SAMPLE SIZE CALCULATION FOR COMPARING PROPORTIONS

HANSHENG WANG
Guanghua School of Management,
Peking University
Department of Business Statistics
& Econometrics
Beijing, P. R. China

SHEIN-CHUNG CHOW
Department of Biostatistics and
Bioinformatics
Duke University
School of Medicine
Durham, New Caledonia, USA

1 INTRODUCTION

In clinical research, in addition to continuous responses, primary clinical endpoints for assessment of efficacy and safety of a drug product under investigation could be binary responses. For example, in cancer trials, patients' clinical reaction to the treatment is often classified as response (e.g., complete response or partial response) or non-response. Based on these binary responses, the proportions of the responses between treatment groups are then compared to determine whether a statistical/clinical difference exists. Appropriate sample size is usually calculated based on a statistical test used to ensure that a desired power exists for detecting such a difference when the difference truly exists. Statistical test procedures employed for testing the treatment effect with binary response are various chi-square or Z-type statistics, which may require a relatively large sample size for the validity of the asymptotic approximations. This article will focus on sample size calculation for binary responses based on asymptotic approximations.

In practice, the objective of sample size calculation is to select the minimum sample size in such a way that the desired power can be obtained for detection of a clinically meaningful difference at the prespecified significance level. Therefore, the selection of the sample size depends on the magnitude of the clinically meaningful difference, the desired power, the prespecified significance level, and the hypotheses of interest under the study of choice (e.g., a parallel design or a crossover design). According to the study objective, the hypotheses could be one of testing for equality, testing for superiority, or testing for equivalence/non-inferiority. Under each null hypothesis, different formulas or procedures for calculation can be derived under different study design accordingly.

The rest of the article is organized as follows. In Section 2, sample size calculation procedures for one-sample design are derived. In Section 3, formulas/procedures for sample size calculation for two-sample parallel design are studied. Considerations for crossover design are given in Section 4. Procedures for sample size calculation based on relative risk, which is often measured by odds ratio in log-scale, are studied under both a parallel design and a crossover design in Section 5 and Section 6, respectively. Finally, the article is concluded with a discussion in Section 7.

2 ONE-SAMPLE DESIGN

This section considers the sample size formula for one-sample test for comparing proportions. More specifically, let $x_i, i = 1, \ldots, n$ be independent and identically distributed (i.i.d) binary observations. It is assumed that $P(x_i = 1) = p$, where p is the true response probability, which can be estimated by

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The parameter of interest is given by $\epsilon = p - p_0$, where p_0 is some reference value (e.g., the response rate of a standard treatment).

2.1 Test for Equality

To test whether a clinically meaningful difference does exist between the test drug and the reference value, the hypotheses of interest are given by

$$H_0 : \epsilon = 0 \text{ versus } H_a : \epsilon \neq 0$$

Under the null hypothesis, the following test statistic

$$T = \frac{\sqrt{n}\hat{\epsilon}}{\sqrt{\hat{p}(1-\hat{p})}}$$

where $\hat{\epsilon} = \hat{p} - p_0$, is asymptotically distributed as a standard normal random variable. Thus, for a given significance level α, the null hypothesis would be rejected if $|T| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$th lower quantile of a standard normal distribution. On the other hand, if the alternative hypothesis $\epsilon \neq 0$ is true, the power of the above test can be approximated by

$$\Phi\left(\frac{\sqrt{n}|\epsilon|}{\sqrt{p(1-p)}} - z_{\alpha/2}\right)$$

Hence, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$\frac{\sqrt{n}|\epsilon|}{\sqrt{p(1-p)}} - z_{\alpha/2} = z_\beta$$

which leads to

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 p(1-p)}{\epsilon^2} \qquad (1)$$

2.2 Test for Non-Inferiority/Superiority

As indicated in Chow et al. (1), the problem of testing non-inferiority and superiority can be unified by the following statistical hypotheses:

$$H_0 : \epsilon \leq \delta \text{ versus } H_a : \epsilon > \delta$$

where δ is the non-inferiority or superiority margin. Under the null hypothesis, the following test statistic

$$\frac{\sqrt{n}(\hat{\epsilon} - \delta)}{\sqrt{\hat{p}(1-\hat{p})}}$$

is asymptotically distributed as a standard normal random variable. Thus, for a given significance level α, the null hypothesis would be rejected if

$$\frac{\sqrt{n}(\hat{\epsilon} - \delta)}{\sqrt{\hat{p}(1-\hat{p})}} > z_\alpha$$

