Simulation-based validation of sample size calculations for comparing proportion in clinical study

Student Name: TingWei, Chen[a]
Supervisor Name: Einbeck, Jochen[b]

[a]*Faculty of Science, Durham University, UK, kftk66@durham.ac.uk*
[b]*Department of Mathematical Science, Durham University, UK, Jochen.einbeck@durham.ac.uk*

Abstract

The objective of this paper is to facilitate user-friendly access to sample size calculations prior to conducting experiments, enabling a comparative assessment of various methodologies integrated into the application. With a focus on contemporary medical studies, which often involve comparing proportions between two groups, this study undertakes the replication of sample size estimations from a selected paper available on the British Medical Journal (BMJ). This replication adheres to fundamental tenets and statistical methodologies such as Z-statistics, Chi-squared test, and Fisher's exact test. Armed with a foundational grasp of sample size calculations, this paper undertakes Monte Carlo simulations, generating random proportional data through binomial distribution to assess inter-group equality. The deployment of simulation, emulating real-world populations, is intended to fortify the stability and robustness of the sample size estimations across multiple simulation iterations. An introduction of random variation, achieved by varying the random seed in each iteration, discerns the variance of power yielded by the chosen equality tests. As a culmination, an interactive R Shiny application is curated, spotlighting power analysis and sample size calculation. This amalgamation seamlessly integrates the general Z-statistics formula with simulation methodologies for comparing two proportional subjects. This application not only furnishes users with an overarching panorama of power distribution across sample sizes but also facilitates a comparative examination of power curves between the general formula and simulation-based approaches. This endeavor offers users the essential flexibility to customize inputs according to their specific experimental requisites.

Keywords: sample size calculation, power analysis, Z statistic, Monte Carlo simulation, binomial distribution, Fisher's exact test

1. Introduction

Before clinical research initiation, an objective is aimed and set to resolve certain illnesses. While new drugs are under development, their effectiveness and potency need to be optimized before becoming available to the public, which helps to prevent the risk of unknown side effects and fatal consumption. Due to the rapid evolution of new diseases and viruses, researchers are continuously developing drugs for both novel harmful bacteria and unresolved deadly strains of viruses. Well-developed drugs can only be hoped to deliver fast enough across the globe with the changing environment. A fully-developed drug takes time to test for its effectiveness and side effects; therefore, sample size calculation, the first step in starting clinical research, becomes an undeniably crucial step in the clinical research process.

Sample size in clinical research represents the number of patients or other units of participants who will be enrolled in a study [1]. In any clinical study, sample size calculation or estimation serves as an essential preliminary stage to ensure an effective experiment outcome, because appropriate sample size calculation estimates the proper number of subjects needed to achieve the treatment effect in hypothesis testing under several statistical factors. In general clinical research, the sample size mainly depends on factors such as the statistical significance level, the study's power, the expected effect size, and variability [2], which will be further elaborated upon in the subsequent section.

1.1 Motivation

Estimating sample size in an analytic study or experiment involves various approaches. These approaches can vary depending on the type of predictor and involved outcome variables. An inaccurate sample size estimation in the initial phase of clinical experimental design will lead to a lack of clinical significance in the desired effect difference, which will waste resources in the clinical research process and valuable time. While surveying research papers published over the past two years on the British Medical Journal website, numerous papers can be observed with a trend of inadequate presentation or the lack of acknowledgment in their sample size calculations. Nonetheless, a prevailing trend has become noticeable in recent research or cohort studies that emphasize the investigation of treatment efficacy, disease prevalence, and risk factors [3][4]. These studies typically involve the investigation of comparing the proportions of subjects in two groups, each having a dichotomous outcome. In this research paper, it will attempt to reproduce the sample size calculation of the surveyed selected paper, which will serve as the pivotal statistical parameter for reproduction. Additionally, this research paper will propose an application to validate the sample size calculation, which is calculated by the method of comparing proportions. This paper aims to provide the tool to validate, reproduce, and determine the sample size population for any future clinical research studies.

In this paper, the application introduced will illustrate its visual functionality in calculating sample size. The idea behind the construction of this application comes from a power calculator application that helps its user understand the power of simple experimental designs which detect the average treatment effects. By providing an application, this paper offers users the essential flexibility to customize inputs according to their specific experimental requisites.

1.2 Structure

In the ensuing sections, the paper is delineated as follows:

First, we will introduce the foundational principles of sample size planning calculation for comparing two proportional subject groups, drawing from biostatistics textbooks, and we will present the statistical approaches used to estimate the sample size. Second, we will apply these methods and knowledge in a step-by-step manner to replicate the sample size calculation based on parameters from a selected journal. This replication will elucidate how researchers estimate sample size before collecting data for their experimental designs. Third, we will present the simulation methodology used for comparing two proportions. This simulation involves generating data from the binomial distribution for two proportions and conducting comparisons using two equality of proportions hypothesis test approaches: Fisher’s exact test and the Chi-squared test. Subsequently, we will construct power curves and conduct power analysis by integrating the Z-statistics formula and the simulation using two hypothesis testing methods into an interactive application in R Shiny. Finally, we will devise a robustness test to assess the impact of random effects on the variability of statistical power.

2. Statistical Foundational Principles

2.1 Sample Size Calculation for Comparing Two Proportions

2.1.1 The Underlying Sample Size Calculation Principles

Before applying the statistical approaches to estimate sample size, it is essential to be acquainted with the underlying statistical principles and gather relevant clinical research information in advance. The determination of sample size in research studies varies, depending on the targeted effect of the study design, but many of these methods share similarities. To analyze the effect of comparing two proportional subjects using Chi-squared or Z-statistics methods, the following steps should be followed:

(1) Formulate hypothesis: Clearly state the null hypothesis and decide whether the alternative hypothesis should be one-tailed (directional) or two-tailed (non-directional).

(2) Calculate Effect Size and Variability: Assess the effect size and variability by estimating the proportions P_1 (representing the outcome in one group) and P_2 (representing the outcome in the other group).

(3) Determine Significance Level and Power: Set the desired significance level (α) and power for the statistical test to control Type I and Type II errors, respectively.

2.1.2 Hypothesis

Accurate formulation of hypotheses serves as a critical foundation in the initiation of research endeavors, particularly when investigating differences among groups. A well-constructed hypothesis possesses essential, unambiguous attributes of defining variables. Also, under scrutiny, the hypothesis will anticipate the relational dynamics between the variables. In the realm of medical research, the null hypothesis typically postulates the absence of any noteworthy or substantial variance, effect, variables, treatments, conditions or interrelationship within the studied groups. Simplistically, the null hypothesis conjectures that any discernible disparities or effects are solely products of chance or random fluctuations. This conjecture also implies the absence of any intrinsic veritable effect or relationship within the sampled population. On the other hand, the alternative hypothesis propels inquiry by asserting the existence of a true effect or relationship within the studied population. Researchers wield the alternative hypothesis as a tool to scrutinize whether the amassed data furnish compelling substantiation for the proposed effect or relationship. Notably, two distinct forms of the alternative hypothesis, namely two-sided and one-sided, are admissible choices contingent upon the envisioned direction of the association. The intricate interplay between the null and alternative hypotheses underpins the fabric of hypothesis testing within scientific inquiry. By systematically evaluating the data and considering both hypotheses, researchers can draw meaningful conclusions about the effects or relationships that they are studying in the context of medical research.

2.1.3 Effect Size

Effect size is a statistical measure that quantifies the magnitude of a difference or relationship between two groups or conditions in a study. However, in practice, investigators often cannot ascertain the exact size of the association in the entire population. This uncertainty arises from various factors, including variations in research measurement devices, detailed procedures, or other evaluation methods [8]. Consequently, researchers rely on estimating the effect size from samples. Effect size is also influenced by variability, which reflects the differences among subjects. As measurement errors contribute to overall variability, less precise measurements necessitate larger sample sizes [9]. Therefore, it is essential for researchers to carefully consider variability and conduct a meticulous examination of pivotal studies in the related field before estimating the effect size. By doing so, they ensure a robust study design and derive a reasonable estimate, enhancing the validity and interpretation of research findings. When comparing the proportions of subjects in two groups with a dichotomous outcome, the effect size is the difference between P_1 and P_2, where P_1 and P_2 respectively represent the expected proportions of subjects in the two groups [5] (pp. 57–59). The proportion of subjects adheres to a binomial distribution, thus encapsulating variability within the other parameters integrated into the sample size formulas. As a result, explicit specification of variability is not required.

