Choosing Confidence Intervals in Bioequivalence Studies: $100(1 - 2\alpha)\%$ and $100(1 - \alpha)\%$ Approaches

Kexuan Li [a*], Susie Sinks [b], Peng Sun [b] and Lingli Yang [c]

[a]*Global Biometrics and Data Sciences Bristol Myers Squibb, Cambridge, Massachusetts, US.*
[b]*Global Analytics and Data Sciences, Biogen, Cambridge, Massachusetts, US.*
[c]*Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts, US.*

Original Research Article

Abstract

A bioequivalence study is a type of clinical trial designed to compare the biological equivalence of two different formulations of a drug. Such studies are typically conducted in controlled clinical settings with human subjects, who are randomly assigned to receive two formulations. The two formulations are then compared with respect to their pharmacokinetic profiles, which encompass the absorption, distribution, metabolism, and elimination of the drug. Under the guidance from Food and Drug Administration (FDA), for a size-α bioequivalence test, the standard approach is to construct a $100(1 - 2\alpha)\%$ confidence interval and verify if the confidence interval falls with the critical region. In this work, we clarify that $100(1 - 2\alpha)\%$ confidence interval approach for

Corresponding author: E-mail: kexuan.li.77@gmail.com;

bioequivalence testing yields a size-α test only when the two one-sided tests in TOST are "equal-tailed". Furthermore, a $100(1-\alpha)\%$ confidence interval approach is also discussed in the bioequivalence study.

Keywords: Bioequivalence study; two one-sided tests; confidence interval.

2010 Mathematics Subject Classification: 53C25, 83C05, 57N16.

1 Introduction

Bioequivalence studies are a type of clinical trial designed to compare the biological equivalence of two different formulations of a drug. The objective of these studies is to establish that the generic drug demonstrates the same rate and extent of absorption as the reference drug and that its therapeutic effects are comparable. Such studies are typically conducted in controlled clinical settings with human subjects, who are randomly assigned to receive either the generic drug or the reference drug. The two formulations are then compared with respect to their pharmacokinetic profiles, which encompass the absorption, distribution, metabolism, and elimination of the drug. Acceptance criteria for bioequivalence are generally based on statistical methods that assess the similarities between the pharmacokinetic profiles of the two drug products. These comparisons often involve evaluating the area under the concentration-time curve (AUC) and the maximum concentration (Cmax) of the drug in the bloodstream. By ensuring that both AUC or Cmax values fall within pre-defined equivalence ranges, researchers can determine if the two formulations are bioequivalent. The examination of drug product bioequivalence and the statistical foundation governing their design, analysis, and interpretation has undergone significant evolution in the past two decades. A preference for a crossover design has emerged over a parallel-group design, given its ability to segregate inter-subject variation (non-product-dependent) from intra-subject variation (product-dependent). In recent years, bioequivalence studies have garnered significant attention in the literature, reflecting an increasing focus on refining methodologies and addressing nuanced challenges. Notably, [1] delved into the multiple testing problem within the realm of equivalence testing for food safety. Their work critically compared various methods aimed at controlling the family-wise error rate, contributing valuable insights to this crucial area of study. [2] conducted a comprehensive review of simultaneous confidence regions, particularly in the context of multi-parameter bioequivalence studies, providing a nuanced exploration of diverse approaches. [3] made a noteworthy contribution by proposing a ratio of means, grounded in the original bioavailability measure, as a novel definition for average bioequivalence. This innovative perspective adds depth to the ongoing discourse surrounding bioequivalence assessment. [4] introduced a method for constructing simultaneous confidence intervals for a parameter vector, employing a series of randomization tests (RT). Notably, their approach utilizes an efficient multivariate Robbins–Monro procedure, which incorporates correlation information from all components, enhancing the robustness of the simultaneous confidence intervals. Furthermore, [5] addressed the vital aspect of finite sample corrections in the context of average equivalence testing. Their work provides valuable considerations for improving the precision and reliability of bioequivalence assessments in practical, finite-sample scenarios. These diverse studies collectively contribute to the evolving landscape of bioequivalence research, offering innovative perspectives, methodological advancements, and solutions to contemporary challenges.

Suppose we want to access the bioequivalence of two drugs or formulations. Under the guidance of Food and Drug Administration (FDA), the bioequivalence is claimed if a 90% two-sided confidence interval of the geometric mean ratio falls within 80-125%. Under the context of statistical hypothesis testing, to demonstrate bioequivalence, the following significance level α hypothesis testing problem is considered:

$$H_0 : \frac{\eta_T}{\eta_R} \leq \delta_L \text{ or } \frac{\eta_T}{\eta_R} \geq \delta_U \text{ versus } H_a : \delta_L < \frac{\eta_T}{\eta_R} < \delta_U, \tag{1.1}$$

where η_T, η_R are the population geometric mean of the test product and reference product respectively, δ_L, δ_U are the lower and upper error bounds determined by the regulatory. Under the theory of intersection-union test, [6, 7, 8] proposed two one-sided tests (TOST), which now has been the standard approach to test (1.1). As

its name implies, TOST decomposes the interval hypothesis into two size-α one-side hypotheses and rejects H_0 if and only if both two one-sided hypotheses are rejected. In general, when multiple size-α tests are combined together, the overall size of the combined test is no longer α. Fortunately, due to the theory of interaction-union test, the size of TOST is still α, which will be discussed in Section 2.1. It has also been showed that, for a size-α TOST procedure, it is identical to construct a $100(1-2\alpha)\%$ confidence interval for $\frac{\eta_T}{\eta_R}$ and reject H_0 if the confidence interval falls entirely between δ_L and δ_U. Even though the procedure seems simple, there are still some questions that need to be clarified, for example:

- Why is the geometric mean preferred over the arithmetic mean?
- Why are the BE limits (0.8, 1.25) not symmetric about 1?
- Why is the product of BE limit equal to 1 ($0.8 \times 1.25 = 1$), i.e., symmetric about 1 on the ratio scale?
- In general, the combination of two size-α tests could not get a size-α test, but why TOST works? Why is a size-α TOST identical to a $100(1-2\alpha)\%$ confidence interval, not a $100(1-\alpha)\%$ confidence interval? Is TOST the most powerful?

The purpose of this paper is to answer these questions theoretically and provide a clear and in-depth understanding of the bioequivalence study. For example, the geometric mean is preferred in bioequivalence studies because pharmacokinetic parameters, such as AUC and Cmax, typically follow a log-normal distribution. Using the geometric mean ensures that the ratio of the means remains unaffected by extreme values or outliers, leading to a more accurate and robust comparison of the test and reference formulations. The geometric mean is calculated by taking the exponential of the average of the natural logarithms of individual observations. This ensures that extreme values or outliers do not unduly influence the mean, providing a robust measure that accurately represents the central tendency of log-normally distributed data. However, the geometric mean is not suitable for data with negative values and can be influenced by small sample sizes. The bioequivalence limits (0.8, 1.25) are not symmetric around 1 because they account for potential differences in variability and the possibility of type I or type II errors. The limits are based on log-transformed data, and when back-transformed to the original scale, they yield asymmetric confidence intervals. The asymmetry helps to accommodate the potential differences in variability between the two drug formulations. The product of the BE limits is equal to 1 to ensure that the ratio of the test and reference products' geometric means is unbiased on the ratio scale. When the two limits are multiplied, they effectively cancel out any deviation from the true ratio of 1, allowing for a more accurate comparison of the drug formulations. The specific bioequivalence limits (0.8, 1.25) are strategically chosen to accommodate variability and control error rates. The product of these limits equals 1 to ensure an unbiased ratio on the scale. This condition significantly impacts the overall design and interpretation of bioequivalence studies. TOST works because of the intersection-union test theory. While it may not be the most powerful test in all situations, it maintains the overall size-α test by decomposing the interval hypothesis into two size-α one-sided hypotheses and rejecting the null hypothesis only if both one-sided hypotheses are rejected. The intersection-union test theory ensures that the overall type I error rate remains controlled at the desired level (α) when combining the two one-sided tests. In general, combining two size-α tests does not guarantee a size-α test because the overall type I error rate may be inflated due to multiple testing. However, the TOST procedure, based on the intersection-union test theory, accounts for this issue and successfully maintains the overall type I error rate at the desired level, making it a suitable method for bioequivalence testing.

