12-2017

Statistical and Clinical Equivalence of Measurements

Puntipa Wanitjirattikal
Western Michigan University, puntipa.w@hotmail.com

STATISTICAL AND CLINICAL EQUIVALENCE OF MEASUREMENTS

by

Puntipa Wanitjirattikal

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Statistics
Western Michigan University
December 2017

Doctoral Committee:

Joshua Naranjo, Ph.D., Chair
Joseph McKean, Ph.D.
Georgiana Onicescu, Ph.D.
Karen Villarente Rosales, Ph.D.

STATISTICAL AND CLINICAL EQUIVALENCE OF MEASUREMENTS

Puntipa Wanitjirattikal, Ph.D.

Western Michigan University, 2017

This study proposes a test for statistical equivalence of two measurements. Typically, a new measurement process $\boldsymbol{Y}$ is compared to an existing or standard measurement process $\boldsymbol{X}$. We are assuming that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are measurements on the same scale. The paired t-test may be used to check for significant difference between $(\boldsymbol{X}, \boldsymbol{Y})$ pairs. However, the paired t-test is intended to detect shift-type relationships of the form $\boldsymbol{Y} = \boldsymbol{X} + \delta\mathbf{1}$ and may have low power for scale-type relations of the form $\boldsymbol{Y} = \gamma\boldsymbol{X}$.

We propose a test that has reasonable power to detect either shift or scale-type relationships. Secondly, we propose a bioequivalence testing approach to swap the hypotheses so that statistical equivalence of the two measurements is the alternative hypothesis and bears the burden of proof. Rather than being the default conclusion in the absence of sufficient evidence, we conclude "clinical equivalence" only if there is evidence to support the claim that the magnitude of disagreement between the two measurements lies within specified limits.

Acknowledgements

I would first like to express my gratitude to my advisor, Dr. Joshua Naranjo. His selfless support, encouragement, and help gave me the courage to overcome the obstacles, so that I could finish this dissertation.

I would also like to thank my committee members, Dr. Joseph McKean, Dr. Georgiana Onicescu, and Dr. Karen Villarente Rosales. Their suggestions are invaluable.

I would like to thank The Royal Thai Government and King Mongkut's Institute of Technology Ladkraban to sponsor me.

I would like to thank my family. Without their support, I would not have a chance to study in American and finish my Ph.D. program.

Lastly, I would like to thank Dr. Chenyang Shi, my dearest friend. His support and encouragement will be with me in my entire life.

Puntipa Wanitjirattikal

Table of Contents

Table of Contents—Continued

List of Tables

List of Tables—Continued

List of Figures

Chapter 1

Introduction

Erythrocyte sedimentation rate (ESR) is a blood test that can reveal inflammatory activity in your body. Inflammation causes red blood cells to clump, and the dense clumps settle to the bottom more quickly. Westergren ESR and STATplus ESR are two popular measurements of sedimentation rate. These two measurements are used to monitor disease severity in patients with rheumatoid arthritis and other inflammatory rheumatologic conditions. Westergren ESR was developed by R. S. Fahraeus and A.V.A. Westergren in 1921. STATplus ESR or Boycott is a new measurement to accelerate turnaround time to less than 30 minutes for Westergren ESR. Compared with Westergren ESR, the result from STATplus ESR is easier to understand. Since these two measurements can be used for the same testing, it is necessary to explore that what the best situation is to use STATplus.

Our study focuses on testing for statistical equivalence of two different measurements. Typically, a new measurement process $\boldsymbol{Y}$ is compared to an existing (or standard) measurement process $\boldsymbol{X}$. Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$, the paired t-test may be used to check the difference between pairs of measurement. However, the paired t-test is intended to detect shift-type relationships of the form

$\boldsymbol{Y} = \boldsymbol{X} + \delta\mathbf{1}$ and may have low power for scale-type relations of the form $\boldsymbol{Y} = \gamma\boldsymbol{X}$. In this dissertation, we aim to propose a test that has reasonable power to detect either shift or scale-type relationships. The second part of our research will use principles of clinical equivalence testing to *invert* the hypothesis so that statistical equivalence of the two measurements is the alternative hypothesis and bears the burden of proof. Rather than being the default conclusion in the absence of sufficient evidence, we will conclude "clinical equivalence" only if there is evidence to support the claim that the magnitude of disagreement between the two measurements lies within specified limits.

1.1 Research Objectives

1. Since the difference between $\boldsymbol{Y}$ (new measurement) and $\boldsymbol{X}$ (standard measurement) may be either a shift-type or scale type, we propose the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ and propose the general Shift-Scale test: $H_0 : (\beta_0, \beta_1) = (0, 1)$

2. We investigate the properties of a two-df test for Shift-Scale test $H_0 : (\beta_0, \beta_1) = (0, 1)$ compared to, e.g. the paired t-test in Shift-test, or a test for slope based on Scale test.

3. We investigate ways to extend the 2-df test for Shift-Scale test $H_0 : (\beta_0, \beta_1) = (0, 1)$ to be used in the framework of bioequivalence testing where equivalence of the two measurements is treated as the alternative instead of the null hypothesis.

4. More specifically, we investigate how to conduct a 2-df test for Shift-Scale-Equivalence test $H_1 : (\beta_0, \beta_1) = (0, 1)$ using extensions of the "two one-sided tests (TOST)"

5. We investigate the properties of the 2-df clinical equivalence test for $H_1 : (\beta_0, \beta_1) = (0, 1)$ and compare it with the 1-df clinical equivalence test for $H_1 : \beta_0 = 0$ or $H_1 : \beta_1 = 1$.

1.2 Outline

Chapter 1 contents the background and motivation, statement of the problem and research objectives. Related studies in the literature will be discussed in Chapter 2 for clinical equivalence testing. In Chapter 3, the statistical hypothesis testing will be discussed using three modeling approaches:

- Shift test $H_0 : \beta_0 = 0$, where $\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$ (Shift model)
- Scale test $H_0 : \beta_1 = 1$, where $\boldsymbol{Y} = \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$ (Scale model)
- Shift-Scale test $H_0 : (\beta_0, \beta_1) = (0, 1)$, where $\boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$ (Shift-scale model),

and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ and $\varepsilon_i \sim N(0, \sigma^2)$ for all these models. The paired t-test is optimal for Shift test. However, it performs badly when the data is generated from Scale-type relationships. Scale test is equivalent to fit a simple regression through the origin, and conducting a t-test on the slope parameter. However, this test performs poorly when the data is generated from Shift model. Shift-Scale test is equivalent to fitting a simple regression and conducting a 2-degree-of-freedom test. Our simulation results show that it performs relatively well under either Shift-type or Scale-type relationships. This means that practitioners can conduct a test that has reasonable power without first conducting a pretest for whether the alternative is shift or scale.

In chapter 4, we use clinical equivalence testing to Shift test and Scale test. In other words, clinical equivalence will be treated as the alternative rather the null hypothesis, and will thus bear the burden of proof. Chapter 5 uses the modeling framework of Shift-Scale test but apply the principles of clinical equivalence testing.

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

This approach is equivalent to testing that the magnitude of disagreement between the two measurements is within specified limits. Then we will compare it with Shift-Equivalence test and Scale-Equivalence test. In Chapter 6, we will compare the performance of Shift-E, Scale-E and Shift-Scale-E with simulations. Chapter 7 will conclude our results, discuss our work, and propose some studies in the future.

Chapter 2

Literature Review

2.1 Clinical Equivalence Testing

In clinical trials and biostatistics studies, the clinical equivalence testing aids us to interprete 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 12 each subject. Based on normality assumptions, $\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s\sqrt{1/n_1+1/n_2}}$ has the t distribution with 10 degrees of freedom, where s^2 is the mean squared error, and the standard error is $s\sqrt{1/n_1+1/n_2}$. Then the 95% probability inequality of the mean total urinary excretion is $k_2 s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq (\bar{X}-\bar{Y})-(\mu_X-\mu_Y) \leq k_1 s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$. It is rearranged to be $\mu_X + k_2 s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} - (\bar{X}-\bar{Y}) \leq \mu_Y \leq \mu_X + k_1 s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} - (\bar{X}-\bar{Y})$. Adapting to paired t-test, $\mu_X + k_2 s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} - \bar{d} \leq \mu_Y \leq \mu_X + k_1 s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} - \bar{d}$, where $d_i = y_i - x_i$.

Hence the confidence interval for μ_Y would be $\mu_X - \Delta \leq \mu_Y \leq \mu_X + \Delta$, where $\Delta = k_1 s\sqrt{2/12} - (\bar{X}-\bar{Y}) = -k_2 s\sqrt{2/12} - (\bar{X}-\bar{Y})$ and $n = 12$. This implies that

$2(\bar{X} - \bar{Y}) = (k_1 + k_2)s\sqrt{2/12}$. 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 $2(\bar{X} - \bar{Y}) = (k_1 + k_2)s\sqrt{2/12}$. His 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 their 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

$$k_2 s\sqrt{2/n} - (\bar{X} - \bar{Y}) \leq \log_{10}(\mu_Y/\mu_X) \leq k_1 s\sqrt{2/n} - (\bar{X} - \bar{Y})$$

"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 $k_2 < \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s\sqrt{2/n}} < k_1$, 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 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 α=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 measure-

ments 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 $T = \frac{\bar{Y}-\bar{X}-(A+B)/2}{s\sqrt{2/n}}$, where $\bar{Y}$ and $\bar{X}$ are the sample mean 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 $\pm 20\%$ criteria, the alternative hypothesis is $H_1 : \ 0.8 < \mu_Y/\mu_X < 1.2$ They approximated the p-value with $\rho = F_\nu(|T| - \delta) - F_\nu(-|T| - \delta)$, where $\delta = \frac{B-A}{2s\sqrt{2/n}}$. In 1982, Blackwelder published similar work as Hauck & Anderson (1984).

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 $H_1 : \ \theta_1 < \mu_Y - \mu_X < \theta_2, \quad (\theta_1 < \theta_2)$, 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:

$$H_{0A} : \mu_Y - \mu_X \leq \theta_1 \quad \text{vs.} \quad H_{1A} : \mu_Y - \mu_X > \theta_1$$

and

$$H_{0B} : \mu_Y - \mu_X \geq \theta_2 \quad \text{vs.} \quad H_{1B} : \mu_Y - \mu_X < \theta_2$$