On the other hand, under the alternative hypothesis $\epsilon > \delta$, the power of the above test can be approximated by

$$\Phi\left(\frac{\sqrt{n}(\epsilon - \delta)}{\sqrt{p(1-p)}} - z_\alpha\right)$$

Thus, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$\frac{\sqrt{n}(\epsilon - \delta)}{\sqrt{p(1-p)}} - z_\alpha = z_\beta$$

which leads to

$$n = \frac{(z_\alpha + z_\beta)^2 p(1-p)}{(\epsilon - \delta)^2} \qquad (2)$$

2.3 Test for Equivalence

In order to establish equivalence, the following hypotheses are considered

$$H_0 : |\epsilon| \geq \delta \text{ versus } H_a : |\epsilon| < \delta$$

where δ is the equivalence limit. The above hypotheses can be decomposed into the following two one-sided hypotheses:

$$H_{01} : \epsilon \geq \delta \text{ versus } H_{a1} : \epsilon < \delta$$

and

$$H_{01} : \epsilon \leq -\delta \text{ versus } H_{a1} : \epsilon > -\delta$$

The equivalence between p and p_0 can be established if

$$\frac{\sqrt{n}(\hat{\epsilon} - \delta)}{\sqrt{\hat{p}(1-\hat{p})}} < -z_\alpha \quad \text{and} \quad \frac{\sqrt{n}(\hat{\epsilon} + \delta)}{\sqrt{\hat{p}(1-\hat{p})}} > z_\alpha$$

When the sample size is sufficiently large, the power of the above testing procedure can be approximated by

$$\Phi\left(\frac{\sqrt{n}(\delta-\epsilon)}{\sqrt{p(1-p)}} - z_\alpha\right) + \Phi\left(\frac{\sqrt{n}(\delta+\epsilon)}{\sqrt{p(1-p)}} - z_\alpha\right) - 1$$

Based on similar argument as given in Chow and Liu (2, 3), the sample size needed for achieving the power of $1-\beta$ is given by

$$n = \frac{(z_\alpha + z_{\beta/2})^2 p(1-p)}{\delta^2} \quad \text{if } \epsilon = 0$$

$$n = \frac{(z_\alpha + z_\beta)^2 p(1-p)}{(\delta - |\epsilon|)^2} \quad \text{if } \epsilon \neq 0$$

2.4 An Example

For illustration purposes, consider an example concerning a cancer study. Suppose that it is estimated that the true response rate of the study treatment is about 50% (i.e., $p = 0.50$). On the other hand, suppose that a response rate of 30% ($p = 0.30$) is considered a clinically meaningful response rate for treating cancer of this kind. The initial objective is to select a sample size so that 80% (i.e., $\beta = 0.20$) can be assured for establishing a superiority of the test treatment as compared with the reference value with a superiority margin of 5% (i.e., $\delta = 0.05$). The significance level is set to be 5% (i.e., $\alpha = 0.05$). According to Equation (2), the sample size needed is given by

$$n = \frac{(z_\alpha + z_\beta)^2 p(1-p)}{(p - p_0 - \delta)^2} = \frac{(1.64 + 0.84)^2 0.5(1-0.5)}{(0.5 - 0.3 - 0.05)^2} \approx 69$$

That is, a total of 69 patients are needed for achieving the desired power for demonstration of superiority of the test treatment at the 5% level of significance. However, the investor feels difficult to recruit so many patients within the limited budget and relatively short time frame. Therefore, it may be of interest to consider the result if the study objective is only to detect a significant difference. In this case, according to Equation (1), the sample size needed is given by

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 p(1-p)}{(p - p_0)^2} = \frac{(1.96 + 0.84)^2 0.5(1-0.5)}{(0.5 - 0.3)^2} = 49$$

Therefore, the total number of sample size required reduced from 69 to 49 if the hypotheses of interest for testing superiority (with a superiority margin of 5%) is changed to the hypotheses for detecting a significant difference.

