Simulation Study to Optimize Cluster Size and Number of Clusters Given Budget Constraints for a Cluster Randomized Controlled Trial

Rachel Yost

2024-12-05

Abstract

Cluster Randomized Trials are often a useful study design, but can be costly to implement. Outcomes for patients within a cluster are not independent like they are in randomized controlled trials, so researchers must carefully consider their study design. The number of clusters and the size of each cluster must be selected to attain enough statistical power to identify a treatment effect, while also staying within their budgets. It is likely that adding a new cluster to a trial will be more expensive than adding observations to an existing cluster. In this study, we simulate a cluster randomized trial with a fixed budget, and evaluate how the optimal number of clusters and observations per cluster change as data generation parameters are varied. We generate data from a hierarchical model under different values for the within and between cluster variance. We also evaluate how the optimal study design changes based on the relative costs of adding an observation in a new cluster compared to an existing cluster. Data is initially generated following a normal distribution, and then we repeat out simulations on data with a poisson outcome. We found that as the within cluster variance increases relative to the between cluster variance, the optimal number of clusters decreases.

Introduction

Randomized controlled trials (RCTs) are often considered the gold-standard in research studies (Hariton and Locascio, 2018) since randomization evenly distributes participant characteristics between the treatment groups. If participant characteristics are very similar between groups, differences in the outcome variable can be attributed to the intervention. However, sometimes a cluster randomized controlled trial is the preferable option for study design. A cluster randomized controlled trial randomizes entire groups of participants rather than individuals (Eldridge and Wiley, 2012). One reason to use a cluster RCT rather than an RCT is if the target intervention is meant to be applied to an entire group, such as with a hospital-wide or system-wide change that could affect patient outcomes (Heagerty, 2017). Another reason to use clusters is if there is a significant risk of contamination in the study (Heagerty, 2017), which is defined as "when members of the 'control' group inadvertently receive the treatment or are exposed to the intervention" (Dron et al., 2021).

However, in a cluster randomized trial, outcomes for participants are no longer independent, as they are in RCTs, since individuals within a cluster are more likely to have similar results. An example is if an intervention is applied at a hospital level (Campbell and Grimshaw, 1998). Patients within one hospital might be more similar to one another due to socioeconomic status, so patient outcomes within a hospital might be more similar than patient outcomes between different hospitals.

When designing a cluster randomized trial, we must consider the number of clusters and number of observations per cluster to include in the study. It is likely that the first observation in a cluster c_1 will cost more than the subsequent observations c_2, since this could involve adding another hospital or practice into the study. It could also be the case that the cluster represents a series of repeated measures on a single participant. Adding another participant to a study is likely more costly than taking another measurement on the same individual.

Given a certain budget B, we can optimize the study design based on c_1/c_2, by finding the number of clusters G and number of observations per cluster R that minimizes the variance of the treatment effect estimate.

In this simulation study, we aim to evaluate how varying R and G affects estimates of treatment effect β_1 using a hierarchical model. Furthermore, we investigate how the data generation parameters (β_1, cluster level variation, individual level variation) and c_1 and c_2 affect the optimal study design given a certain budget. For these first analyses, the outcome Y_{ij} will follow a normal distribution, and then we extend our simulation to evaluate results when Y_{ij} follows a poisson distribution.

Methods

Aims

The aims of our simulation study are as follows: 1. Evaluate how changing the number of clusters G and number of observations in a cluster R affects the estimates of the average treatment effect β in a cluster RCT when the outcome $Y_i j$ follows a normal distribution. 2. Evaluate how changing the parameters used to generate data, (treatment effect and ratio of cluster level variance to individual level variance) affect the variance of $\hat{\beta}$, and how varying the cost of adding an observation in a new cluster c_1, and the cost of adding an observation to an existing cluster c_2, affect the optimal choice of G and R under budget B constraints. 3. Extend our previous analyses to a scenario where Y_{ij} follows a poisson distribution.

Data Generation and Estimands

In our simulation we generate data from a hierarchical model in two different settings. The first is when the observed outcome, Y_{ij} follows a normal distribution.

For each observation j ($j = 1, ..., R$) in cluster i ($i = 1, ..., G$), let X_i be a binary variable indicating if cluster i is assigned to the treatment group ($X_i = 1$) or control group ($X_i = 0$). We randomly allocate half of the clusters to be in the control and half to be in the treated group. The hierarchical model of Y_{ij} is the following:

$\mu_{i0} = \alpha + \beta X_i$, which represents the fixed effect of treatment. This results in $\mu_{i0} = \alpha + \beta$ for the treatment group and $\mu_{i0} = \alpha$ for the control group.

$\mu_i|\epsilon_i = \mu_{i0} + \epsilon_i$ with $\epsilon_i \sim N(0, \gamma^2)$, so $u_i \sim N(\mu_{i0}, \gamma^2)$

$Y_{ij}|\mu_i = \mu_i + e_{ij}$ with $e_{ij} \sim N(0, \sigma^2)$, so $Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2)$.

For the poisson scenario, we have

$log(u_i) \sim N(\alpha + \beta X_i, \gamma^2)$

and $Y_{ij}|\mu_i \sim Poisson(\mu_i)$.

Our estimand of interest in both settings is $\hat{\beta}$, which represents the average treatment effect. We look at two settings for β, one with $\beta = .05$ and one with $\beta = 4$. We keep α constant at $\alpha = 5$.