The test statistics $t_1 = \frac{(\bar{Y}-\bar{X})-\theta_1}{s\sqrt{2/n}}$ and $t_2 = \frac{\theta_2-(\bar{Y}-\bar{X})}{s\sqrt{2/n}}$. If $t_1 \geq t_{1-\alpha,\nu}$ and $t_2 \geq t_{1-\alpha,\nu}$, 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 which 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 $H_1 : \mu_Y - \mu_X \neq 0$. Their rejection region for the power approach under the null hypothesis is $-t_{0.975,\nu} \leq \frac{\bar{Y}-\bar{X}}{s\sqrt{2/n}} \leq t_{0.975,\nu}$. 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 $\left(\frac{\mu_Y}{\mu_X}\right)$ and took logarithmic transformation to get the testing: $H_0 : \eta_Y - \eta_X \leq \theta_l \quad \text{or} \quad \eta_Y - \eta_X \geq \theta_u$, 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. By United States Food and Drug Administration: FDA (1992), $\delta_u = 1.25,\ \delta_l = 0.8 = 1/1.25$ for AUC (area under curve) and C_{max} (maximum concentration). By the European Community, $\delta_u = 1.43,\ \delta_l = 0.7 = 1/1.43$ for C_{max}. They got $\theta_u = \log(\delta_u),\ \theta_l = \log(\delta_l)$, so $\theta_u = -\theta_l$. Then they showed the concept of intersection-union tests to clarify, simplify and unify bioequivalence testing. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the standard measurement and the new measurement in logarithmic scale. Then $\bar{D} = \bar{Y} - \bar{X},\ s_{\bar{D}} = s_p\sqrt{1/n_1 + 1/n_2},\ s_p$ is the pooled estimate of σ. The two statistics based on two one-sided tests (TOST) are $T_u = \frac{\bar{D}-\theta_u}{s_{\bar{D}}}$ and $T_l = \frac{\bar{D}-\theta_l}{s_{\bar{D}}}$. They will reject H_0 if $T_u < -t_{\alpha,r}$ and $T_l > t_{\alpha,r}$ at the level α, where $t_{\alpha,r}$ is the upper 100α percentile of a t distribution with the degrees of freedom $r = m + n - 2$. 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 $\left[D^- = \bar{D} - t_{\alpha,r} s_{\bar{D}},\ D^+ = \bar{D} + t_{\alpha,r} s_{\bar{D}}\right]$. 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 $\theta = \log(\rho) = \mu_Y - \mu_X,\ \mu_Y = \log(m_Y),\ \mu_X = \log(m_X)$. When m_Y and m_X are the parameters for the new treatment and the standard treatment. $\Delta = \log(1.25) = 0.223$. Let $\bar{D} \sim N(\theta, \sigma^2),\ \bar{D}$ is an estimated value of θ. By Schuirmann (1987), the rejection region of the two one-sided tests at the level α is $\Delta \geq |\bar{D}| + t_\alpha s_{\bar{D}}$, where t_α is the upper quantile α of the t distribution with degrees of freedom ν 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 $\theta = \delta + s^*[t_{(1-\alpha,2n-2)} + t_{(1-\beta/2,2n-2)}]\sqrt{2/n}$, 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, $D_i \sim N(\mu_D, \sigma_D^2)$, $\forall i = 1, ..., n$, and s_D is the estimated value of σ_D. Therefore, the test statistic is $T = \frac{\bar{D}}{s_D/\sqrt{n}}$, which is a noncentral t distribution with the degrees of freedom $n-1$ and the noncentral parameter $\delta = \frac{\mu_D}{s_D/\sqrt{n}}$. 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. Hence, the rejection region is $|T| < C_{\alpha,\delta}$. 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 $|T| = \left|\frac{\bar{D}}{s_D/\sqrt{n}}\right| \sim \sqrt{F_{1,n-1}}$ with noncentral parameter $\delta = \frac{\mu_D}{s_D/\sqrt{n}}$.

Jones & Kenward (2014) published a book "the design and analysis of cross-over trials." They studied testing of average bioequivalence by using TOST in chapter 7.

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.

Chapter 3

Test of Hypotheses

In this chapter we compare the performance of three modeling approaches to testing for dissimilarity between $\boldsymbol{Y}$ and $\boldsymbol{X}$.

1. $H_0 : \beta_0 = 0$, where $\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$ (Shift model)

2. $H_0 : \beta_1 = 1$, where $\boldsymbol{Y} = \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$ (Scale model)

3. $H_0 : (\beta_0, \beta_1) = (0, 1)$, where $\boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$ (Shift-scale model),

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ and $\mathbf{1}$ denotes the vector all of whose components are one. $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ and $\varepsilon_i \sim N(0, \sigma^2)$

All three hypotheses can be testing using the General Linear F-statistic. We show that F-test for shift (1) is the same as the paired-t test. We show using simulations that the F-test for shift performs badly under scale, and the F-test for scale performs badly under shift. This motivates a 2-df approach that tests simultaneously for shift and scale.

3.1 General Linear Test (F-statistic)

The general linear test starts from specification of the full model. Next, we set up a test for investigating whether a reduced model adequately fits the data. This test help us to ascertain whether full model or reduced model is good for our data. We use **F-test** to test our null hypothesis by Kutner et al. (2005) or Stapleton (2009). The general linear test approach requires the reduced model under the null hypothesis H_0.

$$H_0 : \theta \in V_0 \text{ which is the reduced model}$$
$$H_1 : \theta \notin V_0 \text{ which is the full model}$$

Where V_0 is a proper subspace of V of dimension $k_0 < k = \dim(V)$.

Then the test statistic is:

$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

$$\text{Where} \quad SSE(R) = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\boldsymbol{R}}\|^2,\ df_R = n - k_0$$

$$SSE(F) = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\boldsymbol{F}}\|^2,\ df_F = n - k.$$

$$\text{That is} \quad F = \frac{\|\hat{\boldsymbol{Y}}_{\boldsymbol{F}} - \hat{\boldsymbol{Y}}_{\boldsymbol{R}}\|^2/(k - k_0)}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\boldsymbol{F}}\|^2/(n - k)}$$

Where $\hat{\boldsymbol{Y}}_{\boldsymbol{F}} = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1\boldsymbol{X}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{R}} = \boldsymbol{X}$. The degrees of freedom for the reduced model and the full model are $n - k$ and $n - k_0$, respectively. After that, we assure the value of the F-statistic. The test will reject H_0 if the value of $F \geq F_{(k-k_0,n-k,1-\alpha)}$, and conclude H_1 at level α. Otherwise, we will fail to reject null hypothesis and conclude that the model $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$.

We can also calculate the p-value from our F-statistic. The p-value is the probability that we would observe a more extreme test statistic in the direction of the alternative

hypothesis than the observed value if the null hypothesis is actually true. Then p-value $=P(F_{(k-k_0,n-k)} > F\text{-statistic})$. After that, we compare our p-value to the significance level $\alpha = 0.05$. We will reject null hypothesis if our p-value is less than α. If not, we fail to reject null hypothesis.

3.1.1 Shift test $H_0 : \beta_0 = 0$

Here we will use the general linear F-test for

$$H_0 : \boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon} \quad \text{vs.} \quad H_1 : \boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}, \text{ where } \beta_0 \neq 0$$

The objective is to test equivalence between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The full model is $\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$. We propose the reduced model $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$.

Therefore, the F-statistic is

$$F = \frac{(SSE(R) - SSE(F))/(k - k_0)}{SSE(F)/(n-k)}$$

$$\begin{aligned} \text{where} \quad & SSE(R) = \|\boldsymbol{Y} - \boldsymbol{X}\|^2, \ df_R = n \\ & SSE(F) = \|\boldsymbol{Y} - (\hat{\beta}_0 \mathbf{1} + \boldsymbol{X})\|^2, \ df_F = n - 1 \\ & \hat{\beta}_0 = \bar{Y} - \bar{X}, \ \text{ by the least squares estimate} \end{aligned}$$

$$\begin{aligned} \text{That is} \quad F &= \frac{\|(\hat{\beta}_0 \mathbf{1} + \boldsymbol{X}) - \boldsymbol{X}\|^2/(1)}{\|\boldsymbol{Y} - (\hat{\beta}_0 \mathbf{1} + \boldsymbol{X})\|^2/(n-1)} \\ &= \frac{\|\hat{\beta}_0 \mathbf{1}\|^2}{\|\boldsymbol{Y} - ((\bar{Y} - \bar{X})\mathbf{1} + \boldsymbol{X})\|^2/(n-1)} \\ &= \frac{\|(\bar{Y} - \bar{X})\mathbf{1}\|^2}{\|(\boldsymbol{Y} - \bar{Y}\mathbf{1}) - (\boldsymbol{X} - \bar{X}\mathbf{1})\|^2/(n-1)} \end{aligned} \tag{3.1}$$

After calculation, we will reject H_0 at level α and declare the full model is adequate for our data if $F \geq F_{(1,n-1,1-\alpha)}$. Otherwise, we will fail to reject null hypothesis and conclude that the reduced model $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$. That is two measurements are equivalence.

Also, we can calculate the p-value which is $P(F_{1,n-1} > F\text{-statistic})$. If the p-value is less than significance level $\alpha = 0.05$, we will reject null hypothesis.

We would like to show that F test is same as the Paired-t test.

Proof:

$$\text{Full model:} \quad \hat{\boldsymbol{Y}}_{\boldsymbol{F}} = \hat{\beta}_0 \mathbf{1} + \boldsymbol{X} = (\bar{Y} - \bar{X})\mathbf{1} + \boldsymbol{X}, \qquad \text{by LS}$$

$$\text{Reduced model:} \quad \hat{\boldsymbol{Y}}_{\boldsymbol{R}} = \boldsymbol{X}$$

We know that

$$\begin{aligned}
SSE(R) - SSE(F) &= \|\hat{\boldsymbol{Y}}_{\boldsymbol{F}} - \hat{\boldsymbol{Y}}_{\boldsymbol{R}}\|^2 \\
&= \|((\bar{Y} - \bar{X})\mathbf{1} + \boldsymbol{X}) - \boldsymbol{X}\|^2 \\
&= \|(\bar{Y} - \bar{X})\mathbf{1}\|^2 \\
&= n(\bar{Y} - \bar{X})^2 \\
\text{and} \qquad MSE &= SSE(F)/(n-1) \\
&= \|(\boldsymbol{Y} - \bar{Y}\mathbf{1}) - (\boldsymbol{X} - \bar{X}\mathbf{1})\|^2/(n-1) \\
&= \sum((Y_i - \bar{Y}) - (X_i - \bar{X}))^2/(n-1) \\
&= \sum(d_i - \bar{d})^2/(n-1) \\
&= s_d^2 \\
\text{Then} \qquad F &= \frac{(SSE(R) - SSE(F))}{SSE(F)/(n-1)} \\
&= \frac{n(\bar{Y} - \bar{X})^2}{MSE} \\
&= \left(\frac{\bar{d}}{s_d/\sqrt{n}}\right)^2 \\
\text{Therefore} \qquad F &= (\text{Paired-t})^2
\end{aligned}$$

3.1.2 Scale test $H_0 : \beta_1 = 1$

Here we will use the general linear F-test for

$$H_0 : \boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon} \quad \text{vs.} \quad H_1 : \boldsymbol{Y} = \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}, \text{ where } \beta_1 \neq 1$$

Our full model is $\boldsymbol{Y} = \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$, and our reduced model is $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$.

Then, the calculation of F-statistic is

$$F = \frac{(SSE(R) - SSE(F))/(k - k_0)}{SSE(F)/(n-k)},$$

$$\begin{aligned} \text{where} \quad & SSE(R) = \|\boldsymbol{Y} - \boldsymbol{X}\|^2, \ df_R = n - 0 \\ & SSE(F) = \|\boldsymbol{Y} - (\hat{\beta}_1 \boldsymbol{X})\|^2, \ df_F = n - 1 \\ & \hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2}, \quad \text{by the least squares estimate} \end{aligned}$$

$$\text{That is} \quad F = \frac{\|\hat{\beta}_1 \boldsymbol{X} - \boldsymbol{X}\|^2/(1)}{\|\boldsymbol{Y} - \hat{\beta}_1 \boldsymbol{X}\|^2/(n-1)} \tag{3.2}$$

We calculate the value of the F-statistic and reject null hypothesis if the value of $F \geq F_{(1,n-1,1-\alpha)}$. Then we conclude alternative hypothesis or the full model satisfies for our data. Otherwise, we will fail to reject null hypothesis and conclude that the model $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$ is appropriate for our data.

3.1.3 Shift-Scale test $H_0 : (\beta_0, \beta_1) = (0, 1)$

Here we will use the general linear F-test for

$$H_0 : \boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon} \quad \text{vs.} \quad H_1 : \boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}, \text{ where } \beta_0 \neq 0 \text{ or } \beta_1 \neq 1$$

Our full model is $\boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$, and our reduced model is $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$.

The F-statistic can be calculated as

$$F = \frac{(SSE(R) - SSE(F))/(k - k_0)}{SSE(F)/(n-k)},$$

$$\text{where } SSE(R) = \|\boldsymbol{Y} - \boldsymbol{X}\|^2,\ df_R = n - k_0 = n - 0$$

$$SSE(F) = \|\boldsymbol{Y} - (\hat{\beta}_0\mathbf{1} + \hat{\beta}_1\boldsymbol{X})\|^2,\ df_F = n - 2$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}, \text{ and } \hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\text{That is} \quad F = \frac{\|(\hat{\beta}_0\mathbf{1} + \hat{\beta}_1\boldsymbol{X}) - \boldsymbol{X}\|^2/(2)}{\|\boldsymbol{Y} - (\hat{\beta}_0\mathbf{1} + \hat{\beta}_1\boldsymbol{X})\|^2/(n-2)} \tag{3.3}$$

We check the value of the F-statistic whether the value of $F \geq F_{(2,n-2,1-\alpha)}$. If yes, we will reject null hypothesis and conclude the alternative hypothesis: at least one is not equal, or the full model. Otherwise, we will fail to reject H_0 and conclude that our data does not contradiction model $\boldsymbol{Y} = \boldsymbol{X} + \varepsilon$

3.2 The Westergren/STATplus Data

Now, we apply the (3.1), (3.2) and (3.3) using the data in Table 3.1. This is real data that came from an inquiry to the WMU Statistical Consulting Center.

$$\text{Let } \boldsymbol{X} = \text{Westergren measurement}$$

$$\boldsymbol{Y} = \text{STATplus measurement}$$

3.2.1 Shift test

We test the following null hypothesis against a shift alternative hypotheses

$$H_0 : \boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon} \quad \text{vs.} \quad H_1 : \boldsymbol{Y} = \beta_0\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$$

Table 3.1: The Westergren/STATplus data.

