Journal of Pharmacokinetics and Pharmacodynamics, Vol. 28, No. 2, 2001

On Sample Size Calculation in Bioequivalence Trials

Shein-Chung Chow[1] and Hansheng Wang[2]

Received August 1, 2000—Final January 2, 2001

Sample size calculation plays an important role in bioequivalence trials. In practice, a bioequivalence study is usually conducted under a crossover design or a parallel design with raw data or log-transformed data. In this paper, we discuss the differences in sample size calculation between a crossover design and a parallel design with raw data or log-transformed data. Formulas for sample size calculation under a crossover design and a parallel design with raw data or log-transformed data are derived. A brief discussion for the relationship among these formulas is given.

Keywords: crossover design; parallel design; normal distribution; lognormal distribution; log transformation; robust.

1. INTRODUCTION

Sample size calculation plays an important role in pharmaceutical research and development. In practice, a prestudy power analysis and a precision analysis based on confidence interval (1,2) may be conducted for sample size calculation. For a selected sample size, it is desirable to have a sufficient power (say 80%) for the detection of a scientifically meaningful difference, if such a difference truly exists. A typical approach for a prestudy power analysis for sample size calculation is performed based on an appropriate test statistic, which is derived under a valid design for addressing clinical questions or hypotheses of interest. In practice, however, it is not uncommon to observe discrepancies among study objectives or hypotheses (e.g., equality or equivalence), study design (e.g., crossover or parallel), and statistical analysis (e.g., based on raw data or log-transformed data) in

[1]StatPlus, Inc., 1790 Yardley-Langhorne Road, Yardley, Pennsylvania 19067.
[2]Department of Statistics, University of Wisconsin-Madison, 1210 W. Dayton Street, Madison, Wisconsin 53706.

1567-567X/01/0400-0155$19.50/0

sample size calculation. These inconsistencies often result in (i) the wrong test for right hypotheses, (ii) the right test for wrong hypotheses, or (iii) the wrong test for wrong hypotheses (2).

For the assessment of bioequivalence (BE) in drug absorption between drug products, the United States Food and Drug Administration (FDA) indicates that the primary pharmacokinetic (PK) parameters, such as the area under the blood or plasma concentration–time curve (AUC) and the maximum concentration ($Cmax$), should be log-transformed before analysis (3,4). In addition to the crossover design, a parallel group may also be considered as an alternative, especially when the drug product under investigation has a relatively long half-life. Thus, sample size calculation for bioequivalence studies may be performed in each of the following situations: (i) a crossover design with raw data; (ii) a crossover design with log-transformed data, (iii) a parallel design with raw data, and (iv) a parallel design with log-transformed data. The objective of this paper is to clarify the statistical concepts, such as intersubject and intrasubject variability and the statistical property of log-transformed data in each of the above different situations. In addition, a formula for sample size calculation in each situation is provided. Since the FDA (3,4) recommends log transformation before bioequivalence study, we give a detailed discussion on the log scale sample size calculation in the section "Samples Size Estimation." A similar formula for raw scale is given in Appendix B. However, it should be pointed out that BE study based on raw scale is not recommended by the FDA (3,4). Because of limited sample size, researchers are not encouraged to test for normality of data distribution after log-transformation, nor should they use normality of data distribution as a reason for carrying out the statistical analysis on the original scale. If the researcher believes that the BE study should be analyzed on the original scale rather than log scale, justification should be provided.

In the next section, the difference between intrasubject variability and intersubject variability and the difference between a crossover design and a parallel-group design are briefly outlined. Also included in this section is the use of intersubject and/or intrasubject variability for sample size calculation. "Log-transformed Data versus Raw Data" examines the statistical properties of log-transformed data as compared to the raw data. A general formula for sample size calculation is derived in "Sample Size Estimation." Detailed information regarding the formula for sample size calculation for each of the four situations is also provided in this section. A comparison of sample size calculation in each of the four situations is made in the section "Comparison." A brief concluding remark is given in the last section, and some proof and sample size calculation formulas based on raw scale are given in Appendix B.

