Optimizing Study Design: Balancing Clusters and Measurements for Efficient Estimation in Normal and Poisson Distributions

Zhaoxiang Ding

December 6, 2024

Summary

This study examines the factors influencing the standard error of the estimator under normal and Poisson distributions, focusing on the interplay between the number of clusters, measurements per cluster, and resource constraints. Results show that the optimal number of clusters depends on the cost ratio and variance components, with higher cost ratios and outcome variances requiring fewer clusters. For the normal distribution, the standard error generally decreases as the number of clusters increases, aligning with theoretical expectations. However, for the Poisson distribution, while the standard error follows a similar trend, its sensitivity to parameters such as α and γ deviates from theory. Specifically, $\gamma = 1$ unexpectedly resulted in fewer optimal clusters compared to other values, and α, despite its theoretical link to variance, had no discernible effect in simulations. These discrepancies suggest that additional simulations may be necessary to fully understand the underlying dynamics. Across all settings, the standard error decreases as the budget increases, as higher budgets allow for more data collection.

Introduction

In studies where the primary interest lies in estimating the population mean of certain features, challenges often arise in determining how to collect the data effectively. Consider a study aimed at estimating the treatment effect in a population. Measurement error, denoted as ϵ, is assumed to follow a normal distribution with mean 0 and variance γ^2 for each measurement. Additionally, the outcome of each subject follows a normal distribution with mean μ_i and variance σ^2, where $\mu_i \sim N(\alpha + \beta X_i, \gamma^2)$. Consequently, the outcome for a subject can be expressed as:

$$
\begin{aligned}
\mu_{i0} &= \alpha + \beta X_i \\
\mu_i &\sim N(\mu_{i0}, \gamma^2) \\
Y_{ij}|\mu_i &\sim N(\mu_i, \sigma^2) \\
Y_{ij} &\sim N(\mu_{i0}, \sigma^2 + \gamma^2)
\end{aligned}
$$

When $\sigma^2 \gg \gamma^2$ —indicating that the measurement error is significantly smaller than the variance of the outcome—the measurement error can be ignored, and each subject needs to be measured only once. However, when this condition does not hold, the measurement error becomes significant, requiring multiple measurements per subject to mitigate its impact.

An ideal study design in such cases would include as many subjects as possible, with each subject being measured multiple times. However, due to resource constraints, this approach is often impractical. Researchers must therefore decide on an optimal balance between the number of subjects sampled and the number of measurements taken per subject to achieve cost-efficient yet accurate results.

Assuming the estimator follows a linear model, the goal of the optimal study design is to minimize the standard error of the estimator under the given resource constraints while maintaining its unbiasedness within the stated hypothesis. This research question can be generalized as follows: Given resource constraints (B), the cost of measuring a new cluster (c_1), and the cost of taking an additional measurement within a cluster (c_2), how many clusters (G) and how many samples per cluster (R) should be selected to minimize the estimator's variance?

This project aims to investigate the optimal combination of G and R that minimizes the variance of the estimator under different data generation mechanisms and c_1/c_2 ratios. Additionally, we will extend our analysis to cases where the outcome follows a Poisson distribution:$Y_{ij}|\mu_i \sim \text{Poisson}(\exp(\mu_i))$, examining the differences between the normal and Poisson cases.

Simulation

Aims: To investigate the optimal combination of G (number of clusters) and R (number of measurements per cluster) that minimizes the standard error of the estimator under different data generation mechanisms and c_1/c_2 ratios.

Data Generation Mechanisms: As in the previous section, the data follows a hierarchical model where the treatment is binary, with half of the clusters assigned a treatment value of 1

and the other half assigned a treatment value of 0. The model is defined as:

$$\begin{aligned} X_i &\in (0,1) \\ \mu_{i0} &= \alpha + \beta X_i \\ \mu_i &\sim N(\mu_{i0}, \gamma^2) \\ Y_{ij}^N|\mu_i &\sim N(\mu_i, \sigma^2) \\ Y_{ij}^P|\mu_i &\sim \text{Poisson}(\exp(\mu_i)) \end{aligned}$$

Where Y^N and Y^P are the outcome under normal and poisson distribution respectively.

Under the normal distribution, the estimator is unbiased, and its variance is influenced by G, R, B (total resources), the ratio of c_1/c_2 (cluster cost vs. measurement cost), and the ratio of σ/γ (outcome variance vs. measurement error). For the Poisson distribution, the variance of the estimator is also affected by α and β, as the mean and variance of the Poisson distribution are equal. In both cases, the variance decreases with increasing B, as larger resource allocations allow for more data points.