Patient ID	x_i	y_i
1	0	5
2	16	14
3	6	5
4	21	10
5	15	14
6	19	14
7	4	5
8	11	12
9	10	8
10	1	3
11	2	3
12	25	14
13	82	104
14	83	108
15	13	22
16	39	46
17	8	36
18	52	66
19	69	77
20	56	36
21	40	52
22	35	25
23	57	60
24	41	53
25	5	6
26	30	29
27	1	6
28	31	37
29	1	1
30	77	74
31	105	120
32	48	67
33	70	80
34	123	130
35	108	115
36	27	46
37	67	92
38	7	10
39	19	16
40	19	19
41	69	108
42	35	92
43	40	34
44	94	140
45	8	9
46	58	81
47	40	52
48	8	10
Mean	37.40	45.13
sd	32.330	40.196

By least squares regression, we get $\hat{\beta}_0 = \bar{Y} - \bar{X} = 45.13 - 37.4 = 7.73$.

The F-statistic is

$$\begin{aligned} F &= \frac{\|\hat{\beta}_0 \mathbf{1}\|^2/(1)}{\|\boldsymbol{Y} - (\hat{\beta}_0 \mathbf{1} + \boldsymbol{X})\|^2/(n-1)} \\ &= \frac{\|(\bar{Y} - \bar{X})\mathbf{1}\|^2}{\|(\boldsymbol{Y} - \bar{Y}\mathbf{1}) - (\boldsymbol{X} - \bar{X}\mathbf{1})\|^2/(n-1)} \\ &= \frac{48(45.13 - 37.4)^2}{\sum(Y_i - 45.13 - X_i - 37.40)^2/(48-1)} \\ &= \frac{2867.52}{9981.47/47} \\ &= 13.502 \end{aligned}$$

Since $F = 13.502 > F_{1-\alpha,1,n-1} = 4.047$, so reject null hypothesis of $H_0 : \beta_0 = 0$. We conclude that the two measurements are not equivalent.

3.2.2 Scale test

Now consider a scale alternative:

$$H_0 : \boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon} \quad \text{vs.} \quad H_1 : \boldsymbol{Y} = \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$$

where $\beta_1 \neq 1$. By least squares regression, we get $\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{138541}{116251} = 1.1917$.

The F-statistic is

$$\begin{aligned} F &= \frac{\|(\hat{\beta}_1 - 1)\boldsymbol{X}\|^2}{\|\boldsymbol{Y} - \hat{\beta}_1 \boldsymbol{X}\|^2/(n-1)} \\ &= \frac{\sum((\hat{\beta}_1 - 1)X_i)^2}{\sum(Y_i - \hat{\beta}_1 X_i)^2/(48-1)} \\ &= \frac{\sum((1.1917 - 1)X_i)^2}{8575.11/47} \\ &= \frac{4273.89}{182.45} = 23.425 \end{aligned}$$

We reject $H_0 : \beta_1 = 1$ because $F = 23.425 > F_{1-\alpha,1,n-1} = 4.047$. We conclude that the two measurements are not equivalent.

3.2.3 Shift-Scale test

This time, we consider a shift-scale alternative:

$$H_0 : \boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon} \quad \text{vs.} \quad H_1 : \boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{X} + \boldsymbol{\varepsilon}$$

where $\beta_0 \neq 0$ or $\beta_1 \neq 1$. By least squares regression, we get

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{57541.63}{49125.48} = 1.1713$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 45.13 - 1.1713(37.4) = 1.3225$$

The F-statistic is

$$\begin{aligned} F &= \frac{\|(\hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \boldsymbol{X}) - \boldsymbol{X}\|^2/(2)}{\|\boldsymbol{Y} - (\hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \boldsymbol{X})\|^2/(n-2)} \\ &= \frac{\sum(1.3225 + (1.1713 - 1)X_i)^2/(2)}{\sum(Y_i - (1.3225 + 1.1713X_i))^2/(48-2)} \\ &= \frac{4309.37/2}{8539.63/46} \\ &= 11.6065 \end{aligned}$$

The value of $F = 11.6065 > F_{1-\alpha,2,n-2} = 3.199$. So, we reject H_0 and conclude H_1 : at least one is not equal. We conclude that the two measurements are not equivalent.

All these three models give the same conclusion. They are reject null hypotheses which are the reduce model: $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\varepsilon}$. That is two measurements are not equivalent.

3.3 Simulations

Now we investigate the performance of our three modeling approaches through simulation. We study power under various conditions. We kept the X-values the same as the Westergren data, and simulated Y-values under each of the following alternatives 10000 times:

- Shifted simulation: $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$
- Scaled simulation: $\boldsymbol{Y} = \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$
- Shift-Scaled simulation: $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$,

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ and $\varepsilon_i \sim N(0, \sigma^2)$, $\sigma^2 = 7^2$,

$\delta = 0, 1, 2, ..., 10$, and $\gamma = 0.8, 0.85, 0.9, ..., 1.2$

3.3.1 Shifted simulation: $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$

Table 3.2: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$, where δ=0, 1, 2,..., 10 and $\sigma^2 = 7^2$.

δ	Shift test	Scale test	Shift-Scale test
0	0.0509	0.0499	0.0468
1	0.1625	0.1075	0.1245
2	0.5004	0.2915	0.3826
3	0.8308	0.5644	0.7316
4	0.9720	0.8005	0.9397
5	0.9974	0.9350	0.9921
6	1.0000	0.9846	0.9997
7	1.0000	0.9980	1.0000
8	1.0000	0.9996	1.0000
9	1.0000	1.0000	1.0000
10	1.0000	1.0000	1.0000

For shift alternatives, the Shift test has the higest power, as expected. The Scale test performs the worst, with the Shift-Scale somewhere in between.

3.3.2 Scaled simulation: $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$

Table 3.3: The power of the test using the data simulated from $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\gamma = 0.8, 0.85, 0.9, \ldots, 1.2$ and $\sigma^2 = 7^2$.

γ	Shift test	Scale test	Shift-Scale test
0.80	1.0000	1.0000	1.0000
0.85	0.9992	1.0000	1.0000
0.90	0.9256	0.9978	0.9934
0.95	0.4154	0.6710	0.5517
1.00	0.0509	0.0499	0.0468
1.05	0.4266	0.6632	0.5451
1.10	0.9295	0.9975	0.9917
1.15	0.9982	1.0000	1.0000
1.20	1.0000	1.0000	1.0000

The Scale test performs the best, as expected. The worst performer is the Shift test, with the Shift-Scale somewhere in between.

3.3.3 Shift-Scaled simulation: $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$

Table 3.4: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + 0.9\boldsymbol{X} + \boldsymbol{\varepsilon}$, where δ=0, 1, 2,..., 10, and $\sigma^2 = 7^2$.

δ	Shift test	Scale test	Shift-Scale test
0	0.9256	0.9978	0.9934
1	0.6845	0.9829	0.9631
2	0.3137	0.9060	0.8912
3	0.0780	0.7034	0.8170
4	0.0404	0.3925	0.8008
5	0.1737	0.1438	0.8538
6	0.5205	0.0346	0.9320
7	0.8399	0.0258	0.9816
8	0.9742	0.0920	0.9967
9	0.9978	0.2339	0.9998
10	1.0000	0.4558	1.0000

Note that since $\gamma = 0.9$, the null hypothesis $H_0 : (\beta_0, \beta_1) = (0, 1)$ is always false, so rejecting H_0 is always the correct conclusion. The Shift-Scale test performs the best overall, with hgh power regardless of the shift parameter δ. On the other hand, there are cases where the Shift test has very low power (e.g. $\delta = 4$) and cases where the Scale test has low power (e.g. $\delta = 7$).

3.3.4 Summary

Shift test is optimal for detecting shift alternative models, but performs badly when the data is generated from scale alternatives. The scale test performs best under scale alternatives, and poorly when the data is shifted. The Shift-Scale test is equivalent to fit a simple regression and conduct a 2-df test on both intercept and slope. Our simulation results show that it performs relatively well under either Shift test or Scale test. This means that practitioners can conduct a test that has reasonable power without first conducting a pretest for whether the alternative is shift or scale.

Chapter 4

Clinical Equivalence Testing

The weakness of the approach in Chapter 3 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 as 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 $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$. More details can be found in Lehmann & Romano (2005).

The hypothesis for the two one-sided tests are

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

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:

$$H_{0A} : \theta \leq \theta_1 \quad \text{vs.} \quad H_{1A} : \theta > \theta_1$$

$$\text{and} \qquad H_{0B} : \theta \geq \theta_2 \quad \text{vs.} \quad H_{1B} : \theta < \theta_2$$

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

$$H_0 : |\theta| \geq \Delta \quad \text{vs.} \quad H_1 : |\theta| < \Delta \tag{4.1}$$

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 $H_1 : (\beta_0, \beta_1) = (0, 1)$

4.1 Shift-Equivalence Test $H_1 : \beta_0 = 0$

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

$$H_1 : \mu_d = 0, \quad \text{or}$$

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

We propose two different approaches:

1. Shift-E test: Using TOST (Schuirmann, 1987)
2. Shift-E* test: Using TOST (Westlake, 1972)

4.1.1 Shift-E test: Using TOST (Schuirmann, 1987)

Schuirmann (1987) proposed the two one-sided tests (TOST). They studied the crossover design and set hypotheses to be two separate tests.

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

Under the normality assumption, the statistics are

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

Where $\bar{Y} - \bar{X}$ is the difference of means and s is the square root of the mean squared error from the crossover design. $t_{1-\alpha,\nu}$ is the point that the probability α in the upper tail of the student's t distribution with ν degrees of freedom.

We will reject H_0 if $t_1 \geq t_{1-\alpha,\nu}$ and $t_2 \geq t_{1-\alpha,\nu}$. Then our conclusion for the alternative hypothesis states that both measurements are equivalent.

However, our study is the paired-observations study and our hypothesis is

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

Then our two one-sided tests are

$$H_{0A} : \mu_d \leq -\Delta \quad \text{vs.} \quad H_{1A} : \mu_d > -\Delta$$
$$\text{and} \qquad H_{0B} : \mu_d \geq \Delta \qquad \text{vs.} \quad H_{1B} : \mu_d < \Delta$$

and the statistics are

$$t_1 = \frac{\bar{d} - (-\Delta)}{s_d/\sqrt{n}} \qquad \text{and} \qquad t_2 = \frac{\Delta - \bar{d}}{s_d/\sqrt{n}}.$$

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, $t_1 \geq t_{1-\alpha,\nu}$ and $t_2 \geq t_{1-\alpha,\nu}$.

For example

Now, we apply the Shift-E test using TOST by Schuirmann to test our data. To compare two measurements which are the Westergren ESR ("gold standard") and the STATplus ESR from 48 patients.

