SAMPLE SIZE CALCULATION FOR COMPARING VARIABILITIES

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

In practice, the variabilities of responses involved in clinical research can be roughly divided into three categories. They are, namely, intra-subject, inter-subject, and total variabilities (1). More specifically, the intra-subject variability is the variability observed by repeated measurements on the same subject under exactly the same experiment conditions. This type of variability is very often due to measurement error, biological variability, and so on. In an ideal situation, the intra-subject variability can be eliminated by averaging infinite number of repeated observations from the same subject under the same experiment conditions. However, even if this averaging can be done, one can still expect difference to be observed in terms of the mean responses between different subjects, who receive exactly the same treatment. This difference is referred to as inter-subject variability, which is caused by the difference between subjects. The total variability is simply the sum of the intra-subject and inter-subject variabilities, which is the most often observable variability in a parallel design.

Statistical procedures for comparing intra-subject variabilities are well studied by Chichilli and Esinhart (2). The problem of comparing inter-subject and total variabilities are studied by Lee et al. (3, 4). A comprehensive summarization can be found in Lee et al. (5) and Chow et al. (6).

The rest of this entry is organized as follows. In the next section, sample size formulas for comparing intra-subject variabilities will be derived under both a replicated parallel and a crossover design. In Section 3, the problem of sample size calculation for comparing inter-subject variabilities will be studied. In Section 4, the formula for sample size determination based on comparing total variabilities of response between treatment groups will be presented. Finally, the entry is concluded with a discussion in Section 5.

2 COMPARING INTRA-SUBJECT VARIABILITIES

In order to be able to assess intra-subject variability, repeated measurements obtained from the same subject under the same experiment conditions are necessarily obtained. Thus, in practice, a simple parallel design without replicates or a standard 2×2 crossover design (TR, RT) with replicates but under different experiment conditions are often considered. In this section, only sample size calculation for comparing intra-subject variabilities of response between treatment groups under a replicated parallel design and a replicated crossover design are considered.

2.1 Parallel Design with Replicates

Consider a two-arm parallel trial with m replicates. For the jth subject in the ith treatment group, let x_{ijk} be the value obtained in the kth replicate. In practice, the following mixed effects model is usually considered:

$$x_{ijk} = \mu_i + S_{ij} + e_{ijk} \tag{1}$$

where μ_i is the mean response under the i treatment, S_{ij} is the inter-subject variability, and e_{ijk} is the intra-subject variability. It is also assumed that for a fixed i, S_{ij} are independent and identically distributed (i.i.d) as $N(0, \sigma_{Bi}^2)$, and e_{ijk}, are i.i.d $N(0, \sigma_{Wi}^2)$. Under this model, an unbiased estimator for σ_{Wi}^2 can be obtained as follows

$$\hat{\sigma}_{Wi}^2 = \frac{1}{n_i(m-1)} \sum_{j=1}^{n_i} \sum_{k=1}^{m} (x_{ijk} - \bar{x}_{ij\cdot})^2,$$

$$\text{where } \bar{x}_{ij\cdot} = \frac{1}{m} \sum_{k=1}^{m} x_{ijk} \tag{2}$$

2.1.1 Test for Equality. For testing equality in intra-subject variabilities of responses between treatment groups, the following hypotheses are usually considered:

$$H_0 : \sigma_{WT}^2 = \sigma_{WR}^2 \quad \text{versus} \quad H_a : \sigma_{WT}^2 \neq \sigma_{WR}^2$$

Based on Equation (2), the null hypothesis would be rejected at the α level of significance if

$$\frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} > F_{\alpha/2,n_T(m-1),n_R(m-1)} \quad \text{or}$$
$$\frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} < F_{1-\alpha/2,n_T(m-1),n_R(m-1)}$$

On the other hand, under the alternative hypothesis that $\sigma_{WT}^2 < \sigma_{WR}^2$, the power of the above test can be approximated by

$$P\left(F_{n_R(m-1),n_T(m-1)} > \frac{\sigma_{WT}^2}{\sigma_{WR}^2} F_{\alpha/2,n_R(m-1),n_T(m-1)}\right)$$

where $F_{n_R(m-1),n_T(m-1)}$ denotes an F-distributed random variable with $(n_R(m-1), n_T(m-1))$ degrees of freedom. Thus, under the assumption that $n = n_T = n_R$, the sample size needed for achieving the power of $1-\beta$ can be obtained by solving the following equation:

$$\frac{\sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F_{1-\beta,n(m-1),n(m-1)}}{F_{\alpha/2,n(m-1),n(m-1)}} \tag{3}$$

2.1.2 Test for Non-Inferiority/Superiority. From a statistical point of view, the problem of testing non-inferiority and superiority can be unified by the following hypotheses

$$H_0 : \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \geq \delta^2 \quad \text{versus} \quad H_a : \frac{\sigma_{WT}^2}{\sigma_{WR}^2} < \delta^2$$

where δ is the non-inferiority or superiority margin. As a result, the null hypothesis would be rejected at the α level of significance

$$\frac{\hat{\sigma}_{WT}^2}{\delta^2 \hat{\sigma}_{WR}^2} < F_{1-\alpha,n_T(m-1),n_R(m-1)}$$

On the other hand, under the alternative hypothesis that $\sigma_{WT}^2/\sigma_{WR}^2 < \delta$, the power of the above test procedure is given by

$$P\left(F_{n_R(m-1),n_T(m-1)} > \frac{\sigma_{WT}^2}{\delta^2 \sigma_{WR}^2} F_{\alpha,n_R(m-1),n_T(m-1)}\right)$$

Thus, assuming that $n = n_T = n_R$, the sample size required for achieving the power of $1-\beta$ can be obtained by solving the following equation:

$$\frac{\sigma_{WT}^2}{\delta^2 \sigma_{WR}^2} = \frac{F_{1-\beta,n(m-1),n(m-1)}}{F_{\alpha,n(m-1),n(m-1)}}$$

2.1.3 Test for Similarity. Similarity or equivalence between treatment groups can be established by testing the following hypotheses:

$$H_0 : \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \geq \delta^2 \text{ or } \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \leq 1/\delta^2 \quad \text{versus}$$
$$H_a : \frac{1}{\delta^2} < \frac{\sigma_{WT}^2}{\sigma_{WR}^2} < \delta^2$$

where $\delta > 1$ is the similarity limit. The above interval hypotheses can be partitioned into the following two one-sided hypotheses:

$$H_{01} : \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \geq \delta^2 \quad \text{versus} \quad H_{a1} : \frac{\sigma_{WT}^2}{\sigma_{WR}^2} < \delta^2$$
$$H_{02} : \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \leq 1/\delta^2 \quad \text{versus} \quad H_{a2} : \frac{\sigma_{WT}^2}{\sigma_{WR}^2} > 1/\delta^2$$

As a result, the null hypothesis of dissimilarity would be rejected and similarity would be concluded at the α level of significance if

$$\frac{\hat{\sigma}_{WT}^2}{\delta^2 \hat{\sigma}_{WR}^2} < F_{1-\alpha,n_T(m-1),n_R(m-1)} \quad \text{and}$$
$$\frac{\delta^2 \hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} > F_{\alpha,n_T(m-1),n_R(m-1)}$$

On the other hand, under the alternative hypothesis of similarity, a conservative approximation for the power can be obtained as follows (see, for example, Chow et al. (6))

$$1 - 2P\left(F_{n(m-1),n(m-1)} > \frac{\delta^2 \sigma_{WT}^2}{\delta_{WR}^2} F_{1-\alpha,n(m-1),n(m-1)}\right) \tag{4}$$

Hence, the sample size needed for achieving the power of $1-\beta$ for establishment of similarity or equivalence in intra-subject variabilities between treatment groups at the α level of significance can be obtained by solving the following equation:

$$\frac{\delta^2\sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F_{\beta/2,n(m-1),n(m-1)}}{F_{1-\alpha,n(m-1),n(m-1)}}$$

Note that detailed derivation of the above procedures for sample size calculation for comparing intra-subject variabilities can be found in Lee et al. (5) and Chow et al. (6).

