Bioequivalence Trials, Intersection-Union Tests, and Equivalence Confidence Sets

Roger L. Berger
Department of Statistics
North Carolina State University
Raleigh, NC 27595-8203

Jason C. Hsu
Department of Statistics
The Ohio State University
Columbus, OH 43210-1247

October 9, 1996 – Final version

Abstract

The bioequivalence problem is of practical importance because the approval of most generic drugs in the United States and the European Community (EC) requires the establishment of bioequivalence between the name brand drug and the proposed generic version. The problem is theoretically interesting because it has been recognized as one for which the desired inference, instead of the usual *significant difference*, is *practical equivalence*. The concept of intersection-union tests will be shown to clarify, simplify, and unify bioequivalence testing. A test more powerful than the one currently specified by the FDA and EC guidelines will be derived. The claim that the bioequivalence problem defined in terms of the ratio of parameters is more difficult than the problem defined in terms of the difference of parameters will be refuted. The misconception that size-α bioequivalence tests generally correspond to $100(1-2\alpha)\%$ confidence sets will be shown to lead to incorrect statistical practices, and should be abandoned. Techniques for constructing $100(1-\alpha)\%$ confidence sets that correspond to size-α bioequivalence tests will be described. Finally, multiparameter bioequivalence problems will be discussed.

Key words and phrases: Bioequivalence, bioavailability, hypothesis test, confidence interval, intersection-union, size, level, equivalence test, pharmacokinetic, unbiased.

1 Bioequivalence Problem

Two different drugs or formulations of the same drug are called *bioequivalent* if they are absorbed into the blood and become available at the drug action site at about the same rate and concentration. Bioequivalence is usually studied by administering dosages to subjects and measuring concentration of the drug in the blood just before and at set times after the administration. These data are then used to determine if the drugs are absorbed at the same rate.

The determination of bioequivalence is very important in the pharmaceutical industry because regulatory agencies allow a generic drug to be marketed if its manufacturer can demonstrate that the generic product is bioequivalent to the brand-name product. The assumption is that bioequivalent drugs will provide the same therapeutic effect. If the generic drug manufacturer can demonstrate bioequivalence, it does not need to perform costly clinical trials to demonstrate the safety and efficacy of the generic product. Yet, this bioequivalence must be demonstrated in a statistically sound way to protect the consumer from ineffective or unsafe drugs.

These concentration by time measurements are connected with a polygonal curve and several variables are measured. The common measurements are AUC (Area Under Curve), $\mathrm{C_{max}}$ (maximum concentration), and $\mathrm{T_{max}}$ (time until maximum concentration). The two drugs are bioequivalent if the population means of AUC and $\mathrm{C_{max}}$ are sufficiently close. Descriptive statistics for $\mathrm{T_{max}}$ are usually provided, but formal tests are not required.

For example, let μ_T denote the population mean AUC for the generic (Test) drug and μ_R denote the population mean AUC for the brand-name (Reference) drug. To demonstrate bioequivalence, the following hypotheses are tested:

$$\begin{array}{c} \mathrm{H}_0 : \frac{\mu_T}{\mu_R} \leq \delta_L \text{ or } \frac{\mu_T}{\mu_R} \geq \delta_U \\ \text{versus} \\ \mathrm{H}_a : \delta_L < \frac{\mu_T}{\mu_R} < \delta_U . \end{array} \tag{1}$$

The values δ_L and δ_U are standards set by regulatory agencies that define how “close” the drugs must be to be declared bioequivalent. Currently, both the United States Food and Drug Administration (1992) and the European Community uses $\delta_U = 1.25$ and $\delta_L = .80 = 1/1.25$ for AUC. For $\mathrm{C_{max}}$, the United States again uses $\delta_U = 1.25$ and $\delta_L = .80$, but Europe uses the less restrictive limits $\delta_U = 1.43$ and $\delta_L = .70 = 1/1.43$ (Hauck et al. (1995)). Note that these limits for AUC and $\mathrm{C_{max}}$ are symmetric about one in the ratio scale.

Often, logarithms are taken and the hypotheses (1) are stated as

$$\begin{array}{c} \mathrm{H}_0 : \eta_T - \eta_R \leq \theta_L \text{ or } \eta_T - \eta_R \geq \theta_U \\ \text{versus} \\ \mathrm{H}_a : \theta_L < \eta_T - \eta_R < \theta_U . \end{array} \tag{2}$$

Here, $\eta_T = \log(\mu_T)$, $\eta_R = \log(\mu_R)$, $\theta_U = \log(\delta_U)$ and $\theta_L = \log(\delta_L)$. With $\delta_U = 1.25$ and $\delta_L = .80$ or $\delta_U = 1.43$ and $\delta_L = .70$, $\theta_U = -\theta_L$, and the standards are symmetric about zero.

In a hypothesis test of (1) or (2), the Type I error rate is the probability of declaring the drugs to be bioequivalent, when in fact they are not. By setting up the hypotheses as in (1) or (2) and controlling the Type I error rate at a specified small value, say, $\alpha = .05$, the consumer's risk is being controlled. That (1) or (2) is the proper formulation in problems like these was recognized early on by some authors. For example, Lehmann (1959, p. 88), not specifically discussing bioequivalence, says, “One then sets up the (null) hypothesis that [the parameter] does not lie within the required limits so that an error of the first kind consists in declaring [the parameter] to be satisfactory when in fact it is not.” But not until Schuirmann (1981, 1987), Westlake (1981) and Anderson and Hauck (1983) were hypotheses correctly formulated as in (1) or (2) in bioequivalence problems.

Despite the fact that bioequivalence testing problems are now correctly formulated as (1) or (2), many inappropriate statistical procedures are still used in this area. Tests that claim to have a specified size α, but are either liberal or conservative, are used. Liberal tests compromise the consumer's safety, and conservative tests put an undo burden on the generic drug manufacturer. Tests are often defined in terms of confidence intervals in statistically unsound ways. These tests, again, do not properly control the consumer's risk.

In this paper, we will describe current bioequivalence tests that have incorrect error rates. We will offer new tests that correctly control the consumer's risk. In several cases, the tests we propose are uniformly more powerful than the existing tests while still controlling the Type I error rate at the specified rate α. We will examine and criticize the current practice of defining tests in terms of $100(1-2\alpha)\%$ confidence sets. We will show that this only works in special cases and gives poor results in other cases. We will discuss how properly to construct $100(1-\alpha)\%$ confidence sets that correspond to size-α tests. And we will discuss how our methods can be applied to complicated, multiparameter bioequivalence problems that have received only slight attention in the literature. The intersection-union method of testing will be found to be very useful in understanding and constructing bioequivalence tests. Section 2 provides a more detailed outline to our discussions.

Hypotheses such as (1) and (2) that specify only that population means should be close are called *average bioequivalence* hypotheses. Hypotheses that state that the whole distribution of bioavailabilities is the same for the test and reference populations are called *population bioequivalence* hypotheses. If a parametric form of these populations is assumed, then hypotheses such as (25) that specify that all population parameters, e.g., variances as well as means, should be close are population bioequivalence hypotheses. Sometimes bioequivalence is defined in terms of parameters that more directly measure equivalence of response within an individual. Good introductions to *individual bioequivalence* are given by Anderson and Hauck (1990), Hauck and Anderson (1992), Sheiner (1992), Schall and Luus (1993), and Anderson (1993). Although we do not explicitly consider individual bioequivalence in this paper, many of the concepts and techniques we describe should be applicable in that area also.

In this paper, our discussion will be entirely in terms of bioequivalence testing. But our comments and techniques apply to other problems, such as in quality assurance, in which the aim is to show that two parameters are close or that a parameter is between two specification limits. Because of this wider applicability, the methods we will discuss might more properly be referred to as *equivalence tests* and *equivalence confidence intervals.*

2 Tests, Confidence Sets and Curiosities

Various experimental designs are used to gather data for bioequivalence trials. Chow and Liu (1992) describe parallel designs (two independent samples), and two-period and multiperiod crossover designs. The issues we discuss apply to all these different designs. For brevity, we will discuss only the simple parallel design and two period crossover design.

2.1 Difference hypotheses

It is customary to employ lognormal models in bioequivalence studies of AUC and $\mathrm{C}_{\mathrm{max}}$. See Section 2.2 for rationales for this model.

Let X^* denote a lognormal measurement from the test drug in the original scale, and let $X = \log(X^*)$. Similarly, let Y^* denote an original measurement and $Y = \log(Y^*)$ for the reference drug. Let (η_T, σ^2) denote the lognormal parameters for X^* and (η_R, σ^2) denote the lognormal parameters for Y^*. Then the test and reference drug means are $\mu_T = e^{\eta_T + \sigma^2/2}$

and $\mu_R = e^{\eta_R+\sigma^2/2}$, respectively. Therefore, the condition

$$\delta_L < \frac{\mu_T}{\mu_R} = e^{\eta_T-\eta_R} < \delta_U$$

is equivalent to

$$\theta_L < \eta_T - \eta_R < \theta_U, \tag{3}$$

where $\theta_L = \log(\delta_L)$ and $\theta_U = \log(\delta_U)$ are known constants. Thus, the hypothesis to be tested in this lognormal model can be stated as either (1) or (2). Usually the hypotheses are stated as (2) and the test is based on log transformed data that is normally distributed with means η_T and η_R and common variance σ^2. The equivalence of (1) and (2) is dependent on the assumption of equal variances. On the other hand, if μ_T and μ_R represent the medians of X^* and Y^* and $\eta_T = \log(\mu_T)$ and $\eta_R = \log(\mu_R)$, then η_T and η_R are the medians of X and Y, respectively. So, in terms of medians, (1) and (2) are always equivalent, and the analysis can be carried out in either the original or log transformed scale. But, bioequivalence is almost always defined in terms of means rather than medians.

Westlake (1981) and Schuirmann (1981) proposed what has become the standard test of (2). It is called the "two one-sided tests" (TOST). The TOST has this general form. Let D be an estimate of $\eta_T - \eta_R$ that has a normal distribution with mean $\eta_T - \eta_R$ and variance σ_D^2. Let $\mathrm{SE}(D)$ be an estimate of σ_D that is independent of D and such that $r(\mathrm{SE}(D))^2/\sigma_D^2$ has a χ^2 distribution with r degrees of freedom. Then

$$t = \frac{D - (\eta_T - \eta_R)}{\mathrm{SE}(D)}$$

has a Student's t distribution with r degrees of freedom. The TOST is based on the two statistics

$$T_U = \frac{D - \theta_U}{\mathrm{SE}(D)} \quad \text{and} \quad T_L = \frac{D - \theta_L}{\mathrm{SE}(D)}. \tag{4}$$

The TOST tests (2) using the ordinary, one-sided, size-α t-test based on T_L for

$$\begin{array}{c} \mathrm{H}_{01} : \eta_T - \eta_R \le \theta_L \\ \text{versus} \\ \mathrm{H}_{a1} : \eta_T - \eta_R > \theta_L \end{array} \tag{5}$$

and the ordinary, one-sided, size-α t-test based on T_U for

$$\begin{array}{c} \mathrm{H}_{02} : \eta_T - \eta_R \ge \theta_U \\ \text{versus} \\ \mathrm{H}_{a2} : \eta_T - \eta_R < \theta_U. \end{array} \tag{6}$$

It rejects H_0 at level α and declares the two drugs to be bioequivalent if both tests reject, that is, if

$$T_U < -t_{\alpha,r} \quad \text{and} \quad T_L > t_{\alpha,r}, \tag{7}$$

where $t_{\alpha,r}$ is the upper 100α percentile of a Student's t distribution with r degrees of freedom. For testing (2), all the tests we will discuss are functions of $(D, \mathrm{SE}(D))$. The distribution of $(D, \mathrm{SE}(D))$ is determined by the parameter $(\eta_T, \eta_R, \sigma_D^2)$.

In the simple parallel design, let $X_1^*, \ldots, X_m^*$ denote the independent lognormal(η_T, σ^2) measurements on m subjects from the test drug in the original scale, and let $X_1, \ldots, X_m$ denote the logarithms of these measurements. Similarly, let $Y_1^*, \ldots, Y_n^*$ and $Y_1, \ldots, Y_n$ denote the original measurements (lognormal(η_R, σ^2)) and logarithms for an independent sample of n subjects on the reference drug. If $\overline{X}$ denotes the sample mean of $X_1, \ldots, X_m$, $\overline{Y}$ denotes the sample mean of $Y_1, \ldots, Y_n$, and S^2 denotes the pooled estimate of σ^2, computed from both samples, then

$$D = \overline{X} - \overline{Y}$$

and

$$\mathrm{SE}(D) = S\sqrt{\frac{1}{m} + \frac{1}{n}}.$$

The degrees of freedom are $r = m + n - 2$.

In bioequivalence studies, much more common than simple parallel designs are two-period, crossover designs. In a two-period, crossover design, a group of m subjects (Sequence 1) receives the reference drug and observations on the pharmacokinetic response are made. After a washout period to remove any carryover effect, this group receives the test drug and observations are again made. A second group of n subjects (Sequence 2) receives the drugs in the opposite order. After log transformation, the response of the kth subject in the jth period of the ith sequence is modeled as

$$Y_{ijk} = \gamma + S_{ik} + P_j + F_{(i,j)} + \epsilon_{ijk},$$

where γ is the overall mean; P_j is the fixed effect of period j; $F_{(i,j)}$ is the fixed effect of the formulation administered in period j of sequence i, that is, $F_{(1,1)} = F_{(2,2)} = F_R$ and $F_{(1,2)} = F_{(2,1)} = F_T$; S_{ik} is the random effect of subject k in sequence i; and ϵ_{ijk} is the random error. It is assumed that $P_1 + P_2 = F_T + F_R = 0$. The S_{ik}s and the ϵ_{ijk}s are all independent normal random variables with mean 0. The variance of S_{ik} is σ_S^2 and the variance of ϵ_{ijk} is σ_T^2 and σ_R^2 for the test and reference formulations, respectively. For this design,

$$D = \frac{\overline{Y}_{12\cdot} - \overline{Y}_{11\cdot} + \overline{Y}_{21\cdot} - \overline{Y}_{22\cdot}}{2}$$

is a normally distributed unbiased estimate of $F_T - F_R = \eta_T - \eta_r$ with variance

$$\sigma_D^2 = (\sigma_R^2 + \sigma_T^2)\frac{1}{4}\left(\frac{1}{m} + \frac{1}{n}\right).$$

The standard error of D is

$$\mathrm{SE}(D) = S\frac{1}{2}\sqrt{\frac{1}{m} + \frac{1}{n}},$$

where

$$\begin{aligned} S^2 &= \frac{1}{m+n-2}\left(\sum_{k=1}^{m}\left(Y_{12k} - Y_{11k} - (\overline{Y}_{12\cdot} - \overline{Y}_{11\cdot})\right)^2\right. \\ &\qquad \left. + \sum_{k=1}^{n}\left(Y_{21k} - Y_{22k} - (\overline{Y}_{21\cdot} - \overline{Y}_{22\cdot})\right)^2\right). \end{aligned}$$

The estimate D is the average of the averages of the intrasubject differences for the two sequences, and S^2 is a pooled estimate of the variance of an intrasubject difference. For this crossover design, also, the degrees of freedom are $r = m + n - 2$.

