SAMPLE SIZE CALCULATION FOR COMPARING MEANS

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

Shein-chung Chow
Millennium Pharmaceuticals, Inc.
Cambridge, Massachusetts

1 INTRODUCTION

Sample size calculation plays an important role in clinical research. In clinical research, a sufficient number of patients is necessary to ensure the validity and the success of an intended trial. From a statistical point of view, if a clinically meaningful difference between a study treatment and a control truly exists, such a difference can always be detected with an arbitrary power as long as the sample size is large enough. However, from the sponsor's point of view, it is not cost-effective to have an arbitrary sample size because of limited resources for a given timeframe. As a result, the objective of sample size calculation in clinical research is to obtain the minimum sample size needed for achieving a desired power for detecting a clinically meaningful difference at a given level of significance. For good clinical practice, it is suggested that sample size calculation/justification should be included in the study protocol before conducting a clinical trial (1).

In practice, the objective of a clinical trial can be classified into three categories: testing treatment effect, establishing equivalence/noninferiority, and demonstrating superiority. More specifically, a clinical trial could be conducted to evaluate the treatment effect of a study drug, it could be conducted to establish therapeutic equivalence/noninferiority of the study drug as compared with an active control agent currently available in the marketplace, or it could be conducted to demonstrate the superiority of the study drug over a standard therapy or an active control agent. Treatment effect, therapeutical equivalence/noninferiority, or superiority are usually tested in terms of some primary study endpoints, which could be either a continuous variable (e.g., blood pressure or bone density) or a discrete variable (e.g., binary response). In a parallel design, patients are randomly assigned to one of several prespecified treatment groups in a double-blind manner. The merit of the parallel design is that it is relatively easy to conduct. In addition, it can be completed in a relatively short period of time as compared with that of the crossover design. The analysis of variance (ANOVA) model is usually considered for analysis of the collected clinical data. For a crossover design, each patient is randomly assigned to a treatment sequence. Within each sequence, one treatment (e.g., treatment or control) is first applied to the patient. After a sufficient length of *washout*, the patient will be crossovered to receive another treatment. The major advantage of the crossover design is that each patient can serve as his/her own control. For a fixed sample size, a valid crossover design usually provides a higher statistical efficiency as compared with a parallel design. However, crossover designs also suffer from a drawback. A crossover design may have potential carryover effect, which may contaminate the treatment effect. For more details regarding the comparison between a parallel design and a crossover design, readers may find the reference by Chow and Liu (2) useful.

The rest of the entry is organized as follows. In Section 2, testing in one-sample problems is considered. Procedures for sample size calculation in two-sample problems under a parallel design and a crossover design are discussed in Sections 3 and 4, respectively. Sections 5 and 6 present procedures in multiple-sample problems under a parallel design (one-way analysis of variance) and a crossover design (Williams design), respectively. A concluding remark is given in the last section.

Wiley Encyclopedia of Clinical Trials, Copyright © 2007, John Wiley & Sons, Inc.

2 ONE-SAMPLE DESIGN

Let x_i be the response of interest from the ith patient, $i = 1, \ldots, n$. It is assumed that x_i's are independent and identically distributed (i.i.d.) normal random variables with mean 0 and variance σ^2. Define

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $\bar{x}$ and s^2 are the sample mean and sample variance, respectively. Let $\epsilon = \mu - \mu_0$ be the true mean difference between a treatment and a reference. Without loss of generality, assume $\epsilon > 0$ ($\epsilon < 0$) an indication of *improvement* (*worsening*) of the treatment as compared with the reference.

2.1 Test for Equality

To test whether a mean difference between the treatment and the reference value truly exists, the following hypotheses are usually considered:

$$H_0 : \epsilon = 0 \quad \text{versus} \quad H_a : \epsilon \neq 0 \qquad (1)$$

For a given significance level α, the null hypothesis H_0 is rejected if

$$\left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > t_{\alpha/2,n-1}$$

where $t_{\alpha/2,n-1}$ is the upper $(\alpha/2)$th quantile of the t-distribution with $n - 1$ degrees of freedom. Under the alternative hypothesis (i.e., $\epsilon \neq 0$), the power of the above test can be approximated by