CROSSOVER VERSUS PARALLEL

There are two types of variations involved in a BE study. They are, namely, intrasubject and intersubject variation. The clarification of these two types of variability is essential for a thorough understanding of crossover and parallel design.

Intrasubject variability is the variability which could be observed by repeating experiments on the same subject under the same experiment condition. The source of intrasubject variability could be multifold. One important source is biological variability. Exactly the same results cannot be obtained even if they are from the same subject under the same experimental condition. Another important source is measurement or calculation error. For example, in a BE study, it could be the error when measuring the plasma concentration curve, it could be the error when calculating AUC, it could be the error of rounding after log-transformation, etc.

Intrasubject variability could be eliminated if we could repeat the experiment many times (in practice, this just means the average of a large number of times) on the same subject under the same experimental condition. The reason is that intrasubject variability tends to cancel out each other on average in a large scale.

If we could repeat the experiment on different subjects many times. It is possible that we would still see that the averages of the responses from different subjects are different from each other even if the experiments were carried out under the exactly the same condition. Then, what causes this difference or variation? It is not due to intrasubject variability, which have been eliminated by averaging infinitely repeated experiments. It is not due to experimental condition, which is exactly the same for different subjects. Therefore, this difference or variation can only be due to the unexplained difference between the two subjects. This is the so called intersubject variability, or pure intersubject variability as compared to the total intersubject variability we are going to introduce in the next paragraph.

It should be pointed out that sometimes people may call the variation observed from different subjects under the same experimental condition as intersubject variability, which is different from our intersubject variability. The variability observed from different subjects under the same experimental condition could be due to unexplained differences among subjects (pure intersubject variability); it also could due to the biological variability, or measurement error associated with a different experiment on a different subject (intrasubject variability). Therefore, it is clear that the observed variability from different subjects incorporates two components. They are, namely, pure intersubject variability and intrasubject variability. We call this total intersubject variability. For simplicity, it is also called total variability, which is exactly the variability one could observe from a parallel design.

In practice, no experiment can be carried out an infinite number of times. It is also not always true that an experiment can be carried out repeatedly on the same subject under the same experimental condition. But, we can still assess these two variability component (intra- and inter-) under a certain kind of statistical model assumption, e.g., the mixed effects model in BE study.

In a crossover design, subjects are randomly assigned to receive a sequence of treatments, which contain all the treatments in the study. For example, for a standard two-sequence, two-period (2×2) crossover design, subjects are randomly assigned to receive one of the two sequences of treatments (say, RT and TR), where T and R represent the test product and the reference product, respectively. Subjects who are assigned randomly to the sequence of RT, receive the reference product first and then receive the test product after a sufficient length of washout. The merits of a crossover design in bioequivalence studies is that it allows a within-subject (or intrasubject) comparison between treatments, since each subject serves as its own control, by removing the between-subject (or intersubject) variability from the comparison.

In a parallel-group design each subject receives one and only one treatment in a random fashion. The parallel-group design does not provide independent estimates for the intrasubject variability for each treatment. As a result, the assessment of BE is made based on the total variability, which include the intersubject variability and the intrasubject variability.

For the assessment of bioequivalence, the FDA indicates that a crossover design is the design of choice (21 CFR 320) because the assessment is made based on the intrasubject variability rather than the total variability. However, in some PK studies, a parallel design may be more appropriate than a crossover design, especially for drug products with relatively long half-lives or for dosing regimens that require a long time to complete (e.g., studies in which the PK of various dosing regimens of an intramuscular formulation are tested). The parallel design is preferred over a crossover design because the long washout period between dose administrations and the inevitable high dropout rate in a crossover design makes it highly impractical. Besides, the long exposure time of a depot makes it difficult to have it studied in crossover trials.