2.1.4 Determine Significance Level and Power

When conducting hypothesis testing, there is a possibility that the sample may not be fully representative of the population due to inherent uncertainty and variability in the sample data. This can result in erroneous inferences and incorrect conclusions. Type I error (false-positive) occurs when an investigator rejects a null hypothesis that is actually true in the population. Type II error (false-negative) occurs if an investigator fails to reject the null hypothesis when it is actually false in the population. To control these errors, the investigator set the maximum allowable probability of Type I and Type II errors in advance of the study, respectively called α (level of significance) and β [6]. The quantity (1-β) represents statistical power, indicating the probability of correctly rejecting a null hypothesis when it is false. With these values, researchers can estimate the necessary sample size to achieve adequate statistical power, ensuring that the study has a reasonable chance of detecting meaningful effects. In the design of clinical experiments, it is customary to set the α lower than β [7]. The primary purpose of hypothesis testing in clinical trials is to evaluate the effectiveness of a new drug or intervention. Researchers prioritize protecting the null hypothesis to avoid false conclusions about an effect that doesn't exist. Committing a Type I error could lead to erroneous conclusions and potentially harmful actions, such as adopting an ineffective treatment or intervention. Conversely, a Type II error occurs when researchers fail to detect a real effect that exists. While missing an effect may require further investigation, it is generally considered less detrimental than a Type I error. Striking a balance between the two types of errors is crucial to ensure reliable conclusions from clinical trial data and to make informed decisions in medical research and practice.

3. Sample Size Calculation Analysis

3.1 Reproduce Sample Size Calculation

In the realm of clinical trials, the process of reproducing sample size calculations holds paramount importance in ensuring the reliability and credibility of research outcomes. By independently validating the sample size determination of a clinical trial, researchers can verify the accuracy of the original statistical planning. This practice promotes transparency and replicability, allowing other investigators to follow the same methodology and achieve consistent results. Therefore, we will present the reproduction of sample size calculation for comparing the proportion of subjects based on the parameters provided in the research paper [3] titled 'Effect of oral antimicrobial prophylaxis on surgical site infection after elective colorectal surgery: multicentre, randomized, double-blind, placebo-controlled trial with two common calculation approaches.

3.1.1 Background on Selected Research Paper

We selected a research paper from the BMJ website to replicate the sample size calculation for comparing proportional subjects between two groups. The focus of this paper is to investigate whether oral antimicrobial prophylaxis, in addition to intravenous antibiotic prophylaxis, reduces surgical site infections following elective colorectal surgery. Prior to commencing the research, the authors employed statistical methodology to determine the required number of participants. In the "Statistical analysis" section of the cited paper, the authors assume a placebo-associated surgical site infection rate of 15%. Their intention is to identify a 40% relative difference between groups in the incidence of the primary outcome involving both placebo and oral ornidazole treatments. The specified parameters encompass a 5% two-sided Type I error, an 80% power, and the assumption of equal sample sizes for the two groups. As a result of their analysis, the authors determined that a total of 920 patients would be necessary to detect the stated difference within the given power effectively. The exact calculation method employed is not expounded upon in the paper. In this section, we will perform a sample size calculation based on the provided parameters and research objectives. Our aim is to elucidate the methodology employed in the cited paper's sample size determination.

3.1.2 State Hypothesis Testing

The formulation of the hypothesis should elucidate the research question and outcome variables. The experiment's design aims to explore whether incorporating oral antimicrobial prophylaxis with intravenous antibiotic prophylaxis diminishes surgical site infections following elective colorectal surgery. According to the reviews of the literature surveyed by the author, suggests a 15% rate of surgical site infections with a placebo. The study aims to detect a 40% relative difference in the incidence of the primary outcome, which corresponds to a 9% rate of surgical site infections in the oral ornidazole group, with a power of 80% and a 5% two-sided Type I error. Accordingly, the hypothesis testing was constructed as follows:

H_0: There is no difference in the incidence of surgical site infections between the placebo group and the oral ornidazole group.

H_1: There exists a difference in the incidence of surgical site infections between the placebo group and the oral ornidazole group.

3.1.3 Calculate Effect Size

In this paper, $p_1 = 0.15$ is the expected incidence of surgical site infections with placebo, and $p_2 = 0.09$ is the expected incidence of surgical site infections with oral ornidazole, resulting in a relative difference of 40%. Hence, the effect size is the absolute value of group difference $|p_1 - p_2| = 0.06$.

3.1.4 Choose α and Power

In this study, the researcher set a $\alpha = 0.05$, and a statistical power of 0.8, which corresponds to $\beta = 0.2$.

3.1.5 Calculation: Chi-squared table

Upon gathering the necessary elements from the paper, we can determine the sample size using the Chi-squared table, as presented in Table 1 [5] (pp. 75-78). Looking at the table in the leftmost column, where the smaller value between p_1 and p_2 is 0.09, and the expected difference is 0.06, with a two-sided significance level of 0.05 and a power of 0.8 (β = 0.2), the sample size required for each of the two groups is 491 patients.

3.1.6 Calculation: Z-statistics formula

The test of difference of proportions p_1 and p_2 can be conducted using Z-statistics. The total number of sample size n required for study can be expressed as

$$n = \frac{\left[Z_a\sqrt{p(1-p)(1/q_1)} + Z_\beta\sqrt{p_1(1-p_1)(1/q_1) + p_2(1-p_2)(1/q_2))}\right]^2}{(p_1 - p_2)^2} \tag{1}$$

, where q_1 and q_2 are the proportion of subjects in two respective groups, the researcher assumes the number of subjects is equal in two groups ($q_1 = q_2 = 0.5$), P is the expected proportion ($p = p_1 * q_1 + p_2 * q_2 = 0.15 * 0.5 + 0.09 * 0.5 = 0.12$), Z_α is the standard normal deviation for α and Z_β is the standard normal distribution for β. The calculation of Z_α and Z_β can be obtained from the Standard Normal (Z) Table or directly computed using the functions qnorm(1 - α/2) for a two-sided alternative hypothesis, and qnorm(1 - β) in R respectively. $Z_\alpha = 1.96$ when $\alpha = 0.05$; $Z_\beta = 0.84$ when $\beta = 0.2$. After inputting all the relevant values into the equation above, we obtained a total sample size of 920, indicating that there should be at least 460 participants in each group to achieve 80% power. The obtained sample size of 920 aligns with the sample size estimations presented in the referenced paper. Therefore, the reproduction of the sample size calculation process reveals that the researchers in the referenced study employed the Z-statistics methodology to determine the requisite sample size for discerning the relative proportional difference between placebo and oral ornidazole groups.

4. Simulation

As demonstrated in the preceding section, there are a couple of approaches to estimating sample size under the given power by using the formula and generated table. Online sample size calculator tools have proliferated, employing these formulas to aid researchers in their estimations [10], and these calculators have become prevalent in research practice. A majority of these calculators primarily rely on the application of formulas. However, an alternative approach for sample size and power calculation is through the utilization of Monte Carlo simulation. This method entails the generation of numbers of randomized samples according to the probability distributions of desired effect and relevant parameters. These simulated samples are then analyzed to understand the behavior of a statistical model and evaluate

Table 1 sample size per group for comparing two proportions

Upper number: $\alpha = 0.025$ (one sided) OR $\alpha = 0.05$ (two-sided); $\beta = 0.2$
Lower number: $\alpha = 0.025$ (one sided) OR $\alpha = 0.05$ (two-sided); $\beta = 0.1$

	Expected difference between P_1 and P_2						
Smaller of P_1 and P_2	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.06	1272	769	526	388	301	243	202
	1684	1014	691	508	395	318	263
0.07	1419	850	577	423	327	263	217
	1880	1123	760	555	429	343	283
0.08	1562	930	627	457	352	282	232
	2072	1229	827	602	463	369	303
0.09	1702	1007	676	491	377	300	246
	2259	1333	893	647	495	393	322
0.10	1838	1083	724	523	401	318	260
	2441	1434	957	690	527	417	341

properties. Particularly in cases where established formulas are unavailable, Monte Carlo simulation can be a feasible approach for sample size estimation. Furthermore, there are instances where sample size formula might be accessible, yet these formulas could impose constraining assumptions about population parameters that may not hold true within a specific context. For instance, the Chi-squared test adheres to its rule of thumb that renders it unsuitable for small sample sizes. Within this section, we will demonstrate the process of calculating sample size with a designated power through the application of Monte Carlo simulation within a two-proportion group framework. We aim to establish alignment between the application of general formulas and the utilization of simulated data to compute power and determine sample size. The simulation process will involve sampling data generated based on the distribution characteristics of the target population. Subsequently, the Chi-squared hypothesis equality test will be employed to compute power. Furthermore, we will implement the Fisher's exact test to assess hypothesis equality when dealing with scenarios of limited sample sizes. Lastly, we will conduct an uncertainty analysis, leveraging a diverse range of random seed selections for data generation. This will enable the calculation of power and the construction of confidence intervals to evaluate power reliability.

4.1 Principles of sample size calculations by simulation

The basic idea of Monte Carlo simulation is to replicate the process that leads to statistical inferences.

The common approach involves following steps [11]:

Step 1. Model Specification:

Researchers define the statistical model and parameters that underlie the study. This includes variables, distributions, and relationships of interest. For example: if the researcher aims to detect differences in two subgroups from population, we would have to specify the distribution of the outcome variable in each subpopulation (e.g., a normal distribution with mean and standard deviation; a binomial distribution with proportion) and the relative size of two subgroups.

Step 2. Sampling:

Random samples are generated based on specified distributions and parameters. These samples represent hypothetical data that could have been observed in the actual study. In probability sampling, the process initiates with a comprehensive sampling frame encompassing all eligible individuals, ensuring that every qualified individual holds an equal opportunity of being selected for the sample. This approach enhances result generalization in your study. The selection of probability sampling techniques, including but not limited to simple random sampling, systematic sampling, and cluster sampling, is contingent upon the characteristic of the population under consideration. Regardless of the chosen method, representative selection is vital. After establishing a complete population model and sampling strategy, we can generate n sample size by random number.

