Sample Size and Power Calculation for High Order Crossover Designs

Roger P. Qu, Ph.D
Department of Biostatistics
Forest Research Institute, New York, NY, USA

1. Introduction

Sample size and power calculation is one of the most common procedures in statistical applications. For most common designs, such task can be performed readily with existing statistical software or by simple formulae presented in standard statistics books. With complex designs, it is not straightforward.

Crossover designs (Jones and Kenward, 1989) are the primary statistical designs for bioavailability and bioequivalence studies (Chow and Liu, 2000; Westlake, 1986). Unlike parallel designs in which each subject (or patient) receives one and only one of several treatments studied, in crossover studies each subject receives two or more treatments during the entire study. The order in which each subject receives the treatments depends on the particular design of a study. The simplest crossover design is the well known 2x2 design. In this design, each subject receives two different treatments, denoted as A and B usually. Half of the subjects receive A first and then B, and the other half receive B first and then A. Crossover designs allow comparison of individual treatments using within subject variation, and thus increase the power of the study. The disadvantage of crossover designs is the potential carry over effect which requires special designs for proper statistical inferences, such as Williams designs. In general, the power of equivalence test involves non-central t-distribution (Johnson and Kotz, 1970), and therefore numerical integration(Owen, 1965) or approximation has to be used. For a 2x2 crossover design with two treatments and two periods, the sample size can be obtained using the statistical software nQuery or the results of Diletti, Hauschke and Steinijans (DHS, 1991).

This paper introduces methods of Qu and Zheng (2001 and 2002) for the sample size and power determination for high order crossover designs . The methods are for Williams designs, which provide the minimum variance unbiased estimate (Williams, 1949; Jones and Kenward, 1989) even with the presence of carry over effect in the model, and presented assuming no carry over effect in the model. Although unbiased statistical inferences can still be performed with the presence of carry over effect, in most applications of crossover designs, an adequate wash-out period is embedded in the designs to eliminate any possible carry over effect. However, the methods can be easily extended to models with first order carry over effect. The only difference between models with and without carry over effect is the non-centrality parameter in the non-central t-distribution and the degrees of freedom in the estimate of the within subject variation (Yeh and Chen, 1995).

The underlying models for the primary response in pharmacokinetic studies such as *AUC* or *Cmax* of concentration of the study drug are often multiplicative, and thus logarithmic transformation is used first for statistical inference (FDA Guidance, 2001). Such transformation also circumvents the dependence of the equivalence criteria on the unknown true reference treatment effect, and thus allows treatment comparisons directly to be based on the treatment mean differences. The results in this paper are presented for such transformed data.

To demonstrate equivalence, one can either use the 90% confidence interval (CI) approach (Westlake, 1972 and 1976), or the two one-sided hypothesis testing approach (Shuirmann, 1981 and 1987; Anderson and Hauck, 1983) which is equivalent to the CI approach. Let Δ be the true difference between two treatments used for sample size calculation, Δ_l, Δ_u be the pre-specified boundaries for equivalence of two treatment effects. For the 90% CI approach, equivalence is demonstrated if the 90% confidence interval for Δ falls into this interval. For the hypothesis approach, two one-sided null hypotheses are considered

$$H01: \Delta \le \Delta_l \quad \text{vs} \quad Ha1: \Delta > \Delta_l,$$

and

$$H02: \Delta \ge \Delta_u \quad \text{vs} \quad Ha2: \Delta < \Delta_u.$$

Equivalence is accepted if both null hypotheses are rejected at a significance level 0.05.

In most bioequivalence studies, Δ_l= log(0.8)=–0.223 and Δ_u =log(1.25)=0.223, or Δ_l= log(0.7)= –0.36 and Δ_u =log(1.43)=0.36, depending on the primary criteria of the studies (FDA Guidance, 2001). The boundaries –0.223 and 0.223 are chosen based on the requirement of being within (0.8, 1.25) for the ratio of the means of the un-transformed data for the reference and test formulations to be accepted as equivalence. Similarly, -0.36 and 0.36 correspond to range (0.7, 1.43) for the ratio for un-transformed data.