Table 4.1: The Westergren/STATplus data with their difference d_i

Patient ID	1	2	3	4	5	6	...	47	48	Mean	s_d
Westergren = x_i	0	16	6	21	15	19	...	40	8	37.40	32.330
STATplus = y_i	5	14	5	10	14	14	...	52	10	45.13	40.196
Difference: $d_i = y_i - x_i$	5	-2	-1	-11	-1	-5	...	12	2	7.73	14.573

Our hypothesis is:

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

Choosing Δ:

In the rest of the paper, we will choose the clinical equivalence criterion to be $\Delta = 0.10\bar{X}$ (which is an estimator of $0.10\mu_X$). This is analogous to defining $\boldsymbol{Y}$ and $\boldsymbol{X}$ as clinically equivalent if $|\mu_Y - \mu_X| \leq 0.10\mu_X$, or $0.90 \leq \frac{\mu_Y}{\mu_X} \leq 1.10$. For the Westergren data, we thus have

$$\Delta = 0.10\overline{X} = 0.10(37.40) = 3.74$$

Thus our equivalence hypothesis is

$$H_0 : |\mu_d| \geq 3.74 \quad \text{vs.} \quad H_1 : |\mu_d| < 3.74$$

and the rejection region is the intersection of rejection regions for the two one-sided tests

$$H_{0A} : \mu_d \leq -3.74 \quad \text{vs.} \quad H_{1A} : \mu_d > -3.74$$
$$\text{and} \qquad H_{0B} : \mu_d \geq 3.74 \qquad \text{vs.} \quad H_{1B} : \mu_d < 3.74$$

Under the normality assumption, our statistics are

$$t_1 = \frac{\bar{d} - (-\Delta)}{s_d/\sqrt{n}} = \frac{7.73 - (-3.74)}{14.573/\sqrt{48}} = 5.452$$
$$t_2 = \frac{\Delta - \bar{d}}{s_d/\sqrt{n}} = \frac{3.74 - 7.73}{14.573/\sqrt{48}} = -1.897.$$

The first statistic $t_1 = 5.452$ is greater than the critical value $t_{1-0.025,47} = 2.011$, so we reject H_{0A}. However, the second statistic $t_2 = -1.897$ is not greater than $t_{1-0.025,47} = 2.011$. So we cannot reject H_{0B}. There is not enough evidence to conclude clinical equivalence.

We would like to show that the shift test does not have size $\alpha = 0.05$.

Proof:

$$\begin{aligned} P[\text{Type I error}] &= P[\text{Rejection region}|H_0] \\ &= P[RR_1|H_0] + P[RR_2|H_0] \\ &= \underbrace{P\left[\frac{\bar{d} - (-3.74)}{s_d/\sqrt{n}} > 2.011 \middle| H_0\right]}_{1} + \underbrace{P\left[\frac{\bar{d} - 3.74}{s_d/\sqrt{n}} < -2.011 \middle| H_0\right]}_{2} \end{aligned}$$

If $\mu_d = 3.74$, the first term will be central t distribution and has value =0.025. The second term will be non-central t distribution. That means non-central t. If $\mu_d = -3.74$, , the second term will be central t distribution and has value =0.025. The first term will be non-central t distribution. If $|\mu_d| > 3.74$, both terms will be non-central t distribution. Then P[Type I error] cannot be equal to 0.05

4.1.2 Shift-E* test: Using TOST (Westlake, 1972)

The Shift test has unknown $\alpha = P$[Type I error]. Westlake (1972) proposed an approach to fix the size α. Based on normality assumptions, $\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$ has the t distribution with $n-1$ degrees of freedom. The 95% probability inequality is $k_2 s_d/\sqrt{n} \leq \bar{d} - \mu_d \leq k_1 s_d/\sqrt{n}$. Where $k_1,\ k_2$ can be evaluated from the integral of the t distribution from k_2 to k_1 is 0.95. He claimed that "In normal statistical practice, k_1 is chosen to be equal to $-k_2$." Let $k_1 = -k_2 = t_\gamma$. Then the confidence interval for $\bar{d}$ would be

$$\left(\bar{d} - t_\gamma \frac{s_d}{\sqrt{n}}\ ,\ \bar{d} + t_\gamma \frac{s_d}{\sqrt{n}}\right)$$

Westlake described the confidence interval approach, but he did not explain how to calculate the rejection region. Here, we describe one way to calculate t_γ and the rejection region so that P[Type I error]=0.05. The difficulty lies in the fact that the size of the test is evaluated under the null region $|\mu_d| \geq \Delta$, and does not contain $\mu_d = 0$. So, the power function under the null always involves noncentral t. Below, we show that (i) the power function of μ_d is symmetric about 0, (ii) the maximum power occurs at 0, and (iii) the supremum of the power function under H_0 occurs at $\mu_d = \Delta$. Then we can calculate t_γ from the integral of the t distribution.

(i) We prove that the power function is symmetric about 0. The rejection region is

of the form $|\bar{d}| < C_0$, where $C_0 = t_\gamma s_{\bar{d}}$. So the power function of the test is

$$\gamma(\mu_d) = P_{\mu_d}\left[|\bar{d}| < C_0\right], \quad \mu_d \in \Omega = \Omega_0 \cup \Omega_1$$

The size of the test is the maximum power under the null region. Hence

$$\alpha = \sup_{\mu_d \in \Omega_0} \gamma(\mu_d) = \sup_{\mu_d \in \Omega_0} P\left[|\bar{d}| < C_0\right]$$

We prove that the power function is symmetric about 0 by showing $\gamma(\mu_d) = \gamma(-\mu_d)$. Let Δ be an arbitrary nonnegative value.

$$\begin{aligned}
\gamma(\Delta) &= P_\Delta\left[|\bar{d}| < C_0\right] \\
&= P\left(\frac{-C_0 - \Delta}{s_{\bar{d}}} < \frac{\bar{d} - \Delta}{s_{\bar{d}}} < \frac{C_0 - \Delta}{s_{\bar{d}}}\right) \\
&= P\left(\frac{-t_\gamma s_{\bar{d}} - \Delta}{s_{\bar{d}}} < T < \frac{t_\gamma s_{\bar{d}} - \Delta}{s_{\bar{d}}}\right) \\
&= P\left(-t_\gamma - \frac{\Delta}{s_{\bar{d}}} < T < t_\gamma - \frac{\Delta}{s_{\bar{d}}}\right) \\
\text{Let } a = \frac{\Delta_0}{s_{\bar{d}}} &\text{ and } b = t_\gamma \\
\gamma(\Delta_0) &= P\left(-a - b < T < -a + b\right) = F(-a+b) - F(-a-b),
\end{aligned}$$

where F is the cumulative distribution of standard t distribution.

If $\mu_d = -\Delta$, the power is given by

$$\gamma(-\Delta) = P\Big(a - b < T < a + b\Big) = F(a+b) - F(a-b).$$

Since $F(a+b) - F(a-b) = F(-a+b) - F(-a-b)$, then $\gamma(\Delta) = \gamma(-\Delta)$. Therefore,

the power function is symmetric about 0.

(ii) To prove the maximum power occurs at 0.

$$\begin{aligned}
\gamma(0) &= F(b) - F(-b) = 1 - 2F(-b) \\
\gamma(\Delta) &= F(-a+b) - F(-a-b) \\
&= \underbrace{1 - F(a-b)}_{\leq 1-F(-b)} - \underbrace{F(-a-b)}_{\leq F(-b)} \\
&\leq 1 - F(-b) - F(-b) \\
&= 1 - 2F(-b)
\end{aligned}$$

So, the maximum power occurs at 0.

(iii) If μ_d moves away from 0, the power will decrease. Then the size of our test can be appraised at Δ or $-\Delta$.

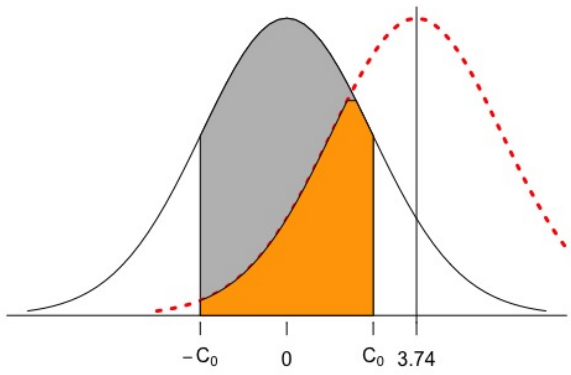

Figure 4.1: P(Reject H_0) when $\mu_d = 3.74$

Figure 4.1 shows that when μ_d shifts to $\Delta = 3.74$. The orange area is the probability of rejection region under $H_0 : \mu_d = 3.74$. If we move μ_d far from 3.74, the

orange area will be smaller. The figure shows that the size of the equivalence test will be evaluated at $\Delta = 3.74$. Moreover, our hypothesis is symmetric about 0. So, the size of our test will be evaluated at $-\Delta = -3.74$ also. Therefore, we will consider only the size at $\Delta = 3.74$.

The supremum of the probability of rejection region under H_0 occurs at $\mu_d = \Delta$

$$\begin{aligned}
\alpha &= \sup_{\mu_d \in \Omega_0} \gamma(\mu_d) = P_\Delta \left[|\bar{d}| < C_0 \right] \\
&= P\left(\frac{-C_0 - \Delta}{s_{\bar{d}}} < \frac{\bar{d} - \Delta}{s_{\bar{d}}} < \frac{C_0 - \Delta}{s_{\bar{d}}} \right), \text{ let } C_0 = t_\gamma s_{\bar{d}} \\
&= P\left(\frac{-t_\gamma s_{\bar{d}} - \Delta}{s_{\bar{d}}} < T < \frac{t_\gamma s_{\bar{d}} - \Delta}{s_{\bar{d}}} \right) \\
&= P\left(-t_\gamma - \frac{\Delta}{s_{\bar{d}}} < T < t_\gamma - \frac{\Delta}{s_{\bar{d}}} \right) \\
&= 0.05
\end{aligned} \tag{4.2}$$

We solve equation (4.2) to find the critical value (t_γ). We know that T has the t distribution with degrees of freedom $\nu = n - 1$. Therefore,

$$\alpha = \int_{-t_\gamma - \frac{\Delta}{s_{\bar{d}}}}^{t_\gamma - \frac{\Delta}{s_{\bar{d}}}} f(t)dt = 0.05, \qquad \text{where}$$

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \; -\infty < t < \infty$$

We get critical value t_γ and our critical region is $(-C_0,\ C_0) = (-t_\gamma s_{\bar{d}},\ t_\gamma s_{\bar{d}})$. If $\bar{d}$ is in the critical region, we can conclude that the two measurements are equivalent.

For example

To illustrate the Shift-E* test using TOST by Westlake approach, we consider a previous example. We have $\bar{d} = 7.73,\ s_d = 14.573$ and $s_{\bar{d}} = s_d/\sqrt{n} = 2.103$. After

solving equation (4.2), we get $t_\gamma = 0.292$ and our critical region is $(-t_\gamma s_{\bar{d}},\ t_\gamma s_{\bar{d}}) = (-0.6143, 0.6143)$. Our $\bar{d}$ is not in the critical region, so we cannot reject null hypothesis. It means that two measurements are not equivalent.

4.2 Scale-Equivalence Test $H_1 : \beta_1 = 1$

The equivalence testing can be constructed the alternative hypothesis as

$$H_1 : \mu_Y/\mu_X = 1, \qquad \text{or}$$

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

where Δ_1 is small interval about 1. We have two different approaches to study.

1. Scale-E test: Using log TOST by Berger et al. approach.

2. Scale-E* test: Using log TOST by Westlake approach to fix size of the test α.

4.2.1 Scale-E test: Using log TOST (Berger et al., 1996)

Berger et al. (1996) developed the Schuirmann (1987) approach (TOST) by using logarithmic transformation. The original alternative hypothesis is

$$H_1 :\ A_0 < \mu_Y/\mu_X < B_0,$$

where μ_Y and μ_X are the population means of the experimental measurement and the standard measurement, respectively.

After the logarithmic transformation, the alternative hypothesis is

$$H_1 :\ A < \eta_Y - \eta_X < B,$$

where $A = \log A_0,\ B = \log B_0,\ \eta_Y = \log \mu_Y$ and $\eta_X = \log \mu_X$.

Then it can be divided into two hypotheses.

