Sample-Size Calculations for A/B Testing using Delta Method: Investigation

Abstract

This project evaluates the method suggested by Zhou et al. (2023) for estimating sample size when the unit of analysis is smaller than the unit of randomization. To account for data correlation, Zhou et al. (2023) recommend using the Delta Method, arguing that it improves statistical significance in assessments of Type I error and power compared to the "standard" sample size method used for independent data. This study replicates and extends the original comparative analysis, focusing on absolute lift in eleven different scenarios. Our findings support those of Zhou et al. (2023).

Keywords: A/B testing, Online Controlled Experiments, Sample Size Calculation, Delta Method, Correlated Data, Absolute Lift

1. Introduction

The rapid evolution of information technology has propelled A/B testing, or online controlled experimentation, to the forefront as the preferred empirical methodology for assessing the impact of various modifications on user behavior, system performance, and feature enhancements. This method entails the random allocation of participants into two distinct groups: a control group (A), which interacts with the existing system configuration, and a treatment group (B), which experiences a novel variant. Subsequent analysis involves measuring and statistically evaluating key performance indicators (KPIs) to ascertain if there are statistically significant differences between the outcomes of these groups.

Critical to the integrity of A/B testing are two statistical concepts: sufficient power and controlled Type I error. Power, from a statistical standpoint, is defined as the likelihood of accurately rejecting a false null hypothesis, signifying the test's capacity to identify an actual effect when it is present. Conversely, Type I error, represented as α, refers to the probability of erroneously rejecting a true null hypothesis, leading to the false identification of a difference where none exists. Striking a balance between high statistical power and controlled Type I error is imperative for ensuring the accuracy and dependability of conclusions drawn from A/B tests, an essential aspect of data-driven decision-making.

Addressing a significant void in power analysis research, particularly in scenarios where the unit of analysis surpasses the granularity of the unit of randomization, Zhou et al. (2023) proposed a novel methodology for sample size calculation that integrates the Delta method. This advancement offers more precise sample size estimations in complex experimental frameworks, thereby augmenting the robustness of statistical analyses.

The remainder of the paper is organized as follows. Section 2 elaborates on the methodology delineated by Zhou et al. (2023) for sample size calculation. Section 3 delves into the simulation of data, describing the varied scenarios employed in each simulation and examining the results. Section 4 presents concluding observations and proposes potential avenues for future research in this field. To aid in the practical comprehension of the discussed concepts, examples of R code are interspersed throughout the paper.

2. Sample Size Calculation: Absolute Lift

Absolute lift is the difference in the metric (such as conversions rate) between the control group and treatment groups. This section is divided by sample size calculation methodology for independent data, followed by a subsection with introduction to Delta method used for correlated data and statistical significance evaluation.

2.1 Independent Data

Assume $X_1, ..., X_n$ and $Y_1, ..., Y_n$ as independent and identically distributed observations from control and treatment groups, respectively. The unknown population means for these groups are denoted by $\mu_x = \mathbb{E}(X)$ for control and $\mu_y = \mathbb{E}(Y)$ for treatment.

In the context of continuous outcomes, assuming equal-sized groups, the sample size per group is calculated as follows:

$$n = 2\sigma^2 \cdot \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 / \delta^2 \tag{1}$$

where σ represents the standard deviation for both control and treatment observations, α and β are the allowable Type I/II error rates, z_k is the k^{th} percentile of the standard normal distribution, and δ is the average treatment effect (ATE) $\delta = \mu_y - \mu_x$

For binary outcomes, a two-sample proportion test is utilized with hypotheses $H_0 : p_x = p_y$ versus $H_1 : p_x \neq p_y$, where $\delta = p_y - p_x$. The sample size per group, under the assumption of equal sample size per arm, is calculated as follows:

$$n = 2p_{\text{pool}}(1 - p_{\text{pool}}) \cdot \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 / \delta^2 \tag{2}$$

where $p_{pool} = (p_x + p_y)/2$. The primary difference between equations (1) and (2) resides in the approach to calculating standard deviation. Zhou et al. (2023) refer to to these equations as the ”standard” formulas for sample size calculation.

