beta X_i, \gamma^2)$$

We will rewrite the simulation functions from the linear model case and revisit the results. We will make use of Poisson regression to model the count data. On a preliminary note, we want to point out that the variance and mean are now related, since for a Poisson random variable Y_i, we have $\text{Var}(Y_i) = \mathbb{E}(Y_i)$. This is in contrast to the normal distribution and out previous example, where the sample variance is independent of the sample mean.

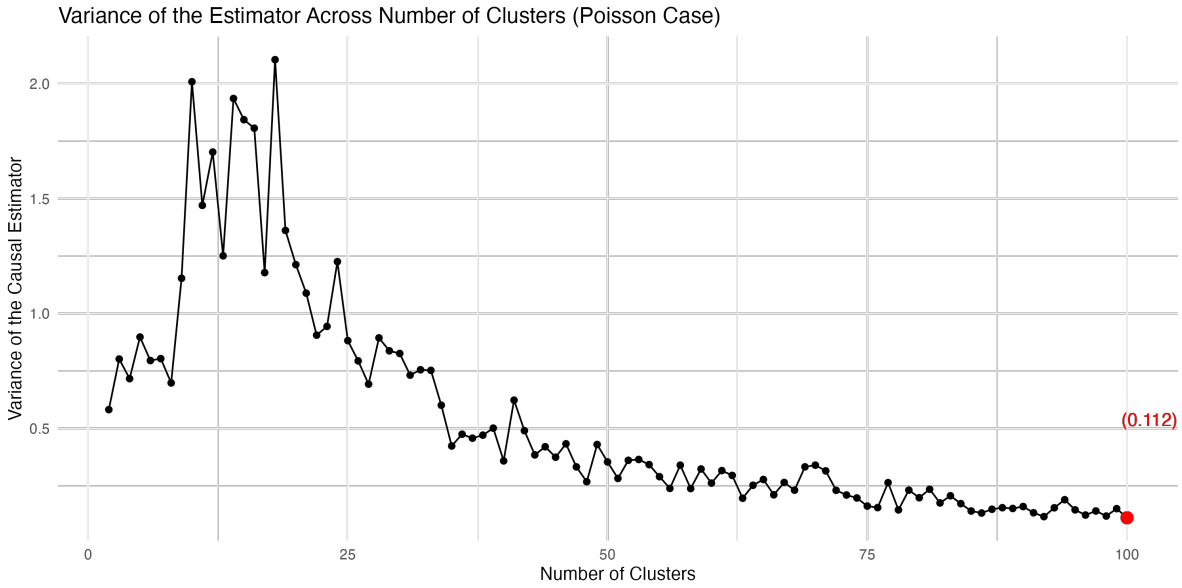

Figure 8: Variance of the Estimator Across Number of Clusters (Poisson Case)

We observe a similar pattern to the normal case, where the variance of the causal estimator decreases as the number of clusters increases. The variance of the causal estimator is minimized at $G^* = 100$ with $\text{Var}(\hat{\beta}) = 0.112$. Although this suggests that the optimal number of clusters is not very sensitive to count data in the outcome variable, we do see that overall variance is significantly higher in the Poisson case (compare 0.112 in the Poisson case with 0.021 in the normal case).

Furthermore, given our previous findings in the sensitivity analysis, we want to ask how sensitive the variance-minimizing number of clusters is to the cost of sampling individuals within a cluster. The plot below shows the variance of the causal estimator across different costs of sampling individuals within a cluster.