Instead of varying both γ and σ, we just varied γ and held σ constant, which varies the intracluster correlation coefficient (ICC) : $\rho = \gamma^2/(\sigma^2 + \gamma^2)$, which describes how the data within a cluster are related. When $\rho =$ 1, it represents a scenario where all observations in a cluster look the same, so the sample size is just the number of clusters. When ρ is near 0, that means that the variance within the cluster is much greater than the variance between the clusters (Killip, 2014). A setting where ρ is high could be when the cluster is a series of repeated measures on a single individual. We would expect measurements within the cluster to be more similar to each other than measurements between the clusters. A setting where ρ is low could be when the clusters are schools with similar teaching methods and socioeconomic conditions, and the outcome is a standardized test score. We would expect that the variance between individuals would be higher than the variance between schools. We tested a scenario with ICC =.99, ($\gamma = 5, \sigma = 0.5$), a scenario with ICC = 0.01 ($\gamma = 0.05, \sigma = 0.5$), and ICC = 0.0001 ($\gamma = 0.005, \sigma = 0.5$) for the normal distribution setting. For the poisson setting, we tested $\gamma = 0.05$ and $\gamma = 1$.

The budget B that we use in all of our simulations is $19,000,000, which is the median cost of a clinical trial between 2015-2016 (Moore, 2018). We tested three different scenarios for c_1 and c_2, representing settings where c_1 is not much greater than c_2 and then increasing the difference between c_1 and c_2. We held c_1 constant at $40000, and tested values of $c_2 \in \{18000, 10000, 500\}$.

If we know the budget, c_1, c_2, and G, we can calculate R based on the following formula:

$$R = (B - (c_1 * G))/(c_2 * G) + 1$$

Therefore, we can vary G, and then calculate R, assuming that we want to use our entire budget. We tested values of G between 2 and 80, only considering even values of G.

We simulated results for each combination of c_2 and γ, setting $\beta = 4$. We also tested $\beta \in (.05, 4)$, varying c_2 the same as above and setting $\gamma = 0.5$ and $\sigma = 0.5$ for the normal distribution setting. We did not test different β for the poisson setting.

For the settings where the outcome is normally distributed, we have 300 repetitions of the simulation when varying γ, and when the outcome is poisson, we have 100 repetitions. For the simulations varying β, we use 30 simulations. A larger number of simulations for each measure would have been preferable, but due to time constraints, we selected small n_sim. All data generation and subsequent analyses were implemented in R version 4.4.1. The input seed is "2431".

Data Analysis Methods

We analyzed our generated data with linear mixed models and generalized linear mixed models using the lme4 R package (Bates et al., 2015). We set X as a fixed effect and specified a random intercept for cluster. If lmer or glmer had convergence issues, we set $\hat{\beta}$ = NA.

Performance Measures

Given that we do not expect the estimates of β to be biased since the model is correctly specified, we will just focus on assessing empirical standard errors, defined as $\sqrt{Var(\hat{\beta})}$.

Results

Table 1: Simulation Results for Normal Distribution

Gamma	C2	Optimal G	R	Min Empirical SE	G/R
0.005	c2 = 18000	54	18	0.0286097	3.0000000
0.050	c2 = 18000	54	18	0.0328372	3.0000000
5.000	c2 = 18000	74	13	1.0814930	5.6923077
0.005	c2 = 10000	18	102	0.0221914	0.1764706
0.050	c2 = 10000	64	26	0.0249839	2.4615385
5.000	c2 = 10000	80	20	1.0599871	4.0000000
0.005	c2 = 500	66	496	0.0051701	0.1330645
0.050	c2 = 500	80	396	0.0113733	0.2020202
5.000	c2 = 500	70	463	1.0929631	0.1511879

