Finite Sample Corrections for Average Equivalence Testing

Younes Boulaguiem[1], Julie Quartier[2,3], Maria Lapteva[2,3], Yogeshvar N Kalia[2,3], Maria-Pia Victoria-Feser[1], Stéphane Guerrier[1,2,3] & Dominique-Laurent Couturier[4,5,*]

[1]Geneva School of Economics and Management, University of Geneva, Switzerland; [2]School of Pharmaceutical Sciences, University of Geneva, Switzerland; [3]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Switzerland; [4]Medical Research Council Biostatistics Unit, University of Cambridge, England; [5]Cancer Research UK – Cambridge Institute, University of Cambridge, England; *Corresponding author: dominique.couturier@mrc-bsu.cam.ac.uk

Abstract: Average (bio)equivalence tests are used to assess if a parameter, like the mean difference in treatment response between two conditions for example, lies within a given equivalence interval, hence allowing to conclude that the conditions have 'equivalent' means. The *Two One-Sided Tests* (TOST) procedure, consisting in testing whether the target parameter is respectively significantly greater and lower than some pre-defined lower and upper equivalence limits, is typically used in this context, usually by checking whether the confidence interval for the target parameter lies within these limits. This intuitive and visual procedure is however known to be conservative, especially in the case of highly variable drugs, where it shows a rapid power loss, often reaching zero, hence making it impossible to conclude for equivalence when it is actually true. Here, we propose a finite sample correction of the TOST procedure, the α-TOST, which consists in a correction of the significance level of the TOST allowing to guarantee a test size (or type-I error rate) of α. This new procedure essentially corresponds to a finite sample and variability correction of the TOST procedure. We show that this procedure is uniformly more powerful than the TOST, easy to compute, and that its operating characteristics outperform the ones of its competitors. A case study about econazole nitrate deposition in porcine skin is used to illustrate the benefits of the proposed method and its advantages compared to other available procedures.

Keywords: similarity test, bioequivalence, scaled average bioequivalence, two one-sided test, interval inclusion principle.

Acknowledgements: S. Guerrier is supported by the SNSF Grants #176843 and # 211007 as well as by the Innosuisse Grants #37308.1 IP-ENG and #53622.1 IP-ENG. Y. Boulaguiem and M.-P. Victoria-Feser are partially supported by the SNSF Grant #182684. D.-L. Couturier is partially supported by the Cancer Research UK grant C9545/A29580. Y. N. Kalia thanks the University of Geneva for teaching assistantships for J. Quartier and M. Lapteva, for providing financial support for the purchase of the Waters Xevo TQ-MS detector, and also thanks the Fondation Ernst and Lucie Schmidheiny and the Société Académique de Genève for providing equipment grants.

Conflict of Interest: None declared.

Data availability statement: The econazole nitrate deposition data as well as an implementation of the method proposed in this paper are available in the `cTOST` R package accessible on the GitHub repository: `yboulag/cTOST`.

1. Introduction

Equivalence tests, also known as similarity or parity tests, have gained significant attention during the past two decades. They originated from the field of pharmacokinetics (Metzler, 1974; Westlake, 1976), where they are called bioequivalence tests and have numerous applications in both research and production (Pallmann and Jaki, 2017). They find their most common application in the manufacturing of generic medicinal drugs, where, by proving that the generic version has a similar bioavailability to its well-studied brand-name counterpart, the manufacturer can considerably shorten the approval process for the generic drug (for an overview, see e.g., Senn, 2021). Moreover, equivalence tests have attracted growing interest in other domains and for other types of purposes, such as in production when, for example, the mode of administration is altered or when the production site is changed (see e.g., Patterson and Jones, 2006), or in the social and behavioral sciences for the evaluation of replication results and corroborating risky predictions (Lakens, 2017). Very recent literature reflects the expanding use of equivalence tests across a growing range of domains. Examples include the investigation of the equivalence of virtual reality imaging measurements by feature (Sureshkumar et al., 2022), of cardiovascular responses to stimuli by sex (O'Brien and Kimmerly, 2022), of children neurodevelopment (Wehrle et al., 2022) or chemotherapy efficacy and safety (Sansone et al., 2022) by treatment, of post-stroke functional connectivity patterns by patient group (Branscheidt et al., 2022), of risk-taking choices by moral type (Feri et al., 2023), of 2020 US presidential election turnout by political advertising condition (Aggarwal et al., 2023). Review articles have also appeared, for example, in food sciences (Meyners, 2012), in psychology (Lakens et al., 2018) in sport sciences (Mazzolari et al., 2022), and in pharmaceutical sciences (Wang et al., 2022).

Equivalence testing implies defining an equivalence region within which the parameter of interest, such as the difference between outcome means measured under two conditions, would lie for these conditions to be considered equivalent. When comparing two treatments, for example, deviations belonging to this region would typically be deemed irrelevant as they are too small to affect the similarity of their therapeutic effects. This is different from standard equality-of-means hypothesis tests in which the null and the alternative hypothesis are interchanged and the null hypothesis states that both means are equal rather than equivalent.

Formally, a canonical form for the average equivalence problem consists of two independent random variables $\widehat{\theta}$ and $\widehat{\sigma}_\nu$ having the distributions

$$\widehat{\theta} \sim \mathcal{N}\left(\theta, \sigma_\nu^2\right) \quad \text{and} \quad \frac{\nu \widehat{\sigma}_\nu^2}{\sigma_\nu^2} \sim \chi_\nu^2, \tag{1}$$

where θ and σ_ν^2 respectively denote the target equivalence parameter and its variance, depending on the number of the degrees of freedom ν which is a function of the sample size and total number of parameters. This setting is very general. It covers cases where, for example, the bioequivalence parameter corresponds to the difference in means of the (logarithm of pharmacokinetic) responses between two experimental conditions, or to an element of the parameter vector of a (generalised) linear mixed effect model, like the difference between the slopes of two conditions in a longitudinal study. The hypotheses of interest are given by

$$\mathrm{H}_0: \ \theta \notin \Theta_1 \quad \text{vs.} \quad \mathrm{H}_1: \ \theta \in \Theta_1 := (\theta_L, \theta_U), \tag{2}$$

where θ_L and θ_U are known constants. Without loss of generality, it can be assumed that the equivalence limits are symmetrical around zero. In this case, $c := \theta_U = -\theta_L$ so that $\Theta_1 = (-c, c)$. Equivalence is typically investigated via the *Two One-Sided Tests* (TOST) procedure (Schuirmann, 1987), consisting in testing whether the target parameter is respectively significantly greater than $-c$ and lower than c, with the test size, or type-I error rate, controlled at the significance (or nominal) level α (see Berger, 1982), usually chosen as 5%. More precisely, the TOST is level-α, meaning that its size is smaller or equal to α (see (4) in Section 3.1, for its formal definition). The most common way of assessing equivalence is to use the *Interval Inclusion Principle* (IIP) and check whether the 100(1 - 2α)% Confidence Interval (CI) for the target parameter falls within the equivalence margins $(-c, c)$ (see e.g., Muñoz et al., 2016; Pallmann and Jaki, 2017). This strategy has been shown to lead to the same test decision as the TOST procedure if the CI is equi-tailed (Hsu et al., 1994; Berger and Hsu, 1996).

However, it is well known that this procedure can be conservative as the size of the TOST can be considerably lower than the (specified) significance level α. This induces a drop in power and therefore to a lower probability of detecting truly equivalent mean effects as such. This problem is particularly noticeable in cases where σ_ν is relatively large. Such situations may occur, for example, when the sample size is determined using an underestimated standard deviation value obtained from a prior experiment, or with studies involving highly variable drugs and in which the sample size that would be needed to achieve reasonable values for σ_ν is unrealistic. For that purpose, Anderson and Hauck (1983) proposed a test that has greater power than the TOST for situations where σ_ν is relatively large. This test, referred to here as the AH-test, can be liberal and therefore doesn't control the size (see Berger and Hsu, 1996). In some cases, it can also lead to the acceptance of equivalence when θ, the target parameter of interest, falls outside the equivalence interval (see Schuirmann, 1987). Brown et al. (1997) constructed an unbiased test that is uniformly more powerful than the TOST, however, it is computationally intensive and its rejection region may exhibit rather irregular shapes in some cases (see e.g., Berger and Hsu, 1996). Berger and Hsu (1996) therefore proposed a smoothed version. These tests cannot be assessed using the IIP and the last two are difficult to interpret due to the use of polar coordinates (see e.g., Liu and Chow, 1996).