Step 3. Analysis:

The generated samples are subjected to its inferential problem and the follow up statistical analysis. This process can entail task such as conducting hypothesis test, estimating effect sizes, and evaluating statistical power.

Step 4. Repeat:

Steps 2 and 3 are repeated many times (often thousands or more) to create a large collection of simulated outcomes.

Step 5. Result:

The results of the simulations are analyzed to understand the variability in sample size estimates, and power. This information provides insights into the performance and reliability of the sample size calculation method under different scenarios.

4.2 Implementation of Simulation on Comparing Two Proportions

Upon establishing a foundational understanding of fundamental simulation concepts and procedures, the subsequent objective is to structure the simulation framework for the comparative analysis of two groups characterized by proportional variables. Figure 1 illustrates the sequential stages of the simulation, spanning from the initial sampling to the estimation of power. In the first step, we acquired estimates $\hat{p}_1$ and $\hat{p}_2$ from pilot studies conducted prior to the research. These estimates originate from separate groups, and as the occurrences within each group follow a Binomial distribution, random samples are drawn from this distribution. It is assumed that both groups have the same sample size denoted as n. The second step is to conduct the equality of proportions hypothesis test, which may encompass either the Chi-squared test or Fisher's exact test. In this context, the parameter, $\hat{\theta}$, signifies the p-value of the hypothesis, reflecting the likelihood of observing the test statistic under the assumption that the null hypothesis holds true. Revisiting the established concept, the null hypothesis generally posits the absence of evidence indicating a difference between the two groups. Subsequently, we iterate through steps 1 and 2 multiple times (e.g., 1,000 iterations or more), denoted as i, to generate a collection of p-values from each instance of sampling and hypothesis testing, denoted as $\widehat{\theta}_1, \ldots, \widehat{\theta}_i$. Power is defined as the probability of accurately rejecting the null hypothesis when it is, in fact, false. To assess power, we compute the likelihood of observing a series of p-values that are less than or equal to the predetermined α. Here, α acts as a threshold used to decide whether to reject the null hypothesis, indicating an inequality between two groups in the generated data. This computation involves determining the probability that each p-value, $\widehat{\theta}_i \leq \alpha$ signifies a correct rejection of the null hypothesis in accordance with the hypothesis test.

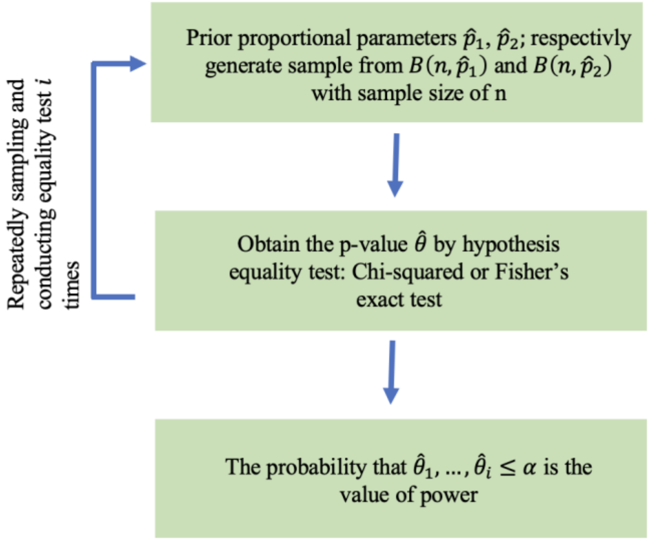

Repeatedly sampling and conducting equality test i times
Prior proportional parameters $\hat{p}_1, \hat{p}_2$; respectivly generate sample from $B(n, \hat{p}_1)$ and $B(n, \hat{p}_2)$ with sample size of n
Obtain the p-value $\hat{\theta}$ by hypothesis equality test: Chi-squared or Fisher's exact test
The probability that $\hat{\theta}_1, \ldots, \hat{\theta}_i \leq \alpha$ is the value of power

Figure 1 The Process of Simulation for Sampling and Power Analysis

4.2.1 Equality of proportions hypothesis test

Three common hypothesis tests arise when comparing dichotomous outcomes of two proportions from independent groups: the Chi-squared test, the Z-statistics test, and Fisher's exact test. The Chi-squared test is frequently employed for categorical data analysis based on the Chi-squared distribution. In its application to proportion comparison across two or more groups, specific

assumptions must be met to ensure the validity of the test results. Specifically, within the context of comparing proportions, the independence of each observation in a contingency table formed by the groups is crucial, along with the expected frequency within each cell of the table exceeds 5, as recommended by Cochran [12] – commonly known as the Chi-squared rule of thumb. This guideline guarantees the Chi-squared distribution's validity as an approximation, with its behavior asymptotically aligning with degrees of freedom (df) equal to $n-1$. This property allows accurate p-value calculations using the chi-square distribution primarily for larger sample sizes – this is where the "asymptotically" concept comes into play. Nevertheless, the test's reliability diminishes when expected frequencies are exceptionally low. Other guidelines have been proposed for the least expected counts [13]. However, all of these guidelines converge in their suggestion that the Chi-squared test is unsuitable for scenarios characterized by small sample sizes. The Z-statistics test finds broad application across various null hypothesis scenarios, as it accommodates both continuous variables and binomial data, provided the sample size is sufficiently large for accurate approximations. Upon observing larger samples, we anticipate the power curve closely aligning with that derived from the Z-statistics general formula. Fisher's exact test is an alternative approach for comparing proportions in small sample sizes. Fisher's exact test is particularly applicable in cases with small sample sizes compared to the Chi-squared test due to its computational approach. While the Chi-squared test relies on approximations that become less reliable with low expected cell counts in a contingency table, Fisher's exact test directly calculates the probability of the observed data under the null hypothesis using the hypergeometric distribution. This makes it more appropriate for situations with limited sample sizes where the assumptions of the Chi-squared test might not be met. Detailed descriptions of the statistical algorithms underpinning both the Chi-squared test and Fisher's exact test when applied to a 2x2 table can be found in established literature [14, 15]. Fisher's exact test considers all possible arrangements of the data (number of combinations of cell counts) that could produce a table as extreme as or more extreme than the observed table, while maintaining the marginal totals fixed.
Although Fisher's exact test is applicable to both small and large samples, its exhaustive consideration of all possibilities can significantly increase computational time, leading to a combinatorial explosion.

4.3 Power Analysis

Before calculating the optimal sample size, researchers define an expected proportion of power they aim to achieve in order to detect the desired effect size based on their hypothesis. However, the process of simulation sampling in Figure 1 involves using samples generated under a specific sample size n to estimate power. The challenge arises because the power achieved by sampling data may not align precisely with the desired power set by researchers for a given sample size n. This discrepancy is addressed through the utilization of a power curve. The power curve is a visual representation that illustrates how the statistical power of a hypothesis test changes as the sample size varies. This curve effectively demonstrates the interplay between sample size and the probability of correctly rejecting a null hypothesis when it is indeed false. By leveraging the power curve, researchers can gain insight into the trade-off between sample size and the capacity to identify a true effect or relationship within the data. To facilitate the simulation process and generate the power curve, we establish a range of sample sizes denoted as r. Therefore, each point on the power curve representing a specific sample size undergoes the simulation process depicted in Figure 1. The total number of iterations needed to construct the power curve will be $i * r$. In the case of general z-statistics hypothesis testing, the power corresponding to each sample size point can be calculated using the formula detailed in section 3.1.5. Z_β can be represented by the equation

$$Z_\beta = \frac{\sqrt{n*(p_1-p_2)^2}-Z_\alpha\sqrt{p(1-p)(1/q_1+1/q_2)}}{\sqrt{p_1(1-p_1)(1/q_1)+p_2(1-p_2)(1/q_2)}} \quad (2)$$

through the rearrangement of the formula. The parameter definitions remain consistent with those in section 3.1.5. The value of Z_α represents the standard normal deviation, which can be computed using the inverse cumulative distribution function of the standard normal distribution or directly using the R function qnorm(1-α/2). Following the computation, the power can be determined using the probability density function of the standard normal distribution at the Z_β quantile.

The power curve can be observed in our exemplified paper, illustrating how it is visually represented. Generally, as the sample size increases, the power tends to improve. This improvement signifies an elevated likelihood of detecting a statistically significant outcome if a true effect is indeed present.