The rest of the paper is organized as follows. Section 2 provides the theoretical background of bioequivalence study for a univariate PK parameter. We summarize the paper in Section 3 and technical proofs are provided in Appendix.

2 Theoretical Background

In this section, we delve into the theoretical background of bioequivalence studies for a univariate pharmacokinetic (PK) parameter. In a typical pharmacokinetic bioequivalence study, the univariate response variables such as $\log(\text{AUC}), \log(\text{C}_{\max})$ are often assumed to follow a normal distribution, or equivalently, $\text{AUC}, \text{C}_{\max}$ follow a lognormal distribution ([9], [10]). Under FDA guidelines, bioequivalence is treated as a statistical hypothesis

testing problem as defined in (1.1), which states that the population geometric mean ratio for the test and reference products should fall between 80-125%. At first glance, it may seem odd that the geometric mean ratio is used instead of the arithmetic mean difference. Moreover, it might appear strange that 80% and 125% are not symmetric about 100%, but $80\% \times 125\% = 100\%$, meaning that the lower and upper bounds are symmetric about 100% on a ratio scale. To explain the reason, we need to take the logarithm of (1.1) and obtain the following hypothesis testing for the difference:

$$H_0 : \mu_T - \mu_R \leq \theta_L \text{ or } \mu_T - \mu_R \geq \theta_U \text{ versus } H_a : \theta_L < \mu_T - \mu_R < \theta_U, \tag{2.1}$$

where $\mu_T = \log(\eta_T), \mu_R = \log(\eta_R), \theta_L = \log(\delta_L), \theta_R = \log(\delta_R)$. The following theorem explains some of the reasons why FDA chooses geometric mean ratio as the test statistic and 80-125% as the BE limits.

Theorem 2.1. *Suppose $X_1, \ldots, X_n$ are independent and identically distributed from normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ and $X_i^* = \exp(\mu + \sigma X_i)$, that is, $X_1^*, \ldots, X_n^*$ are independent and identically distributed from lognormal distribution X^* with corresponding mean μ, variance σ^2 on the log transformed scale. We let $GM(X^*) = (\prod_{i=1}^n X_i^*)^{\frac{1}{n}}$ be the geometric mean of $X_1^*, \ldots, X_n^*$. Then the following statements hold true*

1. $\exp(\bar{X}) = GM(X^*)$;

2. $Median(X^*) = \exp(\mu)$;

3. $[GM(X^*)] = \exp(\mu + \frac{\sigma^2}{2n})$.

The proof is given in the supplementary material. According to Theorem 2.1, it is straightforward to obtain the following facts:

Remark 2.1.

1. Since the PK parameters are lognormal distributed, which is right-tailed, it is more natural to compare the median between treatment and reference products rather than the arithmetic mean;

2. According to (1) in Theorem 2.1, we know that the log-transformed geometric mean of PK data is equivalent to the arithmetic mean of the log-transformed PK data. In other words, comparing the geometric mean ratio between two products is equivalent to comparing the arithmetic mean difference between the log-transformed PK data;

3. According to (2) and (3) in Theorem 2.1, we can make a conclusion that the geometric mean of PK data is a nearly unbiased estimator of PK median with an error rate of $O(1/n)$. To be more specific, as n approaches infinity, $[GM(X^*)] = Median(X^*)$. This implies that, for large sample sizes, the geometric mean of PK data provides an accurate estimate of the PK median.

Combining everything above, we can identify the advantages of using the geometric mean ratio as the test statistic: (1) The comparison of the geometric mean ratio can be converted to the comparison of the arithmetic mean difference; (2) The geometric mean ratio is a "good" estimator of the median ratio, which can be naturally applied to the distribution of PK parameters. Following the above analysis, we can further clarify the procedure of bioequivalence testing for univariate PK parameters in practice:

- Step 1: Log-transform the PK data (e.g., AUC and Cmax) to convert the lognormal distribution into a normal distribution.

- Step 2: Test $H_0 : \mu_T - \mu_R \leq \theta_L$ or $\mu_T - \mu_R \geq \theta_U$ versus $H_a : \theta_L < \mu_T - \mu_R < \theta_U$, where μ_T, μ_R are the log-transformed PK data from Step 1 and $\theta_L = \log(0.8) = -0.223, \theta_U = \log(1.25) = 0.223$.

- Step 3: Find a $100(1-2\alpha)\%$ confidence interval for $\mu_T - \mu_R$ and reject H_0 if and only if the $100(1-2\alpha)\%$ confidence interval falls in $[\theta_L, \theta_R]$ and the significance level of the corresponding procedure is α.

It is important to note that after taking the logarithm, the original BE limits (0.8, 1.25) become symmetric about zero, and this fact plays a crucial role in the $100(1 - 2\alpha)\%$ confidence interval approach. Without the "equal-tail" property, the $100(1 - 2\alpha)\%$ confidence interval approach is no longer valid. This is because the symmetric nature of the limits on the log scale simplifies the comparison of the test and reference products,

allowing for a balanced evaluation of bioequivalence that accounts for both type I and type II errors. For a more in-depth exploration of this concept and its practical implications, refer to Section 2.3.2. In that section, we delve into the intricate details of the $100(1 - 2\alpha)\%$ confidence interval approach, elucidating how the symmetric nature of the log-transformed limits enhances the precision and reliability of bioequivalence assessments.

2.1 Union-interaction tests and interaction-union tests

In the section, we proceed to review the TOST, or two one-sided tests. The TOST involves performing two one-sided tests, one to test if the difference between the treatments or populations is greater than a predetermined non-inferiority margin and another to test if the difference is less than a predetermined equivalence margin. If the results of both tests are significant, it suggests that the treatments or populations are equivalent within the defined margins. Before proceeding, we first review the union-interaction test and intersection-union test, which provide the theoretical foundation for the TOST. Suppose the parameter space of a hypothesis testing is $\Theta = \Theta_0 \bigcup \Theta_0^c$, where Θ_0, Θ_0^c are the parameter space of null and alternative hypothesis, respectively. Furthermore, let R be the rejection region of the test, then the power function of the test, as a function of $\theta \in \Theta$, is defined as $\beta(\theta) = \mathbb{P}_{\theta \in \Theta}(T \in R)$, where T is the test statistics. Before moving on, we first clarify two definitions that are commonly confused with one another.: the *size* and the *level* of a test.

Definition 2.1. For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size-α test if $\sup_{\Theta_0} \beta(\theta) = \alpha$.

Definition 2.2. For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a level-α test if $\sup_{\Theta_0} \beta(\theta) \leq \alpha$.