LOG-TRANSFORMED DATA VERSUS RAW DATA

For the assessment of average BE, the FDA indicates that a bioequivalence between two drug products can be claimed if the 90% confidence interval (CI) of the ratio of means of the primary PK parameters, such as AUC and $Cmax$, is entirely within the BE limits of (80%, 125%). As a result, the

ratio of means of the primary PK parameters is considered the bioequivalence measure. Let μ_T and μ_R be the population means of the test product and the reference product, respectively. FDA suggests that a log transformation be made before analysis. Let X and Y be the PK responses for the test product and the reference product. After log transformation, we assume that X and Y follow normal distributions with means μ_X^* and μ_Y^* and variance σ^2. Then

$$\mu_T = E(X) = e^{\mu_X^* + \delta^2/2}$$

$$\mu_T = E(X) = e^{\mu_X^* + \delta^2/2}$$

$$\Rightarrow \log\left(\frac{\mu_T}{\mu_R}\right) = \log(e^{\mu_X^* - \mu_Y^*} = \mu_X^* - \mu_Y^*$$

Under both the crossover and parallel design, an exact 90% CI for $\mu_X^* - \mu_Y^*$ can be obtained based on the log-transformed data. Hence, an exact 90% CI for $\mu_T - \mu_R$ can be obtained after the back transformation.

SAMPLE SIZE ESTIMATION

The following interval hypotheses are usually considered for the assessment of BE

$$H_0: |\mu_T - \mu_R| \geq \Delta \quad \text{vs.} \quad H_a: |\mu_T - \mu_R| < \Delta$$

where Δ is the BE limit. The interval hypotheses can be decomposed into the following two one-sided hypothesis [see also Chow and Liu (5,6)]

$$H_{01}: \mu_T - \mu_R \leq -\Delta \quad \text{vs.} \quad H_{a1}: \mu_T - \mu_R > -\Delta$$

and

$$H_{02}: \mu_T - \mu_R \geq \Delta \quad \text{vs.} \quad H_{a2}: \mu_T - \mu_R < \Delta$$

For a parallel design, let Y_{Ti} $i = 1, \ldots, n_1$ and Y_{Ri} $i = 1, \ldots, n_2$ be observations from the test product and the reference product, respectively. Here, n_1 is the sample size of the test drug group and n_2 is the sample size of the reference drug group. Then, we can reject the null hypothesis of bioinequivalence and conclude bioequivalence if

$$T_L = \frac{(\bar{Y}_T - \bar{Y}_R) + \Delta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} > t(\alpha, n_1 + n_2 - 2)$$

and

$$T_U = \frac{(\bar{Y}_T - \bar{Y}_R) - \Delta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} < t(\alpha, n_1 + n_2 - 2)$$

where $\bar{Y}_T$ and $\bar{Y}_R$ are the sample mean for the test product and the reference product, respectively, $\hat{\sigma}^2$ is the pooled estimate for $\sigma^2 = \text{Var}(Y_{Ti}) = \text{Var}(T_{Ri})$, which is given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(Y_{Ti} - \bar{Y}_{Ti})^2 + \sum_{i=1}^{n_1}(Y_{Ri} - \bar{Y}_{Ri})^2}{n_1 + n_2 - 2}$$

$t(a, b)$ is the $(1 - a)$th quantile of a t-distribution with b degrees of freedom. Let $\theta = \mu_T - \mu_R \in (-\Delta, \Delta)$, the power of this test procedure can be evaluated at μ_T-$\mu_R = \theta$, which is given by

$$\begin{aligned} P(t(\alpha, n_1 + n_2 - 2) - \frac{\Delta + \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} < \frac{\bar{Y}_T - \bar{Y}_R - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} \\ < \frac{\Delta - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} - t(\alpha, n_1 + n_2 - 2)) \\ \approx P(t(\alpha, n_1 + n_2 - 2) - \frac{\Delta + \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} < \frac{\bar{Y}_T - \bar{Y}_R - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} \\ < \frac{\Delta - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} - t(\alpha, n_1 + n_2 - 2)) \end{aligned}$$

Under the assumption of equal sample size per arm, i.e., $n_1 = n_2 = n$, the required sample size n for achieving the desired power of $(1 - \beta)$ is given by

$$\begin{aligned} n &\geq \frac{2\sigma^2(t(\alpha, 2n - 2) + t(\beta, 2n - 2))^2}{(\Delta - |\theta|)^2} \qquad \text{if } \theta \neq 0 \\ n &\geq \frac{2\sigma^2(t(\alpha, 2n - 2) + t(\beta, 2n - 2))^2}{\Delta^2} \qquad \text{if } \theta \neq 0 \end{aligned} \tag{1}$$

It should be pointed out that the formula for $\varepsilon \neq 0$ is based on certain kind of approximation. When $\sqrt{n}|\theta| \gg 0$, this is a good approximation. But, when this is close to 0, the formula should be used with caution. Similar things also happen for crossover design. The detailed proof for this formula can be found in Appendix A.