4.4 Uncertainty Analysis

The simulation process is based on the random number (random seed), enabling a simulation to include the variability that occurs in real life. In simulations, different sets of random numbers are applied at different stages. Each set of random numbers serves its own purpose and contributes to the overall randomness of the simulation results. By using separate sets of random numbers, simulations can effectively replicate the various sources of uncertainty and randomness present in real-life situations. This approach helps create a wide range of potential outcomes and accurately represents the complexity of the system under study. The process depicted in Figure 1 involves utilizing a single random seed chosen randomly to generate samples. Our aim here is to evaluate the level of randomness inherent in simulations. To achieve this, we replicate the simulation process N times using N distinct randomly selected random seeds, resulting total $N * i * r$ times iterations. As shown in Figure 2, the red line outlines an additional loop that serves this purpose. In the subsequent application, we utilize N =100 different random seeds to assess the impact of uncertainty on power, considering computation time constraints. By employing different random seeds, the power corresponding to each sample size point is generated N times. This yields a set of N power values denoted as $\hat{P}_1, \ldots, \hat{P}_N$. Consequently, we can construct a confidence interval using these N power values under each sample point. The aim of performing the uncertainty analysis in simulation, by estimating power under various random seeds, is to assess the extent of uncertainty and variability inherent in the random elements produced by the simulation process. This endeavor seeks to establish the stability of power analysis outcomes and provide robust sample size calculations. By achieving this, researchers can utilize this application as a reliable reference when planning experiments, thus enhancing the reliability of their subsequent analyses.

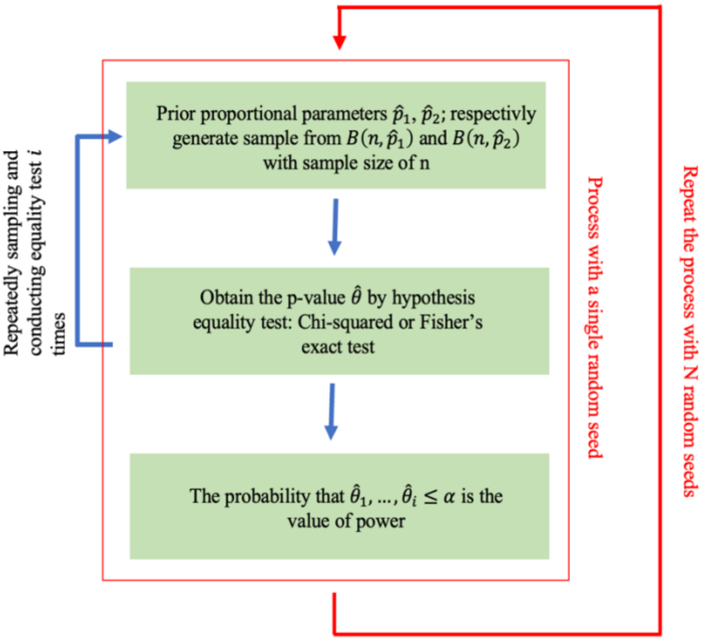

Repeatedly sampling and conducting equality test i times
Prior proportional parameters $\hat{p}_1, \hat{p}_2$; respectivly generate sample from $B(n, \hat{p}_1)$ and $B(n, \hat{p}_2)$ with sample size of n
Obtain the p-value $\hat{\theta}$ by hypothesis equality test: Chi-squared or Fisher's exact test
The probability that $\hat{\theta}_1, \ldots, \hat{\theta}_i \leq \alpha$ is the value of power
Process with a single random seed
Repeat the process with N random seeds

Figure 2 The Process of Uncertainty Analysis in Simulation

4.5 Example from Selected Case Study

Here, we consistently utilize the aforementioned paper to exemplify the simulation procedure for comparing proportions, adhering to the delineated steps and the inferred methodology. The parameters previously utilized to perform sample and power calculations remain consistent with those presented in the preceding section. In this section, we will utilize elements from the paper to demonstrate how we practically conduct the simulation process. We will also perform power analysis by generating power curve plots and implementing the uncertainty analysis. Through hands-on application, we aim to concretely illustrate the methodology outlined in the previous sections.

4.5.1 Population model

The study postulates the existence of two discrete groups: one exhibiting a 15% prevalence of surgical site infections following placebo administration, and the other manifesting a 9% incidence of infections upon oral ornidazole treatment. The assignment of patient proportions within each group conforms to a binomial distribution. Additionally, the author assumes uniformity in sample sizes, denoted as n, for both groups. Let the random variable X represents the number of individuals contracting infections under placebo, and Y signifies the number of individuals acquiring infections with oral ornidazole; these variables can be described as follows:

$$X \sim B(n, 0.15)$$
$$Y \sim B(n, 0.09)$$

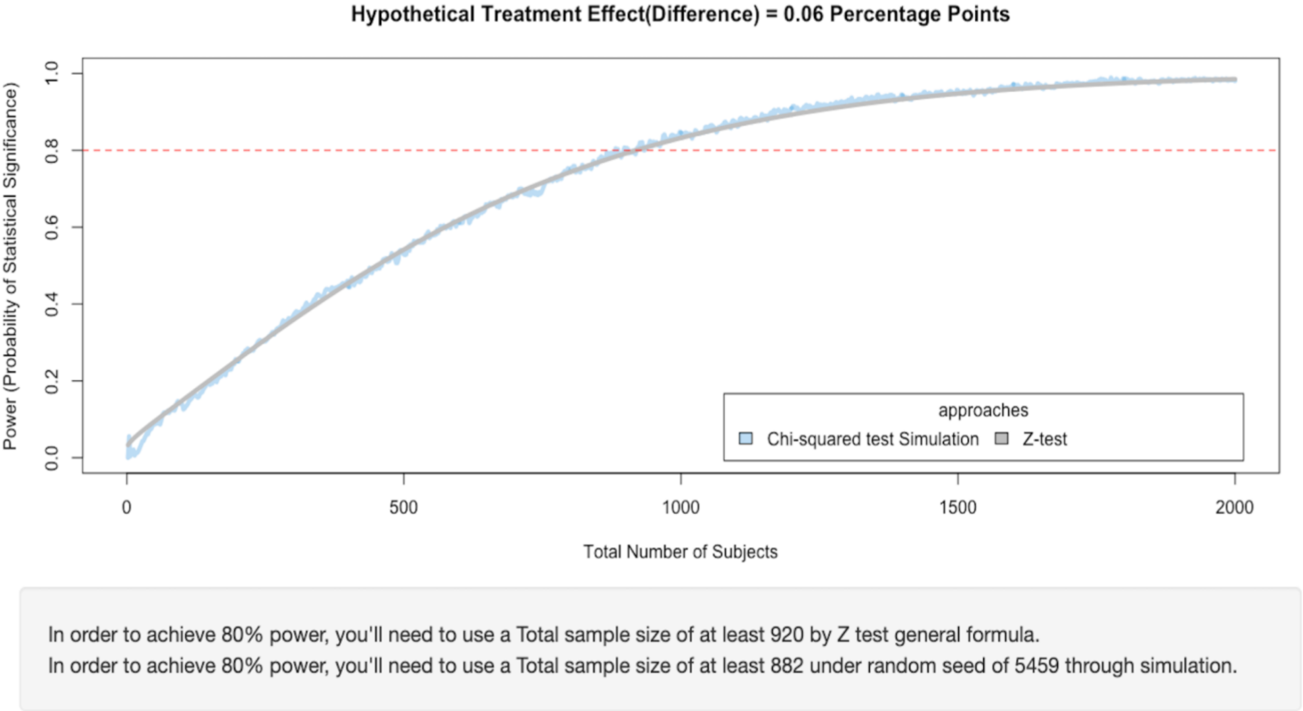
4.5.2 Sampling and Power Analysis

The objective of the study is to determine the sample size that achieves a power of 80% while maintaining a Type I error of 0.05. However, the exact optimal sample size n for each group that best achieves an initial power of 80% cannot be known at the outset of sampling. Consequently, the power curve is utilized to address this situation. A range of sample sizes is defined, ranging from 1 to 1000 per group, denoted as r. While you have the flexibility to choose your preferred range, it's important to note that each sample size point undergoes a simulation process as depicted in Figure 1. Selecting a larger range of sample sizes will result in longer computation times, while a smaller range might lead to sample sizes that fail to attain the desired power. Here, we will use the Chi-squared hypothesis test as an example to illustrate the simulation process. Initially, we define all the necessary parameters as follows:

1. Risk in group 1 (p_1): 0.15
2. Risk in group 2 (p_2): 0.09
3. Type I error (α): 0.05
4. Power target ($1 - \beta$): 0.8
5. The number of simulation iterations (i): 1000
6. maximum number of subjects per group (r):1000

We configure the number of iterations to 1000 and input these values into the algorithms for the simulation process, resulting in the generation of the power curve.