Due to the overly conservative TOST's size with highly variable drugs, i.e., for cases with relatively large σ_ν, regulatory authorities have recommended an alternative approach based on the linear scaling of the bioequivalence limits according to the value of the standard deviation within the reference group, called Scaled Average BioEquivalence (SABE), with the constraint that $\widehat{\theta}$ lies within the bioequivalence margins $(-c, c)$. These recommendations were issued by the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) (see e.g., Muñoz et al., 2016; Davit et al., 2012, for more details). The amount of expansion is limited by the authorities, and several recent publications have shown that the size of the SABE can be larger than the significance level α (see e.g., Wonnemann et al., 2015; Muñoz et al., 2016; Endrenyi and Tothfalusi, 2019; Molins et al., 2021; Schütz et al., 2022, and the references therein) and therefore have proposed different ways to correct for it (see also Labes and Schütz, 2016; Tothfalusi and Endrenyi, 2016; Ocaña and Muñoz, 2019; Deng and Zhou, 2020). These corrections ensure that the size is smaller than or equal to α and lead to acceptance regions that change more smoothly with σ_ν.

In this paper, as an alternative to previous methods, we propose a finite sample correction of the TOST procedure that simply consists in a correction of the TOST's significance level to guarantee a size-α test when σ_ν is known.

The corrected significance level α^* is straightforward to compute and allows to define 100(1 - 2α^*)% CIs used in the classical TOST. Hence, the α-TOST essentially corresponds to a finite sample continuous variability correction of the TOST procedure, that leads to an increased probability of accepting equivalence when it is true for large values of σ_ν while maintaining a size of exactly α. Indeed, the α-TOST is shown to be uniformly more powerful than the TOST and, for small to moderate values of σ_ν, to be nearly equivalent to the TOST with a comparable power as $\alpha^* \approx \alpha$ in such cases. This method also provides a more powerful test than the corrected SABE. Finally, since in practice σ_ν needs to be estimated from the data, a straightforward estimator for α^* is also proposed and it is shown, through simulations, that the estimator remains level-α and its size stays close to α.

The paper is organized as follows. In Section 2, we consider a case study for which we apply the TOST and the α-TOST, as well as other available methods, in order to showcase the advantage of our proposed approach. The α-TOST is presented in Section 3 from a suitable formulation of the TOST. Its statistical properties as well as a simple algorithm to compute α^* are also provided. In particular, we provide a power comparison with alternative available methods. In Section 4, a simulation study is used to compare the empirical performances of the α-TOST, of the TOST and of the standard and corrected implementations of the SABE. Finally, Section 5 discusses some potential extensions.

2. Evaluation of Bioequivalence for Econazole Nitrate Deposition in Porcine Skin

Quartier et al. (2019) studied the cutaneous bioequivalence of two topical cream products: a reference medicinal product (RMP) and an approved generic containing econazole nitrate (ECZ), an antifungal medication used to treat skin infections. The evaluation of the putative bioequivalence is based on the determination of the cutaneous biodistribution profile of ECZ observed after application of the RMP and the generic product. The dataset we analyse in this section consists in 17 pairs of comparable porcine skin samples on which measurements of ECZ deposition were collected using both creams. The data are presented in Figure 1. In order to assess bioequivalence of both topical treatments, the TOST and α-TOST procedures, based on a paired t-test statistic considering ECZ levels on the logarithmic scale, are conducted using $c = \theta_U = -\theta_L = \log(1.25) \approx 0.223$. These bioequivalence limits correspond to the recommendation of regulatory agencies such as the Food and Drugs Administration (2001) and the European Medicine Agency (2010). Figure 2 shows the CIs corresponding to both approaches. The $100(1-2\alpha)\%$ TOST confidence interval for the mean of the paired differences in ECZ levels equals $[-0.204, 0.250]$, given that $\widehat{\theta} = 0.023$, $\widehat{\sigma}_\nu = 0.130$, $\nu = 16$ and $\alpha = 5\%$. As its upper bound exceeds the upper bioequivalence limit, the classical TOST procedure does not allow one to conclude that the topical products are (on average) equivalent. To reach a size of 5%, the α-TOST procedure uses in this case a significance level of $\widehat{\alpha}^* = 7.48\%$ leading to a confidence interval of $[-0.166, 0.211]$. This CI being strictly embedded within the $(-c, c)$ bioequivalence limits, the α-TOST procedure allows to accept bioequivalence, hence illustrating the increase in power induced by the increased significance level considered to reach a size of 5%. Note that in this case, given $\widehat{\sigma}_\nu^2$ and ν, the empirical power of the TOST is zero (regardless of $\widehat{\theta}$) as $t_{0.05,16}\ \widehat{\sigma}_\nu^2 > c$, where $t_{\alpha,\nu}$ denotes the upper quantile α of a t-distribution with ν degrees of freedom; see Section 3.3 for a detailed discussion. Since, as will be shown in Section 3.2, the α-TOST guarantees a size of α (for all sample sizes), the conclusion brought in by the α-TOST is more trustworthy.

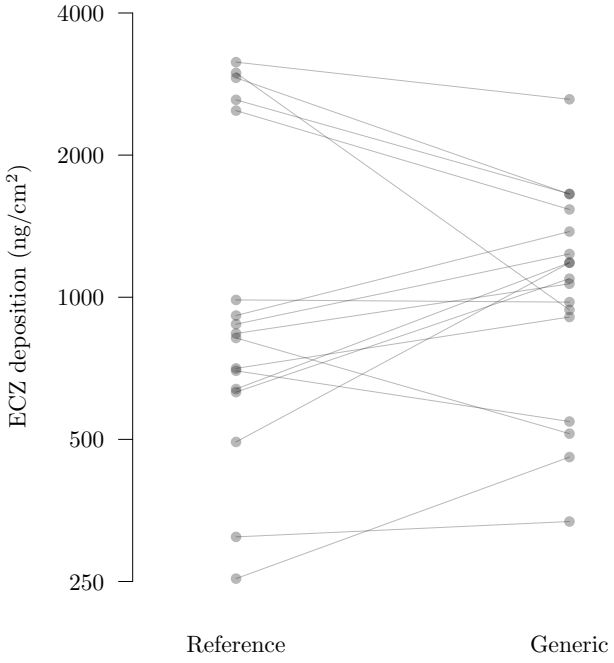

Figure 1: Econazole nitrate deposition levels (y-axis) measured using the reference and generic creams (x-axis) on 17 pairs of comparable porcine skin samples (lines).

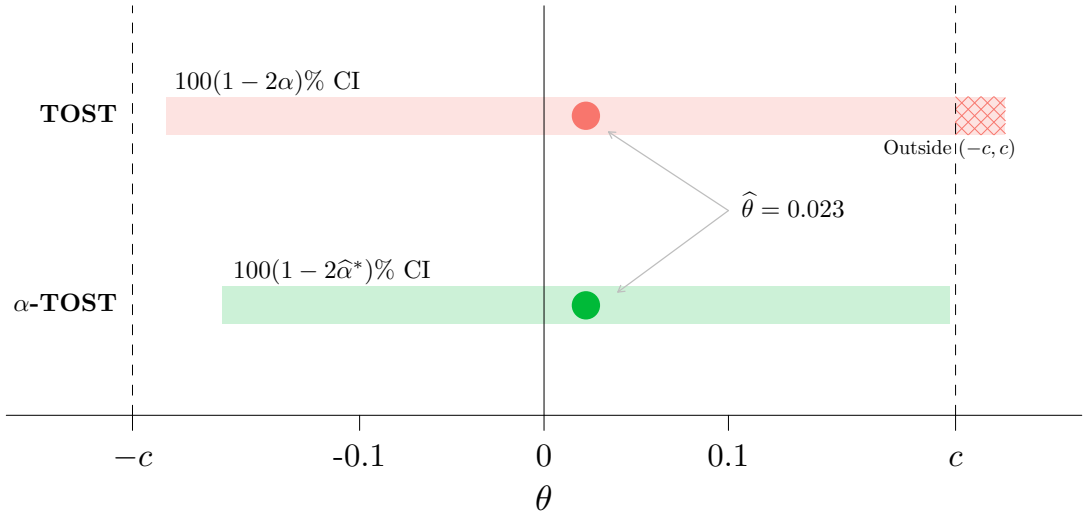

Figure 2: $100(1-2\alpha)\%$ and $100(1-2\widehat{\alpha}^*)\%$ confidence intervals of the TOST and α-TOST procedures for the mean of the paired *log* differences in ECZ levels obtained with the reference and generic creams with $\alpha = 5\%$ and $\widehat{\alpha}^* = 7.48\%$. The dashed vertical lines correspond to the used lower and upper bioequivalence limits with $c = \log(1.25)$. Comparison of the CI of each approach to the bioequivalence limits leads to the acceptance of bioequivalence for the α-TOST procedure and not for the classic TOST approach due to its CI upper limit exceeding c (hatched area).