2.2 Correlated Data

In the context of online controlled experiments, correlated data often emerge when analyzing more granular data, such as session-level information, against user-level randomization. This situation, typically referred to as cluster randomization in statistical and econometric studies, involves clusters (the units of randomization) containing multiple units that are correlated, like multiple sessions from one user, challenging the assumption of data independence.

The Delta method is applied in this setting as follows. Consider k users in a treatment arm, each user i having N_i observations X_{ij} $(i = 1, ..., k; j = 1, ..., N_i)$. The average metric is calculated as $\bar{X} = \sum_{i,j} X_{ij} / \sum_i N_i$. It is generally reasonable to assume independence among users, but observations within a user are likely correlated. Ignoring these intra-user correlations and treating observations as independent leads to an underestimation of the variance of $\bar{X}$ (Zhou et al., 2023). To address this, Deng et al. (2018) redefined $\bar{X}$ using two i.i.d variables S_i and N_i, and normalized by the number of users k:

$$\bar{X} = \frac{\sum_{i,j} X_{ij}}{\sum_i N_i} = \frac{\sum_i S_i/k}{\sum_i N_i/k} = \frac{\bar{S}}{\bar{N}} \tag{3}$$

where $S_i = \sum_j X_{ij}$ signifies the sum of observations for the ith user.

Subsequently, the bivariate Delta method was employed to determine:

$$\text{Var}\left(\frac{\bar{S}}{\bar{N}}\right) \approx \frac{1}{k\mu_N^2}\left(\sigma_S^2 - 2\frac{\mu_S}{\mu_N}\sigma_{SN} + \frac{\mu_S^2}{\mu_N^2}\sigma_N^2\right) \tag{4}$$

where $\mu_S = E(S_i), \mu_N = E(N_i), \sigma_S^2 = Var(S_i), \sigma_N^2 = Var(N_i)$, and $\sigma_{SN} = Cov(S_i, N_i)$. This formula approximates the variance of $\bar{X}$ through the bivariate distribution of (S_i, N_i).

To properly calculate sample size from correlated data, we can rewrite (1) as $\left(\sigma^2/n\right)-1 = 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2/\sigma^2$, then replace σ^2/n with the non i.i.d. version of Var$(\bar{X})$ in (4), and solve for k, resulting in the required number of users:

$$k = 2h \cdot \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \cdot \sigma^2, \tag{5}$$

where

$$h = \frac{1}{\mu_N^2}\left(\sigma_S^2 - 2\frac{\mu_S}{\mu_N}\sigma_{SN} + \frac{\mu_S^2}{\mu_N^2}\sigma_N^2\right), \tag{6}$$

This sets the stage for a thorough evaluation of the proposed methodology's effectiveness via detailed simulation studies.

3. Data Simulation

We aim to conduct a comparative analysis between the sample size calculation method utilizing the Delta method and the conventional 'standard' approach, assessing their robustness in maintaining power and controlling Type I error across various scenarios (refer to Table 1).

Prior to utilizing the proposed sample size calculation methodology, it is imperative to generate simulated data that would mirror as the historical trends observed before an experimental intervention, that would usually be used for initial calculations. We set the initial conversion rate for the control group as $p_x = 0.6$. To emulate the correlation seen in historical data, a clustering structure akin to a random effects model is employed. Each user i is modeled to adhere to its own binary distribution with a mean p_i, thereby introducing correlation. Each treatment arm is thus centered around a common mean p_x, aggregated from the means p_i within the arm. Assuming a historical sample size of $K = 5,000$ users:

1. The number of sessions for each user i is generated, with $N_i \sim \text{Poisson}(\lambda = 5)$.

```
    sess <- rpois(n = 5000, lambda = 5)
```