2.1.4 An Example. Consider a two-arm parallel trial with three replicates ($m = 3$) comparing intra-subject variabilities of bioavailability of a test formulation with a reference formulation of a drug product. Based on a pilot study, it is estimated that the standard deviation for the test formulation is about 30% ($\sigma_{WT} = 0.30$) whereas the standard deviation for the reference formulation is about 45% ($\sigma_{WR} = 0.45$). The investigator is interested in selecting a sample size so that a significance difference in the intra-subject variability between the test and reference formulations can be detected at the 5% ($\alpha = 0.05$) level of significance with an 80% ($\beta = 0.20$) power. Thus, according to Equation (3), the sample size needed can be obtained by solving

$$\frac{0.20^2}{0.45^2} = \frac{F_{0.80,2n,2n}}{F_{0.025,2n,2n}}$$

which leads to $n = 25$. Hence, a total of 50 subjects (25 per arm) are needed in order to achieve the desired power for detecting such a difference in intra-subject variability between treatment groups.

2.2 Replicated Crossover Design

In this section, consider a $2 \times 2m$ crossover design comparing two treatments (Test and Reference) with m replicates. Let n_i be the number of subjects assigned to the ith sequence and x_{ijkl} be the response from the jth subject in the ith sequence under the lth replicate of the kth treatment ($k = T, R$). The following mixed effects model is usually considered for data from a $2 \times 2m$ crossover trial:

$$x_{ijkl} = \mu_k + \gamma_{ikl} + S_{ijk} + \epsilon_{ijkl} \qquad (5)$$

where μ_k is mean response of the kth formulation, γ_{ikl} is the fixed effect of the lth replicate under the kth treatment in the ith sequence with constraint

$$\sum_{i=1}^{2}\sum_{l=1}^{m}\gamma_{ikl} = 0$$

and S_{ijT} and S_{ijR} are the subject random effects of the jth subject in the ith sequence. $(S_{ijT}, S_{ijR})'$s are assumed i.i.d bivariate normal random vectors with mean $(0,0)'$. As S_{ijT} and S_{ijR} are two observations taken from the same subject, they are not independent from each other. The following covariance matrix between S_{ijT} and S_{ijR} is usually assumed to describe their relationship:

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

ϵ_{ijkl}'s are assumed i.i.d as $N(0, \Sigma_{Wk}^2)$. It is also assumed that $(S_{ijT}, S_{ijR})'$ and ϵ_{ijkl} are independent. Note that σ_{WT}^2 and σ_{BR}^2 are the inter-subject variances and σ_{WT}^2 and σ_{WR}^2 are intra-subject variances.

In order to obtain estimators for intra-subject variances, a new random variable z_{ijkl} is defined by an orthogonal transformation $\mathbf{z}_{ijk} = \mathbf{P}'\mathbf{x}_{ijk}$, where

$$\mathbf{x}_{ijk}' = (x_{ijk1}, x_{ijk2}, \ldots, x_{ijkm}),$$
$$\mathbf{z}_{ijk}' = (z_{ijk1}, z_{ijk2}, \ldots, z_{ijkm})$$

and P is an $m \times m$ orthogonal matrix with the first column given by $(1, 1, \ldots, 1)'/\sqrt{m}$. It can be verified that for a fixed i and any $l > 1, z_{ijkl}$ are i.i.d normal random variable with variance σ_{Wk}^2. Therefore, σ_{Wk}^2 can be estimated by

$$\hat{\sigma}_{Wk}^2 = \frac{1}{(n_1 + n_2 - 2)(m-1)} \times \sum_{l=2}^{m}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(z_{ijkl} - \bar{z}_{i\cdot kl})^2 \qquad \text{and}$$

$$\bar{z}_{i\cdot kl} = \frac{1}{n_i}\sum_{j=1}^{n_i} z_{ijkl} \tag{6}$$

It should be noted that $\hat{\sigma}^2_{Wk}/\sigma^2_{Wk}$ is a χ^2-distributed with $d = (n_1 + n_2 - 2)(m-1)$ degrees of freedom, and $\hat{\sigma}^2_{WT}$ and $\hat{\sigma}^2_{WR}$ are mutually independent. More details can be found in Chichilli and Esinhart (2) and Chow et al. (6).

2.2.1 Test for Equality. Similarly, consider the following hypotheses for testing equality in intra-subject variabilities of responses between treatment groups:

$$H_0 : \sigma^2_{WT} = \sigma^2_{WR} \quad \text{versus} \quad H_a : \sigma^2_{WT} \neq \sigma^2_{WR}$$

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

$$\frac{\hat{\sigma}^2_{WT}}{\hat{\sigma}^2_{WR}} > F_{\alpha/2,d,d} \text{ or } \frac{\hat{\sigma}^2_{WT}}{\hat{\sigma}^2_{WR}} < F_{1-\alpha/2,d,d}$$

On the other hand, under the alternative hypothesis that $\sigma^2_{WT} < \sigma^2_{WR}$, the power of the above test is given by

$$P\left(F_{d,d} > \frac{\sigma^2_{WT}}{\sigma^2_{WR}} F_{\alpha/2,d,d}\right)$$

Thus, assuming that $n = n_1 = n_2$, the sample size needed for achieving the power of $1 - \beta$ can be obtained by solving the following equation:

$$\frac{\sigma^2_{WT}}{\sigma^2_{WR}} = \frac{F_{1-\beta,(2n-2)(m-1),(2n-2)(m-1)}}{F_{\alpha/2,(2n-2)(m-1),(2n-2)(m-1)}} \tag{7}$$

2.2.2 Test for Non-Inferiority/Superiority. For testing non-inferiority and superiority, similarly, consider the following hypotheses:

$$H_0 : \frac{\sigma^2_{WT}}{\sigma^2_{WR}} \geq \delta^2 \quad \text{versus} \quad H_a : \frac{\sigma^2_{WT}}{\sigma^2_{WR}} < \delta^2$$

then reject the null hypothesis at the α level of significance if

$$\frac{\hat{\sigma}^2_{WT}}{\delta^2\hat{\sigma}^2_{WR}} < F_{1-\alpha,d,d}$$

On the other hand, under the alternative hypothesis that $\sigma^2_{WT}/\sigma^2_{WR} < \delta$, the power of the above test is given by

$$P\left(F_{d,d} > \frac{\sigma^2_{WT}}{\delta^2\sigma^2_{WR}} F_{\alpha,d,d}\right)$$

Hence, assuming that $n = n_1 = n_2$, the sample size required for achieving the power of $1 - \beta$ can be obtained by solving

$$\frac{\sigma^2_{WT}}{\delta^2\sigma^2_{WR}} = \frac{F_{1-\beta,(2n-2)(m-1),(2n-2)(m-1)}}{F_{\alpha,(2n-2)(m-1),(2n-2)(m-1)}}$$