Method	IIP	Level-α	Size-α	Decision
AH-test	no	no	no	Accept
SABE	yes	no	no	Accept
SABE corrected	yes	yes*	no	Reject
TOST	yes	yes	no	Reject
α-TOST	yes	yes*	yes*	Accept

Table 1: Bioequivalence acceptance/rejection (Decision) for the econazole nitrate deposition in porcine skin data using the AH-test, EMA implementation of the SABE (SABE), level-corrected EMA implementation of the SABE (SABE corrected) of Labes and Schütz (2016), TOST and α-TOST. The estimated parameter values are $\widehat{\sigma}_\nu = 0.130$, $\nu = 16$, $\widehat{\theta} = 0.0227$ and $\alpha = 5\%$. The columns IIP, Level-α and Size-α, respectively indicate if each method satisfies the IIP, if its size is bounded by α and if its size is exactly α. The symbol * specifies that the property is valid when the standard error σ_ν is known.

To gain additional insight into the benefits conferred by our approach, we also compare the characteristics and conclusion of the α-TOST to other available methods in Table 1 as well as their rejection region as a function of $\widehat{\theta}$ and $\widehat{\sigma}_\nu$ in Figure 3. We considered here the AH-test, the EMA implementation of the SABE, and its level-corrected version provided by Labes and Schütz (2016) (SABE corrected), together with the TOST and α-TOST. The AH-test does not satisfy the IIP, but represents a good proxy for the other tests without this property and is relatively easy to implement. The standard SABE, which is not level-α and therefore untrustworthy, is mainly considered here for comparison with its corrected version. Note that, among the level-α tests, the α-TOST is the only one leading to the acceptance decision.

Figure 3 shows the combinations of values of $\widehat{\theta}$ and $\widehat{\sigma}_\nu$ leading to accepting bioequivalence in the setting of the porcine skin dataset, i.e., with $c = \log(1.25)$ and $\nu = 16$. The rejection regions of the different methods almost perfectly overlap for values of $\widehat{\sigma}_\nu$ below 0.09 and differ for larger values. Regardless of $\widehat{\sigma}_\nu$, the TOST and EMA corrected version of the SABE cannot accept bioequivalence for large values of $\widehat{\sigma}_\nu$ (greater than approximately 0.15 here) while the α-TOST and AH-test can, with the rejection region of the α-TOST embedded in the too liberal one of the AH-test. The SABE method shows an atypical rejection area leading to paradoxical cases where, for a given value of $\widehat{\theta}$ and two (reasonable) values of $\widehat{\sigma}_\nu$, $\widehat{\sigma}_\nu^{(a)}$ and $\widehat{\sigma}_\nu^{(b)}$, with $\widehat{\sigma}_\nu^{(a)} > \widehat{\sigma}_\nu^{(b)}$, bioequivalence is accepted with $\widehat{\sigma}_\nu^{(a)}$ but not with $\widehat{\sigma}_\nu^{(b)}$. Such irregular rejection regions were also noted for other tests briefly mentioned above (see e.g., the Figures 1 of Berger and Hsu, 1996, Brown et al., 1997 and Cao and Mathew, 2008). In Figure 3, the symbol $\times$ shows the values of $\widehat{\theta}$ and $\widehat{\sigma}_\nu$ obtained in the porcine skin dataset. These coordinates lead to the acceptance of (average) bioequivalence of the two topical products with the standard SABE, AH-test and α-TOST procedures among which only the α-TOST is size-α.

A more comprehensive description of the power of the selected methods is provided in Section 3.3, and a simulation study comparing the empirical performances is given in Section 4.

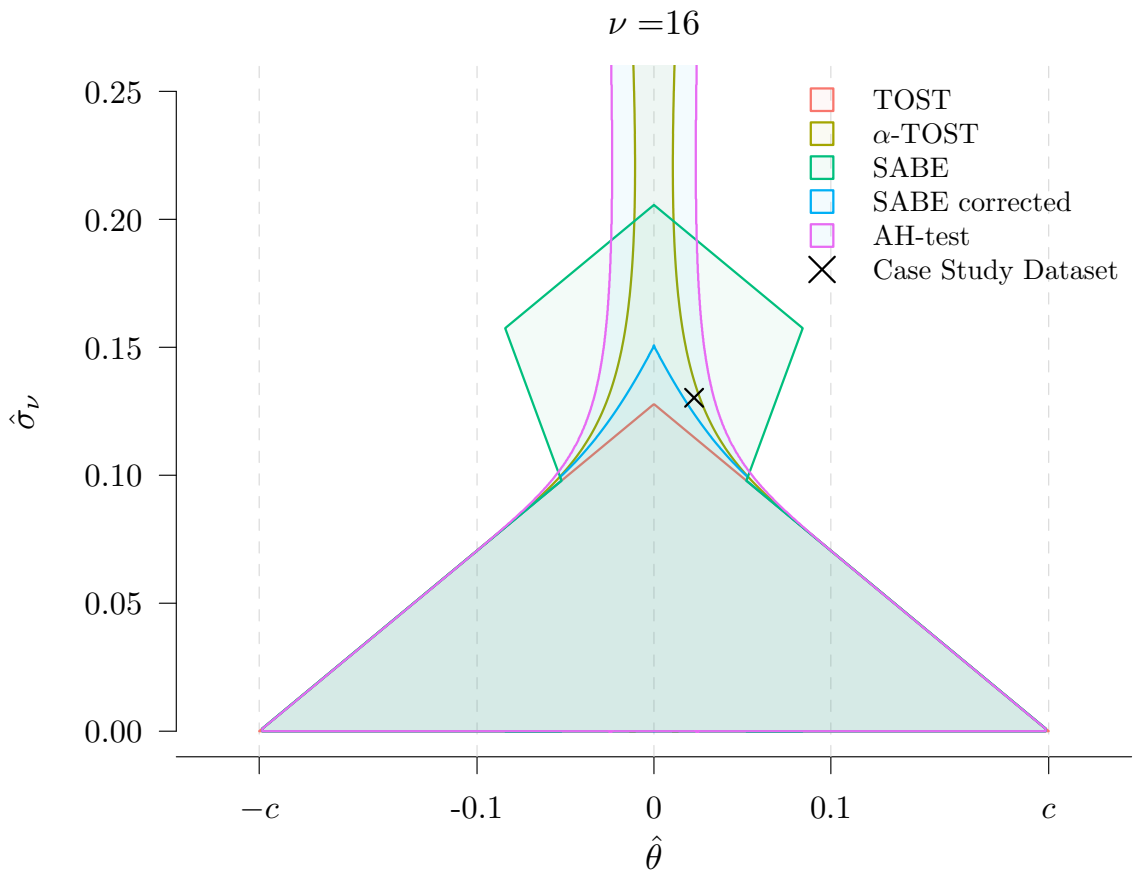

Figure 3: Bioequivalence test rejection regions as a function of $\widehat{\theta}$ (x-axis) and $\widehat{\sigma}_\nu$ (y-axis) per method considered in Table 1 (coloured areas) showing combinations of values of $\widehat{\theta}$ and $\widehat{\sigma}_\nu$ leading to accepting equivalence in the setting of the porcine skin dataset, i.e., with $c = \log(1.25)$ and $\nu = 16$. The rejection regions of the different methods almost perfectly overlap for values of $\widehat{\sigma}_\nu$ below 0.09 and differ for larger values. Regardless of $\widehat{\sigma}_\nu$, the TOST and the corrected SABE can't accept bioequivalence for large values of $\widehat{\sigma}_\nu$ (greater than approximately 0.15 here) while the α-TOST and AH-test can, with the rejection region of the α-TOST embedded in the too liberal one of the AH-test. The SABE method has an atypical rejection area. The symbol $\times$ represents the analysed data set in the acceptance/rejection regions where $\widehat{\theta} = 0.023$ and $\widehat{\sigma}_\nu = 0.130$.

3. Equivalence Testing

In this section, we present the methodology for deriving a corrected statistical equivalence test. We first present the TOST and its properties. We then define the α-TOST procedure through a natural correction of the TOST, derive its statistical properties, propose an iterative procedure to compute the corrected level α^* and show that this procedure is exponentially fast. We also show that the α-TOST is uniformly more powerful than the TOST and the corrected version of the SABE.