To explore these relationships, we consider the following parameter settings: - B: 1000, 2000, 5000. Fix $c_1 = 5$, $c_2 = 1$, $\alpha = 1$, $\beta = 1$, $\sigma = 1$, $\gamma = 1$ - β: 1, 2, 5. Fix c1 = 5, c2 = 1, $\alpha = 1$, $\sigma = 1$, $\gamma = 1$, B = 1000 - c_1/c_2: 5, 10, 50, 100. Fix $\beta = 1$, $B = 1000$, $c_2 = 1$ with varying σ (0.25, 1, 5), γ (0.25, 1, 5), and α (0, 1, 5).

Each simulation is replicated 100 times, with input seeds ranging from 1 to 100 for each replication.

Estimands: The estimator is the coefficient of the treatment effect in the model.

Methods: A grid search approach will be used to identify the optimal G and R under different parameter settings. For normal outcomes, the treatment effect is estimated using a linear mixed-effects model when $R > 1$ and a linear model when $R = 1$. For Poisson outcomes, a generalized linear mixed-effects model is used for R > 1, and a generalized linear model is used for R = 1.

Performance measures: The performance of the estimator is evaluated using the empirical standard error (SE), calculated as $\sqrt{Var(\hat{\beta})}$. The empirical SE quantifies the efficiency or precision of the estimator, providing a sufficient basis for comparing estimator performance across different settings, as the estimator is unbiased.

Result

The raw simulation results are presented in Figure 1. The estimated β values are centered around the true values of 1, 2, and 5 for both distributions, indicating that the estimator is unbiased. The variance of the estimator is influenced by the number of clusters and the number

of measurements per cluster. Moreover, the optimal allocation of clusters and measurements per cluster varies depending on the distribution of the outcome.

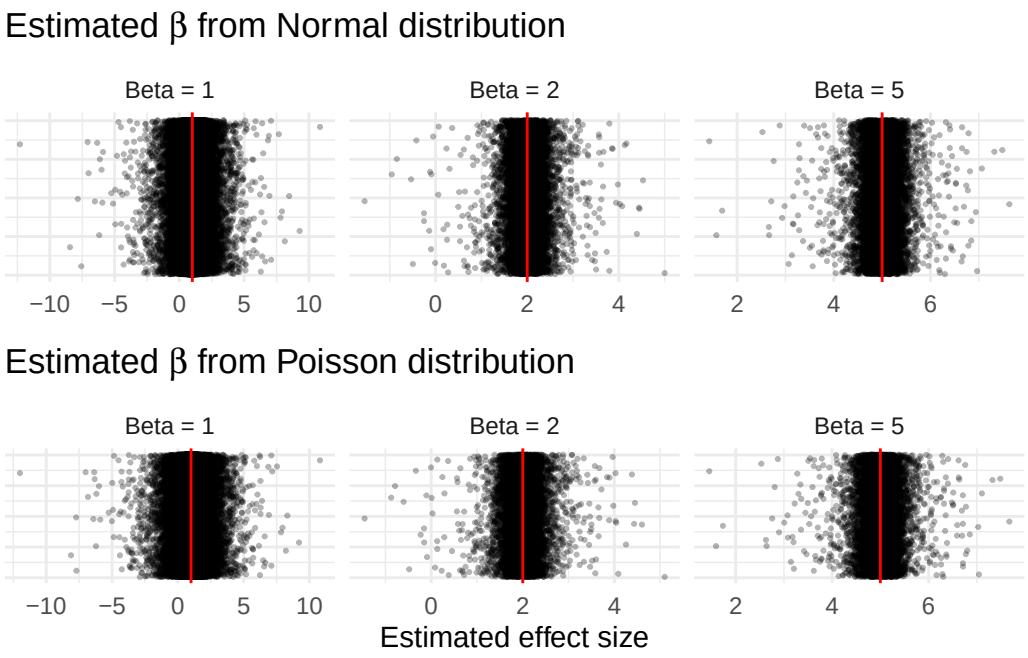

Figure 1: Distribution of estimated beta

The standard error of the estimator is presented in Figure 2. In most cases, the standard error for the Poisson distribution is comparable to that of the normal distribution. However, in some instances, the estimators derived from the Poisson distribution exhibit a higher standard error than those from the normal distribution. Notably, there are no missing values in the dataset.

Optimal number of clusters and measurements per cluster under different settings.

Figure 3 illustrates the standard error of the estimator under various numbers of clusters for the normal distribution, considering different σ values and cost ratio settings. For all settings, the standard error initially decreases as the number of clusters increases. However, in the case of $\sigma = 5$, the standard error begins to rise as the number of clusters increases further.

For settings with other σ values and cost ratios below 100, the standard error continues to decrease, albeit with diminishing returns as the number of clusters grows. When the cost ratio is 100, the standard error increases with additional clusters for $\sigma = 1$ and $\sigma = 5$. The optimal number of clusters for each setting is summarized in Table 1.