2.2.3 Test for Similarity. For testing similarity, similarly, consider the following interval hypotheses:

$$H_0 : \frac{\sigma^2_{WT}}{\sigma^2_{WR}} \geq \delta^2 \text{ or } \frac{\sigma^2_{WT}}{\sigma^2_{WR}} \leq 1/\delta^2 \quad \text{versus}$$

$$H_a : \frac{1}{\delta^2} < \frac{\sigma^2_{WT}}{\sigma^2_{WR}} < \delta^2$$

where $\delta > 1$ is the equivalence limit. Testing the above interval hypotheses is equivalent to testing the following two one-sided hypotheses:

$$H_{01} : \frac{\sigma^2_{WT}}{\sigma^2_{WR}} \geq \delta^2 \quad \text{versus} \quad H_{a1} : \frac{\sigma^2_{WT}}{\sigma^2_{WR}} < \delta^2$$

$$H_{02} : \frac{\sigma^2_{WT}}{\sigma^2_{WR}} \leq 1/\delta^2 \quad \text{versus} \quad H_{a2} : \frac{\sigma^2_{WT}}{\sigma^2_{WR}} > 1/\delta^2$$

Thus, the null hypothesis of dissimilarity would be rejected and similarity would be concluded at the α level of significance if

$$\frac{\hat{\sigma}^2_{WT}}{\delta^2\hat{\sigma}^2_{WR}} < F_{1-\alpha,d,d} \quad \text{and} \quad \frac{\delta^2\hat{\sigma}^2_{WT}}{\hat{\sigma}^2_{WR}} > F_{\alpha,d,d}$$

On the other hand, under the alternative hypothesis of similarity, a conservative approximation to the power as given in Chow et al. (6) is given by

$$1 - 2P\left(F_{d,d} > \frac{\delta^2\sigma^2_{WT}}{\sigma^2_{WR}} F_{1-\alpha,d,d}\right)$$

Hence, assuming that $n = n_1 = n_2$, the sample size required for achieving the power of

$1-\beta$ can be obtained by solving the following equation:

$$\frac{\delta^2\sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F_{\beta/2,(2n-2)(m-1),(2n-2)(m-1)}}{F_{1-\alpha,(2n-2)(m-1),(2n-2)(m-1)}}$$

Note that the detailed derivation of the above procedures for sample size calculation for comparing intra-subject variabilities between treatment groups under a $2 \times 2m$ crossover design can be found in Lee et al. (5) and Chow et al. (6).

2.2.4 An Example. Consider the same example as described in the previous section. However, the study design is now a standard 2×6 crossover design (m = 3). According to Equation (6), the sample size needed for achieving an 80% power ($\beta = 0.20$) at the 5% ($\alpha = 0.05$) level of significance can be obtained by solving

$$\frac{0.30^2}{0.45^2} = \frac{F_{0.80,4(n-1),4(n-1)}}{F_{0.025,4(n-1),4(n-1)}}$$

which gives n = 14. As a result, a total of 28 subjects (14 subjects per sequence) are needed in order to achieve the desired power for detecting a 15% difference in intra-subject variability between treatment groups.

3 COMPARING INTER-SUBJECT VARIABILITIES

For comparing intra-subject variability, because an estimator for inter-subject variability can only be obtained under a replicated design and it usually can be expressed as a linear combination of various variance component estimates, its sampling distribution is relatively difficult to derive.

Howe (7), Graybill and Wang (8), and Hyslop et al. (9) developed various methods for estimation of inter-subject variabilities. One important assumption for the validity of these methods is that the variance component estimators involved in the estimation must be independent from one another. Lee et al. (3) generalized these methods for the situation where some variance components are actually dependent with one another. The sample size formulas introduced in this section are mostly based on the methods by Lee et al. (3).

3.1 Parallel Design with Replicates

Under the model in Equation (1), consider the following statistics:

$$s_{Bi}^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot})^2, \quad \text{where}$$

$$\bar{x}_{i\cdot\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}\bar{x}_{ij\cdot} \tag{8}$$

Thus, σ_{Bi}^2 can be estimated by

$$\hat{\sigma}_{Bi}^2 = s_{Bi}^2 - \frac{1}{m}\hat{\sigma}_{Wi}^2$$

where $\bar{x}_{ij\cdot}$ and $\hat{\sigma}_{Wi}^2$ are as defined in Equation (2).

3.1.1 Test for Equality. For testing equality in inter-subject variabilities of response between treatment groups, consider the following hypotheses:

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

where $\eta = \sigma_{BT}^2 - \sigma_{BR}^2$ and can be estimated by

$$\hat{\eta} = \hat{\sigma}_{BR}^2 - \hat{\sigma}_{BT}^2 = s_{BT}^2 - s_{BR}^2 - \hat{\sigma}_{WT}^2/m + \hat{\sigma}_{WR}^2/m$$

For a given significance level α, a $(1-\alpha)\times 100\%$ confidence interval for η can be obtained as $(\hat{\eta}_L, \hat{\eta}_U) = (\hat{\eta} - \sqrt{\Delta_L}, \hat{\eta} + \sqrt{\Delta_U})$, where

$$\Delta_L = s_{BT}^4\left(1 - \frac{n_T - 1}{\chi^2_{\alpha/2,n_T-1}}\right)^2 + s_{BR}^4\left(1 - \frac{n_R - 1}{\chi^2_{1-\alpha/2,n_R-1}}\right)^2 + \frac{\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_T(m-1)}{\chi^2_{1-\alpha/2,n_T(m-1)}}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_R(m-1)}{\chi^2_{\alpha/2,n_R(m-1)}}\right)^2$$

$$\Delta_U = s_{BT}^4\left(1-\frac{n_T-1}{\chi^2_{1-\alpha/2,n_T-1}}\right)^2 + s_{BR}^4\left(1-\frac{n_R-1}{\chi^2_{\alpha/2,n_R-1}}\right)^2 + \frac{\hat{\sigma}_{WT}^4}{m^2}\left(1-\frac{n_T(m-1)}{\chi^2_{\alpha/2,n_T(m-1)}}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2}\left(1-\frac{n_R(m-1)}{\chi^2_{1-\alpha/2,n_R(m-1)}}\right)^2$$

Thus, the null hypothesis would be rejected at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$. On the other hand, under the alternative hypothesis that $\eta \neq 0$ is true, the power of the above test can be approximated by

$$1-\Phi\left(z_{\alpha/2}-\frac{\sqrt{n}|\sigma_{BT}^2-\sigma_{BR}^2|}{\sigma^*}\right)$$

where

$$\sigma^{*2} = 2\left[\left(\sigma_{BT}^2+\frac{\sigma_{WT}^2}{m}\right)^2+\left(\sigma_{BR}^2+\frac{\sigma_{WR}^2}{m}\right)^2 + \frac{\sigma_{WT}^4}{m^2(m-1)}+\frac{\sigma_{WR}^4}{m^2(m-1)}\right]$$

As a result, the sample size needed for achieving the power of $1-\beta$ for detecting a meaningful difference in the inter-subject variability between treatment groups at the α level of significance is given by

$$n = \frac{\sigma^{*2}(z_{\alpha/2}+z_\beta)^2}{(\sigma_{BT}^2-\sigma_{BR}^2)^2}$$