These definitions help distinguish between the size and the level of a test, which are related but distinct concepts. The size of a test is the maximum probability of making a Type I error (i.e., rejecting the null hypothesis when it is true) under the null hypothesis parameter space Θ_0. In contrast, the level of a test refers to the upper bound on the probability of making a Type I error. When the size of a test equals α, it is considered a size-α test, whereas when the size of a test is less than or equal to α, it is considered a level-α test. For a level-α test, the size of the test may be much less then α, and in such case, if we still use α as the significance level, then the test will be less powerful.

Obviously, the null hypothesis in (2.1) can be viewed as the union of two simple null hypotheses $H_{01} : \mu_T - \mu_R \leq \theta_L$ and $H_{02} : \mu_T - \mu_R \geq \theta_U$. Not only in bioequivalence studies but also in some other applications, such complicated hypotheses can be developed from tests for simpler hypotheses. These hypotheses can be generalized from the so-called *intersection-union tests* (IUT) and *union-intersection test* (UIT), which play an important role in bioequivalence testing and multiple comparison procedures.

Definition 2.3. Assume a hypothesis $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta_0^c$ with reject R. A family of hypotheses $H_{i0} : \theta \in \Theta_i$ versus $H_{ia} : \theta \in \Theta_i^c$ for $i = 1, \ldots, k$ with corresponding rejection region R_i is said to obey the intersection-union principle if

$$H_0 : \theta \in \bigcup_{i=1}^{k} \Theta_i, \text{ and } H_a : \theta \in \bigcap_{i=1}^{k} \Theta_i^c, \tag{2.2}$$

and is said to obey the union-intersection principle if

$$H_0 : \theta \in \bigcap_{i=1}^{k} \Theta_i, \text{ and } H_a : \theta \in \bigcup_{i=1}^{k} \Theta_i^c. \tag{2.3}$$

It is straightforward to show that the rejection region of IUT is $\bigcap_{i=1}^{k} R_i$. The logic behind this is simple. H_0 is false if and only all of the $H_{0i}, i = 1, \ldots, k$ are false. So rejecting $H_0 : \theta \in \bigcup_{i=1}^{k} \Theta_i$ is equivalent as rejecting each individual $H_{0i} : \theta \in \Theta_i$. Similarly, in terms of UIT, using the same logic, if any one of the hypotheses H_{0i} is rejected, then H_0 must also be rejected, and only if each of H_{0i} is accepted will the intersection H_0 be accepted. Thus, the rejection region of UIT is $\bigcup_{i=1}^{k} R_i$. Clearly, the hypothesis testing in 2.1 belongs to IUT. Now, a natural question arises: given the significance of each individual H_{0i}, what is the significance level of IUT? The following theorem gives an upper bound for the size of the IUT.

Theorem 2.2. *Let α_i be the size of the test of H_{0i} with rejection region R_i. Then the IUT with rejection region $R = \bigcap_{i=1}^{k} R_i$ is a level $\alpha = \sup_{i=1,\ldots,k} \alpha_i$ test.*

It should be remarked that Theorem 2.2 only shows that the level of an IUT is $\alpha = \sup_{i=1,\ldots,k} \alpha_i$, that is, an upper bound for the size of the IUT. In fact, the size of the IUT could be much less than α. The following theorem provides the conditions under which the size of the IUT is exactly α, ensuring that the IUT is not too conservative.

Theorem 2.3. *Consider testing $H_0 : \theta \in \bigcup_{j=1}^{k} \Theta_j$ and let R_j be the rejection region such that the level of H_{0j} is α. Suppose that for some $i = 1, \ldots, k$, there exists a sequence of parameters, $\theta_l \in \Theta_i, l = 1, 2, \ldots$, such that*

1. *$\lim_{l \to \infty} \mathbb{P}_{\theta_l}(X \in R_i) = \alpha$,*

2. *for each $j = 1, \ldots, k, j \neq i, \lim_{l \to \infty} \mathbb{P}(X \in R_i) = 1$.*

Then, the IUT with rejection region $R = \bigcap_{j=1}^{k} R_j$ is a size α test.

2.2 Two One-sided Tests (TOST)

We are now in a position to formally introduce the method for bioequivalence testing in (2.1). Albeit good theoretical guarantees for IUT, [6, 7, 8] proposed the following so-called two one-sided tests (TOST), which has now been one of the standard procedures of (2.1). As its name implies, TOST consists two one-sided hypotheses

$$H_{01} : \mu_T - \mu_R \leq \theta_L \text{ versus } H_{a1} : \mu_T - \mu_R > \theta_L \tag{2.4}$$

and

$$H_{02} : \mu_T - \mu_R \geq \theta_U \text{ versus } H_{a2} : \mu_T - \mu_R < \theta_U. \tag{2.5}$$

The $H_0 : \mu_T - \mu_R \leq \theta_L$ or $\mu_T - \mu_R \geq \theta_U$ in (2.1) can be expressed as the union of H_{01} and H_{02}. This procedure establishes bioequivalence at significance level α if both H_{01} and H_{02} are rejected at level-α. The rationale underlying is simple. If one may conclude that $\mu_T - \mu_R > \theta_L$ and also $\mu_T - \mu_R < \theta_U$, then it has in effect been concluded that $\theta_L < \mu_T - \mu_R < \theta_U$. In practice, size-$\alpha$ Student's t-test is used for (2.4) and (2.5) and H_0 is rejected if

$$T_L = \frac{(\bar{X}_T - \bar{X}_R) - \theta_L}{\hat{\sigma}_{\Delta\bar{X}}} > t_{1-\alpha,r}, \quad T_U = \frac{(\bar{X}_T - \bar{X}_R) - \theta_U}{\hat{\sigma}_{\Delta\bar{X}}} < -t_{1-\alpha,r}, \tag{2.6}$$

where $\bar{X}_T, \bar{X}_R$ are the average of log-transformed PK data for treatment/reference product, and $\hat{\sigma}_{\Delta\bar{X}} = S_p\sqrt{(\frac{1}{n_T} + \frac{1}{n_R})}$ is the standard error of $\bar{X}_T - \bar{X}_R$ and $S_p^2 = \frac{1}{n_T+n_R-2}[(n_T - 1)S_T^2 + (n_R - 1)S_R^2]$, with S_T, S_R being the standard deviation of two groups, and $t_{1-\alpha,r} = \mathbb{P}(X \leq t_{1-\alpha,r}) = 1 - \alpha$ is the critical value from t distribution with degree of freedom $r = n_T + n_R - 2$. In addition, it should be mentioned that the size for each individual t-test H_{01}, H_{02} is α, not $\alpha/2$. There is no need for multiplicity adjustment in testing each of the two one-sided null hypotheses for a univariate PK parameter. Suppose the size for each t-test is α, since TOST belongs to IUT, from Theorem 2.2, we know the level of TOST is α, that is, the size is at most α. To prove the size of TOST is exactly α, we need to check the conditions of Theorem 2.3. First, consider a parameter point $\theta_l = \mu_T - \mu_R = \theta_U$, then we have

$$\lim_{l \to \infty} \mathbb{P}_{\theta_l}(X \in R_2) = \mathbb{P}_{\theta_U}(X \in R_2) = \mathbb{P}_{\theta_U}(T_U < -t_{1-\alpha,r}) = \alpha,$$

thus the first condition in Theorem 2.3 holds true. Furthermore,

$$\lim_{l \to \infty} \mathbb{P}_{\theta_l}(X \in R_1) = \mathbb{P}_{\theta_U}(X \in R_1) = \mathbb{P}_{\theta_U}(T_L > t_{1-\alpha,r}) \to 1, \text{ as } \sigma^2 \to 0,$$

therefore, the second condition in Theorem 2.3 also remains valid, from where we summarize the above analysis to the following theorem

