A simulation-based comparison of tests for equivalence in clinical trials with application to tobacco data

by

Junpu Xie

B.Sc., Carleton University, 2021

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

© Junpu Xie 2023
SIMON FRASER UNIVERSITY
Summer 2023

Declaration of Committee

Name:	**Junpu Xie**
Degree:	**Master of Science**
Thesis title:	**A simulation-based comparison of tests for equivalence in clinical trials with application to tobacco data**
Committee:	**Chair:** Haolun Shi Assistant Professor, Statistics and Actuarial Science **Donald Estep** Supervisor Professor, Statistics and Actuarial Science **Joan Hu** Committee member Professor, Statistics and Actuarial Science **Ian Bercovitz** Examiner Director, Statistical Consulting Service, Statistics and Actuarial Science

Abstract

The equivalence problem is concerned with statistical methodology to assess the equivalence between two medications or two formulations in clinical trials. Frequently, researchers incorrectly use difference tests to evaluate equivalence. This study introduces three common equivalence tests (two one-sided tests, power analysis, and the Hauck-Anderson method) and investigates their performance. We conclude by describing appropriate experimental design and testing procedures for assessing equivalence between tobacco composition developed by Health Canada (HC) and the World Health Organization (WHO) and explore their performance through a simulation-based comparison.

Keywords: Tobacco composition; Equivalence test; Simulation analysis; Experimental design.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Tobacco Reporting Regulations (TRR), which are governed by the Tobacco Control Directorate (TCD), establish the official measurement procedure that local tobacco manufacturers must follow in order to submit annual reports on the prescribed emissions and ingredients of tobacco products to Health Canada (HC) [7]. The World Health Organization's (WHO) Tobacco Laboratory Network (TobLabNet) has recently developed a series of standard operating procedures (SOP) that measure the same tobacco's chemical data specified in the TRR. The objectives of both measuring systems are to regulate tobacco corporations as far as the health and environmental impact of their products are concerned and to ensure that the public has access to health and safety information regarding tobacco use.

The TCD states that the smoke nitrosamine methods (T-111B and SOP 03), nicotine and carbon monoxide (CO) methods (T-115 and SOP 10), and whole tobacco humectant methods (T-304 and SOP 06) are comparable in measuring data about cigarette constituents and smoking emission levels. The conclusion is summarized in Table 1.1.

HC Code	WHO Code	Constituents/Emissions
T-111B	SOP03	Determination of Tobacco emission of NAB, NAT, NNK, NNN in Mainstream Intense smoking condition
T-111b	SOP03	Determination of Tobacco emission of NAB, NAT, NNK, NNN in Mainstream ISO smoking condition
T-304	SOP06	Determination of Tobacco constituent of humectants/ glycerol in cigarette tobacco filler
T-115	SOP10	Determination of Tobacco emission of Nicotine and Carbon Monoxide in Mainstream Intense smoking condition

Table 1.1: Table of similar tobacco data collection techniques for HC and WHO.

Although two methods might be evaluated as comparable based on practical experience, it is acknowledged that no two procedures yield statistically identical measurement results (Hauck & Anderson, 1984) [4] due to varying experimental conditions (ISO, 2018) [1]. So, Health Canada and this report aim to review statistical methodologies used to demonstrate statistical equivalence between data collected for tobacco smoke analysis from HC and WHO, as well as to recommend alternative approaches to determining equivalence where applicable.

The traditional two-sample t-test is always considered to compare the two means; it cannot be used to assess the equivalence of two groups because of the following reasons: First, the t-test's conclusion of "do not reject the null hypothesis of equality", only states that the sample is insufficient to infer that two groups differ considerably. It does not suggest that two measurements are identical or equivalent. Secondly, the t-test only compares the sample mean difference to zero. When the practical significance of the mean difference exists, it cannot be determined whether the two groups are within an equivalence range. Moreover, the t-test makes it hard to reject the null hypothesis of equality when the sample measurement precision is high.

This report first (a) illustrates the t-test's inadequacies for assessing the equivalence of two procedures and (b) proposes a series of alternative equivalence tests (power analysis, the two one-sided t-tests (TOST) (1987), and Hauck-Anderson's test (1984)). Then the report c) summarizes equivalence limit and sample size determination methods and further compares three tests' power trends on correctly concluding a true equivalence between two groups with various parameters. The study closes with an application for assessing equivalence between tobacco's emission and constituent levels using HC's method. Figures, tables, and R codes are presented in the appendix.

Chapter 2

Two-sample t-test for equivalence determination

2.1 Two-sample t-test

The two-sample t-test is a conventional approach to determining if the means of two independently normal distributions are substantially different.

Assume that two populations are independently normally distributed and have identical variances ($\sigma_R^2 = \sigma_T^2$). The null hypothesis of the t-test is that the two population means are equal, while its alternative hypothesis states that they are not. The hypotheses are:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0, \\ H_A : \mu_1 - \mu_2 \neq 0. \end{cases} \tag{2.1}$$

By randomly collecting two samples $(y_{11}, ..., y_{1n_1}), (y_{21}, ..., y_{2n_1})$ from their distributions, the t-test is run to determine if there is a significant difference under the null hypothesis of equality. The test statistic is computed based on the sample statistics, containing sample size ($n = n_1 = n_2$), pooled standard deviation ($s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$) sample mean $(\bar{y_1}, \bar{y_2})$ as follows:

$$T = \frac{(\bar{y_1} - \bar{y_2})}{s_p\sqrt{2/n}}. \tag{2.2}$$

Both the critical value approach and the p-value approach employ the critical t-value from the t-distribution of the entire population as a point of reference to assess a significant mean difference between two groups. When $p < \alpha$ or $|T| > t_{1-\alpha/2,2n-2}$, the t-test concludes that the sample data has adequate evidence that there is $(1 - \alpha)\%$ that the two population means significantly differ. The confidence interval (CI) also concludes a significant difference when the zero is not contained within the range of mean difference $(\bar{y_1} - \bar{y_2}) \pm t_{1-\alpha/2,2n-2} s_p \sqrt{2/n}$.