3.1.2 Test for Non-Inferiority/Superiority. Similar to testing for non-inferiority/superiority in the intra-subject variability between treatment groups, the problem of testing non-inferiority and superiority in the inter-subject variability between treatment groups can also be unified by the following hypotheses:

$$H_0: \eta \geq 0 \text{ versus } H_a: \eta < 0$$

where $\eta = \sigma_{BT}^2 - \delta^2\sigma_{BR}^2$. For a given significance level α, its $(1-\alpha)\times 100\%$th upper confidence bound is given by $\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$, where

$$\Delta_U = s_{BT}^4\left(1-\frac{n_T-1}{\chi^2_{1-\alpha,n_T-1}}\right)^2 + \delta^4 s_{BR}^4\left(1-\frac{n_R-1}{\chi^2_{\alpha,n_R-1}}\right)^2 + \frac{\hat{\sigma}_{WT}^4}{m^2}\left(1-\frac{n_T(m-1)}{\chi^2_{\alpha,n_T(m-1)}}\right)^2 + \frac{\delta^4\hat{\sigma}_{WR}^4}{m^2}\left(1-\frac{n_R(m-1)}{\chi^2_{1-\alpha,n_R(m-1)}}\right)^2$$

Therefore, the null hypothesis would be rejected at the α level of significance if $\hat{\eta}_U < 0$. On the other hand, under the alternative hypothesis, the power of the above test can be approximated by

$$\Phi\left(-z_\alpha-\frac{\sqrt{n}(\sigma_{BT}^2-\delta^2\sigma_{BR}^2)}{\sigma^*}\right)$$

where

$$\sigma^{*2} = 2\left[\left(\sigma_{BT}^2+\frac{\sigma_{WT}^2}{m}\right)^2+\delta^4\left(\sigma_{BR}^2+\frac{\sigma_{WR}^2}{m}\right)^2 + \frac{\sigma_{WT}^4}{m^2(m-1)}+\frac{\delta^4\sigma_{WR}^4}{m^2(m-1)}\right]$$

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

$$n = \frac{\sigma^{*2}(z_\alpha+z_\beta)^2}{(\sigma_{BT}^2-\delta^2\sigma_{BR}^2)^2}$$

3.1.3 An Example. Consider a two-arm parallel design with three replicates ($m = 3$) for each patient. Suppose that the investor is interested in comparing the inter-subject variability of the pharmacokinetics parameters collected from the patients. Suppose from a pilot study it is estimated that $\sigma_{BT} = 0.35$,

$\sigma_{BR} = 0.45$, $\sigma_{WT} = 0.25$, and $\sigma_{WR} = 0.20$. It follows that

$$\sigma^{*2} = 2\left[\left(0.35^2 + \frac{0.25^2}{3}\right)^2 + \left(0.45^2 + \frac{0.20^2}{3}\right)^2 + \frac{0.25^4}{3^2(3-1)} + \frac{0.20^4}{3^2(3-1)}\right] = 0.126$$

Thus, the sample size needed for achieving an 80% power ($\beta = 0.20$) at the 5% level of significance ($\alpha = 0.05$) is given by

$$n = \frac{\sigma^{*2}(z_{\alpha/2} + z_\beta)^2}{(\sigma_{BT}^2 - \sigma_{BR}^2)^2} = \frac{0.126(1.96 + 0.84)^2}{(0.35^2 - 0.45^2)^2} \approx 155$$

Therefore, a total of 310 patients (155 patients per treatment group) are needed in order to achieve the desired power for detecting a 10% difference in the inter-subject variability between treatment groups at the α level of significance.

3.2 Replicated Crossover Design

Consider the model in Equation (4), the inter-subject variabilities can be estimated by

$$\hat{\sigma}_{BT}^2 = s_{BT}^2 - \frac{1}{m}\hat{\sigma}_{WT}^2 \text{ and } \hat{\sigma}_{BR}^2 = s_{BR}^2 - \frac{1}{m}\hat{\sigma}_{WR}^2$$

where $\bar{x}_{i.k.} = \sum_{j=1}^{n_i} \bar{x}_{ijk.}/n_i$, $n = n_1 + n_2$,

$$s_{BT}^2 = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{x}_{ijT.} - \bar{x}_{i.T.})^2 \quad \text{and}$$

$$s_{BR}^2 = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{x}_{ijR.} - \bar{x}_{i.R.})^2 \qquad (9)$$

3.2.1 Test for Equality. For testing equality in inter-subject variability of responses between treatment groups, similarly, consider the following hypotheses:

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

where $\eta = \sigma_{BT}^2 - \sigma_{BR}^2$ and can be estimated by

$$\hat{\eta} = \hat{\sigma}_{BT}^2 - \hat{\sigma}_{BR}^2 = s_{BT}^2 - s_{BR}^2 - \hat{\sigma}_{WT}^2/m + \hat{\sigma}_{WR}^2/m$$

According to Lee et al. (3), an approximate $(1-\alpha)$th confidence interval for η is given by $(\hat{\eta}_L, \hat{\eta}_U) = (\hat{\eta} - \sqrt{\Delta_L}, \hat{\eta} + \sqrt{\Delta_U})$, where

$$\Delta_L = \hat{\lambda}_1^2\left(1 - \frac{n_s - 1}{\chi^2_{\alpha/2,n_s-1}}\right)^2 + \hat{\lambda}_2^2\left(1 - \frac{n_s - 1}{\chi^2_{1-\alpha/2,n_s-1}}\right)^2 + \frac{\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi^2_{\alpha/2,n_s(m-1)}}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi^2_{1-\alpha/2,n_s(m-1)}}\right)^2$$

$$\Delta_U = \hat{\lambda}_1^2\left(1 - \frac{n_s - 1}{\chi^2_{1-\alpha/2,n_s-1}}\right)^2 + \hat{\lambda}_2^2\left(1 - \frac{n_s - 1}{\chi^2_{\alpha/2,n_s-1}}\right)^2 + \frac{\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi^2_{1-\alpha/2,n_s(m-1)}}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi^2_{\alpha/2,n_s(m-1)}}\right)^2$$

and $n_s = n_1 + n_2 - 2$. Thus, the null hypothesis would be rejected at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$. On the other hand, under the alternative hypothesis, the power of the above test can be approximated by

$$1 - \Phi\left(z_{\alpha/2} - \frac{\sqrt{n_s}|\sigma_{BT}^2 - \sigma_{BR}^2|}{\sigma^*}\right)$$

where

$$\sigma^{*2} = 2\left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m}\right)^2 + \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m}\right)^2 - 2\rho^2\sigma_{BT}^2\sigma_{BR}^2 + \frac{\sigma_{WT}^4}{m^2(m-1)} + \frac{\sigma_{WR}^4}{m^2(m-1)}\right]$$

Thus, the total sample size needed for achieving the power of $1-\beta$ is given by

$$n = \frac{\sigma^{*2}(z_{\alpha/2}+z_{\beta})^2}{(\sigma_{BT}^2-\sigma_{BR}^2)^2}+2$$

3.2.2 Test for Non-Inferiority/Superiority. For testing non-inferiority and superiority, similarly, consider the following hypotheses:

$$H_0:\eta\geq 0 \text{ versus } \quad H_a:\eta<0$$

where $\eta=\sigma_{BT}^2-\delta^2\sigma_{BR}^2$. For a given significance level of α, an approximate $(1-\alpha)$th upper confidence bound for η can be constructed as $\hat{\eta}_U=\hat{\eta}+\sqrt{\Delta_U}$, where

$$\begin{aligned}\Delta_U = {}& \hat{\lambda}_1^2\left(1-\frac{n_s-1}{\chi^2_{1-\alpha/2,n_s-1}}\right)^2 \\ & +\hat{\lambda}_2^2\left(1-\frac{n_s-1}{\chi^2_{\alpha/2,n_s-1}}\right)^2 \\ & +\frac{\hat{\sigma}_{WT}^4}{m^2}\left(1-\frac{n_s(m-1)}{\chi^2_{1-\alpha/2,n_s(m-1)}}\right)^2 \\ & +\frac{\delta^4\hat{\sigma}_{WR}^4}{m^2}\left(1-\frac{n_s(m-1)}{\chi^2_{\alpha/2,n_s(m-1)}}\right)^2\end{aligned}$$

and

$$\hat{\lambda}_i=\frac{s_{BT}^2-\delta^2 s_{BR}^2\pm\sqrt{(s_{BT}^2+\delta^2 s_{BR}^2)^2-4\delta^2 s_{BTR}^4}}{2}$$