3.1. The TOST Procedure

For testing the hypotheses in (2), the TOST uses the two following test statistics:

$$Z_L := \frac{\widehat{\theta} + c}{\widehat{\sigma}_\nu} \quad \text{and} \quad Z_U := \frac{\widehat{\theta} - c}{\widehat{\sigma}_\nu},$$

where Z_L tests for $\mathrm{H}_{01} : \theta \leqslant -c$ versus $\mathrm{H}_{11} : \theta > -c$, and Z_U tests for $\mathrm{H}_{02} : \theta \geqslant c$ versus $\mathrm{H}_{12} : \theta < c$. At a significance level α, the TOST therefore rejects $\mathrm{H}_0 := \mathrm{H}_{01} \cup \mathrm{H}_{02}$ (i.e., $\theta \notin \Theta_1$) in favour of $\mathrm{H}_1 := \mathrm{H}_{11} \cap \mathrm{H}_{12}$ (i.e., $\theta \in \Theta_1$) if both tests simultaneously reject their marginal null hypotheses, that is, if

$$Z_L \geqslant t_{\alpha,\nu} \quad \text{and} \quad Z_U \leqslant -t_{\alpha,\nu},$$

where $t_{\alpha,\nu}$ denotes the upper α quantile of a t-distribution with ν degrees of freedom. The corresponding rejection region of the TOST is given by

$$C_1 := \left\{ \widehat{\theta} \in \mathbb{R},\, \widehat{\sigma}_\nu \in \mathbb{R}_+ \,\middle|\, c \geqslant \left|\widehat{\theta}\right| + t_{\alpha,\nu}\widehat{\sigma}_\nu \right\}. \tag{3}$$

Consequently and as noted in Figure 3, equivalence cannot be accepted with the TOST for all $\widehat{\sigma}_\nu > \widehat{\sigma}_{\max} := c/t_{\alpha,\nu}$, even for $\widehat{\theta} = 0$.

With respect to the TOST's size, given α, θ, σ_ν, ν and c, the probability of accepting equivalence can be expressed as follows (see also Phillips, 1990):

$$\begin{aligned} p(\alpha, \theta, \sigma_\nu, \nu, c) :=& \Pr\left(Z_L \geqslant t_{\alpha,\nu} \text{ and } Z_U \leqslant -t_{\alpha,\nu}, \mid \alpha, \theta, \sigma_\nu, \nu, c\right) \\ =& Q_\nu\left(-t_{\alpha,\nu},\, \frac{\theta - c}{\sigma_\nu},\, \lambda\right) - Q_\nu\left(t_{\alpha,\nu},\, \frac{\theta + c}{\sigma_\nu},\, \lambda\right), \end{aligned} \tag{4}$$

where $\lambda := \frac{c\sqrt{\nu}}{\sigma_\nu t_{\alpha,\nu}}$ and $Q_\nu(t,\, y,\, z)$ corresponds to a special case of Owen's Q-function (Owen, 1965) defined as

$$Q_\nu(t,\, y,\, z) := \frac{\sqrt{2\pi}}{\Gamma\left(\frac{1}{2}\nu\right) 2^{\frac{1}{2}(\nu-2)}} \int_0^z \Phi\left(\frac{tx}{\sqrt{\nu}} - y\right) x^{\nu-1} \varphi(x) dx,$$

where $\varphi(x)$ and $\Phi(x)$ denote the probability and cumulative distribution functions of a standard normal distribution, respectively.

Then, for given values of α, σ_ν and ν, the TOST's size is defined as the supremum of (4) (see e.g., Lehmann, 1986),

and is given by

$$\omega(\alpha, c, \sigma_\nu, \nu) := \sup_{\theta \notin \Theta_1} p(\alpha, \theta, \sigma_\nu, \nu, c) = p(\alpha, c, \sigma_\nu, \nu, c) = Q_\nu\left(-t_{\alpha,\nu},\, 0,\, \lambda\right) - Q_\nu\left(t_{\alpha,\nu},\, \frac{2c}{\sigma_\nu},\, \lambda\right). \tag{5}$$

We can then deduce that the TOST is level-α, by noting that, for $\sigma_\nu > 0$, we have

$$\begin{aligned} \omega(\alpha, c, \sigma_\nu, \nu) < Q_\nu\left(-t_{\alpha,\nu},\, 0,\, \lambda\right) &= \frac{\sqrt{2\pi}}{\Gamma\left(\frac{1}{2}\nu\right) 2^{\frac{1}{2}(\nu-2)}} \int_0^\lambda G\left(-\frac{t_{\alpha,\nu} x}{\sqrt{\nu}}\right) x^{\nu-1} G'(x) dx \\ &< \frac{\sqrt{2\pi}}{\Gamma\left(\frac{1}{2}\nu\right) 2^{\frac{1}{2}(\nu-2)}} \int_0^\infty G\left(-\frac{t_{\alpha,\nu} x}{\sqrt{\nu}}\right) x^{\nu-1} G'(x) dx = \Pr\left(T_\nu \leqslant -t_{\alpha,\nu}\right) = \alpha, \end{aligned} \tag{6}$$

where T_ν denotes a random variable following a t-distribution with ν degrees of freedom, so that

$$\lim_{\sigma_\nu \to 0} \omega(\alpha, c, \sigma_\nu, \nu) = \alpha.$$

Thus, while the TOST is indeed level-α, it actually never achieves a size of α, except in the theoretical case of $\sigma_\nu = 0$, as already highlighted by several authors (see e.g., Deng and Zhou, 2020, and the references therein). When $\widehat{\sigma}_\nu$ is small, the difference between the size and α is marginal, but as $\widehat{\sigma}_\nu$ approaches or exceeds $c/t_{\alpha,\nu}$, this difference increases, leading to a high probability of the TOST failing to detect equivalence when it exists. As a solution to this issue, we suggest an alternative approach, the α-TOST, that corrects the size of the TOST for a large range of values of $\widehat{\sigma}_\nu$ and still allows to assess equivalence by means of confidence intervals, as depicted in Figure 2.

3.2. The α-TOST

A corrected version of the TOST can theoretically be constructed by adjusting the significance level and using α^* instead of α in the standard TOST procedure, where

$$\alpha^* := \alpha^*(\sigma_\nu) = \underset{\gamma \in [\alpha, 0.5)}{\text{argzero}} \left[\omega(\gamma, c, \sigma_\nu, \nu) - \alpha\right], \tag{7}$$

with $\omega(\gamma, c, \sigma_\nu, \nu)$ defined in (5). The dependence of α^* on α and ν is omitted from the notation as these quantities are known. A similar type of correction was also used to amend the significance level of the SABE procedure by Labes and Schütz (2016) and Ocaña and Muñoz (2019) (see also Palmes et al., 2022, for power adjustment). However, in these cases, the corrected significance level was reduced (instead of increased like in (7)) so that the size does not exceed the significance level of α. The aim of these corrections is therefore not the same as the one proposed here. Furthermore, the size of the α-TOST is guaranteed to be exactly α when σ_ν is known, which is not the case for these competing methods.

In Appendix A, we demonstrate that the existence of α^* relies on a simple condition that is satisfied in most settings of practical importance. In particular, this requirement can be translated into a maximal value for the estimated standard error $\widehat{\sigma}_\nu$, that is $\widehat{\sigma}_\nu < \frac{2c}{\Phi^{-1}(\alpha+0.5)}$. As explained in Section 3.3, this bound allows the α-TOST to have a strictly positive power for larger values of σ_ν, compared to other methods satisfying the IIP.

Since, $\alpha^*(\sigma_\nu)$ is a population size quantity as it depends on the unknown quantity σ_ν, a natural estimator for its sample value is given by

$$\widehat{\alpha}^* := \alpha^*(\widehat{\sigma}_\nu) = \underset{\gamma\in[\alpha,0.5)}{\text{argzero}} \left[\omega(\gamma, c, \widehat{\sigma}_\nu, \nu) - \alpha\right]. \tag{8}$$

Hence, in practice, based on the (estimated) corrected significance level $\widehat{\alpha}^*$, the α-TOST procedure rejects the non-equivalence null hypothesis in favour of the equivalence one at the significance level α, if $Z_L > t_{1-\widehat{\alpha}^*,\nu}$ and $Z_U < -t_{1-\widehat{\alpha}^*,\nu}$. In Section 4, we consider a Monte Carlo simulation to compare the empirical performances of different methods when σ_ν needs to be estimated, and show that the α-TOST outperforms other available methods.