Following Lehmann (1959), we define the size of a test as

$$\text{size} = \sup_{\mathrm{H}_0} P(\text{reject } \mathrm{H}_0).$$

The size of the TOST is exactly equal to α, even though $P(\text{reject } \mathrm{H}_0) < \alpha$ for every $(\eta_T, \eta_R, \sigma_D^2)$ in the null hypothesis. The supremum value of α is attained in the limit as $\eta_T - \eta_R = \theta_L$ (or θ_U) and $\sigma_D^2 \to 0$. Both the FDA bioequivalence guideline (FDA, 1992) and the European Community guideline (EC-GCP, 1993) specify that bioequivalence be established using a 5% TOST.

The TOST is unusual in that two size-α tests are combined to form a size-α test. Often, when multiple tests are combined, some adjustment must be made to the sizes of the individual tests to achieve an overall size-α test. Why this is not necessary for the TOST is best understood through the theory of intersection-union tests (IUTs), which we describe in Section 3. In Sections 4.1 and 4.2 we will show that the IUT theory is useful for understanding the TOST. Also, the IUT theory can guide the construction of tests for (2) that have the same size-α as the TOST but are uniformly more powerful than the TOST.

2.2 Ratio hypotheses

Sometimes, a normal model should be used. In this model, the original measurements are normally distributed with means μ_T and μ_R. This model is different from the lognormal model in that now the hypothesis to be tested concerns the ratio of the means of these normal observations. That is, we wish to test (1). This problem has received less attention than (2). Dealing with the ratio μ_T/μ_R has been perceived as more difficult than dealing with the difference $\eta_T - \eta_R$.

For AUC and $\mathrm{C}_{\max}$, the FDA (1992) strongly recommends logarithmically transforming the data and testing the hypotheses (2). They offer three rationales for their recommendation. Based on these, the FDA (1992, p. 7) states,

> Based on the arguments in the preceding section, the Division of Bioequivalence recommends that the pharmacokinetic parameters AUC and $\mathrm{C}_{\max}$ be log transformed. Firms are *not* encouraged to test for normality of data distribution after log transformation, nor should they employ normality of data distribution as a justification for carrying out the statistical analysis on the original scale.

The emphasis is ours.

The FDA's three rationales for log transformation are labeled Clinical, Pharmacokinetic, and Statistical. The Clinical Rationale is that the real interest is in the ratio μ_T/μ_R rather than the difference $\mu_T - \mu_R$. But, the link between this fact (which we certainly do not dispute) and the log transformation of the data is based on statistical considerations. It is that a linear statistical model can be used for the transformed data to make inferences about the difference $\eta_T - \eta_R$. These inferences then can be restated in terms of μ_T/μ_R. Thus, the justification of the log transformations seems to be based mainly on the perceived difficulty in dealing with the ratio μ_T/μ_R, rather than the difference $\eta_T - \eta_R$. If appropriate statistical procedures can be used to make inferences about the ratio μ_T/μ_R directly, then there seems to be no need for a log transformation.

The Pharmacokinetic Rationale is based on multiplicative compartmental models of Westlake (1973, 1988). The multiplicative model is changed to a linear model by the log transformation. Part of the Statistical Rationale is that, in the original scale, much bioequivalence data is skewed and appears more lognormal than normal. We agree that these two considerations suggest that the first method of analysis to be considered in bioequivalence studies is on the log transformed data, and, in most cases, this analysis will be appropriate.

The Statistical Rationale consists of the previous lognormal justification and two more points. The first is that,

> Standard parametric methods are ill-suited to making inferences about the ratio of two averages, though some valid methods do exist. Log transformation changes the problem to one of making inferences about the difference (on the log scale) of the two averages, for which the standard methods are well suited.

The second is that the small sample sizes used in typical bioequivalence studies (20 to 30) will produce tests for normality that have fairly low power in either the original or log scale. The FDA recommends that no check of normality be made on the log transformed data. But, if a low-power normality test rejects the hypothesis of normality for the log transformed data, then surely some caution is warranted in the use of procedures that assume normality. In this case, tests such as the TOST, based on the Student's t distribution, are inappropriate. If normality of the log transformed data is rejected and the original data appear more normal than the log transformed data, then procedures that assume normality of the original data would seem more appropriate. In Section 4.3, we show that Sasabuchi (1980,1988a,b) described the size-α likelihood ratio test for (1). It is a simple test based on the Student's t distribution. So the FDA's statement about ill-suited standard parametric procedures seems unfounded. We also show that the tests commonly used are liberal and have size greater than the nominal value of α. Furthermore, we show that the IUT method can be used in this problem, also, to construct size-α tests that are uniformly more powerful than the likelihood ratio test. Thus, the FDA's avoidance of (1) because of statistical difficulties is unwarranted.

An alternative test, when normality is in doubt, might be to use a Wilcoxon-Mann-Whitney analogue of the TOST (based on the original logarithmically transformed data for a parallel design, or the intrasubject between-period differences of the logarithmically transformed data, as proposed by Hauschke, Steinijans and Diletti (1990), for a crossover design).

2.3 $100(1-2\alpha)\%$ confidence intervals

One would expect the TOST to be identical to some *confidence interval* procedure: For some appropriate $100(1-\alpha)\%$ confidence interval $[D^-, D^+]$ for $\eta_T - \eta_R$, declare the test drug to be bioequivalent to the reference drug if and only if $[D^-, D^+] \subset (\theta_L, \theta_U)$.

It has been noted (e.g., Westlake, 1981; Schuirmann, 1981) that the TOST is operationally identical to the procedure of declaring equivalence only if the ordinary $100(1-2\alpha)\%$, not $100(1-\alpha)\%$, two-sided confidence interval for $\eta_T - \eta_R$

$$[D - t_{\alpha,r}\mathrm{SE}(D), D + t_{\alpha,r}\mathrm{SE}(D)] \tag{8}$$

is contained in the interval (θ_L, θ_U). In fact, both FDA (1992) as well as EC-GCP (1993) specify that the TOST should be executed in this fashion.

The fact that the TOST seemingly corresponds to a $100(1-2\alpha)\%$, not $100(1-\alpha)\%$, confidence interval procedure initially caused some concern (Westlake 1976, 1981). Recently, Brown, Casella and Hwang (1995) called this relationship an "algebraic coincidence." But many authors (e.g., Chow and Shao, 1990, and Schuirmann, 1989) have defined bioequivalence tests in terms of $100(1-2\alpha)\%$ confidence sets.

Standard statistical results, such as Theorems 3 and 4 in Section 5, give relationships between size-α tests and $100(1-\alpha)\%$ confidence intervals. In Section 5, we discuss a $100(1-\alpha)\%$ confidence interval that corresponds exactly to the size-α TOST. We also explore the relationship between $100(1-2\alpha)\%$ confidence intervals and size-α tests. We describe situations more general than the TOST in which size-α tests can be defined in terms of $100(1-2\alpha)\%$ confidence intervals. But we also give examples from the bioequivalence literature of tests that have been defined in terms of $100(1-2\alpha)\%$ confidence intervals and sets that are not size-α tests. Tests defined by $100(1-2\alpha)\%$ confidence intervals can be either liberal or conservative. Because of these potential difficulties, our conclusion is that the practice of defining bioequivalence tests in terms of $100(1-2\alpha)\%$ confidence intervals should be abandoned. If both a confidence interval and a test are required, a $100(1-\alpha)\%$ confidence interval that corresponds to the given size-α test should be used.

2.4 Multiparameter problems

In Section 6, we discuss multiparameter bioequivalence problems. We discuss two examples in which the IUT theory can be used to define size-α tests that are uniformly more powerful than tests that have been previously proposed. These examples concern controlling the experimentwise error rate when several parameters are tested for equivalence, simultaneously.

3 Intersection-Union Tests

Berger (1982) proposed the use of intersection-union tests in a quality control context closely related to bioequivalence testing. Tests for many different bioequivalence hypotheses are easily constructed using the IUT method. The TOST is a simple example of an IUT. Tests with a specified size are easily constructed using this method, even in complicated problems involving several parameters. And tests that are uniformly more powerful than standard tests can often be constructed using this method.

The IUT method is useful for the following type of hypothesis testing problem. Let θ denote the unknown parameter (θ can be vector valued) in the distribution of the data $\boldsymbol{X}$. Let Θ denote the parameter space. Let $\Theta_1, \ldots, \Theta_k$ denote subsets of Θ. Suppose we wish to test

$$(9) \qquad \mathrm{H}_0 : \theta \in \bigcup_{i=1}^{k} \Theta_i \quad \text{versus} \quad \mathrm{H}_a : \theta \in \bigcap_{i=1}^{k} \Theta_i^c,$$

where A^c denotes the complement of the set A. The important feature in this formulation is the null hypothesis is expressed as a union and the alternative hypothesis is expressed as an intersection. For $i = 1, \ldots, k$, let R_i denote a rejection region for a test of $\mathrm{H}_{0i} : \theta \in \Theta_i$ versus $\mathrm{H}_{ai} : \theta \in \Theta_i^c$. Then an IUT of (9) is the test that rejects H_0 if and only if $\boldsymbol{X} \in \bigcap_{i=1}^k R_i$. The rationale behind an IUT is simple. The overall null hypothesis, $\mathrm{H}_0 : \theta \in \bigcup_{i=1}^k \Theta_i$ can be rejected only if each of the individual null hypotheses, $\mathrm{H}_{0i} : \theta \in \Theta_i$, can be rejected.

Berger (1982) proved the following two theorems.

Theorem 1 *If R_i is a level-α test of H_{0i}, for $i = 1, \ldots, k$, then the intersection-union test with rejection region $R = \bigcap_{i=1}^k R_i$ is a level-α test of H_0 versus H_a in (9).*

An important feature in Theorem 1 is that each of the individual tests is performed at level-α. But the overall test also has the same level α. There is no need for multiplicity adjustment for performing multiple tests. The reason there is no need for such a correction is the special way the individual tests are combined. H_0 is rejected only if every one of the individual hypotheses, H_{0i}, is rejected.

Theorem 1 asserts that the IUT is level-α. That is, its size is at most α. In fact, a test constructed by the IUT method can be quite conservative. Its size can be much less that the specified value α. But, Theorem 2 (a generalization of Theorem 2 in Berger (1982)) provides conditions under which the IUT is not conservative; its size is exactly equal to the specified α.

Theorem 2 *For some $i = 1, \ldots, k$, suppose R_i is a size-α rejection region for testing H_{0i} versus H_{ai}. For every $j = 1, \ldots, k$, $j \neq i$, suppose R_j is a level-α rejection region for testing H_{0j} versus H_{aj}. Suppose there exists a sequence of parameter points $\theta_l, l = 1, 2, \ldots$, in Θ_i such that*

$$\lim_{l \to \infty} P_{\theta_l}(\boldsymbol{X} \in R_i) = \alpha.$$

and, for every $j = 1, \ldots, k$, $j \neq i$,

$$\lim_{l \to \infty} P_{\theta_l}(\boldsymbol{X} \in R_j) = 1.$$

Then the intersection-union test with rejection region $R = \bigcap_{i=1}^k R_i$ is a size-α test of H_0 versus H_a.

Note that in Theorem 2, the one test defined by R_i has size exactly α. The other tests defined by R_j, $j = 1, \ldots, k$, $j \neq i$, are level-α tests. That is, their sizes may be less than α. The conclusion is the IUT has size α. Thus, if rejection regions $R_1, \ldots, R_k$ with sizes $\alpha_1, \ldots, \alpha_k$ are combined in an IUT and Theorem 2 is applicable, then the IUT will have size equal to $\max_i\{\alpha_i\}$. We will discuss bioequivalence examples in which tests of different sizes are combined. The resulting test has size equal to the maximum of the individual sizes.

4 Old and New Tests for Difference and Ratio Hypotheses

4.1 Two one-sided tests

The TOST is naturally thought of as an IUT. The bioequivalence alternative hypothesis $\mathrm{H}_a : \theta_L < \eta_T - \eta_R < \theta_U$, is conveniently expressed as the intersection of the two sets, $\Theta_1^c = \{(\eta_T, \eta_R, \sigma_D^2) : \eta_T - \eta_R > \theta_L\}$ and $\Theta_2^c = \{(\eta_T, \eta_R, \sigma_D^2) : \eta_T - \eta_R < \theta_U\}$. The test that rejects $\mathrm{H}_{01} : \eta_T - \eta_R \leq \theta_L$ in (5) if $T_L \geq t_{\alpha,r}$ is a size-α test of H_{01}. The test that rejects $\mathrm{H}_{02} : \eta_T - \eta_R \geq \theta_U$ in (6) if $T_U \leq -t_{\alpha,r}$ is a size-α test of H_{02}. So, by Theorem 1, the test that rejects H_0 only if both of these tests reject is a level-α test of (2).