Figure 3 displays the generated power curve using two hypothesis testing approaches: the Chi-squared test and the Z-statistics formula. The gray line illustrates the relationship between power and the total number of subjects reflecting the outcomes produced by inputting parameters into formula (2) across a range of sample sizes n. The calculated sample size for achieving 80% power remains consistent at 920, aligning with the results obtained



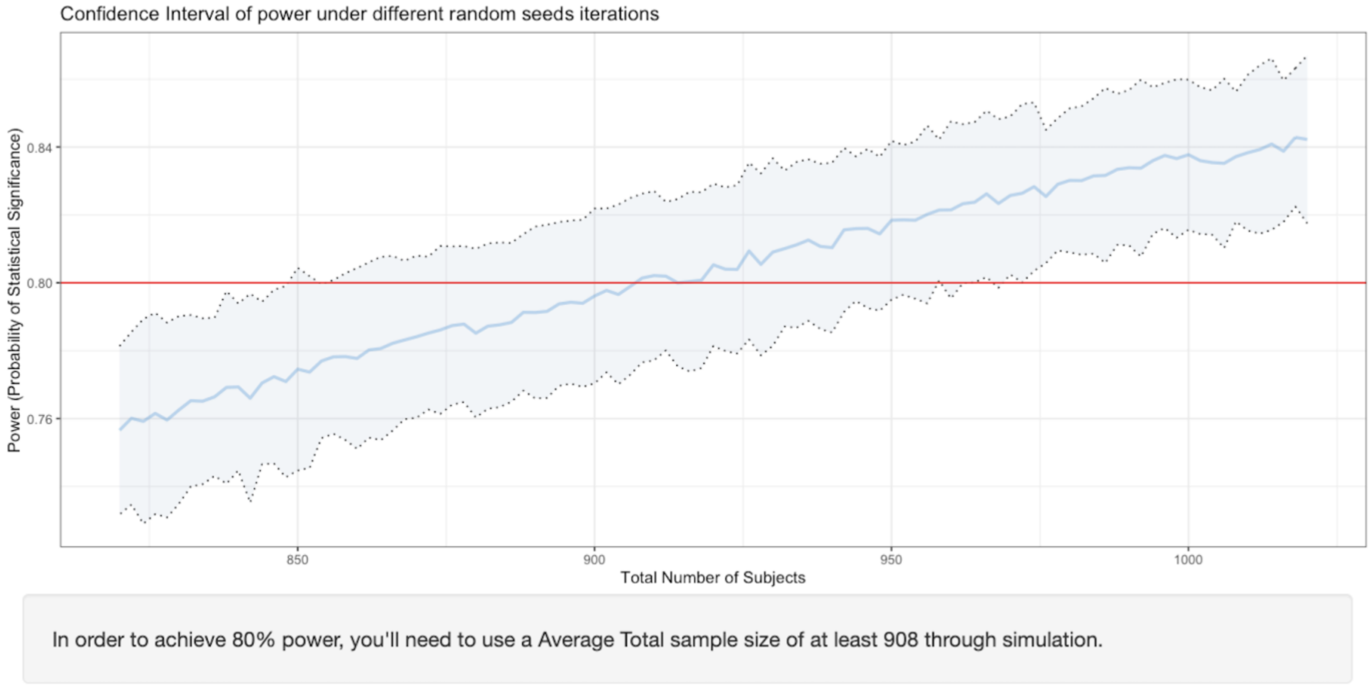
```
In order to achieve 80% power, you'll need to use a Total sample size of at least 920 by Z test general formula.
In order to achieve 80% power, you'll need to use a Total sample size of at least 882 under random seed of 5459 through simulation.
```

Figure 3 The power curve resulting from the Z-statistics formula and the Chi-squared testing of simulated data

in the Sample Size Calculation Analysis section. On the other hand, the light blue line represents power estimates obtained by the Chi-squared test for each sample size point through simulation (using a single random seed). We notice a strong alignment between the blue and gray lines, affirming the validation of this simulation process. The sample size grows with an increase in the number of iterations, resulting in the power derived from simulated data using the Chi-squared test converging toward the outcomes generated by the Z-statistics formula. After conducting $i = 1000$ rounds of sampling for each sample size and calculating power values using the algorithm outlined in Figure 1, it is determined that an optimal sample size of 441 patients per group should participate in this experiment to detect 80% power. To illustrate the simulation process, let's consider $n = 441$ as an example. Initially, we generated one sample for the placebo group from a binomial distribution $B(441, 0.15)$, and for the oral ornidazole group from $B(441, 0.9)$, respectively. Subsequently, a 2x2 contingency table is constructed for these two independent samples, representing the two groups. The Chi-squared equality test is performed, yielding the value theta1 after the first iteration. This process is repeated for 1000 iterations, resulting in 1000 p-values denoted as $\hat{\theta}_1, \dots, \hat{\theta}_{1000}$. The power is computed as the probability of these 1000 p-values being less than or equal to the significance level of $\alpha = 0.05$. Specifically, a power of 0.8 signifies an 80% likelihood that $\hat{\theta}_1, \dots, \hat{\theta}_{1000}$ will reject the null hypothesis.

4.5.3 Uncertainty Analysis

We observe that the optimal sample size under simulation for the Chi-squared test is 441 per group (the paper assumes equal sample size in each group), based on a specific randomly selected random number. It's important to note that this optimal size can vary depending on different random numbers. Consequently, the uncertainty analysis will be utilized to assess the variability in randomness generated by the simulation. This test provides researchers with an interval illustrating how power simulations are distributed across different numbers of subjects. In Figure 4, we focus on a specific range of the number of subjects to perform an uncertainty analysis centered around the optimal sample size derived from one random seed. As depicted in Figure 2, the approach isn't limited to a single random seed; instead, we employ $N = 100$ random seeds to generate power values for each sample size point. The flexibility exists to choose a substantial number of random seeds for constructing confidence intervals for power. A



In order to achieve 80% power, you'll need to use a Average Total sample size of at least 908 through simulation.

Figure 4 The average power curve and the corresponding confidence interval generated through 100 random seeds of simulation.

larger N encompasses a broader range of possible scenarios, thereby yielding power distributions closer to the true population distribution. However, it's essential to consider computation time, as there's a trade-off between computational efficiency and accuracy.

The process of conducting the uncertainty analysis remains consistent until the power is generated for each sample size point.

$$\begin{pmatrix} \hat{P}_{j,1} & \cdots & \hat{P}_{j,N} \\ \vdots & \ddots & \vdots \\ \hat{P}_{r,1} & \cdots & \hat{P}_{r,N} \end{pmatrix}$$

The matrix shown above represents the data generated for conducting the uncertainty analysis. For each sample point $\hat{P}_j, \dots, \hat{P}_r$, $i = 1000$ iterations are carried out to calculate a power value. Each power value within the range (j, r) is computed using the sample generated by a specific sample size n in two groups. Therefore, each $j = 2n$. After performing computations across the range of sample sizes, confidence intervals can be constructed for each sample size point. The confidence intervals for each specific sample size point are established by utilizing the standard deviation of N power data $\hat{\sigma}_j, \dots, \hat{\sigma}_r$, we compute the mean of these N power data values $\bar{P}_j, \dots, \bar{P}_r$, within the range (j, r) for each sample point. The resulting power curve is then visualized using the light blue line in Figure 4. The value of j, corresponding to $\bar{P}_j$ closest to the target power, signifies the optimal total sample size. To illustrate, we use the example presented in Figure 4. In this case, we illustrate sample points ranging from 850 to 1000 for conducting simulations. After performing repeated simulations under $N = 100$ random seeds, power data is acquired for each sample point as shown in the matrix below.

$$\begin{pmatrix} \hat{P}_{850,1} & \cdots & \hat{P}_{850,100} \\ \vdots & \ddots & \vdots \\ \hat{P}_{1000,1} & \cdots & \hat{P}_{1000,100} \end{pmatrix}$$

As an example, consider a total number of subjects of 850, which implies $n = 425$. For this particular sample size, we have 100 power values $\hat{P}_{850,1}, \dots, \hat{P}_{850,100}$, generated using 100 random seed numbers. This results in a total of 1000*100 iterations conducted for a single sample size point. Then, we compute the mean of these 100 power data values for each sample point, resulting in $\bar{P}_{850}$,…,$\bar{P}_{1000}$. Following this, we ascertain the average power value, ranging from $\bar{P}_{850}$ to $\bar{P}_{1000}$, that is closest to the target power of 0.8. The corresponding sample size point associated with this calculated average power becomes the optimal sample size determined through the uncertainty analysis.

5. Application

We have developed an interactive R Shiny program that effectively visualizes the interplay between sample size and power calculations. The aim of this application is to create a user-friendly interface that facilitates the replication of sample size calculations and simulation-based validation when comparing two proportional subjects. This functionality is designed to be versatile, allowing researchers to apply it across various scenarios rather than being limited to a single case. By displaying the dynamic relationship between power and sample size, both simulation-based and formula-driven, this application is specifically designed to assist researchers and practitioners in efficiently estimating the required sample size for comparing proportions between two groups, while considering the desired power and significance level. Meanwhile, the comparison of power curves generated by simulation and formula serves to validate the efficacy of the simulation method. Specifically, this validation is conducted under the Chi-squared hypothesis equality test, where the power curve derived from the simulation is analogous to that obtained through the formulaic approach. We have also incorporated Fisher's exact hypothesis equality test into the application, designed particularly for scenarios involving small sample sizes. While it remains valid for larger sample sizes, considering computational efficiency, we suggest using the Chi-squared test for simulations when dealing with larger sample sizes. Additionally, this application conducts an uncertainty analysis by introducing randomness to assess the variability of power through simulation and construct corresponding confidence intervals. The functionality of this app is in accordance with the methodologies outlined in the preceding section, encompassing the replication of sample size calculations.