The corrected significance level $\widehat{\alpha}^*$ can easily be computed using the following iterative approach. At iteration k, with $k \in \mathbb{N}$, we define

$$\widehat{\alpha}^{*(k+1)} = \alpha + \widehat{\alpha}^{*(k)} - \omega\left(\widehat{\alpha}^{*(k)}, c, \widehat{\sigma}_\nu, \nu\right), \tag{9}$$

with $\omega(\alpha, c, \sigma_\nu, \nu)$ given in (5) and where the procedure is initialized at $\widehat{\alpha}^{*(0)} = \alpha$. This simple iterative approach converges exponentially fast to $\widehat{\alpha}^*$ as it can be shown that

$$\left|\widehat{\alpha}^{*(k+1)} - \widehat{\alpha}^*\right| < \frac{1}{2}\exp(-bk),$$

for some positive constant b (see Appendix B for more details).

3.3. Power Analysis

Since the conclusion of α-TOST considers an interval computed using a smaller value than $t_{\alpha,\nu}$ compared to the TOST, the α-TOST rejection interval is necessarily larger than its TOST counterpart as $\widehat{\alpha}^* > \alpha$. This implies that the α-TOST is uniformly more powerful than the TOST, and explains cases like the one encountered in the porcine skin case study presented in Section 2, in which equivalence is accepted using the α-TOST but not with the TOST (which has an empirical power of zero given $\widehat{\sigma}_\nu$).

Figure 4 compares the rejection regions of the TOST, α-TOST and corrected EMA implementation of the SABE of Labes and Schütz (2016) for different values of ν ranging from $\nu = 12$ to $\nu = 36$. The corrected SABE was selected here as it respects the IIP and as it is level-α. The corrected SABE using the FDA variability adjustment of the equivalence margins (see e.g., Muñoz et al., 2016) would lead to the same conclusions. We can see that, for all values of ν and $\widehat{\sigma}_\nu$ considered here, the rejection regions of the TOST and corrected SABE are embedded in the α-TOST one, meaning that our approach is uniformly more powerful than both methods. The comparison of the rejection regions of the α-TOST and corrected SABE shows that the difference in power decreases with increasing values of ν and that the α-TOST maintains a power that is non zero for larger values of $\widehat{\sigma}_\nu$, namely up to $\frac{2c}{\Phi^{-1}(\alpha+0.5)} \approx 3.55$ (see Appendix A). It should be noted that allowing a testing procedure to accept equivalence with large standard error values might be questionable (see also the discussion in Choi et al., 2008). Ultimately, the determination of an acceptable rand of standard errors goes beyond statistical considerations and requires the input of practitioners and regularity authorities.

Finally, note that the idea of correcting the CI length by adjusting the critical values is not new and has been used,

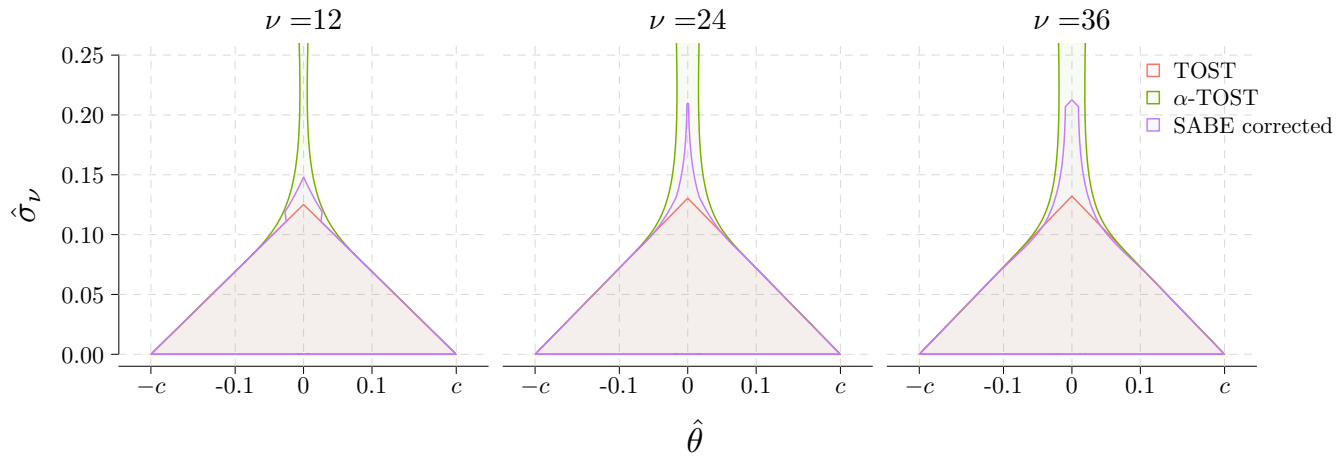

Figure 4: Equivalence test rejection regions of 3 methods, the TOST (red), the α-TOST (green) and the level-corrected EMA implementation of the SABE of Labes and Schütz (2016) (SABE corrected, purple), as a function of $\widehat{\theta}$ (x-axis), $\widehat{\sigma}_\nu$ (y-axis) and ν (panels) with $c = \log(1.25)$ and $\alpha = 5\%$. For all values of ν and $\widehat{\sigma}_\nu$ considered here, the rejection regions of the TOST and corrected SABE are embedded in the α-TOST one, meaning that our approach is more powerful.

for example, by Cao and Mathew (2008) who defined adjustment constants for specific values of $\widehat{\sigma}_\nu$ and used linear interpolation for adjacent values. The lower panel of Figure 7 of Appendix C compares the critical values obtained with the method of Cao and Mathew (2008) to the ones obtained by the α-TOST for different values of $\widehat{\sigma}_\nu$ and ν. We can note that, for all values of ν considered here and for values of $\widehat{\sigma}_\nu$ above 0.1, the critical values of Cao and Mathew (2008) correspond to a piecewise version of the critical values of the α-TOST obtained when ν is large. Therefore, their correction appears to be an approximation of the α-TOST, evaluated asymptotically, i.e., at $\nu \to \infty$.

4. Simulation Study

The analyses shown in Section 3.3 were performed using the true (and unknown) values for σ_ν and θ. In practice, however, these values must be estimated. In this section, we conduct a Monte Carlo simulation study to assess the empirical performances of different procedures when σ_ν and θ are estimated. We selected the same methods as considered in Section 3.3 to which we added the standard SABE to showcase that this approach does not have good properties. For given values of $\theta, c, \nu, \sigma_\nu$ and α, the simulation process used to evaluate the empirical performances of the α-TOST can be described using the following steps for each Monte Carlo sample $b = 1, \ldots, B$:

1. simulate two values for $\widehat{\theta}$, namely $\widehat{\theta}_{b,0}$ and $\widehat{\theta}_{b,c}$, using (1), with $\theta = 0$ and $\theta = c$ to respectively estimate the power and the size,
2. simulate one value for $t = \frac{\nu\widehat{\sigma}_\nu^2}{\sigma_\nu^2}$ using (1) and set $\widehat{\sigma}_{\nu,b} = \sqrt{t \cdot \sigma_\nu^2/\nu}$,
3. compute $\widehat{\alpha}^*$ using (the algorithm associated to) (7) with $\widehat{\sigma}_{\nu,b}$ instead of σ_ν,
4. compute $t_{\widehat{\alpha}^*,\nu,b}$.

The estimated size is then obtained as

$$\frac{1}{B}\sum_{b=1}^{B}\eta\left(c \geqslant |\widehat{\theta}_{b,c}| + t_{\widehat{\alpha}*,\nu,b}\widehat{\sigma}_{\nu,b}\right)$$

and estimated power as

$$\frac{1}{B}\sum_{b=1}^{B}\eta\left(c \geqslant |\widehat{\theta}_{b,0}| + t_{\widehat{\alpha}*,\nu,b}\widehat{\sigma}_{\nu,b}\right),$$

where $\eta(\cdot)$ is the indicator function with $\eta(A) = 1$ if A is true and zero otherwise. For the other methods, Steps 1 and 2 are the same, and the following ones are adapted to the specific methods. The parameters' values for the Monte Carlo simulation are:

- $c = \log(1.25)$,
- $\nu = 20, 30, 40$,
- $\sigma_\nu = 0.01, 0.012, \ldots, 0.6$,
- $\alpha = 5\%$,
- $B = 10^4$, the number of Monte Carlo simulations per combination of parameters.

Figure 5 shows the Monte Carlo finite sample size estimates as a function of σ_ν and ν for the different methods of interest. We can note that, as expected, the SABE procedure can have an empirical size well above the $\alpha = 5\%$ significance level, while the other three methods successfully guarantee an empirical size of at most 5%. Under all scenarios considered here, the α-TOST is the method allowing to maintain the highest non-liberal empirical size, especially when the number of degrees of freedom is large. Both the TOST and the corrected SABE approaches have an empirical size that reaches zero for relatively small values of σ_ν (around 0.274).