3 TWO-SAMPLE PARALLEL DESIGN

In this section, the problem of sample size calculation for a two-sample parallel design is studied. Let x_{ij} be a binary response from the jth subject in the ith treatment group, where $j = 1, \ldots, n_i$ and $i = 1, 2$. For a fixed i, x_{ij}s are assumed to be i.i.d binary responses with $P(x_{ij} = 1) = p_i$. The parameter of interest is then given by $\epsilon = p_1 - p_2$, which can be estimated by $\hat{\epsilon} = \hat{p}_1 - \hat{p}_2$, where

$$\hat{p}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

3.1 Test for Equality

In a two-sample parallel design, usually one treatment group serves as a control whereas the other is an active treatment. In order to test for equality between the control and the treatment group in terms of the response rate, the following hypotheses are considered

$$H_0 : \epsilon = 0 \quad \text{versus} \quad H_a : \epsilon \neq 0$$

For a given significance level α, the null hypothesis would be rejected if

$$\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \right| > z_{\alpha/2}$$

On the other hand, if the alternative hypothesis $\epsilon \neq 0$ is true, the power of the above test can be approximated by

$$\Phi\left(\frac{|\epsilon|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_{\alpha/2} \right)$$

Thus, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$\frac{|\epsilon|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_{\alpha/2} = z_\beta$$

which leads to

$$n_1 = \kappa n_2$$
$$n_2 = \frac{(z_{\alpha/2} + z_\beta)^2}{\epsilon^2}\left[p_1(1-p_1)/\kappa + p_2(1-p_2)\right]$$

3.2 Test for Non-Inferiority/Superiority

Similarly, the problem of testing non-inferiority and superiority can be unified by the following hypotheses:

$$H_0 : \epsilon \le \delta \quad \text{versus} \quad H_a : \epsilon > \delta$$

where δ is the superiority or non-inferiority margin. For a given significance level α, the null hypothesis should be rejected if

$$\frac{\hat{p}_1 - \hat{p}_2 - \delta}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} > z_\alpha$$

On the other hand, under the alternative hypothesis that $\epsilon > \delta$, the power of the above test can be approximated by

$$\Phi\left(\frac{\epsilon - \delta}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_\alpha\right)$$

Hence, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$\frac{\epsilon - \delta}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_\alpha = z_\beta$$

which leads to

$$n_1 = \kappa n_2$$
$$n_2 = \frac{(z_\alpha + z_\beta)^2}{(\epsilon - \delta)^2}[p_1(1-p_1)/\kappa + p_2(1-p_2)] \tag{3}$$

3.3 Test for Equivalence

Equivalence between the treatment and the control can be established by testing the following hypotheses

$$H_0 : |\epsilon| \ge \delta \text{ versus } H_a : |\epsilon| < \delta$$

where δ is the equivalence limit. The above hypotheses can be decomposed into the following two one-sided hypotheses:

$$H_{01} : \epsilon \ge \delta \text{ versus } H_{a1} : \epsilon < \delta$$

and

$$H_{01} : \epsilon \le -\delta \text{ versus } H_{a1} : \epsilon > -\delta$$

For a given significance level α, the null hypothesis of *inequivalence* would be rejected if

$$\frac{\hat{p}_1 - \hat{p}_2 - \delta}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} < -z_\alpha$$

and

$$\frac{\hat{p}_1 - \hat{p}_2 + \delta}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} > z_\alpha$$

On the other hand, under the alternative hypothesis $|\epsilon| > \delta$ (equivalence), the power of the above test can be approximated by

$$\Phi\left(\frac{\delta - \epsilon}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_\alpha\right) + \Phi\left(\frac{\delta + \epsilon}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_\alpha\right) - 1$$

Based on similar argument as given in Chow and Liu (2, 3), the sample size required for achieving the desired power of $1-\beta$ is given by

$$n_1 = \kappa n_2$$
$$n_2 = \frac{(z_\alpha + z_{\beta/2})^2}{\delta^2}[p_1(1-p_1)/\kappa + p_2(1-p_2)] \quad \text{if } \epsilon = 0 \tag{4}$$
$$n_2 = \frac{(z_\alpha + z_\beta)^2}{(\delta - |\epsilon|)^2}[p_1(1-p_1)/\kappa + p_2(1-p_2)] \quad \text{if } \epsilon \ne 0$$

3.4 An Example

Consider an example concerning the evaluation of two anti-infective agents in the treatment of patients with skin structure infections. The two treatments to be compared include a standard therapy (active control) and a test treatment. After the treatment, the patient's skin is evaluated as cured or not. The parameter of interest is the post-treatment cure rates.

