COMPARING THE EQUIVALENCE TESTING WITH USING TWO ONE-SIDED TEST, SQUARE ROOT OF F DISTRIBUTION AND 2-DF FOR SHIFT-SCALE-EQUIVALENCE TEST

Puntipa Wanitjirattikal
Department of Statistics
King Mongkut's Institute of Technology Ladkrabang
1 Ladkrabang, Bangkok, Thailand 10520
E-mail: puntipa.wa@kmitl.ac.th

KEYWORDS
Equivalence Testing, Paired t-test, Two One-sided Test.

ABSTRACT

In pharmaceutical and medical studies, we would like to show any formulations or two treatments are equivalent. For example, Westergren ESR and STATplus ESR are two popular measurements of sedimentation rate, which are used to monitor disease severity in patients with rheumatoid arthritis and other inflammatory rheumatologic conditions. Westergren ESR is a well-known measurement that was developed by R. S.Fahraeus and A.V.A. Westergren in 1921, while STATplus ESR is an innovative measurement to accelerate turnaround time. Compared with Westergren ESR, the result from STATplus ESR is easier to understand. Since these two measurements can be used to test the same study, it is necessary to know if they can be switched.

Typically, a new measurement process is compared with an existing measurement process. Paired data of these two measurements occur because they are used on the same subject. Usually, paired t- test is appropriate for paired data, but it does not fit well for some situations because paired t-test can only be used to check significant differences from paired data. If the paired data have positive or negative association, the result from the paired t-test might be the same. For example, one paired dataset has positive correlation, and the other one paired dataset has negative correlation. But paired t- tests give the same conclusion because they have the same differences .Moreover, the paired t- test might have low power for scale-type relationships.

In this paper, we propose a test that has reasonable power for both shift and scale-type relationships, which is based on shift- scale type relationships. We consider an equivalence testing for hypothesis. It is an approach to swap the hypotheses so that statistical equivalence of the two measurements is the alternative hypothesis and bears the burden of proof. We conclude "equivalence" only if there is evidence to support the claim that the magnitude of disagreement between the two measurements lie within specified limits.

LITERATURE REVIEW

In clinical trials and biostatistics studies, the clinical equivalence testing aids us to interpret the equality of two different measurements. Westlake (1972) used the equality to decide that the new drug would be essentially equivalent to the current drug. They analyzed in a crossover trial with two independent drugs, where μ_X and μ_Y are the true population means of the mean total urinary excretion of drug for standard and new formulations respectively, with sample size of $n_1 = n_2 = n$ each subject. Based on normality assumptions, the difference sample has the t distribution with $n-2$ degrees of freedom, where S^2 is the mean squared error. Then the 95% probability inequality of the mean total urinary excretion is $\mu_X + \Delta \le \mu_Y \le \mu_X + \Delta.$, where $\Delta = k_1 S\sqrt{2/n}$ $-\left(\bar{X}-\bar{Y}\right) = -k_2 S\sqrt{2/n} - \left(\bar{X}-\bar{Y}\right)$. This implies that $2\left(\bar{X}-\bar{Y}\right) = \left(k_1+k_2\right)S\sqrt{2/n}$. Therefore, k_1, k_2 can be evaluated from solving two equations. The first equation is the integral of the t distribution from k_2 to k_1, which equals 0.95. The second equation is $2\left(\bar{X}-\bar{Y}\right)$ $= \left(k_1+k_2\right)S\sqrt{2/n}.$ He will conclude that both are equivalent, if the mean total urinary excretion of new drug is within 15% of the standard drug.

Four year later, Westlake (1976) extended his previous paper and found alternative ways to evaluate k_1 and k_2. In the conventional approach, k_1+k_2 = 0 then k_1 = $-k_2$ =$t_{0.975,n-2}$. Under the logarithmic transformed data, his confident interval is $k_2 S\sqrt{2/n} - \left(\bar{X}-\bar{Y}\right) \le \log_{10}\left(\mu_Y / \mu_X\right)$ $\le k_1 S\sqrt{2/n} - \left(\bar{X}-\bar{Y}\right)$

"Bioequivalence trials is the comparative trials to test two different measurements are equivalent in *vivo*," is said by Westlake (1979). The 95% confidence interval on the difference between μ_X and μ_Y is formed as follows inequality $k_2 \le \dfrac{\left(\bar{X}-\bar{Y}\right)-\left(\mu_X-\mu_Y\right)}{S\sqrt{2/n}} \le k_1$, where $\bar{X}$ and $\bar{Y}$ are the mean of the standard measurement and the new measurement. The probability from k_2 to k_1 of t

Communications of the ECMS, Volume 33, Issue 1,
Proceedings, ©ECMS Mauro Iacono, Francesco Palmieri,
Marco Gribaudo, Massimo Ficco (Editors)
ISBN: 978-3-937436-65-4/978-3-937436-66-1(CD) ISSN 2522-2414

distribution with the degrees of freedom $n-2$ equals to 0.95. Conventionally, $k_1 + k_2 = 0$ then $k_1 = -k_2 = t_{0.975,n-2}$.