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

where Φ is the cumulative standard normal distribution function. As a result, the sample size needed to achieve the desired power of $1 - \beta$ is given by

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

2.2 Test for Noninferiority/Superiority

The following hypotheses are usually considered to test noninferiority or superiority:

$$H_0 : \epsilon \leq \delta \quad \text{versus} \quad H_a : \epsilon > \delta \qquad (2)$$

where δ is the superiority or noninferiority margin. When $\delta > 0$, the rejection of the null hypothesis indicates superiority over the reference value. When $\delta < 0$, the rejection of the null hypothesis implies noninferiority against the reference value. For a given significance level α, the null hypothesis H_0 is rejected if

$$\frac{\bar{x} - \mu_0 - \delta}{s/\sqrt{n}} > t_{\alpha,n-1}$$

Similarly, the power of the above test can be approximated by

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

Therefore, the sample size needed to achieve the desired power of $1 - \beta$ is given by

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

2.3 Test for Equivalence

Equivalence between the treatment and the reference value can be established by testing the following hypotheses:

$$H_0 : |\epsilon| \geq \delta \quad \text{versus} \quad H_a : |\epsilon| < \delta \qquad (3)$$

where δ is the equivalence limit. Equivalence between the treatment and the reference can be established by testing the following two one-sided hypotheses:

$$\begin{array}{c} H_{01} : \epsilon \geq \delta \quad \text{versus} \quad H_{a1} : \epsilon < \delta \\ \text{and} \\ H_{02} : \epsilon \leq -\delta \quad \text{versus} \quad H_{a2} : \epsilon > -\delta \end{array} \qquad (4)$$

In other words, H_{01} and H_{02} are rejected and equivalence at the α level of significance is concluded if

$$\frac{\sqrt{n}(\bar{x} - \mu_0 - \delta)}{s} < -t_{\alpha,n-1} \quad \text{and}$$
$$\frac{\sqrt{n}(\bar{x} - \mu_0 + \delta)}{s} > t_{\alpha,n-1}$$

The power of the above test can be approximated by

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

Based on a similar argument as given in Chow and Liu (3, 4), the sample size needed to achieve the desired power of $1 - \beta$ is given by

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

2.4 An Example

Consider a clinical study for evaluation of a study drug for treatment of patients with hypertension. Each patient's diastolic blood pressure was measured at baseline and at post-treatment. The primary endpoint is post-treatment blood pressure change from baseline. Assuming the normal range of diastolic blood pressure is 80 mm Hg or below, a patient with a diastolic blood pressure higher than 90 mm Hg is considered as hypertension. The objective of the study is to show whether the study drug will decrease diastolic blood pressure from 90 mm Hg to the normal range of 85 mm Hg or below. Therefore, a minimum decrease of 5 mm Hg is considered as a clinically meaningful difference. Suppose from a pilot study that it is estimated that the standard deviation of the study drug is about 10 mm Hg ($\sigma = 10$). Thus, the sample size needed for achieving an 80% power at the 5% level of significance for detection of a clinically meaningful difference of 5 mm Hg is given by

$$n = \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\epsilon^2} = \frac{(1.96 + 0.84)^2 \times 10^2}{5^2}$$
$$= 31.36 \approx 32$$

On the other hand, suppose a standard therapy exists for treatment of hypertension in the marketplace. To show the superiority of the study drug as compared with the standard therapy, the sample size needed for achieving an 80% power at the 5% level of significance assuming a superiority margin of 2.5 mm Hg is given by

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\epsilon - \delta)^2} = \frac{(1.64 + 0.84)^2 \times 10^2}{(5 - 2.5)^2}$$
$$= 98.41 \approx 99$$

3 TWO-SAMPLE PARALLEL DESIGN

For testing two-sample from a parallel design, let x_{ij} be the responses of interest, which are obtained from the jth patient in the ith treatment group, $j = 1, \ldots, n_i,\ i = 1, 2$. It is assumed that $x_{ij},\ j = 1, \ldots, n_i,\ i = 1, 2$, are independent normal random variables with mean μ_i and variance σ^2. Define

$$\bar{x}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \quad \text{and}$$
$$s^2 = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\cdot})^2$$

where $\bar{x}_{i\cdot}$ and s^2 are the sample mean of the ith treatment and sample variance, respectively. Let $\epsilon = \mu_2 - \mu_1$ be the true mean difference between the treatment and the control. In practice, it is not uncommon to have an unequal sample size allocation between treatment groups. Let $n_1/n_2 = \kappa$ for some κ. When $\kappa = 2$, there is a 2:1 ratio between the treatment group and the control group.