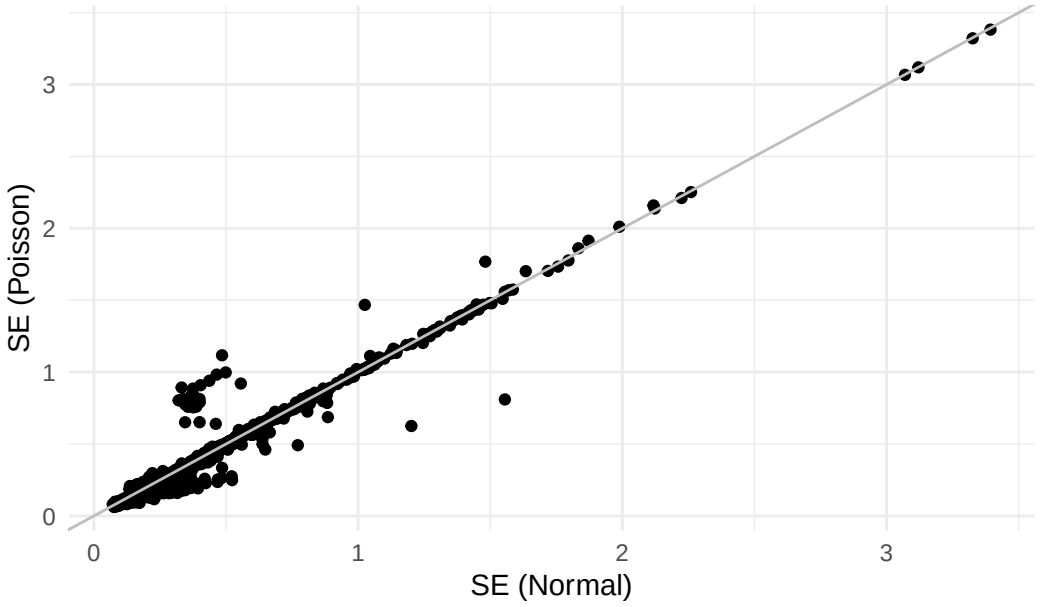

Figure 2: Comparison of the standard error of estimand between Normal and Poisson Model

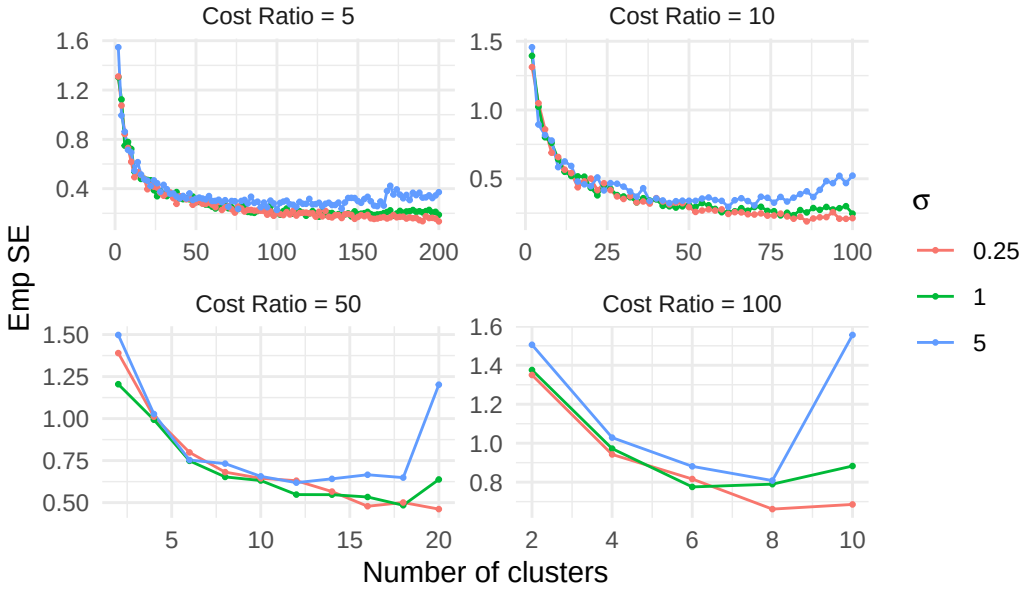

Figure 3: Estimator's performance with different number of clusters , cost ratios and σ under normal distribution. $B = 1000$, $\gamma = 1$, $\alpha = 1$, $\beta = 1$, $c_2 = 1$, Normal distribution

Table 1: Optimal number of clusters, cost ratios and σ under normal distribution, $B = 1000$, $\gamma = 1$, $\alpha = 1$, $\beta = 1$, $c_2 = 1$