However, Kirkwood & Westlake (1981) were still rethinking about the bioequivalence. They claimed that "If this approach is used it will be seen that the use of a conventional $1-\alpha$ confidence interval with $\alpha = 0.05$ is unduly conservative since the probability that the interval falls within the $\pm\Delta$ limits, where the difference in means is Δ. It can be shown to be $< \alpha/2$, or 0.025." Then the estimate $\hat{\delta}$ is equivalent to acceptance, where $-\Delta < \delta < \Delta$ exceeds $1-\alpha$. Conventionally, the probability that $\delta < -\Delta$ is less than $\alpha/2$ and the probability that $\delta > \Delta$ is less than $\alpha/2$.

In 1984, Hauck & Anderson claimed that the ANOVA testing was to test the equality of two measurements, but it could not test for equivalence testing, which two measurements differ by less than the specified limits. So, they constructed the equivalence testing by setting the alternative hypothesis as $H_1 : A_0 < \mu_Y/\mu_X < B_0$, where μ_Y and μ_X are the population means of the experimental measurement and the standard measurement, respectively. Next, they took logarithmic transformation and got $H_1 : A < \eta_Y - \eta_X < B$, where $A = \log_{10}(A_0)$, $B = \log_{10}(B_0)$, $\eta_Y = \log_{10}(\mu_Y)$ and $\eta_X = \log_{10}(\mu_X)$. The test statistic was $T = \dfrac{(\bar{Y} - \bar{X}) - (A+B)/2}{S\sqrt{2/n}}$, where $\bar{Y}$ and $\bar{X}$ are the sample means for the new measurement and for the standard measurement (in the logarithmic scale) and S is the mean squared error from the ANOVA table under the logarithmic transformation. They quested that the difference in means was from the center of the equivalence interval $(A + B)/2$. Under testing ±20% criteria, the alternative hypothesis is $H_1 : 0.8 < \mu_Y/\mu_X < 1.2$ They approximated the p-value with $\rho = F_\upsilon(|T| - \delta) - F_\upsilon(-|T| - \delta)$, where $\delta = \dfrac{B-A}{2S\sqrt{2/n}}$.

Another well-known bioequivalent paper was published by Schuirmann (1987). They extended the Hauck & Anderson (1984) by using the "two one-sided tests" (TOST) with the alternative hypothesis

$$H_1 : \theta_1 < \mu_Y - \mu_X < \theta_2,$$

where $(\theta_1 < \theta_2)$, θ_1 and θ_2 are the lower and upper bounds specified in the two one-sided tests (TOST). They consider it to be two separate tests:

$$H_{0A} : \mu_Y - \mu_X \le \theta_1 \quad \text{vs.} \quad H_{1A} : \mu_Y - \mu_X > \theta_1$$
$$\text{and} \quad H_{0B} : \mu_Y - \mu_X \ge \theta_2 \quad \text{vs.} \quad H_{1B} : \mu_Y - \mu_X < \theta_2$$

The test statistics

$$t_1 = \frac{(\bar{Y} - \bar{X}) - \theta_1}{S\sqrt{2/n}} \text{ and } t_2 = \frac{\theta_2 - (\bar{Y} - \bar{X})}{S\sqrt{2/n}}.$$

If $t_1 \ge t_{1-\alpha,\upsilon}$ and $t_2 \ge t_{1-\alpha,\upsilon}$, they will reject H_0. Then they concluded the alternative hypothesis which states that both measurements are equivalent. Next, they also added another approach is called the power approach that is a statistical test for no difference between the average of two measurements at the level 0.05. The power approach for the alternative hypothesis of no difference is $H_1 : \mu_Y - \mu_X \ne \theta$. Their rejection region for the power approach under the null hypothesis is $-t_{0.975,\upsilon} \le \dfrac{(\bar{Y} - \bar{X})}{S\sqrt{2/n}} \le t_{0.975,\upsilon}$. They compared these two approaches and the results showed that "the power of two one-sided tests (TOST) is superior to the power approach as a test of the interval hypothesis H_0."

Stegner et al. (1996) explained "the equivalence testing for use in psychosocial and services research: an introduction with examples." They use the two one-sided test to test the hypothesis. Their method is the same as Schuirmann (1987)'s method. However, they showed the example of the original data and the logarithmic transformation data.

Berger et al. (1996) developed their testing from Hauck & Anderson (1984) method by using the two one-sided tests (TOST) which were proposed by Schuirmann (1987). The bioequivalence of proportion of the population mean was (μ_Y/μ_X) and took logarithmic transformation to get the testing: $H_0 : \eta_Y - \eta_X \le \theta_l$ or $\eta_Y - \eta_X \ge \theta_u$, where $\eta_Y = \log(\mu_Y)$, $\eta_X = \log(\mu_X)$. Let μ_X and μ_Y denote the true population means of AUC for the standard drug and for new drug, respectively. If they reject H_0 for both tests, then they will declare that two measurements are equivalent. They also showed the misconception of size α in equivalence testing. The $100(1-2\alpha)$% two-sided confidence interval for $\eta_Y - \eta_X$ is $[D^- = \bar{D} - t_{\alpha,\upsilon}S\sqrt{2/n}, D^+ = \bar{D} + t_{\alpha,\upsilon}S\sqrt{2/n}]$ when $\bar{D} = \bar{X} - \bar{Y}$ From this confidence interval, they concluded the test drug is equivalent to the standard measurement if and only if $[D^-, D^+] \subset (\theta_l, \theta_u)$.