3.1 Test for Equality

For testing equality between treatment groups, consider the hypotheses given in Equation (1). For a given significance level α, the null hypothesis H_0 of (1) is rejected if

$$\left| \frac{\bar{x}_{1\cdot} - \bar{x}_{2\cdot}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{\alpha/2, n_1+n_2-2}$$

The power of the above test can be approximated by

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

As a result, the sample size needed to achieve the desired power of $1 - \beta$ at the α level of significance is given by

$$n_1 = \kappa n_2 \quad \text{and} \quad n_2 = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2 (1 + 1/\kappa)}{\epsilon^2}$$

3.2 Test for Noninferiority/Superiority

For testing noninferiority/superiority, consider the hypotheses given in Equation (2). For a given significance level α, the null hypothesis H_0 is rejected if

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha, n_1+n_2-2}$$

Under the alternative hypothesis (i.e., $\epsilon > \delta$), the power of the above test can be approximated by

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

Hence, the sample size needed to achieve the desired power of $1 - \beta$ at the α level of significance is given by

$$n_1 = \kappa n_2 \quad \text{and} \quad n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2 (1 + 1/\kappa)}{(\epsilon - \delta)^2}$$

3.3 Test for Equivalence

For testing therapeutic equivalence, consider the two one-sided hypotheses given in Equation (4). For a given significance level α, the null hypothesis of inequivalence at the α level of significance is rejected if

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{\alpha, n_1+n_2-2} \quad \text{and}$$
$$\frac{\bar{x}_1 - \bar{x}_2 + \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha, n_1+n_2-2}$$

Under the alternative hypothesis (i.e., $|\epsilon| < \delta$), the power of this test can be approximated by

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

As a result, the sample size needed to achieve the desired power of $1 - \beta$ is given by

$$n_1 = \kappa n_2$$
$$n_2 = \frac{(z_\alpha + z_{\beta/2})^2 \sigma^2 (1 + 1/\kappa)}{\delta^2} \quad \text{if} \quad \epsilon = 0$$
$$n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2 (1 + 1/\kappa)}{(\delta - |\epsilon|)^2} \quad \text{if} \quad \epsilon \neq 0$$

3.4 An Example

Consider the same example as discussed in Section 3.3. Now assume that a parallel design will be conducted to compare the study drug with an active control agent. The objective is to establish the superiority of the study drug over the active control agent. If this objective cannot be achieved, then the sponsor will then make an attempt to establish equivalence between the study drug and the active control agent. According to historical data, the standard deviations for both the study drug and the active control agent are approximately 5 mm Hg ($\sigma = 5$). Assuming that the superiority margin is 3.5 mm

Hg ($\delta = 3.5$), the true difference between treatments is 3.0 mm Hg ($\epsilon = 3.0$). Then the sample size needed to achieve the desired power of 80% ($\beta = 0.20$) at the 5% ($\alpha = 0.05$) level of significance is given by

$$n = \frac{2\sigma^2(z_\alpha + z_\beta)^2}{(\epsilon - \delta)^2} = \frac{2 \times 5^2(1.64 + 0.84)^2}{(3.5 - 2.5)^2}$$
$$= 307.52 \approx 308$$

On the other hand, the sponsor may suspect that the study drug will not provide such a significant improvement of 3.0 mm Hg. Alternatively, the sponsor may want to show that the study drug is at least as good as the active control agent ($\delta = 0$). Therefore, instead of demonstrating superiority, the sponsor may want to establish equivalence between the study drug and the active control agent. Assume that a difference of 2.0 mm Hg ($\delta = 2.0$) is not considered of clinical importance. Therefore, the sample size needed to achieve the desired power of 80% ($\beta = 0.20$) at the 5% ($\alpha = 0.05$) level of significance for establishment of therapeutical equivalence between the study drug and the active control agent is given by

$$n = \frac{2\sigma^2(z_\alpha + z_{\beta/2})^2}{(\epsilon - \delta)^2} = \frac{2 \times 5^2(1.64 + 1.28)^2}{(2.0 - 0.0)^2}$$
$$= 106.58 \approx 107$$