2.2 Problems that arise in using the t-test to determine equivalence

Although the two-sample t-test is an adequate hypothesis test to assess if the means of two normal groups are significantly different from one another, attempting to use it to demonstrate equivalence between two groups can lead to a variety of complications. The following simulation study reveals some difficulties.

We consider simulations with two samples with a range of sample sizes (n=5, 7, 10, 12, 15, 20, 25, 30). For each sample size, the first sample is drawn from a normal distribution with parameters $N(0, 1)$, whereas the second sample is drawn from a normal distribution with the same variance but varying population means of 0, 0.5, 1, 2, or 3. For each parameter combination of sample size and population mean difference, 1000 paired samples are generated. The t-test is conducted on every pair of them.

Figure 2.1 contains five curves showing the average probabilities that the t-test fails to infer the significant difference versus sample sizes for a specific level of mean difference.

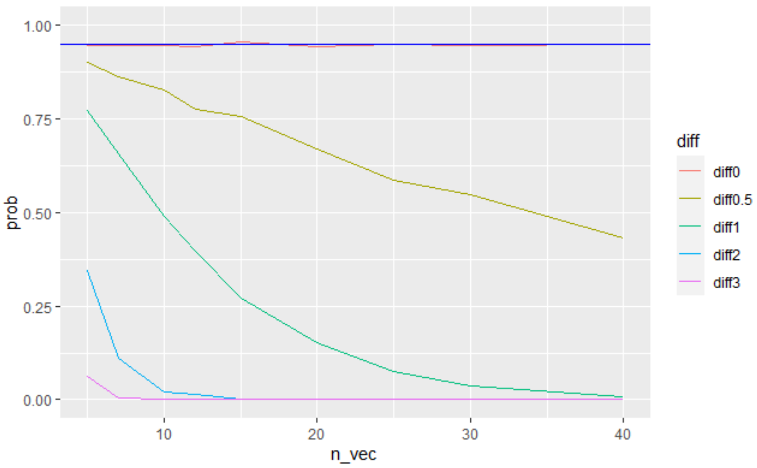


Figure 2.1: Plot of probabilities (prob) that a paired t-test fails to determine a significant difference versus sample size ($\mathrm{n}_{\mathrm{vec}}$) for 5 difference mean differences.

Problem 1: Practical significance of the mean difference

The red curve in Figure 2.1 illustrates that only when the t-test is conducted on two populations with no actual mean difference, the likelihood of not rejecting the null hypothesis of equality is 95%. As the mean difference increases, its probability decreases dramatically.

So, the first issue occurs when the magnitude of the difference has little practical significance. On the basis of prior experience and professional expertise, subject-matter specialists sometimes consider that a difference between two means is acceptable. Yet, the t-test cannot distinguish between statistically significant and scientifically relevant differences.

Problem 2: Sample measurement precision

Figure 2.1 shows that the t-test has a great chance of failing to conclude a significant mean difference when the sample size is small. For instance, we expect the t-test to strongly reject the initial assumption when the mean difference between two groups is 2. However, around 40% of simulated pairs with five data points do not conclude a significant mean difference.

Limentani (2005) observes that higher sample measurement precision (smaller sample size and/or large sample variance) yields a smaller test statistic value and a greater p-value. It results in the t-test becoming more likely to conclude that the null hypothesis of equality should not be rejected.

Problem 3: The t-test's objective

In the absence of strong evidence that the t-test concludes two mean values are dissimilar (prob > 0.05 in the Figure 2.1), many analysts confidently state that two populations are equal [2]. This leads to another problem. Failure to reject the null hypothesis of equality under the t-test only implies that the sample data has insufficient evidence to make a reliable conclusion.

Chapter 3

Equivalence Tests

This chapter introduces three equivalence tests, they are investigated to compensate for the shortcomings associated with using the t-test to determine equivalence. These three tests are:

1. Power analysis emphasizes the t-test's power to assess an acceptably significant difference as equivalence in terms of measurement precision [6];
2. Two one-sided t-tests (TOST) with Limentani's formula present a suitable limit and sample size determination for equivalence tests [5];
3. An alternative equivalence test for log-normally distributed data was suggested by Hauck and Anderson (1984) [4].

3.1 Power analysis

3.1.1 Introduction to power analysis

Power analysis is generally employed in experimental design, especially for sample size determination prior to a study. The following subsection uses formulas and simulation studies to present the relationships between power and the other parameters. The next subsection demonstrates how the power analysis penalizes samples with high measurement precision and concludes that a significant mean difference is within an equivalence range.