Brown et al. (1997) developed Schuirmann (1987) with the alternative hypothesis is $H_1 : |\theta| < \Delta$, where $\theta = \log(\rho) = \eta_Y - \eta_X$, $\eta_Y = \log(\mu_Y)$, $\eta_X = \log(\mu_X)$. When μ_Y and μ_X are the parameters for the new treatment and the standard treatment. $\Delta = \log(1.25) = 0.223$. Let $D \sim N(\theta, \sigma^2)$, $\bar{D}$ is an estimated value of θ. By Schuirmann (1987), the rejection region of the two one-sided tests at the level α is $\Delta \ge |\bar{D}| + t_\alpha S_{\bar{D}}$, where t_α is the upper quantile α of the t distribution with degrees of freedom υ and $S_{\bar{D}}$ is the standard error of the mean difference.

Chambers et al. (2005) used the same method as Schuirmann (1987)'s method. Then the confidence

interval for the difference mean is $(\bar{Y}-\bar{X})\pm t_{\alpha,n_1+n_2-2}$ $S_p\sqrt{1/n_1+1/n_2}$, where S_p is the pooled estimate of σ. We will reject H_0 if the confidence interval is within the acceptance limits, and then we conclude that these two measurements are equivalent.

Limentani et al. (2005) used the two sample t test to calculate the critical value $\theta=\delta+S^*[t_{1-\alpha,n-2}+t_{1-\beta/2,n-2}]$ $\sqrt{2/n}$, where δ is a hypothetical value, α is a given significance level, β is the type 2 error, $S^*=S\sqrt{\frac{n-1}{\chi^2_{\gamma,n-1}}}$, and $\chi^2_{\gamma,n-1}$ is the $(100\gamma)^{th}$ percentile of the chi-square distribution with $n-1$ degrees of freedom. They will reject the hypothesis H_0 if the confidence interval for the difference mean is contained within $(-\theta,\theta)$. This concludes that these two measurements are nonequivalent.

Nandakumar (2009) developed the study from Stefanescu & Mehrotra (2007) with cross-over design by comparing the test and reference drug formulation's effect on a subject for the small sample from. Moreover, they proposed the robust procedures to test the small sample population bioequivalence hypothesis. In addition, he propounded the multivariate bioequivalence hypothesis by comparing the Least squares procedure and the component-wise rank method.

Wellek (2010) proposed the alternative hypothesis of the paired t-test for equivalent testing is $H_1:\theta_1<\mu_D/\sigma_D<\theta_2$. Let D_i be the differences between the pair measurements, $D_i\sim N(\mu_D,\sigma_D^2)$, $\forall_i=1,2,\ldots,n$, and S_D is the estimated value of σ_D. If $\Delta=-\theta_1=\theta_2$, then our alternative hypothesis will be $H_1:|T|<\Delta$, where $T=\frac{\bar{D}}{S_D/\sqrt{n}}$. Then $|T|$ is a folded t distribution. He showed that $|T|=\sqrt{T^2}$ and T^2 has the F distribution with degrees of freedom 1 and $n-1$. Then, the statistic that he used for equivalence testing is $|T|\sim\sqrt{F_{1,n-1}}$ with noncentral parameter $\delta=\frac{\mu_D}{S_D/\sqrt{n}}$.

These are some literature reviews that related to the clinical equivalence testing. However, we focuses on the paired-observations study. Also, the purpose of our study is to detect shift or scale type relationships on bioequivalence testing.

CLINICAL EQUIVALENCE TESTING

The weakness of an approach is that equivalence is stated under the null hypothesis and is default conclusion when evidence is lacking to say otherwise. Here, we propose to use a clinical equivalence approach where equivalence is the alternative and require the burden of proof.

Our study focuses on testing for statistical equivalence of two different measurements. A new measurement may be proclaimed equivalent to the current measurement if the difference is small. So we set an alternative hypothesis that $\theta=\mu_1-\mu_2$ is in a small interval about 0, i.e. $H_1:\theta_1<\theta<\theta_2$ where θ_1 and θ_2 are limits specified by the investigator. The null hypothesis is that θ does not lie within the required limits, or $H_0:\theta\le\theta_1$ or $\theta\ge\theta_2$. More details can be found in Lehmann & Romano (2005).

The hypothesis for the two one-sided tests are

$$H_0:\theta\le\theta_1 \text{ or } \theta\ge\theta_2 \quad \text{vs} \quad H_1:\theta_1<\theta<\theta_2.$$

Where $\theta_1<\theta_2$, θ_1 and θ_2 are the lower and upper bounds specified in the two one-sided tests (TOST). We can consider it to be two separate tests:

$$H_{0A}:\theta\le\theta_1 \quad \text{vs.} \quad H_{1A}:\theta>\theta_1$$
$$\text{and} \quad H_{0B}:\theta\ge\theta_2 \quad \text{vs.} \quad H_{1B}:\theta<\theta_2$$

Suppose, in addition, that $\theta_1=-\theta_2=\Delta$. Then our hypothesis is

$$H_0:|\theta|\ge\Delta \quad \text{vs} \quad H_1:|\theta|<\Delta.$$

where Δ is selected by the investigator or physician and is a benchmark for clinical equivalence. In this paper, we will compare three equivalence tests.