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

$$\text{and} \qquad H_{0B} : \eta_Y - \eta_X \geq B \quad \text{vs.} \quad H_{1B} : \eta_Y - \eta_X < B$$

Under the normality assumption, the statistics are

$$t_l = \frac{\bar{d} - A}{s_{\bar{d}}} \qquad \text{and} \qquad t_u = \frac{\bar{d} - B}{s_{\bar{d}}},$$

where $\bar{d}$ is an estimate of $\eta_Y - \eta_X$ and $s_{\bar{d}}$ is standard error of difference between y and x (in logarithmic scale). It rejects H_0 at level α and declares the two measurements to be equivalent if both tests reject, that is, $t_l \geq t_{\alpha,\nu}$ and $t_u \leq -t_{\alpha,\nu}$.

However, we consider the paired testing. Then our hypothesis is

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

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

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

It is two one-sided tests.

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

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

$$\text{where} \qquad A = \log \Delta_1 \text{ and } \ B = -A = \log 1/\Delta_1$$

Under the normality assumption that the statistics are

$$t_u = \frac{\bar{d} - A}{s_d/\sqrt{n}} \qquad \text{and} \qquad t_l = \frac{\bar{d} - B}{s_d/\sqrt{n}}.$$

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, $t_l \geq t_{\alpha,\nu}$ and $t_u \leq -t_{\alpha,\nu}$.

For example

We transform our data in the previous example with logarithmic scale.

Table 4.2: Log-transformated data.

Patient ID	1	2	3	4	...	47	48	Mean	s_d
$\log X = \log x_i$	-0.69	2.77	1.79	3.04	...	3.69	2.08	3.026	1.370
$\log Y = \log y_i$	1.61	2.64	1.61	2.30	...	3.95	2.30	3.257	1.212
$d_i = \log y_i - \log x_i$	2.30	-0.13	-0.18	-0.74	...	0.26	0.22	0.232	0.5453

To test Scale-E, our hypothesis is

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

We can consider it to be two separate tests:

$$H_{0A} : \eta_d \geq A \quad \text{vs.} \quad H_{1A} : \eta_d < A$$
$$\text{and} \qquad H_{0B} : \eta_d \leq B \quad \text{vs.} \quad H_{1B} : \eta_d > B,$$

where $A = \log \Delta_1 = \log 1.1 = 0.0953$ and $B = -A = -0.0953$.

To illustrate Scale-E test, we have $\bar{d} = 0.232,\ s_d = 0.5453$ (in logarithmic scale).

Then we calculate our statistics:

$$t_u = \frac{\bar{d} - A}{s_d/\sqrt{n}} = \frac{0.232 - 0.0953}{0.5453/\sqrt{48}} = 1.736$$
$$t_l = \frac{\bar{d} - B}{s_d/\sqrt{n}} = \frac{0.232 - (-0.0953)}{0.5453/\sqrt{48}} = 4.158$$

Since $t_u = 1.736$ is not less than $t_{0.025,47} = -2.011$ and $t_l = 4.158$ is greater than $-t_{0.025,47} = 2.011$. So we cannot reject H_{0B}. There is not enough evidence to conclude clinical equivalence.

4.2.2 Scale-E* test: Using log TOST (Westlake, 1972) after logarithmic transformation

We modify the Shift-E* with the logarithmic transformated data, when our alternative hypothesis is

$$H_1 : |\eta_d| < \log \Delta_1$$

The power function of the hypothesis test with rejection is

$$\gamma(\eta_d) = P_{\eta_d}\left[|\bar{d}| < C_1\right], \quad \eta_d \in \Omega = \Omega_0 \cup \Omega_1$$

We can find the critical region based on the type I error is 0.05. That is the size

$$\alpha = \sup_{\eta_d \in \Omega_0} \gamma(\mu_d) = \sup_{\eta_d \in \Omega_0} P\left[|\bar{d}| < C_1\right]$$

By the similar proof, we know that the power of η_d is symmetric about 0 and the maximum power occurs at 0. Then the supremum of the probability of rejection region

under H_0 at $\eta_d = \Delta^*$, where $\Delta^* = \log \Delta_1$, is

$$\begin{aligned}
\alpha &= \sup_{\eta_d \in \Omega_0} \gamma(\mu_d) = P_{\Delta^*}\left[|\bar{d}| < C_1\right] \\
&= P\left(\frac{-C_1 - \Delta^*}{s_{\bar{d}}} < \frac{\bar{d} - \Delta}{s_{\bar{d}}} < \frac{C_1 - \Delta^*}{s_{\bar{d}}}\right), \text{ let } C_1 = t_\gamma s_{\bar{d}} \\
&= P\left(-t_\gamma - \frac{\Delta^*}{s_{\bar{d}}} < T < t_\gamma - \frac{\Delta^*}{s_{\bar{d}}}\right) \\
&= 0.05
\end{aligned} \tag{4.3}$$

We solve equation (4.3) to find the critical value (t_γ). We know that T has the t distribution with degrees of freedom $\nu = n - 1$. Therefore,

$$\alpha = \int_{-t_\gamma - \frac{\Delta^*}{s_{\bar{d}}}}^{t_\gamma - \frac{\Delta^*}{s_{\bar{d}}}} f(t)dt = 0.05$$

We will get critical value t_γ and our critical region is $(-C_1,\ C_1) = (-t_\gamma s_{\bar{d}},\ t_\gamma s_{\bar{d}})$. If $\bar{d}$ is in the critical region, we can conclude that the two measurements are equivalent.

For example

We apply the Scale-E* test with is logarithmic transformation data. We get $t_\gamma = 0.1525$, after solving equation (4.3). Here our critical region is $(-t_\gamma s_{\bar{d}},\ t_\gamma s_{\bar{d}}) = (-0.0109, 0.0109)$. Our $\bar{d}$ is not in the critical region, so we cannot reject H_0. It means that two measurements are not equivalent.

Chapter 5

Shift-Scale-Equivalence Test

$H_1 : (\beta_0, \beta_1) = (0, 1)$

Recall that the Shift-Scale model is equivalent to fitting a simple regression and conducting a 2-df test on both intercept and slope. Here we propose a procedure for doing clinical equivalence hypothesis testing:

$$H_0 : |\beta_0| \geq \Delta_0 \ \cup \ |\beta_1 - 1| \geq \Delta_1 \quad \text{vs.} \quad H_1 : |\beta_0| < \Delta_0 \ \cap \ |\beta_1 - 1| < \Delta_1$$

where Δ_0 and Δ_1 are investigator-specified values that represent limits of allowable clinical dissimilarities. In the examples and simulations that follow, we used $\Delta_0 = .10\overline{X} = 3.74$ and $\Delta_1 = 0.10$ so that our alternative hypothesis test representing equivalence is effectively

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

This *alternative region* may be represented by a rectangle drawn in Figure 5.1. The null region is the region outside the rectangle.

We propose a rejection region of the form

$$\begin{aligned} RR &\equiv \left\{ |\hat{\beta}_0| < C_0 \ \cap \ |\hat{\beta}_1 - 1| < C_1 \right\} \\ &\equiv \left\{ |\hat{\beta}_0| < t_\gamma s_{b0} \ \cap \ |\hat{\beta}_1 - 1| < t_\gamma s_{b1} \right\} \end{aligned}$$

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

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

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

$$\begin{aligned} \alpha &= \sup_{\beta \in \Omega_0} \gamma(\beta) = \sup_{\beta \in \Omega_0} P \left[|\hat{\beta}_0| < C_0, \ |\hat{\beta}_1 - 1| < C_1 \right], \quad \text{where } C_0 = t_\gamma s_{b0}, \ C_1 = t_\gamma s_{b1} \\ &= \sup_{\beta \in \Omega_0} P \left[|\hat{\beta}_0| < t_\gamma s_{b0}, \ |\hat{\beta}_1 - 1| < t_\gamma s_{b1} \right] \end{aligned}$$

Since the rejection region is symmetric about $(0, 1)$ which is the center of the rectangle in Figure 5.1, 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 simulations in the next chapter shows that the test does achieve the desired size, and confirms that the maximum power occurs at the boundary.

The steps to find the t_γ to satisfy the size of the test α =0.05.

Our alternative hypothesis is $H_1 : |\beta_0| < \Delta_0 \ \cap \ |\beta_1 - 1| < \Delta_1$.

1. Calculate statistics: $\hat{\beta}_0,\ s_{b0},\ \hat{\beta}_1,\ s_{b1}, cov(\hat{\beta}_0, \hat{\beta}_1) = s_{b01}$

2. Rejection region is a rectangle/intersection

$$\begin{aligned} RR &\equiv \left\{ |\hat{\beta}_0| < C_0 \ \cap \ |\hat{\beta}_1 - 1| < C_1 \right\} \\ &\equiv |\hat{\beta}_0| < t_\gamma s_{b0},\ \ |\hat{\beta}_1 - 1| < t_\gamma s_{b1} \end{aligned}$$

3. Find the t_γ so that

$$\begin{aligned} \alpha &= \sup_{(\beta_0,\, \beta_1) \in H_0} P(RR) \leq 0.05, \ \text{ where } H_0 : |\beta_0| \geq \Delta_0 \ \cup \ |\beta_1 - 1| \geq \Delta_1 \\ P(RR) &= P\left[|\hat{\beta}_0| < t_\gamma s_{b0},\ |\hat{\beta}_1 - 1| < t_\gamma s_{b1} \right] \\ &= P\left[-t_\gamma s_{b0} < \hat{\beta}_0 < t_\gamma s_{b0}, \quad 1 - t_\gamma s_{b1} < \hat{\beta}_1 < 1 + t_\gamma s_{b1} \right] \end{aligned}$$

We know that $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{bmatrix} \sigma_{b0}^2 & \sigma_{b01} \\ \sigma_{b01} & \sigma_{b1}^2 \end{bmatrix} \right)$ and $\rho_b = \frac{\sigma_{b01}}{\sigma_{b0}\sigma_{b1}}$

$$f\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{1}{2\pi\sigma_{b0}\sigma_{b1}\sqrt{1-\rho_b^2}} e^{-\frac{1}{2(1-\rho_b^2)}\left[\left(\frac{y_0-\beta_0}{\sigma_{b0}}\right)^2 + \left(\frac{y_1-\beta_1}{\sigma_{b1}}\right)^2 - 2\rho\left(\frac{y_0-\beta_0}{\sigma_{b0}}\right)\left(\frac{y_1-\beta_1}{\sigma_{b1}}\right)\right]},$$

where $\sigma_{b0}, \sigma_{b1}, \sigma_{b01}$ and ρ_b are unknown. So we use their estimators: s_{b0}, s_{b1}, s_{b01} and r_b.

Then

$$P(RR) = \int_{1-t_\gamma s_{b1}}^{1+t_\gamma s_{b1}} \int_{-t_\gamma s_{b0}}^{t_\gamma s_{b0}} f\left(\hat{\beta}_0, \hat{\beta}_1\right) dy_0 dy_1$$

$$= \int_{1-t_\gamma s_{b1}}^{1+t_\gamma s_{b1}} \int_{-t_\gamma s_{b0}}^{t_\gamma s_{b0}} \frac{1}{2\pi s_{b0} s_{b1}\sqrt{1-r_b^2}} e^{-\frac{1}{2(1-r_b^2)}\left[\left(\frac{y_0-\beta_0}{s_{b0}}\right)^2+\left(\frac{y_1-\beta_1}{s_{b1}}\right)^2-2\rho\left(\frac{y_0-\beta_0}{s_{b0}}\right)\left(\frac{y_1-\beta_1}{s_{b1}}\right)\right]} dy_0 dy_1$$

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

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

From the previous section, we know that the supremum occurs at the boundary of the test since the distribution of the test is symmetric and unimodal. We imply the previous section, and find the $t_{\gamma 1}$ and $t_{\gamma 2}$.