For a standard two-sequence two-period (2×2) crossover design (TR, RT), let Y_{ijk} be the observation of the ith subject in kth sequence at the jth dosing period, and suppose there are n_1 subjects are randomly assigned to sequence (TR) and n_2 subjects are randomly assigned to sequence (RT). Similarly, we can reject the null hypothesis of bioinequivalence and conclude

bioequivalence if

$$T_{\mathrm{L}} = \frac{(\bar{Y}_{\mathrm{T}} - \bar{Y}_{\mathrm{R}}) + \Delta}{\hat{\sigma}_d \sqrt{1/n_1 + 1/n_2}} > t(\alpha, n_1 + n_2 - 2)$$

and

$$T_{\mathrm{U}} = \frac{(\bar{Y}_{\mathrm{T}} - \bar{Y}_{\mathrm{R}}) - \Delta}{\hat{\sigma}_d \sqrt{1/n_1 + 1/n_2}} < t(\alpha, n_1 + n_2 - 2)$$

where $\bar{Y}_{\mathrm{T}} = \frac{1}{2}(\bar{Y}_{\cdot 11} + \bar{Y}_{\cdot 22})$, $\bar{Y}_{\mathrm{R}} = \frac{1}{2}(\bar{Y}_{\cdot 21} + \bar{Y}_{\cdot 12})$.Therefore

$$\begin{aligned}\bar{Y}_{\mathrm{T}} - \bar{Y}_{\mathrm{R}} &= \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{Y_{i11} - Y_{i12}}{2}\right) + \frac{1}{n_2}\sum_{i=1}^{n_2}\left(\frac{Y_{i22} - Y_{i21}}{2}\right) \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1} d_{i1} + \frac{1}{n_2}\sum_{i=1}^{n_2} d_{i2}\end{aligned}$$

where $d_{i1} = \frac{1}{2}(\bar{Y}_{i11} - Y_{i21})$ and $d_{i2} = \frac{1}{2}(\bar{Y}_{i22} - Y_{i12})$ It follows that d_{ik} are independent and identically distributed as $N(\theta, \sigma_d^2)$, where $\sigma_d^2 = \mathrm{Var}(d_{ik})$. Therefore, an estimate for σ_d^2 can be obtained by

$$\hat{\sigma}_d^2 = \frac{1}{n_1 + n_2 - 2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(d_{ik} - \bar{d}_{\cdot k})^2$$

$\bar{Y}_{\mathrm{T}}$, instead of pooled mean based on Y_{i11} and Y_{i22} is used because the use of the weighted mean $\bar{Y}_{\mathrm{T}}$ eliminates a possible period effect.

Thus, the power at $\mu_{\mathrm{T}} - \mu_{\mathrm{R}} = \theta$ is given by

$$\begin{aligned}P\Big(t(\alpha, n_1 + n_2 - 2) - \frac{\theta + \Delta}{\hat{\sigma}_d \sqrt{1/n_1 + 1/n_2}} &< \frac{\bar{Y}_{\mathrm{T}} - \bar{Y}_{\mathrm{R}} - \theta}{\hat{\sigma}_d \sqrt{1/n_1 + 1/n_2}} \\ &< \frac{\Delta - \theta}{\hat{\sigma}_d \sqrt{1/n_1 + 1/n_2}} - t(\alpha, n_1 + n_2 - 2)\Big)\end{aligned}$$

where $\sigma_e^2 = 2\sigma_d^2 = 2E(\sigma_d^2)$ and σ_e^2 is the intrasubject variability. The sample size calculation formula is then given by (7)

$$\begin{aligned} n &\geq [t(\alpha, 2n-2) + t(\beta, 2n-2)]^2\left(\frac{\sigma_e^2}{\Delta - |\theta|}\right)^2 \qquad \text{if } \theta \neq 0 \\ n &\geq [t(\alpha, 2n-2) + t(\beta/2, 2n-2)]^2\left(\frac{\sigma_e^2}{\Delta}\right)^2 \qquad \text{if } \theta \neq 0 \end{aligned} \tag{2}$$