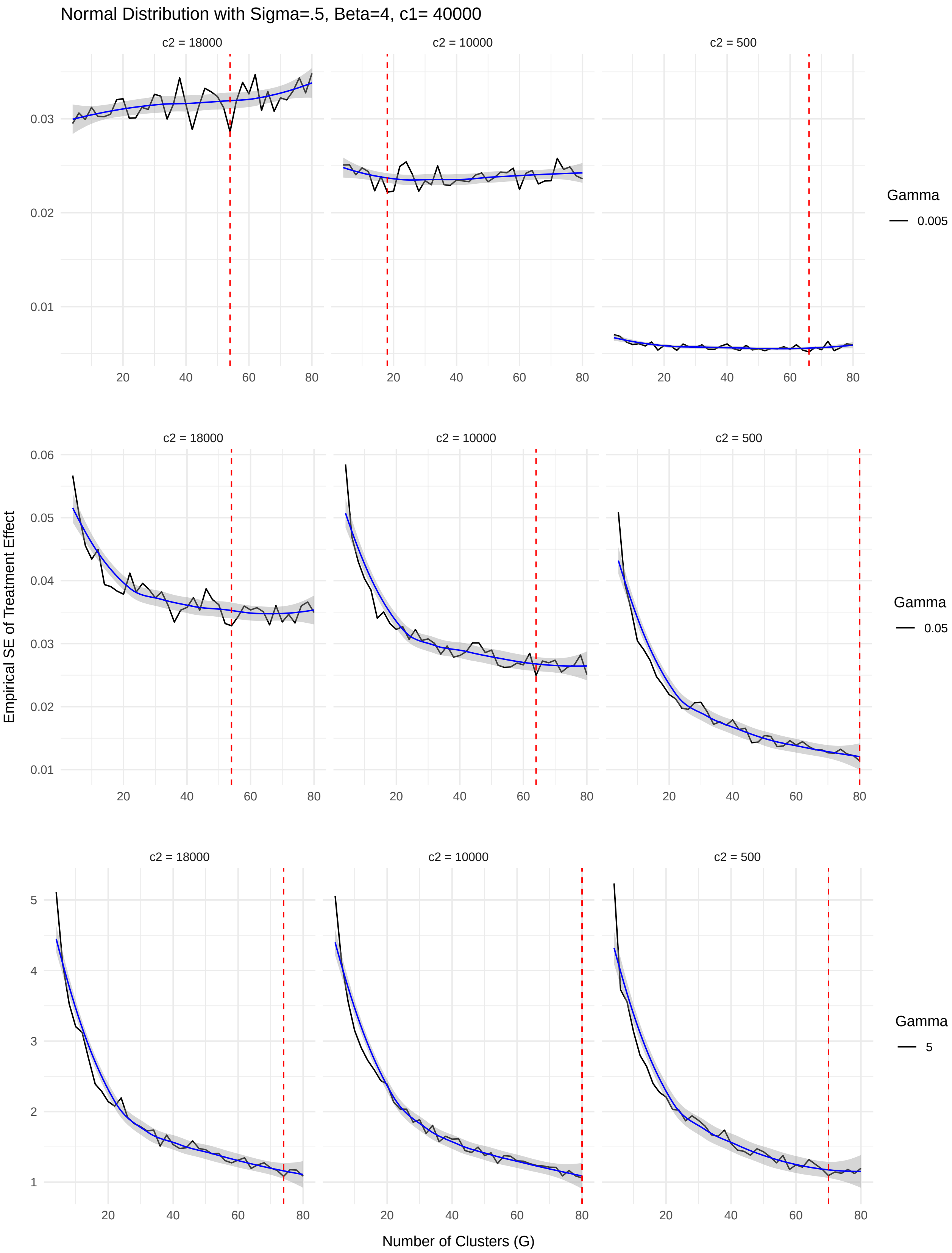

Figure 1: Empirical Standard Error of Estimates For Different Gamma and C2-Normally Distributed Outcome

Figure 1 displays the empirical SE of the treatment effect estimate on the y-axis, with increasing G on the x-axis. The outcome in this generated data is distributed normally. Given our budget of \$19000000 and c_1 of \$40000, we get the empirical SE of $\hat{\beta}$ for $\gamma = .005$, $.05$, and 5 with $\sigma = 0.5$. The rows represent different values of γ, while the columns are different values of c_2 (18000, 10000, 500). The simulation was only repeated 300 times per value of G tested and setting, which led to some volatility in the variance of $\hat{\beta}$. To visualize the trends more clearly, we added a loess fit line. The red dashed lines show the value of G that produces the lowest empirical standard error. Table 1 displays these values and the associated R.