Power formula

The statistical power $B(\theta)$ measures the probability that the test correctly rejects the incorrect null hypothesis of equality. The hypothesis is conducted under the population distribution of mean difference centered at the preset equivalence limit θ in terms of measurement precision. The following calculation illustrates a trade-off between statistical power $B(\theta)$ and the other parameters like sample size n, significance level α, and effect size D.

$$\begin{aligned}
B(\theta) &= 1-\beta \\
&= P\left(\text{reject } H_0 | \text{true } H_1\right) \\
&= P\left(\left|N\left(\frac{\theta}{\sigma\sqrt{2/n}},1\right)\right| > t_{1-\alpha/2,2n-2} | \mu_D = \theta\right) \\
&= P\left(N\left(\frac{\theta}{\sigma\sqrt{2/n}},1\right) > t_{1-\alpha/2,2n-2}\right) I(\theta \geq 0) \\
&\qquad\qquad + P\left(N\left(-\frac{\theta}{\sigma\sqrt{2/n}},1\right) > t_{1-\alpha/2,2n-2}\right) I(\theta < 0) \\
&= P\left(N(0,1) > t_{1-\alpha/2,2n-2} - \frac{\theta}{\sigma\sqrt{2/n}}\right) I(\theta \geq 0) \\
&\qquad\qquad + P\left(N(0,1) > t_{1-\alpha/2,2n-2} + \frac{\theta}{\sigma\sqrt{2/n}}\right) I(\theta < 0) \\
&= F_v\left(\frac{\theta}{\sigma\sqrt{2/n}} - t_{1-\alpha/2,2n-2}\right) I(\theta \geq 0) + F_v\left(t_{\alpha/2,2n-2} - \frac{\theta}{\sigma\sqrt{2/n}}\right) I(\theta < 0) \\
&= F_v\left(\frac{D}{\sqrt{2/n}} - t_{1-\alpha/2,2n-2}\right) I(\theta \geq 0) + F_v\left(t_{\alpha/2,2n-2} - \frac{D}{\sqrt{2/n}}\right) I(\theta < 0)
\end{aligned} \tag{3.1}$$

The term F_v in the equation (3.1) stands for the distribution function for all values that are smaller than the observed one under the student's t distribution with $v = 2n-2$ degrees of freedom. The effect size is a standardized division of the standard deviation to mean, and its formula is written as $D = \frac{\mu_D}{\sigma} = \frac{\theta}{\sigma}$.

Simulation analysis

In order to demonstrate how parameters affect the power, a simulation analysis is performed with two normal populations of the same variances of 1. The intention is to investigate the variation of the power $B(\theta)$ as every statistical parameters, including the significance level α, sample size n and effect size d vary. Figure 3.2 shows the curve of calculated power versus influential component.

Figure 3.2 a) specifies the plot of calculated power versus significance level ranging from 0.01 to 0.20, while mean difference and effect size are 1, and sample size is at 30. Figure 3.2 b) depicts how power fluctuates with sample sizes ranging from 5 to 40, assuming $\alpha = 0.05$ and $d = \mu_D = 1$. Plot c) depicts the relationship between power and effect size, ranging from 0 to 3, with $\alpha = 0.05$ and $n = 30$.

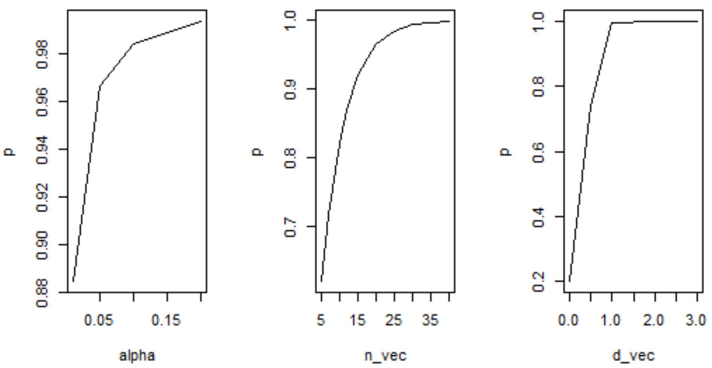


Figure 3.1: Plots of average power (p) changes with various a) significance level (alpha), b) sample size (n_vec) and c) effect size (d_vec).

a. Power versus significance level

Figure 3.1 a) demonstrates that increasing the significance level raises the statistical power. In order to understand this trend, the following example emphasizes the link between two levels.

The following example plots two normal distributions of population mean difference: $N(\mu_1 = 0, s_1^2 = 6)$ and $N(\mu_2 = 1, s_2^2 = 6)$. In Figure 3.2, the blue area shows the standard normal distribution of the population mean difference centered at zero under the null hypothesis of the test ($\mu_1 - \mu_2 = 0$). The orange curve depicts the alternative hypothesis that the population mean of the test group is one unit greater than that of the reference group. In hypothesis testing, there is only one value for the critical t-value; it reflects the intended threshold for accepting or rejecting the null hypothesis.

Under the different assumptions regarding the distribution of mean differences, the left and right panels present two error rates. The area on the left displays the expected significance level for the t-test, which is the acceptable probability that the sample data incorrectly leads to the conclusion that there is a significant difference assuming the two population means are the same. On the other hand, the shaded region on the right indicates the desired degree of power, which represents the chance that the sample data is properly collected leads to the inference that a significant difference between two means.