- Shift-Equivalence test $H_1:\beta_0=0$
- Scale-Equivalence test $H_1:\beta_1=1$
- Shift-Scale-Equivalence test $H_1:(\beta_0,\beta_1)=(0,1)$

SHIFT-EQUIVALENCE TEST

To test $H_1:\beta_0=0$, we can consider our paired testing as

$$H_1:\mu_d=0$$
$$\text{or} \quad H_0:|\mu_d|\ge\Delta \ \text{vs} \quad H_1:|\mu_d|<\Delta.$$

Where Δ is a small interval about 0.
We propose two different approaches
1. Shift-E test: Using TOST by Schuirmann approach.
2. Shift-E* test: Using square-root F-test by Wellek approach.

Shift-E test: Using TOST by Schuirmann approach

Schuirmann (1987) proposed the two one-sided tests (TOST). We apply his approach with our study which has the paired-observations study and our hypothesis as

$$H_0 : |\mu_d| \geq \Delta \quad \text{vs} \quad H_1 : |\mu_d| < \Delta$$

Then our two one-sided tests are

$$H_{0A} : \mu_d \leq -\Delta \quad \text{vs.} \quad H_{1A} : \mu_d > -\Delta$$

and $$H_{0B} : \mu_d \geq \Delta \quad \text{vs.} \quad H_{1B} : \mu_d \leq \Delta$$

and the statistics are

$$t_1 = \frac{\bar{D} - \Delta}{S_d / \sqrt{n}} \quad \text{and} \quad t_2 = \frac{\Delta - \bar{D}}{S_d / \sqrt{n}}$$

Where $\bar{D}$ and S_d are the mean and standard deviation of the difference between Y and X, respectively. It rejects H_0 at level α and declares the two measurements to be equivalent if both tests reject, that is, $t_1 \geq t_{1-\alpha,\upsilon}$ and $t_2 \geq t_{1-\alpha,\upsilon}$.

Shift-E* test: Using Square-Root F-test by Wellek approach.

Our alternative hypothesis will be

$$H_1 : \mu_d = 0$$

$$\text{or} \quad H_0 : |\mu_d| \geq \Delta \quad \text{vs} \quad H_1 : |\mu_d| < \Delta$$

Under the normality assumption, the statistic is

$$T = \frac{\bar{D}}{S_d / \sqrt{n}}$$

Where $\bar{D}$ and S_d are the mean and standard deviation of the difference between Y and X, respectively. From Wellek (2010) studied, $|T|$ is a folded t distribution. He showed that $|T| = \sqrt{T^2}$ and T^2 has the F distribution with degrees of freedom 1 and $n-1$. Hence, the rejection region is $|T| \geq \sqrt{F_{1-\alpha,1,n-1}}$

SCALE-EQUIVALENCE TEST

To test $H_1 : \beta_1 = 1$, the equivalence testing can be constructed the alternative hypothesis as

$$H_1 : \mu_Y / \mu_X = 1$$

$$\text{or} \quad H_0 : |\mu_Y / \mu_X| \geq \Delta_1 \quad \text{vs} \quad H_1 : |\mu_Y / \mu_X| < \Delta_1,$$

where Δ_1 is small interval about 1. We have to study Scale-E test: Using TOST by Berger approach.

Scale-E test:Using TOST by Berger approach

Berger et al. (1996) developed the Schuirmann (1987) approach (TOST) by using logarithmic transformation. Applying his method with our study, our hypothesis is

$$H_0 : |\eta_d| \geq \log \Delta_1 \quad \text{vs} \quad H_1 : |\eta_d| < \log \Delta_1$$

Where $\eta_d = \log \mu_Y - \log \mu_X$

Let $\Delta_1 > 1$, $A = \log \Delta_1$, $B = -A$. Then

$$H_1 : A < \eta_d < B,$$

It is two one-sided tests.

$$H_{0A} : \eta_d \geq A \quad \text{vs.} \quad H_{1A} : \eta_d < A$$

$$\text{or} \quad H_{0B} : \eta_d \leq B \quad \text{vs.} \quad H_{1B} : \eta_d > B$$

where $A = \log \Delta_1$, and $B = -A = \log 1/\Delta_1$

Under the normality assumption that the statistics are

$$t_u = \frac{\bar{D} - A}{S_d / \sqrt{n}} \quad \text{and} \quad t_l = \frac{\bar{D} - B}{S_d / \sqrt{n}}.$$

Where $\bar{D}$ and S_d are the mean and standard deviation of difference between Y and X (in logarithmic scale). It rejects null hypothesis at level α and declares the two measurements to be equivalent if both tests reject, that is, $t_l \geq t_{\alpha,\upsilon}$ and $t_u \leq -t_{\alpha,\upsilon}$.

2-DF FOR SHIFT-SCALE-EQUIVALENCE TESTING

Shift-Scale-Equivalence test

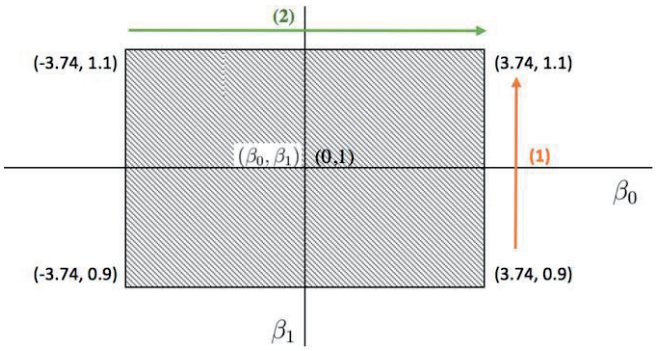
Shift-Scale model is equivalent to fitting a simple regression and conducting a 2-df test on both intercept and slope. We propose a procedure for doing clinical equivalence hypothesis testing: $H_1 : (\beta_0, \beta_1) = (0,1)$ or