To use Theorem 2 to see that the size of the TOST is exactly α, consider parameter points with $\eta_T - \eta_R = \theta_U$ and take the limit as $\sigma_D^2 \to 0$. Such parameters are on the boundary of H_{02}. Therefore,

$$P(\boldsymbol{X} \in R_2) = P(T_U \leq -t_{\alpha,r}) = \alpha,$$

for any $\sigma_D^2 > 0$. But,

$$P(\boldsymbol{X} \in R_1) = P(T_L \geq t_{\alpha,r}) \to 1, \text{ as } \sigma_D^2 \to 0,$$

because the power of a one-sided t test converges to one as $\sigma_D^2 \to 0$ for any point in the alternative. The value $\eta_T - \eta_R = \theta_U$ is in the alternative, H_{a1}.

The advantage of considering bioequivalence problems in an IUT format is not limited to verifying properties of the TOST. Rather, other bioequivalence hypotheses, such as (1), state an interval as the alternative hypothesis. This interval can be expressed as the intersection of two one-sided intervals. So two one-sided, size-α tests can be combined to obtain a level-α (typically, size-α) test. Furthermore, as we will see in Section 6, even more complicated forms of bioequivalence can be expressed in the IUT format. This allows the easy construction of tests with guaranteed size-α for these problems.

4.2 More powerful tests

Despite its simplicity and intuitive appeal, the TOST suffers from a lack of power. The line labeled TOST in the top part of Table 1 shows the power function, $P(\text{reject } \mathrm{H}_0)$, for parameter points with $\eta_T - \eta_R = \theta_U$ (or θ_L), points on the boundary between H_0 and H_a. The power function is near α for σ_D^2 near 0, but decreases as σ_D^2 grows. An unbiased test would have power equal to α for all such parameter points. The TOST is clearly biased. The bottom part of Table 1 shows the power function when the two drugs are exactly equal, $\eta_T = \eta_R$. The power is near one for σ_D^2 near zero, but decreases to zero as σ_D^2 increases. Despite these shortcomings, Diletti, Hauschke and Steinijans (1991) declared that the TOST maximizes the power among all size-α tests. This is incorrect.

Anderson and Hauck (1983) proposed a test with higher power that the TOST. Whereas the TOST does not reject H_0 if $\mathrm{SE}(D)$ is sufficiently large, the Anderson and Hauck test always rejects H_0 if D is near enough to zero, even if $\mathrm{SE}(D)$ is large. This provides an improvement in power. However the Anderson and Hauck test does not control the Type I error probability at the specified level α. It is liberal and the size is somewhat greater than α. Shortly after Anderson and Hauck proposed their test, Patel and Gupta (1984) and

Table 1: Powers of three bioequivalence tests. $r = 30$, $\alpha = .05$, and $\theta_U = \log(1.25) = -\theta_L$.

	σ_D							
	.00	.04	.08	.12	.16	.20	.30	∞
	$\eta_T - \eta_R = \theta_U$ or θ_L							
TOST	.050	.050	.050	.031	.003	.000	.000	.000
BHM	.050	.050	.050	.050	.050	.050	.050	.050
new	.050	.050	.050	.047	.049	.050	.050	.050
	$\eta_T - \eta_R = 0$							
TOST	1.000	1.000	.720	.158	.007	.000	.000	.000
BHM	1.000	1.000	.721	.260	.131	.093	.066	.050
new	1.000	1.000	.720	.247	.128	.092	.066	.050

Rocke (1984) proposed the same test. This scientific coincidence was commented upon by Anderson and Hauck (1985) and Martin Andrés (1990).

Due to the seriousness of a Type I error, declaring two drugs to be equivalent when they are not, the search for a size-α test that was uniformly more powerful than the TOST continued. Munk (1993) proposed a slightly different test. Munk claims that this test is a size-α test that is uniformly more powerful than the TOST. But this claim is supported by numerical calculations, not analytic results.

Brown, Hwang and Munk (1995) constructed an unbiased, size-α test of (2) that is uniformly more powerful than the TOST. Their construction is recursive. To determine if a point $(d, \mathrm{se}(D))$ is in the rejection region of the Brown, Hwang and Munk test, a good deal of computing can be necessary. This may limit the practical usefulness of the Brown, Hwang and Munk test. Also, sometimes the Brown, Hwang and Munk rejection region has a quite irregular shape. An example of this is shown in Figure 1.

We will now describe a new test of the hypotheses (2). This test is uniformly more powerful than the TOST. Unlike the Anderson and Hauck and Munk tests, our test is a size-α test. Our test is nearly unbiased. It is simpler to compute than the Brown, Hwang and Munk test. It will not have the irregular boundaries that the Brown, Hwang and Munk test sometimes possesses. The construction of this new test again illustrates the usefulness of the IUT method.

To simplify the notation in describing our test, we assume, without loss of generality, that $\theta_L = -\theta_U$ and call $\theta_U = \Delta$. Following Brown, Hwang and Munk, define $S_*^2 = r(\mathrm{SE}(D))^2$. It is simpler to define our test in terms of the polar coordinates, centered at $(\Delta, 0)$,

$$v^2 = (d - \Delta)^2 + s_*^2$$

and

$$b = \cos^{-1}\left((d - \Delta)/v\right).$$

In the (d, s_*) space, v is the distance from $(\Delta, 0)$ to (d, s_*), and b is the angle between the d axis and the line segment joining $(\Delta, 0)$ and (d, s_*). To define a size-α test, we need

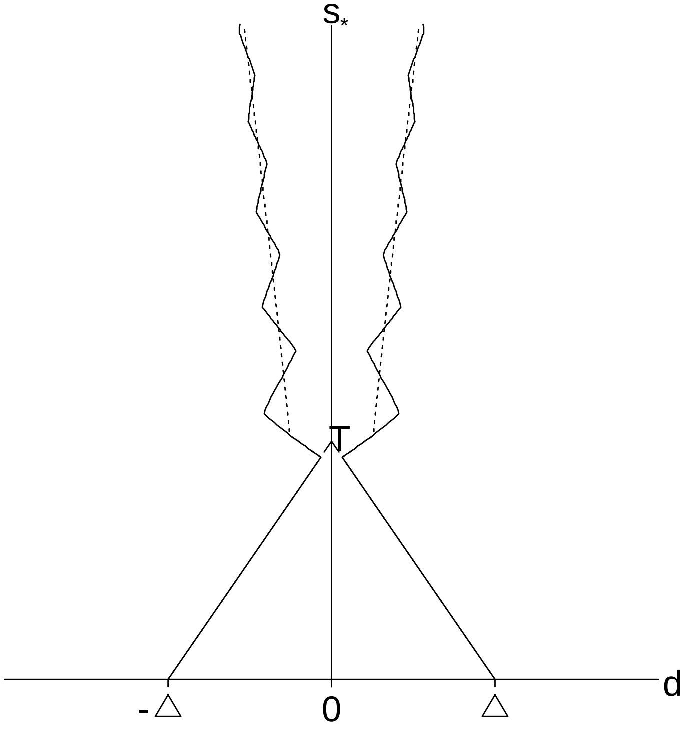

Figure 1: Irregular boundary of Brown, Hwang and Munk test (solid line) and smoother boundary of test from Section 4.2 (dashed line). The TOST rejection region is bounded by the triangle with vertices at $-\Delta$, Δ, and T. Here $r = 3$, $\alpha = .16$ and $-\theta_L = \theta_U = 1$.

the distribution of (V, B) when $\theta = \Delta$, In this case, it is easy to verify that V and B are independent. The probability density function of B is

$$f(b) = \frac{\Gamma((r+1)/2)}{\Gamma(r/2)\sqrt{\pi}}(\sin(b))^{r-1}, \quad 0 < b < \pi,$$

which does not depend on σ_D^2. To implement our test, it is useful to note that the cumulative distribution function of B has a closed form given by

$$F(b) = \frac{b}{\pi} - \frac{1}{2\sqrt{\pi}} \sum_{k=1}^{(r-1)/2} (\sin(b))^{2k-1}\cos(b)\frac{\Gamma(k)}{\Gamma(k+\frac{1}{2})},$$

if r is odd, and

$$F(b) = \frac{1}{2} - \frac{1}{2\sqrt{\pi}} \sum_{k=1}^{r/2} (\sin(b))^{2k-2}\cos(b)\frac{\Gamma(k-\frac{1}{2})}{\Gamma(k)},$$

if r is even. The probability density function of V will be denoted by $g_{\sigma_D}(v)$.

We will describe the rejection region of the new test geometrically here. Exact formulas are in the Appendix. The new test will be an IUT. We will define a size-α, unbiased rejection region, R_2, for testing (6). This R_2 will contain the rejection region of the size-α TOST and will be approximately symmetric about the line $d = 0$. Then we will define $R_1 = \{(d, s_*) : (-d, s_*) \in R_2\}$. R_1 is R_2 reflected across the line $d = 0$. R_1 is a size-α, unbiased rejection region for testing (5). Then $R = R_1 \cap R_2$ is the rejection region of the new test. Because R_2 is approximately symmetric about the line $d = 0$, R_1 is almost the same as R_2, and not much is deleted when we take the intersection. This foresight in choosing the individual rejection regions so that the intersection is not much smaller is always useful when using the IUT method.

The set $\{V = v\}$ is a semicircle in (d, s_*) space. For each value of v, $R_2(v) \equiv \{V = v\} \cap R_2$ is either one or two intervals of b values, that is, one or two arcs on $\{V = v\}$. These arcs will be chosen so that, for every $v > 0$,

$$\int_{R_2(v)} f(b)\, db = \alpha. \tag{10}$$

Then the rejection probability

$$P(R_2) = \int_0^\infty \int_{R_2(v)} f(b)\, db\; g_{\sigma_D}(v)\, dv = \int_0^\infty \alpha g_{\sigma_D}(v)\, dv = \alpha,$$

for every $\sigma_D > 0$ if $\eta_T - \eta_R = \Delta$. This will ensure that R_2 is a size-α, unbiased rejection region for testing (6).

We now define the arc(s) that make up $R_2(v)$. Refer to Figure 2 in this description. The rejection region of the size-α TOST, call it R_T, is the triangle bounded by the lines $s_* = 0$, $d = \Delta - t_{\alpha,r} s_*/\sqrt{r}$ (call this line l_U), and $d = -\Delta + t_{\alpha,r} s_*/\sqrt{r}$ (call this line l_L). Let v_0 denote the distance from $(\Delta, 0)$ to l_L. In this description, we assume $1/2 > \alpha > \alpha_* \equiv 1 - F(3\pi/4)$. Brown, Hwang and Munk (1995) in their Table 1 show that if $r \geq 4$, then $\alpha = .05 > \alpha_*$. The new test for $\alpha \leq \alpha_*$ is given in the Appendix. Brown, Hwang and Munk did not propose any test for $\alpha \leq \alpha_*$. The condition $\alpha > \alpha_*$ ensures that the point on l_L closest to $(\Delta, 0)$ is on the boundary of R_T, as shown.

Let b_0 denote the angle between the d axis and l_U. For $0 < v \leq v_0$, $R_2(v) = \{b : b_0 < b < \pi\}$. The arc A_0 in Figure 2 is an example of such an arc. So, for $v < v_0$, $R_2(v)$ is exactly the points in the TOST.

For $v_0 < v$, the semicircle $V = v$ intersects l_L at two points. Let $b_1 < b_2$ denote the angles corresponding to these two points. If $v_0 < v < 2\Delta$, let $A_2(v) = \{b : b_2 < b < \pi\}$. These are the points in R_T adjacent to the d axis, and A_2 in Figure 2 is an example of such an arc. If $2\Delta \leq v$, let $A_2(v)$ be the empty set. Let $\alpha(v)$ denote the probability content of $A_2(v)$ under F. That is,

$$\alpha(v) = \begin{cases} 1 - F(b_2), & v_0 < v < 2\Delta, \\ 0, & 2\Delta \leq v. \end{cases}$$

For $v_0 < v$, $R_2(v) = A_1(v) \cup A_2(v)$, where to ensure that (10) is true, $A_1(v)$ must satisfy

$$\int_{A_1(v)} f(b)\, db = \alpha - \alpha(v). \tag{11}$$

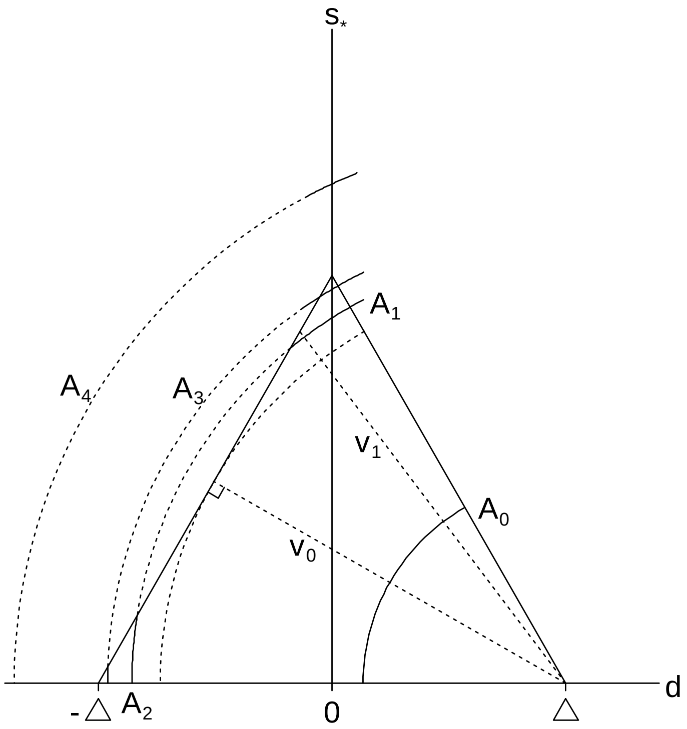

Figure 2: Arcs that define the rejection region R_2.

Let (d_1, s_{*1}) denote the point where the $\{V = v_0\}$ semicircle intersects l_U, and let v_1 denote the radius corresponding to $(-d_1, s_{*1})$. For $v_0 < v < v_1$, let b_{L1} be the angle defined by

$$F(b_1) - F(b_{L1}) = \alpha - \alpha(v), \tag{12}$$

where b_1 is as defined in the previous paragraph. Then $A_1(v) = \{b : b_{L1} < b < b_1\}$ is the arc that satisfies (11) whose endpoint is on l_L. For $v_0 < v < v_1$, $R_2(v) = A_1(v) \cup A_2(v)$, using this $A_1(v)$. The arcs labeled A_1 and A_2 in Figure 2 comprise such an $R_2(v)$. For $v < v_1$, the cross sections $R_2(v)$ we have defined are the same as the cross sections for the Brown, Hwang and Munk (1995) test. They now define the remainder of their rejection region recursively in terms of these arcs. We define our rejection region in a nonrecursive manner.