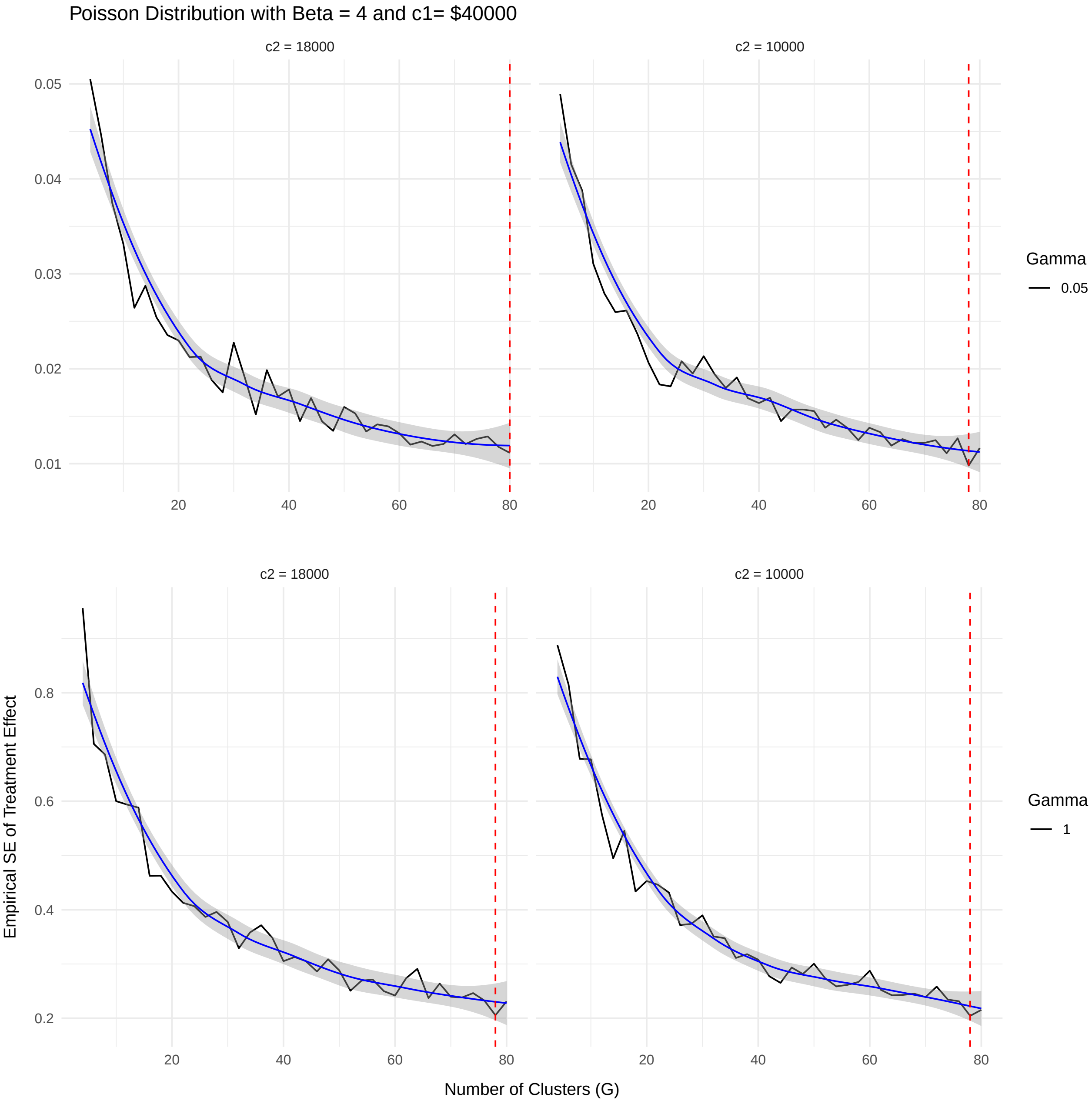

Figure 2: Empirical Standard Error of Estimates For Different Gamma and C2-Poisson Distributed Outcome

Table 2: Simulation Results for Poisson Distribution

Gamma	C2	Optimal G	R	Min Empirical SE	G/R
0.05	c2 = 18000	80	11	0.0111397	7.272727
1.00	c2 = 18000	78	12	0.2056270	6.500000
0.05	c2 = 10000	78	21	0.0098042	3.714286
1.00	c2 = 10000	78	21	0.2044399	3.714286

Figure 2 shows the empirical SE of $\hat{\beta}$ on the y-axis, and increasing G on the x-axis for the poisson setting. Here, we also have a budget of \$19000000 and c_1 for \$40000. We test $\gamma = .005$, .05, and 5 with $\sigma = 0.5$ We tested values of c_2: (18000, 10000). The simulation was repeated 100 times per value of G tested and setting of γ. The red dashed lines show the value of G that produces the lowest empirical standard error. Table 2 displays these values and the associated R.

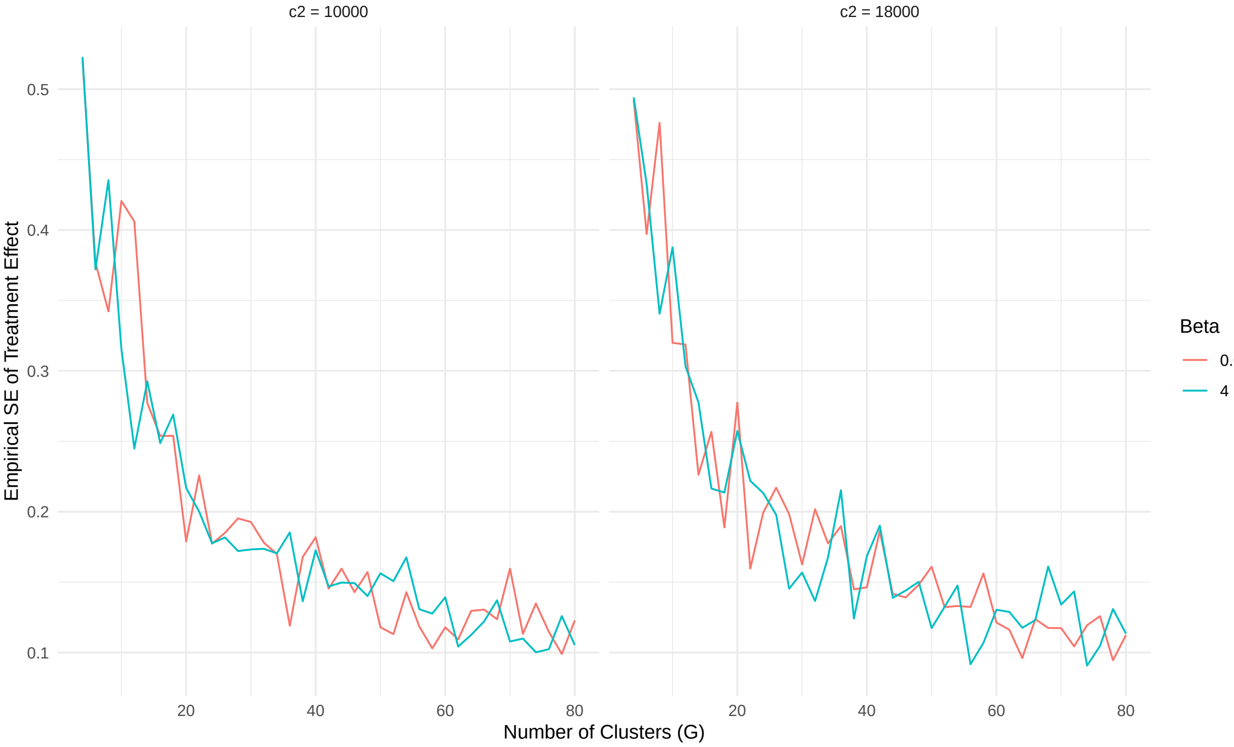

Figure 3: Empirical SE of Treatment Effect Under Different Beta by Number of Clusters

Figure 3 shows the empirical SE as G increases for $\beta = 0.05$ and 4. The budget was \$19000000, $c1$ was \$40000, and $\sigma = 0.5$. The outcome is normally distributed and the simulation was ran 30 times per value of G and setting tested.