$$H_0 : |\beta_0| \geq \Delta_0 \cup |\beta_1 - 1| \geq \Delta_1$$

$$\text{vs.} \quad H_1 : |\beta_0| < \Delta_0 \cap |\beta_1 - 1| < \Delta_1$$

Where Δ_0 and Δ_1 are investigator-specified values that represent limits of allowable clinical dissimilarities. We would like to call this testing as "**2-df for Shift-Scale-Equivalence Testing**" or **"Shift-Scale-Equivalence tesing."** In our simulations, we used $\Delta_0 = 0.10\bar{X} = 3.74$ and $\Delta_1 = 0.10$ so that our alternative hypothesis test representing equivalence is effectively

$$H_1 : \{|\beta_0| < 3.74\} \cap \{|\beta_1 - 1| < 0.10\}$$



Figures 1: An Alternative Region

This *alternative region* may be represented by a rectangle drawn in Figure 1. The null region is the region outside the rectangle. We propose a rejection region of the form

$$RR \equiv \{|\hat{\beta}_0| < C_0\} \cap \{|\hat{\beta}_1 - 1| < C_1\}$$

$$\equiv \{|\hat{\beta}_0| < t_\gamma S_{b_0}\} \cap \{|\hat{\beta}_1 - 1| < t_\gamma S_{b_1}\},$$

where t_γ is chosen so that the shift-scale test has desired size. The power function of the test is

$$\gamma(\beta) = P_\beta\left[\left|\hat{\beta}_0\right| < C_0, \left|\hat{\beta}_1 - 1\right| < C_1\right],$$
$$\beta = (\beta_0, \beta_1) \in \Omega = \Omega_0 \cup \Omega_1$$

We want to find the critical region based on the type I error is 0.05. That is the size α

$$\alpha = \sup_{\beta\in\Omega_0} \gamma(\beta) = \sup_{\beta\in\Omega_0} P\left[\left|\hat{\beta}_0\right| < C_0, \left|\hat{\beta}_1 - 1\right| < C_1\right],$$
where $C_0 = t_\gamma S_{b_0}, C_1 = t_\gamma S_{b_1}$
$$= \sup_{\beta\in\Omega_0} P\left[\left|\hat{\beta}_0\right| < t_\gamma S_{b_0}, \left|\hat{\beta}_1 - 1\right| < t_\gamma S_{b_1}\right]$$

Since the rejection region is symmetric about (0, 1) which is the center of the rectangle in Figure 1, then the power function decreases as we move farther away in either direction. Then the size of the test (or supremum of the power function under the null region) should occur on the boundary of the rectangle. This is the rationale behind the testing procedure/algorithm that we propose below. First we search for the maximum along the vertical sides of the rectangle (with its corresponding t_γ), then along the horizontal sides of the rectangle (and its corresponding t_γ). The smaller of the two t_γ 's gives the correctly-sized rejection region.

The steps to find the t_γ to satisfy the size of the test $\alpha = 0.05$.

Our alternative hypothesis is

$$H_1 : |\beta_0| < \Delta_0 \cap |\beta_1 - 1| < \Delta_1.$$

1. Calculate statistics:

$$\hat{\beta}_0,\ S_{b_0},\ \hat{\beta}_1,\ S_{b_1},\ \mathrm{cov}(\hat{\beta}_0,\ \hat{\beta}_1) = S_{b_{01}}$$

2. Set a rejection region:

$$RR \equiv \left\{\left|\hat{\beta}_0\right| < C_0\right\} \cap \left\{\left|\hat{\beta}_1 - 1\right| < C_1\right\}$$
$$\equiv \left\{\left|\hat{\beta}_0\right| < t_\gamma S_{b_0}\right\} \cap \left\{\left|\hat{\beta}_1 - 1\right| < t_\gamma S_{b_1}\right\}$$

3. Find the t_γ so that

$$\alpha = \sup_{\beta\in\Omega_0} \gamma(\beta) = \sup_{\beta\in\Omega_0} P(RR) \le 0.05,$$
where $H_0 : |\beta_0| \ge \Delta_0 \cup |\beta_1 - 1| \ge \Delta_1$

$$P(RR) = P\left[\left|\hat{\beta}_0\right| < t_\gamma S_{b_0}, \left|\hat{\beta}_1 - 1\right| < t_\gamma S_{b_1}\right]$$
$$= P\left[-t_\gamma S_{b_0} < \hat{\beta}_0 < t_\gamma S_{b_0}, 1 - t_\gamma S_{b_1} < \hat{\beta}_1 < 1 + t_\gamma S_{b_1}\right]$$

We know that

$$f(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{2\pi\sigma_{b_0}\sigma_{b_1}\sqrt{1-\rho_b^2}} e^{-\frac{1}{2(1-\rho_b^2)}\left[\left(\frac{y_0-\beta_0}{\sigma_{b_0}}\right)^2 + \left(\frac{y_1-\beta_1}{\sigma_{b_1}}\right)^2 - 2\rho_b\left(\frac{y_0-\beta_0}{\sigma_{b_0}}\right)\left(\frac{y_1-\beta_1}{\sigma_{b_1}}\right)\right]}$$

where $\sigma_{b_0}, \sigma_{b_1}, \sigma_{b_{01}}$ and ρ_b are unknown. So we use their estimators: Sb_0, Sb_1, Sb_{01} and r_b.