The finite sample correction provided by the α-TOST has a direct consequence on the power, as can be seen in Figure 6 showing the Monte Carlo finite sample power estimates as a function of σ_ν and ν for the different methods of interest. Indeed, for the TOST and corrected SABE, the empirical power is zero when the empirical size is zero, which is achieved at values of σ_ν around 0.274 and above. The seemingly large power of the SABE is a consequence of its empirical size being well above the significance level α. When its empirical size stops being liberal with growing values of σ_ν, its power is dominated by the one of the α-TOST. The empirical performances of the α-TOST remains above the ones of its competitors with non-liberal empirical size under all configurations and an empirical power greater than zero for all values of σ_ν considered here.

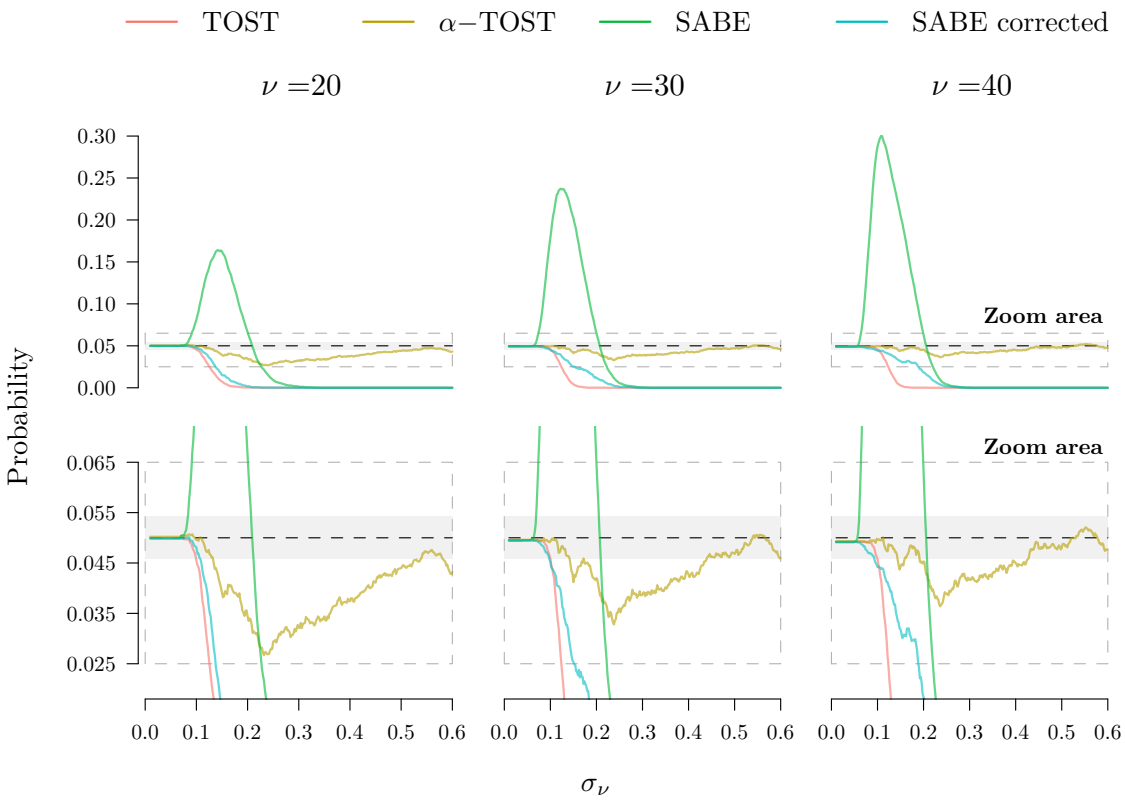

Figure 5: Empirical size (y-axis) computed using the simulation setting of Section 4 as a function of σ_ν (x-axis), and ν (panels) for the different methods of interest (colours). Top panels: complete results. Bottom panels: focus over the range $(0.025; 0.065)$. The shaded gray areas correspond to the pointwise 95% Monte Carlo tolerance band. We can note that the empirical size of the TOST and the corrected SABE approaches tend to be overly conservative for values of $\sigma_\nu > 0.1$, with a slight advantage of the corrected SABE approach compared to the TOST for values of σ_ν below 0.2, especially when ν is large. As expected, the original SABE approach is overly liberal for some range of σ_ν. Under all scenarios considered here, the α-TOST is the method allowing to maintain the highest non-liberal empirical size.

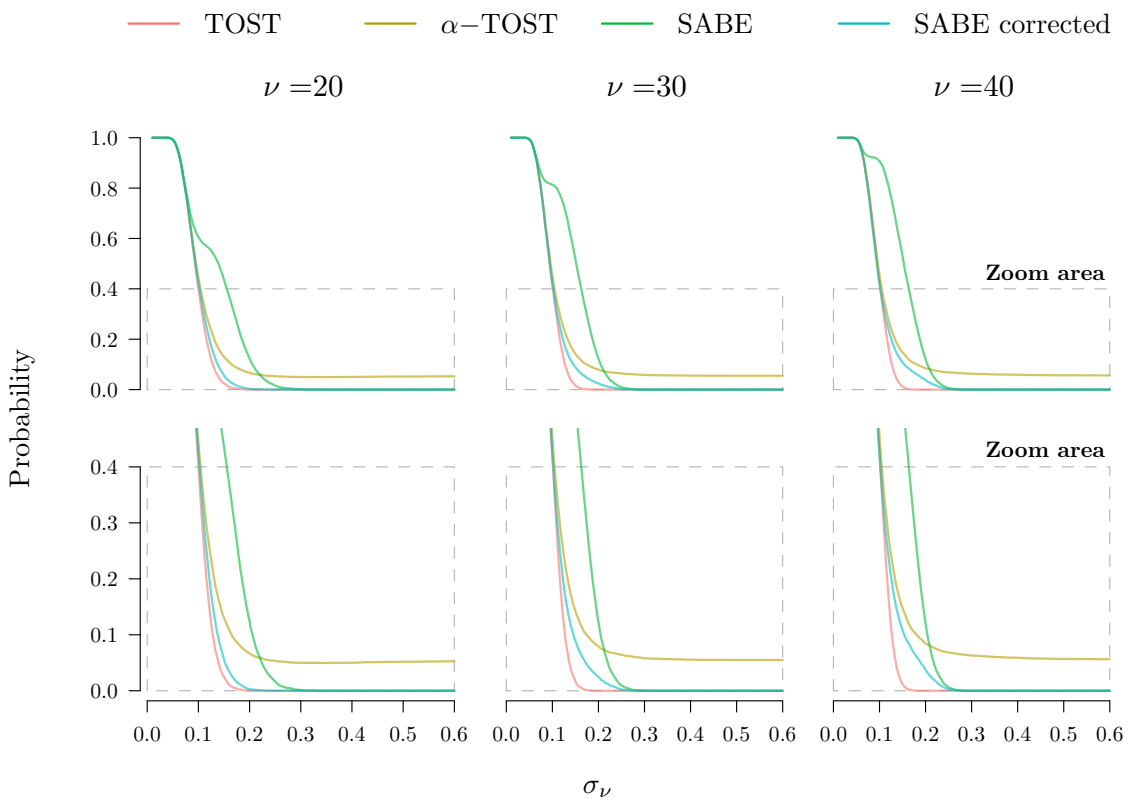

Figure 6: Finite sample power (y-axis) computed using the simulation setting of Section 4 as a function of σ_ν (x-axis), and ν (columns) for the different methods of interest (colours). Top panels: complete results. Bottom panels: focus on values below 0.4. The power of the α-TOST not only remains above the one of its competitors with non-liberal empirical size under all configurations, it is also greater than zero for all values of σ_ν considered here.

5. Discussion

The canonical framework treated in this paper is given in (1) and therefore concerns differences that can be assumed to be normally distributed (in finite samples), with a known finite sample distribution for $\widehat{\sigma}_\nu$. This framework covers a quite large spectrum of data settings, such as the standard two-period crossover experimental design (see e.g., Jones and Kenward, 2014), and could be extended to include covariates to possibly reduce residual variance. Extensions to non-linear cases, such are for example binary outcomes (see e.g., Dunnett and Gent, 1977; Tu, 1998; Schouten and Kester, 2010; Lui and Chang, 2011; Ostrovski, 2022, and the references therein), would follow the same logic, but would require a specific treatment due to the nature of the responses and to the use of link functions. Such extensions also deserve some attention but are left for further research.