Parallel Design with Log-Transformed Data

Under a parallel design with log-transformed data, then the BE limit is $\Delta = \log 1.25$. Formula (1) becomes

$$\begin{aligned} n &\geq \frac{2\sigma^2 t(\alpha, 2n-2) + t(\beta, 2n-2)]^2}{(\log 1.25 - |\theta'|)^2} \qquad \text{if } \theta' \neq 0 \\ n &\geq \frac{2\sigma^2 t(\alpha, 2n-2) + t(\beta/2, 2n-2)]^2}{\log 1.25^2} \qquad \text{if } \theta' \neq 0 \end{aligned} \tag{3}$$

where $\theta' = \log(\mu_T/\mu_R)$.

Crossover Design with Log-Transformed Data

Under a crossover design with log-transformed data, the BE limit is $\Delta = \log 1.25$. If we let $\theta' = |\log(\mu_T/\mu_R)|$, Formula (2) becomes

$$\begin{aligned} n &\geq \frac{\sigma_e^2 t(\alpha, 2n-2) + t(\beta, 2n-2)]^2}{(\log 1.25 - |\theta'|)^2} \qquad \text{if } \theta' \neq 0 \\ n &\geq \frac{\sigma_e^2 t(\alpha, 2n-2) + t(\beta/2, 2n-2)]^2}{\log 1.25^2} \qquad \text{if } \theta' \neq 0 \end{aligned} \tag{4}$$

where $\sigma_e^2 = 2\sigma_d^2$ is the intrasubject variability after log transformation. Note that Formula (2) is similar to that given in Chen *et al.* (8).

To provide a better understanding, the sample sizes required for various powers at the $\alpha = 0.05$ nominal level of significance under different designs (crossover or parallel) with raw data or log-transformed data are given in Tables I–IV.

COMPARISON

As it can be seen from the above discussion and the discussion for raw data in Appendix B, for different designs (parallel or crossover) and different data (raw or log-transformed), the formulas for sample size calculation are very similar, but slightly different. In practice, scientists often confuse one with another. The following discussion may be helpful for clarification.

Coefficient of Variation for Raw Data and σ^2 for Log-Transformed Data

We note that the sample size derivation is based on normality assumption for the raw data and lognormality assumption for the transformed data. Thus, it is of interest to study the distribution of log X when X is normally distributed with mean μ and variance σ^2. Under the assumption

Table I. Total Sample Size for Schuirmann's Two One-Sided *t*-Tests Procedure at $\alpha = 0.05$ Nominal Level—Parallel Design with Log-Transformed Data

	$\lvert\theta'\rvert$ (Power 80%)				$\lvert\theta'\rvert$ (Power 90%)			
σ%	0%	5%	10%	15%	0%	5%	10%	15%
10	10	10	18	48	12	14	24	66
12	12	14	26	68	14	18	34	94
14	16	18	34	92	20	24	46	128
16	20	22	44	120	24	32	60	166
18	24	28	54	152	30	38	74	210
20	30	34	66	186	36	48	92	258
22	36	42	80	226	44	58	110	312
24	42	50	96	268	52	68	132	370
26	48	58	112	314	60	78	154	434
28	56	66	130	364	70	92	178	504
30	64	76	148	418	80	104	204	578
32	72	86	168	474	90	118	232	658
34	82	96	190	536	102	134	262	742
36	90	108	212	600	114	150	294	832
38	102	120	238	670	128	166	328	926
40	112	134	262	742	140	184	364	1026

Table II. Total Sample Size for Schuirmann's Two One-Sided *t*-Tests Procedure at $\alpha = 0.05$ Nominal Level—Crossover Design with Log-Transformed Data