4 TWO-SAMPLE CROSSOVER DESIGN

For testing two-sample from a crossover design, without loss of generality, consider a standard $2 \times 2m$ replicated crossover design comparing mean responses of a treatment and a control. Let y_{ijkl} be the lth response of interest ($l = 1, \ldots, m$) observed from the jth patient ($j = 1, \ldots, n$) in the ith sequence ($i = 1, 2$) under the kth treatment ($k = 1, 2$). The following linear mixed effects model is usually considered:

$$y_{ijkl} = \mu_k + \gamma_{ik} + s_{ijk} + e_{ijkl}$$

where μ_k is the effect due to the kth treatment, γ_{ik} is the fixed effect of the ith sequence under the kth treatment, and s_{ijk} is the random effect of the jth patient in the ith sequence under treatment k. It is further assumed that (s_{ij1}, s_{ij2}), $i = 1, 2$, $j = 1, \ldots, n$ i.i.d. bivariate normal random vectors with mean 0 and covariance matrix given by

$$\sum = \begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix}$$

where σ_{BT}^2 and σ_{BR}^2 are the intersubject variabilities under the test and reference, respectively; ρ is the intersubject correlation coefficient; and e_{ij1l} and e_{ij2l} are assumed to be independent normal random variables with mean 0. Depending on the treatment, e_{ij1l} and e_{ij2l} may have variance σ_{WT}^2 or σ_{WR}^2. Further define

$$\sigma_D^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR}$$

which is usually referred to as the variability from the subject-by-treatment interaction. Let $\epsilon = \mu_2 - \mu_1$ be the true mean difference between the treatment and the control. Define

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

An unbiased estimator for ϵ is given by

$$\hat{\epsilon} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} d_{ij}$$

Under our model assumption, $\hat{\epsilon}$ follows a normal distribution with mean ϵ and variance $\sigma_m^2/(2n)$, where

$$\sigma_m^2 = \sigma_D^2 + \frac{1}{m}(\sigma_{WT}^2 + \sigma_{WR}^2)$$

To estimate σ_m^2, the following estimator is useful:

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

where

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

4.1 Test for Equality

For testing equality, consider the hypothesis given in Equation (1). For a given significance level α, the null hypothesis H_0 of (1) is rejected if

$$\left| \frac{\hat{\epsilon}}{\hat{\sigma}_m/\sqrt{2n}} \right| > t_{\alpha/2,2n-2}$$

Under the alternative hypothesis (i.e., $\epsilon \neq 0$), the power of this test can be approximated by

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

Therefore, the sample size needed to achieve the desired power of $1-\beta$ at the α level of significance is given by

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

4.2 Test for Noninferiority/Superiority

For testing noninferiority/superiority, consider the hypotheses given in Equation (2). The null hypothesis H_0 of (2) at the α level of significance is rejected if

$$\frac{\hat{\epsilon} - \delta}{\hat{\sigma}_m/\sqrt{2n}} > t_{\alpha,2n-2}$$

Under the alternative hypothesis (i.e., $\epsilon > \delta$), the power of this test can be approximated by

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

As a result, the sample size needed to achieve the desired power of $1-\beta$ at the α level of significance is given by

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

4.3 Test for Equivalence

Similarly, therapeutic equivalence can be established by testing the two one-sided hypotheses given in Equation (4). For a given significance level α, the null hypothesis H_0 of inequivalence is rejected if

$$\frac{\sqrt{2n}(\hat{\epsilon} - \delta)}{\hat{\sigma}_m} < -t_{\alpha,2n-2} \quad \text{and}$$

$$\frac{\sqrt{2n}(\hat{\epsilon} + \delta)}{\hat{\sigma}_m} > t_{\alpha,2n-2}$$

Under the alternative hypothesis (i.e., $|\epsilon| < \delta$), the power of this test can be approximated by

$$\Phi\left(\frac{\sqrt{2n}(\delta - \epsilon)}{\sigma_m} - z_\alpha \right) + \Phi\left(\frac{\sqrt{2n}(\delta + \epsilon)}{\sigma_m} - z_\alpha \right) - 1$$