Theorem 2.4. *Suppose the size of each individual test in TOST is α, then the size of TOST equals α exactly.*

To compute the power of TOST, [11] derived the explicit form of the power, as a function of $(\mu_T, \mu_R, n_T, n_R, \sigma^2, \theta_L, \theta_U)$ given by

$$\text{Power}\left(\mu_T, \mu_R, n_T, n_R, \sigma^2, \theta_L, \theta_U\right) = Q_r\left(-t_{1-\alpha,r}, \frac{\mu_T - \mu_R - \theta_U}{\sigma\sqrt{\frac{1}{n_T} + \frac{1}{n_T}}}; 0, \frac{(\theta_U - \theta_L)\sqrt{r}}{2\sigma\sqrt{\frac{1}{n_T} + \frac{1}{n_T}}t_{1-\alpha,r}}\right)$$

$$- Q_r\left(t_{1-\alpha,r}, \frac{\mu_T - \mu_R - \theta_L}{\sigma\sqrt{\frac{1}{n_T} + \frac{1}{n_T}}}; 0, \frac{(\theta_U - \theta_L)\sqrt{r}}{2\sigma\sqrt{\frac{1}{n_T} + \frac{1}{n_T}}t_{1-\alpha,r}}\right)$$

where

$$Q_v(t, \delta; a, b) = \frac{\sqrt{2\pi}}{\Gamma\left(\frac{v}{2}\right) \cdot 2^{\frac{v-2}{2}}} \int_a^b \Phi\left(\frac{tx}{\sqrt{v}} - \delta\right) x^{v-1}\phi(x)dx, r = n_1 + n_2 - 2$$

is referred as the Owen's Q function and, $\Phi(\cdot), \phi(\cdot)$ are the cdf and pdf of standard normal distribution, respectively.

In the previous sections, the theoretical foundations of the TOST method for bioequivalence testing have been established. However, the relationship between the TOST and $100(1-\alpha)\%$ and $100(1-2\alpha)\%$ confidence intervals is not yet clear. In the next section, we will explore the connections between the TOST and these two types of confidence intervals, shedding light on their similarities, differences, and implications for bioequivalence testing. Understanding the connections between the TOST and different confidence intervals will provide insights into how these statistical methods relate to each other, and how they can be effectively applied in bioequivalence testing for univariate pharmacokinetic parameters. This will enhance the interpretation and application of these methods in practice, contributing to more accurate and reliable conclusions about bioequivalence.

2.3   Confidence sets

Intuitively, a hypothesis is associated with an equivalent confidence interval approach. Traditionally, bioequivalence is claimed if the $100(1 - 2\alpha)\%$ two-sided confidence interval of the geometric mean ratio, or equivalently, the arithmetic mean difference of the log-transformed data, falls within the BE limits. The use of a $100(1 - 2\alpha)\%$ confidence interval for bioequivalence testing, rather than a $100(1 - \alpha)\%$ confidence interval, has been a subject of debate among statisticians. The relationship between the size-α TOST procedure and the $100(1-2\alpha)\%$ confidence interval approach has been questioned by some researchers, who argue that the similarity is based more on an algebraic coincidence than a true statistical equivalence. [12] suggested that the association between the TOST procedure and the $100(1-2\alpha)\%$ confidence interval approach may be somewhat of a fiction, while [13] pointed out that using a $100(1-2\alpha)\%$ confidence interval for bioequivalence testing can be conservative and may only work in specific cases. Other discussions on the use of the $100(1 - 2\alpha)\%$ confidence interval procedure in bioequivalence testing can be found in works by [14],[6], [15], [16]. These discussions highlight the ongoing debate about the appropriateness of using a $100(1 - 2\alpha)\%$ confidence interval for bioequivalence testing, and the importance of understanding the underlying statistical concepts and assumptions in order to make informed decisions about the most suitable methods for a given study.

2.3.1 $100(1 - \alpha)\%$ confidence interval

As we mentioned before, there are many different formulations of the bioequivalence hypothesis that lead to alternate tests and confidence intervals. In this section, we will discuss a $100(1 - \alpha)\%$ confidence interval approach that corresponds exactly to a size-α TOST. It is well-known that there is a closed relationship between the level-α hypothesis and $100(1 - \alpha)\%$ confidence set, that is, rejecting H_0 if and only if the intersection of the $100(1 - \alpha)\%$ confidence set and the null hypothesis is empty. We summarize the property into the following theorem

Theorem 2.5. *Let Θ be the parameter space. For each $\theta_0 \in \Theta$, let T_{θ_0} be a test statistic for $H_0 : \theta = \theta_0$ with significance level α and acceptance region $A(\theta_0)$, then the set $C(X) = \{\theta : X \in A(\theta)\}$ is a level-α confidence set for θ.*

In other words, suppose $[L(X), R(X)]$ is a confidence interval of parameter $\theta \in \Theta$ with condidence coefficient equal to $1 - \alpha$, that is, $\inf_\theta \mathbb{P}_{\theta \in \Theta}(\theta \in [L(X), R(X)]) = 1 - \alpha$. Consider the hypothesis testing $H_0 : \theta \in \Theta_0$ verse $H_a : \theta \in \Theta_0^c$. Then from Theorem 2.5, we know that the test that rejects H_0 if and only if $[L(X), R(X)] \bigcap \Theta_0^c = \varnothing$ is a level-α test. In this section, we will apply Theorem 2.5 to show that a size-α TOST is associated with a $100(1 - \alpha)\%$ confidence interval.

Theorem 2.6. *Consider the hypothesis testing $H_0 : \mu_T - \mu_R \leq \theta_L$ or $\mu_T - \mu_R \geq \theta_U$ versus $H_a : \theta_L < \mu_T - \mu_R < \theta_U$ in (2.1) and define the following confidence interval for $\mu_T - \mu_R$*

$$[L(X), U(X)] = \left[\min\{0, (\bar{X}_T - \bar{X}_R) - t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}, \max\{0, (\bar{X}_T - \bar{X}_R) + t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\} \right], \quad (2.7)$$

where

$$\hat{\sigma}_{\Delta\bar{X}} = S_p\sqrt{\frac{1}{m} + \frac{1}{n}}, S_p = \sqrt{\frac{(m-1)S_{X_T}^2 + (n-1)S_{X_R}^2}{m + n - 2}}, r = m + n - 2, \quad (2.8)$$

and S_{X_T}, S_{X_R} are the sample standard derivation of the two groups. Then the following two statements hold true

label=(0) If $\mu_T - \mu_R \neq 0$, then the coverage probability of $[L(X), U(X)]$ is $100(1 - \alpha)\%$, otherwise, the coverage probability equals 1.

lbbel=(0) $[L(X), U(X)]$ is associated with the size-α TOST for (2.1).

Following Theorem 2.6, we could make the following remarks.

Remark 2.2.

1. From (1) in Theorem 2.6, we know $[L(X), U(X)]$ is a level-α confidence interval.

2. If $[L(X), U(X)]$ contains zero, then it is identical to the $100(1 - 2\alpha)\%$ confidence interval which will be introduced in the next section.