Discussion and Conclusion

The results shown in Figure 1 show that as γ increases while σ is held constant, we see an overall increase in empirical standard error of β. This makes sense since as the overall variance increases, the estimates for β will vary more. When γ is 5, we notice that as the number of clusters increases, the empirical standard error decreases, suggesting that the optimal study design has a high number of clusters. Since we use our entire budget in each scenario, as the number of clusters increases, the number of observations per cluster decreases. When the variation between the clusters is high compared to the variation within the clusters, we want to maximize the number of clusters in the study design to reduce the variance of the treatment effect estimate. We do not see a large difference in this trend between different c_1/c_2 in the plots for $\gamma = 5$. However, in table 1. we can see that in the scenario where c_2 is very cheap relative to c_1, not as many clusters are needed since we can afford to have such a large amount of observations in each cluster. However, further simulations should be run to confirm this.

When $\gamma = 0.05$, the variance between the clusters is less than the variance within the clusters, so we would expect that the optimal study design has a smaller number of clusters than for $\gamma = 5$. We notice that the downwards trend shown for $\gamma = 5$ is reduced for $\gamma = 0.05$, which matches our expectations. When $c_2 = 500$, the downwards trend is the steepest, compared to the others with the same γ and higher c_2. Table 1 shows that for cheap c_2, the ratio of G/R is the lowest.

When $\gamma = .005$, the variance is increasing with G or stable, which is the opposite trend as for larger γ. When γ is low compared to σ, the variation between clusters is low, so it makes sense to maximize the number of observations per cluster. In Table 1, we see that the ratio G/R generally increases as γ increases.

In Figure 2, we see the results of the the simulations with the poisson distribution. Here, the setting of γ does not affect the optimal combinations of G and R in the way it does for the normal distribution shown in Figure 1. This may just be due to not testing large enough values of G, since G was near the maximum number of clusters tested for each setting. Future simulations should consider larger G.

We did not see any significant difference in empirical SE between the values of β tested (.05 and 4), across any of the values for c_2 (Figure 3). Due to this, we decided not to test values of β for the poisson distributed outcome.

A limitation of our study is that we only tested combinations that included values of G between 2 and 80. Studies where a cluster represents repeated measures on a person may want to include many more clusters than this. Additionally, the tested values for c_1 and c_2 may not be realistic or applicable to all research settings. We found study results suggesting that the cost per participant in a clinical trial has a median of $41,117, but we were unsure about how costs would vary in a cluster randomized trial, and how much it might cost to initiate a new cluster. Future research could apply our methods to other possible values of c_1,c_2, budgets and treatment effect estimates. Additionally, we tested rather extreme settings for the ICC, to clearly demonstrate how the best combination of G and R was affected by changes in ICC. Furthermore, we did not perform many replications of the simulation due to time constraints, although we can still identify the trends in the results. In a future study, we could identify the number of simulations that achieves an acceptable Monte Carlo SE.

Overall, our results show that when the between cluster variance is relatively large, the optimal number of clusters is high, and when the between cluster variance is relatively small, the optimal number of clusters is low. Choice of G and R also depends on the relative costs of adding an observation to a new cluster versus to an existing cluster. If c_1 is relatively large, then the optimal number of clusters to include may be less than it would be for a smaller c_1.

References

Bates D, Maechler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. BMJ. (1998) Oct 31;317(7167):1171-2. doi: 10.1136/bmj.317.7167.1171. PMID: 9794847; PMCID: PMC1114151.

Dron L, Taljaard M, Cheung YB, Grais R, Ford N, Thorlund K, … Mills EJ (2021). The role and challenges of cluster randomised trials for global health. The Lancet Global Health, 9(5), e701–e710. doi: 10.1016/S2214-109X(20)30541-6

Eldridge S, Kerry S. Wiley; Chichester (2012). A practical guide to cluster randomized trials in health services research

Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. BJOG. (2018) Dec;125(13):1716. doi: 10.1111/1471-0528.15199. Epub 2018 Jun 19. PMID: 29916205; PMCID: PMC6235704.

Heagerty PD, ER. (2017). Cluster Randomized Trials Designing with Implementation and Dissemination in Mind: Introduction. In: Rethinking Clinical Trials: A Living TextBook of Pragmatic Clinical Trials. Retrieved from https://rethinkingclinicaltrials.org/chapters/design/experimental-designs-randomization-schemes-top/cluster-randomized-trials/

Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Ann Fam Med. 2004 May-Jun;2(3):204-8. doi: 10.1370/afm.141. PMID: 15209195; PMCID: PMC1466680.

Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. JAMA Intern Med. (2018) Nov 1;178(11):1451-1457. doi: 10.1001/jamainternmed.2018.3931. PMID: 30264133; PMCID: PMC6248200.

Mullard, A. How much do phase III trials cost?. Nat Rev Drug Discov 17, 777 (2018). https://doi.org/10.1038/nrd.2018.198

Code Appendix

```r
knitr::opts_chunk$set(echo=FALSE, message = FALSE, warning = FALSE)

library(tidyverse)
library(kableExtra)
library(lme4)
library(knitr)
library(gridExtra)
generate_data <- function(g,r,beta, gamma, sigma, distribution,
                          output_file = "simultated_data.csv") {
#' Simulate data based on given g,r,beta,gamma,sigma, and distribution
#'
#' @param g Integer. Number of clusters
#' @param r Integer. Number of observations in a cluster
#' @param beta Numeric. Treatment effect
#' @param gamma Numeric. Cluster specific sd
#' @param sigma Numeric. Participant specific sd
#' @param distribution String. Poisson or Normal
#' @param output_file Name for output file
#' @return Dataframe with x and y, g, and r
#'

  #initialize df to store generated data
  data <- data.frame(group= NA, x = NA, y=NA, r=NA)

  #assign treatment
    treated <- sample(seq(1,g,1), g/2, replace=FALSE) #assign treatment
    assignment <- rep(0, g)
    assignment[treated] <- 1

  #for each cluster, assign a treatment and generate y
  for(i in 1:g) {
    x <- assignment[i]
    if(distribution == "normal"){
        u_i <- rnorm(1, 5 + beta*x, gamma)
        y <- rnorm(r, u_i, sigma)
      }
    else{
      logu_i <- rnorm(1, 5 + beta*x, gamma)
      y <- rpois(r, exp(logu_i))
    }

    #create results df
    df <- data.frame(group = i,
                     r = 1:r,
                     x = x,
                     y = y)
    data <- rbind(data, df)
  }

    # save data to a CSV file
    write.csv(na.omit(data), file = output_file, row.names = FALSE)

    return(na.omit(data))
  }
#given a budget B, c1,c2 we want to calcualte r
calculate_R <- function(B, c1, c2, num_clusters) {
#' Calculate total cost given c1,c2,number of clusters and observations per cluster
#'
#' @param B Numeric. Total Budget
#' @param c1 Numeric. Cost of initial participant in new cluster
#' @param c2 Numeric. Cost of participant in existing cluster
#' @param num_clusters Integer. Number of clusters
#' @return cost


  num_per_cluster <- (B- (c1*num_clusters))/(c2*num_clusters) + 1

  return(floor(num_per_cluster))
}
run_sim <- function(g,r,beta,gamma,sigma,distribution,c1,c2,
                    results_file = "sim_results.csv"){
#' Analyzes a dataset with a mixed effects model, gets estimates for beta,
#' calculates cost, and save results to a csv file
#' @param g Integer. Number of clusters
#' @param r Integer. Number of observations in a cluster
#' @param beta Numeric. Treatment effect
#' @param gamma Numeric. Cluster specific sd
#' @param sigma Numeric. Participant specific sd
#' @param distribution String. Poisson or Normal
#' @param c1 Numeric. Cost of initial participant in new cluster
#' @param c2 Numeric. Cost of participant in existing cluster
#' @param results_file Name for output file
#' @return Dataframe including estimates for beta

  #generate data
  data <- generate_data(g,r,beta,gamma,sigma,distribution)
  #use glmer to get coefficients
  if (distribution == "normal"){
    summary <- summary(lmer(y~x+(1|group),data = data))
    beta_hat <- tryCatch(summary[["coefficients"]][2,1], error=function(err) NA)
  }else{
    summary <- summary(glmer(y~x+(1|group),data = data, family= "poisson"))
    beta_hat <- tryCatch(summary[["coefficients"]][2,1], error=function(err) NA)
  }

  #create dataframe for results
  result <- data.frame(beta_hat, beta, gamma, sigma, distribution, g, r)

  # save results to a file
  if (file.exists(results_file)) {
    write.table(result, file = results_file, append = TRUE,
                sep = ",", col.names = FALSE, row.names = FALSE)
  } else {
    write.csv(result, file = results_file, row.names = FALSE)
  }

  return(result)
}
run_experiments_varying_g <- function(B,c1,c2, values_for_num_clusters,beta,gamma,
                                      sigma,distribution,reps=100, file_name_additions=NA) {
#' Runs simulations for varying values of g and r
#'
#' @param B Numeric. Total Budget
#' @param c1 Numeric. Cost of initial participant in new cluster
#' @param c2 Numeric. Cost of participant in existing cluster
#' @param values_for_num_clusters. Values to test for g
#' @param beta Numeric. Treatment effect
#' @param gamma Numeric. Cluster specific sd
#' @param sigma Numeric. Participant specific sd
#' @param distribution String. Poisson or Normal
#' @param reps Integer. Number of repetitions of simulation
#' @return

  #name results file with the experiment specifics
  results_file <- paste("sim_results", B,c1,c2,beta,gamma,sigma,distribution,reps,