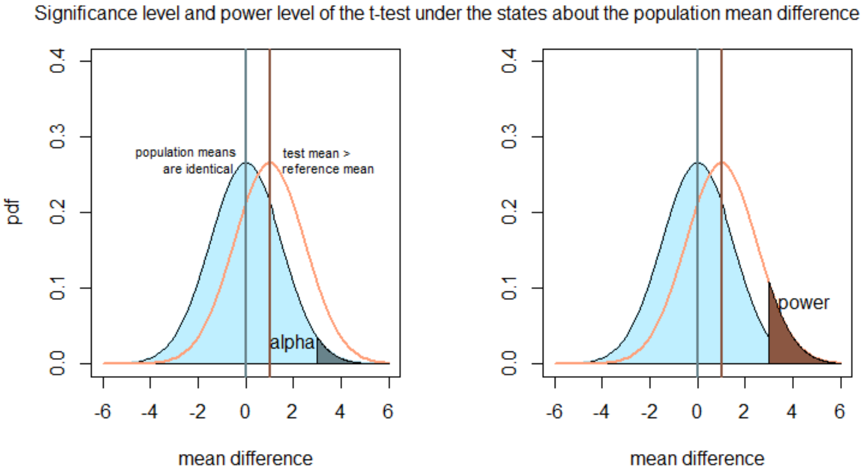


Figure 3.2: Plots of a) significance and b) power of the upper one-sided t-test under the two states about a given population mean difference.

Therefore, one strategy to increase the power of a test is to use a higher significance threshold. As the level of significance rises, the optical t-value drops down the x-axis of the population distribution. So, the chance of a type I error decreases, resulting in an increase in the test's power. This explains why increasing the level of importance raises power.

b. Power versus sample size

Figure 3.1 b) illustrates the link between sample size and the power of a test. Both graph and calculation demonstrate that a increasing sample size increases the power. Their positive association explains why power analysis penalizes situations with a limited sample size.

c. Power versus effect size

Figure 3.1 c) demonstrates a positive correlation between power and effect size in the range from 0 to 3. When the effect size grows, the likelihood of detecting a significant mean difference increases. Because both population variances equal 1, the mean difference equals the effect size. It is clear that when the effect size or significance of the mean difference increases, the power becomes close to 1.

In summary, this simulation study demonstrates the trade-offs between various parameters and statistical power. As the sample size, degree of significance, or effect size increase, the power increases.

3.1.2 Power analysis in equivalence tests

When the t-test shows the data cannot support a conclusion of a significant mean difference, the power analysis is used to validate that the conclusion results because the magnitude of the mean difference is within a preset equivalence range and is not due to high measurement precision (Schuirmann, 1987). [6]

The following equation converts the formula of power to the formula of standard deviation and the calculated standard deviation $\sigma_{1-\beta}$ based on the desired power represents the most acceptable measurement precision value.

$$\begin{aligned} B(\theta) &= F_v\left(\frac{D}{\sqrt{2/n}} - t_{1-\alpha/2,2n-2}\right) I(\theta \geq 0) + F_v\left(t_{\alpha/2,2n-2} - \frac{D}{\sqrt{2/n}}\right) I(\theta < 0) \\ 1-\beta &= F_v\left(\frac{\theta}{\sigma\sqrt{2/n}} - t_{1-\alpha/2,2n-2}\right) I(\theta \geq 0) + F_v\left(t_{\alpha/2,2n-2} - \frac{\theta}{\sigma\sqrt{2/n}}\right) I(\theta < 0) \\ \sigma_{1-\beta} &\approx \frac{\theta\sqrt{n/2}}{(z_{1-\beta} + z_{1-\alpha/2})} I(\theta \geq 0) + \frac{\theta\sqrt{n/2}}{(z_{\alpha/2} - z_{1-\beta})} I(\theta < 0) \\ \sigma_{1-\beta} &= \frac{\theta\sqrt{n/2}}{(z_{1-\beta} + z_{1-\alpha/2})} (I(\theta \geq 0) - I(\theta < 0)) \end{aligned} \tag{3.2}$$

Consequently, any samples satisfies the following requirements generates more than $(1-\beta)100\%$ power to infer that the mean difference is within a specific equivalence range about θ and also carries out the test of hypothesis of insignificant difference at a significance level of α.

$$\begin{cases} t_{\alpha/2,2n-2} \leq \frac{\bar{y_1}-\bar{y_2}}{s_p\sqrt{2/n}} \leq t_{1-\alpha/2,2n-2}, \\ s_p \leq \frac{\theta\sqrt{n/2}}{(z_{1-\beta}+z_{1-\alpha/2})} (I(\theta \geq 0) - I(\theta < 0)). \end{cases} \tag{3.3}$$

3.2 The two one-sided test (TOST)

3.2.1 Introduction to the TOST

The Two One-Sided Test (TOST) was proposed by Schuirmann (1987) [6] and its application to assessing equivalence has become widespread recently.