Optimal number of clusters under normal distribution

Cost Ratio	Sigma = 0.25	Sigma = 1	Sigma = 5
5	200	126	100
10	86	78	62
50	20	18	12
100	8	6	8

In summary, the optimal number of clusters under the normal distribution is influenced by the cost ratio and the variance of the outcome. As both the cost ratio and the variance of the outcome increase, fewer clusters are needed to achieve the optimal standard error of the estimator. With γ fixed at 1, the results indicate that when the variance of the sampling error is higher, the optimal number of clusters decreases, implying that more measurements per cluster are necessary to achieve the desired standard error. Additionally, when the cost ratio is higher, the optimal number of clusters decreases because adding new clusters becomes less effective in reducing the standard error, as it significantly limits the total number of samples that can be collected.

The results for the empirical standard error (Emp SE) of the Poisson distribution are omitted, as σ does not affect the variance of the estimator in the Poisson distribution.

Figure 4 presents the standard error of the estimator across different numbers of clusters for both the normal and Poisson distributions, under various γ values and cost ratio settings. Unlike previous results, for all γ values when the cost ratio is below 10, the standard error decreases monotonically as the number of clusters increases in the normal distribution. However, for cost ratios of 50 and 100, the standard error initially increases before decreasing as the number of clusters grows.

For the Poisson distribution, the trend in the standard error closely resembles the pattern observed in Figure 3. The optimal number of clusters for each setting is detailed in Table 3.

The optimal number of clusters for each setting is summarized in Table 2 and Table 3. Similar to previous results, the optimal number of clusters under the Poisson distribution is influenced by the cost ratio, with higher ratios requiring fewer clusters to achieve the optimal standard error of the estimator.

However, the results indicate that when $\gamma = 1$, the optimal number of clusters is lower than when $\gamma = 0.25$ or $\gamma = 5$. This finding is inconsistent with theoretical expectations, as a higher γ should necessitate more clusters to achieve the optimal standard error. This discrepancy may be attributed to the relatively low number of simulations conducted.

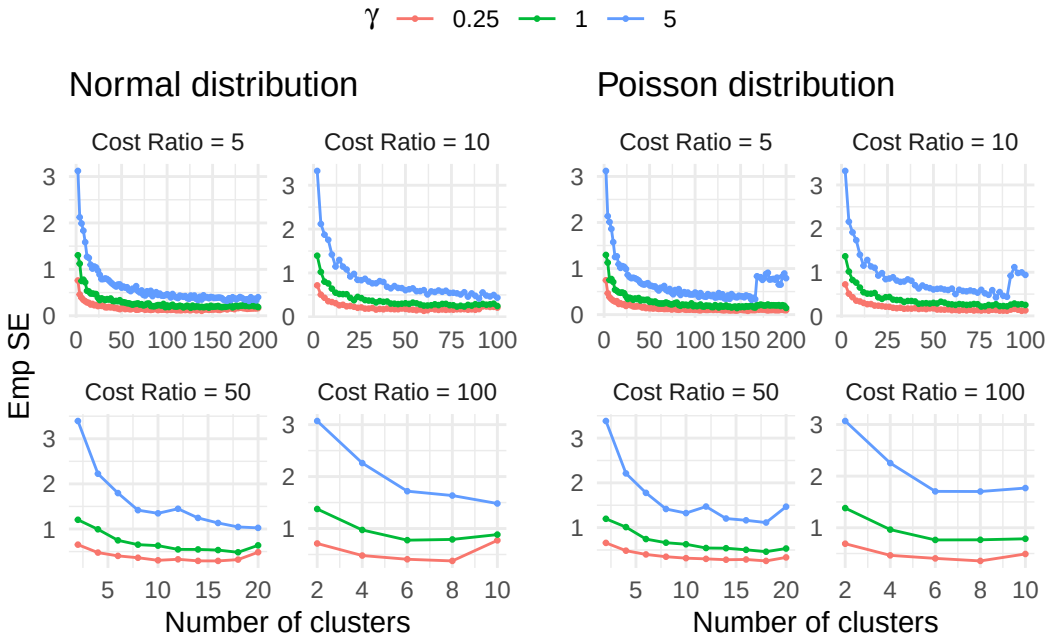

Figure 4: Estimator's performance with different number of clusters, cost ratios and γ. $B = 1000$, $\sigma = 1$, $\alpha = 1$, $\beta = 1$, $c_2 = 1$

Table 2: Optimal number of clusters under normal distribution, cost ratios and γ. $B = 1000$, $\sigma = 1$, $\alpha = 1$, $\beta = 1$, $c_2 = 1$

Cost Ratio	Gamma = 0.25	Gamma = 1	Gamma = 5
5	138	126	190
10	60	78	90
50	16	18	20
100	8	6	10

Table 3: Optimal number of clusters under poisson distribution, cost ratios and γ. $B = 1000$, $\sigma = 1$, $\alpha = 1$, $\beta = 1$, $c_2 = 1$

Cost Ratio	Gamma = 0.25	Gamma = 1	Gamma = 5
5	158	146	162

10	98	76	84
50	18	18	18
100	8	6	8

Figure 5 illustrates the standard error of the estimator across different numbers of clusters for both the normal and Poisson distributions, under various cost ratios and α settings. Both distributions show no clear differences across α values.