Thus, the null hypothesis is rejected at the α level of significance if $\hat{\eta}_U<0$. On the other hand, under the alternative hypothesis, the power of the above test can be approximated by

$$\Phi\left(-z_{\alpha}-\frac{\sqrt{n_s}(\sigma_{BT}^2-\delta^2\sigma_{BR}^2)}{\sigma^*}\right)$$

where

$$\begin{aligned}\sigma^{*2} = 2\Bigg[& \left(\sigma_{BT}^2+\frac{\sigma_{WT}^2}{m}\right)^2+\delta^4\left(\sigma_{BR}^2+\frac{\sigma_{WR}^2}{m}\right)^2 \\ & -2\delta^2\rho^2\sigma_{BT}^2\sigma_{BR}^2+\frac{\sigma_{WT}^4}{m^2(m-1)} \\ & +\frac{\delta^4\sigma_{WR}^4}{m^2(m-1)}\Bigg]\end{aligned}$$

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

$$n=\frac{\sigma^{*2}(z_{\alpha}+z_{\beta})^2}{(\sigma_{BT}^2-\delta^2\sigma_{BR}^2)^2}+2$$

3.2.3 An Example. To illustrate the use of sample size formula derived above, consider a standard $2\times 4(m=2)$ crossover design (ABAB, BABA). The objective is to compare inter-subject variabilities of a test treatment and a control. From a pilot study, it is estimated that $\rho=0.65$, $\sigma_{BT}=0.35$, $\sigma_{BR}=0.45$, $\sigma_{WT}=0.25$, and $\sigma_{WR}=0.35$. Based on this information, it follows that

$$\begin{aligned}\sigma^{*2} = 2\Bigg[& \left(0.35^2+\frac{0.25^2}{2}\right)^2 \\ & +\left(0.45^2+\frac{0.35^2}{2}\right)^2 \\ & -2(0.65\times 0.35\times 0.45)^2 \\ & +\frac{0.25^4}{2^2}+\frac{0.35^4}{2^2}\Bigg]=0.115\end{aligned}$$

Hence, the sample size needed for achieving an 80% power ($\beta=0.20$) at the 5% ($\alpha=0.05$) level of significance is given by

$$n=\frac{0.115(1.96+0.84)^2}{(0.35^2-0.45^2)^2}+2\approx 143$$

As a result, a total of 143 subjects are needed for achieving the desired power.

4 COMPARING TOTAL VARIABILITIES

In practice, in addition to the intra-subject and inter-subject variability, the total variability is also of interest to researchers. The total variability is defined as the sum of the intra-subject and inter-subject variabilities. As the total variability is observable even in an experiment without replicates, in this section, both replicated and nonreplicated designs (parallel and crossover) will be discussed.

4.1 Parallel Designs Without Replicates

Consider a parallel design without replicates. In this case, the model in Equation (1) reduces to

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

where ϵ_{ij} is assumed to be i.i.d. as $N(0, \sigma_{Ti}^2)$. In this case, the total variability σ_{Ti}^2 can be estimated by

$$\hat{\sigma}_{Ti}^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\cdot})^2, \quad \text{where}$$

$$\bar{x}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

4.1.1 Test for Equality. For testing equality in total variability between treatment groups, the hypotheses become

$$H_0 : \sigma_{TT}^2 = \sigma_{TR}^2 \quad \text{versus} \quad H_a : \sigma_{TT}^2 \neq \sigma_{TR}^2$$

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

$$\frac{\hat{\sigma}_{TT}^2}{\hat{\sigma}_{TR}^2} > F_{\alpha/2, n_T-1, n_R-1} \quad \text{or}$$

$$\frac{\hat{\sigma}_{TT}^2}{\hat{\sigma}_{TR}^2} < F_{1-\alpha/2, n_T-1, n_R-1}$$

Under the alternative hypothesis that $\sigma_{TT}^2 < \sigma_{TR}^2$, the power of the above test is given by

$$P\left(F_{n_R, n_T} > \frac{\sigma_{TT}^2}{\sigma_{TR}^2} F_{\alpha/2, n_R-1, n_T-1}\right)$$

Thus, assuming that $n = n_R = n_T$, the sample size needed for achieving the power of $1 - \beta$ can be obtained by solving the following equation:

$$\frac{\sigma_{TT}^2}{\sigma_{TR}^2} = \frac{F_{1-\beta, n-1, n-1}}{F_{\alpha/2, n-1, n-1}}$$

4.1.2 Test for Non-Inferiority/Superiority. For testing non-inferiority and superiority, consider the following unified hypotheses:

$$H_0 : \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \geq \delta^2 \quad \text{versus} \quad \frac{\sigma_{TT}^2}{\sigma_{TR}^2} < \delta^2$$

where δ is the non-inferiority or superiority margin. Thus, the null hypothesis would be rejected at the α level of significance if

$$\frac{\hat{\sigma}_{TT}^2}{\delta^2 \hat{\sigma}_{TR}^2} < F_{1-\alpha, n_T, n_R}$$

On the other hand, under the alternative hypothesis, the power of the above test is given by

$$P\left(F_{n_R, n_T} > \frac{\sigma_{TT}^2}{\delta^2 \sigma_{TR}^2} F_{\alpha, n_R-1, n_T-1}\right)$$

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

$$\frac{\sigma_{TT}^2}{\delta^2 \sigma_{TR}^2} = \frac{F_{1-\beta, n-1, n-1}}{F_{\alpha, n-1, n-1}}$$

4.1.3 Test for Similarity. Similarly, similarity between treatment groups can be established by testing the following interval hypotheses:

$$H_0 : \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \geq \delta^2 \text{ or } \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \leq 1/\delta^2 \quad \text{versus}$$

$$H_a : \frac{1}{\delta^2} < \frac{\sigma_{TT}^2}{\sigma_{TR}^2} < \delta^2$$

where $\delta > 1$ is the similarity limit. As indicated earlier, testing the above interval hypotheses is equivalent to testing the following two one-sided hypotheses:

$$H_{01} : \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \geq \delta^2 \quad \text{versus} \quad H_{a1} : \frac{\sigma_{TT}^2}{\sigma_{TR}^2} < \delta^2$$

$$H_{02} : \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \leq 1/\delta^2 \quad \text{versus} \quad H_{a2} : \frac{\sigma_{TT}^2}{\sigma_{TR}^2} > 1/\delta^2$$

Thus, for a given significance level α, the null hypothesis of dissimilarity is rejected and the alternative hypothesis of similarity is accepted if

$$\frac{\hat{\sigma}_{TT}^2}{\delta^2 \hat{\sigma}_{TR}^2} < F_{1-\alpha, n_T, n_R} \quad \text{and} \quad \frac{\delta^2 \hat{\sigma}_{TT}^2}{\hat{\sigma}_{TR}^2} > F_{\alpha, n_T, n_R}$$

On the other hand, under the alternative hypothesis of similarity, a conservative approximation to the power is given by (see, for example, Chow et al. (6))

$$1 - 2P\left(F_{n-1,n-1} > \frac{\delta^2\sigma_{TT}^2}{\sigma_{TR}^2}F_{1-\alpha,n-1,n-1}\right)$$

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

$$\frac{\delta^2\sigma_{TT}^2}{\sigma_{TR}^2} = \frac{F_{\beta/2,n-1,n-1}}{F_{1-\alpha,n-1,n-1}}$$

4.1.4 An Example. Consider a two-arm parallel design comparing total variabilities of a test treatment with a reference treatment. It is estimated from a pilot study that $\sigma_{TT} = 0.55$ and $\sigma_{TR} = 0.60$. Suppose that the investigator wishes to establish noninferiority of the test treatment as compared with the reference treatment with a noninferiority margin of 10% ($\delta = 1.10$). The sample size needed for achieving an 80% ($\beta = 0.20$) power at the 5% ($\alpha = 0.05$) level of significance can be obtained by solving the following equation:

$$\frac{0.55^2}{1.10^2 \times 0.60^2} = \frac{F_{0.20,n-1,n-1}}{F_{0.05,n-1,n-1}}$$

which gives $n = 22$. Hence, a total of 44 subjects (22 subjects per treatment group) are needed.