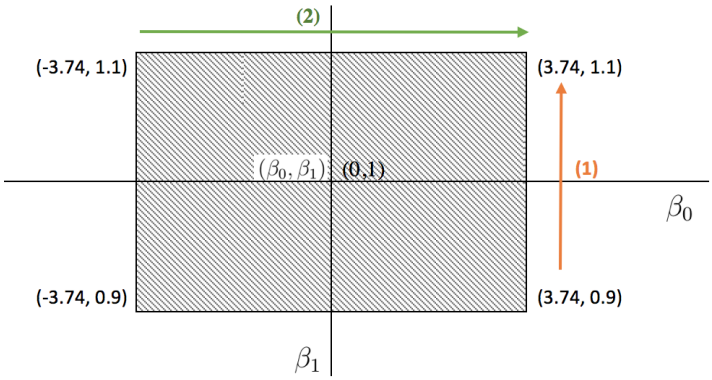

Figure 5.1: The rejection region

Figure 5.1 shows that $t_{\gamma 1}$ is the critical value when we fix $\beta_0 = \Delta_0$ to have size 0.05 (along the red line). $t_{\gamma 2}$ is the critical value when we fix $\beta_1 = 1 + \Delta_1$ to have size 0.05 (along the green line). t_γ is the minimum of ($t_{\gamma 1}$ and $t_{\gamma 2}$) in order to satisfy the size α=0.05

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

We fix $\beta_0 = \Delta_0$ and evaluate $P_{\beta_1}[RR] = P_{\beta_1}[RR|\beta_0 = \Delta_0] \equiv h(\beta_1)$

$$h(\beta_1) = \int_{1-t_\gamma s_{b1}}^{1+t_\gamma s_{b1}} \int_{-t_\gamma s_{b0}}^{t_\gamma s_{b0}} \frac{1}{2\pi s_{b0} s_{b1}\sqrt{1-r_b^2}} e^{-\frac{1}{2(1-r_b^2)}\left[\left(\frac{y_0-\Delta_0}{s_{b0}}\right)^2+\left(\frac{y_1-\beta_1}{s_{b1}}\right)^2-2\rho\left(\frac{y_0-\Delta_0}{s_{b0}}\right)\left(\frac{y_1-\beta_1}{s_{b1}}\right)\right]} dy_0 dy_1$$
$$\leq 0.05 \tag{5.1}$$

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

$$g(\beta_1) = \frac{\partial}{\partial \beta_1} h(\beta_1)$$
$$= \int_{1-t_\gamma s_{b1}}^{1+t_\gamma s_{b1}} \int_{-t_\gamma s_{b0}}^{t_\gamma s_{b0}} \frac{1}{2\pi s_{b0} s_{b1}\sqrt{1-r_b^2}} e^{-\frac{1}{2(1-r_b^2)}\left[\left(\frac{y_0-\Delta_0}{s_{b0}}\right)^2+\left(\frac{y_1-\beta_1}{s_{b1}}\right)^2-2r_b\left(\frac{y_0-\Delta_0}{s_{b0}}\right)\left(\frac{y_1-\beta_1}{s_{b1}}\right)\right]}$$
$$\frac{1}{(1-r_b^2)s_{b1}}\left[\left(\frac{y_1-\beta_1}{s_{b1}}\right) - r_b\left(\frac{y_0-\Delta_0}{s_{b0}}\right)\right] dy_0\, dy_1$$
$$\equiv 0 \tag{5.2}$$

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

(i) Start with $\beta_1^{(0)} = \beta_1$. Calculate $t_{\gamma 1}^{(1)}$ from equation (5.1).

(ii) Use $t_{\gamma 1}^{(1)}$ and equation (5.2) 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 $P_{\beta_0}[RR] = P_{\beta_0}[RR|\beta_1 = \Delta_1] \equiv h(\beta_0)$

$$h(\beta_0) = \int_{1-t_\gamma s_{b1}}^{1+t_\gamma s_{b1}} \int_{-t_\gamma s_{b0}}^{t_\gamma s_{b0}} \frac{1}{2\pi s_{b0} s_{b1}\sqrt{1-r_b^2}} e^{-\frac{1}{2(1-r_b^2)}\left[\left(\frac{y_0-\beta_0}{s_{b0}}\right)^2+\left(\frac{y_1-\Delta_1}{s_{b1}}\right)^2-2r_b\left(\frac{y_0-\beta_0}{s_{b0}}\right)\left(\frac{y_1-\Delta_1}{s_{b1}}\right)\right]} dy_0 dy_1$$
$$\leq 0.05. \tag{5.3}$$

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

$$\begin{aligned} g(\beta_0) &= \frac{\partial}{\partial \beta_0} h(\beta_0) \\ &= \int_{1-t_\gamma s_{b1}}^{1+t_\gamma s_{b1}} \int_{-t_\gamma s_{b0}}^{t_\gamma s_{b0}} \frac{1}{2\pi s_{b0} s_{b1}\sqrt{1-r_b^2}} e^{-\frac{1}{2(1-r_b^2)}\left[\left(\frac{y_0-\beta_0}{s_{b0}}\right)^2+\left(\frac{y_1-\Delta_1}{s_{b1}}\right)^2-2r_b\left(\frac{y_0-\beta_0}{s_{b0}}\right)\left(\frac{y_1-\Delta_1}{s_{b1}}\right)\right]} \\ &\quad \frac{1}{(1-r_b^2)s_{b1}}\left[\left(\frac{y_0-\beta_0}{s_{b0}}\right) - r_b\left(\frac{y_1-\Delta_1}{s_{b1}}\right)\right] dy_0\, dy_1 \\ &\equiv 0 \end{aligned} \tag{5.4}$$

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

(i) Start with $\beta_0^{(0)} = \beta_0$. Calculate $t_{\gamma 2}^{(1)}$ from equation (5.3).

(ii) Use $t_{\gamma 2}^{(1)}$ and equation (5.4) 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 $C_0 = t_\gamma s_{\hat{\beta}_0},\ C_1 = t_\gamma s_{\hat{\beta}_1}$. 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.

For example

To illustrate this method, we get $\hat{\beta} = (1.3225,\ 1.1713)^T$, $\hat{\sigma}^2 = 185.64$, $s_{b0} = 3.0252$, $s_{b1} = 0.0614$, $s_{b01} = -0.1413$, $r_b = \frac{-0.1413}{3.0252 * 0.0614} = -0.75988$.

Suppose that we choose $\Delta_0 = 0.10 * \bar{x} = 3.74$, (which is an estimator of $0.10\mu_X$) that means Δ_0 is 10% of the Westergren, and $\Delta_1 = 0.1$.

Calculate the critical value $t_\gamma = \min(t_{\gamma 1}, t_{\gamma 2})$

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

- Start with $\beta_1^{(0)} = 1 - \Delta_1 = 0.9$.

  Then calculate $t_{\gamma 1}^{(1)}$ from equation (5.1). We get $t_{\gamma 1}^{(1)} = 0.4525$.
- Use $t_{\gamma 1}^{(1)} = 0.4525$ and equation (5.2) to get $\beta_1^{(1)} = 0.9445$.
- Repeat (i) and (ii) until $\left|t_{\gamma 1}^{(i+1)} - t_{\gamma 1}^{(i)}\right| < 0.001$. We get $t_{\gamma 1} = 0.3408$.

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

- Start with $\beta_0^{(0)} = \Delta_0 = 3.74$.

  Then calculate $t_{\gamma 2}^{(1)}$ from equation (5.3). We get $t_{\gamma 2}^{(1)} = 0.4525$.
- Use $t_{\gamma 2}^{(1)} = 0.4525$ and equation(5.4) to get $\beta_0^{(1)} = 3.8117$.
- Repeat (i) and (ii) until $\left|t_{\gamma 2}^{(i+1)} - t_{\gamma 2}^{(i)}\right| < 0.001$. We get $t_{\gamma 2} = 0.4536$.

We get critical value $t_\gamma = \min(t_{\gamma 1},\ t_{\gamma 2}) = \min(0.3408, 0.4536) = 0.3408$ and $C_0 = t_\gamma s_{\hat{\beta}_0} = 0.3408 * 3.0252 = 1.031$, $C_1 = t_\gamma s_{\hat{\beta}_1} = 0.3408 * 0.0614 = 0.021$. We found

that the value of the test statistic $\hat{\beta}_0 = 1.323$ does not fall in the rejection region $(-C_0,\ C_0) = (-1.031, 1.031)$, and also $\hat{\beta}_1 = 1.713$ does not fall in the rejection region $(1 - C_1,\ 1 + C_1) = (0.979, 1.021)$. Then we cannot reject null hypothesis and conclude that the two measurements are not equivalent.

Chapter 6

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 $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the Westergren data, for each of the following cases:

1. Shifted simulation: $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$ (See table 6.1)

2. Scaled simulation: $\boldsymbol{Y} = \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$ (See table 6.2)

3. Shift-Scaled simulation: $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$ (See table 6.3)

   where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ and $\varepsilon_i \sim N(0, \sigma^2)$, $\sigma^2 = 3^2, 5^2, 7^2, 9^2$ and 14^2

   δ = -6, -5, -4, ..., 6. and γ = 0.7, 0.8, 0.9, 1, 1.1, 1.2.

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

   - Shift-Equivalence test $H_1 : \beta_0 = 0$
     - Shift-E test: Using TOST (Schuirmann, 1987).
     - Shift-E* test: Using TOST (Westlake, 1972).

- Scale-Equivalence test $H_1 : \beta_1 = 1$
  - Scale-E test: Using log TOST (Berger et al., 1996).
  - Scale-E* test: Using log TOST (Westlake, 1972).
- Shift-Scale-Equivalence test: $H_1 : (\beta_0, \beta_1) = (0, 1)$

6.1 Shifted Simulation: $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$

Table 6.1: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$, where δ=-4, -3.90, -3.80, ..., 4, and $\sigma^2 = 3^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-4.00	0.9163	0.0123	0.1512	0.0000	0.0240
	-3.90	0.9439	0.0222	0.1846	0.0000	0.0350
	-3.80	0.9662	0.0384	0.2270	0.0000	0.0510
H_1	-3.70	0.9791	0.0604	0.2704	0.0000	0.0640
	-3.60	0.9875	0.0941	0.3237	0.0000	0.0860
	-3.50	0.9920	0.1379	0.3774	0.0000	0.1060
	-3.00	0.9997	0.5256	0.6691	0.0000	0.3050
	-2.00	1.0000	0.9883	0.9687	0.0138	0.8330
	-1.00	1.0000	1.0000	0.9997	0.1568	0.9890
	0.00	1.0000	1.0000	1.0000	0.2705	1.0000
	1.00	1.0000	1.0000	0.9985	0.0828	0.9920
	2.00	1.0000	0.9908	0.9277	0.0081	0.8260
	3.00	0.9999	0.5109	0.4802	0.0003	0.2960
	3.50	0.9935	0.1328	0.2404	0.0000	0.0990
	3.60	0.9888	0.0919	0.2096	0.0000	0.0820
	3.70	0.9802	0.0617	0.1813	0.0000	0.0690
H_0	3.80	0.9650	0.0370	0.1549	0.0000	0.0540
	3.90	0.9455	0.0235	0.1292	0.0000	0.0340
	4.00	0.9130	0.0144	0.1043	0.0000	0.0260

The Shift-E* test achieves the correct size, and the $\sup P[\text{RR}]$ occurs at $\delta = 3.74$ and -3.74 which are on the boundary of the H_0 region. In comparison, the Shift-E test

does not hold its size. The Shift-Scale-E test holds its size but has lower power than the Shift-E* test.

We show more tables where $\sigma^2 = 5^2, 7^2, 9^2$ and 14^2 in an appendix A.1 Shift simulation for clinical equivalence testing.

6.2 Scaled Simulations: $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$

Table 6.2: The power of the test using the data simulated from $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$ where $\gamma = 0.8, 0.85, 0.9, ..., 1.2$, and $\sigma^2 = 3^2$.