Let n be the number of subjects per sequence, t the number of periods, k the number of sequences and N=k×n the total sample size. The general form of a statistical model for a crossover design without carry over effect is a mixed effect model. Let σ_s^2 and σ_e^2 be the inter-subject and intra-subject variance respectively; and $\hat{\Delta}$ be the least squares estimate of Δ. Then $\hat{\Delta}$ is normally distributed with mean Δ and variance $2\sigma_e^2$ /N. Let $\hat{\sigma}_e^2$ be the estimate of σ_e^2, α=0.05 and $t_{\alpha,\nu}$ the upper α percentile of t-distribution with degrees of freedom ν=(N-2)(t-1).

2. Sample Size for 2x2 Crossover Designs

For a 2x2 crossover design, sample size calculation can be performed using existing software such as nQuery (1999), or results in the literature such as DHS (1991).

2.1 Using Statistical Software nQuery

Table 1 displays the layout of the output for sample size calculation from nQuery for logarithmically transformed data, where $\sigma=\sigma_e/\sqrt{2}$. For example, with Δ_l =–0.223 and Δ_u=0.223, and assuming Δ=0, then 8 and 16 subjects are required per sequence respectively for σ_e=0.2 and 0.3 (σ=0.14 and 0.21) to have 80% power of demonstrating equivalence.

On the other hand, when provided with all other information in Table 1, nQuery can perform the power calculation also. For instance, if all information other than the power in the first configuration is entered, then nQuery will present 80% power as the result. We will use this feature below for sample size calculation for high order crossover designs.

2.2 Using Tabulated Results of Diletti, Hauschke and Steinijans

Using statistical software such as nQuery is certainly a very convenient and powerful tool for sample size and power calculation. Nevertheless, one can also use the tabulated results of DHS (1991) for the calculation, albeit with the price of incompleteness for all possible configurations. With Δ_l = -0.223 and Δ_u=0.223, DHS (1991) presented the power and total sample size for selected configurations of the parameters. For example, for the first configuration in Table 1, their calculation shows a total of 16 subjects (thus 8 subjects per group) for 80% power, same as that given by nQuery.

Table 1. nQuery Output (Module MTE1)
Two one-sided equivalence tests (TOST) for two-group or crossover design

Configuration No.	1	2
Test significance levels, α (one-sided)	0.05	0.05
Lower equivalence limit, Δ_l	-0.223	-0.223
Upper equivalence limit, Δ_u	0.223	0.223

Expected difference, Δ	0	0
Common standard deviation, σ	0.14	0.21
Power (%)	80	80
n per group	8	16

3. Sample Size for High Order Crossover Designs

With high order crossover designs, both nQuery and DHS' results are not applicable directly for sample size and power calculation. Although one could start from the very basics for such calculation using numeric integration for non-central t-distribution, such task can not be easily carried out. Moreover, the needed package perhaps is not existent for most applicants. Qu and Zheng (2001) introduced an exact method and two approximate methods using the foregoing results for the 2x2 designs for sample size and power calculation.

The key to their approach is the underlying probability expression for the power of the equivalence test.

3.1 *Calculation of Exact Power*

In contrast to the traditional sample size calculation by equating the required sample size explicitly to a function of the treatment effect and power etc., the exact method calculates the power (probability of having a significant test result) given the sample size and other information needed.

For high order designs, the power expression is similar to that for 2x2 designs except that the degrees of freedom of the chi-squared distribution is $\nu=(N-2)(t-1)$ instead of N-2 for 2x2 designs. However, using transformation $N^*=(t-1)N-2(t-2)$ and $\sigma^*=\sigma_e\sqrt{N^*/N}$, the power expression for the high order designs is exactly the same as for 2x2 designs. Therefore, the sample size for the high order designs can be calculated using the methods for 2x2 designs as described following.