4.2 Parallel Design with Replicates

In this section, focus is placed on the replicated parallel design as described in the model in Equation (1). Under the model in Equation (1), the total variabilities can be estimated by

$$\hat{\sigma}_{Ti}^2 = s_{Bi}^2 + \frac{m-1}{m}\hat{\sigma}_{Wi}^2$$

where s_{Bi}^2 is as defined in Equation (7) in Section 3.

4.2.1 Test for Equality. For testing equality in total variabilities of responses between treatment groups, consider the following statistical hypotheses that are usually considered:

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

where $\eta = \sigma_{TT}^2 - \sigma_{TR}^2$. η can be estimated by

$$\hat{\eta} = \hat{\sigma}_{TT}^2 - \hat{\sigma}_{TR}^2$$

For a given significance level α, an approximate $(1-\alpha)\times 100\%$ confidence interval of η can be constructed as $(\hat{\eta}_U, \hat{\eta}_L) = (\hat{\eta} - \sqrt{\Delta_L}, \hat{\eta} + \sqrt{\Delta_U})$, where

$$\begin{aligned}\Delta_L = {} & s_{BT}^4\left(1 - \frac{n_T - 1}{\chi^2_{\alpha/2,n_T-1}}\right)^2 \\ & + s_{BR}^4\left(1 - \frac{n_R - 1}{\chi^2_{1-\alpha/2,n_R-1}}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_T(m-1)}{\chi^2_{1-\alpha/2,n_T(m-1)}}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_R(m-1)}{\chi^2_{\alpha/2,n_R(m-1)}}\right)^2\end{aligned}$$

and

$$\begin{aligned}\Delta_U = {} & s_{BT}^4\left(1 - \frac{n_T - 1}{\chi^2_{1-\alpha/2,n_T-1}}\right)^2 \\ & + s_{BR}^4\left(1 - \frac{n_R - 1}{\chi^2_{\alpha/2,n_R-1}}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_T(m-1)}{\chi^2_{\alpha/2,n_T(m-1)}}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_R(m-1)}{\chi^2_{1-\alpha/2,n_R(m-1)}}\right)^2\end{aligned}$$

Thus, the null hypothesis is rejected at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$. On the other hand, under the alternative hypothesis, assuming that $n = n_T = n_R$, the power of the above test can be approximated by

$$1 - \Phi\left(z_{\alpha/2} - \frac{\sqrt{n}|\sigma_{TT}^2 - \sigma_{TR}^2|}{\sigma^*}\right)$$

where

$$\sigma^{*2} = 2\left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m}\right)^2 + \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m}\right)^2 + \frac{(m-1)\sigma_{WT}^4}{m^2} + \frac{(m-1)\sigma_{WR}^4}{m^2}\right]$$

Hence, the sample size needed for achieving the desired power of $1-\beta$ is given by

$$n = \frac{\sigma^{*2}(z_{\alpha/2} + z_\beta)^2}{(\sigma_{TT}^2 - \sigma_{TR}^2)^2}$$

4.2.2 Test for Non-Inferiority/Superiority. For testing non-inferiority/superiority, similarly, consider the following unified hypotheses:

$$H_0 : \eta \geq 0 \text{ versus } \quad H_a : \eta < 0$$

where $\eta = \sigma_{TT}^2 - \delta^2\sigma_{TR}^2$. For a given significance level of α, an approximate $(1-\alpha)$th upper confidence bound of η can be constructed as $\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$, where $\hat{\eta} = \hat{\sigma}_{TT}^2 - \delta^2\hat{\sigma}_{TR}^2$ and Δ_U is given by

$$\begin{aligned}\Delta_U = {} & s_{BT}^4\left(1 - \frac{n_T - 1}{\chi_{1-\alpha,n_T-1}^2}\right)^2 + \delta^4 s_{BR}^4\left(1 - \frac{n_R - 1}{\chi_{\alpha,n_R-1}^2}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_T(m-1)}{\chi_{\alpha,n_T(m-1)}^2}\right)^2 \\ & + \frac{\delta^4(m-1)^2\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_R(m-1)}{\chi_{1-\alpha,n_R(m-1)}^2}\right)^2\end{aligned}$$

Thus, the null hypothesis is rejected at the α level of significance if $\hat{\eta}_U < 0$. The power of the above test can be approximated by

$$\Phi\left(-z_\alpha - \frac{\sqrt{n}(\sigma_{TT}^2 - \delta^2\sigma_{TR}^2)}{\sigma^*}\right)$$

where

$$\sigma^{*2} = 2\left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m}\right)^2 + \delta^4\left(\sigma_{BR}^2 + \delta^4\frac{\sigma_{WR}^2}{m}\right)^2 + \frac{(m-1)\sigma_{WT}^4}{m^2} + \delta^4\frac{(m-1)\sigma_{WR}^4}{m^2}\right]$$

Hence, the sample size needed for achieving the desired power of $1-\beta$ is given by

$$n = \frac{\sigma^{*2}(z_\alpha + z_\beta)^2}{(\sigma_{TT}^2 + \delta^2\sigma_{TR}^2)^2}$$

4.2.3 An Example. Consider a two-arm parallel design with three replicates ($m = 3$) comparing total variabilities of responses between a test treatment and a control. It is assumed that $\sigma_{BT} = 0.35$, $\sigma_{BR} = 0.45$, $\sigma_{WT} = 0.25$, and $\sigma_{WR} = 0.35$. Suppose that one of the primary objective is to claim that a significance difference exists in total variabilities of responses between the test treatment and the control. It follows that

$$\sigma^{*2} = 2\left[\left(0.35^2 + \frac{0.25^2}{3}\right)^2 + \left(0.45^2 + \frac{0.35^2}{3}\right)^2 + \frac{(3-1)0.25^4}{3^2} + \frac{(3-1)0.35^4}{3^2}\right] = 0.168$$

Hence, the sample size needed for achieving an 80% power ($\beta = 0.20$) at the 5% ($\alpha = 0.05$) level of significance can be obtained as

$$n = \frac{0.168(1.96 + 0.84)^2}{(0.35^2 + 0.25^2 - (0.45^2 + 0.35^2))^2} \approx 68$$

As a result, a total of 136 subjects (68 subjects per treatment group) are needed.