For $v_1 \le v$, define two values $b_L(v) < b_U(v)$ such that $F(b_U(v)) - F(b_L(v)) = \alpha - \alpha(v)$, and the angle between the line joining $(0, 0)$ and $(v, b_L(v))$ and the s_* axis is the same as the angle between the line joining $(0, 0)$ and $(v, b_U(v))$ and the s_* axis. This equal angle condition is what we meant earlier by the phrase "approximately symmetric about the line $d = 0$." If $b_U(v) \ge b_1$, then $A_1(v) = \{b : b_L(v) < b < b_U(v)\}$. But, if $b_U(v) < b_1$, then this arc does not contain all the points in the TOST. So, if $b_U(v) < b_1$, $A_1(v) = \{b : b_{L1} < b < b_1\}$, where b_{L1} is defined by (12). For $v_1 \le v$, $R_2(v) = A_1(v) \cup A_2(v)$. Recall, if $2\Delta \le v$, $A_2(v)$ is empty, and $R_2(v)$ is the single arc $A_1(v)$. Also, for $v^2 \ge \max\{4\Delta^2, \Delta^2 + \Delta^2 r/t^2_{\alpha,r}\}$, the semicircle $\{V = v\}$ does not intersect R_T, and $R_2(v)$ is the arc defined by $b_L(v)$ and $b_U(v)$. The b_1 condition never applies in this case. In Figure 2, the solid parts of the arcs A_3 and A_4 are examples of $R_2(v)$ for $v_1 \le v$.

The cross-sections $R_2(v)$ have been defined for every $v > 0$, and this defines R_2. R_1 is the reflection of R_2 across the s_* axis, and the rejection region of the new test is $R = R_1 \cap R_2$. This construction is illustrated in Figure 3.

In Figure 1, the rejection region R with the same size as the Brown, Hwang and Munk test is the region between the dotted lines. The boundary of R is smooth compared to the irregular boundary of the Brown, Hwang and Munk test. This smoothness results from the attempt in the construction of R to center arcs around the s_* axis. To determine if a sample point (d, s_*^2) is in R, two arcs, $R_2(v)$ and $R_1(v) = R_2(v')$ ($v' = (-d - \Delta)^2 + s_*^2$, computed from $(-d, s_*^2)$), must be constructed. If (d, s_*^2) is on both arcs, $(d, s_*^2) \in R$. But, to determine if (d, s_*^2) is in the rejection region of the Brown, Hwang and Munk test, a starting point is selected. Then a sequence of arcs is constructed until (d, s_*^2) is passed. Then another sequence of arcs is constructed from a new starting point. This process is continued until enough arcs in the vicinity of (d, s_*^2) are obtained to approximate the boundary of the rejection region. From this it is determined if (d, s_*^2) is in the rejection region. Thus, a good deal more computation is needed to implement the Brown, Hwang and Munk test. Also, the Brown, Hwang and Munk test is not defined for $\alpha \le \alpha_*$. This smoothness, general applicability, and simplicity of computation recommends R as a reasonable alternative to the Brown, Hwang and Munk test. But R is slightly biased whereas the Brown, Hwang and Munk test is unbiased.

A small power comparison of the TOST, Brown, Hwang and Munk test, and our new test is given in Table 1 for $\alpha = .05$ and $r = 30$. In the top block of numbers, $\eta_T - \eta_R = \Delta$. For these boundary values, the power is exactly $\alpha = .05$ for the unbiased Brown, Hwang and Munk test. The power is also very close to .05 for our test, indicating it has only slight

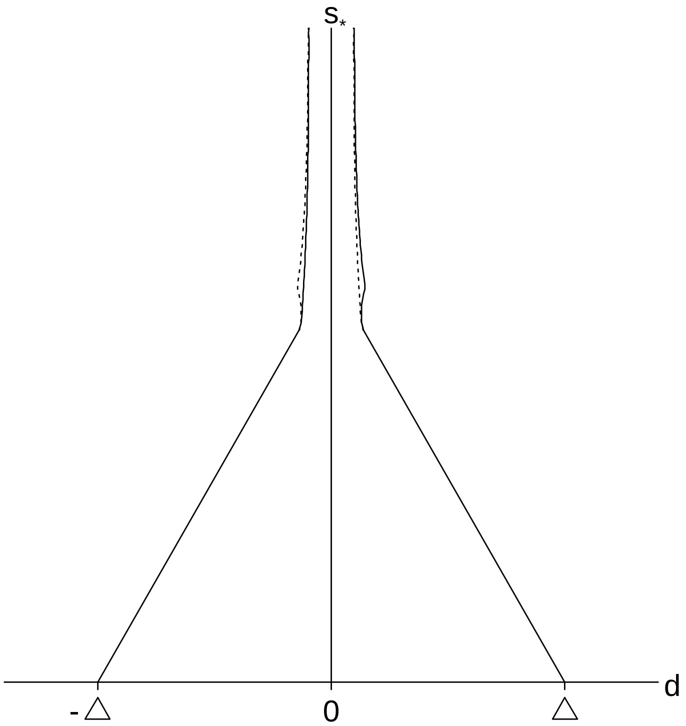

Figure 3: Rejection region of new test. Region R_2 (between solid lines) and region R_1 (between dashed lines). Rejection region $R = R_1 \cap R_2$. $r = 10$ and $\alpha = .05$.

bias. But the TOST is highly biased with power much less than .05 for moderate and large σ_D. In the bottom block of numbers, $\eta_T - \eta_R = 0$. The drugs are equivalent. Our test and the Brown, Hwang and Munk test have very similar powers. Their powers are much greater than the TOST's power for all but small σ_D. For example, it can be seen that the power improvement is about 60% when $\sigma_D = .12$ and about 85% when $\sigma_D = .16$. Sample sizes for bioequivalence tests are often chosen so that the test has power of about .8 when $\eta_T = \eta_R$. In this case, Table 1 indicates there is no advantage to using the new tests over the TOST. But if the variability turns out to be larger than expected in the planning stage, the new tests offer significant power improvements.

The tests of Anderson and Hauck (1983), Brown, Hwang and Munk (1995), and our new test all have the property that, as $s_* \to \infty$, the width of the rejection region increases, eventually containing values of (d, s_*) with d outside the interval (θ_L, θ_U). There will be values (d, s_{*1}) and (d, s_{*2}) with $s_{*1} < s_{*2}$, but (d, s_{*1}) is not in the rejection region while (d, s_{*2}) is in the rejection region. This "flaring out" of the rejection region is evident in Figures 1 and 5. This counterintuitive shape was pointed out by Rocke (1984). The rejection region of any bioequivalence test that is unbiased, or approximately unbiased, must eventually contain sample points with d outside the interval (θ_L, θ_U). Some have suggested that such procedures should be truncated in the sense that the narrowest point of the rejection region be determined and then the rejection region is extended along the s_* axis only of this width. Brown, Hwang and Munk suggest this as a possible modification of their test, although the resulting test will no longer be unbiased. We believe that notions of size, power, and unbiasedness are more fundamental than "intuition" and do not recommend truncation. But for those who disagree, our new test could be truncated in this same way. The narrowest point will need to be determined numerically for all these tests, and the smoother shape of our rejection region will make this determination easier. Referring to Figure 1, a numerical routine might be fooled by the irregular shape of the Brown, Hwang and Munk test.

4.3 Tests for ratios of parameters

Usually, data from a bioequivalence trial is logarithmically transformed before analysis. This leads to a test of the hypotheses (2), as described in the previous section. In the model we will consider now, the original data are normally distributed. Let $X_1, \ldots, X_m$ form a random sample from a normal population with mean μ_T and variance σ^2, and $Y_1, \ldots, Y_n$ form an independent random sample from a normal population with mean μ_R and variance σ^2. In this section, we will present our comments in terms of this simple parallel design. Yang (1991) and Liu and Weng (1995) describe models for this normally distributed data in crossover experiments.

The bioequivalence hypothesis to be tested in this case is (1), namely,

$$\begin{array}{c} \mathrm{H}_0 : \frac{\mu_T}{\mu_R} \le \delta_L \text{ or } \frac{\mu_T}{\mu_R} \ge \delta_U \\ \text{versus} \\ \mathrm{H}_a : \delta_L < \frac{\mu_T}{\mu_R} < \delta_U . \end{array} \tag{13}$$

In the past, the values of $\delta_L = .80$ and $\delta_U = 1.20$ were commonly used (called the ±20 rule). But, the FDA Division of Bioequivalence (1992) now uses $\delta_L = .80$ and $\delta_L = 1.25$.

These limits are symmetric about one in the ratio scale since $.80 = 1/1.25$.

The parameter μ_R is positive because the measured variable, AUC or $\mathrm{C}_{\max}$, is positive. Therefore the hypotheses (13) can be restated as

$$\text{(14)} \qquad \begin{array}{ll} & \mathrm{H}_0 : \mu_T - \delta_L\mu_R \leq 0 \text{ or } \mu_T - \delta_U\mu_R \geq 0 \\ \text{versus} & \\ & \mathrm{H}_a : \mu_T - \delta_L\mu_R > 0 \text{ and } \mu_T - \delta_U\mu_R < 0. \end{array}$$

The testing problem (14) was first considered by Sasabuchi (1980, 1988a,b). Let $\overline{X}$, $\overline{Y}$, and S^2 denote the two sample means and the pooled estimate of σ^2. Sasabuchi showed that the size-α likelihood ratio test of (14) rejects H_0 if and only if

$$T_1 \geq t_{\alpha,r} \quad \text{and} \quad T_2 \leq -t_{\alpha,r}$$

where

$$T_1 = \frac{\overline{X} - \delta_L\overline{Y}}{S\sqrt{\frac{1}{m} + \frac{\delta_L^2}{n}}} \quad \text{and} \quad T_2 = \frac{\overline{X} - \delta_U\overline{Y}}{S\sqrt{\frac{1}{m} + \frac{\delta_U^2}{n}}}.$$

This will be called the T_1/T_2 test.

The T_1/T_2 test is easily understood as an IUT. The usual, normal theory, size-α t-test of $\mathrm{H}_{01} : \mu_T - \delta_L\mu_R \leq 0$ versus $\mathrm{H}_{a1} : \mu_T - \delta_L\mu_R > 0$ is the test that rejects H_{01} if $T_1 \geq t_{\alpha,r}$. Similarly, the usual, normal theory, size-α t-test of $\mathrm{H}_{02} : \mu_T - \delta_U\mu_R \geq 0$ versus $\mathrm{H}_{a2} : \mu_T - \delta_U\mu_R < 0$ is the test that rejects H_{02} if $T_2 \leq -t_{\alpha,r}$. Because H_a is the intersection of H_{a1} and H_{a2}, these two t-tests can be combined, using the IUT method, to get a level-α test of H_0 versus H_a. Using an argument like in Section 3, Theorem 2 can be used to show that the size of this test is α.

Yang (1991) and Liu and Weng (1995) proposed tests closely related to the T_1/T_2 test for the bioequivalence problem of testing (13) in a crossover experiment. Hauck and Anderson (1992) also discuss the hypotheses in the form (14). But no reference to Sasabuchi's earlier work is given. The derivation of the confidence set for μ_T/μ_R in Hsu, Hwang, Liu, and Ruberg (1994) contains a mistake in the standardization. Properly corrected, their rather complicated confidence set would lead to the rejection of (14) when the simple test described above does. So, somehow, the value of this simple, size-α test seems to have been completely overlooked in the bioequivalence literature. Rather, Chow and Liu (1992) and Liu and Weng (1995) both report that the following is the standard analysis. Rewrite the hypotheses (13) or (14) as

$$\text{(15)} \qquad \begin{array}{ll} & \mathrm{H}_0 : \mu_T - \mu_R \leq (\delta_L - 1)\mu_R \text{ or } \mu_T - \mu_R \geq (\delta_U - 1)\mu_R \\ \text{versus} & \\ & \mathrm{H}_a : (\delta_L - 1)\mu_R < \mu_T - \mu_R < (\delta_U - 1)\mu_R. \end{array}$$

These hypotheses look like (2), but there is an important difference. In (2), θ_L and θ_U are known constants. In (15), $(\delta_L - 1)\mu_R$ and $(\delta_U - 1)\mu_R$ are unknown parameters. Nevertheless, the standard analysis proceeds to use the TOST with $(\delta_L - 1)\overline{Y}$ replacing θ_L in T_L and $(\delta_U - 1)\overline{Y}$ replacing θ_U in T_U. The standard analysis ignores the fact that a constant has been replaced by a random variable and compares these two test statistics to standard t percentiles as in the TOST. This test will be called the T_1^*/T_2^* test.

The statistics that are actually used in this analysis are

$$T_1^* = \frac{\overline{X} - \overline{Y} - (\delta_L - 1)\overline{Y}}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{\overline{X} - \delta_L \overline{Y}}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} = T_1\sqrt{\frac{n+m\delta_L^2}{n+m}},$$

and

$$T_2^* = \frac{\overline{X} - \overline{Y} - (\delta_U - 1)\overline{Y}}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{\overline{X} - \delta_U \overline{Y}}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} = T_2\sqrt{\frac{n+m\delta_U^2}{n+m}}.$$

The statistics T_1 and T_2 are properly scaled to have Student's t distributions, but T_1^* and T_2^* are not. The T_1^*/T_2^* test is an IUT in which the two tests have different sizes. The test that rejects H_{01} if $T_1^* > t_{\alpha,r}$ has size

$$\begin{aligned} P_{\mu_T=\delta_L\mu_R}(T_1^* > t_{\alpha,r}) &= P_{\mu_T=\delta_L\mu_R}\left(T_1 > \sqrt{\frac{n+m}{n+m\delta_L^2}}t_{\alpha,r}\right) \\ &= \alpha_1 < \alpha, \end{aligned}$$

because

$$\sqrt{\frac{n+m}{n+m\delta_L^2}} > 1.$$

On the other hand, the test that rejects H_{02} if $T_2^* < -t_{\alpha,r}$ has size

$$\begin{aligned} P_{\mu_T=\delta_U\mu_R}(T_2^* < -t_{\alpha,r}) &= P_{\mu_T=\delta_U\mu_R}\left(T_2 < -\sqrt{\frac{n+m}{n+m\delta_U^2}}t_{\alpha,r}\right) \\ &= \alpha_2 > \alpha, \end{aligned}$$

because

$$\sqrt{\frac{n+m}{n+m\delta_U^2}} < 1.$$

Theorem 2 can be used to show that, as a test of the hypothesis (13), the T_1^*/T_2^* test has size $\alpha_2 > \alpha$. It is a liberal test.