3. From Bayesian point of view, if the prior distribution of $\mu_T - \mu_R$ is noninformative, that is, $\pi(\mu_T - \mu_R) \propto 1$, then the posterior credible probability of $[L(X), U(X)]$ is exactly $(1 - 2\alpha)$ for $-t_{1-\alpha,r} \leq \mu_T - \mu_R \leq t_{1-\alpha,r}$ and converges to $(1 - \alpha)$ as $|\mu_T - \mu_R| \to \infty$.

2.3.2 $100(1 - 2\alpha)\%$ confidence interval

Following the guidance of the FDA, in practice, bioequivalence is claimed if the $100(1-2\alpha)\%$ two-sided confidence interval of the geometric mean ratio falls within the BE limits. However, unfortunately, from Theorem 2.5, we could only see the connection between TOST and $100(1-\alpha)\%$ confidence interval. The reason why $100(1-2\alpha)\%$ confidence interval yields a size-α test is still unclear. In this section, owing to the "equal-tail" property, we will see that is not only an "algebraic coincidence", but also theoretically guaranteed. In general, the size of a test associated with a $100(1 - 2\alpha)\%$ confidence interval is not α. For example, consider the following $100(1 - 2\alpha)\%$ confidence interval

$$[L(X), U(X)] = \left[(\bar{X}_T - \bar{X}_R) - t_{1-\alpha_1,r}\hat{\sigma}_{\Delta\bar{X}}, (\bar{X}_T - \bar{X}_R) + t_{1-\alpha_2,r}\hat{\sigma}_{\Delta\bar{X}} \right], \quad (2.9)$$

where $\alpha_1 + \alpha_2 = 2\alpha$. It is obvious that $[L(X), U(X)]$ defined above is a $100(1-2\alpha)\%$ confidence interval. As $\alpha_1 \to 0$, the confidence interval reduces to $\left(-\infty, (\bar{X}_T - \bar{X}_R) + t_{1-2\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\right]$, and the size associated with this confidence interval is 2α, not α. In fact, the lower confidence interval $[L(x), \infty) = [(\bar{X}_T - \bar{X}_R) - t_{1-\alpha_1,r}\hat{\sigma}_{\Delta\bar{X}}, \infty)$ in (2.9) defines a size-α_1 test, and similarly, the upper confidence interval $(-\infty, U(x)] = (\infty, (\bar{X}_T - \bar{X}_R) + t_{1-\alpha_2,r}\hat{\sigma}_{\Delta\bar{X}}]$ in (2.9) defines a size-α_2 test. From the IUT theory in Theorem 2.2 and Theorem 2.3, we know the size of the defined by $[L(X), U(X)]$ in (2.9) is $\max\{\alpha_1, \alpha_2\}$, which equals α only if $\alpha_1 = \alpha_2$. We summarize the above analysis into the following theorem.

Theorem 2.7. *Let X be a random sample from a probability distribution with statistical parameter θ. Suppose $[L(X), \infty)$ is a $100(1-\alpha_1)\%$ lower confidence interval for θ and $(-\infty, U(X)]$ is a $100(1-\alpha_2)\%$ upper confidence interval for θ and Then $[L(x), U(x)]$ is a $100(1-\alpha_1-\alpha_2)\%$ confidence interval for θ.*

Remark 2.3. The above theorem reveals the importance of "equal-tail", that is $\alpha_1 = \alpha_2$, without this property, the TOST will not yields a size-α test. In other words, the total uncertainty of 2α should be spent half above and half below the observed mean. Moreover, this theorem also provides another reason why θ_L, θ_U in (2.1) should be symmetric about zero. Otherwise, for example, if $\theta_L + \theta_U < 0$, it is reasonable to put more weights on α_1 than α_2, in which case, the size of $100(1-2\alpha)\%$ procedure is no longer α.

2.4 Illustration

In this section, we present an illustrative example to showcase the application of two distinct confidence interval approaches. We utilized data from the study conducted by [17] on Ticlopidine hydrochloride—a platelet aggregation inhibitor employed in managing and preventing thromboembolic disorders. The study aimed to assess the bioequivalence of a new formulation of ticlopidine hydrochloride (test) against the established market product Tiklid. This evaluation was pivotal for registering the new formulation as a generic through the Abbreviated New Drug Application (ANDA) procedure. Essential pharmacokinetic (PK) parameters are detailed in Table 1. Notably, the dataset exhibits a right-skewed and heavy-tailed distribution. Post-logarithmic transformation, the average values for Cmax stand at 6.64 and 6.72 in each group, with a corresponding estimated standard deviation of $\hat{\sigma}\Delta\bar{X} = 0.059$. Similarly, for the second PK parameter, AUC, the averages are 5.67 and 5.76, with $\hat{\sigma}\Delta\bar{X} = 0.066$. Employing both the $100(1-2\alpha)\%$ and $100(1-\alpha)\%$ confidence interval methods, both approaches conclude the bioequivalence of the treatment drug.

Table 1. Pharmacokinetic parameters of ticlopidine measured in healthy volunteers after treatment with test and reference drug

Subject ID	Cmax Treatment	Cmax Reference	AUC Treatment	AUC Reference
1	784.3	878.2	2131.4	3030.1
2	304.2	211.7	1107.9	798.6
3	307.3	259.6	806.2	696.1
4	156.7	307.8	509.9	833.7
5	745.6	1036.2	2784.0	3015.6
6	295.1	443.1	1391.2	1159.5
7	89.6	120.1	248.1	584.8
8	321.1	329.8	672.4	960.7
9	310.8	316.1	1105.8	1048.4
10	475.4	617.8	1321.8	1797.4
11	228.6	249.2	638.1	725.1
12	241.3	314.1	908.2	1003.9
13	78.7	166.8	225.7	280.2
14	351.5	455.0	860.0	960.9
15	341.7	406.5	1020.8	1208.3
16	221.3	240.9	588.4	715.8
17	443.5	378.8	1026.8	910.6
18	275.2	278.4	984.0	801.8
19	273.6	209.6	629.0	577.5
20	341.1	212.4	1056.1	662.9
21	262.3	162.1	694.6	414.0
22	178.1	212.1	420.9	572.0
23	470.1	394.2	1304.7	1101.4
24	372.6	567.0	1133.2	1433.7

2.5 More powerful tests

From the previous analysis, we know that there are various procedures for testing (2.1), and a natural question arises, which one is better? Unfortunately, even though the FDA suggests using $100(1-2\alpha)\%$ to test bioequivalence, the procedure is not the best. In fact, it should be remarked that TOST has been criticized by many authors. On the one hand, TOST yields a biased test, on the other hand, the power of TOST is quite low. In practice, we would like a hypothesis test to be more likely to reject H_0 if $\theta \in \Theta_0^c$ than if $\theta \in \Theta_0$, that is, $\beta(\theta_2) \geq \beta(\theta_1)$ for every $\theta_1 \in \Theta_0$ and $\theta_2 \in \Theta_0^c$. The following definition summarizes the tests summarizing this property.

Definition 2.4. Let α be the significance level. A hypothesis test $H_0 : \theta \in \Theta_0$ versus $H_a : \theta \in \Theta_0^c$ with power function $\beta(\cdot)$ is said to be unbiased if and only if

$$\beta(\theta) \leq \alpha, \theta \in \Theta_0 \text{ and } \beta(\theta) \geq \alpha \in \Theta_0^c.$$

Unfortunately, even using a $100(1-2\alpha)\%$ confidence interval approach, TOST is still a biased test. Furthermore, it has also been shown that TOST is conservative and inefficient under an asymptotic setup. Much attention has been paid to equivalence testing problem in (2.1) for decades, to mention but a few, [18], [19], [20], [16], [21], [22], [23]. See [24], [25] for a comprehensive review. Even though many methodologies have been proposed to test bioequivalence, the theoretical results are limited. Except some special models, like normal distribution with known variance, no finite sample optimality theory is available for tests of equivalence. To the best of our knowledge, the work by [26] is the first theoretical result where the asymptotic optimality theory is established. Since the optimality theory is beyond the scope of the paper, we only show two examples to give the author some intuitions about optimal testing of equivalence.