4.3 The Standard 2 × 2 Crossover Design

Under the standard 2×2 crossover design, the notations defined in the model in Equation (4) can still be used. However, the subscript l is omitted as no replicate exists. Under the model in Equation (4), the total variability can be estimated by

$$\hat{\sigma}^2_{TT} = \frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ijT}-\bar{x}_{i\cdot T})^2 \quad \text{and}$$

$$\hat{\sigma}^2_{TR} = \frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ijR}-\bar{x}_{i\cdot R})^2$$

where

$$\bar{x}_{i\cdot T} = \frac{1}{n_i}\sum_{j=1}^{n_i}x_{ijT}, \quad \text{and} \quad \bar{x}_{i\cdot R} = \frac{1}{n_i}\sum_{j=1}^{n_i}x_{ijR}$$

4.3.1 Test for Equality. For testing equality in total variabilities between treatment groups, similarly, consider the following hypotheses:

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

where $\eta = \sigma^2_{TT} - \sigma^2_{TR}$. η be estimated by $\hat{\eta} = \hat{\sigma}^2_{TT} - \hat{\sigma}^2_{TR}$. Define

$$\sigma^2_{BTR} = \frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ijT}-\bar{x}_{i\cdot T}) \times (x_{ijR}-\bar{x}_{i\cdot R})$$

and

$$\hat{\lambda}_i = \frac{\hat{\sigma}^2_{TT} - \hat{\sigma}^2_{TR} \pm \sqrt{(\hat{\sigma}^2_{TT}+\hat{\sigma}^2_{TR})^2 - 4\hat{\sigma}^4_{BTR}}}{2}$$

Assume that $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$. According to Lee et al. (3), an approximate $(1-\alpha)\times 100\%$ confidence interval for η can be constructed by $(\hat{\eta}_L, \hat{\eta}_U) = (\hat{\eta} - \sqrt{\Delta_L}, \hat{\eta} + \sqrt{\Delta_U})$, where

$$\Delta_L = \hat{\lambda}_1^2\left(1 - \frac{n_1+n_2-2}{\chi^2_{1-\alpha/2,n_1+n_2-2}}\right)^2 + \hat{\lambda}_2^2\left(1 - \frac{n_1+n_2-2}{\chi^2_{\alpha/2,n_1+n_2-2}}\right)^2$$

$$\Delta_U = \hat{\lambda}_1^2\left(1 - \frac{n_1+n_2-2}{\chi^2_{\alpha/2,n_1+n_2-2}}\right)^2 + \hat{\lambda}_2^2\left(1 - \frac{n_1+n_2-2}{\chi^2_{1-\alpha/2,n_1+n_2-2}}\right)^2$$

Thus, the null hypothesis is rejected at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$. Under the alternative hypothesis, the power of the above test can be approximated by

$$1 - \Phi\left(z_{\alpha/2} - \frac{\sqrt{n_s}|\sigma^2_{TT} - \sigma^2_{TR}|}{\sigma^*}\right)$$

where

$$\sigma^{*2} = 2(\sigma^4_{TT} + \sigma^4_{TR} - 2\rho^2\sigma^2_{BT}\sigma^2_{BR})$$

The sample size needed for achieving the desired power of $1-\beta$ is then given by

$$n = \frac{\sigma^{*2}(z_{\alpha/2}+z_\beta)^2}{(\sigma^2_{TT}-\sigma^2_{TR})^2} + 2$$

4.3.2 Test for Non-Inferiority/Superiority. For testing non-inferiority/superiority, similarly, consider the following unified hypotheses:

$$H_0 : \eta \geq 0 \text{ versus } H_a : \eta < 0$$

where $\eta = \sigma^2_{TT} - \delta^2\sigma^2_{TR}$. For a given significance level of α, an approximate $(1-\alpha)$th upper confidence bound of η can be constructed as $\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$, where

$$\Delta_U = \hat{\lambda}_1^2\left(\frac{n_1+n_2-2}{\chi^2_{\alpha,n_1+n_2-2}} - 1\right)^2 + \hat{\lambda}_2^2\left(\frac{n_1+n_2-2}{\chi^2_{1-\alpha,n_1+n_2-2}} - 1\right)^2$$

and

$$\hat{\lambda}_i = \frac{\hat{\sigma}^2_{TT} - \delta^4\hat{\sigma}^2_{TR} \pm \sqrt{(\hat{\sigma}^2_{TT}+\delta^4\hat{\sigma}^2_{TR})^2 - 4\delta^2\hat{\sigma}^4_{BTR}}}{2}$$

Thus, the null hypothesis would be rejected at the α level of significance if $\hat{\eta}_U < 0$. On the other hand, under the alternative hypothesis, the power of the above test can be approximated by

$$\Phi\left(-z_\alpha - \frac{\sqrt{n}(\sigma^2_{TT} - \delta^2\sigma^2_{TR})}{\sigma^*}\right)$$

where

$$\sigma^{*2} = 2(\sigma_{TT}^4 + \delta^4\sigma_{TR}^4 - 2\delta^2\rho^2\sigma_{BT}^2\sigma_{BR}^2)$$

As a result, the sample size needed for achieving the power of $1-\beta$ can be obtained as

$$n = \frac{\sigma^{*2}(z_\alpha + z_\beta)^2}{(\sigma_{TT}^2 - \delta^2\sigma_{TR}^2)^2} + 2$$

4.3.3 An Example. Consider a standard 2×2 crossover design (TR, RT) comparing total variabilities of responses between a test treatment with a control. It is assumed that $\rho = 0.60$, $\sigma_{BT} = 0.35$, $\sigma_{BR} = 0.45$, $\sigma_{WT} = 0.25$, and $\sigma_{WR} = 0.35$. Suppose that one of the primary objectives is to detect a clinically significant difference. It follows that

$$\sigma^{*2} = 2\Big[(0.35^2 + 0.25^2)^2 + (0.45^2 + 0.35^2)^2 - 2 \times 0.60^2 \times 0.35^2 \times 0.45^2\Big] = 0.124$$

Hence, the sample size needed for achieving an 80% ($\beta = 0.20$) power at the 5% ($\alpha = 0.05$) level of significance is given by

$$n = \frac{0.124(1.96 + 0.84)^2}{(0.35^2 + 0.25^2 - (0.45^2 + 0.35^2))^2} + 2 \approx 52$$

As a result, a total of 52 subjects (e.g., 26 subjects per sequence) are needed for achieving the desired power.

4.4 Replicated 2 × 2*m* Crossover Design

Under the model in Equation (4), the total variabilities can be estimated by

$$\hat{\sigma}_{Tk}^2 = s_{Bk}^2 + \frac{m-1}{m}\hat{\sigma}_{Wk}^2 \quad k = T, R$$

where σ_{Wk}^2 is as defined in Equation (5) and s_{Bk}^2 is as defined in Equation (8).