This outcome is expected in the normal distribution setting, as the mean of the normal distribution does not influence the variance of the estimator. However, in the Poisson distribution, where the mean is equal to the variance, a higher α would theoretically impact the variance of the estimator. Surprisingly, our simulations suggest otherwise, showing no significant effect of α on the variance.

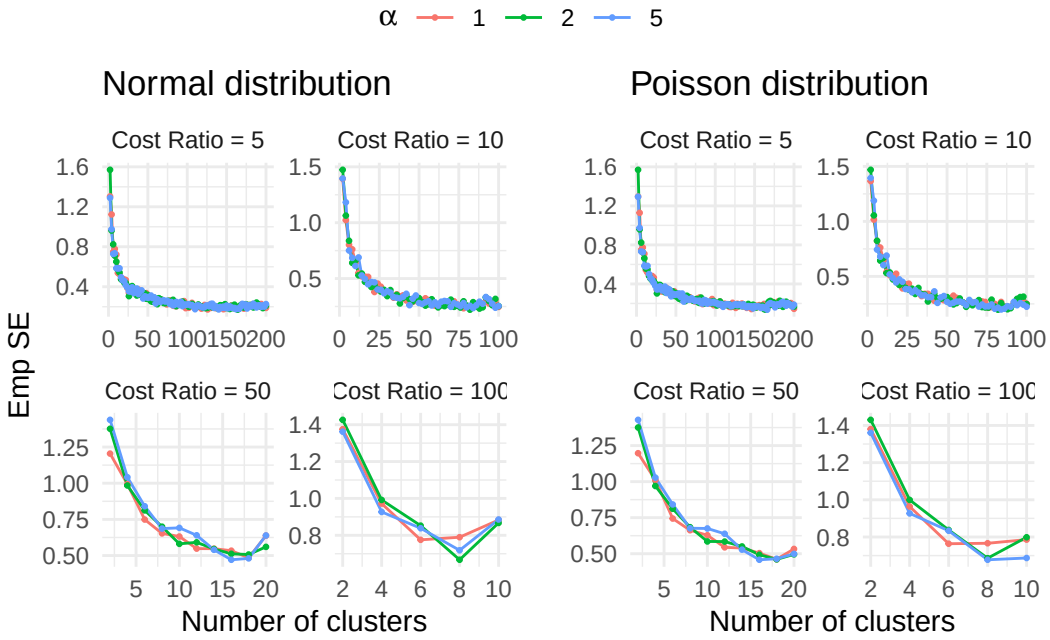

Figure 5: Estimator's performance with different number of clusters, cost ratios and α. $B = 1000$, $\sigma = 1$, $\gamma = 1$, $\beta = 1$, $c_2 = 1$

Figure 5 displays the standard error of the estimator across different numbers of clusters for both the normal and Poisson distributions under various budget settings. The standard error consistently decreases as the number of clusters increases across all budget levels. When the number of clusters reaches its maximum allowable value for a given budget, the standard error for the same number of clusters is lower under a higher budget. This result is expected, as a higher budget enables more samples to be collected.