Then

$$P(RR) = \int_{1-t_\gamma S_{b_1}}^{1+t_\gamma S_{b_1}} \int_{-t_\gamma S_{b_0}}^{t_\gamma S_{b_0}} f(\hat{\beta}_0, \hat{\beta}_1)\, dy_0\, dy_1 ,$$

and $\alpha = \sup_{(\beta_0,\beta_1)\in H_0} P(RR)$

We can find the $t_\gamma = \min(t_{\gamma 1}, t_{\gamma 2})$.

(a) Find $t_{\gamma 1}$

We fix $\beta_0 = \Delta_0$ and evaluate

$$P_{\beta_1}(RR) = P_{\beta_1}(RR \mid \beta_0 = \Delta_0) \equiv h(\beta_1)$$

$$h(\beta_1) = \int_{1-t_\gamma S_{b_1}}^{1+t_\gamma S_{b_1}} \int_{-t_\gamma S_{b_0}}^{t_\gamma S_{b_0}} f(\Delta_0, \hat{\beta}_1) dy_0\, dy_1 \le 0.05 \qquad (1)$$

Where $f(\Delta_0, \hat{\beta}_1) = \frac{1}{2\pi S_{b_0} S_{b_1}\sqrt{1-r_b^2}} e^{-\frac{1}{2(1-r_b^2)}\left[\left(\frac{y_0-\Delta_0}{S_{b_0}}\right)^2 + \left(\frac{y_1-\beta_1}{S_{b_1}}\right)^2 - 2r_b\left(\frac{y_0-\Delta_0}{S_{b_0}}\right)\left(\frac{y_1-\beta_1}{S_{b_1}}\right)\right]}$

Then we calculate β_1^* by derivative $h(\beta_1)$ with respect to β_1 and set it equal to 0.

$$g(\beta_1) = \frac{\partial}{\partial\beta_1} h(\beta_1)$$
$$= \int_{1-t_\gamma S_{b_1}}^{1+t_\gamma S_{b_1}} \int_{-t_\gamma S_{b_0}}^{t_\gamma S_{b_0}} f(\Delta_0, \hat{\beta}_1) \frac{1}{(1-r_b^2)S_{b_1}}\left[\left(\frac{y_1-\beta_1}{S_{b_1}}\right) - r_b\left(\frac{y_0-\Delta_0}{S_{b_0}}\right)\right] dy_0\, dy_1$$
$$\equiv 0 \qquad (2)$$

An algorithm to find the $t_{\gamma 1}$.

(i) Start with $\beta_1^{(0)} = \beta_1$. Calculate $t_{\gamma 1}^{(1)}$ from equation (1).

(ii) Use $t_{\gamma 1}^{(1)}$ and equation (2) to get $\beta_1^{(1)}$.

(iii) Repeat (i) and (ii) until $\left|t_{\gamma 1}^{(i+1)} - t_{\gamma 1}^{(i)}\right| < 0.001$. We will get $t_{\gamma 1}$.

(b) Find $t_{\gamma 2}$

We set $\beta_1 = \Delta_1$, and calculate

$$P_{\beta_0}(RR) = P_{\beta_0}(RR \mid \beta_1 = \Delta_1) \equiv h(\beta_0)$$

$$h(\beta_0) = \int_{1-t_\gamma S_{b_1}}^{1+t_\gamma S_{b_1}} \int_{-t_\gamma S_{b_0}}^{t_\gamma S_{b_0}} f(\hat{\beta}_0, \Delta_1) dy_0\, dy_1 \le 0.05 \qquad (3)$$

Where $f(\hat{\beta}_0, \Delta_1) = \frac{1}{2\pi S_{b_0} S_{b_1}\sqrt{1-r_b^2}} e^{-\frac{1}{2(1-r_b^2)}\left[\left(\frac{y_0-\beta_0}{S_{b_0}}\right)^2 + \left(\frac{y_1-\Delta_1}{S_{b_1}}\right)^2 - 2r_b\left(\frac{y_0-\beta_0}{S_{b_0}}\right)\left(\frac{y_1-\Delta_1}{S_{b_1}}\right)\right]}$

Then we evaluate the β_0^* by derivative $h(\beta_0)$ with respect to β_0 and set equal to 0.

$$g(\beta_0)=\frac{\partial}{\partial\beta_0}h(\beta_0)$$
$$=\int_{1-t_\gamma S_{b_1}}^{1+t_\gamma S_{b_1}}\int_{-t_\gamma S_{b_0}}^{t_\gamma S_{b_0}} f(\hat{\beta}_0,\Delta_1)\frac{1}{(1-r_b^2)S_{b_0}}\left[\left(\frac{y_0-\beta_0}{S_{b_0}}\right)-r_b\left(\frac{y_1-\Delta_1}{S_{b_1}}\right)\right]dy_0\,dy_1$$
$$\equiv 0 \qquad (4)$$

An algorithm to find the $t_{\gamma 2}$.

(i) Start with $\beta_0^{(0)}=\beta_0$. Calculate $t_{\gamma 2}^{(1)}$ from equation (3).