The implementation of this application requires the utilization of R packages "ggplot2" and "shiny" for plotting and program development. The R code can be accessed on GitHub via the following link at https://github.com/vivianC99/SampleSizeProportionApp or directly through the URL at https://vvnc.shinyapps.io/Simulation_based_sample_size_calculation/.
Users are prompted to define input variables to engage in sample size calculations, power estimation, and simulation-based validation processes. Subsequently, by simply activating the "run" button, the application generates the power curve and conducts the uncertainty analysis. It also enables the generation of a summary containing the results of the sample size calculations performed based on the chosen simulation settings.

5.1 Input Control

The control widgets serve as interactive elements that users can manipulate by selecting various inputs to generate distinct results. These input widgets are tailored to meet the necessary parameters for sample size calculation and simulation processes. Within this application, users can seamlessly perform both sample size reproduction and simulation-based validation by selecting the desired input values corresponding to their research hypotheses. In Figure 5, we illustrate the user interface featuring input control options for two equality hypothesis tests: the Chi-squared test and Fisher's exact test. Notably, the default parameter settings dynamically adjust based on the selected hypothesis test. This set of input controls is positioned on the left side of the application interface. A notable distinction in the default parameter settings is observed in the "number of simulation iterations." Given the differing computational characteristics, we establish a notably smaller iteration count for Fisher's exact test compared to the Chi-squared test. Users have the flexibility to increase this count if their computing system permits, or alternatively, they can adjust the maximum number of subjects (provided in the bottom widget) to achieve desired settings. This feature provides users with options for tailored and efficient simulations, considering the inherent computational demands of each hypothesis test. Each input widget represents crucial parameters that should be defined in advance according to the research context. The included input widgets are as follows:

1. Number of Simulation Iterations: This input allows users to define the number of iterations used for sampling during the simulation process. It is represented by the parameter i in the simulation process. The range of values for this parameter is set from 10 to 100000. The default number of iterations is 1000 for the Chi-squared test and 200 for Fisher's exact test. The chosen value directly influences the computational time of the simulation.

2. Significance Level: This input corresponds to the Type I error rate, denoted as α. Users can select values of 0.01, 0.025, 0.05, or 0.1 based on the specific research context. The default value for alpha is 0.05. This value serves as the critical threshold for determining the rejection of the null hypothesis.

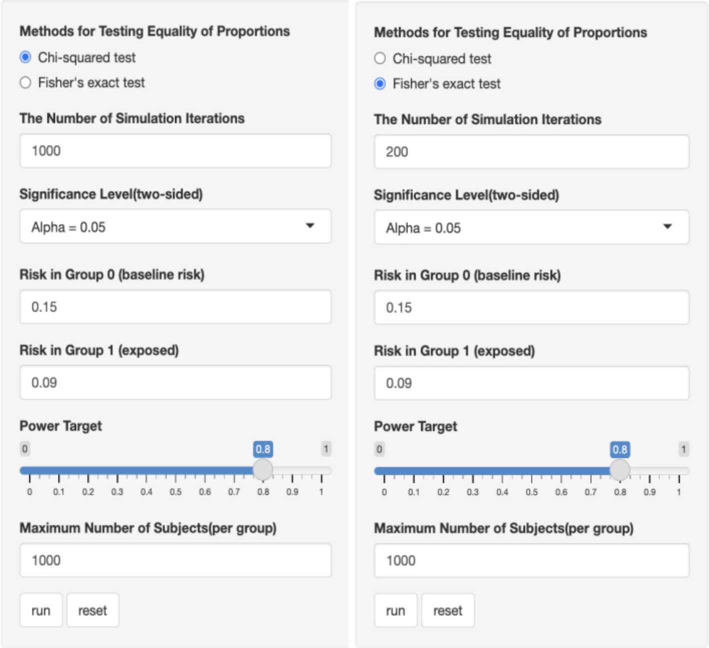
3. Baseline Risk in Group 0: This input represents the parameter P_1, which signifies the baseline risk proportion in one group. The permissible range for this value is between 0 and 1. The default value is set at 0.15, as drawn from the example study.

4. Exposed Risk in Group 1: This input represents the parameter P_2, which signifies the risk proportion in the exposed group, typically applicable to new medical tests or drug trials. The valid range for this parameter is also between 0 and 1. The default value is 0.09, derived from the example study.

5. Power Target: Users can conveniently adjust the desired power using a slider widget. The default value for the target power is 0.8. In medical research, a power value of 0.8 is commonly chosen, as it corresponds to a beta value of 0.2. This beta value signifies the acceptance of a 20% probability of failing to detect a true effect, thereby establishing a balance between sensitivity and specificity in the analysis.

6. Maximum Number of Subjects (per group): This input corresponds to the parameter r, which defines the range of sample sizes from 1 to r on the x-axis of the power curve. The default maximum number is set to 1000. It's important to note that the range

should be chosen carefully. If the selected range is too narrow, it might result in the power curve not adequately covering the necessary sample sizes for achieving the desired power. Conversely, if the range is excessively large, it can significantly impact the computation time. If the maximum number of subjects is set too small to achieve the desired power, errors may arise within the application.



Methods for Testing Equality of Proportions
● Chi-squared test
○ Fisher's exact test

The Number of Simulation Iterations
1000

Significance Level(two-sided)
Alpha = 0.05

Risk in Group 0 (baseline risk)
0.15

Risk in Group 1 (exposed)
0.09

Power Target
0.8

Maximum Number of Subjects(per group)
1000

Methods for Testing Equality of Proportions
○ Chi-squared test
● Fisher's exact test

The Number of Simulation Iterations
200

Significance Level(two-sided)
Alpha = 0.05

Risk in Group 0 (baseline risk)
0.15

Risk in Group 1 (exposed)
0.09

Power Target
0.8

Maximum Number of Subjects(per group)
1000

Figure 5 Input options corresponding to the selected hypothesis tests

Whenever users configure the input values, they should proceed by clicking the “run” button located at the bottom. It's important to be aware that due to the substantial computational iterations involved, pressing the "Run" button initiates the program, and repeatedly pressing it could potentially lead to system instability. The “reset” button can be employed after the previous computation has concluded and users wish to execute the program with different input values. The application provides introductory information and reminders to guide users through the process.

5.2 Analysis Output

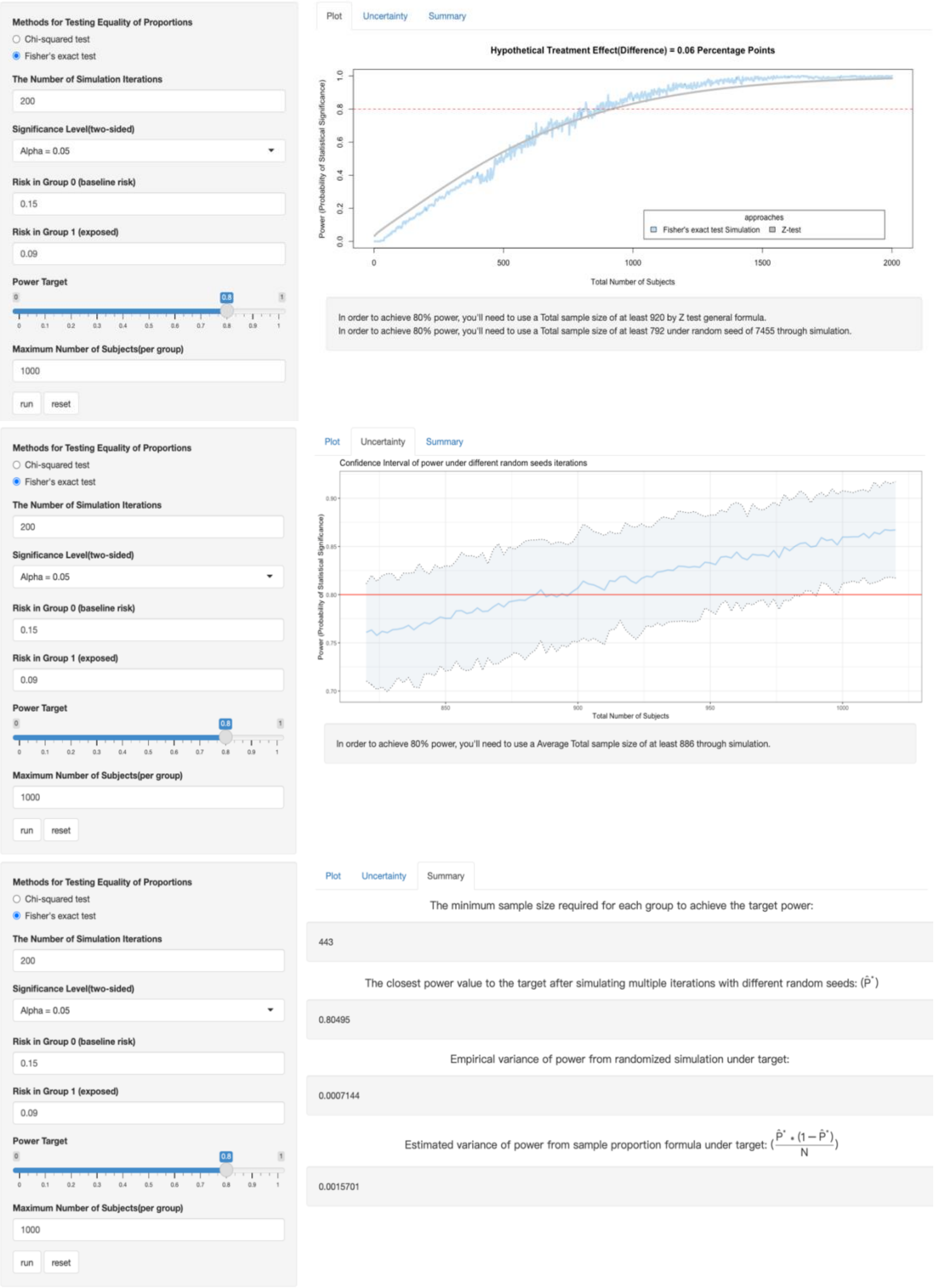
Once users have selected the desired input values and clicked the "Run" button, the application will generate output plots and accompanying text. The output is presented through three tabs: "Plot," "Uncertainty," and "Summary," each containing relevant plots and summary statistics. Further information about the functionality of the buttons and the content available within each tab will be provided in following sections:

5.2.1 Plot tab

The "Plot" tab features a power plot generated through the simulation process outlined in Figure 1. Additionally, this tab provides the output regarding the optimal sample size under a randomly selected random seed. In Figure 6, the first plot illustrates the outcomes derived from the provided input values, as displayed on the left side, in accordance with the example study. The power plot graphically showcases the relationship between the total sample size across two groups, ranging from 1 to 2000, and the power computed using the Fisher’s exact equality test. The label of the light blue line on the plot changes based on the chosen hypothesis test for the simulation process. Furthermore, the light blue power curve varies even when inputting the same values, as the program randomly selects a new random seed for each simulation run. The red line on the plot adapts according to the user's specified target power. The gray line remains consistent regardless of whether the Chi-squared test or Fisher’s exact test is chosen. It remains unaltered for the same input value, as it directly derives from the Z-statistics formula. The primary purpose of this power plot under a single random seed is to provide a comparison between the power curve generated by the Z-statistics formula and the power curves derived from tests using simulated data. This comparison is achieved without incurring excessive computational time. Meanwhile, this plot offers users with an overview and insights into the simulation process, illustrating how the pattern of the power curve evolves across varying sample sizes. More precise analysis for selecting the optimal sample size will be conducted through the uncertainty analysis.

5.2.2 Uncertainty tab

The second plot within Figure 6 showcases the outcome of the uncertainty analysis, presenting a magnified view of the power plot and the calculated optimal average total sample size that achieves the target power. We allocate 10% of the "maximum number of subjects per group" to define a new range for the sample size along the x-axis. The red horizontal line maintains its significance as an indicator of the target power level. The confidence interval is constructed based on power data generated from a range of N different random seeds, with algorithmic details provided in Section 4.5.3. Complementary to the plot, the textual result displays the optimal total sample size corresponding to the desired power level. Incorporated into the system are 100 random seeds, which are not modifiable through the input control widget. However, users can adapt the code on GitHub and execute it in their R application if they



Methods for Testing Equality of Proportions

○ Chi-squared test

● Fisher's exact test

The Number of Simulation Iterations

200

Significance Level(two-sided)

Alpha = 0.05

Risk in Group 0 (baseline risk)

0.15

Risk in Group 1 (exposed)

0.09

Power Target

0.8

Maximum Number of Subjects(per group)

1000

run
reset
0
1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Plot
Uncertainty
Summary
Hypothetical Treatment Effect(Difference) = 0.06 Percentage Points
Power (Probability of Statistical Significance)
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Total Number of Subjects
approaches
Fisher's exact test Simulation
Z-test

In order to achieve 80% power, you'll need to use a Total sample size of at least 920 by Z test general formula.
In order to achieve 80% power, you'll need to use a Total sample size of at least 792 under random seed of 7455 through simulation.

Methods for Testing Equality of Proportions

○ Chi-squared test

● Fisher's exact test

The Number of Simulation Iterations

200

Significance Level(two-sided)

Alpha = 0.05

Risk in Group 0 (baseline risk)

0.15

Risk in Group 1 (exposed)

0.09

Power Target

0.8

Maximum Number of Subjects(per group)

1000

In order to achieve 80% power, you'll need to use a Average Total sample size of at least 886 through simulation.

Methods for Testing Equality of Proportions

○ Chi-squared test

● Fisher's exact test

The Number of Simulation Iterations

200

Significance Level(two-sided)

Alpha = 0.05

Risk in Group 0 (baseline risk)

0.15

Risk in Group 1 (exposed)

0.09

Power Target

0.8

Maximum Number of Subjects(per group)

1000

The minimum sample size required for each group to achieve the target power:

443

The closest power value to the target after simulating multiple iterations with different random seeds: ($\hat{P}^*$)

0.80495

Empirical variance of power from randomized simulation under target:

0.0007144

Estimated variance of power from sample proportion formula under target: ($\frac{\hat{P}^* \cdot (1-\hat{P}^*)}{N}$)

0.0015701

Figure 6 The analysis tab results correspond to the example study conducted using Fisher's exact test

wish to adjust the number of random seeds. The choice of random seed quantity is also contingent on the computational system's capabilities. The primary goal of conducting the uncertainty analysis and generating the power plot complete with its associated confidence interval is to provide users with an understanding of how power trends may vary across different real-world scenarios. This process aims to foster user confidence in employing the application for estimating the optimal sample size in their research experiments.

5.2.3 Summary tab

The “Summary” tab in Figure 6 serves to present the statistical outcomes obtained through the uncertainty analysis. The initial result displayed indicates the minimum sample size required for each group to attain the target power, which is determined by the intersection point of the light blue line and the red line. The second outcome is the power value that comes closest to the target after conducting multiple iterations using different random seeds, denoted as P*. Since power estimations in the uncertainty analysis are carried out through simulation, the average power may not precisely align with the user-set target. To address this, we select the nearest estimated power to the target, specifically the closest value greater than the target power. The third reported outcome is the empirical variance of power obtained from randomized simulations under the target conditions. This represents the squared variance ($\sigma^2_{\hat{P}^*}$) of the nearest estimated power as calculated from the power matrix. The fourth outcome pertains to the estimated variance of power derived from the sample proportion formula under the target settings. The power value itself adheres to the Binomial distribution, wherein the variance of this distribution is denoted by $\frac{\hat{P}^*\times(1-\hat{P}^*)}{N}$. N represents the data length of power for a specific sample point, which is set to 100.

5.3 Computation Time

Having gained a comprehensive understanding of the simulation process, our focus now shifts towards examining the computational aspects associated with generating the power plot and conducting the uncertainty analysis. Within our program, we have conducted simulations involving three instances of the Chi-squared test and two instances of Fisher's exact test. The tabulated data in Table 2 provides insights into the iteration counts for each analytical scenario, along with their corresponding computational time requirements. For the power analysis, the total number of iterations is derived by multiplying the “Input for Iterations” (i) with the “Maximum Sample Size per Group” (r). When considering the uncertainty analysis, the aggregate number of iterations is calculated as the product of the “Input for Iterations” (i), the “Number of Random Seeds” (N), and 10% of the “Maximum Sample Size per Group” (r). We can observe that in the case of performing a total of 10^5 iterations for power analysis, utilizing the Chi-squared test for simulation takes approximately 180 seconds, while employing the Fisher’s exact test results in a longer computation time of around 960 seconds, which is roughly five times longer than the former. This discrepancy in computation time arises from the distinct methodologies employed in conducting the respective hypothesis tests. This test involves calculating the exact probability of observing the given distribution of data under the null hypothesis, which requires considering all possible permutations that could lead to the observed or

Table 2 Iterations and computation time for possible input values corresponding to the example study

Hypothesis Test	Input for Iterations (i)	Maximum sample size per group (r)	Number of random seeds (N)	Total Iterations for Power Analysis (i*r)	Total iterations for unterationty test (i*r*N*0.1)	Computation times
Chi-squared Test	100	1000	100	10^5	10^6	180 sec
Chi-squared Test	500	1000	100	5×10^5	5×10^6	780 sec
Chi-squared Test	1000	1000	100	10^6	10^7	1500 sec
Fisher’s exact Test	100	1000	100	10^5	10^6	960 sec
Fisher’s exact Test	200	1000	100	2×10^5	2×10^6	1800 sec