	γ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	0.80	0.0001	0.0000	0.5979	0.0002	0.0000
	0.85	0.3196	0.0000	0.8801	0.0044	0.0000
	0.90	0.9978	0.0083	0.9874	0.0597	0.0002
H_1	0.92	1.0000	0.3187	0.9961	0.1130	0.0256
	0.94	1.0000	0.9260	0.9991	0.1771	0.3260
	0.96	1.0000	0.9994	1.0000	0.2313	0.8423
	0.98	1.0000	1.0000	1.0000	0.2696	0.9945
	1.00	1.0000	1.0000	1.0000	0.2705	0.9997
	1.02	1.0000	1.0000	1.0000	0.2368	0.9943
	1.04	1.0000	0.9998	0.9998	0.1847	0.8548
	1.06	1.0000	0.9242	0.9982	0.1320	0.3402
	1.08	0.9999	0.3177	0.9930	0.0813	0.0286
H_0	1.10	0.9982	0.0088	0.9786	0.0489	0.0004
	1.15	0.3061	0.0000	0.8243	0.0097	0.0000
	1.20	0.0000	0.0000	0.5413	0.0032	0.0000

The Scale-E* test achieve the correct size, and the sup $P[\text{RR}]$ occurs at $\gamma = 0.90$ and 1.10 which is on the boundary of the H_0 region. But it does not perform well as it has low power. In comparison, the Shift-E* test hold its size and power. The Shift-Scale-E test has lower power than the Shift-E* test.

We show more tables where $\sigma^2 = 5^2, 7^2, 9^2$ and 14^2 in an appendix A.2 Scaled simulation for clinical equivalence testing.

6.3 Shift-Scaled Simulations: $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$

Table 6.3: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\gamma = 0.98$, $\delta = -4, -3, -2, ..., 4$ and $\sigma^2 = 3^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-4.00	0.3950	0.0000	0.0710	0.0000	0.0260
	-3.90	0.4980	0.0000	0.0900	0.0000	0.0430
	-3.80	0.5950	0.0000	0.1060	0.0000	0.0530
H_1	-3.70	0.6740	0.0000	0.1260	0.0000	0.0710
	-3.60	0.7500	0.0000	0.1670	0.0000	0.0910
	-3.50	0.8200	0.0000	0.2090	0.0000	0.1140
	-3.00	0.9690	0.0380	0.4460	0.0000	0.3020
	-2.00	1.0000	0.7390	0.9010	0.0050	0.8030
	-1.00	1.0000	0.9980	0.9950	0.1000	0.9750
	0.00	1.0000	1.0000	1.0000	0.2780	0.9900
	1.00	1.0000	1.0000	1.0000	0.1470	0.9820
	2.00	1.0000	1.0000	0.9780	0.0110	0.7930
	3.00	1.0000	0.9650	0.6650	0.0020	0.3040
	3.50	1.0000	0.7120	0.3680	0.0000	0.1250
	3.60	1.0000	0.6320	0.3230	0.0000	0.0990
	3.70	1.0000	0.5310	0.2800	0.0010	0.0790
H_0	3.80	1.0000	0.4290	0.2430	0.0000	0.0610
	3.90	1.0000	0.3480	0.2160	0.0000	0.0470
	4.00	1.0000	0.2680	0.1950	0.0000	0.0300

For this situation, Shift-Scale-E test can hold its size, while other tests cannot do. Even though, Shift-Scale-E test has lower power than others tests.

We show more tables in an appendix A.3 Shift-Scaled simulation.

Chapter 7

Conclusion and Future Work

Our simulations show that the uncorrected shift-E test not have correct P[Type I error], even under shift-type alternatives for which it was designed. The corrected Shift-E* version of the test does better and has correct size and reasonable power. The Shift-Scale-E tests also have the correct size, but is not as powerful as Shift-E* under shift-type distributions.

For simulations under scaled alternatives, the Shift-E*, Scale-E* test and Shift-Scale-E test have good power and size. However, for simulations under simultaneous shift and scale alternatives, only the Shift-Scale-E works, as expected.

The Shift-Scale-E test remains valid under all conditions. The Shift-E* performs best under shift-type alternatives, but can do really badly under shift-scale alternatives (See table 6.3)

Future work

We have done simulations under limited distribution patterns and alternative structures. We need to expand the simulations to account for more patterns. Also we can do a study of required sample size to achieve certain power of the tests. Clinical equiv-

alence approach may not be workable or practical for some combinations of sample size (or standard error) and Δ, because the resulting rejection regions may be too small. Also, we can investigate diagnostics to use as pretest for shift or scale type relationships.

Appendix A

Clinical Equivalence Testing Simulations

The below tables illustrate the performance of five different approaches:

- Shift-E: Shift-Equivalence test using TOST (Schuirmann, 1987)
- Shift-E*: Shift-Equivalence test using TOST (Westlake, 1972)
- Scale-E: Scale-Equivalence test using TOST (Berger et al., 1996)
- Scale-E*: Scale-Equivalence test using TOST (Westlake, 1972)
- Shift-Scale-Equivalence test

A.1 Shifted Simulations

Table A.1: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$, where δ=-6, -5, -4, ..., 6, and $\sigma^2 = 5^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-6.00	0.1374	0.0000	0.0331	0.0000	0.0001
	-5.00	0.6049	0.0001	0.1840	0.0000	0.0006
	-4.00	0.9453	0.0226	0.5201	0.0002	0.0191
H_1	-3.00	0.9985	0.2630	0.8356	0.0026	0.1378
	-2.00	1.0000	0.7666	0.9714	0.0253	0.4347
	-1.00	1.0000	0.9807	0.9981	0.0871	0.7483
	0.00	1.0000	0.9998	0.9998	0.1230	0.8628
	1.00	1.0000	0.9824	0.9969	0.0784	0.7348
	2.00	1.0000	0.7625	0.9655	0.0237	0.4192
	3.00	0.9991	0.2601	0.8042	0.0047	0.1337
	3.50	0.9891	0.0938	0.6570	0.0015	0.0562
	3.60	0.9844	0.0735	0.6222	0.0015	0.0462
	3.70	0.9776	0.0575	0.5901	0.0016	0.0372
H_0	3.80	0.9690	0.0424	0.5553	0.0012	0.0292
	3.90	0.9583	0.0312	0.5199	0.0006	0.0236
	4.00	0.9461	0.0238	0.4885	0.0006	0.0182
	5.00	0.5958	0.0004	0.2017	0.0001	0.0008
	6.00	0.1339	0.0000	0.0651	0.0000	0.0000

Table A.2: The power of the test using the data simulated from $\boldsymbol{Y} = \delta + \boldsymbol{X} + \varepsilon$, where δ=-6, -5, -4, ..., 6, and $\sigma^2 = 7^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-6.00	0.4112	0.0001	0.1870	0.0000	0.0000
	-5.00	0.7750	0.0013	0.4278	0.0003	0.0006
	-4.00	0.9559	0.0301	0.6932	0.0013	0.0071
H_1	-3.00	0.9955	0.1786	0.8831	0.0081	0.0505
	-2.00	0.9997	0.5311	0.9695	0.0323	0.1826
	-1.00	1.0000	0.8485	0.9942	0.0695	0.3783
	0.00	1.0000	0.9536	0.9988	0.0886	0.4654
	1.00	1.0000	0.8433	0.9964	0.0746	0.3699
	2.00	0.9999	0.5172	0.9803	0.0409	0.1750
	3.00	0.9963	0.1783	0.9138	0.0155	0.0534
	3.50	0.9866	0.0784	0.8428	0.0081	0.0235
	3.60	0.9823	0.0665	0.8275	0.0079	0.0195
	3.70	0.9770	0.0550	0.8069	0.0071	0.0158
H_0	3.80	0.9706	0.0441	0.7868	0.0050	0.0124
	3.90	0.9634	0.0354	0.7679	0.0053	0.0104
	4.00	0.9559	0.0292	0.7472	0.0043	0.0090
	5.00	0.7699	0.0020	0.5102	0.0010	0.0008
	6.00	0.4012	0.0001	0.2824	0.0002	0.0000

Table A.3: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$, where δ=-6, -5, -4, ..., 6, and $\sigma^2 = 9^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-6.00	0.6065	0.0001	0.3586	0.0000	0.0001
	-5.00	0.8468	0.0050	0.5754	0.0005	0.0012
	-4.00	0.9614	0.0338	0.7690	0.0040	0.0088
H_1	-3.00	0.9922	0.1411	0.9063	0.0140	0.0357
	-2.00	0.9995	0.3709	0.9688	0.0335	0.1009
	-1.00	1.0000	0.6492	0.9913	0.0625	0.1925
	0.00	1.0000	0.7662	0.9978	0.0763	0.2416
	1.00	0.9999	0.6393	0.9965	0.0707	0.1941
	2.00	0.9999	0.3655	0.9854	0.0474	0.0981
	3.00	0.9936	0.1371	0.9531	0.0302	0.0372
	3.50	0.9840	0.0721	0.9191	0.0182	0.0185
	3.60	0.9811	0.0638	0.9103	0.0177	0.0163
	3.70	0.9765	0.0543	0.9011	0.0167	0.0145
H_0	3.80	0.9714	0.0450	0.8897	0.0153	0.0119
	3.90	0.9663	0.0386	0.8769	0.0152	0.0101
	4.00	0.9594	0.0323	0.8673	0.0128	0.0087
	5.00	0.8436	0.0061	0.7172	0.0041	0.0016
	6.00	0.5973	0.0004	0.5281	0.0008	0.0000

Table A.4: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \boldsymbol{X} + \boldsymbol{\varepsilon}$, where δ=-6, -5, -4, ..., 6, and $\sigma^2 = 14^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-6.00	0.8076	0.0037	0.5970	0.0021	0.0017
	-5.00	0.9128	0.0155	0.7326	0.0050	0.0074
	-4.00	0.9668	0.0444	0.8427	0.0092	0.0179
H_1	-3.00	0.9885	0.0972	0.9189	0.0217	0.0365
	-2.00	0.9974	0.1680	0.9605	0.0350	0.0629
	-1.00	0.9995	0.2441	0.9825	0.0466	0.0911
	0.00	0.9997	0.2682	0.9929	0.0643	0.1035
	1.00	0.9999	0.2344	0.9956	0.0733	0.0918
	2.00	0.9981	0.1687	0.9930	0.0622	0.0632
	3.00	0.9895	0.0915	0.9826	0.0574	0.0373
	3.50	0.9816	0.0642	0.9736	0.0443	0.0273
	3.60	0.9790	0.0587	0.9708	0.0463	0.0261
	3.70	0.9762	0.0535	0.9686	0.0430	0.0244
H_0	3.80	0.9730	0.0480	0.9663	0.0409	0.0222
	3.90	0.9691	0.0436	0.9637	0.0392	0.0206
	4.00	0.9657	0.0390	0.9615	0.0367	0.0196
	5.00	0.9090	0.0150	0.9182	0.0199	0.0081
	6.00	0.8041	0.0046	0.8467	0.0134	0.0033

A.2 Scaled Simulations

Table A.5: The power of the test using the data simulated from $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$, where γ=0.8, 0.85, 0.9, ..., 1.2, and $\sigma^2 = 5^2$.

	γ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	0.80	0.0348	0.0000	0.7535	0.0016	0.0000
	0.85	0.5860	0.0000	0.9191	0.0127	0.0000
H_1	0.90	0.9881	0.0270	0.9843	0.0466	0.0012
	0.92	0.9990	0.2040	0.9930	0.0710	0.0200
	0.94	1.0000	0.6320	0.9940	0.0930	0.1540
	0.95	1.0000	0.7959	0.9974	0.1016	0.2705
	0.96	1.0000	0.9230	0.9960	0.1030	0.4230
	0.98	1.0000	0.9940	0.9990	0.1130	0.7110
	1.00	1.0000	0.9998	0.9998	0.1230	0.8628
	1.02	1.0000	0.9930	1.0000	0.1390	0.7230
	1.04	1.0000	0.9200	0.9970	0.1150	0.4230
	1.05	1.0000	0.7920	0.9975	0.0986	0.2811
	1.06	1.0000	0.6010	0.9960	0.0880	0.1310
	1.08	1.0000	0.1830	0.9890	0.0720	0.0120
	1.10	0.9902	0.0269	0.9810	0.0523	0.0009
H_0	1.15	0.5686	0.0000	0.9087	0.0210	0.0000
	1.20	0.0325	0.0000	0.7575	0.0094	0.0000

Table A.6: The power of the test using the data simulated from $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$, where γ=0.8, 0.85, 0.9, ..., 1.2, and $\sigma^2 = 7^2$.