Suppose that, based on a pilot study, it is estimated that the cure rate for the active control is about 50% ($p_1 = 0.50$) whereas the cure rate for the test treatment is about 55% ($p_2 = 0.55$). Only a 5% ($\epsilon = 0.05$) improvement is observed for the test treatment as compared with the control, which may not be considered of any clinical importance. Therefore, the investigator is interested in establishing equivalence between the test treatment and the control. According to Equation (4) and assuming equal sample size allocation ($\kappa = 1$), 5% ($\alpha = 0.05$) level of significance, and 80% ($\beta = 0.20$) power, the sample size needed for establishment of equivalence with an equivalence margin of 15% ($\delta = 0.15$) is given by

$$\begin{aligned} n_1 = n_2 &= \frac{(z_\alpha + z_\beta)^2 (p_1(1-p_1) + p_2(1-p_2))}{(\delta - |\epsilon|)^2} \\ &= \frac{(1.64+0.84)^2 (0.50(1.00-0.50) + 0.55(1.00-0.55))}{(0.15-0.05)^2} \\ &\approx 306 \end{aligned}$$

Thus, a total of 306 patients per treatment group are needed for establishment of equivalence between the test treatment and the active control. On the other hand, suppose the investigator is also interested in showing non-inferiority with a non-inferiority margin of 15% ($\delta = 0.15$), the sample size needed is then given by

$$\begin{aligned} n_1 = n_2 &= \frac{(z_\alpha + z_\beta)^2 (p_1(1-p_1) + p_2(1-p_2))}{(\epsilon - \delta)^2} \\ &= \frac{(1.64+0.84)^2 (0.50(1.00-0.50) + 0.55(1-0.55))}{(0.15+0.05)^2} \approx 77 \end{aligned}$$

Hence, testing non-inferiority with a non-inferiority margin of 15% only requires 77 subjects per treatment group.

4 TWO-SAMPLE CROSSOVER DESIGN

In clinical research, a crossover design is sometimes considered to remove intersubject variability for treatment comparison (4). This section focuses on the problem of sample size determination with binary responses under an $a \times 2m$ replicated crossover design. Let x_{ijkl} be the lth replicate of a binary response ($l = 1, \ldots, m$) observed from the jth subject ($j = 1, \ldots, n$) in the ith sequence ($i = 1, \ldots, a$) under the kth treatment ($k = 1, 2$). It is assumed that $(x_{ij11}, \ldots, x_{ij1m}, \ldots, x_{ijk1}, \ldots, x_{ijkm})$, $i = 1, 2$, and $j = 1, \ldots, n$ are i.i.d. random vectors with each component's marginal distribution specified by $P(x_{ijkl} = 1) = p_k$. Note that by not specifying the joint distribution, the observations from the same subject are correlated with each other in an arbitrary manner. On the other hand, specifying that $P(x_{ijkl} = 1) = p_k$ implies that no sequence, period, and carryover effects exist.

Let $\epsilon = p_2(\text{test}) - p_1(\text{reference})$ be the parameter of interest and define

$$\bar{x}_{ijk\cdot} = \frac{1}{m}(x_{ijk1} + \cdots + x_{ijkm}) \quad \text{and}$$
$$d_{ij} = \bar{x}_{ij1\cdot} - \bar{x}_{ij2\cdot}$$

An unbiased estimator of ϵ can be obtained as

$$\hat{\epsilon} = \frac{1}{an} \sum_{i=1}^{a} \sum_{j=1}^{n} d_{ij}$$

It can be verified that $\hat{\epsilon}$ is asymptotically distributed as $N(0, \sigma_d^2)$, where $\sigma_d^2 = \text{var}(d_{ij})$ and can be estimated by

$$\hat{\sigma}_d^2 = \frac{1}{a(n-1)} \sum_{i=1}^{a} \sum_{j=1}^{n} (d_{ij} - \bar{d}_{i\cdot})^2 \text{ and}$$

$$\bar{d}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} d_{ij}$$

4.1 Test for Equality

For testing equality, the following hypotheses are considered:

$$H_0 : \epsilon = 0 \quad \text{versus} \quad H_a : \epsilon \neq 0$$

Then, the null hypothesis would be rejected at the α level of significance if

$$\left| \frac{\hat{\epsilon}}{\hat{\sigma}_d/\sqrt{an}} \right| > z_{\alpha/2}$$

On the other hand, under the alternative hypothesis that $\epsilon \neq 0$, the power of the above test can be approximated by

$$\Phi\left(\frac{\sqrt{an}|\epsilon|}{\sigma_d} - z_{\alpha/2} \right)$$

Thus, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$\frac{\sqrt{an}|\epsilon|}{\sigma_d} - z_{\alpha/2} = t_\beta$$

which leads to

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma_d^2}{2\epsilon^2} \tag{5}$$