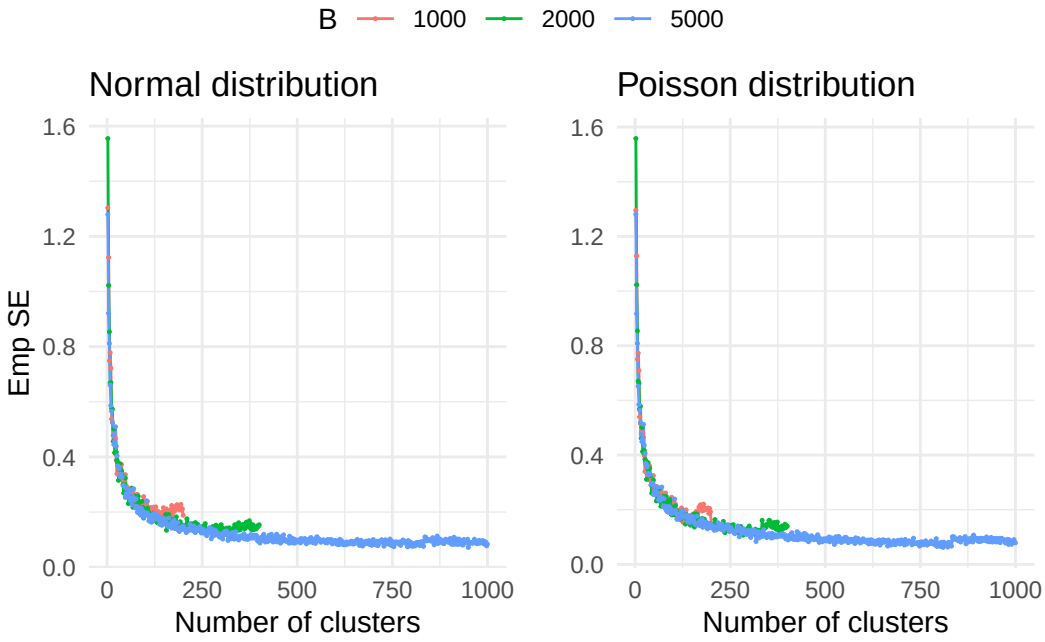

Figure 6: Estimator's performance with different number of clusters and budget. $\sigma = 1$, $\gamma = 1$, $\alpha = 1$, $c_2 = 1$, Cost ratio $= 5$

Figure 6 shows results similar to Figure 5, indicating that the standard error of the estimator is not affected by the value of β in either distribution.

Conclusion

The results of this project provide insights into the factors influencing the standard error of the estimator under both normal and Poisson distributions. The optimal number of clusters is primarily influenced by the cost ratio and the variance components. Higher cost ratios and larger outcome variances generally require fewer clusters to minimize the standard error, emphasizing the trade-off between the number of clusters and measurements per cluster.

Interestingly, the simulation results for the Poisson distribution deviate slightly from theoretical expectations, particularly with respect to γ and α. For example, the unexpected relationship between $\gamma = 1$ and cluster requirements suggests that further investigation with a higher number of simulations may be necessary to confirm these findings. Similarly, the lack of variance changes with α in the Poisson distribution, despite its theoretical link to variance, warrants additional exploration.

Overall, the result demonstrates that the optimal design of clusters and measurements depends not only on the resource constraints but also on the distributional properties of the outcome.

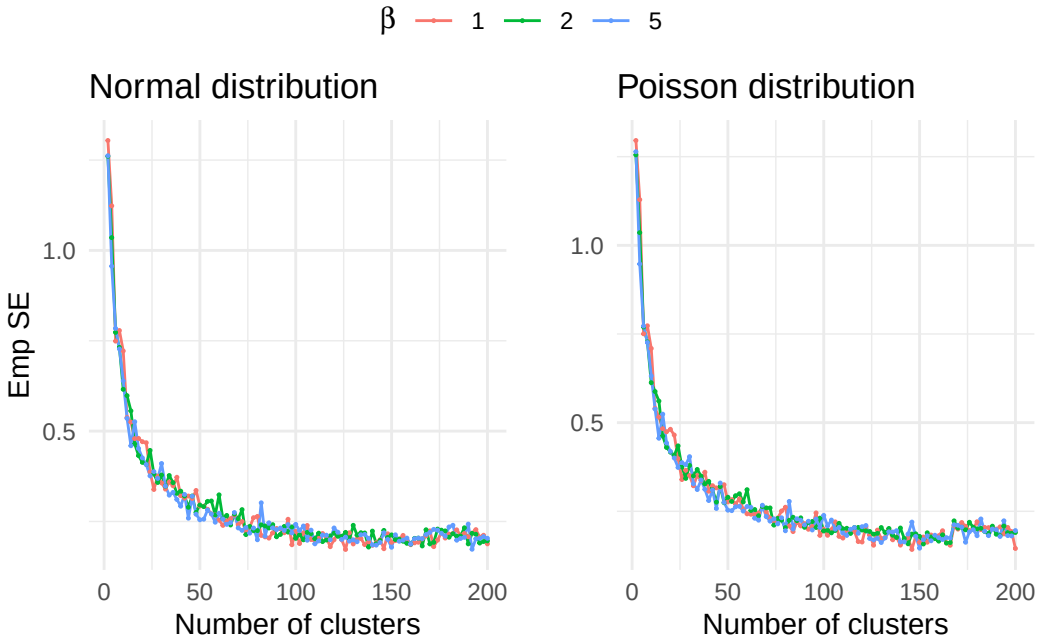

Figure 7: Estimator's performance with different number of clusters and β. $B = 1000$, $\sigma = 1$, $\gamma = 1$, $\alpha = 1$, Cost ratio $= 5$, $c_2 = 1$

These findings can guide researchers in allocating resources effectively to achieve unbiased and efficient estimators while balancing practical constraints.