Recall that given all information in Table 1, nQuery performs power calculation as illustrated in Section 2.1. This amounts to providing all information (σ, n(=N/2), Δ, Δ_l and Δ_u), and nQuery can then calculate power. All nQuery does is to calculate the probability (power) given the information needed for the calculation. Because the power expression for the high order designs is exactly the same as for 2x2 designs once using N^* and σ^*, one can use nQuery to calculate the probability given (σ^*, $N^*/2$, Δ, Δ_l and Δ_u), and thus sample size can be examined based on power calculation.

Consider as an example a 4x4 Williams design with 80% power at $\Delta=0$. Suppose again that $\Delta_l=$ -0.223 and $\Delta_u=0.223$ and $\sigma_e=0.20$. Then Table 2 below shows the result of the sample size determination. For this particular example, ten subjects per sequence (n=10) gives way too high power, four subjects per sequence gives slightly more power than the required, and power for three subjects per sequence is too low. Thus four subjects per sequence is required, which leads to a total sample size 16.

Table 2. Summary of Sample Size Determination by Exact Method for 2 4x4 Williams Design ($\alpha=0.05$, power=80% at $\Delta=0$, $\Delta_l=$ -0.223 and $\Delta_u=0.223$; $\sigma_e=0.20$)

n	N	N*	σ^*	N*/2 (number of subjects per sequence for nQuery calculation)	$\sigma=\sigma^*/\sqrt{2}$ (for nQuery calculation)	Power
10	40	116	0.338	58	0.238	>99%
8	32	92	0.339	46	0.239	>99%
4	16	44	0.332	22	0.234	83%
3	12	32	0.326	16	0.231	64%

3.2 Approximate Methods

The forgoing sample size determination is based on nQuery or other statistical software which can calculate the exact power for 2x2 designs. When such software is not available, the exact method can not be used. Instead, Qu and Zheng (2002) introduced two approximate conservative methods for sample size calculation, which can be used in conjunction with the published tabulated results for 2x2 designs, such as in DHS (1991).

3.2.1 N Approach*

This approach ignores the difference in degrees of freedom between a 2x2 design and a high-order design and replaces N* by N, that is to replace ν by N-2 in the power expression of the high order design. The resulting sample size is a conservative approximation. Intuitively, replacing N* (> N) by N amounts to ignoring extra degrees of freedom resulting from the higher order than 2 of the design and thus leads to some loss of information.

3.2.2 σ Approach*

Instead of approximating the degrees of freedom, the second approach approximates the intra-subject variation by replacing σ_e using $\sigma^* = \sigma_e\sqrt{t-1}$ in the power expression. For large N, N*/N=(t-1)-2(t-2)/N $\approx$ (t-1), and thus this approach gives about the same result as the exact method when N is large. With this replacement, the power probability for the high order designs has exactly the same form as the power for a 2x2 design. Therefore any existing results such as DHS (1991) for 2x2 designs can be used to find out N* and in turn the required sample size N= [N*+2(t–2)]/(t–1).

This approach transforms σ_e to σ^* by multiple of $\sqrt{t-1}$. When the value of σ_e is not small, σ^* could be out of the value range for the use of the tabulated results, such as in DHS (1991) which has a range of about 0.05 to 0.30. When this happens, the N* approach can be considered instead.

3.2.3 Examples

Table 3 presents the approximation results for several configurations for Williams designs with three treatments. For each design, the top panel shows the results of the σ^* approach in conjunction with the use of tabulated results of DHS (1991); the bottom half presents the approximation results using the N* approach. As can be seen from the table, for the examples considered, the two approximation approaches yield exactly the same results when both applicable. The sample size is always overestimated with a varying degree of overestimation. With two or three subjects per sequence, overestimation should not be a serious concern compared to being under power; with four or more subjects per sequence, the approximation performs well. As indicated in Table 3, the study could be very likely underpowered if it is run with the estimated sample size minus one (n-1) subjects per sequence.

Note that the approximate methods should be applied only when the needed tool, nQuery or other statistical software, for their exact method is not available. The two approximate methods complement each other in the sense that one method may be usable while the other is not when the value of intra-subject variation is out of value range for tabulated results. In addition, when both are usable, one can compare the calculated sample sizes and choose the minimum.