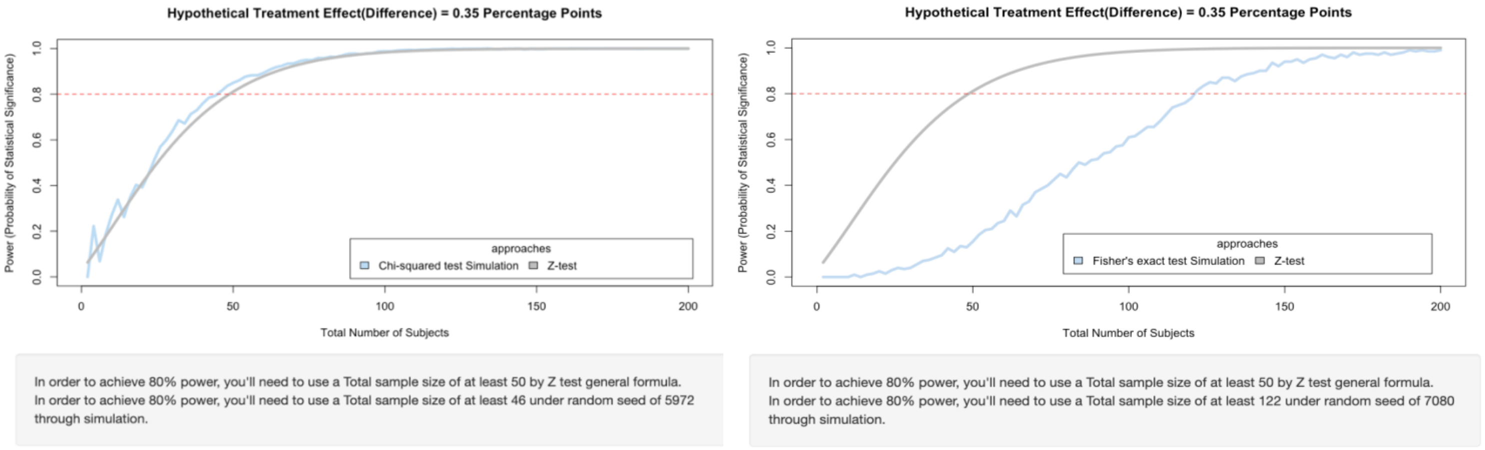
more extreme data. This exhaustive consideration of all possible permutations can become computationally intensive as the sample size and the number of categories in the contingency table increase. Users can refer to this table to assess the trade-off between input values for iterations and the associated computation time.

6. Results and Discussion

This paper is organized into three main sections: (1) Reproduction of Sample Size Calculation, (2) Simulation-Based Validation, and (3) Power and Sample Size Calculation Application Program. The contents encompass both the methodology and its application in a selected example study. Beginning with the fundamental concept of sample size calculation, the paper progresses to address more intricate scenarios. Ultimately, the paper introduces an interactive application that integrates the various methodologies, presenting them through visual plots and summaries. This application simplifies the process of sample size calculation, providing users with a comprehensive tool for their research endeavors. This paper demonstrates accurate validation on two fronts. Firstly, we successfully replicate the sample size calculation of a selected paper that did not detail its calculation process. Our approach encompasses various methods such as Chi-squared table and Z-statistics formula. Secondly, we construct a simulation program for comparing two proportions across two groups. The power pattern depicted in Figure 3 reaffirms the correctness of our simulation. The alignment between the power curve generated by the simulation's Chi-squared test and the one produced by the Z-statistics formula illustrates the statistical accuracy of our simulation, particularly in scenarios involving large sample sizes.

There are several points of discussion within this paper. In terms of simulation, we can discern the disparities between the power curves generated by the Chi-squared and Fisher's exact tests for small sample sizes, as depicted in Figure 7. We initiated the sample size calculation and generated the power plot utilizing the example of proportions, namely 0.55 and 0.9, for two distinct groups. The left plot corresponds to the chi-squared test, while the right one corresponds to Fisher's exact test. Upon closer inspection, we observe that the light blue line in the chi-squared test plot appears relatively irregular. This could be attributed to the sample size being insufficient to satisfy the assumptions of the chi-squared test adequately. On the contrary, the power curve in the case of Fisher's exact test seems smoother when compared to the left plot. While it might be notably distant from the gray line, representing the power curve by z-statistics, this discrepancy is permissible. This is due to the fact that the proportional variable is discrete, and as the sample size increases sufficiently, it approximates the standard normal distribution. In this case, we will adopt the power curve and the optimal sample size calculation conducted by Fisher’s exact test. In the context of the uncertainty analysis, our attention shifts to Figure 4 and Figure 6, specifically plot 2. In this context, we adopt the mean of the power data generated by employing 100 different random seeds as the basis for determining the optimal sample size under the desired target power. However, an element of uncertainty quantification is introduced through the presence of the confidence interval. This uncertainty factor addresses a level of instability, as it implies that for certain cases, the achieved power

Figure 7 Power analysis with Chi-squared and Fisher's exact tests for small sample sizes (proportions: 0.55 and 0.9 in two groups)



In order to achieve 80% power, you'll need to use a Total sample size of at least 50 by Z test general formula.
In order to achieve 80% power, you'll need to use a Total sample size of at least 46 under random seed of 5972 through simulation.

In order to achieve 80% power, you'll need to use a Total sample size of at least 50 by Z test general formula.
In order to achieve 80% power, you'll need to use a Total sample size of at least 122 under random seed of 7080 through simulation.

might fall below the average indicated by the optimal sample size under the target power. Consequently, researchers may opt to exercise greater caution when conducting power analysis. One approach to enhance the safeguarding of achieving the target power involves selecting the optimal sample size based on the intersection of the lower bound of the confidence interval with the red line. This precautionary strategy serves to fortify the likelihood of attaining the desired power level.

7. limitation

When it comes to assessing proportions or dichotomous outcome variables between different groups, there exist various equality tests aside from the Chi-squared test and Fisher’s exact test. However, our application is confined to the utilization of three common hypothesis testing methodologies. Therefore, individuals seeking to employ alternative hypothesis tests will find limitations within this application. Regarding computation time, the topic has long been a point of discussion and deliberation among researchers. Employing a substantial number of simulation iterations (e.g., 10,000 or more) or random seeds aims to emulate real-world population scenarios. In our case, we've chosen a moderate number of iterations (1000 for Chi-squared test and 200 for Fisher’s exact test) and 100 random seeds for the simulation process. This choice is constrained by the computational capabilities of our computer system, and this serves as one of the limitations of our study. In terms of application functionality, we did not implement a real-time progress function due to the complexities of coding. Users are advised to consult Table 2 for an estimation of computation time.

Reference

1. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterol Hepatol Bed Bench. 2013 Winter;6(1):14-7. PMID: 24834239; PMCID: PMC4017493.

2. Kadam P, Bhalerao S. Sample size calculation. Int J Ayurveda Res. 2010 Jan;1(1):55-7. doi: 10.4103/0974-7788.59946. PMID: 20532100; PMCID: PMC2876926.

3. Futier E, Jaber S, Garot M, Vignaud M, Panis Y, Slim K et al. Effect of oral antimicrobial prophylaxis on surgical site infection after elective colorectal surgery: multicentre, randomised, double blind, placebo controlled trial BMJ 2022; 379 :e071476 doi:10.1136/bmj-2022-071476

4. G. Pichler et al., “Cerebral regional tissue Oxygen Saturation to Guide Oxygen Delivery in preterm neonates during immediate transition after birth (COSGOD III): multicentre randomised phase 3 clinical trial,” BMJ, vol. 380, p. e072313, Jan. 2023, doi: https://doi.org/10.1136/bmj-2022-072313.

5. S. B. Hulley, S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, Designing clinical research, 4th ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins, 2013.

6. Daniel, W. W., & Cross, C. L. Biostatistics: A Foundation for Analysis in the Health Sciences, 10th Edition. Hoboken, NJ, Wiley, 2013, ch. 7, pp. 272-277.

7. Das, S., Mitra, K. and Mandal, M. (2016) ‘Sample size calculation: Basic principles’, Indian Journal of Anaesthesia, 60(9), p. 652. doi:10.4103/0019-5049.190621.

8. Hedges, L.V. (2008) ‘What are effect sizes and why do we need them?’, *Child Development Perspectives*, 2(3), pp. 167–171. doi:10.1111/j.1750-8606.2008.00060.x.

9. McKeown-Eyssen, G.E. and Tibshirani, R. (1994) ‘Implications of measurement error in exposure for the sample sizes of case-control studies’, *American Journal of Epidemiology*, 139(4), pp. 415–421. doi:10.1093/oxfordjournals.aje.a117014.

10. J. S. Kohn Michael, “Sample size – Survival analysis | Sample Size Calculators.” https://sample-size.net/sample-size-survival-analysis/(accessed Aug. 14, 2023)

11. S. Landau and D. Stahl, “Sample size and power calculations for medical studies by simulation when closed form expressions are not available,” *Statistical Methods in Medical Research*, vol. 22, no. 3, pp. 324–345, Apr. 2012, doi: https://doi.org/10.1177/0962280212439578.

12. W. G. Cochran, “The χ^2 Test of Goodness of Fit,” *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 315–345, Sep. 1952, doi: https://doi.org/10.1214/aoms/1177729380.

13. W. G. Cochran, “Some Methods for Strengthening the Common χ^2 Tests,” *Biometrics*, vol. 10, no. 4, p. 417, Dec. 1954, doi: https://doi.org/10.2307/3001616.

14. J. E. Overall and R. R. Starbuck, “F-Test Alternatives to Fisher’s Exact Test and to the Chi-Square Test of Homogeneity in 2 × 2 Tables,” *Journal of Educational Statistics*, vol. 8, no. 1, pp. 59–73, Mar. 1983, doi: https://doi.org/10.3102/10769986008001059.

15. I. Campbell, “Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations,” *Statistics in Medicine*, vol. 26, no. 19, pp. 3661–3675, 2007, doi: https://doi.org/10.1002/sim.2832.