Thus, the sample size needed to achieve the desired power of $1-\beta$ at the α level of significance is given by

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

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

4.4 An Example

In the previous example, instead of using a parallel design, use a standard two-sequence two-period crossover design to compare the study drug and the active control agent. Then the standard deviation of intrasubject comparison is about 2.5 mm Hg ($\sigma_2 = 2.5$). And the mean difference between the study drug and the active control agent is about 1 mm Hg ($\delta = 1$). Then the sample size required for achieving an 80% power at the 5% ($\alpha = 0.05$) level of significance is given by

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma_2^2}{2\epsilon^2} = \frac{(1.96 + 0.84)^2 \times 2.5^2}{2 \times 1.0^2} = 24.5 \approx 25$$

5 MULTIPLE-SAMPLE ONE-WAY ANOVA

In this section, multiple samples from a parallel design comparing more than two treatments are considered. More specifically, let x_{ij} be the jth patient from the ith treatment group, $i = 1, \ldots, k, j = 1, \ldots, n$. It is assumed that

$$x_{ij} = A_i + \epsilon_{ij}$$

where A_i is the fixed effect of the ith treatment and ϵ_{ij} is a random error in observing x_{ij}. It is assumed that ϵ_{ij} are i.i.d. normal random variables with mean 0 and variance σ^2. Define

$$\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{x}_{i\cdot})^2$$

$$\text{SSA} = \sum_{i=1}^{k}(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2$$

where

$$\overline{x}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \quad \text{and} \quad \overline{x}_{\cdot\cdot} = \frac{1}{k}\sum_{i=1}^{k}\overline{x}_{i\cdot}$$

Then σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{\text{SSE}}{k(n-1)}$$

5.1 Pairwise Comparison

It is common practice to compare means between the pairs of treatments of interest. This type of problem can be formulated as the following hypotheses:

$$H_0 : \mu_i = \mu_j \quad \text{versus} \quad H_a : \mu_i \neq \mu_j \tag{5}$$

for some pairs (i, j). If all possible pairwise comparisons, are considered a total of $k(k-1)/2$ possible comparisons exists. It should be noted that multiple comparison will inflate the type I error. As a result, appropriate adjustment, such as the Bonferroni adjustment, should be made for the purpose of controlling the overall type I error rate at the desired significance level. Assume that τ is the number of pairwise comparisons of interest. The null hypothesis is rejected if

$$\left|\frac{\sqrt{n}(\overline{x}_{i\cdot} - \overline{x}_{j\cdot})}{\sqrt{2}\hat{\sigma}}\right| > t_{\alpha/(2\tau),k(n-1)}$$

The power of this test can be approximated by

$$\Phi\left(\frac{\sqrt{n}|\epsilon_{ij}|}{\sqrt{2}\sigma} - z_{\alpha/(2\tau)}\right)$$

where $\epsilon_{ij} = \mu_i - \mu_j$ is the true mean difference between the treatment i and j. As a result, the sample size needed to achieve the desired power of $1 - \beta$ is given by

$$n = \max\{n_{ij}, \text{for all interested comparison}\}$$

where n_{ij} is given by

$$n_{ij} = \frac{2(z_{\alpha/(2\tau)} + z_\beta)^2\sigma^2}{\epsilon_{ij}^2}$$

5.2 Simultaneous Comparison

Situations also exist where the interest is to detect any clinically meaningful difference between any possible treatment comparisons. Thus, the following hypotheses are usually considered:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$\text{versus} \quad H_a : \mu_i \neq \mu_j \text{ for some } 1 \leq i < j \leq k \tag{6}$$