In the equivalence test, the null hypothesis assumes that the groups are distinct, i.e., the mean difference between two groups lies outside the preset equivalence range $[\theta_1, \theta_2]$ and its hypotheses are as follows:

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq \theta_1 \text{ or } \mu_1 - \mu_2 \geq \theta_2, \\ H_A : \theta_1 < \mu_1 - \mu_2 < \theta_2. \end{cases} \tag{3.4}$$

The TOST concludes equivalence when the following requirements are met:

$$\begin{cases} T_U = \left|\frac{\theta_2-(\bar{y_1}-\bar{y_2})}{s_p\sqrt{2/n}}\right| \geq t_{1-\alpha/2,2n-2}, \\ T_L = \left|\frac{(\bar{y_1}-\bar{y_2})-\theta_1}{s_p\sqrt{2/n}}\right| \geq t_{1-\alpha/2,2n-2}. \end{cases} \tag{3.5}$$

The equation shows that the TOST involves two one-sided t-tests. One evaluates whether the mean difference is significantly less than the upper equivalence limit, while the other examines if it is greater than the lower limit. Because lower and upper t-tests are mutually exclusive, both test statistics are compared to the same critical t-value $t_{1-\alpha/2,2n-2}$, which is an upper tail percentile of the central t distribution at the point of the half of significance level (Westlake, 1981) [8]. So, Schuirmann (1987) indicates that the test concludes equivalence when $(1-\alpha)100\%$ CI for the population mean difference is completely contained inside the equivalence interval $[\theta_1, \theta_2]$ [6].

The test statistic formula demonstrates that greater measurement accuracy results in a smaller test statistic value and a lower chance of rejecting the null hypothesis of inequivalence. It ensures the conclusion is solely based on the magnitude of the mean difference as the power analysis.

Taking the error rates as potential risks into account improves the experiment design. Lakens D. (2017) stated that the equivalence test is a Newman-Pearson hypothesis testing approach that allows the researchers to regulate the equivalence tests based on statistical and practical experiences [2]. The preset equivalence boundaries are within reasonable values most of the time, the equivalence test has sufficient statistical power to avoid two error rates (Tan D., Feng G., Zhu R., & Yang H., 2017). Because the TOST is vulnerable to two types of mistakes, the beta value is generally set to 0.05 rather than 0.20.

3.2.2 Simulation analysis of the TOST

The following simulation shows the testing procedure for the TOST. Assuming that two samples of equal size, $n_1 = n_2 = 30$ are randomly drawn from $N(0,1)$ and $N(-5,1)$. By considering two equivalence limits, $\theta_1 = -6$ and $\theta_2 = -4$, there are 1,000 paired simulations generated.

Figure 3.3 a) displays the histogram of the sample mean difference; it is approximately normally distributed and centered at the actual population mean difference of -5. By setting the significance level to $\alpha = 0.05$, the CI for the mean difference is computed (Green vertical lines), and the blue vertical lines represent the equivalence range.

The simulated upper and lower test statistics t_U and t_L are computed as well. Figures 3.3 b) and c) depict their histograms, and the red vertical lines present the critical upper t-values $t_{\alpha/2,n-1}$.

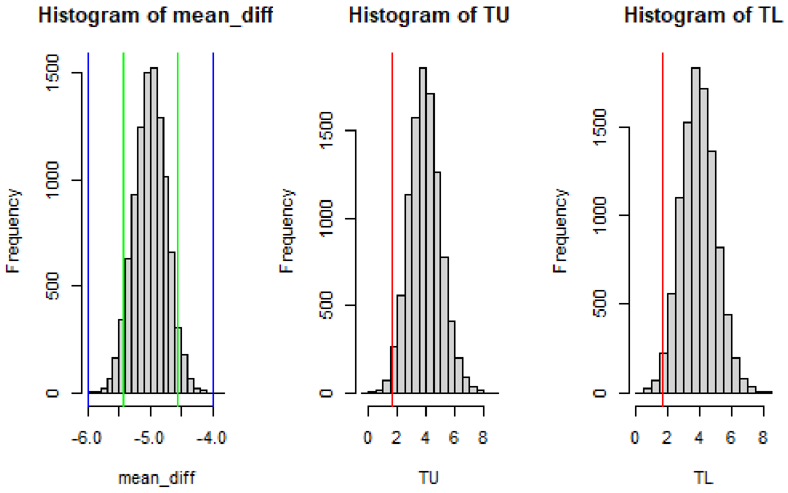


Figure 3.3: Histogram of a) sample mean difference (mean_diff); b) upper test statistics (TU); c) lower test statistics (TL) over 1,000 simulations under the TOST.

i. Confidence interval approach

As with the standard t-test, the equivalence test can be conducted in three ways. We first determine whether the predetermined equivalence range $[\theta_1, \theta_2]$ covers the CI of their mean difference, as illustrated by the vertical lines in the Plot 3.3 a).

Table **??** displays the average interval limits for the population mean difference for 1,000 simulations. The value of $p = 0.972$ shows that the CI of the mean difference falls within the equivalence range 97.2% of the time.

$\bar{y_D}$	LCI	UCI	θ_1	θ_2	p
5.01	-5.426	-4.566	-6	-4	0.972

Table 3.1: Table of confidence interval approach under the TOST.

ii. Critical value approach

Under the critical t value approach, we conclude equivalence when both the upper test statistic of the equivalence limit is greater than the critical t-value of the upper limit of the population mean difference, $T_U \geq t_{1-\alpha/2,2n-2}$, and the lower test statistic is smaller than its lower bound, $T_L \leq -t_{1-\alpha/2,2n-2}$, simultaneously. Table **??** shows that both average test statistics over 1,000 runs are greater than the threshold t-value, and 97.2% of TOST rejects the null hypothesis of non-equivalence and concludes the equivalence between two groups.