For sample size calculations, we could, in principle, proceed with the α-TOST, for given values of c, θ and σ_ν. However, when considering high levels of power, the correction is negligible and we have $\alpha^* \approx \alpha$ as shown in Section 4, so that the sample size can be computed using the TOST, as implemented in standard packages. The α-TOST approach would then be used to assess equivalence and show its benefits when the observed value of σ_ν is unexpectedly large compared to the one considered in the sample size calculation either due to (lack of) chance or to an underestimated value obtained from a prior experiment.

6. Bibliography

Aggarwal, M., J. Allen, A. Coppock, D. Frankowski, S. Messing, K. Zhang, J. Barnes, A. Beasley, H. Hantman, and S. Zheng (2023). A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. Nature Human Behaviour, 1–10.

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

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. Technometrics 24(4), 295–300.

Berger, R. L. and J. C. Hsu (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science 11, 283–319.

Branscheidt, M., N. Ejaz, J. Xu, M. Widmer, M. D. Harran, J. C. Cortés, T. Kitago, P. Celnik, C. Hernandez-Castillo, J. Diedrichsen, A. Luft, and J. W. Krakauer (2022). No evidence for motor-recovery-related cortical connectivity changes after stroke using resting-state fmri. Journal of Neurophysiology 127, 637–650.

Brown, L. D., J. G. Hwang, and A. Munk (1997). An unbiased test for the bioequivalence problem. The Annals of Statistics, 2345–2367.

Cao, L. and T. Mathew (2008). A simple numerical approach towards improving the two one-sided test for average bioequivalence. Biometrical Journal 50, 205–211.

Choi, L., B. Caffo, and C. Rohde (2008). A survey of the likelihood approach to bioequivalence trials. Statistics in Medicine 27, 4874–4894.

Davit, B. M., M. L. Chen, D. P. Conner, S. H. Haidar, S. Kim, C. H. Lee, R. A. Lionberger, F. T. Makhlouf, P. E. Nwakama, D. T. Patel, D. J. Schuirmann, and L. X. Yu (2012). Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the US Food and Drug Administration. American Association of Pharmaceutical Scientists Journal 14, 915–924.

Deng, Y. and X.-H. Zhou (2020). Methods to control the empirical type I error rate in average bioequivalence tests for highly variable drugs. Statistical Methods in Medical Research 29, 1650–1667.

Dunnett, C. W. and M. Gent (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. Biometrics 33, 593–602.

Endrenyi, L. and L. Tothfalusi (2019). Bioequivalence for highly variable drugs: regulatory agreements, disagreements, and harmonization. Journal of Pharmacokinetics Pharmacodynamics 46, 117–126.

European Medicine Agency (2010). Guideline on the investigation of bioequivalence-cpmp. Technical report, EWP/QWP/1401/98 Rev. 1.

Federer, H. (2014). Geometric Measure Theory. Springer.

Feri, F., C. Giannetti, and P. Guarnieri (2023). Risk-taking for others: An experiment on the role of moral discussion. Journal of Behavioral and Experimental Finance 37, 100735.

Food and Drugs Administration (2001). Guidance for industry, statistical approaches to establishing bioequivalence. http://www.fda.gov/cder/guidance/index.htm.

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

Jones, B. and M. G. Kenward (2014). Design and Analysis of Cross-over Trials. CRC press.

Labes, D. and H. Schütz (2016). Inflation of type I error in the evaluation of scaled average bioequivalence, and a method for its control. Pharmaceutical Research 33, 2805–2814.

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. Social Psychological and Personality Science 8, 355–362.

Lakens, D., A. M. Scheel, and P. M. Isager (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science 1, 259–269.

Lehmann, E. L. (1986). Testing Statistical Hypothesis, 2nd edition. New York: Wiley.

Liu, J. P. and S. C. Chow (1996). Bioequivalence trials, intersection-union tests and equivalence confidence set: Comment. Statistical Science 11, 306–312.

Lui, K.-J. and K.-C. Chang (2011). Test non-inferiority (and equivalence) based on the odds ratio under a simple crossover trial. Statistics in Medicine 30, 1230–1242.

Mazzolari, R., S. Porcelli, D. J. Bishop, and D. Lakens (2022). Myths and methodologies: The use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. Experimental Physiology 107, 201–212.

Metzler, C. (1974). Bioavailability – a problem in equivalence. Journal of Pharmaceutical Sciences 30, 309–317.

Meyners, M. (2012). Equivalence tests – a review. Food Quality and Preference 26, 231–245.

Molins, E., D. Labes, H. Schütz, E. Cobo, and J. Ocaña (2021). An iterative method to protect the type I error rate in bioequivalence studies under two-stage adaptive 2×2 crossover designs. Biometrical Journal 63, 122–133.

Muñoz, J., D. Alcaide, and J. Ocaña (2016). Consumer's risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs. Statistics in Medicine 35, 1933–1943.

O'Brien, M. W. and D. S. Kimmerly (2022). Is "not different" enough to conclude similar cardiovascular responses across sexes? American Journal of Physiology-Heart and Circulatory Physiology 322, H355–H358.

Ocaña, J. and J. Muñoz (2019). Controlling type I error in the reference-scaled bioequivalence evaluation of highly variable drugs. Pharmaceutical Statistics 18, 583–599.

Ostrovski, V. (2022). Testing equivalence to binary generalized linear models with application to logistic regression. Statistics & Probability Letters 191, 109658.

Owen, D. B. (1965). A special case of a bivariate non-central t-distribution. Biometrika 52, 437–446.

Pallmann, P. and T. Jaki (2017). Simultaneous confidence regions for multivariate bioequivalence. Statistics in Medicine 36(29), 4585–4603.

Palmes, C., T. Bluhmki, B. Funke, and E. Bluhmki (2022). Asymptotic properties of the two one-sided t-tests - new insights and the schuirmann-constant. The International Journal of Biostatistics 18, 19–38.

Patterson, S. and B. Jones (2006). Bioequivalence and Statistics in Clinical Pharmacology. Boca Raton, FL: Chapman & Hall/CRC.

Phillips, K. (1990). Power of the two one-sided tests procedure in bioequivalence. Journal of Pharmacokinetics and Biopharmaceutics 18, 137–144.

Quartier, J., N. Capony, M. Lapteva, and Y. N. Kalia (2019). Cutaneous biodistribution: A high-resolution methodology to assess bioequivalence in topical skin delivery. Pharmaceutics 11, 484.

Sansone, P., L. G. Giaccari, C. Aurilio, F. Coppolino, M. B. Passavanti, V. Pota, and M. C. Pace (2022). Comparative efficacy of Tapentadol versus Tapentadol plus Duloxetine in patients with chemotherapy-induced peripheral neuropathy. Cancers 14, 4002.

Schouten, H. and A. Kester (2010). A simple analysis of a simple crossover trial with a dichotomous outcome measure. Statistics in Medicine 29, 193–198.

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.

Schütz, H., D. Labes, and M. J. Wolfsegger (2022). Critical remarks on reference-scaled average bioequivalence. Journal of Pharmacy & Pharmaceutical Sciences 25, 285–296.

Senn, S. (2021). Statistical Issues in Drug Development, 3rd edition. Wiley.

Sureshkumar, H., R. Xu, N. Erukulla, A. Wadhwa, and L. Zhao (2022). "snap on" or not? a validation on the measurement tool in a virtual reality application. Journal of Digital Imaging 35(3), 692–703.

Tothfalusi, L. and L. Endrenyi (2016). An exact procedure for the evaluation of reference-scaled average bioequivalence. American Association of Pharmaceutical Scientists Journal 18, 476–489.

Tu, D. (1998). On the use of the ratio or the odds ratio of cure rates in establishing therapeutic equivalence of non-systemic drugs with binary clinical endpoints. Journal of Biopharmaceutical Statistics 8, 263–282.

Wang, K., Y. Li, B. Chen, H. Chen, D. E. Smith, D. Sun, M. R. Feng, and G. L. Amidon (2022). In vitro predictive dissolution test should be developed and recommended as a bioequivalence standard for the immediate-release solid oral dosage forms of the highly variable mycophenolate mofetil. Molecular Pharmaceutics 19, 2048–2060.

Wehrle, F. M., T. Bartal, M. Adams, D. Bassler, C. F. Hagmann, O. Kretschmar, G. Natalucci, and B. Latal (2022). Similarities and differences in the neurodevelopmental outcome of children with congenital heart disease and children born very preterm at school entry. The Journal of Pediatrics 250, 29–37.e1.