4.2 Test for Non-Inferiority/Superiority

Similarly, the problem of testing for non-inferiority/superiority can be tested based on the following hypotheses

$$H_0 : \epsilon \leq \delta \quad \text{versus} \quad H_a : \epsilon > \delta$$

where δ is the non-inferiority or superiority margin. For a given significance level α, the null hypothesis would be rejected if

$$\frac{\hat{\epsilon} - \delta}{\hat{\sigma}_d/\sqrt{an}} > z_\alpha$$

On the other hand, under the alternative hypothesis that $\epsilon > \delta$, the power of the above test procedure can be approximated by

$$\Phi\left(\frac{\epsilon - \delta}{\sigma_d/\sqrt{an}} - z_{\alpha/2} \right)$$

Hence, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$\frac{\epsilon - \delta}{\sigma_d/\sqrt{an}} - z_{\alpha/2} \geq z_\beta$$

which gives

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma_d^2}{a(\epsilon - \delta)^2} \tag{6}$$

4.3 Test for Equivalence

The equivalence between two treatments can be established by testing the following hypotheses

$$H_0 : |\epsilon| \geq \delta \quad \text{versus} \quad H_a : |\epsilon| < \delta$$

where δ is the equivalence limit. The above hypotheses can be decomposed into the following two one-sided hypotheses:

$$H_{01} : \epsilon \geq \delta \text{ versus } H_{a1} : \epsilon < \delta$$

and

$$H_{01} : \epsilon \leq -\delta \text{ versus } H_{a1} : \epsilon > -\delta$$

Thus, the null hypotheses of *inequivalence* at the α level of significance would be rejected if

$$\frac{\sqrt{an}(\hat{\epsilon} - \delta)}{\hat{\sigma}_d} < -z_\alpha \quad \text{and} \quad \frac{\sqrt{an}(\hat{\epsilon} + \delta)}{\hat{\sigma}_d} > z_\alpha$$

On the other hand, under the alternative hypothesis that $|\epsilon| < \delta$, the power of the above test can be approximated by

$$\Phi\left(\tfrac{\sqrt{an}(\delta - \epsilon)}{\sigma_d} - z_\alpha \right) + \Phi\left(\tfrac{\sqrt{an}(\delta + \epsilon)}{\sigma_d} - z_\alpha \right) - 1$$

Hence, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$\frac{\sqrt{an}(\delta - |\epsilon|)}{\sigma_d} - z_\alpha \geq z_{\beta/2}$$

which leads to

$$n \geq \frac{(z_\alpha + z_{\beta/2})^2 \sigma_d^2}{a(\delta - |\epsilon|)^2}$$

4.4 An Example

Suppose that an investigator is interested in conducting a clinical trial with a crossover design for comparing two formulations of a drug product. The design used is a standard 2×4 crossover design (i.e., ABAB,BABA) $(a = 2, m = 2)$. Based on a pilot study, it is estimated that $\sigma_d = 50\backslash\%$ and the clinically meaningful difference is about $\epsilon = 0.10$. Thus, the sample size needed for achieving an 80% ($\beta = 0.10$) power at the 5% ($\alpha = 0.05$) level of significance can be computed according to Equation (5). It is given by

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma_d^2}{a\epsilon^2} = \frac{(1.96 + 0.84)^2 \times 0.50^2}{2 \times 0.10^2} = 98$$

Thus, a total of 98 patients per sequence are needed for achieving the desired power. If the investigator is interested in showing non-inferiority with a margin of 5% ($\delta = -0.05$), then according to Equation (6), the sample size required can be obtained as

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma_d^2}{a(\epsilon - \delta)^2} = \frac{(1.64 + 0.84)^2 \times 0.5^2}{2 \times (0.10 + 0.05)^2} \approx 35$$

Hence, only 35 subjects per sequence are needed in order to show the non-inferiority of the test formulation as compared with the reference formulation of the drug product.

5 RELATIVE RISK—PARALLEL DESIGN

In addition to the response rate, in clinical trial, it is often of interest to examine the relative risk between treatment groups (5). Odds ratio is probably one of the most commonly considered parameters for evaluation of the relative risk. Given the resposne rates of p_1 and p_2 for the two treatment groups, the odds ratio is defined as

$$\mathrm{OR} = \frac{p_1(1 - p_2)}{p_2(1 - p_1)}$$

It can be estimated by replacing p_i with $\hat{p}_i$, where $\hat{p}_i$ is as defined in Section 3. In practice, the log-scaled odds ratio, defined as $\epsilon = \log(\mathrm{OR})$, is often considered. An estimator of ϵ can be obtained as $\hat{\epsilon} = \log(\hat{\mathrm{OR}})$, where

$$\hat{\mathrm{OR}} = \frac{\hat{p}_1(1 - \hat{p}_2)}{\hat{p}_2(1 - \hat{p}_1)}$$