The true size of the T_1^*/T_2^* test, for a nominal size of $\alpha = .05$, is shown in Table 2. In Table 2 it is assumed that the sample sizes from the test and reference drugs are equal, $m = n$. In this case, the size of the T_1^*/T_2^* test is simply

$$\alpha_2 = P\left(T < -\sqrt{\frac{2}{1+\delta_U^2}}t_{\alpha,r}\right),$$

where T has a students t distribution with $r = 2n-2$ degrees of freedom. It can be seen that the size of the T_1^*/T_2^* test is about .07 for all sample sizes. The liberality worsens slightly as the sample size increases.

On the other hand, the T_1/T_2 test has size exactly equal to the nominal α. It is just as simple to implement as the T_1^*/T_2^* test. Therefore the T_1/T_2 test should replace the T_1^*/T_2^* test for testing (13).

Table 2: Actual size of T_1^*/T_2^* test for nominal $\alpha = .05$

$m = n$	5	10	15	20	30	∞
size	.070	.071	.072	.072	.073	.073

In Section 4.2, the IUT method was used to construct a size-α test that is uniformly more powerful than the TOST. For the known σ^2 case, Berger (1989) and Liu and Berger (1995) used the IUT method to construct size-α tests that are uniformly more powerful than the T_1/T_2 test. In Figure 4, the cone shaped region labeled R_o is the rejection region of the T_1/T_2 test for $\alpha = .05$. The region between the dashed lines is the rejection region of Liu and Berger's size-α test that is uniformly more powerful. We refer the reader to Berger (1989) and Liu and Berger (1995) for the details about these tests. We believe that for the σ^2 unknown case, size-α tests that are uniformly more powerful than the T_1/T_2 test will be found.

5 Confidence Sets and Bioequivalence Tests

5.1 A $100(1-\alpha)\%$ confidence interval

We will show that the $100(1-\alpha)\%$ confidence interval $[D_1^-, D_1^+]$ given by

$$\left[(D - t_{\alpha,r}\mathrm{SE}(D))^-, (D + t_{\alpha,r}\mathrm{SE}(D))^+\right] \tag{16}$$

corresponds to the size-α TOST for (2). Here $x^- = \min\{0, x\}$ and $x^+ = \max\{0, x\}$. The $100(1-\alpha)\%$ interval (16) is equal to the $100(1-2\alpha)\%$ interval (8) when the interval (8) contains zero. But, when the interval (8) lies to the right (left) of zero, the interval (16) extends from zero to the upper (lower) endpoint of interval (8).

The confidence interval (16) has been derived by Hsu (1984), Bofinger (1985), and Stefansson, Kim, and Hsu (1988) in the multiple comparisons setting, and by Müller-Cohrs (1991), Bofinger (1992), and Hsu, Hwang, Liu, and Ruberg (1994) in the bioequivalence setting. Our derivation follows Stefansson, Kim, and Hsu (1988) and Hsu, Hwang, Liu, and Ruberg (1994), which makes the correspondence to TOST more explicit.

To see this correspondence, we use the standard connection between tests and confidence sets. Most often in statistics, this connection is used to construct confidence sets from tests via a result such as the following.

Theorem 3 (Lehmann, 1986, p. 90) *Let the data $\boldsymbol{X}$ have a probability distribution that depends on a parameter $\boldsymbol{\theta}$. Let Θ denote the parameter space. For each $\boldsymbol{\theta}_0 \in \Theta$, let $A(\boldsymbol{\theta}_0)$ denote the acceptance region of a level-α test of $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. That is, for each $\boldsymbol{\theta}_0 \in \Theta$, $P_{\theta=\theta_0}(\boldsymbol{X} \in A(\boldsymbol{\theta}_0)) \geq 1 - \alpha$. Then, $C(\boldsymbol{x}) = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{x} \in A(\boldsymbol{\theta})\}$ is a level $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$.*

But in bioequivalence testing in the past, tests have often been constructed from confidence sets. A result related to this practice follows.

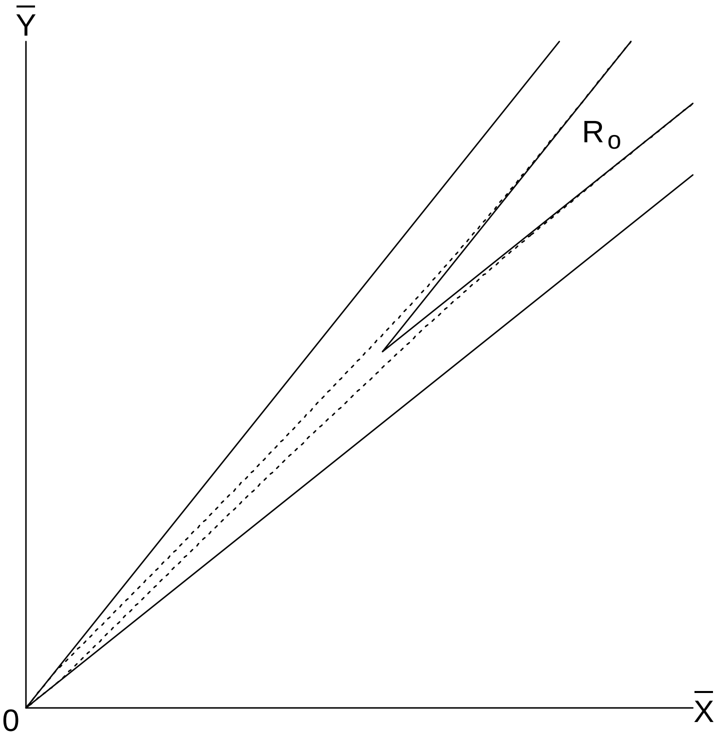

Figure 4: Rejection region for T_1/T_2 test is cone shaped R_o. Region between dashed lines is rejection region of uniformly more powerful Liu and Berger (1995) test. The estimates $\overline{X}$ and $\overline{Y}$ satisfy $\delta_L < \overline{X}/\overline{Y} < \delta_U$ in the larger cone shaped region.

Theorem 4 *Let the data $\boldsymbol{X}$ have a probability distribution that depends on a parameter $\boldsymbol{\theta}$. Suppose $C(\boldsymbol{X})$ is a $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$. That is, for each $\boldsymbol{\theta} \in \Theta$, $P_\theta(\boldsymbol{\theta} \in C(\boldsymbol{X})) \geq 1-\alpha$. Consider testing $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$ versus $\mathrm{H}_a : \boldsymbol{\theta} \in \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$. Then, the test that rejects H_0 if and only if $C(\boldsymbol{X}) \cap \Theta_0 = \emptyset$ is a level-α test of H_0.*

Proof. Let $\boldsymbol{\theta}_0 \in \Theta_0$. Then

$$P_{\theta_0}(\text{reject } H_0) \leq 1 - P_{\theta_0}(\boldsymbol{\theta}_0 \in C(\boldsymbol{X})) \leq \alpha.$$

Unfortunately, Theorem 4 has not always been carefully applied in the bioequivalence area. Commonly, $100(1-2\alpha)\%$ confidence sets are used in an attempt to define level-α tests. Theorem 4 guarantees only that a level-2α test will result from a $100(1-2\alpha)\%$ confidence set. Sometimes, the size of the resulting test is, in fact, α, but this is not generally true. In this subsection we use Theorem 3 to show the correspondence between the $100(1-\alpha)\%$ confidence interval (16) and the size-α TOST. In the next subsection, we criticize the practice of using $100(1-2\alpha)\%$ confidence sets to define bioequivalence tests.

Let $\theta = \eta_T - \eta_R$. The family of size-α tests with acceptance regions

$$A(\theta_0) = \Big\{(d, \mathrm{se}(D)) : |d - \theta_0| \leq t_{\alpha/2,r}\mathrm{se}(D)\Big\} \tag{17}$$

leads to usual equivariant confidence interval, which is of the form (8) but with $t_{\alpha,r}$ replaced by $t_{\alpha/2,r}$.

However, no current law or regulation states one must employ confidence sets that are equivariant over the entire real line. Using Theorem 3 and inverting the family of size-α tests defined by, for $\theta_0 \geq 0$,

$$A(\theta_0) = \{(d, \mathrm{se}(D)) : d - \theta_0 \geq -t_{\alpha,r}\mathrm{se}(D)\} \tag{18}$$

and for $\theta_0 < 0$,

$$A(\theta_0) = \{(d, \mathrm{se}(D)) : d - \theta_0 \leq t_{\alpha,r}\mathrm{se}(D)\} \tag{19}$$

yields the $100(1-\alpha)\%$ confidence interval (16). Technically, when inverting (18) and (19), the upper confidence limit will be open when $D + t_{\alpha,r}\mathrm{SE}(D) < 0$. This point is inconsequential in bioequivalence testing. The only value of the upper bound with positive probability is 0, and, in bioequivalence testing, the inference $\eta_T \neq \eta_R$ is not of interest. In terms of operating characteristics, the confidence interval with the possibly open endpoint has coverage probability $100(1-\alpha)\%$ everywhere. The confidence interval (16) also has coverage probability $100(1-\alpha)\%$ except at $\eta_T - \eta_R = 0$ where it has 100% coverage probability.

Note that the family of tests (18) contains the one-sided size-α t-test for (6), and the family of tests (19) contains the one-sided size-α t-test for (5), in contrast to the family of tests (17). The 5% TOST is equivalent to asserting bioequivalence, $\theta_L < \eta_T - \eta_R < \theta_U$, if and only if the 95% confidence interval $[D_1^-, D_1^+] \subset (\theta_L, \theta_U)$. Therefore, as pointed out by Hsu, Hwang, Liu, and Ruberg (1994), it is more consistent with standard statistical theory to say that the $100(1-\alpha)\%$ confidence interval $[D_1^-, D_1^+]$, instead of the ordinary $100(1-2\alpha)\%$ confidence interval (8), corresponds to the TOST.

Pratt (1961) showed that for the $r = \infty$ case (i.e. $\mathrm{SE}(D) = \sigma_D$), when $\eta_T = \eta_R$, that is, when the test drug is indeed equivalent to the reference drug, $[D_1^-, D_1^+]$ has the smallest expected length among all $100(1-\alpha)\%$ confidence intervals for $\eta_T - \eta_R$. On the other hand, when $\eta_T - \eta_R$ is far from zero, $[D_1^-, D_1^+]$ has larger expected length than the equivariant confidence interval (8). So the bioequivalence confidence interval $[D_1^-, D_1^+]$ can be thought of as specifically constructed from Theorem 3 for more precise inference when it is expected that η_T is close to η_R. One multi-parameter extension of this construction, utilized by Stefansson, Kim, and Hsu (1988), gives rise to the multiple comparison with the best (MCB) confidence intervals of Hsu (1984), which eliminate treatments that are not the best and identify treatments close to the true best. In fact, the bioequivalence confidence interval (16) is an MCB confidence interval because, when only two treatments are being compared, a treatment close to the other treatment is either the true best treatment or close to the true best treatment.

This ability of MCB confidence interval to give *practical equivalence* inference is useful in another problem. Ruberg and Hsu (1992) pointed out that whether to include certain parameters in a regression model should sometimes be formulated as a practical equivalence problem rather than a significant difference problem. In modeling the stability of a drug, for example, given the clear intent of the FDA (1987) Guideline that data from batches of a drug can be pooled only if they have practically equivalent degradation rates, the decision of which *time* × *batch* interaction terms to include in the model can logically be based on MCB confidence intervals comparing the degradation rate of each batch with the true worst degradation rate. Another problem which has not been but should be formulated as one of practical equivalence is the establishment of safety of substances such as bovine growth hormone in toxicity studies (e.g., Juskevich and Guyer, 1990), since the desired inference is practical equivalence between the treated groups and the (negative) control group (cf. Hsu, 1996, Chapter 2).

A different multiparameter extension of the same construction was utilized by Brown, Casella, and Hwang (1995) to obtain the confidence region for a vector parameter $\boldsymbol{\theta}$ which has the smallest expected volume when $\boldsymbol{\theta} = \mathbf{0}$, generalizing Pratt's result. The confidence set is constructed through Theorem 3 using the family of size-α Neyman-Pearson likelihood ratio tests for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_a : \boldsymbol{\theta} = \mathbf{0}$. When $\hat{\boldsymbol{\theta}}$ is multivariate normal with unknown mean vector $\boldsymbol{\theta}$ and known variance-covariance matrix Σ, the acceptance regions are

$$A(\boldsymbol{\theta}_0) = \left\{\hat{\boldsymbol{\theta}} : \boldsymbol{\theta}_0'\Sigma^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/\sqrt{\boldsymbol{\theta}_0'\Sigma^{-1}\boldsymbol{\theta}_0} > -t_{\alpha,\infty}\right\},$$

which leads to the confidence region

$$C(\hat{\boldsymbol{\theta}}) = \left\{\boldsymbol{\theta} : \boldsymbol{\theta}'\Sigma^{-1}\hat{\boldsymbol{\theta}}/\sqrt{\boldsymbol{\theta}'\Sigma^{-1}\boldsymbol{\theta}} + t_{\alpha,\infty} > \sqrt{\boldsymbol{\theta}'\Sigma^{-1}\boldsymbol{\theta}}\right\}. \tag{20}$$

Their paper describes and illustrates interesting geometric properties of $C(\hat{\boldsymbol{\theta}})$.