Westlake, T. (1976). Symmetrical confidence intervals for bioequivalence trials. Biometrics 32, 741–744.

Wonnemann, M., C. Frömke, and A. Koch (2015). Inflation of the type I error: Investigations on regulatory recommendations for bioequivalence of highly variable drugs. Pharmaceutical Research 32, 135–143.

Appendix

A. Existence of α^*

In this section, we state the conditions for α^*, defined in (7), to be a singleton. Fixing α, c, σ_ν, and ν, we simplify our notation so that $\omega(\gamma) := \omega(\gamma, c, \sigma_\nu, \nu)$ and let

$$\mathcal{A} := \left\{ x \in [\alpha,\, 0.5) \,\middle|\, \omega(x) > 0 \right\}.$$

The function $\omega(\gamma)$, defined in (5), is continuously differentiable and strictly increasing in γ in $\mathcal{A}$. From (6), we have that $\alpha \geqslant \omega(\alpha)$. Thus, it is sufficient to show that $\alpha < \alpha_{\max} := \lim_{\alpha \to 0.5^-} \omega(\alpha)$, where $\alpha \to 0.5^-$ denotes the limit from below 0.5, to ensure that α^* is a singleton. Let $T_{\nu,\delta}$ denote a random variable following a non-central t-distribution with ν degrees of freedom and noncentrality parameter $\delta = 2c/\sigma_\nu$. Then, we have

$$\begin{aligned}\alpha_{\max} &= \lim_{\gamma \to 0.5^-} \omega(\gamma) = \lim_{\gamma \to 0.5^-} \left\{ Q_\nu\left(-t_{\gamma,\nu},\, 0,\, \frac{c\sqrt{\nu}}{\sigma_\nu t_{\gamma,\nu}}\right) - Q_\nu\left(t_{\gamma,\nu},\, \delta,\, \frac{c\sqrt{\nu}}{\sigma_\nu t_{\gamma,\nu}}\right) \right\} \\ &= \Pr\left(T_{\nu,0} \leqslant 0\right) - \Pr\left(T_{\nu,\delta} \leqslant 0\right) = 0.5 - \Pr\left(T_{\nu,\delta} \leqslant 0\right) = 0.5 - \Phi(-\delta) = \Phi(\delta) - 0.5.\end{aligned}$$

Thus, the condition $\alpha < \alpha_{\max}$ can expressed as follows

$$\Phi(\delta) - 0.5 > \alpha \quad \Longleftrightarrow \quad \frac{2c}{\sigma_\nu} > \Phi^{-1}(\alpha + 0.5) \quad \Longleftrightarrow \quad \sigma_\nu < \frac{2c}{\Phi^{-1}(\alpha + 0.5)}.$$

Therefore, the condition $\sigma_\nu < \frac{2c}{\Phi^{-1}(\alpha+0.5)}$ implies that $\alpha < \alpha_{\max}$ and consequently that α^* is a singleton. The existence of $\widehat{\alpha}^*$ follows the same argument but replacing σ_ν by $\widehat{\sigma}_\nu$ in our condition.

B. Convergence Rate of the Iterative Approach for α^*

Using the notation of Appendix A and for $\gamma \in \mathcal{A}$, we have that $\omega(\gamma)$ is continuously differentiable and such that $0 < \dot{\omega}(\gamma) < 2$, where

$$\dot{\omega}(\gamma) := \left.\frac{\partial}{\partial x}\omega(x)\right|_{x=\gamma}.$$

Next, we define

$$T(\gamma) := \alpha + \gamma - \omega(\gamma).$$

For all $\alpha_1, \alpha_2 \in \mathcal{A}$, we have the mean value theorem stating that

$$T(\alpha_1) - T(\alpha_2) = \alpha_1 - \alpha_2 - \omega(\alpha_1) + \omega(\alpha_2) = \alpha_1 - \alpha_2 + \dot{\omega}(\alpha^*)(\alpha_2 - \alpha_1),$$

where $\alpha^* = \tau\alpha_1 + (1-\tau)\alpha_2$ for some $\tau \in [0,1]$. Thus, we obtain

$$\left|T(\alpha_1) - T(\alpha_2)\right| = \left|\left\{1 - \dot{\omega}(\alpha^*)\right\}(\alpha_1 - \alpha_2)\right| = \left|1 - \dot{\omega}(\alpha^*)\right| \left|\alpha_1 - \alpha_2\right| < \left|\alpha_1 - \alpha_2\right|.$$

Then, using Kirszbraun theorem (see Federer, 2014), we can extend the function $T(\gamma)$ with respect to $\gamma \in \mathcal{A}$ to a contraction map from $\mathbb{R}$ to $\mathbb{R}$. Thus, Banach fixed point theorem ensures that $T\left(\alpha^{*(k)}\right)$ converges as $k \to \infty$. We then define the limit of the sequence $\left(\alpha^{*(k+1)}\right)_{k\in\mathbb{N}}$ as α^*, which is the unique fixed point of the function $T(\gamma)$. Indeed, we have

$$\alpha^* = T\left(\alpha^*\right) = \alpha + \alpha^* - \omega\left(\alpha^*\right).$$

By rearranging terms, we have

$$\alpha^* = \underset{\gamma\in\mathcal{A}}{\text{argzero}}\ \omega\left(\gamma\right) - \alpha = \underset{\gamma\in[\alpha,0.5)}{\text{argzero}}\ \omega\left(\gamma\right) - \alpha,$$

concluding the convergence of the sequence $\left(\alpha^{*(k+1)}\right)_{k\in\mathbb{N}}$. As a result, there exists some $0 < \epsilon < 1$ such that for $k \in \mathbb{N}$ we have

$$\left|\alpha^{*(k+1)} - \alpha^*\right| < \epsilon^k \left|\alpha^* - \alpha\right| < \frac{1}{2}\exp(-bk),$$

for some positive constant b.

C. Comparison of the α-TOST with Cao and Mathew (2008) Method

In Figure 7, the critical values for different values of ν, obtained by Cao and Mathew (2008) (Table 1) and the ones obtained using the α-TOST, are compared. One can see that the method of Cao and Mathew (2008) appears to be an approximation of the α-TOST, evaluated asymptotically, i.e., at $\nu \to \infty$.

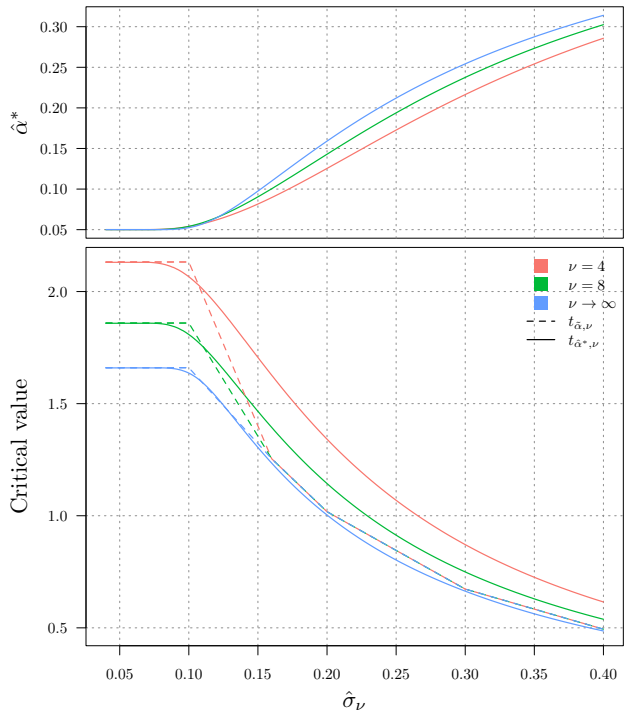

Figure 7: Upper panel: values of $\widehat{\alpha}^*$ of the α-TOST (y-axis) as a function of $\widehat{\sigma}_\nu$ (x-axis) for different values of ν (coloured lines). Lower panel: comparison of the critical values (y-axis) obtained by the method of Cao and Mathew (2008)(dashed lines showing $t_{\tilde{\alpha},\nu}$) and of the α-TOST (solid lines showing $t_{\widehat{\alpha}*,\nu}$)as a function of $\widehat{\sigma}_\nu$ (x-axis) for different values of ν (coloured lines). Note that for all values of ν considered here and for values of $\widehat{\sigma}_\nu$ above 0.1, the critical values of Cao and Mathew (2008) correspond to a piecewise version of the critical values of the α-TOST obtained when ν is large. Therefore, their correction appears to be an approximation of the α-TOST, evaluated asymptotically, i.e., at $\nu \to \infty$.