$\bar{y}_D$	$\bar{t}_L$	$\bar{t}_U$	t	p
5.01	-3.934	3.907	1.672	0.972

Table 3.2: Table of critical value approach under the TOST.

iii. P-value approach

The p-value measures the probability that the critical t-values are greater than the computed test statistics. Figure 3.3 b) and c) show that the majority of test statistic values exceed the threshold t-value $t_{1-\alpha/2,2n-2}$. Table **??** further shows 97.2% of the simulated pairs reject the initial hypothesis of non-equivalence as their respective p-values are significantly smaller than $\alpha/2 = 0.025$.

$\bar{y}_D$	$\bar{p}_L$	$\bar{p}_U$	α	p
5.01	0.0035	0.0037	0.05	0.972

Table 3.3: Table of p-value approach under the TOST.

3.3 The Hauck-Anderson test

3.3.1 Introduction to the Hauck-Anderson test

Hauck and Anderson (1983) state that the majority of data used in biological applications is log-normally distributed [4]. Hence, they propose a new equivalence test, its main feature is to compare the equivalence between two groups on a logarithmic scale.

We assume the two populations are log-normally distributed. By setting the upper and lower equivalence limits (θ_1, θ_2) and its hypothesis is as follows:

$$\begin{cases} H_0 : \log(\mu_1) - \log(\mu_2) \leq \log(\theta_1) \text{ or } \log(\mu_1) - \log(\mu_2) \geq \log(\theta_2), \\ H_A : \log(\theta_1) < \log(\mu_1) - \log(\mu_2) < \log(\theta_2). \end{cases} \tag{3.6}$$

To simplify the above formula, we consider $M_1 = \log(\mu_1)$ and $M_2 = \log(\mu_2)$ as the population mean values on logarithmic scales and set $A = \log(\theta_1)$ and $B = \log(\theta_2)$ to be the logarithms of the lower and upper equivalence limits of θ_1 and θ_2. The hypotheses simplify to:

$$\begin{cases} H_0 : M_1 - M_2 \leq A \text{ or } M_1 - M_2 \geq B, \\ H_A : A \leq M_1 - M_2 \leq B. \end{cases} \tag{3.7}$$

Assuming that two samples with the same sizes are drawn from their independent normal distributions with the same variances, the test statistic in the Hauck-Anderson approach measures the distance between the sample mean difference m_D and the midpoint of the equivalence range μ_D in terms of measurement precision s' shown as follows:

$$T = \left| \frac{(m_1 - m_2) - \frac{1}{2}(A+B)}{s_p\sqrt{2/n}} \right| = \left| \frac{m_D - \mu_D}{s'} \right|. \tag{3.8}$$

Rather than comparing the test statistic T to the critical t value, the Hauck-Anderson test approaches the problem by measuring the standardized probability of the mean difference at the edge of the equivalence boundaries.

By taking $\delta = \frac{B-A}{2s'}$ as the width of the standardized equivalence range, the observed p-value is $p = F_v(T-\delta) - F_v(-T-\delta)$. It shows a decreasing δ raises the significance of the resulting p-value, which makes it more challenging to demonstrate equivalence.

The following calculation reveals the probability that the observed test statistic is smaller than the critical value at the limit of the equivalence limit:

$$\begin{aligned} p &= F_v(T-\delta) - F_v(-T-\delta) \\ &= F_v\left(\left| \frac{(m_1 - m_2) - \frac{1}{2}(A+B)}{s'} \right| - \frac{B-A}{2s'} \right) - F_v\left(- \left| \frac{(m_1 - m_2) - \frac{1}{2}(A+B)}{s'} \right| - \frac{B-A}{2s'} \right) \\ &= \left[F_v\left(\frac{m_D - B}{s'} \right) - F_v\left(\frac{-m_D + A}{s'} \right) \right] (I(2m_D \geq A+B) - I(2m_D < A+B)) \end{aligned} \tag{3.9}$$

The first term in the equation (3.5) denotes the probability that the areas for all statistical values smaller than the observed term lie under the sampling distribution of the mean difference centered at the upper equivalence limit. The second term represents the probability that the areas for all statistical values are greater than the observed terms under the distribution of mean difference centered at the lower limit. This suggests that the observed p-value is the probability that all observed values are within the equivalence range under the t-distribution of the mean difference. Unlike taking the error rates at the tails of the t-distribution of the mean difference, it takes them at the center of the t-distribution.

3.3.2 Simulation analysis of the Hauck-Anderson approach

The following simulation shows the testing procedure of the Hauck-Anderson method. We first assume two samples $(y_{11}, ..., y_{1n_1})$, $(y_{21}, ..., y_{2n_2})$ with the sizes $n_1 = n_2 = 30$ are generated from log-normal distributions logNormal$(0, 1)$ and logNormal$(0.5, 1)$. The logarithmic equivalence limits are set at $A = 0$ and $B = 1$. 1,000 simulations are generated, and their distributions are shown in the upper panel of Figure 3.5.

The logarithmic sample mean difference and its test statistic are obtained based on the formula. The left-bottom plot depicts the distributions of the logarithmic sample mean difference, with the red vertical lines denoting the upper and lower equivalence bounds.