It should be pointed out that the utility of Theorem 3 is not restricted to the construction of confidence sets which give better practical equivalence inference. Stefansson, Kim, and Hsu (1988) and Hayter and Hsu (1994) used Theorem 3 to construct confidence sets associated with step-down and step-up multiple comparison methods, which are usually thought of as specifically constructed to give better significant difference inference than single-step methods.

5.2 $100(1-2\alpha)\%$ confidence intervals

Bioequivalence tests are often defined in terms of $100(1-2\alpha)\%$ confidence sets. That is, if $\boldsymbol{\theta}$ denotes the parameter of interest, Θ_0^c denotes the set of parameter values for which the drugs are bioequivalent, and $C(\boldsymbol{X})$ is a $100(1-2\alpha)\%$ confidence set for $\boldsymbol{\theta}$, then the drugs are declared bioequivalent if and only if $C(\boldsymbol{X}) \subset \Theta_0^c$. This practice seems to be based entirely on the perceived equivalence between the $100(1-2\alpha)\%$ confidence interval (8) and the size-α TOST of (2). This practice is encouraged by the fact that both FDA (1992) and EC-GCP (1993) specify that the $\alpha = .05$ TOST should be executed by constructing a 90% confidence interval. In the bioequivalence literature, when used in this way, the 90% is called the *assurance* of the confidence set.

The intent of the regulating agencies is clearly to use a test with size $\alpha = .05$. Unfortunately, bioequivalence tests have been proposed using $100(1-2\alpha)\%$ confidence sets without any verification that the resulting tests have size-α. Theorem 4 guarantees that the resulting test is a level-2α test, not size-α. In this section, we will explore the usage of $100(1-2\alpha)\%$ confidence sets. We shall show that the usual $100(1-2\alpha)\%$ confidence interval (8) results in a size-α TOST of (2) because (8) is "equal-tailed." So the relationship is deeper than the "algebraic coincidence" mentioned by Brown, Casella and Hwang (1995). Hauck and Anderson (1992) discuss this fact without proof. But we shall see in examples that the use of $100(1-2\alpha)\%$ confidence sets can result in both liberal and conservative bioequivalence tests. Because there is no general guarantee that a $100(1-2\alpha)\%$ confidence set will result in a size-α test, we believe it is unwise to attempt to define a size-α test in terms of a $100(1-2\alpha)\%$ confidence set. Rather, a test with the specified Type I error probability of α should be used. Theorem 3 might be used to construct the corresponding $100(1-\alpha)\%$ confidence set.

Let $[C^-, C^+]$ denote (8), the usual $100(1-2\alpha)\%$ confidence interval for $\eta_T - \eta_R$. Why does rejecting H_0 in (2) if and only if $[C^-, C^+] \subset (\theta_L, \theta_U)$ result in a size-α testΓ The superficial answer is that, obviously, $C^+ < \theta_U$ is equivalent to $T_U < -t_{\alpha,r}$ and $C^- > \theta_L$ is equivalent to $T_L > t_{\alpha,r}$. Thus, the test based on $[C^-, C^+]$ is equivalent to the size-α TOST. But a more thorough understanding of this is suggested by the following result (Exercise 9.1, Casella and Berger, 1990).

Theorem 5 *Let the data $\boldsymbol{X}$ have a probability distribution that depends on a real-valued parameter θ. Suppose $(-\infty, U(\boldsymbol{X})]$ is a $100(1-\alpha_1)\%$ upper confidence bound for θ. Suppose $[L(\boldsymbol{X}), \infty)$ is a $100(1-\alpha_2)\%$ lower confidence bound for θ. Then, $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is a $100(1-\alpha_1-\alpha_2)\%$ confidence interval for θ.*

Now consider the $100(1-2\alpha)\%$ confidence interval $[C^-, C^+]$ for $\theta = \eta_T - \eta_R$. The interval $(-\infty, C^+]$ is a $100(1-\alpha)\%$ upper confidence bound for θ. From Theorem 4, the test that rejects H_{02} in (6) if and only if $C^+ < \theta_U$ is a level-α test of H_{02}. Likewise, $[C^-, \infty)$ is a $100(1-\alpha)\%$ lower confidence bound for θ and the test that rejects H_{01} in (5) if and only if $C^- > \theta_L$ is a level-α test of H_{01}. Forming an IUT from these two level-α tests yields a level-α test of H_0 in (2), by Theorem 1. Thus, we see that it is not so important that $[C^-, C^+]$ is a $100(1-2\alpha)\%$ confidence interval for θ. Rather, it is the fact that $(-\infty, C^+]$ and $[C^-, \infty)$ are both $100(1-\alpha)\%$ confidence intervals that yields a level-α test. That is, it is important that $[C^-, C^+]$ is an "equal-tailed" confidence interval.

It is easy to see that $100(1-2\alpha)\%$ confidence intervals will not always yield size-α tests. Consider an "unequal-tailed" $100(1-2\alpha)\%$ confidence interval for $\theta = \eta_T - \eta_R$, $[C_1^-, C_1^+]$, defined by

$$[D - t_{\alpha_2,r}\mathrm{SE}(D),\ D + t_{\alpha_1,r}\mathrm{SE}(D)]\,, \tag{21}$$

where $\alpha_1 + \alpha_2 = 2\alpha$. Using $(-\infty, C_1^+]$ to define a test of H_{02} yields a size-α_1 test. And using $[C_1^-, \infty)$ to define a test of H_{01} yields a size-α_2 test. Therefore, by Theorem 1, the IUT that rejects H_0 if and only if $[C_1^-, C_1^+] \subset (\theta_L, \theta_U)$ has level $\max\{\alpha_1, \alpha_2\}$. That this test has size equal to $\max\{\alpha_1, \alpha_2\}$ can be verified using Theorem 2. This relationship between the size of the test and the maximum of the one-sided error probabilities is alluded to by equation (1) in Yee (1986). The size of this test can be made arbitrarily close to 2α by choosing α_1 close to zero and α_2 close to 2α. In this problem, the only $100(1-2\alpha)\%$ confidence interval of the form (21) that defines a size-α test happens to be the usual, equal-tailed confidence interval, $[C^-, C^+]$.

The preceding example using an unequal-tailed test simply illustrates that defining a bioequivalence test in terms of a $100(1-2\alpha)\%$ confidence interval can lead to a liberal test with size greater than α. But, no one has proposed using the interval (21) to define a bioequivalence test. So we now discuss two other examples that have been proposed in the bioequivalence literature. Both examples concern testing (1) about the ratio μ_T/μ_R.

Tests based on $100(1-2\alpha)\%$ Fieller-type confidence intervals provide examples of tests that are sometimes liberal. Mandallaz and Mau (1981), Locke (1984) and Kinsella (1989) all propose using a Fieller-type (1940, 1954) confidence interval to estimate μ_T/μ_R. Neither Locke nor Kinsella propose constructing a bioequivalence test using this interval. But Mandallaz and Mau (1981), Yee (1986,1990), Metzler (1991) and Schuirmann (1989) all propose defining a test of (1) using these Fieller confidence intervals, and all suggest that a $100(1-2\alpha)\%$ confidence interval should be used. A test defined in this way using the Locke $100(1-2\alpha)\%$ confidence interval is, in fact, a size-α tests because the Locke interval is equal-tailed. But, Metzler (1991) and Schuirmann (1989) give graphs of the power function of the Mandallaz and Mau (1981) test that show that the test has size greater than the specified α. For example, Figures 3 through 9 in Metzler (1991) are graphs of 1 − (power function) based on the Mandallaz and Mau (1981) confidence interval. At $\delta_U = 1.2$, the rejection probability is about .07 for the $\alpha = .05$ test, and, the power is about .15 for the $\alpha = .10$ test. These figures cover a variety of sample sizes and variances. But in all cases the rejection probability exceeds the nominal α at $\delta_U = 1.2$. The same liberality of the Mandallaz and Mau test is illustrated by Figures 3–13 of Schuirmann (1989).

On the other hand, a test defined in terms of a $100(1-2\alpha)\%$ confidence set might be very conservative. An example is the test proposed by Chow and Shao (1990) for testing (1) about the ratio μ_T/μ_R. Specifically, Chow and Shao considered a two period crossover design with no carry-over, period or sequence effects. Let $\overline{\boldsymbol{X}}$ denote the sample mean vector with mean $\boldsymbol{\mu} = (\mu_T, \mu_R)'$ and let $\boldsymbol{S}$ denote the sum of cross-products matrix. Let m patients receive the first sequence, n patients receive the second sequence and $n^* = n + m$. Then, $C = \{\boldsymbol{\mu} : T_1 \le F_{\alpha,2,n^*-2}\}$ defines a $100(1-\alpha)\%$ confidence ellipse for $\boldsymbol{\mu}$, where $T_1 = n^*(n^*-2)(\overline{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu})/2$ and $F_{\alpha,2,n^*-2}$ is the upper 100α percentile of an F-distribution with 2 and n^*-2 degrees of freedom. Chow and Shao propose rejecting H_0 in (1) and concluding $\mathrm{H}_a : \delta_L < \mu_T/\mu_R < \delta_U$ is true if and only if the 90% confidence ellipse is contained in the cone defined by H_a. They do not comment on the actual size of this test, but we assume 90% was chosen to be $100(1-2\alpha)\%$ where $\alpha = .05$.

Chow and Shao's test can be described much more simply by recalling the relationship between the confidence ellipse, C, and simultaneous confidence intervals for all linear functions $\boldsymbol{l}'\boldsymbol{\mu}$ (Scheffé, 1959). $\boldsymbol{\mu} \in C$ if and only if $\boldsymbol{l}'\overline{\boldsymbol{X}} - \sqrt{2F_{\alpha,2,n^*-2}\boldsymbol{l}'\boldsymbol{S}\boldsymbol{l}/(n^*(n^*-2))} \leq \boldsymbol{l}'\boldsymbol{\mu} \leq \boldsymbol{l}'\overline{\boldsymbol{X}} + \sqrt{2F_{\alpha,2,n^*-2}\boldsymbol{l}'\boldsymbol{S}\boldsymbol{l}/(n^*(n^*-2))}$ for every vector $\boldsymbol{l}$. But, in fact, the only two vectors needed to define Chow and Shao's test are $\boldsymbol{l}_L = (1, -\delta_L)'$ and $\boldsymbol{l}_U = (1, -\delta_U)'$. The hypotheses in (1) or (14) can be written as $\mathrm{H}_0 : \boldsymbol{l}_L'\boldsymbol{\mu} \leq 0$ or $\boldsymbol{l}_U'\boldsymbol{\mu} \geq 0$ and $\mathrm{H}_a : \boldsymbol{l}_U'\boldsymbol{\mu} < 0 < \boldsymbol{l}_L'\boldsymbol{\mu}$. Furthermore, the ellipse C is below the line $\boldsymbol{l}_U'\boldsymbol{\mu} = 0$ if and only if $\boldsymbol{l}_U'\overline{\boldsymbol{X}} + \sqrt{2F_{\alpha,2,n^*-2}\boldsymbol{l}_U'\boldsymbol{S}\boldsymbol{l}_U/(n^*(n^*-2))} < 0$, that is, the upper endpoint of the confidence interval for $\boldsymbol{l}_U'\boldsymbol{\mu}$ is negative. Similarly, the ellipse C is above the line $\boldsymbol{l}_L'\boldsymbol{\mu} = 0$ if and only if $\boldsymbol{l}_L'\overline{\boldsymbol{X}} - \sqrt{2F_{\alpha,2,n^*-2}\boldsymbol{l}_L'\boldsymbol{S}\boldsymbol{l}_L/(n^*(n^*-2))} > 0$. If we define

$$T_L = \frac{\boldsymbol{l}_L'\overline{\boldsymbol{X}}}{\sqrt{\boldsymbol{l}_L'\boldsymbol{S}\boldsymbol{l}_L/(n^*(n^*-2))}} \quad \text{and} \quad T_U = \frac{\boldsymbol{l}_U'\overline{\boldsymbol{X}}}{\sqrt{\boldsymbol{l}_U'\boldsymbol{S}\boldsymbol{l}_U/(n^*(n^*-2))}},$$

then Chow and Shao's test rejects H_0 if and only if

$$T_L > \sqrt{2F_{\alpha,2,n^*-2}} \quad \text{and} \quad T_U < -\sqrt{2F_{\alpha,2,n^*-2}}. \tag{22}$$

This simple description of Chow and Shao's test has not appeared before. In this form, it is apparent that this test can be viewed as an IUT. A reasonable test of $\mathrm{H}_{0L} : \boldsymbol{l}_L'\boldsymbol{\mu} \leq 0$ versus $\mathrm{H}_{aL} : \boldsymbol{l}_L'\boldsymbol{\mu} > 0$ is the test that rejects H_{0L} if $T_L > \sqrt{2F_{\alpha,2,n^*-2}}$. A reasonable test of $\mathrm{H}_{0U} : \boldsymbol{l}_U'\boldsymbol{\mu} \geq 0$ versus $\mathrm{H}_{aU} : \boldsymbol{l}_U'\boldsymbol{\mu} < 0$ is the test that rejects H_{0U} if $T_U < -\sqrt{2F_{\alpha,2,n^*-2}}$. Thus, Chow and Shao's test is the IUT of H_0 versus H_a formed by combining these two tests. Theorems 1 and 2 then tell us that the actual size of this test is $\alpha' = P(T > \sqrt{2F_{\alpha,2,n^*-2}})$, where T has a Student's t distribution with n^*-1 degrees of freedom. This is because T_L has this t-distribution if $\boldsymbol{l}_L'\boldsymbol{\mu} = 0$, and T_U has this t-distribution if $\boldsymbol{l}_U'\boldsymbol{\mu} = 0$. That is, α' is the size of each of the two individual tests. We computed α' using a 90% confidence ellipse as suggested by Chow and Shao. We found that $\alpha' = .017$ for $m = n = 5$, 10, and 15, and $\alpha' = .016$ for $m = n = 20$, 30, and ∞. Thus, if the intent of using a $100(1-2\alpha)\% = 90\%$ confidence ellipse was to produce a bioequivalence test with type I error probability of $\alpha = .05$, the result was very conservative.