Code Appendix

```
library(ggplot2)
library(gtsummary)
library(ggpubr)
library(dplyr)
library(tidyr)
library(latex2exp)
library(gtsummary)
library(gt)
res <- read.csv("../Result/estimation_result.csv")
res$B <- as.factor(res$B)

res_sum <- read.csv("../Result/estimation_summary.csv")
beta_names <- c('1' = "Beta = 1", "2" = "Beta = 2", "5" = "Beta = 5")
```

```r
p1 <- ggplot(res, aes(x = beta_hat_n, y = m)) +
  geom_jitter(size = 0.4, alpha = 0.3) +
  geom_vline(data = res[res$beta == 1,], aes(xintercept = 1), color = 'red') +
  geom_vline(data = res[res$beta == 2,], aes(xintercept = 2), color = 'red') +
  geom_vline(data = res[res$beta == 5,], aes(xintercept = 5), color = 'red') +
  facet_wrap(~beta, scales = 'free_x', labeller = as_labeller(beta_names) ) +
  labs(title = TeX("Estimated $\\beta$ from Normal distribution")) +
  theme_minimal() +
  theme(axis.line.y = element_blank(), axis.text.y = element_blank(), axis.ticks.y = element_

p2 <- ggplot(res, aes(x = beta_hat_p, y = m)) +
  geom_jitter(size = 0.4, alpha = 0.3) +
  geom_vline(data = res[res$beta == 1,], aes(xintercept = 1), color = 'red') +
  geom_vline(data = res[res$beta == 2,], aes(xintercept = 2), color = 'red') +
  geom_vline(data = res[res$beta == 5,], aes(xintercept = 5), color = 'red') +
  facet_wrap(~beta, scales = 'free_x', labeller = as_labeller(beta_names)) +
  labs(title = TeX("Estimated $\\beta$ from Poisson distribution")) +
  xlab("Estimated effect size") +
  theme_minimal() +
  theme(axis.line.y = element_blank(), axis.text.y = element_blank(), axis.ticks.y = element_

ggarrange(p1, p2, ncol = 1)

ggplot(res_sum, aes(x = Var_n, y = Var_p)) +
  geom_point()+
  geom_abline(intercept = 0, slope = 1, color = 'grey') +
  ylab(("SE (Poisson)")) +
  xlab(("SE (Normal)")) +
  labs(title = '') +
  theme_minimal()

ratio_label <- c('5' = "Cost Ratio = 5", "10" = "Cost Ratio = 10", "50" = "Cost Ratio = 50",
res_sum_n_sigma <- res_sum[res_sum$alpha == 1 &
                       res_sum$beta == 1 &
                       res_sum$gamma == 1 &
                       res_sum$B == 1000,]

ggplot(res_sum_n_sigma, aes(y = Var_n, x = G, color = as.factor(sigma))) +
  geom_line() +
  geom_point(size = 0.5) +
  facet_wrap(~ratio, scales = 'free', labeller = as_labeller(ratio_label)) +
  scale_color_discrete(name = TeX("$\\sigma$")) +
```

```
  labs(title ='') +
  ylab('Emp SE')+ xlab('Number of clusters') +
  theme_minimal()
sigma_sum <- res_sum_n_sigma %>%
  group_by(ratio, sigma) %>%
  summarise(G = G[which.min(Var_n)]) %>%
  pivot_wider(names_from = sigma, values_from = G)

sigma_sum <- as.data.frame(sigma_sum)
  gt(sigma_sum) %>%
  tab_header(title = "Optimal number of clusters under normal distribution") %>%
  cols_label(`ratio` = 'Cost Ratio' ,`0.25` = "Sigma = 0.25", `1` = "Sigma = 1", `5` = "Sigma


res_sum_n_gamma <- res_sum[res_sum$alpha == 1 &
                        res_sum$beta == 1 &
                        res_sum$sigma == 1 &
                        res_sum$B == 1000,]

p1 <- ggplot(res_sum_n_gamma, aes(y = Var_n, x = G, color = as.factor(gamma))) +
  geom_line() +
  geom_point(size = 0.5) +
  facet_wrap(~ratio, scales = 'free', labeller = as_labeller(ratio_label)) +
  scale_color_discrete(name = TeX("$\\gamma$")) +
  labs(title ='Normal distribution') +
  ylab('Emp SE')+ xlab('Number of clusters') +
  theme_minimal()
p2 <- ggplot(res_sum_n_gamma, aes(y = Var_p, x = G, color = as.factor(gamma))) +
  geom_line() +
  geom_point(size = 0.5) +