	γ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	0.80	0.1784	0.0000	0.7999	0.0054	0.0000
	0.85	0.7228	0.0001	0.9264	0.0176	0.0000
H_1	0.90	0.9828	0.0354	0.9776	0.0454	0.0005
	0.92	0.9940	0.1570	0.9880	0.0570	0.0100
	0.94	1.0000	0.4050	0.9920	0.0690	0.0610
	0.95	1.0000	0.5618	0.9955	0.0744	0.1066
	0.96	1.0000	0.7190	0.9940	0.0850	0.2000
	0.98	1.0000	0.8960	0.9950	0.0860	0.3700
	1.00	1.0000	0.9536	0.9988	0.0886	0.4654
	1.02	1.0000	0.8950	0.9960	0.0950	0.3580
	1.04	1.0000	0.6950	0.9970	0.0970	0.1890
	1.05	1.0000	0.5515	0.9967	0.0827	0.1093
	1.06	1.0000	0.3820	0.9960	0.0900	0.0440
	1.08	1.0000	0.1370	0.9930	0.0640	0.0040
	1.10	0.9843	0.0357	0.9847	0.0578	0.0010
H_0	1.15	0.7138	0.0001	0.9463	0.0324	0.0000
	1.20	0.1723	0.0000	0.8644	0.0167	0.0000

Table A.7: The power of the test using the data simulated from $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$, where γ=0.8, 0.85, 0.9, ..., 1.2, and $\sigma^2 = 9^2$.

	γ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	0.80	0.3496	0.0000	0.8244	0.0065	0.0000
	0.85	0.7950	0.0002	0.9238	0.0185	0.0000
	0.90	0.9804	0.0413	0.9733	0.0393	0.0015
	0.92	0.9910	0.1290	0.9850	0.0460	0.0110
	0.94	0.9990	0.2830	0.9900	0.0580	0.0490
	0.95	0.9996	0.3989	0.9921	0.0651	0.0689
	0.96	1.0000	0.5310	0.9920	0.0680	0.1190
	0.98	1.0000	0.7250	0.9940	0.0610	0.1850
H_1	1.00	1.0000	0.7662	0.9978	0.0763	0.2416
	1.02	1.0000	0.6900	0.9940	0.0700	0.1940
	1.04	1.0000	0.4950	0.9940	0.0700	0.1020
	1.05	0.9999	0.3916	0.9963	0.0752	0.0692
	1.06	1.0000	0.2750	0.9940	0.0810	0.0310
	1.08	0.9980	0.1140	0.9920	0.0710	0.0120
	1.10	0.9804	0.0410	0.9884	0.0651	0.0009
H_0	1.15	0.7896	0.0006	0.9644	0.0423	0.0000
	1.20	0.3393	0.0000	0.9106	0.0229	0.0000

Table A.8: The power of the test using the data simulated from $\boldsymbol{Y} = \gamma \boldsymbol{X} + \boldsymbol{\varepsilon}$, where γ=0.8, 0.85, 0.9, ..., 1.2, and $\sigma^2 = 14^2$.

	γ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	0.80	0.6411	0.0000	0.8281	0.0103	0.0000
	0.85	0.8799	0.0062	0.9104	0.0191	0.0005
H_1	0.90	0.9700	0.0430	0.9590	0.0340	0.0070
	0.92	0.9850	0.1090	0.9700	0.0400	0.0210
	0.94	0.9930	0.1450	0.9780	0.0520	0.0460
	0.95	0.9970	0.1710	0.9820	0.0460	0.0570
	0.96	0.9980	0.1980	0.9840	0.0500	0.0650
	0.98	1.0000	0.2680	0.9900	0.0590	0.0800
	1.00	1.0000	0.2840	0.9910	0.0660	0.0970
	1.02	1.0000	0.2430	0.9920	0.0560	0.0920
	1.04	1.0000	0.1820	0.9920	0.0810	0.0640
	1.05	1.0000	0.1620	0.9910	0.0760	0.0520
	1.06	0.9990	0.1350	0.9910	0.0720	0.0370
	1.08	0.9930	0.0840	0.9930	0.0710	0.0140
	1.10	0.9790	0.0410	0.9920	0.0650	0.0080
H_0	1.15	0.8750	0.0090	0.9860	0.0550	0.0000
	1.20	0.6314	0.0004	0.9582	0.0466	0.0000

A.3 Shift-Scaled Simulations

Table A.9: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\gamma = 0.98$, $\delta = -4, -3, -2, ..., 4$ and $\sigma^2 = 3^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-4.00	0.3950	0.0000	0.0710	0.0000	0.0260
	-3.90	0.4980	0.0000	0.0900	0.0000	0.0430
	-3.80	0.5950	0.0000	0.1060	0.0000	0.0530
H_1	-3.70	0.6740	0.0000	0.1260	0.0000	0.0710
	-3.60	0.7500	0.0000	0.1670	0.0000	0.0910
	-3.50	0.8200	0.0000	0.2090	0.0000	0.1140
	-3.00	0.9690	0.0380	0.4460	0.0000	0.3020
	-2.00	1.0000	0.7390	0.9010	0.0050	0.8030
	-1.00	1.0000	0.9980	0.9950	0.1000	0.9750
	0.00	1.0000	1.0000	1.0000	0.2780	0.9900
	1.00	1.0000	1.0000	1.0000	0.1470	0.9820
	2.00	1.0000	1.0000	0.9780	0.0110	0.7930
	3.00	1.0000	0.9650	0.6650	0.0020	0.3040
	3.50	1.0000	0.7120	0.3680	0.0000	0.1250
	3.60	1.0000	0.6320	0.3230	0.0000	0.0990
	3.70	1.0000	0.5310	0.2800	0.0010	0.0790
H_0	3.80	1.0000	0.4290	0.2430	0.0000	0.0610
	3.90	1.0000	0.3480	0.2160	0.0000	0.0470
	4.00	1.0000	0.2680	0.1950	0.0000	0.0300

Table A.10: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\gamma = 0.96$, $\delta = -4, -3, -2, ..., 5$ and $\sigma^2 = 3^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-4.00	0.0280	0.0000	0.0260	0.0000	0.0130
H_0	-3.90	0.0500	0.0000	0.0300	0.0000	0.0230
H_0	-3.80	0.0790	0.0000	0.0430	0.0000	0.0280
	-3.70	0.1290	0.0000	0.0580	0.0000	0.0370
	-3.60	0.1740	0.0000	0.0820	0.0000	0.0460
	-3.50	0.2480	0.0000	0.1060	0.0000	0.0570
	-3.00	0.6810	0.0000	0.2780	0.0000	0.1850
	-2.00	0.9930	0.1100	0.7840	0.0010	0.6180
	-1.00	1.0000	0.8600	0.9850	0.0450	0.7830
H_0	0.00	1.0000	1.0000	1.0000	0.2170	0.7980
	1.00	1.0000	1.0000	1.0000	0.1820	0.7980
	2.00	1.0000	1.0000	0.9940	0.0300	0.7250
	3.00	1.0000	1.0000	0.8240	0.0020	0.3040
	3.50	1.0000	0.9850	0.5710	0.0000	0.1250
	3.60	1.0000	0.9760	0.5120	0.0030	0.0990
	3.70	1.0000	0.9640	0.4420	0.0000	0.0790
	3.80	1.0000	0.9270	0.3950	0.0000	0.0610
H_0	3.90	1.0000	0.8960	0.3420	0.0000	0.0470
H_0	4.00	1.0000	0.8400	0.2970	0.0000	0.0300
H_0	5.00	0.9970	0.0950	0.0380	0.0000	0.0020

Table A.11: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\gamma = 0.94$, $\delta = -4, -3, -2..., 6$ and $\sigma^2 = 3^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-4.00	0.0000	0.0000	0.0110	0.0000	0.0010
	-3.00	0.1490	0.0000	0.1460	0.0000	0.0290
	-2.00	0.8880	0.0000	0.6060	0.0000	0.2200
	-1.00	1.0000	0.2070	0.9550	0.0170	0.3350
	0.00	1.0000	0.9330	0.9970	0.1600	0.3450
H_1	1.00	1.0000	1.0000	1.0000	0.2250	0.3450
	2.00	1.0000	1.0000	1.0000	0.0480	0.3430
	3.00	1.0000	1.0000	0.9390	0.0040	0.2400
	3.50	1.0000	1.0000	0.7640	0.0000	0.1180
	3.60	1.0000	1.0000	0.7150	0.0010	0.0930
	3.70	1.0000	1.0000	0.6680	0.0000	0.0740
H_0	3.80	1.0000	0.9990	0.6000	0.0000	0.0580
	3.90	1.0000	0.9970	0.5470	0.0010	0.0460
	4.00	1.0000	0.9930	0.4890	0.0000	0.0300
	5.00	1.0000	0.6020	0.0870	0.0000	0.0020
	6.00	0.9930	0.0310	0.0070	0.0000	0.0000

Table A.12: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\gamma = 0.92$, $\delta = -4, -3, -2, ..., 6$ and $\sigma^2 = 3^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-4.00	0.0000	0.0000	0.0030	0.0000	0.0000
	-3.00	0.0080	0.0000	0.0790	0.0000	0.0000
	-2.00	0.4150	0.0000	0.4430	0.0000	0.0110
	-1.00	0.9690	0.0030	0.8900	0.0080	0.0310
	0.00	1.0000	0.3100	0.9960	0.1070	0.0370
	1.00	1.0000	0.9620	1.0000	0.2310	0.0370
H_1	2.00	1.0000	1.0000	1.0000	0.0670	0.0370
	3.00	1.0000	1.0000	0.9890	0.0050	0.0370
	3.50	1.0000	1.0000	0.9100	0.0040	0.0340
	3.60	1.0000	1.0000	0.8890	0.0000	0.0330
	3.70	1.0000	1.0000	0.8450	0.0030	0.0310
	3.80	1.0000	1.0000	0.7960	0.0000	0.0250
	3.90	1.0000	1.0000	0.7420	0.0000	0.0200
H_0	4.00	1.0000	1.0000	0.7140	0.0000	0.0160
	5.00	1.0000	0.9680	0.1610	0.0000	0.0020
	6.00	1.0000	0.2900	0.0170	0.0000	0.0000

Table A.13: The power of the test using the data simulated from $\boldsymbol{Y} = \delta\mathbf{1} + \gamma\boldsymbol{X} + \boldsymbol{\varepsilon}$, where $\gamma = 0.90$, $\delta = -4, -3, -2, ..., 9$ and $\sigma^2 = 3^2$.

	δ	Shift-E	Shift-E*	Scale-E	Scale-E*	Shift-Scale-E
H_0	-4.00	0.0000	0.0000	0.0000	0.0000	0.0000
	-3.00	0.0000	0.0000	0.0310	0.0000	0.0000
	-2.00	0.0500	0.0000	0.2940	0.0000	0.0000
	-1.00	0.7440	0.0000	0.7950	0.0020	0.0010
	0.00	0.9940	0.0060	0.9880	0.0570	0.0010
	1.00	1.0000	0.4510	1.0000	0.2260	0.0010
H_1	2.00	1.0000	0.9800	1.0000	0.1060	0.0010
	3.00	1.0000	1.0000	0.9980	0.0100	0.0010
	3.50	1.0000	1.0000	0.9770	0.0040	0.0010
	3.60	1.0000	1.0000	0.9700	0.0050	0.0010
	3.70	1.0000	1.0000	0.9610	0.0020	0.0010
	3.80	1.0000	1.0000	0.9370	0.0030	0.0010
	3.90	1.0000	1.0000	0.9120	0.0010	0.0010
	4.00	1.0000	1.0000	0.8880	0.0010	0.0010
	5.00	1.0000	1.0000	0.3240	0.0000	0.0000
H_0	6.00	1.0000	0.8190	0.0330	0.0000	0.0000
	7.00	1.0000	0.0890	0.0030	0.0000	0.0000
	8.00	0.9620	0.0000	0.0000	0.0000	0.0000
	9.00	0.2720	0.0000	0.0000	0.0000	0.0000

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.

Blackwelder, W. C. (1982). "proving the null hypothesis" in clinical trials. *Controlled clinical trials*, *3*(4), 345–353.

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.