2. For each user i, a conversion probability p_i is drawn from a normal distribution $N(p_x = 0.6, \sigma = 0.175)$, truncated at two standard deviations to ensure p_i is constrained between 0 and 1.

```r
  pi <- rtruncnorm(n = 1, a = max(0, 0.6 - 0.175*2), b = min(1, 0.6 + 0.175*2), mean = 0.6, sd = 0.175/2)
```

3. Each session j for user i is modeled as $X_{ij} \sim \text{Bernoulli}(p_i)$ for $j = 1, \ldots, N_i$, which results in the total number of converted sessions per user i being $S_i = \sum_j X_{ij}$.

```r
sess <- rpois(n = 5000, lambda = 5)
conv <- numeric(5000) # Placeholder for converted sessions

# Simulate user behavior for session conversions
for (i in 1:N) {
  pi <- rtruncnorm(n = 1, a = max(0, 0.6 - 0.175*2), b = min(1, 0.6 + 0.175*2), mean = 0.6, sd = 0.175/2)

  conv[i] <- sum(rbinom(n = sess[i], size = 1, prob = pi))
}
```

Following data generation, all requisite components as specified in equation (6) are computed, leading to an h value of 0.071. Targeting a power of 80%, a Type I error rate of 0.05, and an anticipated actual treatment effect $\delta = p_y - p_x = 0.05$, the computed requisite sample size per arm is $k = 448$, as derived from equation (5).

```r
h <- 1 / mean(sess)^2 *(var(conv) - 2*mean(conv)/mean(sess)*cov(sess,conv) + mean(conv)^2/mean(sess)^2*var(sess))

#number of users (correlated)
k <-  2 * h * (qnorm(1 - alpha / 2) + qnorm(1 - beta))^2 / (0.05^2)
k*mean(sess) #number of sessions
```

In comparison, using i.i.d method (2) we get n = 1,472, which is less than $k \cdot \mu(sess) =$ 2,247 by proposed method (Table 2, Case II).

```r
# Number of sessions (i.i.d)
p_pool <- (0.6 + 0.65) / 2
ate <- abs(0.65 - 0.6) #absolute treatment effect

n <- 2 * p_pool * (1 - (p_pool)) * (qnorm(1 - alpha / 2) + qnorm(1 - beta))^2 / ate^2
```

Table 1: Simulation Setup

Case	True p_x	True δ in H_1	λ	p_i distribution (range)
I	0.6	0.02	5	Trunc $N(p \pm 0.35)$
II	0.6	0.05	5	Trunc $N(p \pm 0.35)$
III	0.6	0.1	5	Trunc $N(p \pm 0.35)$
IV	0.6	0.15	5	Trunc $N(p \pm 0.35)$
V	0.6	0.2	5	Trunc $N(p \pm 0.35)$
VI	0.6	0.05	2	Trunc $N(p \pm 0.35)$
VII	0.6	0.05	10	Trunc $N(p \pm 0.35)$
VIII	0.6	0.05	15	Trunc $N(p \pm 0.35)$
IX	0.6	0.05	30	Trunc $N(p \pm 0.35)$
X	0.6	0.05	5	Trunc $N(p \pm 0.18)$
XI	0.6	0.05	5	Trunc $N(p \pm 0.27)$