A test of H_0 versus H_a with the desired size of α can be obtained by replacing $\sqrt{2F_{\alpha,2,n^*-2}}$ with the t percentile, t_{α,n^*-1} in (22). Then each of the individual tests is size-α and the combined IUT also has size-α. This test is uniformly more powerful than Chow and Shao's test because the rejection region of Chow and Shao's test is a proper subset of this one. This test is the analogue of the TOST for this crossover model. In fact, Yang (1991) proposed this test for this problem as an alternative to Chow and Shao's test. But Yang did not state that this test was uniformly more powerful nor quantify the conservativeness of Chow and Shao's test.

Our conclusions from the results and examples in this subsection are simple. The usage of $100(1-2\alpha)\%$ confidence sets to define bioequivalence tests should be abandoned. This practice produces tests with the appropriate size only when special, "equal-tailed" confidence intervals are used, and offers no intuitive insight. The mixture of $100(1-2\alpha)\%$ confidence sets and size-α tests is only confusing. Rather, a test with the specified Type I error probability of α should be used. The IUT method can usually be used to construct such a test. Then, Theorem 3 might be used to construct the corresponding $100(1-\alpha)\%$ confidence set.

6 Multiparameter Equivalence Problems

Until now, we have discussed bioequivalence testing in terms of only one parameter. In this section, we discuss two problems in which the desired inference is equivalence in terms of two parameters. These results immediately generalize to situations in which bioequivalence is defined in terms of more than two parameters.

These two examples have been discussed as multiparameter bioequivalence problems by several authors. But, in some cases, the tests that have been proposed do not have the correct size α. The proposed tests do not properly account for the multiple testing aspect of this problem. These two multiparameter examples vividly illustrate that the IUT method can provide a simple mechanism for constructing tests with the correct size α, even in seemingly complicated bioequivalence problems. Size-α tests can be combined to obtain an overall size-α test. No adjustment for multiple testing is needed if the IUT method is used.

6.1 Simultaneous AUC and $\mathrm{C}_{\max}$ bioequivalence

Sections 4 and 5 discussed bioequivalence testing in terms of only one parameter. That is, the test and reference drugs are to be compared with respect to either average AUC or average $\mathrm{C}_{\max}$. FDA (1992) and EC-GCP (1993) consider two drugs are bioequivalent only if they are similar in both parameters. Westlake (1988) and Hauck et al. (1995) have considered the problem of comparing AUC and $\mathrm{C}_{\max}$ simultaneously. (Westlake actually compares three parameters, including $\mathrm{T}_{\max}$ also. But this does not conform to current FDA guidelines.)

Assume the measurements are lognormal so that, after log transformation, we wish to consider hypotheses like (2). Let the superscripts A and C refer to the variables AUC and $\mathrm{C}_{\max}$, respectively. For example, η_R^C denotes the mean of $\log(\mathrm{C}_{\max})$ for the reference drug. The test and reference drugs are to be considered bioequivalent only if

$$\mathrm{H}_a^m : \begin{array}{c} \theta_L < \eta_T^A - \eta_R^A < \theta_U \\ \text{and} \\ \theta_L < \eta_T^C - \eta_R^C < \theta_U \end{array} . \tag{23}$$

Using current FDA guidelines, $\theta_U = \log(1.25) = -\log(.80) = -\theta_L$. If one variable is deemed more important than another, the limits could be different for the different variables. For example, if AUC was considered more important than $\mathrm{C}_{\max}$, then the limits θ_L^A and θ_U^A for AUC could be chosen to be narrower than the limits θ_L^C and θ_U^C for $\mathrm{C}_{\max}$, as they are in Europe.

The statement H_a^m in (23) should be the alternative hypothesis in this multivariate bioequivalence test. The null hypothesis, H_0^m should be the negation of H_a^m. That is, H_0^m

states that one or more of the four inequalities in H_a^m is false. Westlake proposed testing H_0^m versus H_a^m by doing two separate tests, one for each variable. Specifically, he proposed using the TOST to test (2) for each variable. The drugs will be declared bioequivalent only if each of the tests rejects its hypothesis. Furthermore, Westlake said a Bonferroni correction should be used, and each TOST should be performed at the $\alpha/2$ level to account for the multiple testing. (Westlake actually said $\alpha/3$ since he was considering three tests.)

Westlake's procedure is conservative. The size of Westlake's test is $\alpha/2$, not α. This is true because, although he did not use this terminology, he has proposed an IUT. The alternative H_a^m is the intersection of two statements, one about each variable. Computing two separate TOSTs and concluding H_a^m is true only if both TOSTs reject, is an IUT. By Theorem 1, this test has level $\alpha/2$ if each TOST is performed at level $\alpha/2$. In fact, Theorem 2 can be used to show that this test has size equal to $\alpha/2$.

Therefore, to test H_0^m versus H_a^m, Westlake's procedure can be used except that each of the two TOSTs should be performed at size-α. The resulting test has probability at most α of declaring the drugs to be bioequivalent, if they are bioinequivalent.

Hauck et al. (1995) propose testing (23) using two size-α TOSTs. They recognize that the Bonferroni adjustment recommended by Westlake is unnecessary. But they come to the opposite conclusion. Based on a simulation study, they conclude that this test is too conservative and suggest that the two TOSTs might be performed using a higher error rate than α, and the resulting test of (23) would be size-α. (They admit that more simulations are needed to confirm this conjecture.) But, if the two TOSTs are each size-α, then the test of (23) is exactly size-α. To see this, use Theorem 2 by setting $\boldsymbol{\theta}_L = \eta_T^A - \eta_R^A$, $\eta_T^C = \eta_R^C$, and considering the limit as $\sigma_{D^A} \to 0$ and $\sigma_{D^C} \to 0$. Here, D^A and D^C are the estimates of $\eta_T^A - \eta_R^A$ and $\eta_T^C - \eta_R^C$, respectively. In this limit, three of the four one-sided tests will have rejection probability converging to 1, because these parameter points are in the alternative hypothesis and the corresponding standard deviations are converging to 0. The forth one-sided test will have rejection probability exactly equal to α, for all such parameter points, because $\boldsymbol{\theta}_L = \eta_T^A - \eta_R^A$ is on the boundary.

A test that is uniformly more powerful, but still has size-α will be obtained if the test we propose in Section 4.2 is used to perform the two tests, rather than using the two TOSTs. Again, both of these tests would be performed at size-α.

An alternative way of assessing the simultaneous bioequivalence of AUC and $\mathrm{C}_{\max}$ is to inspect the Brown, Casella, and Hwang (1995) confidence set (20), generalized to the Σ unknown case. Suppose $(X_i^A, X_i^C)', (Y_i^A, Y_i^C)', i = 1, \ldots, n$, are log-transformed *i.i.d.* observations on AUC and $\mathrm{C}_{\max}$ under the test and reference drugs, respectively. Let $\boldsymbol{Z}_i = (X_i^A, X_i^C)' - (Y_i^A, Y_i^C)', i = 1, \ldots, n$, which are assumed to be multivariate normal with mean $\boldsymbol{\theta} = (\eta_T^A - \eta_R^A, \eta_T^C - \eta_R^C)'$ and unknown variance-covariance matrix Σ. Let $\hat{\boldsymbol{\theta}} = (\overline{Z}^A, \overline{Z}^C)'$ and $\hat{\Sigma}$ be the sample mean vector and variance-covariance matrix of the $\boldsymbol{Z}_i$s. Then $\boldsymbol{\theta}'\hat{\boldsymbol{\theta}}$ is univariate normal with mean $\boldsymbol{\theta}'\boldsymbol{\theta}$ and variance $\boldsymbol{\theta}'\Sigma\boldsymbol{\theta}/n$, while $(n-1)\boldsymbol{\theta}'\hat{\Sigma}\boldsymbol{\theta}/\boldsymbol{\theta}'\Sigma\boldsymbol{\theta}$ is independent of $\boldsymbol{\theta}'\hat{\boldsymbol{\theta}}$ and has a χ^2 distribution with $n-1$ degrees of freedom. Thus, a size-α test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is obtained using the acceptance region

$$A(\boldsymbol{\theta}_0) = \left\{ (\hat{\boldsymbol{\theta}}, \hat{\Sigma}) : \boldsymbol{\theta}_0'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/\sqrt{\boldsymbol{\theta}_0'\hat{\Sigma}\boldsymbol{\theta}_0/n} > -t_{\alpha,n-1} \right\},$$

which leads to the confidence region

$$C(\hat{\boldsymbol{\theta}}, \hat{\Sigma}) = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta}'\hat{\boldsymbol{\theta}} / \sqrt{\boldsymbol{\theta}'\hat{\Sigma}\boldsymbol{\theta}/n} + t_{\alpha,n-1} > \boldsymbol{\theta}'\boldsymbol{\theta} / \sqrt{\boldsymbol{\theta}'\hat{\Sigma}\boldsymbol{\theta}/n} \right\}. \tag{24}$$

Brown, Casella, and Hwang (1995) applied (20) to the simultaneous AUC and $\mathrm{C_{max}}$ problem for illustration, assuming Σ is known. In practice, this assumption is perhaps unrealistic considering the moderate sample size typical in bioequivalence trials.

6.2 Mean and variance bioequivalence

Anderson and Hauck (1990) and Liu and Chow (1992) discuss another type of multiparameter bioequivalence. They point out that bioequivalence should not be defined only in terms of the mean responses for the two drugs. Rather, the variances of the two drugs' responses should also be considered. If two drugs have bioequivalent means but different variances, the drug with the smaller variance might be preferred. This kind of multiparameter bioequivalence is often called population bioequivalence.

Consider a single variable, e.g., AUC. Let η_T and η_R denote the means of log(AUC). Let σ_T^2 and σ_R^2 denote the intrasubject variances of the test and reference drugs, respectively. The test and reference drugs will be considered bioequivalent only if η_T and η_R are similar and σ_T^2 and σ_R^2 are similar. To demonstrate bioequivalence, we wish to test

$$\begin{array}{ll} \mathrm{H}_0^m : & \begin{array}{c} \eta_T - \eta_R \le \theta_L \ \text{ or } \ \eta_T - \eta_R \ge \theta_U \\ \text{or} \\ \sigma_T^2/\sigma_R^2 \le \kappa_L \ \text{ or } \ \sigma_T^2/\sigma_R^2 \ge \kappa_U \end{array} \\ \text{versus} & \\ \mathrm{H}_a^m : & \begin{array}{c} \theta_L < \eta_T - \eta_R < \theta_U \\ \text{and} \\ \kappa_L < \sigma_T^2/\sigma_R^2 < \kappa_U \end{array} . \end{array} \tag{25}$$

The constants θ_L, θ_U, κ_L, and κ_U would be chosen to define clinically important differences.

Liu and Chow (1992) propose a size-α test of

$$\begin{array}{ll} & \mathrm{H}_0^\sigma : \sigma_T^2/\sigma_R^2 \le \kappa_L \ \text{ or } \ \sigma_T^2/\sigma_R^2 \ge \kappa_U \\ \text{versus} & \\ & \mathrm{H}_a^\sigma : \kappa_L < \sigma_T^2/\sigma_R^2 < \kappa_U \end{array} .$$

Their test is an IUT composed of two size-α tests, one for testing each inequality. Wang (1994) describe an unbiased, size-α test that is uniformly more powerful than the Liu and Chow test.

The hypotheses

$$\begin{array}{ll} & \mathrm{H}_0^\eta : \eta_T - \eta_R \le \theta_L \ \text{ or } \ \eta_T - \eta_R \ge \theta_U \\ \text{versus} & \\ & \mathrm{H}_a^\eta : \theta_L < \eta_T - \eta_R < \theta_U \end{array}$$

can be tested with a TOST. Because H_a^m is the intersection of H_a^η and H_a^σ, the IUT method can be used to construct a test of H_0^m versus H_a^m. The test that rejects H_0^m only if the size-α

Liu and Chow test rejects H_0^σ and the size-α TOST rejects H_0^η is a size-α test of H_0^m versus H_a^m.

Liu and Chow, however, propose a more conservative combination of these two tests. Let α denote the desired size of the test of H_0^m. Let α_1 denote the size of the TOST and let α_2 denote the size of the Liu and Chow test. They say to choose α_1 and α_2 so that

$$\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2). \tag{26}$$

Liu and Chow note that the test statistics use for the TOST are independent of the test statistics used in their test. But they give no further explanation of (26). The probability that H_0^η is accepted, given that H_0^η is true, is bounded below by $1 - \alpha_1$. The probability that H_0^σ is accepted, given that H_0^σ is true, is bounded below by $1 - \alpha_2$. So the quantity α in (26) is an upper bound for the probability that at least one of the two tests rejects its null hypothesis, given that both H_0^η and H_0^σ are true. This is not the error probability of the proposed test. The error probability is the probability the both tests reject, given that either H_0^η or H_0^σ is true. H_0^m is the union of H_0^η and H_0^σ, not the intersection.

Again, it should be noted that a more powerful size-α test of H_0^m will be obtained if the test from Section 4.2, rather than the TOST, is used to test H_0^η and Wang's (1994) test is used to test H_0^σ.

7 Concluding Remarks

We have shown that the theory of intersection-union tests is central to bioequivalence studies. We have demonstrated the danger of incorrect association of confidence sets with such tests. Due to the traditional emphasis on *significant difference* inference in statistics, many *practical equivalence* problems have not been recognized as such, we believe. It is our hope (and anticipation) that the concepts and techniques discussed in this article will, in time, prove to be useful not only in bioequivalence studies, but in other practical equivalence problems as well.

8 Acknowledgment

We thank Dr. Hans Frick and Dr. Volker Rahlfs for references on European bioequivalence guidelines.

A Details of New Test in Section 4.2

A size-α, nearly unbiased test for (2) was described geometrically in Section 4.2. In Section A.1, formulas and computational suggestions are given for the quantities that define that test. The construction in Section 4.2 is valid for $\alpha > \alpha_*$. In Section A.2 a similar construction yields a size-α, nearly unbiased test for $\alpha \leq \alpha_*$. Brown, Hwang and Munk did not propose any test for $\alpha \leq \alpha_*$.