5.1 Test for Equality

For testing equality, the following hypotheses are considered:

$$H_0 : \epsilon = 0 \quad \text{versus} \quad H_a : \epsilon \neq 0$$

Based on Taylor's expansion, the following test statistic can be obtained:

$$T = \hat{\epsilon} \left[\frac{1}{n_1 \hat{p}_1(1 - \hat{p}_1)} + \frac{1}{n_2 \hat{p}_2(1 - \hat{p}_2)} \right]^{-1/2}$$

Under the null hypothesis, T is asymptotically distributed as a standard normal random variable. Thus, for a given significance level α, the null hypothesis would be rejected if $|T| > z_{\alpha/2}$. Under the alternative hypothesis, the power of such a testing procedure can be approximated by

$$\Phi\left(|\epsilon| \left[\frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)} \right]^{-1/2} - z_{\alpha/2} \right)$$

Hence, the sample size needed for achieving the power of $1 - \beta$ can be obtained by solving the following equation

$$|\epsilon| \left[\frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)} \right]^{-1/2} - z_{\alpha/2} = z_\beta$$

Assuming $n_1/n_2 = \kappa$, it follows that

$$n_1 = \kappa n_2$$

$$n_2 = \frac{(z_{\alpha/2} + z_\beta)^2}{\epsilon^2} \left(\frac{1}{\kappa p_1(1 - p_1)} + \frac{1}{p_2(1 - p_2)} \right)$$

5.2 Test for Non-Inferiority/Superiority

As indicated earlier, the problem of testing non-inferiority and superiority can be unified by the following statistical hypotheses

$$H_0 : \epsilon \leq \delta \quad \text{versus} \quad H_a : \epsilon > \delta$$

where δ is the non-inferiority or superiority margin. Define the following test statistic

$$T = (\hat{\epsilon} - \delta) \left[\frac{1}{n_1 \hat{p}_1(1-\hat{p}_1)} + \frac{1}{n_2 \hat{p}_2(1-\hat{p}_2)} \right]^{-1/2}$$

For a given significance level α, the null hypothesis would be rejected if $T > z_\alpha$. On the other hand, under the alternative hypothesis, the power of the above test can be approximated by

$$\Phi\left((\epsilon-\delta)\left[\frac{1}{n_1p_1(1-p_1)}+\frac{1}{n_2p_2(1-p_2)}\right]^{-1/2}-z_\alpha\right)$$

Hence, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation

$$(\epsilon-\delta)\left[\frac{1}{n_1p_1(1-p_1)}+\frac{1}{n_2p_2(1-p_2)}\right]^{-1/2}-z_{\alpha/2}=z_\beta$$

Assuming $n_1/n_2=\kappa$, it follows that

$$n_1=\kappa n_2$$
$$n_2=\frac{(z_\alpha+z_\beta)^2}{(\epsilon-\delta)^2}\left(\frac{1}{\kappa p_1(1-p_1)}+\frac{1}{p_2(1-p_2)}\right)$$

5.3 Test for Equivalence

In order to establish equivalence, the following statistical hypotheses are considered

$$H_0:|\epsilon|\geq\delta \quad \text{versus} \quad H_a:|\epsilon|<\delta$$

For a given significance level α, the null hypothesis should be rejected if

$$(\hat{\epsilon}-\delta)\left[\frac{1}{n_1\hat{p}_1(1-\hat{p}_1)}+\frac{1}{n_2\hat{p}_2(1-\hat{p}_2)}\right]^{-1/2}<-z_\alpha$$

and

$$(\hat{\epsilon}+\delta)\left[\frac{1}{n_1\hat{p}_1(1-\hat{p}_1)}+\frac{1}{n_2\hat{p}_2(1-\hat{p}_2)}\right]^{-1/2}>z_\alpha$$

On the other hand, under the alternative hypothesis $-\epsilon- < \delta$ is true, the power of the above testing procedure can be approximated by

$$\Phi\left((\delta-\epsilon)\left[\frac{1}{n_1p_1(1-p_1)}+\frac{1}{n_2p_2(1-p_2)}\right]^{-1}-z_{\alpha/2}\right)+$$
$$\Phi\left((\delta+\epsilon)\left[\frac{1}{n_1p_1(1-p_1)}+\frac{1}{n_2p_2(1-p_2)}\right]^{-1}-z_{\alpha/2}\right)-1$$