Example 2.8 (Normal Distribution with Known Variance). *Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where σ is known. Consider the hypothesis testing: $H_0 : |\mu| \geq \theta$ versus $H_a : |\mu| < \theta$. Then the uniformly most powerful level-α test is rejecting H_0 is $n^{1/2}|\bar{X}_n| \leq \psi(\alpha, n^{1/2}\theta, \sigma)$, where $\psi(\alpha, \theta, \sigma)$ satisfies*

$$\phi(\frac{\psi - \theta}{\sigma}) - \phi(\frac{-\psi - \theta}{\sigma}) = \alpha,$$

and $\phi(\cdot)$ is the cdf of standard normal distribution.

However, the uniformly most powerful test does not exist is σ is unknown. The next example generalizes the results to the exponential family.

Example 2.9 (One-parameter Exponential Family). *Suppose $X_1, \ldots, X_n$ are i.i.d. generated from the following one-parameter exponential family*

$$f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x),$$

where $\eta(\theta)$ is an increasing function of θ only, $h(x)$ is a function of x only, and $Y(\cdot)$ has monotone likelihood ratio in $(X_1, \ldots, X_n)$. Then the uniformly most powerful level-α test for the equivalence hypothesis testing $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ versus $H_1 : \theta_1 \leq \theta \leq \theta_2$ is

$$T(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } c_1 < Y(X_1, \ldots, X_n) < c_2 \\ \gamma_i & \text{if } Y(X_1, \ldots, X_n) = c_i, i = 1, 2 \\ 0 & \text{if } Y(X_1, \ldots, X_n) < c_1 \text{ or } Y(X_1, \ldots, X_n) > c_2, \end{cases}$$

where $\gamma_1, \gamma_2, c_1, c_2$ are determined by $\beta(\theta_1) = \beta(\theta_2) = \alpha$.

3 Conclusion

The paper discusses the significance of bioequivalence studies, which compare the biological equivalence of two different formulations of a drug. These studies are conducted in clinical settings with human subjects to compare

the pharmacokinetic profiles of the two formulations. Acceptance criteria for bioequivalence are generally based on statistical methods that evaluate the similarities between the two formulations. The paper explains the theory of intersection-union tests and how they can be used to determine bioequivalence. The TOST (two one-sided tests) approach is the standard approach to bioequivalence testing and has been shown to maintain a size-α test by rejecting the null hypothesis only if both one-sided hypotheses are rejected. The paper provides insights into the theory behind bioequivalence testing, including the use of geometric mean, the non-symmetry of bioequivalence limits around 1, and the product of bioequivalence limits equaling 1. The paper also discusses the connection between the $100(1-2\alpha)\%$ confidence interval approach and $100(1-\alpha)\%$ confidence interval approach. Future work will address the imperative issue of multiplicity control in bioequivalence studies. The demonstration of bioequivalence between two drugs typically involves two pharmacokinetic parameters: the area under the concentration-time curve (AUC) and maximum concentration (Cmax). since the decision rule for bioequivalence often requires equivalence to be achieved simultaneously on both parameters, this consideration becomes paramount in advancing the robustness and reliability of bioequivalence testing methodologies.

Acknowledgment

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Leday, Gwenaël GR, Hemerik, Jesse, Engel, Jasper, van der Voet, Hilko. Improved family-wise error rate control in multiple equivalence testing. Food and Chemical Toxicology. Elsevier. 2023;178:113928.

[2] Pallmann, Philip, Jaki, Thomas. Simultaneous confidence regions for multivariate bioequivalence. Statistics in Medicine. Wiley Online Library. 2017;36(29):4585-4603.

[3] He, Yingdong, Deng, Yuhao, You, Chong, Zhou, Xiao-Hua. Equivalence tests for ratio of means in bioequivalence studies under crossover design. Statistical Methods in Medical Research. SAGE Publications Sage UK: London, England. 2022;31(7):1405-1419.

[4] Dufera, Abdisa G, Xiong, Cui, Xu, Jin. Simultaneous confidence intervals from randomization tests with application in testing bioequivalence with multiple endpoints. Biometrical Journal. Wiley Online Library. 2023;2200082.

[5] Boulaguiem, Younes, Quartier, Julie, Lapteva, Maria, Kalia, Yogeshvar N, Victoria-Feser, Maria-Pia, Guerrier, Stéphane, Couturier, Dominique-Laurent. Finite sample adjustments for average equivalence testing. bioRxiv. Cold Spring Harbor Laboratory. 2023;2023-03.

[6] Westlake, WJ. Bioequivalence testing-a need to rethink. Biometrics. JSTOR. 1981;37(3):589-594.

[7] Schuirmann, DL. On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval. Biometrics. International Biometric SOC 808 17th ST NW Suite 200, Washington, DC 20006-3910. 1981;37(3):617-617.

[8] Schuirmann, Donald J. 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. Springer. 1987;15(6):657-680.

[9] Shen, Haipeng, Brown, Lawrence D, Zhi, Hui. Efficient estimation of log-normal means with application to pharmacokinetic data. Statistics in Medicine. Wiley Online Library. 2006;25(17):3023-3038.

[10] Shen, Meiyu, Russek-Cohen, Estelle, Slud, Eric V. Checking distributional assumptions for pharmacokinetic summary statistics based on simulations with compartmental models. Journal of Biopharmaceutical Statistics. Taylor & Francis. 2017;27(5):756-772.

[11] Phillips, Kem F. Power of the two one-sided tests procedure in bioequivalence. Journal of Pharmacokinetics and Biopharmaceutics. Springer. 1990;18(2):137-144.

[12] Brown, Lawrence D, Casella, George, Gene Hwang, JT. Optimal confidence sets, bioequivalence, and the limacon of Pascal. Journal of the American Statistical Association. Taylor & Francis. 1995;90(431):880-889.

[13] Berger, Roger L, Hsu, Jason C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science. Institute of Mathematical Statistics. 1996;11(4):283-319.

[14] Westlake, Wilfred J. Symmetrical confidence intervals for bioequivalence trials.Biometrics. JSTOR. 1976;741-744.

[15] Rani, Shubha, Pargal, A, others. Bioequivalence: An overview of statistical concepts. Indian Journal of Pharmacology. Indian Pharmacological Society. 2004;36:209-216.

[16] Choi, Leena, Caffo, Brian, Rohde, Charles. A survey of the likelihood approach to bioequivalence trials. Statistics in Medicine. Wiley Online Library. 2008;27(24):4874-4894.

[17] Marzo, Antonio, Dal Bo, Lorenzo, Rusca, Antonio, Zini, Pierangelo. Bioequivalence of ticlopidine hydrochloride administered in single dose to healthy volunteers. Pharmacological Research. Elsevier. 2002;46(5):401-407.

[18] Anderson, Sharon, Hauck, Walter W. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Communications in Statistics-Theory and Methods. Taylor & Francis. 1983;12(23):2663-2692.