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

$$F_A = \frac{n\text{SSA}/(k-1)}{\text{SSE}/[k(n-1)]} > F_{\alpha,k-1,k(n-1)}$$

where $F_{\alpha,k-1,k(n-1)}$ denote the α upper quantile of the F-distribution with $k-1$ and $k(n-1))$ degrees of freedom. As demonstrated by Chow et al. (5), under the alternative hypothesis, the power of the test can be approximated by

$$P(F_A > F_{\alpha,k-1,k(n-1)}) \approx P(n\text{SSA} > \sigma^2\chi^2_{\alpha,k-1})$$

where $\chi^2_{\alpha,k-1}$ represents the αth upper quantile for a χ^2 distribution with $k-1$ degrees of freedom. Under the alternative hypothesis, $n\text{SSA}/\sigma^2$ is distributed as a noncentral χ^2 distribution with degrees of freedom $k-1$ and noncentrality parameter $\lambda = n\Delta$, where

$$\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{k}(\mu_i - \overline{\mu})^2, \qquad \overline{\mu} = \frac{1}{k}\sum_{j=1}^{k}\mu_j$$

As a result, the sample size needed to achieve the desired power of $1-\beta$ can be obtained by solving

$$\chi^2_{k-1}(\chi^2_{\alpha,k-1}|\lambda) = \beta \tag{7}$$

where $\chi^2_{k-1}(\cdot|\lambda)$ is the cumulative distribution function of the noncentral χ^2 distribution with degrees of freedom $k-1$ and noncentrality parameter λ.

5.3 An Example

Assume that the sponsor wants to conduct a parallel trial to compare three drugs ($k = 3$) for treatment of patients with hypertension. The three treatments are a study drug, an active control agent, and a placebo. The primary endpoint is the diastolic blood pressure decreases from baseline. It is assumed that mean decrease for the three treatments are given by 5 mm Hg, 2.5 mm Hg, and 1.0 mm Hg, respectively ($\mu_1 = 5$, $\mu_2 = 2.5$, and $\mu_3 = 1.0$). A constant standard deviation of 5 mm Hg ($\sigma = 5$) is assumed for the three treatments. Then the sample size needed to achieve the desired power of 80% ($\beta = 0.20$) at the 5% ($\alpha = 0.05$) level of significance can be obtained by first finding the value of λ according to Equation (7), which is given by $\lambda = 9.64$. Therefore, the sample size required per treatment group is given by

$$n = \frac{\lambda}{\Delta} = \frac{9.64}{0.33} = 29.21 \approx 30$$

6 MULTIPLE-SAMPLE WILLIAMS DESIGN

As discussed, one advantage for adopting a crossover design in clinical research is that each patient can serve as his/her own control. As a result, intersubject variability can be removed during pairwise comparisons under appropriate assumptions. The U.S. Food and Drug Administration (FDA) has identified crossover design as the design of choice for bioequivalence trials. In practice, the standard two-sequence two-period crossover design is often used. However, it may not be useful when more than two treatments are compared. When more than two treatments are compared, it is desirable to compare pairwise treatment effects with the same degrees of freedom. In such a situation, Williams design is recommended [see, e.g., Chow and Liu, (3, 4)]. Under a Williams design, the following model is commonly used:

$$y_{ijl} = P_{j'} + \gamma_i + \mu_l + e_{ijl}, \quad i,l = 1,\ldots,k,$$
$$j = 1,\ldots,n$$

where y_{ijl} is the response of interest from the jth patient in the ith sequence under the lth treatment, $P_{j'}$ represents the fixed effect for the j' period, j' represents the number of the period for the ith sequence's lth treatment, $\sum_{j=1}^{a} P_j = 0$, γ_i is the fixed sequence effect, μ_j is the fixed treatment effect, and e_{ijl} is a normal random variable with mean 0 and variance σ^2_{il}. For fixed i and l, $e_{ijl}, j = 1,\ldots,n$ are independent and identically distributed. For fixed i and j, e_{ijl}, $l = 1,\ldots,a$ are usually correlated because they all come from the same patient.