                        file_name_additions,
                        ".csv", sep = "_")

  #if the file already exists, replace it
  if (file.exists(results_file)) file.remove(results_file)

  #for each value of g, run the experiment "reps" times
    for (g in values_for_num_clusters) {

      #get the value of r based on budget
      r <- calculate_R(B, c1, c2, g)

      for (rep in 1:reps) {
        data <- generate_data(g,r,beta, gamma, sigma, distribution)
        run_sim(g,r,beta,gamma,sigma,distribution,c1,c2,results_file)
    }
  }
}

summarize_results_varying_g <- function(results_file) {
#' Summarizes simulation results varying g
#'
#' @param results_file csv file containing the results to summarize
#' @return list with summary of results for each combination of g and r and plot
  results <- read.csv(results_file)

  #for each g, get the lowest variance available

  #summarize results file by getting the empirical SE of beta_hat across the simulations
  #for each r and g along with the cost
  summary <- results %>%
    group_by(g,r) %>%
    summarize(
      empSE = sqrt(var(beta_hat, na.rm=TRUE)),
      gamma = mean(gamma),
      beta = mean(beta),
      c1 = mean(c1),
      c2 = mean(c2)
    )

  # plot results
  p <- ggplot(summary, aes(x = g, y = empSE)) +
    geom_line() +
    labs(title = paste("Simulation Results") ,
         x = "G",
         y = "Empirical SE") +
    theme_minimal()

  return(summary)
}
##set seed
set.seed(2431)

#values of c1 and c2 to test
c_1_test <- 40000
c_2_test <- c(18000,  10000, 500)

#values of gamma to test
gamma_to_test <- c(.005, .05,5)

#initialize summary df
#summary <- data.frame(g=NA,r=NA,empSE=NA,gamma=NA, beta=NA, c1 = NA, c2 = NA)

#for each possible c2:
for (i in 1:3){
  c1 <- c_1_test
  c2 <- c_2_test[i]

  for (gam in gamma_to_test){ #run the experiment for each gamma
   # run_experiments_varying_g(19000000,c1,c2,values_for_num_clusters=seq(4,80,2),
    #4,gam,.5,"normal",reps=300, file_name_additions=NA)

    #get the file name
    file_name = paste("sim_results_1.9e+07_",c1,"_",c2,"_4_",gam,"_0.5_normal_300_NA_.csv", sep="")
    #get the results summary for this scenario
    res <- summarize_results_varying_g(file_name)
    #combine the results of all simulation settings into one df
    summary <- rbind(summary, res)
  }
}

#write the summary df to a csv file
#write.csv(summary, file = "summary.csv", row.names = FALSE)

#poisson
#values of c1 and c2 to test
c_1_test <- 40000
c_2_test <- c(18000,  10000)

#values of gamma to test
gamma_to_test <- c(.05,1)

#initialize summary df
summary_poisson <- data.frame(g=NA,r=NA,empSE=NA,gamma=NA, beta=NA, c1 = NA, c2 = NA)

#for each possible c2:
for (i in 1:2){
  c1 <- c_1_test
  c2 <- c_2_test[i]

  for (gam in gamma_to_test){ #run the experiment for each gamma
   # run_experiments_varying_g(19000000,c1,c2,
    #values_for_num_clusters=seq(4,80,2),4,gam,.5,"poisson",reps=100, file_name_additions=NA)

    #get the file name
    file_name = paste("sim_results_1.9e+07_",c1,"_",c2,"_4_",gam,"_0.5_poisson_100_NA_.csv", sep="")
    #get the results summary for this scenario
    res <- summarize_results_varying_g(file_name)
    #combine the results of all simulation settings into one df
    summary_poisson <- rbind(summary_poisson, res)
  }
}

#write the summary df to a csv file
#write.csv(summary_poisson, file = "summary_poisson.csv", row.names = FALSE)

#initialize df to vary beta
summary_beta <- data.frame(g=NA,r=NA,empSE=NA,gamma=NA,beta=NA, c1 = NA, c2 = NA)
```

```
#different beta values to test
beta_to_test <- c(.05,4)

#for each setting of c2
for(i in 1:2) {
    c1 <- c_1_test
    c2 <- c_2_test[i]

  for (beta in beta_to_test){ #run experiment for each setting of beta
    #run_experiments_varying_g(19000000,c1,c2,    values_for_num_clusters=seq(4,80,2),
                              #beta,.5,.5,"normal",reps=30, file_name_additions=NA)

    #get file name
    file_name = paste("sim_results_1.9e+07_",c1,"_",c2,"_",beta,"_0.5_0.5_normal_30_NA_.csv", sep="")

    #get summary table for this setting
    res <- summarize_results_varying_g(file_name)

    #combine summary dfs from each setting
    summary_beta <- rbind(summary_beta, res)
  }
}

#write the summary df to a csv file
#write.csv(summary_beta, file = "summary_beta.csv", row.names = FALSE)
#create plot for different settings of gamma and c2

#read data
summary <- read.csv("summary.csv")

summary <- summary %>% mutate(c2 = case_when(c2 == 18000 ~ "c2 = 18000",
                        c2 == 10000 ~ "c2 = 10000",
                        c2 == 500 ~ "c2 = 500"))

summary$c2 <- factor(summary$c2, levels = c("c2 = 18000", "c2 = 10000", "c2 = 500"))

min_se_data <- summary %>%
  na.omit() %>%
  group_by(gamma, c2) %>%
  summarize(g_min = g[which.min(empSE)], r = r[which.min(empSE)], empSE_min = min(empSE))