Based on a similar argument as given in Chow and Liu (2, 3), the sample size needed to achieve the desired power of $1-\beta$ is given by

$$n_1=\kappa n_2$$
$$n_2=\frac{(z_\alpha+z_{\beta/2})^2}{\delta^2}\left[\frac{1}{kp_1(1-p_1)}+\frac{1}{p_2(1-p_2)}\right]^{-1} \quad \text{if } \epsilon=0 \qquad (7)$$
$$n_2=\frac{(z_\alpha+z_\beta)^2}{(\delta-|\epsilon|)^2}\left[\frac{1}{kp_1(1-p_1)}+\frac{1}{p_2(1-p_2)}\right]^{-1} \quad \text{if } \epsilon\neq 0$$

5.4 An Example

Consider a clinical trial for evaluating the safety and efficacy of a test treatment for treating patients with schizophrenia. The objective of the trial is to establish the equivalence between a test treatment with an active control in terms of the odds ratio of the relapse rates. Data from a pilot study indicates that the relapse rates for both the test treatment and the active control are about 50% ($p_1=p_2=0.50$). Assuming a 20% equivalence limit ($\delta=0.20$), equal sample size allocation ($\kappa=1$), 5% significance level ($\alpha=0.05$), and an 80% power ($\beta=0.20$), according to Equation (7), the sample size needed is given by

$$\begin{aligned} n &= \frac{(z_\alpha+z_{\beta/2})^2}{\delta^2}\left[\frac{1}{p_1(1-p_1)}+\frac{1}{p_2(1-p_2)}\right] \\ &= \frac{(1.64+1.28)^2}{0.2^2}\left[\frac{1}{0.50(1-0.50)}+\frac{1}{0.50(1-0.50)}\right] \\ &\approx 1706 \end{aligned}$$

Therefore, a total of 1706 patients per treatment group is needed for achieving the desired power for establishment of equivalence in terms of the odds ratio.

6 RELATIVE RISK—CROSSOVER DESIGN

For the purpose of simplicity, consider a 1×2 crossover design with no period effects. Without loss of generality, it is assumed that every patient will receive the test treatment first

and then crossover to receive the control. Let x_{ij} be the binary response from the jth subject in the ith period, $j = 1, \ldots, n$. The parameter of interest is still the log-scale odds ratio between the test and the control, for example,

$$\epsilon = \log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)$$

which can be estimated by replacing p_i with its estimator $\hat{p}_i = \sum_j x_{ij}/n_i$. According to Taylor's expansion, it can be verified that

$$\sqrt{n}(\hat{\epsilon} - \epsilon) \to_d N(0, \sigma_d^2)$$

where

$$\sigma_d^2 = \text{var}\left(\frac{x_{1j} - p_1}{p_1(1-p_1)} - \frac{x_{2j} - p_2}{p_2(1-p_2)}\right)$$

which can be estimated by $\hat{\sigma}_d^2$, the sample variance based on

$$d_j = \left(\frac{x_{1j}}{\hat{p}_1(1-\hat{p}_1)} - \frac{x_{2j}}{\hat{p}_2(1-\hat{p}_2)}\right), \quad \text{where } j = 1, \cdots, n$$

6.1 Test for Equality

In order to test for equality, the following statistical hypotheses are considered

$$H_0 : \epsilon = 0 \quad \text{versus} \quad H_a : \epsilon \neq 0$$

Under the null hypothesis, the following test statistic

$$T = \frac{\sqrt{n}\hat{\epsilon}}{\hat{\sigma}_d}$$

is asymptotically distributed as a standard normal distribution. Thus, the null hypothesis at the α level of significance would be rejected if $|T| > z_{\alpha/2}$. On the other hand, under the alternative hypothesis that $\epsilon \neq 0$, the power of the above testing procedure can be approximated by

$$\Phi\left(\frac{\sqrt{n}|\epsilon|}{\sigma_d} - z_{\alpha/2}\right)$$

Thus, the sample size needed for achieving the power of $1 - \beta$ can be obtained by solving the following equation

$$\frac{\sqrt{n}|\epsilon|}{\sigma_d} - z_{\alpha/2} = z_\beta$$

which gives

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma_d^2}{\epsilon^2}$$

6.2 Test for Non-Inferiority/Superiority

Similarly, the problem of testing non-inferiority and superiority can be unified by the following hypotheses:

$$H_0 : \epsilon \leq \delta \quad \text{versus} \quad H_a : \epsilon > \delta$$

where δ is the non-inferiority or superiority margin in log-scale. Under the null hypothesis, the following test statistic