	$\lvert\theta'\rvert$ (Power 80%)				$\lvert\theta'\rvert$ (Power 90%)			
σ%	0%	5%	10%	15%	0%	5%	10%	15%
10	6	6	10	24	8	8	14	34
12	8	8	14	34	8	10	18	48
14	10	10	18	46	10	14	24	64
16	10	12	22	60	14	16	30	84
18	14	16	28	76	16	20	38	106
20	16	18	34	94	20	24	46	130
22	18	22	42	114	24	30	56	156
24	22	26	48	134	28	34	66	186
26	26	30	56	158	32	40	78	218
28	28	34	66	182	36	46	90	252
30	32	38	74	210	42	54	104	290
32	36	44	86	238	46	60	118	330
34	42	50	96	268	52	68	132	372
36	46	56	108	302	58	76	148	416
38	52	62	120	336	64	84	164	464
40	56	68	132	372	72	94	182	514

that X is normally distributed, we have

$$\log X - \log \mu = \log\left(1 + \frac{X - \mu}{\mu}\right)$$

Table III. Total Sample Size for Schuirmann's Two One-Sided *t*-Tests Procedure at $\alpha = 0.05$ Nominal Level—Parallel Design with Raw Data

	$\lvert\theta'\rvert$ (Power 80%)				$\lvert\theta'\rvert$ (Power 90%)			
σ%	0%	5%	10%	15%	0%	5%	10%	15%
10	10	12	26	100	14	18	36	138
12	14	18	38	144	18	24	52	198
14	18	24	50	196	24	32	68	270
16	24	30	64	254	30	40	90	352
18	30	38	82	322	38	52	112	446
20	36	46	100	398	46	62	138	550
22	44	54	122	480	54	76	168	664
24	52	64	144	572	64	90	198	790
26	60	76	168	670	76	104	234	928
28	68	88	196	778	86	122	270	1076
30	78	100	224	892	100	138	310	1234
32	90	114	254	1014	112	158	352	1404
34	100	128	288	1146	126	178	398	1586
36	112	144	322	1284	142	198	446	1778
38	126	160	358	1430	158	222	496	1980
40	138	178	398	1584	176	246	550	2194

Table IV. Total Sample Size for Schuirmann's Two One-Sided *t*-Tests Procedure at $\alpha = 0.05$ Nominal Level—Crossover Design with Raw Data

	$\lvert\theta'\rvert$ (Power 80%)				$\lvert\theta'\rvert$ (Power 90%)			
σ%	0%	5%	10%	15%	0%	5%	10%	15%
10	6	8	14	52	8	10	20	70
12	8	10	20	72	10	12	26	100
14	10	12	26	98	12	16	36	136
16	12	16	34	128	16	22	46	178
18	16	20	42	162	20	26	58	224
20	20	24	52	200	24	32	70	276
22	22	28	62	240	28	38	84	334
24	26	34	72	286	34	46	100	396
26	30	38	86	336	38	54	118	464
28	36	44	98	390	44	62	136	538
30	40	52	112	446	50	70	156	618
32	46	58	128	508	58	80	178	704
34	52	66	144	574	64	90	200	794
36	58	72	162	642	72	100	224	890
38	64	80	180	716	80	112	248	990
40	70	90	200	792	88	124	276	1098

Since

$$\mathrm{Var}\left(\frac{X-\mu}{\mu}\right) = \frac{\sigma^2}{\mu^2} = CV^2$$

if CV is sufficiently small, $(X-\mu)/\mu$ is close to 0. As a result, by Taylor's expansion

$$\log\left(1+\frac{X-\mu}{\mu}\right)\approx\frac{X-\mu}{\mu}$$

Then, it follows

$$\log X\approx\log\mu+\frac{X-\mu}{\mu}\sim N(\log\mu,\, CV^2).$$

The above indicates that when CV is small, X is still approximately normally distributed, even if X is from a normal population. Therefore, the procedure based on log-transformed data is robust in some sense. In addition, the CV observed from the raw data is very similar to the variance obtained from the log-transformed data.