[19] Brown, Lawrence D, Hwang, JT Gene, Munk, Axel. An unbiased test for the bioequivalence problem. The annals of Statistics. Institute of Mathematical Statistics. 1997;25(6):2345-2367.

[20] Martín Andrés, A, Herranz Tejedor, I. Asymptotical tests on the equivalence, substantial difference and non-inferiority problems with two proportions. Biometrical Journal: Journal of Mathematical Methods in Biosciences. Wiley Online Library. 2004;46(3):305-319.

[21] Fogarty, Colin B, Small, Dylan S. Equivalence testing for functional data with an application to comparing pulmonary function devices. The Annals of Applied Statistics. JSTOR. 2014;2002-2026.

[22] Pesarin, Fortunato, Salmaso, Luigi, Carrozzo, Eleonora, Arboretti, Rosa. Union–intersection permutation solution for two-sample equivalence testing. Statistics and Computing. Springer. 2016;26(3):693-701.

[23] Li, Kexuan, Yang, Lingli, Zhao, Shaofei, Sinks, Susie, Lin, Luan, Sun, Peng. Multivariate rank-based analysis of multiple endpoints in clinical trials: A global test approach; 2023. arXiv preprint arXiv:2306.15380

[24] Wellek, Stefan. Testing statistical hypotheses of equivalence. Chapman and Hall/CRC; 2002.

[25] Meyners, Michael. Equivalence tests-A review. Food Quality and Preference. Elsevier. 2012;26(2):231-245.

[26] Romano, Joseph P. Optimal testing of equivalence hypotheses. The Annals of Statistics. Institute of Mathematical Statistics. 2005;33(3):1036-1047.

Appendix

A   Proof of Theorem 2.1

Proof. The proof of (1) can be accomplished by direct calculations. It can be shown that

$$\exp(\bar{X}) = \exp(n^{-1}\sum_{i=1}^{n} X_i) = \exp\{\log[\exp(\sum_{i=1}^{n} X_i/n)]\} = \exp\{\log[\prod_{i=1}^{n} \exp(X_i/n)]\}$$

$$= (\prod_{i=1}^{n} \exp(X_i))^{\frac{1}{n}} = (\prod_{i=1}^{n} X^*)^{\frac{1}{n}} = \mathrm{GM}(X^*).$$

To prove (2), we let $F_{X^*}(x^*) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\log(x^*)-\mu}{\sqrt{2}\sigma}\right)\right]$ be the cumulative distribution function of the lognormal distribution, where $\mathrm{erf}(x^*) = \frac{2}{\pi}\int_{0}^{\frac{\pi}{2}}\exp\left(-\frac{x^{*2}}{\sin^2\theta}\right)d\theta$ is the complementary error function. For simplicity, we let the median of X^* equals ϱ, that is, $Median(X^*) = \varrho$. Then, from the definition of median, we have $F_{X^*}(\varrho) = 0.5$, thus $\varrho = F_{X^*}^{-1}(0.5)$. Therefore, it follows that

$$\varrho = F_{X^*}^{-1}(0.5) = \exp\left[\sigma\sqrt{2}\cdot\mathrm{erf}^{-1}(2p-1) + \mu\right]\Big|_{p=0.5},$$

where $\mathrm{erf}^{-1}(\cdot)$ is the inverse of the complementary error function satisfying $\mathrm{erf}^{-1}(0) = 0$, thus, the median of X^* equals $\exp(\mu)$. In fact, it can be also seen that the logarithm function is a monotonous increaseing function, therefore, the median of X^* is the exponential of the median of $\log(X^*)$. Since $\log(X^*)$ is normal distributed, thus, median of X^* equals $\exp(\mu)$.

Let $f(x^*) = \frac{1}{x^*\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(x^*)-\mu)^2}{2\sigma^2}\right)$ be the pdf of lognormal distribution. Consider the expected value of $X_i^{*\frac{1}{n}}$, $[X_i^{*\frac{1}{n}}]$. It follows that

$$
\begin{aligned}
[X_i^{*\frac{1}{n}}] &= \int_0^{\infty} \frac{1}{x^*\sigma\sqrt{2\pi}} x^{*\frac{1}{n}} \exp\left(-\frac{(\ln(x^*)-\mu)^2}{2\sigma^2}\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{\exp(t)} \exp(t)^{1/n} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \exp(t) dt \quad (\text{A.1}) \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\{(t-\mu)^2 n - 2\sigma^2 t\}}{2n\sigma^2}\right) dt \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left\{n\left(t-\mu-\frac{\sigma^2}{n}\right)^2 - \frac{\sigma^4}{n} - 2\sigma^2\mu\right\}}{2n\sigma^2}\right) dt \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{n\left(t-\mu-\frac{\sigma^2}{n}\right)^2}{2n\sigma^2}\right) \exp\left(\frac{\frac{\sigma^4}{n} + 2\mu\sigma^2}{2n\sigma^2}\right) dt \\
&= \exp\left(\frac{\mu}{n} + \frac{\sigma^2}{2n^2}\right) \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(t-\mu-\frac{\sigma^2}{n}\right)^2}{2\sigma^2}\right) dt, \quad (\text{A.2})
\end{aligned}
$$

where (A.1) uses the change of variable technique, i.e., $x = \exp(t)$. Note that $\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(t-\mu-\frac{\sigma^2}{n}\right)^2}{2\sigma^2}\right)$ in (A.2)

is the pdf of a normal distribution with mean $\mu + \frac{\sigma^2}{n}$ variance σ^2, thus $\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(t-\mu-\frac{\sigma^2}{n}\right)^2}{2\sigma^2}\right)dt = 1$,

and therefore, $[X_i^{*\frac{1}{n}}] = \exp\left(\frac{\mu}{n} + \frac{\sigma^2}{2n^2}\right)$.

The result of (3) follows from

$$[\text{GM}(X^*)] = [\prod_{i=1}^{n}X_i^*)^{\frac{1}{n}}] = \prod_{i=1}^{n}[X_i^*)^{\frac{1}{n}}] = \prod_{i=1}^{n}\exp\left(\frac{\mu}{n} + \frac{\sigma^2}{2n^2}\right) = \exp\left(\mu + \frac{\sigma^2}{2n}\right).$$

$\square$

B   Proof of Theorem 2.2

Proof. Let $\theta \in \bigcup_{i=1}^{k}\Theta_i$. Then $\theta \in \Theta_i$, for some $i = 1, \ldots, k$ and

$$\mathbb{P}_\theta(X \in R) = \mathbb{P}_\theta(X \in \bigcap_{i=1}^{k}R_i) \leq \mathbb{P}_\theta(X \in R_i) \leq \alpha_i \leq \sup_{i=1,\ldots,k}\alpha_i = \alpha.$$

Since the above equation holds true for arbitrary $\theta \in \bigcup_{i=1}^{k}\Theta_i$, the defined IUT test is a level α test. $\square$

C   Proof of Theorem 2.3

Proof. By Theorem 2.2, we know rejection region $R = \bigcap_{j=1}^{k}R_j$ yields a level-α test, that is

$$\sup_{\theta\in\bigcup_{j=1}^{k}\Theta_j}\mathbb{P}_\theta(X \in R) \leq \alpha. \tag{C.1}$$