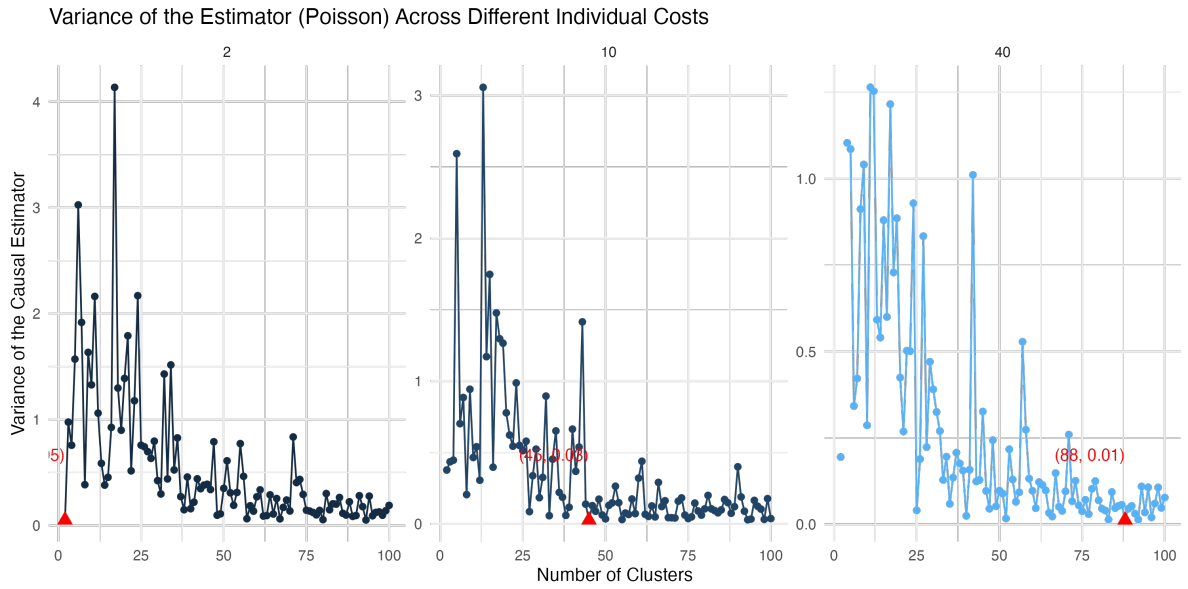

Figure 9: Variance of the Estimator Across Different Costs of Sampling Individuals (Poisson Case)

We can observe a trend that is very similar to the sensitivity analysis in the normal case: Very low c_2 leads a low number of clusters that minimize the variance of the treatment effect estimator. We want to acknowledge that due to time constraints, we were only able to perform $n = 50$ simulations in the Poisson case. This is why we see generally noisier results in the Poisson case.

Before we conclude, we want to state a theoretical result from the literature. If we are interested in the combination of (R^*, G^*) that optimizes our study design, "multiplication of the sample size calculation for ordinary Poisson regression by the standard design effect" (Rutterford et al., 2015, p. 1055) can be used to calculate the number of individuals in a cluster:

$$R^* = \lfloor \frac{\left(Z_{\alpha/2}\sqrt{2} + Z_\lambda \sqrt{[1 + e^{-\hat{\beta}}]}\right)^2}{e^{\lambda_0}\hat{\beta}^2} \cdot [1 + (r-1)\cdot\rho] \rfloor$$

where Z_x is the x'th percentage point on the standard normal distribution. λ^0 represents the event rate in the control group and $\hat{\beta}$ is the estimated treatment effect. r is the number of individuals in a cluster and ρ is the intracluster correlation. This method assumes an analysis by Generalised Estimating Equations (GEE) and fixed cluster size. The latter property makes it less suitable for our simulation study, but it warrants inclusion in this report since it is a useful tool for researchers seeking to optimize their study design in the context of count data. This result is derived from the work of Amatya et al. (2013).

References

- Amatya, A., Bhaumik, D., & Gibbons, R. D. (2013). Sample size determination for clustered count data. Statistics in medicine, 32(24), 4162–4179. https://doi.org/10.1002/sim.5819
- Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019; 38: 2074–2102. https://doi.org/10.1002/sim.8086
- Moerbeek, M., Gerard J. P. van Breukelen, & Martijn P. F. Berger. (2000). Design Issues for Experiments in Multilevel Populations. Journal of Educational and Behavioral Statistics, 25(3), 271–284. https://doi.org/10.2307/1165206
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. Psychological Methods, 2(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173
- Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. International journal of epidemiology, 44(3), 1051–1067. https://doi.org/10.1093/ije/dyv113

Code Appendix

```
# Set up knit environment
knitr::opts_chunk$set(echo = F)
knitr::opts_chunk$set(error = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(message = F)

# Libraries
library(knitr)
library(kableExtra)
library(tidyverse)
library(lme4)
library(purrr)

# Source functions
source("/Users/posmikdc/Documents/brown/classes/php2550-pda/php2550-projects/project3/code/0
# Input parameters
rho <- c(0.01, 0.05, 0.1, 0.2, 0.5)
c1 <- 50 # Cost of sampling the first individual in a cluster
c2 <- c(2, 10, 40) # Cost of sampling the i-th individual in a cluster, i = 2, ..., n
B <- 500 # Budget

# Initialize an empty data frame to store results
df.theory <- data.frame(rho = numeric(0),
                        c2 = numeric(0),
                        ROpt = numeric(0),
                        GOpt = numeric(0),
                        var_beta = numeric(0))

# Loop over the parameters
for (r in rho){
  for (c in c2){
    # Calculate results for each iteration
    results <- sample_strategy(rho = r, c1 = c1, c2 = c2, B = B)
  }
  # Bind results to a data frame
  df.theory <- rbind(df.theory, results)
}

# Save the table as .csv
```

```r
write.csv(df.theory,
  "/Users/posmikdc/Documents/brown/classes/php2550-pda/php2550-projects/project3/output/sim-t
  row.names = FALSE)

# Create table
tbl.theory <- df.theory %>%
  kbl(
    escape = FALSE,
    col.names = c("Intracluster Correlation (rho)",
      "Cost of Sampling Individual (c2)",
      "Optimal Number of Individuals (R)",
      "Optimal Number of Clusters (G)",
      "Variance of the Causal Estimator"),
    caption = "Optimal Sampling Strategy for Different Intracluster Correlations and Sampling
  ) %>%
  kable_styling(
    full_width = TRUE,
    font_size = 10, # Adjust font size to make it more compact
    position = "center" # Center-align the table
  ) %>%
  row_spec(0, bold = TRUE) %>%
  row_spec(seq(3, nrow(df.theory), by = 3), hline_after = TRUE)

tbl.theory
```