Inter- and Intrasubject Variation

For the crossover design, since each subject serves as its own control, the intersubject variation is removed from comparison. As a result, Formula (2) only involves the intrasubject variability. On the other hand, for the parallel design, σ^2 in Formula (1) includes both the inter- and intrasubject variabilities.

Bioequivalence Limits

Traditionally, for raw data, BE can be established if the 90% CI for $\mu_R-\mu_R$ is entirely within the interval of $(-0.2\mu_R, 0.2\mu_R)$. This is the reason why 0.2 appears in the formula for raw data. However, FDA guidances (3,4) recommended log transformation before BE study. Then for log-transformed data, the BE can be established if the 90% CI for μ_T/μ_R is entirely located in the interval (80%, 125%). That is why 1.25 appears in the formula for log-transformed data. It should be noted that $\log 1.25 = -\log 0.8 = 0.2331$. In other words, the BE limit for the raw data is symmetric about 0 (i.e., $\pm 0.2\mu_R$), while the BE limit for the log-transformed data is also symmetric about 0 after log transformation.

CONCLUSION

In this paper, formulas for sample size calculation under a crossover design and a parallel-group design with raw data or log-transformed data are provided. The resultant sample size has a desired power of $(1-\beta)\times 100$ for the demonstration of BE in drug absorption between drug products.

The formulas can be applied to the establishment of therapeutic equivalence in clinical development.

It should be noted, however, that the formulas derived for sample size calculation for the assessment of bioequivalence are for average bioequivalence, not population bioequivalence or individual bioequivalence (9). In addition, the sample size calculation discussed in this paper does not account for α-adjustment for multiple comparisons.

APPENDIX A

Proof of Formula (1)

According to our discussion, the two-one side procedure for hypothesis

$$H_0: |\mu_T - \mu_R| > \Delta \quad \text{versus} \quad H_\alpha: |\mu_T - \mu_R| < \Delta$$

would have power approximately equal to

$$P(t(\alpha, n_1 + n_2 - 2) - \frac{\Delta + \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} < \frac{\bar{Y}_T - \bar{Y}_R - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} < \frac{\Delta - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} - t(\alpha, n_1 + n_2 - 2))$$

Note that

$$t^* = < \frac{\bar{Y}_T - \bar{Y}_R - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}}$$

follows a t-distribution with degree of freedom $(n_1 + n_2 - 2)$. Under the assumption $n = n_2$ and $\theta = 0$, this power will equal to

$$P(|t^*| < \frac{\Delta}{\sqrt{2/n}\,\sigma} - t(\alpha, 2n - 2))$$

In order to have desired power $(1 - \beta)$, we have to set

$$\frac{\Delta}{\sqrt{2/n}\sigma} - t(\alpha, 2n - 2) = t(\beta/2, 2n - 2)$$

This implies n, the sample size per group, should satisfy

$$n \geq \frac{2\sigma^2(t(\alpha, 2n - 2) + t(\beta/2, 2n - 2))^2}{\Delta^2}$$

However, if $\theta \neq 0$ (without loss of generality assume $\theta > 0$), the power will be

$$P(t(\alpha, n_1 + n_2 - 2) - \frac{\Delta + \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} < \frac{\bar{Y}_T - \bar{Y}_R - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} < \frac{\Delta - \theta}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} - t(\alpha, n_1 + n_2 - 2))$$

$$= P(t(\alpha, 2n - 2) - \frac{\Delta + \theta}{\sqrt{2/n}\,\sigma} < t^* < \frac{\Delta - \theta}{\sqrt{2/n}\,\sigma} - t(\alpha, 2n - 2))$$

When $\theta > 0$ is relatively big,

$$P\left(t^* < t(\alpha, 2n - 2) - \frac{\Delta + \theta}{\sqrt{2/n}\,\sigma}\right) \approx 0$$

then the power is approximately equal to

$$P\left(t^* < \frac{\Delta - \theta}{\sqrt{2/n}\,\sigma}\right) t(\alpha, 2n - 2)\right)$$

In this case, in order to get the desired power $1 - \beta$, we need to set

$$\frac{\Delta - \theta}{\sqrt{2/n}\,\alpha} - t(\alpha, 2n - 2) = t(\beta, 2n - 2)$$