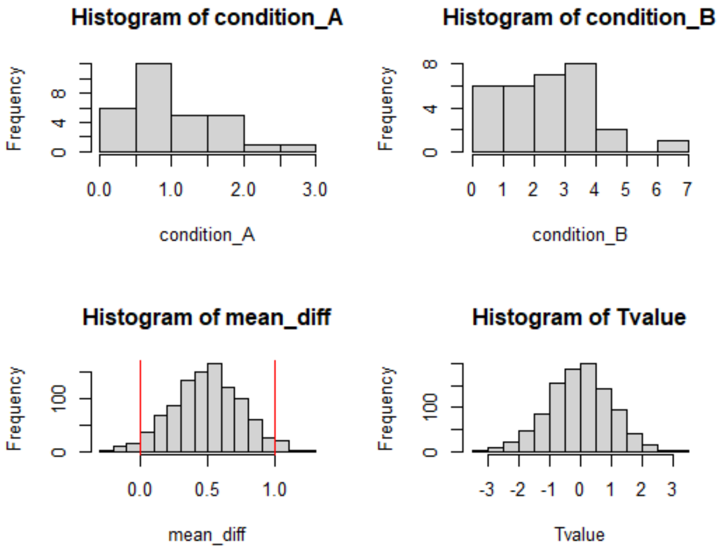


Figure 3.4: Sampling distribution of two sample means, logarithmic sample mean differences and t-test statistics using the Hauck-Anderson approach

The final panel of the four-panel Figure 3.4 shows the histogram of test statistics. Its symmetric feature tells that the usual Student's t distribution is a good approximation for the Hauck-Anderson method. The significance level in this test is determined by maximizing the type I error throughout the interval of the null hypothesis, which is the region located in the center of the t-distribution. So, when a greater proportion of test statistic data is closer to 0, the more sample data supports the statement of group equivalence.

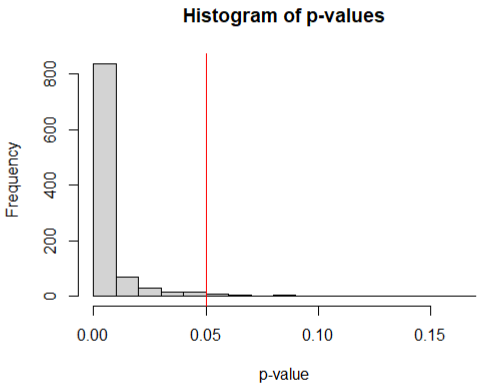


Figure 3.5: Sampling distribution of paired samples, sample mean difference and t test statistics using Hauck-Anderson approach

In conclusion, the Hauck-Anderson approach initially assumes that two groups have a significant mean difference ($\mu_D < A; \mu_D > B$). The measured p-value is the difference between the probability that all values are less than the observed statistic under the null hypothesis of the population mean difference, centered at the upper equivalence limit, and the probability that all values are greater than the observed statistics under the null hypothesis of the population mean difference, centered at the lower equivalence limit. So, the above simulation shows that 96.9% of the simulated pairs are in the rejection region, which concludes that most of them are in the equivalence range.

3.4 Equivalence limit determination

Compared to the t-test, equivalence tests yield a more reliable determination of equivalence. Unfortunately, they are not often employed in practice because specific limitations of experimental designs and equivalence bounds θ must be carefully defined using statistical knowledge and practical experience before a study.

Several public organizations and professional scholars have provided suggestions on equivalence range determination. The U.S. Food and Drug Administration (FDA) defines the set of equivalence limits of the population mean difference as being within 80% to 125% of the reference value. [3] Chow and Liu (1999) observed that most of the mean differences within ±20% of the reference mean are considered equivalent. Moreover, Limentani developed an equivalent limit formula by considering a range of parameters in 2005. [5]

3.4.1 Limentani's Formula

With three key parameters: an expected sample size ($n = n_R = n_T$), a preset mean difference δ and an one-sided upper $(1-\gamma)100\%$ confidence limit of the reference group's sample standard deviation, denoted as s^*, the equivalence limit is computed as follows [5].

$$\theta = \delta + s^*[t_{1-\alpha/2,2n-2} + t_{1-\beta/2,2n-2}]\sqrt{\frac{2}{n}}. \tag{3.10}$$

The first unknown parameter δ is the absolute value of the mean difference between the reference and test groups. It is a threshold variable that has a significant impact on the ability to determine equivalence. For example, if δ increases, θ will likewise rise. When δ is adjusted to be greater than the sample mean difference, it significantly enhances the probability that the mean difference is contained within the equivalence boundary $[-\theta, \theta]$. Therefore, a larger δ value reduces TOST's capacity to distinguish small but statistically significant differences between two means. Usually, δ is assigned zero because a true mean difference is usually unknown.

Another unknown parameter in this formula is $s^* = s_R\sqrt{\frac{n-1}{\chi^2_{\gamma,n-1}}}$. It denotes a one-sided upper $(1-\gamma)100\%$ confidence limit of the sample standard deviation of the reference group, based on the theory that a sampling distribution of the standard deviation for a series of normally distributed groups follows a Chi-square distribution. Compared to the sample standard deviation, it is more reliable to estimate the population standard deviation.

After the sample size is determined, two samples are randomly collected from the population. An equivalence limit θ is computed to assess equivalence between two groups.

3.4.2 Parameter trade-offs on TOST

Limentani's formula (3.10) demonstrates that the equivalence limit is strongly reliant on the preset mean difference δ, sample size n, sample standard deviation s, the rates of type I and type II errors. Hence, this section highlights their trade-offs through tables.