Without loss of generality, suppose that the first two treatments are to be compared (i.e., treatments 1 and 2). Define

$$d_{ij} = y_{ij1} - y_{ij2}$$

Then, the true mean difference between treatments 1 and 2 can be estimated by the following unbiased estimator:

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

It can be shown that $\hat{\epsilon}$ is normally distributed with mean $\epsilon = \mu_1 - \mu_2$ and variance $\sigma^2_d/(kn)$, where σ_d^2 is defined as the variance of d_{ij} and can be estimated by

$$\hat{\sigma}_d^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(d_{ij} - \frac{1}{n} \sum_{j'=1}^{n} d_{ij'} \right)^2$$

6.1 Test for Equality

For testing equality, consider the hypotheses given in Equation (1). For a given significance level α, the null hypothesis H_0 of (1) is rejected if

$$\left| \frac{\hat{\epsilon}}{\hat{\sigma}_d/\sqrt{kn}} \right| > t_{\alpha/2,k(n-1)}$$

Under the alternative hypothesis (i.e., $\epsilon \neq 0$), the power of the test can be approximated by

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

The sample size needed to achieve the desired power of $1-\beta$ at the α level of significance is given by

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

6.2 Test for Noninferiority/Superiority

For testing for noninferiority/superiority, the hypotheses given in Equation (2) are considered. For a given significance level α, the null hypothesis H_0 of Equation (2) is rejected if

$$\frac{\hat{\epsilon} - \delta}{\hat{\sigma}_d/\sqrt{kn}} > t_{\alpha,k(n-1)}$$

Under the alternative hypothesis (i.e., $\epsilon > \delta$), the power of this test can be approximated by

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

As a result, the sample size needed to achieve the desired power of $1-\beta$ at the α level of significance is given by

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

6.3 Test for Equivalence

The equivalence between the treatment and the control can be established by testing the two one-sided hypotheses given in Equation (4). For a given significance level α, the null hypothesis of inequivalence at the α level of significance is rejected if

$$\frac{\sqrt{kn}(\hat{\epsilon} - \delta)}{\hat{\sigma}_d} < t_{\alpha,k(n-1)} \text{ and}$$
$$\frac{\sqrt{kn}(\hat{\epsilon} + \delta)}{\hat{\sigma}_d} > t_{\alpha,k(n-1)}$$

Under the alternative hypothesis (i.e., $|\epsilon| < \delta$), the power of the above test can be approximated by

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

Hence, the sample size needed for achieving the power of $1-\beta$ at the α level of significance is given by

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

6.4 An Example

Suppose a clinical trial is conducted with a standard 6 × 3 Williams design ($k = 6$) to compare a study drug, an active control agent, and a placebo. Assume that the mean difference between the study drug and the placebo is 3 mm Hg ($\epsilon = 3$) with a standard deviation for the intrasubject comparison of 5 mm Hg ($\sigma_d = 5$). Thus, the sample size required per sequence to achieve the desired power of 80% ($\beta = 0.20$) at the 5% ($\alpha = 0.05$) level of significance can be obtained as

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma_d^2}{k\epsilon^2} = \frac{(1.96 + 0.84)^2 \times 5^2}{6 \times 3^2} = 3.63 \approx 4$$

7 DISCUSSION

In clinical research, sample size calculation/ justification plays an important role to ensure the validity, success, and cost-effectiveness of the intended clinical trials. A clinical trial without sufficient sample size may not provide a desired *reproducibility probability* (6–8). In other words, the observed clinical results may not be reproducible at a certain level of significance. From a regulatory point of view, a large sample size is always preferred. However, from the sponsor's point of view, an unnecessarily large sample size is a huge waste of the limited resources in clinical research and development. Therefore, the objective of the sample size calculation is to select a minimum sample size for achieving a desired power at a prespecified significance level.

REFERENCES

1. ICH, *Harmonised Tripartite Guideline: Guideline For Good Clinical Practice*. International Conference on Harmonisation, 1996.
2. S. C. Chow and J. P. Liu, *Design and Analysis of Clinical Trials*. New York: Wiley, 1998.
3. S. C. Chow and J. P. Liu, *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker, 1992.
4. S. C. Chow and J. P. Liu, *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker, 2000.
5. S. C. Chow, J. Shao, and H. Wang, *Sample Size Calculation in Clinical Research*. New York: Marcel Dekker, 2003.
6. S. C. Chow and J. Shao, *Statistics in Drug Research*. New York: Marcel Dekker, 2002.
7. S. C. Chow, J. Shao, and O. Y. P. Hu, Assessing sensitivity and similariy in bridging studies. *Journal of Biopharmaceutical Statistics* 2002: **12**: 385–340.
8. S. C. Chow, Reproducibility probability in clinical research. *Encyclopedia of Biopharmaceutical Statistics* 2003; 838–849.