(ii) Use $t_{\gamma 2}^{(1)}$ and equation (4) to get $\beta_0^{(1)}$.

(iii) Repeat (i) and (ii) until $\left|t_{\gamma 2}^{(i+1)}-t_{\gamma 2}^{(i)}\right|<0.001$. We will get $t_{\gamma 2}$.

We will get critical value $t_\gamma=\min(t_{\gamma 1},t_{\gamma 2})$ and then $C_0=t_\gamma S_{b_0}, C_1=t_\gamma S_{b_1}$. We check whether the value of the test statistic $\hat{\beta}_0$ falls in the rejection region $(-C_0,C_0)$ and $\hat{\beta}_1$ falls in the rejection region $(1-C_1,1+C_1)$. If it does, then we reject the null hypothesis and conclude these two measurements are equivalence. If not, we cannot reject null hypothesis and conclude $H_0:|\beta_0|\geq\Delta_0\cup|\beta_1-1|\geq\Delta_1$ or these two measurements are not equivalent. I would like to call this test as "2-df for Shift-Scale-Equivalence Testing".

SIMULATIONS

We investigate the performance of our equivalence hypotheses through simulation. We study the size and the power under various conditions. It is the proportion of p values that are lower than a specified α -level, which is 0.05. The simulations were replicated 10,000 times with $Y=\beta_0+\beta_1X+\varepsilon$, where X is the Westergren data and $\varepsilon\sim N(0,\sigma^2)$, where σ^2 are 3^2 for each of the following cases:

1. Shifted simulation: $Y=\beta_0+X+\varepsilon$ (See table 1)
2. Scaled simulation: $Y=\beta_1X+\varepsilon$ (See table 2)
3. Shift-Scaled simulation: $Y=\beta_0+\beta_1X+\varepsilon$ (See table 3), where $\varepsilon\sim N(0,\sigma^2)$, $\sigma^2=3^2$
   β_0 = -4, -3.9, -3.8, ..., 4
   β_1 = 0.8, 0.85, 0.9, ..., 1.2.

For each simulation, we compare four different approaches as follows:

- Shift-E test: Using TOST by Schuirmann approach.
- Shift-E* test: Using Square-Root F-test by Wellek approach.
- Scale-E test: Using TOST by Berger approach.
- Shift-Scale-Equivalence test.

Shifted simulation:

Table 1: The Power of the Test
Data Simulated from $Y=\beta_0+X+\varepsilon$,
where $\beta_0=-4,-3.90,-3.80,\ldots,4$, and $\sigma^2=3^2$.

	β_0	Shift-E	Shift-E*	Scale-E	Shift-Scale-E
H_0	-4	0.9197	0.0000	0.1540	0.0277
H_0	-3.9	0.9467	0.0000	0.1930	0.0403
H_0	-3.8	0.9653	0.0000	0.2327	0.0560
H_1	-3.7	0.9787	0.0000	0.2783	0.0740
H_1	-3.6	0.9873	0.0000	0.3230	0.0937
H_1	-3.5	0.9920	0.0000	0.3733	0.1177
H_1	-3	0.9997	0.0143	0.6623	0.3077
H_1	-2	1.0000	0.5130	0.9660	0.8097
H_1	-1	1.0000	0.9843	0.9997	0.9880
H_1	0	1.0000	1.0000	1.0000	0.9993
H_1	1	1.0000	1.0000	0.9990	0.9873
H_1	2	1.0000	1.0000	0.9260	0.8057
H_1	3	1.0000	1.0000	0.4663	0.2787
H_1	3.5	0.9940	1.0000	0.2353	0.1137
H_1	3.6	0.9890	1.0000	0.1967	0.0893
H_1	3.7	0.9807	1.0000	0.1730	0.0700
H_0	3.8	0.9677	1.0000	0.1443	0.0513
H_0	3.9	0.9490	1.0000	0.1153	0.0387
H_0	4	0.9153	1.0000	0.0970	0.0277

The Shift-Scale-E test achieves the size of the test $(\alpha=0.05)$, but it has lower power than Shift-E test. In comparison, the Shift-E test, does not hold its size.

Scaled simulations:

Table 2: The Power of the Test
Data Simulated from $Y=\beta_1X+\varepsilon$,
where β_1 = 0.8, 0.85, 0.9, ..., 1.2, and $\sigma^2=3^2$.

	β_1	Shift-E	Shift-E*	Scale-E	Shift-Scale-E
H_0	0.80	0.0000	0.0000	0.6033	0.0000
H_0	0.85	0.3163	0.0000	0.8803	0.0000
H_1	0.90	0.9970	0.0000	0.9893	0.0010
H_1	0.92	1.0000	0.0003	0.9970	0.0353
H_1	0.94	1.0000	0.1087	0.9993	0.3117
H_1	0.96	1.0000	0.7860	0.9997	0.8183
H_1	0.98	1.0000	0.9940	1.0000	0.9907
H_1	1.00	1.0000	1.0000	1.0000	0.9993
H_1	1.02	1.0000	1.0000	1.0000	0.9867
H_1	1.04	1.0000	1.0000	1.0000	0.8213
H_1	1.06	1.0000	1.0000	0.9990	0.3287
H_1	1.08	1.0000	1.0000	0.9923	0.0340
H_1	1.10	0.9990	1.0000	0.9733	0.0007
H_0	1.15	0.2977	1.0000	0.8203	0.0000
H_0	1.20	0.0003	1.0000	0.5580	0.0000

The Shift-E test holds the size on the boundary of the H_0 region. In contrast, the Shift-E* test, Scale-E test and Shift-Scale-E test do not hold their size and have lower power than Shift-E test.