3.1 Statistical Significance: Proposed Method

Now, using the derived users (k) and (n), the statistical significance for each dataset will be evaluated by examining the t-statistic against the critical value: $\left|\frac{\bar{Y}-\bar{X}}{\sqrt{\text{Var}(\bar{Y})+\text{Var}(\bar{X})}}\right| \geq z_{1-\frac{\alpha}{2}}$. For the control group, we simulate data using a mean $\mu = p_x$ as specified in step (1), and $\mu = p_y$ for the treatment group. This forms a single dataset, which we then replicate to independently generate 10,000 additional datasets for evaluating $H_0 : p_x = p_y = 0.6$ versus $H_1 : p_x \neq p_y$. Table 1 indicate scenarios of simulation setups, where either ATE, $\lambda = 5$, or distribution range is changed.

```
power_correlated <- function(N, lambda, a_mu, b_mu, range, n_simulations, alpha) {
  t_stats <- numeric(n_simulations) #placeholder
  (....) #part of the code is skipped

    # Implementing Delta methods (3) and (4)
    x_mean <- (sum(x_conv)/N) / (sum(x_sess)/N)
    y_mean <- (sum(y_conv)/N) / (sum(y_sess)/N)

    x_var <- 1 / (mean(x_sess)^2 * N) * (var(x_conv) - 2 * mean(x_conv)/mean(x_sess) * cov(x_sess, x_conv) + mean(x_conv)^2/mean(x_sess)^2 * var(x_sess))
    y_var <- 1 / (mean(y_sess)^2 * N) * (var(y_conv) - 2 * mean(y_conv)/mean(y_sess) * cov(y_sess, y_conv) + mean(y_conv)^2/mean(y_sess)^2 * var(y_sess))

    t_stat <- (y_mean - x_mean) / sqrt(x_var + y_var)
    t_stats[sim] <- t_stat
  }
  power <- sum(abs(t_stats) > qnorm(1 - alpha / 2))
  power_rate <-  power / n_simulations
  return(power_rate)
}
```

3.2 Statistical Significance: Standard Method

Calculating sample size using standard method leads to less sessions than when using the proposed method. For example, in case II, n = 1,472 sessions are suggested rather than 2,247 sessions. There are two scenarios, how n sessions can be reached.

The following are excerpts from the code used in the analysis. These excerpts are not in sequential order but have been selected to demonstrate the implementation of each scenario. The approach to determine statistical significance remains unchanged,

3.2.1 Scenario I: reducing k

The number of sessions per user is kept, while the number of users (k) is reduced from proposed method. This adjustment is made by cumulatively counting the sessions for each group until the target number of sessions is reached.

```
target_sessions <- 1472*2 #multiplied by two as it combines both groups

    (...) #parts of code are skipped

   # Initialize cumulative session count and user index
    cumulative_sessions <- 0
    last_included_user <- N

    # Simulate for each user
    for (i in 1:N) {
      # Check if target sessions reached and break loop if so
      if (cumulative_sessions >= target_sessions) {
        last_included_user <- i - 1
        break
      }

      # Update cumulative sessions
      cumulative_sessions <- cumulative_sessions + a_sess[i] + b_sess[i]
    }

    # Limit data to users up to last_included_user
    a_sess <- a_sess[1:last_included_user]
    b_sess <- b_sess[1:last_included_user]
    a_conv <- a_conv[1:last_included_user]
    b_conv <- b_conv[1:last_included_user]
```

3.2.2 Scenario II: reducing sessions

The number of users (k) is kept, while the number sessions per users is reduced from proposed method. This adjustment is made by aligning the number of sessions to n, instead of the product of the number of users and the average number of sessions, $k \cdot \mu(sess)$. This reduction is operationalized through incorporating a reduction factor to the number of sessions.

```r
# Maximum number of sessions per group, from initial k calculations
max_sessions <- n

for (sim in 1:num_simulations) {
  a_sess <- rpois(n = N, lambda = lambda)
  b_sess <- rpois(n = N, lambda = lambda)

  # Proportionally reduce sessions for each group if they exceed the maximum
  if (sum(a_sess) > max_sessions_per_group) {
    reduction_factor_a <- max_sessions_per_group / sum(a_sess)
    a_sess <- round(a_sess * reduction_factor_a)
  }
  if (sum(b_sess) > max_sessions_per_group) {
    reduction_factor_b <- max_sessions_per_group / sum(b_sess)
    b_sess <- round(b_sess * reduction_factor_b)
```