4.4.1 Test for Equality. For testing equality, consider the following hypotheses:

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

where $\hat{\eta} = \hat{\sigma}_{TT}^2 - \hat{\sigma}_{TR}^2$. For a given significance level α, an approximate $(1-\alpha) \times 100\%$ confidence interval of η can be constructed by $(\hat{\eta}_L, \hat{\eta}_U) = (\hat{\eta} - \sqrt{\Delta_L}, \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U})$, where

$$\begin{aligned}\Delta_L = {} & \hat{\lambda}_1^2\left(1 - \frac{n_s - 1}{\chi_{1-\alpha/2,n_s-1}^2}\right)^2 + \hat{\lambda}_2^2\left(1 - \frac{n_s - 1}{\chi_{\alpha/2,n_s-1}^2}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi_{\alpha/2,n_s(m-1)}^2}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha/2,n_s(m-1)}^2}\right)^2\end{aligned}$$

and

$$\begin{aligned}\Delta_U = {} & \lambda_1^2\left(1 - \frac{n_s - 1}{\chi_{\alpha/2,n_s-1}^2}\right)^2 + \hat{\lambda}_2^2\left(1 - \frac{n_s - 1}{\chi_{1-\alpha/2,n_s-1}^2}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WT}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha/2,n_s(m-1)}^2}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WR}^4}{m^2}\left(1 - \frac{n_s(m-1)}{\chi_{\alpha/2,n_s(m-1)}^2}\right)^2\end{aligned}$$

and $\hat{\lambda}_i'$s are the same as those used for the test of equality for inter-subject variabilities. Thus, the null hypothesis would be rejected at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$. Under the alternative hypothesis, the power of the above test can be approximated by

$$1 - \Phi\left(z_{\alpha/2} - \frac{\sqrt{n}|\sigma_{TT}^2 - \sigma_{TR}^2|}{\sigma^*}\right)$$

where

$$\begin{aligned}\sigma^{*2} = 2\Bigg[& \left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m}\right)^2 + \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m}\right)^2 \\ & - 2\rho^2\sigma_{BT}^2\sigma_{BR}^2 \\ & + \frac{(m-1)\sigma_{WT}^4}{m^2} + \frac{(m-1)\sigma_{WR}^4}{m^2}\Bigg]\end{aligned}$$

Hence, the sample size needed for achieving the desired power of $1-\beta$ is given by

$$n = \frac{\sigma^{*2}(z_{\alpha/2}+z_\beta)^2}{(\sigma_{TT}^2-\sigma_{TR}^2)^2}.$$

4.4.2 Test for Non-Inferiority/Superiority. For testing non-inferiority/superiority, similarly, consider the following unified hypotheses:

$$H_0 : \eta \geq 0 \text{ versus } H_a : \eta < 0$$

where $\eta = \sigma_{TT}^2 - \delta^2\sigma_{TR}^2$. η can be estimated by $\hat{\eta} = \hat{\sigma}_{TT}^2 - \delta^2\hat{\sigma}_{TR}^2$. For a given significance level of α, an approximate $(1-\alpha)$th upper confidence bound of η can be constructed as $\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$, where

$$\begin{aligned}\Delta_U = {} & \hat{\lambda}_1^2\left(1-\frac{n_s-1}{\chi^2_{\alpha,n_s-1}}\right)^2 \\ & + \hat{\lambda}_2^2\left(1-\frac{n_s-1}{\chi^2_{1-\alpha,n_s-1}}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WT}^4}{m^2}\left(1-\frac{n_s(m-1)}{\chi^2_{1-\alpha,n_s(m-1)}}\right)^2 \\ & + \frac{(m-1)^2\hat{\sigma}_{WR}^4}{m^2}\left(1-\frac{n_s(m-1)}{\chi^2_{\alpha,n_s(m-1)}}\right)^2\end{aligned}$$

and $\hat{\lambda}_i'$s are same as those used for the test of non-inferiority for inter-subject variabilities. Thus, the null hypothesis would be rejected at the α level of significance if $\hat{\eta}_U < 0$. On the other hand, under the alternative hypothesis, the power of the above test can be approximated by

$$\Phi\left(-z_\alpha - \frac{\sqrt{n_s}(\sigma_{TT}^2-\delta^2\sigma_{TR}^2)}{\sigma^*}\right)$$

where

$$\begin{aligned}\sigma^{*2} = 2\Bigg[& \left(\sigma_{BT}^2+\frac{\sigma_{WT}^2}{m}\right)^2 + \delta^4\left(\sigma_{BR}^2+\frac{\sigma_{WR}^2}{m}\right)^2 \\ & - 2\delta^2\rho^2\sigma_{BT}^2\sigma_{BR}^2 \\ & + \frac{(m-1)\sigma_{WT}^4}{m^2} + \frac{\delta^4(m-1)\sigma_{WR}^4}{m^2}\Bigg]\end{aligned}$$

Hence, the sample size needed for achieving the desired power of $1-\beta$ is given by

$$n = \frac{\sigma^{*2}(z_\alpha+z_\beta)^2}{(\sigma_{TT}^2-\delta^2\sigma_{TR}^2)^2} + 2$$

4.4.3 An Example. Consider a 2×4 crossover design (ABAB,BABA) for comparing total variabilities between two formulations (A and B) of a drug product. It is estimated from a pilot study that $\rho = 0.65$, $\sigma_{BT} = 0.35$, $\sigma_{BR} = 0.45$, $\sigma_{WT} = 0.25$, and $\sigma_{WR} = 0.35$. Suppose the objective is to detect a significant difference in total variabilities between the two formulations. It follows that

$$\begin{aligned}\sigma^{*2} = 2\Bigg[& \left(0.35^2+\frac{0.25^2}{2}\right)^2 \\ & + \left(0.45^2+\frac{0.35^2}{2}\right)^2 \\ & - 2\times(0.65\times 0.35\times 0.45)^2 \\ & + \frac{0.25^4}{2^2}+\frac{0.35^4}{2^2}\Bigg] = 0.150\end{aligned}$$

Hence, the sample size needed for achieving an 80% (β = 0.20) is given by

$$\begin{aligned}n = {} & \frac{(0.150)(1.96+0.84)^2}{(0.35^2+0.25^2-(0.45^2+0.35^2))^2} \\ & + 2 \approx 60\end{aligned}$$

As a result, a total of 60 subjects (e.g., 30 subjects per sequence) are needed for achieving the desired power.

5 DISCUSSION

In clinical research, in addition to comparing mean responses, it is also of interest to compare the variabilities associated with the responses. As indicated in Chow and Shao (10), the treatment with a larger variability may be a safety concern. In addition, the treatment with a larger variability may have a small probability of reproducibility of the clinical responses. In this article, statistical procedures for sample size calculation are derived under a parallel design and a

crossover design with and without replicates. However, it should be noted that if the variance estimator is a linear component of several variance components, how to establish the similarity in variabilities between two treatments is still a challenging problem to researchers. Further research is needed.

REFERENCES

1. S. C. Chow and H. Wang, On sample size calculation in bioequivalence trials. *J. Pharmacokinet. Pharmacodynam.* 2001; **28**: 155–169.
2. V. M. Chichilli and J. D. Esinhart, Design and analysis of intra-subject variability in cross-over experiments. *Stat. Med.* 1996; **15**: 1619–1634.
3. Y. Lee, J. Shao, and S. C. Chow, Confidence intervals for linear combinations of variance components when estimators of variance components are dependent: an extension of the modified large sample method with applications. *J. Amer. Stat. Assoc.*, in press.
4. Y. Lee, J. Shao, S. C. Chow, and H. Wang, Test for inter-subject and total variabilities under crossover design. *J. Biopharmaceut. Stat.* 2002; **12**: 503–534.
5. Y. Lee, H. Wang, and S. C. Chow, Comparing variabilities in clinical research. *Encycl. Biopharmaceut. Stat.* 2003: 214–230.
6. S. C. Chow, J. Shao, and H. Wang, *Sample Size Calculation in Clinical Research*. New York: Marcel Dekker, 2003.
7. W. G. Howe, Approximate confidence limits on the mean of $X + Y$ where X and Y are two tabled independent random variables. *J. Amer. Stat. Assoc.* 1974; **69**: 789–794.
8. F. Graybill and C. M. Wang, Confidence intervals on nonnegative linear combinations of variances. *J. Amer. Stat. Assoc.* 1980; **75**: 869–873.
9. T. Hyslop, F. Hsuan, and D. J. Holder, A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* 2000; **19**: 2885–2897.
10. S. C. Chow and J. Shao, *Statistics in Drug Research*. New York: Marcel Dekker, 2002.