A.1 Formulas for Section 4.2

Define functional notation for the transformation from rectangular to polar coordinates by

$$\begin{aligned} v(d, s_*) &= \sqrt{(d-\Delta)^2 + s_*^2} \\ b(d, s_*) &= \cos^{-1}((d-\Delta)/v(d, s_*)) \end{aligned}$$

for $-\infty < d < \infty$ and $s_* \geq 0$. The inverse transformation is

$$\begin{aligned} d(v, b) &= \Delta + v\cos(b) \\ s_*(v, b) &= v\sin(b), \end{aligned}$$

for $v \geq 0$ and $0 \leq b \leq \pi$. The point $(d, s_*) = (0, \Delta\sqrt{r}/t_{\alpha,r})$ is the vertex of the triangular region R_T. Therefore,

$$\begin{aligned} b_0 &= b(0, \Delta\sqrt{r}/t_{\alpha,r}), \\ v_0 &= 2\Delta\sin(\pi - b_0), \\ (d_1, s_{*1}) &= (d(v_0, b_0), s_*(v_0, b_0)), \\ v_1 &= v(-d_1, s_{*1}). \end{aligned}$$

The line of length v_0 in Figure 2 has $b = 3\pi/2 - b_0$. Therefore,

$$\begin{aligned} b_1 &= 3\pi/2 - b_0 - \cos^{-1}(v_0/v), \\ b_2 &= 3\pi/2 - b_0 + \cos^{-1}(v_0/v). \end{aligned}$$

The angle b_{L1}, defined by (12), is easily found by a numeric root finding method such as bisection.

Finally, for any point (d, s_*) on $\{V = v\}$, $s_* = \sqrt{v^2 - (d-\Delta)^2}$. For any point (d_u, s_{*u}) on $\{V = v\}$ with $d_u \leq 0$, there is a unique point (d_l, s_{*l}) on $\{V = v\}$ with $d_l \geq 0$ such that the line joining (d_l, s_{*l}) and $(0,0)$ and the s_* axis form the same angle as the line joining (d_u, s_{*u}) and $(0,0)$ and the s_* axis. This point satisfies

$$\frac{d_u}{\sqrt{v^2 - (d_u - \Delta)^2}} = -\frac{d_l}{\sqrt{v^2 - (d_l - \Delta)^2}}$$

which has the solution

$$d_l = \frac{d_u(v^2 - \Delta^2)}{v^2 + 2d_u\Delta - \Delta^2}. \tag{27}$$

Using this expression for d_l in terms of d_u, the equation

$$F(b(d_u, s_u)) - F(b(d_l, s_l)) = \alpha - \alpha(v)$$

is a function of the single variable d_u. The unique solution to this equation, in the interval $\Delta - v \leq d_u \leq 0$ is easily found by a numeric root finding method such as bisection. Call the solution d_U. Define d_L by (27) using $d_u = d_U$. The angles $b_U(v)$ and $b_L(v)$ are

$$\begin{aligned} b_U(v) &= b\left(d_U, \sqrt{v^2 - (d_U - \Delta)^2}\right), \\ b_L(v) &= b\left(d_L, \sqrt{v^2 - (d_L - \Delta)^2}\right). \end{aligned}$$

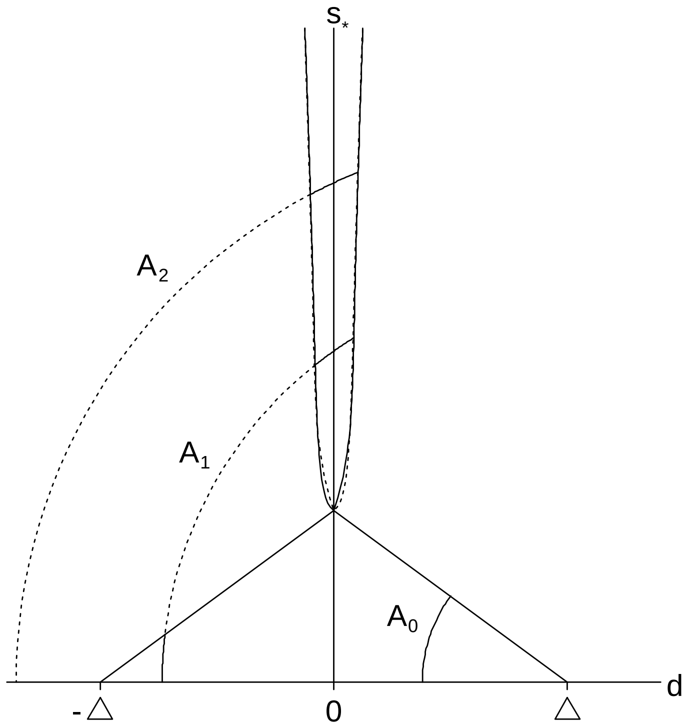

Figure 5: Rejection region of new test for $\alpha \le \alpha_*$. Region R_2 (between solid lines) and region R_1 (between dashed lines). Rejection region $R = R_1 \cap R_2$. $r = 3$ and $\alpha = .05$.

A.2 New test for $\alpha \leq \alpha_*$

For small values of $\alpha \leq \alpha_*$, a size-α, nearly unbiased test of (2), that is uniformly more powerful than the TOST, can be constructed. The construction is very similar, and somewhat simpler, than the construction in Section 4.2. The notation of Section A.1 will be used, and Figure 5 illustrates the construction.

For $\alpha \leq \alpha_*$, the point on l_L closest to $(\Delta, 0)$ is the vertex of R_T, $(d_0, s_{*0}) = (0, \Delta\sqrt{r}/t_{\alpha,r})$. Let $v_0 = v(d_0, s_{*0})$. For $v \leq v_0$, $R_2(v) = \{b : b_0 < b < \pi\}$, exactly the points in the TOST. The arc A_0 is such an arc. For $v_0 < v < 2\Delta$, $R_2(v)$ consists of two arcs. $R_2(v) = \{b : b_L(v) < b < b_U(v)\} \cup \{b : b_2 < b < \pi\}$. $b_L(v)$, $b_U(v)$ and b_2 are defined as before. The two solid pieces of arc A_1 are examples of these arcs. The semicircle $\{V = v\}$ does not intersect R_T near the s_* axis so there is no need to check that $\{b : b_L(v) < b < b_U(v)\}$ covers all the TOST. For $v \geq 2\Delta$, $R_2(v) = \{b : b_L(v) < b < b_U(v)\}$. The solid piece of arc A_3 is such an arc. In Figure 5, R_2 is outlined with a solid line, R_1 is outlined with a dashed line, and the intersection is the rejection region of the IUT.

References

Anderson, S. (1993). Individual bioequivalence: a problem of switchability (with discussion). *Biopharmaceutical Report*, 2(2):1–11.

Anderson, S. and Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics - Theory and Methods*, 12:2663–2692.

Anderson, S. and Hauck, W. W. (1985). Letter to the editor. *Biometrics*, 41:561–563.

Anderson, S. and Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, 18:259–273.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300.

Berger, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *Journal of the American Statistical Association*, 84:192–199.

Bofinger, E. (1985). Expanded confidence intervals. *Communications in Statistics - Theory and Methods*, A14:1849–1864.

Bofinger, E. (1992). Expanded confidence intervals, one-sided tests, and equivalence testing. *Journal of Biopharmaceutical Statistics*, 2:181–188.

Brown, L. D., Casella, G., and Hwang, J. T. G. (1995a). Optimal confidence sets, bioequivalence, and the limaçon of Pascal. *Journal of the American Statistical Association*, 90:880–889.

Brown, L. D., Hwang, J. T. G., and Munk, A. (1995b). An unbiased test for the bioequivalence problem. Technical report, Cornell University.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Duxbury, Belmont, CA.

Chow, S.-C. and Liu, J.-P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies.* Marcel Dekker.

Chow, S.-C. and Shao, J. (1990). An alternative approach for the assessment of bioequivalence between two formulations of a drug. *Biometrical Journal*, 32:969–976.

Diletti, E., Hauschke, D., and Steinijans, V. W. (1991). Sample size determination for bioequivalence assessment by means of confidence intervals. *International Journal of Clinical Pharmacology, Therapy and Toxicology*, 29:1–8.

EC-GCP (1993). *Biostatistical methodology in clinical trials in applications for marketing authorization for medical products.* CPMP Working Party on Efficacy of Medical Products, Commission of the European Communities, Brussels, Draft Guideline edition.

FDA (1987). *Guideline for Submitting Documentation for Stability Studies of Human Drugs and Biologics.* Center for Drugs and Biologics, Food and Drug Administration, Rockville, MD.

FDA (1992). Bioavailability and bioequivalence requirements. In *U. S. Code of Federal Regulations*, volume 21, chapter 320. U. S. Government Printing Office.

Fieller, E. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, 16:175–185.

Fieller, E. C. (1940). The biological standardisation of insulin. *Journal of the Royal Statistical Society, Supplement*, 7:1–64.

Hauck, W. W. and Anderson, S. (1992). Types of bioequivalence and related statistical considerations. *International Journal of Clinical Pharmacology, Therapy and Toxicology*, 30:181–187.

Hauck, W. W., Hyslop, T., Anderson, S., Bois, F. Y., and Tozer, T. N. (1995). Statistical and regulatory considerations for multiple measures in bioequivalence testing. *Clinical Research and Regulatory Affairs*, 12:249–265.

Hauschke, D., Steinijans, V. W., and Diletti, E. (1990). A distribution-free procedure for the statistical analysis of bioequivalence studies. *International Journal of Clinical Pharmacology, Therapy and Toxicology*, 28:72–78.

Hayter, A. J. and Hsu, J. C. (1994). On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association*, 89:128–136.

Hsu, J. C. (1984). Constrained two-sided simultaneous confidence intervals for multiple comparisons with the best. *The Annals of Statistics*, 12:1136–1144.

Hsu, J. C. (1996). *Multiple Comparisons.* Chapman and Hall, London.

Hsu, J. C., Hwang, J. T. G., Liu, H.-K., and Ruberg, S. J. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika*, 81:103–114.

Juskevich, J. C. and Guyer, C. G. (1990). Bovine growth hormone: Human food safety evaluation. *Science*, 249:875–884.

Kinsella, A. (1989). Bootstrapping a bioequivalence measure. *The Statistician*, 38:175–179.

Lehmann, E. L. (1959). *Testing Statistical Hypothesis.* John Wiley, New York.

Lehmann, E. L. (1986). *Testing Statistical Hypothesis.* John Wiley, New York, second edition.

Liu, H. and Berger, R. L. (1995). Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. *Annals of Statistics*, 23:55–72.

Liu, J.-P. and Chow, S.-C. (1992). On the assessment of variability in bioavailability/bioequivalence studies. *Communications in Statistics - Theory and Methods*, 21:2591–2607.

Liu, J.-P. and Weng, C.-S. (1995). Bias of two one-sided tests procedures in assessment of bioequivalence. *Statistics in Medicine*, 14:853–861.

Locke, C. S. (1984). An exact confidence interval from untransformed data for the ratio of two formulation means. *Journal of Pharmacokinetics and Biopharmaceutics*, 12:649–655.

Mandallaz, D. and Mau, J. (1981). Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics*, 20:213–222.

Martin Andrés, A. (1990). On testing for bioequivalence. *Biometrical Journal*, 32:125–126.

Metzler, C. M. (1991). Sample sizes for bioequivalence studies. *Statistics in Medicine*, 10:961–970.

Müller-Cohrs, J. (1991). An improvement of the Westlake symmetric confidence interval. *Biometrical Journal*, 33(3):357–360.

Munk, A. (1993). An improvement on commonly used tests in bioequivalence assessment. *Biometrics*, 49:1225–1230.

Patel, H. I. and Gupta, G. D. (1984). A problem of equivalence in clinical trials. *Biometrical Journal*, 26:471–474.

Pratt, J. W. (1961). Length of confidence intervals. *Journal of the American Statistical Association*, 56:541–567.

Rocke, D. M. (1984). On testing for bioequivalence. *Biometrics*, 40:225–230.

Ruberg, S. J. and Hsu, J. C. (1992). Multiple comparison procedures for pooling batches in stability studies. *Technometrics*, 34:465–472.

Sasabuchi, S. (1980). A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. *Biometrika*, 67:429–439.

Sasabuchi, S. (1988a). A multivariate one-sided test with composite hypotheses when the covariance matrix is completely unknown. *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics*, 42:37–46.

Sasabuchi, S. (1988b). A multivariate test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor. *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics*, 42:9–19.

Schall, R. and Luus, H. G. (1993). On population and individual bioequivalence. *Statistics in Medicine*, 12:1109–1124.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.

Schuirmann, D. J. (1989). Confidence intervals for the ratio of two means from a crossover study. In *American Statistical Association Proceedings of the Biopharmaceutical Section*, pages 121–126. American Statistical Association, Washington, D.C.

Schuirmann, D. L. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, 37:617. Abstract.

Sheiner, L. B. (1992). Bioequivalence revisited. *Statistics in Medicine*, 11:1777–1788.

Stefansson, G., Kim, W., and Hsu, J. C. (1988). On confidence sets in multiple comparisons. In Gupta, S. S. and Berger, J. O., editors, *Statistical Decision Theory and Related Topics IV*, volume 2, pages 89–104. Springer-Verlag, New York.

Wang, W. (1994). Optimal unbiased tests for bioequivalence in variability. Technical report, Cornell University.

Westlake, W. J. (1973). The design and analysis of comparative blood-level trials. In Swarbrick, J., editor, *Current Concepts in the Pharmaceutical Sciences, Dosage Form Design and Bioavailability*, pages 149–179. Lea and Febiger, Philadelphia.

Westlake, W. J. (1976). Symmetric confidence intervals for bioequivalence trials. *Biometrics*, 32:741–744.

Westlake, W. J. (1981). Response to T.B.L. Kirkwood: Bioequivalence testing - a need to rethink. *Biometrics*, 37:589–594.

Westlake, W. J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations. In Peace, K. E., editor, *Biopharmaceutical Statistics for Drug Development*, pages 329–352. Marcel Dekker, New York.

Yang, H.-M. (1991). An extended two one-sided tests procedure. In *American Statistical Association Proceedings of the Biopharmaceutical Section*, pages 157–162. American Statistical Association, Washington, D.C.

Yee, K. F. (1986). The calculation of probabilities in rejecting bioequivalence. *Biometrics*, 42:961–965.

Yee, K. F. (1990). Correspondence to the editor. *The Statistician*, 39:465–466.