Table 3. Examples of Sample Size and Power Calculation --- 3x 6 Williams Designs (Three Treatments, α=0.05, power=80% at Δ=0, Δ_l= –0.223 and Δ_u=0.223)

Configuration No.	1	2	3	4
σ* Approach				
σ_e	0.10	0.15	0.20	0.25[a]
$\sigma^* = \sigma_e\sqrt{2}$	0.14	0.21	0.28	0.35
N*	10	20	32	
N=(N*+2)/2	6	11	17	
Rounding N	6	12	18	
n per sequence	1	2	3	
*N** Approach				
N	6	12	16	24
Rounding N	6	12	18	24
n per sequence	1	2	3	4
Power				
Power with estimated sample size n	88	91	87	81
Power with sample size (n-1)	NA	40	62	63

a. Value of σ* is out of the value range for the use of tabulated results in DHS (1991).

4. Discussion

The approximation methods are conservative, i.e. always over estimate the sample size when power is considered, and such over estimation is exacerbated especially for the final total sample size which has to be a multiple of the number of sequences for Williams designs. However, this over estimation was observed basically only when the total sample size was small. The exact power method is more accurate than the approximation. Therefore, the approximate methods might be used only when one does not have proper software to calculate the exact probability.

The sample size calculation described in the paper is for comparisons of two treatments. However, although there are more than two treatments in a high order Williams design, the primary inferences are most likely for such pair-wise comparisons, and thus sample size calculation should be based on such comparisons.

References:

1. Chow, S.C. and Liu, J.P. (2000*). Design and Analysis of Bioavailability and Bioequivalence studies.* Marcel Dekker, New York.
2. Anderson, S. and Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Commun. Stat. Theory Methods*, 12, 26763- 2692.
3. Diletti, E, Hauschke, E.D. and Steinijans, V.W. (1991). Sample size determination for bioequivalence assessment by means of confidence intervals*. Int. J. Clin. Pharmacol. Ther.Toxicol.*, 29, 1-8.
4. Jones, B and Kenward, M.G. (1989). *Design and Analysis of Crossover Trials*. Chapman and Hall, London, UK.
5. Johnson, N. L. and Kotz, A. (1970). *Continuous Univariate Distributions*. John Wiley, New York.
6. *Nquery Advisor*, Version 3.0. (1999). Statistical Solutions Ltd., Boston, USA.
7. Owen, D.B. (1965). A sepcial case of a bivariate non-central t-distribution, *Biometrrika*, 52, 437-446.

8. Qu, R.P. and Zheng, H. (2001). Exact power and sample size calculation for bioequivalence studies with high order crossover designs. *Proceedings of the Annual Meeting of the American Statistical Association,* August 5-9.
9. Qu, R.P. and Zheng, H. (2002). Sample size calculation for bioequivalence studies with high-order crossover designs. Controlled Clinical Trials, accepted.
10. Schurimann, D.J. (1981). On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval. *Biometrics,* 37, 617 [abstract].
11. Shurimann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence if average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics,* 15, 657-680.
12. US Fodd and Drug Administration (FDA) (2001) Guidance for Industry, Statistical Approaches to Establishing Bioequivalence. Internet: *http://www.fda.gov/cder/guideance/index.htm*.
13. Westlake, W.J. (1972). Use of confidence intervals in analysis of comparative bioavailibity trials. *J. Pharm. Sci*., 61. 1340-1341.
14. Westlake, W.J. (1976) Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, 32, 741-744.
15. Westlake, W.J. (1986). Bioavailability and bioequivalence of pharmaceutical formulations. *Biopharmaceutical Statistics for Drug Development,* K. Peace, ed. Marcel Dekker, New York, 329-352.
16. Williams, E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatment. *Aust. J. Sci. Res*., 2, 149-168.
17. Yeh, C.M. and Chen, M. (1995). Power and sample size for high-order crossover designs. *Proceedings of the Biopharmaceutical section of the American Statistical Association*, 279-184.