Next, we are going to show that $\sup_{\theta\in\bigcup_{j=1}^{k}\Theta_j}\mathbb{P}_\theta(X \in R) \geq \alpha$. Because $\theta_i \in \bigcup_{j=1}^{k}\Theta_j$, therefore,

$$\sup_{\theta\in\bigcup_{j=1}^{k}\Theta_j}\mathbb{P}_\theta(X \in R) \geq \lim_{l\to\infty}\mathbb{P}_{\theta_l}(X \in R)$$

$$= \lim_{l\to\infty}\mathbb{P}_{\theta_l}(X \in \bigcap_{j=1}^{k}R_j)$$

$$\geq \lim_{l\to\infty}\sum_{j=1}^{k}\mathbb{P}_{\theta_l}(X \in R_j) - k + 1,$$

where the last inequality follows from Bonferroni's inequality. By condition (1) and (2), we can get $\sum_{j=1}^{k}\mathbb{P}_{\theta_l}(X \in R_j) = k - 1 + \alpha$. Thus,

$$\sup_{\theta\in\bigcup_{j=1}^{k}\Theta_j}\mathbb{P}_\theta(X \in R) \geq k - 1 + \alpha - k + 1 = \alpha. \tag{C.2}$$

Combine (C.1) and (C.2), we prove that the size of IUT with rejection region $R = \bigcap_{j=1}^{k}R_j$ is exactly α. $\square$

D   Proof of Theorem 2.4

Proof. Because TOST belongs to IUT, by Theorem 2.2, we know the size of TOST is at most α. Consider a parameter point $\theta_l = \mu_T - \mu_R = \theta_U$, then we have

$$\lim_{l \to \infty} \mathbb{P}_{\theta_l}(X \in R_2) = \mathbb{P}_{\theta_U}(X \in R_2) = \mathbb{P}_{\theta_U}(T_U < -t_{1-\alpha,r}) = \alpha,$$

Furthermore,

$$\lim_{l \to \infty} \mathbb{P}_{\theta_l}(X \in R_1) = \mathbb{P}_{\theta_U}(X \in R_1) = \mathbb{P}_{\theta_U}(T_L > t_{1-\alpha,r}) \to 1, \text{ as } \sigma^2 \to 0,$$

therefore, the conditions in Theorem 2.3 hold true. Thus, the size of TOST is exactly α. $\square$

E   Proof of Theorem 2.5

Proof. By the definition of acceptance region, we have

$$\sup_{\theta = \theta_0} \mathbb{P}(X \notin A(\theta_0)) = \sup_{\theta = \theta_0} \mathbb{P}(T_{\theta_0} = 1) \leq \alpha,$$

which is the same as

$$1 - \alpha \leq \inf_{\theta = \theta_0} \mathbb{P}(X \in A(\theta_0)) = \inf_{\theta = \theta_0} \mathbb{P}(\theta_0 \in C(X)).$$

Since the above statements hold true for all $\theta \in \Theta$, thus

$$\inf_{\theta_0 \in \Theta} \inf_{\theta = \theta_0} \mathbb{P}(\theta_0 \in C(X)) \geq 1 - \alpha,$$

which implies $C(X) = \{\theta : X \in A(\theta)\}$ is a level-α confidence set for θ. $\square$

F   Proof of Theorem 2.6

Proof. If $\mu_T - \mu_R = 0$, then $\mu_T - \mu_R \in [L(X), U(X)]$ for all X, thus the coverage probability of $[L(X), U(X)]$ is 1 in this case. So it is only need to consider the case when $\mu_T - \mu_R \neq 0$.

When $\mu_T - \mu_R > 0$, the event $\{\mu_T - \mu_R \geq \min\{0, (\bar{X}_T - \bar{X}_R) - t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\}$ is the whole sample space, thus

$$\mathbb{P}\left(\{\mu_T - \mu_R \geq \min\{0, (\bar{X}_T - \bar{X}_R) - t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\}\right) = 1.$$

And if $\mu_T - \mu_R > 0$, the event $\{\mu_T - \mu_R \leq \max\{0, (\bar{X}_T - \bar{X}_R) + t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\}$ is the same as event $\{\mu_T - \mu_R \leq (\bar{X}_T - \bar{X}_R) + t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}$, so the coverage probability equals

$$\begin{aligned}
\mathbb{P}\left(\mu_T - \mu_R \in [L(X), U(X)]\right) &= \mathbb{P}\left(\mu_T - \mu_R \leq \max\{0, (\bar{X}_T - \bar{X}_R) + t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\right) \\
&= \mathbb{P}\left(\mu_T - \mu_R \leq (\bar{X}_T - \bar{X}_R) + t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\right) \\
&= \mathbb{P}\left(\bar{X}_T - \bar{X}_R - (\mu_T - \mu_R) \geq -t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\right) \\
&= 1 - \alpha.
\end{aligned}$$

Similarly, when $\mu_T - \mu_R < 0$, the event $\{\mu_T - \mu_R \leq \max\{0, (\bar{X}_T - \bar{X}_R) + t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\}$ is the whole sample space, thus

$$\mathbb{P}\left(\{\mu_T - \mu_R \leq \max\{0, (\bar{X}_T - \bar{X}_R) + t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\}\right) = 1.$$

If $\mu_T - \mu_R < 0$, the event $\{\mu_T - \mu_R \geq \min\{0, (\bar{X}_T - \bar{X}_R) - t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\}$ is the same as event $\{\mu_T - \mu_R \geq (\bar{X}_T - \bar{X}_R) - t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}$, so the coverage probability equals

$$
\begin{aligned}
\mathbb{P}\left(\mu_T - \mu_R \in [L(X), U(X)]\right) &= \mathbb{P}\left(\mu_T - \mu_R \geq \min\{0, (\bar{X}_T - \bar{X}_R) - t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\}\right) \\
&= \mathbb{P}\left(\mu_T - \mu_R \geq (\bar{X}_T - \bar{X}_R) - t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\right) \\
&= \mathbb{P}\left(\bar{X}_T - \bar{X}_R - (\mu_T - \mu_R) \leq t_{1-\alpha,r}\hat{\sigma}_{\Delta\bar{X}}\right) \\
&= 1 - \alpha.
\end{aligned}
$$

Combine everything above, the proof is completed. □

G   Proof of Theorem 2.7

Proof. By definition, we have

$$
\mathbb{P}(L(X) \leq \theta) = 1 - \alpha_1, \mathbb{P}(U(X) \geq \theta) = 1 - \alpha_2.
$$

Consider the events $\mathcal{A}$ and $\mathcal{B}$ defined as $\mathcal{A} = \{X : L(X) \leq \theta\}, \mathcal{B} = \{X : U(X) \geq \theta\}$. Therefore, the interaction of $\mathcal{A}$ and $\mathcal{B}$ is $\mathcal{A}\bigcap\mathcal{B} = \{X : L(X) \leq \theta \leq U(X)\}$. It follows that

$$
\mathbb{P}\left(\mathcal{A}\bigcup\mathcal{B}\right) = \mathbb{P}\left(L(X) \leq \theta \text{ or } U(X) \geq \theta\right) \geq \mathbb{P}\left(L(X) \leq \theta \text{ or } L(X) \geq \theta\right) = 1.
$$

Since $\mathbb{P}(\mathcal{A}\bigcup\mathcal{B}) \leq 1$, we have $\mathbb{P}(\mathcal{A}\bigcup\mathcal{B}) = 1$. The result follow from that

$$
\mathbb{P}\left(\mathcal{A}\bigcap\mathcal{B}\right) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) - \mathbb{P}\left(\mathcal{A}\bigcup\mathcal{B}\right) = 1 - \alpha_1 - \alpha_2.
$$

□