Table 2: Sample Size & Performance Using Proposed Method

Case	k	$k \cdot \mu_N$	Power	Type I error
I	2,820	14,046	0.8024	0.0529
II	448	2,247	0.8002	0.0496
III	112	567	0.7639	0.052
IV	49	248	0.7202	0.057
V	28	139	0.7157	0.059
VI	885	1,768	0.8017	0.0518
VII	305	3,038	0.8095	0.054
VIII	224	3,681	0.766	0.0562
IX	194	5,836	0.7867	0.0524
X	353	1,750	0.8239	0.0481
XI	451	2,247	0.7995	0.0491

4. Results

Zhou et al. (2023) have demonstrated that their proposed method outperforms the standard method in all cases when evaluating power and Type I error for absolute lift. However, the expanded case scenarios reveal that while the proposed method generally surpasses the standard method in reliability (with the exception of Type I error in scenario ii), it exhibits a reduction in power when the Average Treatment Effect (ATE) exceeds 0.1 and when the correlation is high ($\lambda > 15$).

Consider, for example, Case IV, characterized by a true $p_x = 0.6$, $\delta = 0.15$, $\lambda = 5$, and a truncated normal distribution of $p \pm 0.35$. In this case, the proposed method's power was below the desired 80% threshold at 72%, coupled with a marginally elevated Type I error rate of 5.7%. However, type I error performance fluctuation could be attributed to variability. Conversely, in the same case, the standard method under scenario (i) exhibited a power of 67% and a Type I error rate of 12%; scenario (ii) demonstrated a power of 63%

with a Type I error of 5.2%. These results highlight that despite some limitations, the proposed method consistently outperforms the standard method in terms of reliability.

Table 2 and Table 3 contains results for each case detailed in Table 1. Red marks metrics that under-performed.

Table 3: Sample Size & Performance Using Standard Method

Case	n	Scenario	Power	Type I error
I	9,336	(i)	0.7529	0.1129
		(ii)	0.6833	0.0489
II	1,472	(i)	0.7583	0.112
		(ii)	0.6991	0.0515
III	357	(i)	0.7096	0.1173
		(ii)	0.6307	0.0518
IV	153	(i)	0.6882	0.1254
		(ii)	0.5923	0.0561
V	82	(i)	0.6707	0.1417
		(ii)	0.5641	0.0573
VI	1,472	(i)	0.777	0.0703
		(ii)	0.7667	0.0488
VII	1,472	(i)	0.7263	0.1669
		(ii)	0.6352	0.0513
VIII	1,472	(i)	0.6782	0.1974
		(ii)	0.5975	0.0482
IX	1,472	(i)	0.6685	0.3323
		(ii)	0.5816	0.0463
X	1,472	(i)	0.7999	0.0694
		(ii)	0.7562	0.0536
XI	1,472	(i)	0.7711	0.0837
		(ii)	0.7274	0.0498

5. Conclusion

In this study, the findings of Zhou et al. (2023) regarding the efficacy of the Delta method over the standard i.i.d sample size calculation approach were validated. However, research also highlighted the Delta method's limitations, that were not found in the original paper. Notably in scenarios III-V with ATE of 10% to 20%, and VIII-IX where users sessions averaged about 15 to 30 sessions. These observations suggest that the Delta method may not be the most suitable in cases of very high user correlation.

An area that warrants further exploration is the comparison of the Delta method with other sample size calculation techniques, such as those available in the R 'pwr' package.

Future research should focus on this comparative analysis to broaden the understanding of the Delta method's performance relative to other methodologies, especially in complex data scenarios.

Overall, while the Delta method is a substantial improvement for correlated data scenarios, its application should be approached with caution in cases of significant treatment effects and high user session counts. The research community would benefit from further studies that compare a wider array of sample size calculation methods.

References

Alex Deng, Ulf Knoblich, and Jiannan Lu. Applying the delta method in metric analytics: A practical guide with novel ideas. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 233–242, 2018.

Jing Zhou, Jiannan Lu, and Anas Shallah. All about sample-size calculations for a/b testing: Novel extensions & practical guide, 2023.