  facet_wrap(~ratio, scales = 'free', labeller = as_labeller(ratio_label)) +
  scale_color_discrete(name = TeX("$\\gamma$")) +
  labs(title ='Poisson distribution') +
  ylab('')+ xlab('Number of clusters') +
  theme_minimal()

ggarrange(p1, p2, ncol = 2, common.legend = TRUE)

gamma_sum_n <- res_sum_n_gamma %>%
  group_by(ratio, gamma) %>%
  summarise(G = G[which.min(Var_n)]) %>%
  pivot_wider(names_from = gamma, values_from = G)
```

```r
gamma_sum_p <- res_sum_n_gamma %>%
  group_by(ratio, gamma) %>%
  summarise(G = G[which.min(Var_p)]) %>%
  pivot_wider(names_from = gamma, values_from = G)

gamma_sum_n <- as.data.frame(gamma_sum_n)
gamma_sum_p <- as.data.frame(gamma_sum_p)
gt(gamma_sum_n) %>%
tab_header(title = "") %>%
cols_label(`ratio` = 'Cost Ratio' ,`0.25` = "Gamma = 0.25", `1` = "Gamma = 1", `5` = "Gamma =
gt(gamma_sum_p) %>%
tab_header(title = "") %>%
cols_label(`ratio` = 'Cost Ratio' ,`0.25` = "Gamma = 0.25", `1` = "Gamma = 1", `5` = "Gamma =

res_sum_n_alpha <- res_sum[res_sum$beta == 1 &
                        res_sum$sigma == 1 &
                        res_sum$gamma == 1 &
                        res_sum$B == 1000,]

p1 <- ggplot(res_sum_n_alpha, aes(y = Var_n, x = G, color = as.factor(alpha))) +
  geom_line() +
  geom_point(size = 0.5) +
  facet_wrap(~ratio, scales = 'free', labeller = as_labeller(ratio_label)) +
  scale_color_discrete(name = TeX("$\\alpha$")) +
  labs(title ='Normal distribution') +
  ylab('Emp SE')+ xlab('Number of clusters') +
  theme_minimal()
p2 <- ggplot(res_sum_n_alpha, aes(y = Var_p, x = G, color = as.factor(alpha))) +
  geom_line() +
  geom_point(size = 0.5) +
  facet_wrap(~ratio, scales = 'free', labeller = as_labeller(ratio_label)) +
  scale_color_discrete(name = TeX("$\\alpha$")) +
  labs(title ='Poisson distribution') +
  ylab('')+ xlab('Number of clusters') +
  theme_minimal()

ggarrange(p1, p2, ncol = 2, common.legend = TRUE)

res_sum_b <- res_sum[res_sum$alpha == 1 &
                        res_sum$beta == 1 &
                        res_sum$sigma == 1 &
                        res_sum$gamma == 1 &
```

```
                      res_sum$ratio == 5,]

p1 <- ggplot(res_sum_b, aes(y = Var_n, x = G, color = as.factor(B))) +
  geom_line() +
  geom_point(size = 0.2) +
  scale_color_discrete(name = "B") +
  labs(title ='Normal distribution') +
  ylab('Emp SE')+ xlab('Number of clusters') +
  theme_minimal()
p2 <- ggplot(res_sum_b, aes(y = Var_p, x = G, color = as.factor(B))) +
  geom_line() +
  geom_point(size = 0.2) +
  scale_color_discrete(name = "B") +
  labs(title ='Poisson distribution') +
  ylab('')+ xlab('Number of clusters') +
  theme_minimal()

ggarrange(p1, p2, ncol = 2, common.legend = TRUE)

res_sum_beta <- res_sum[res_sum$alpha == 1 &
                        res_sum$sigma == 1 &
                        res_sum$gamma == 1 &
                        res_sum$B == 1000 &
                        res_sum$ratio == 5,]

p1 <- ggplot(res_sum_beta, aes(y = Var_n, x = G, color = as.factor(beta))) +
  geom_line() +
  geom_point(size = 0.2) +
  scale_color_discrete(name = TeX("$\\beta$")) +
  labs(title ='Normal distribution') +
  ylab('Emp SE')+ xlab('Number of clusters') +
  theme_minimal()
p2 <- ggplot(res_sum_beta, aes(y = Var_p, x = G, color = as.factor(beta))) +
  geom_line() +
  geom_point(size = 0.2) +
  scale_color_discrete(name = TeX("$\\beta$")) +
  labs(title ='Poisson distribution') +
  ylab('')+ xlab('Number of clusters') +
  theme_minimal()

ggarrange(p1, p2, ncol = 2, common.legend = TRUE)
```