Shift-Scaled simulations:

Table 3: The Power of the Test
Data Simulated from $Y = \beta_0 + \beta_1 X + \varepsilon$,
where $\beta_1 = 0.98$, $\beta_0 = -4, -3.9, -3.8, ..., 4$ and $\sigma^2 = 3^2$

	β_0	Shift-E	Shift-E	Scale-E	Shift-Scale-E
H_0	-4	0.3893	0.0000	0.0633	0.0257
H_0	-3.9	0.4797	0.0000	0.0837	0.0383
H_0	-3.8	0.5757	0.0000	0.1043	0.0537
H_1	-3.7	0.6717	0.0000	0.1387	0.0713
H_1	-3.6	0.7537	0.0000	0.1767	0.0897
H_1	-3.5	0.8163	0.0000	0.2127	0.1133
H_1	-3	0.9763	0.0003	0.4513	0.3010
H_1	-2	1.0000	0.0423	0.9090	0.8007
H_1	-1	1.0000	0.6867	0.9963	0.9790
H_1	0	1.0000	0.9940	1.0000	0.9907
H_1	1	1.0000	1.0000	1.0000	0.9817
H_1	2	1.0000	1.0000	0.9743	0.8060
H_1	3	1.0000	1.0000	0.6570	0.2790
H_1	3.5	1.0000	1.0000	0.3760	0.1140
H_1	3.6	1.0000	1.0000	0.3270	0.0897
H_1	3.7	1.0000	1.0000	0.2813	0.0703
H_0	3.8	1.0000	1.0000	0.2363	0.0517
H_0	3.9	1.0000	1.0000	0.1920	0.0390
H_0	4	1.0000	1.0000	0.1660	0.0280

The Shift-Scale-E test achieves the size of the test, but has lower power than Shift-E test. In comparison, the Shift-E test, does not hold its size.

CONCLUSIONS

Our simulations show that the Shift-E testing performs the best as it has the most power of the test without considering the size of the test. On the other hand, the Shift-Scale-E testing achieves the size of the test, but it has lower power of the test.

If we consider the power of the test only, the Shift-E test (it is well-known as Two One-sided Test: TOST) works the best. To consider the size ot the test, only the Shift-Scale-E testing (it is our proposed testing as 2-df for shift-scale equivalence test) works well as our expected.

ACKNOWLEDGEMENT

This work was financially support by the grant research fund of the Faculty of Science, King Mongkut's Institute of Technology Ladkrabang.

REFERENCES

Berger, R. L., Hsu, J. C., et al. 1996. "Bioequivalence trials, intersection-union tests and equivalence confidence sets." *Statistical Science*, 11 (4), 283–319.

Brown, L. D., Hwang, J. G., & Munk, A. 1997. "An unbiased test for the bioequivalence problem." *The annals of Statistics*, (pp. 2345–2367).

Chambers, D., Kelly, G., Limentani, G., Lister, A., Lung, K. R., & Warner, E. 2005. Analytical method equivalency. Pharmaceutical Technology.

Hauck, W. W., & Anderson, S. 1984. "A new statistical procedure for testing equivalence in two-group comparative bioavailability trials." *Journal of Pharmacokinetics and Biopharmaceutics*, 12 (1), 83–91.

Jones, B., & Kenward, M. G. 2014. Design and analysis of cross-over trials. CRC Press.

Kirkwood, T. B. L., &Westlake, W. J. 1981. "Bioequivalence testing–a need to rethink." *Biometrics*, 37 (3), 589–594.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., et al. 2005. Applied linear statistical models, vol. 103. McGraw-Hill Irwin New York.

Lehmann, E. L., & Romano, J. P. 2005. Testing statistical hypotheses. Springer Science & Business Media.

Limentani, G. B., Ringo, M. C., Ye, F., Bergquist, M. L., & MCSorley, E. O. 2005. Beyond the t-test: statistical equivalence testing.

Nandakumar, S. P. 2009. Statistical procedures for bioequivalence analysis.

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.

Stapleton, J. H. 2009. Linear statistical models, vol. 719. John Wiley & Sons.

Stefanescu, C., & Mehrotra, D. V. 2007. "A more powerful average bioequivalence analysis for the 2× 2 crossover." *Communications in Statistics-Simulation and Computation*, 37 (1), 212–221.

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. 1996. "Equivalence testing for use in psychosocial and services research: An introduction with examples." *Evaluation and Program Planning*, 19 (3), 193–198.

Wellek, S. 2010. Testing statistical hypotheses of equivalence and noninferiority. CRC Press.

Westlake, W. J. 1972. "Use of confidence intervals in analysis of comparative bioavailability trials." *Journal of Pharmaceutical Sciences*, 61 (8), 1340–1341.

Westlake, W. J. 1976. "Symmetrical confidence intervals for bioequivalence trials." *Biometrics*, (pp. 741–744).

Westlake, W. J. 1979. "Design and statistical evaluation of bioequivalence studies in man." *In Principles and Perspectives in Drug Bioavailability*, (pp. 192–210). Karger Publishers