This implies that n, the sample size per treatment group, should satisfy

$$n \geq \frac{2\sigma^2(t(\alpha, 2n - 2) + t(\beta, 2n - 2))^2}{(\Delta - \theta)^2}$$

Similarly, if $\theta < 0$, then

$$n \geq \frac{2\sigma^2(t(\alpha, 2n - 2) + t(\beta, 2n - 2))^2}{(\Delta - \theta)^2}$$

In either case, when $\theta = 0$, the sample size per treatment group can be determined by

$$n \geq \frac{2\sigma^2(t(\alpha, 2n - 2) + t(\beta, 2n - 2))^2}{(\Delta - |\theta|)^2}$$

It should be pointed out, the sample size calculation when $\theta \neq 0$ is based on approximation. When θ is big enough, this is a good approximation, but when θ is close to 0, the derived formula should be used with caution.

APPENDIX B

Sample Size Formula for Raw Scale

Parallel Design with Raw Data

Under a parallel design with raw data and assuming that the BE limit $\Delta = 0.2\mu_R$, the following formula is useful

$$\begin{aligned} n &\geq \frac{2CV^2(t(\alpha, 2n-2) + t(\beta, 2n-2))^2}{(0.2 - |\theta'|)^2} \qquad \text{if } \theta' > 0 \\ n &\geq \frac{2CV^2(t(\alpha, 2n-2) + t(\beta/2, 2n-2))^2}{0.2^2} \qquad \text{if } \theta' = 0 \end{aligned} \tag{B1}$$

where $\theta' = (\mu_T - \mu_R)/\mu_R$, σ^2 is the sum of the inter- and intrasubject variability, and $CV = \sigma/\mu_R$.

Crossover Design with Raw Data

Under a crossover design with raw data and assuming that the BE limit is $\Delta = 0.2\mu_R$, the formula for sample size calculation is given by

$$\begin{aligned} n &\geq \frac{2CV^2(t(\alpha, 2n-2) + t(\beta, 2n-2))^2}{(0.2 - |\theta'|)^2} \qquad \text{if } \theta' > 0 \\ n &\geq \frac{2CV^2(t(\alpha, 2n-2) + t(\beta/2, 2n-2))^2}{0.2^2} \qquad \text{if } \theta' = 0 \end{aligned} \tag{B1}$$

where $\theta' = (\mu_T - \mu_R)/\mu_R$, $\mathrm{CV} = \sigma_e/\mu_R$, and $\sigma_e^2 = 2\sigma_d^2$ is the intrasubject variability. Note that Formula (B2) is consistent with that given in Liu and Chow (7).

REFERENCES

1. E. Diletti, D. Hauschke, and V. W. Steinijans. Sample size determination for bioequivalence assessment by means of confidence intervals. *Int. J. Clin. Pharm. Ther. Toxicol.* **19:**1–8 (1991).
2. S. C. Chow and J. P. Liu. *Design and Analysis of Clinical Trials*. John Wiley & Sons, New York, 1998.
3. FDA. *Guidance on Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*, Division of Bioequivalence, Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD, 1992.
4. FDA. *Guidance for Industry*: *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD, 2000.
5. S. C. Chow and J. P. Liu. *Design and Analysis of Bioavailability and Bioequivalence*, Marcel Dekker New York, 1992.
6. S. C. Chow and J. P. Liu. *Design and Analysis of Bioavailability and Bioequivalence Studies—Revised and Expanded*, 2nd ed., Marcel Dekker, New York, 1999.
7. Liu, J. P. and Chow, S. C.. Sample size determination for the two one-sided tests procedure in bioequivalence. *J. Pharmacokin. & Biopharm.* **20:**101–104 (1992).

8. K. W. Chen, G. Li, and S. C. Chow. A note on sample size determination for bioequivalence studies with high-order crossover designs. *J. Pharmacokin. & Biopharm.* **25:**753–765 (1997).
9. FDA. Average, Population, and Individual Approaches to Establishing Bioequivalence. U.S. Department of Health and Human Services, Food and Drug Administration. Center for Drug Evaluation and Research (CDER), Rockville, MD, 1999.