p1 <- ggplot(na.omit(summary[summary$gamma == 0.005, ]), aes(x = g, y = empSE,
                                                             color = as.factor(gamma))) +
  geom_line() +  geom_smooth(color="blue", size=.5) +
  scale_color_manual(values = c("black", "black", "black")) +
  geom_vline(data = min_se_data %>% filter(gamma == 0.005),
             aes(xintercept = g_min), linetype = "dashed", color = "red") +
  facet_wrap(~c2) +
  theme_minimal() +
  labs(x = "", y = "", color = "Gamma") +
    theme(
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  ) +
  ggtitle("Normal Distribution with Sigma=.5, Beta=4, c1= 40000")


p2 <- ggplot(na.omit(summary[summary$gamma == .05, ]), aes(x = g, y = empSE,
                                                          color = as.factor(gamma))) +
  geom_line() +   geom_smooth(color="blue", size=.5) +
  scale_color_manual(values = c("black", "black", "black")) +
  geom_vline(data = min_se_data %>% filter(gamma == .05), aes(xintercept = g_min),
             linetype = "dashed", color = "red") +
  facet_wrap(~c2) +
  theme_minimal() +
  labs(x = "", y = "Empirical SE of Treatment Effect", color = "Gamma") +
    theme(
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  )


p3 <- ggplot(na.omit(summary[summary$gamma == 5, ]), aes(x = g, y = empSE,
                                                        color = as.factor(gamma))) +
  geom_line() + geom_smooth(color="blue", size=.5) +
  scale_color_manual(values = c("black", "black", "black")) +
  geom_vline(data = min_se_data %>% filter(gamma == 5), aes(xintercept = g_min),
             linetype = "dashed", color = "red") +
  facet_wrap(~c2) +
  theme_minimal() +
  labs(x = "Number of Clusters (G)", y = "", color = "Gamma") +
    theme(
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  )


# Arrange the plots in a grid
grid.arrange(p1, p2, p3)

#create new column for g/r
min_se_data$ratio <- min_se_data$g_min/min_se_data$r

min_se_data <- min_se_data %>% arrange(c2)

colnames(min_se_data) <- c("Gamma", "C2", "Optimal G", "R", "Min Empirical SE", "G/R")
kable(min_se_data, caption = "Simulation Results for Normal Distribution")
#create plot for different settings of gamma and c2

#read data
summary <- read.csv("summary_poisson.csv")


summary <- summary %>% mutate(c2 = case_when(c2 == 18000 ~ "c2 = 18000",
                        c2 == 10000 ~ "c2 = 10000",
                        c2 == 500 ~ "c2 = 500"))
```

```
summary$c2 <- factor(summary$c2, levels = c("c2 = 18000", "c2 = 10000", "c2 = 500"))


min_se_data <- summary %>%
  na.omit() %>%
  group_by(gamma, c2) %>%
  summarize(g_min = g[which.min(empSE)], r = r[which.min(empSE)], empSE_min = min(empSE))


p1 <- ggplot(na.omit(summary[summary$gamma == .05,]), aes(x = g, y = empSE,
                                                          color = as.factor(gamma))) +
  geom_line() +  geom_smooth(color="blue", size=.5) +
  scale_color_manual(values = c("black", "black", "black")) +
  geom_vline(data = min_se_data %>% filter(gamma == .05), aes(xintercept = g_min),
             linetype = "dashed", color = "red") +
  facet_wrap(~c2) +
  theme_minimal() +
  labs(x = "", y = "", color = "Gamma") +
    theme(
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  ) + ggtitle("Poisson Distribution with Beta = 4 and c1= $40000")


p2 <- ggplot(na.omit(summary[summary$gamma == 1,]), aes(x = g, y = empSE,
                                                        color = as.factor(gamma))) +
  geom_line() + geom_smooth(color="blue", size=.5) +
  scale_color_manual(values = c("black", "black", "black")) +
  geom_vline(data = min_se_data %>% filter(gamma == 1), aes(xintercept = g_min),
             linetype = "dashed", color = "red") +
  facet_wrap(~c2) +
  theme_minimal() +
  labs(x = "Number of Clusters (G)", y = "Empirical SE of Treatment Effect",
       color = "Gamma") +
    theme(
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  )


# Arrange the plots in a grid
grid.arrange(p1, p2)
min_se_data$ratio <- min_se_data$g_min/min_se_data$r

min_se_data <- min_se_data %>% arrange(c2)

colnames(min_se_data) <- c("Gamma", "C2", "Optimal G", "R", "Min Empirical SE", "G/R")

kable(min_se_data, caption = "Simulation Results for Poisson Distribution")
summary_beta <- read.csv("summary_beta.csv")

summary_beta <- summary_beta %>% mutate(c2 = case_when(c2 == 18000 ~ "c2 = 18000",
                        c2 == 10000 ~ "c2 = 10000",
                        c2 == 500 ~ "c2 = 50"))

summary$c2 <- factor(summary$c2, levels = c("c2 = 18000", "c2 = 10000", "c2 = 500"))

ggplot(na.omit(summary_beta), aes(x=g,y=empSE,color=as.factor(beta))) +
  geom_line() + facet_wrap(~as.factor(c2)) + theme_minimal() +
  labs(x = "Number of Clusters (G)", y = "Empirical SE of Treatment Effect", color = "Beta")
```