$$T = \frac{\sqrt{n}(\hat{\epsilon} - \delta)}{\hat{\sigma}_d}$$

is asymptotically distributed as a standard normal random variable. Thus, the null hypothesis at α level of significance would be rejected if $|T| > z_{\alpha/2}$. On the other hand, under the alternative hypothesis that $\epsilon > 0$, the power of the above test procedure can be approximated by

$$\Phi\left(\frac{\epsilon - \delta}{\sigma_d} - z_\alpha\right)$$

Hence, the sample size needed for achieving the power of $1 - \beta$ can be obtained by solving the following equation

$$\frac{\epsilon - \delta}{\sigma_d} - z_\alpha = z_\beta$$

which leads to

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma_d^2}{(\epsilon - \delta)^2}$$

6.3 Test for Equivalence

Equivalence between treatment groups can be established by testing the following interval hypotheses

$$H_0 : |\epsilon| \geq \delta \quad \text{versus} \quad H_a : |\epsilon| < \delta$$

or the following two one-sided hypotheses:

$$H_{01} : \epsilon \geq \delta \text{ versus } H_{a1} : \epsilon < \delta$$

and

$$H_{01} : \epsilon \leq -\delta \text{ versus } H_{a1} : \epsilon > -\delta$$

Thus, the null hypothesis of *inequivalence* at the α level of significance would be rejected if

$$\frac{\sqrt{n}(\hat{\epsilon} - \delta)}{\hat{\sigma}_d} < -z_\alpha$$

and

$$\frac{\sqrt{n}(\hat{\epsilon} + \delta)}{\hat{\sigma}_d} > z_\alpha$$

On the other hand, under the alternative hypothesis that $|\epsilon| < \delta$ is true, the power of the above test can be approximated by

$$\Phi\left(\frac{\delta - \epsilon}{\sigma_d} - z_\alpha\right) + \Phi\left(\frac{\delta + \epsilon}{\sigma_d} - z_\alpha\right) - 1$$

Based on similar arguments as given in Chow and Liu (2, 3), the sample size needed for achieving the desired power of $1 - \beta$ is given by

$$\begin{aligned} n &= \frac{(z_\alpha + z_{\beta/2})^2 \sigma_d^2}{\delta^2} \quad \text{if } \epsilon = 0 \\ n &= \frac{(z_\alpha + z_\beta)^2 \sigma_d^2}{(\delta - |\epsilon|)^2} \quad \text{if } \epsilon \neq 0 \end{aligned} \tag{8}$$

6.4 An Example

Consider the previous example regarding schizophrenia with the same objectives. Suppose that the trial is now conducted with a crossover design as discussed in this section. Assuming a 20% equivalence limit ($\delta = 0.20$), equal sample size allocation ($\kappa = 1$), 5% significance level ($\alpha = 0.05$), and an 80% power ($\beta = 0.20$), then according to Equation (8), the sample size needed according to Equation (8) is given by

$$\begin{aligned} n &= \frac{(z_\alpha + z_{\beta/2})^2 \sigma_d^2}{\delta^2} = \frac{(1.64 + 1.28)^2 2.00^2}{0.20^2} \\ &\approx 853 \end{aligned}$$

Hence, a total of 853 subjects are needed in order to achieve the desired power of 80% for establishing equivalence between treatment groups.

7 DISCUSSION

In clinical research, sample size calculation with binary response is commonly encountered. This article provides formulas for sample size calculation based on asymptotic theory under a parallel design and a crossover design. It should be noted that a relatively large sample size is usually required in order to ensure that the empirical Type I and II errors are close to the nominal levels. In practice, however, the sample size is often far too small. In such a situation, various exact tests (e.g., binomial test and Fisher's exact test) should be used for sample size calculation. However, sample size determination based on an exact test usually requires extensive computation. More details can be found in Chow et al. (1).

REFERENCES

1. S. C. Chow, J. Shao and H. Wang, *Sample Size Calculation in Clinical Research*. New York: Marcel Dekker, 2003.
2. S. C. Chow and J. P. Liu, *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker, 1992.

3. S. C. Chow and J. P. Liu, *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker, 2000.
4. S. C. Chow and H. Wang, On sample size calculation in bioequivalence trials. *J. Pharmacokinet. Pharmacodynam.* 2001; **28**: 155–169.
5. H. Wang, S. C. Chow and G. Li, On sample size calculation baseed on odds ratio in clinical trials. *J. Biopharmaceut. Stat.* 2003; **12**: 471–483.