Equivalence limits θ with n and s^*

The following Table 3.4 illustrates how an equivalence limit θ changes for various sample size n and upper limits of measurement precision s^* when $\delta = 0, \alpha = 0.05, \beta = 0.05$ are applied. It is evident that a small sample size and/or a large variance contribute to a relatively high equivalence limit.

	n=5	n=7	n=9	n=11	n=13	n=15	n=17	n=20	n=25	n=30	n=35	n=40
std=0.5	1.19	0.98	0.85	0.77	0.71	0.66	0.62	0.57	0.51	0.47	0.44	0.41
std=0.6	1.43	1.17	1.02	0.92	0.85	0.79	0.74	0.68	0.61	0.56	0.52	0.49
std=0.7	1.67	1.37	1.19	1.07	0.99	0.92	0.86	0.8	0.72	0.66	0.61	0.57
std=0.8	1.91	1.56	1.36	1.23	1.13	1.05	0.99	0.91	0.82	0.75	0.7	0.65
std=0.9	2.15	1.76	1.53	1.38	1.27	1.18	1.11	1.03	0.92	0.85	0.79	0.74
std=1	2.38	1.95	1.7	1.53	1.41	1.31	1.24	1.14	1.02	0.94	0.87	0.82
std=1.5	3.58	2.93	2.55	2.3	2.12	1.97	1.85	1.71	1.54	1.41	1.31	1.23
std=2	4.77	3.9	3.4	3.07	2.82	2.63	2.47	2.28	2.05	1.88	1.74	1.64
std=2.5	5.96	4.88	4.26	3.84	3.53	3.28	3.09	2.85	2.56	2.35	2.18	2.05
std=3	7.15	5.85	5.11	4.6	4.23	3.94	3.71	3.42	3.07	2.82	2.62	2.45
std=3.5	8.34	6.83	5.96	5.37	4.94	4.6	4.32	4	3.59	3.29	3.05	2.86
std=4	9.54	7.8	6.81	6.14	5.64	5.26	4.94	4.57	4.1	3.76	3.49	3.27
std=4.5	10.73	8.78	7.66	6.9	6.35	5.91	5.56	5.14	4.61	4.23	3.93	3.68
std=5	11.92	9.75	8.51	7.67	7.05	6.57	6.18	5.71	5.12	4.7	4.36	4.09

Table 3.4: Table of θ for various sample size n and upper limit of method precision s^*.

Table 3.4 indicates that an unreliable sample size renders the equivalence limits ineffective for distinguishing between truly comparable and non-equivalent procedures. For example, consider a sample that has only 5 data points and whose sample measurement precision is 5. The formula yields a great and unrealistic value of 11.92 because the computation seeks a statistically feasible equivalence limit that infers the equivalence with a certain degree of power. So, the equivalence limit cannot be solely determined by the formula and has to be confirmed by practical experience.

In addition, a large standard error usually generates a wide confidence interval for the mean difference, which results in a greater probability of not rejecting the initial assumption. So, an equivalence limit must be greater than the reference group's measurement error $s/\sqrt{n}$ as it ensures the mean difference is the sole reason to infer that the test fails to reject the null hypothesis of non-equivalence, not the measurement imprecision (Limentani, 2005). [5]

Equivalence Limits θ with α and β

Table 3.5 and Table 3.6 show estimated equivalence limit changes over a variety of error rates, with $s^* = 3$ held constant.

	n=5	n=7	n=9	n=11	n=13	n=15	n=17	n=20	n=25	n=30	n=35	n=40
alpha=0.01	16.73	11.73	9.41	8.02	7.09	6.41	5.89	5.29	4.59	4.1	3.74	3.46
alpha=0.05	13.63	9.77	7.91	6.79	6.02	5.46	5.02	4.52	3.93	3.52	3.22	2.97
alpha=0.1	12.31	8.88	7.21	6.2	5.51	4.99	4.6	4.14	3.61	3.23	2.95	2.73
alpha=0.15	11.52	8.33	6.78	5.83	5.18	4.7	4.33	3.9	3.4	3.04	2.78	2.57
alpha=0.2	10.94	7.92	6.45	5.55	4.93	4.48	4.12	3.72	3.24	2.9	2.65	2.45

Table 3.5: Table of θ for various sample size n and significance level α.

	n=5	n=7	n=9	n=11	n=13	n=15	n=17	n=20	n=25	n=30	n=35	n=40
beta=0.01	16.73	11.73	9.41	8.02	7.09	6.41	5.89	5.29	4.59	4.1	3.74	3.46
beta=0.05	13.63	9.77	7.91	6.79	6.02	5.46	5.02	4.52	3.93	3.52	3.22	2.97
beta=0.1	12.31	8.88	7.21	6.2	5.51	4.99	4.6	4.14	3.61	3.23	2.95	2.73
beta=0.15	11.52	8.33	6.78	5.83	5.18	4.7	4.33	3.9	3.4	3.04	2.78	2.57
beta=0.2	10.94	7.92	6.45	5.55	4.93	4.48	4.12	3.72	3.24	2.9	2.65	2.45

Table 3.6: Table of θ for various sample size n and beta level β.

As expected, θ decreases with increasing type I or type II error rates. So, it increases the difficulty of rejecting the null hypothesis of in-equivalence. In a traditional two-sample t-test, the type I and type II error rates are chosen to be 0.05 and 0.20, respectively. However, reasonable equivalence limits determined by statistical and practical knowledge have extremely low chances of producing type II error (a probability that researchers wrongly conclude that two groups are not equivalent). So, both error rates for equivalence tests are commonly set to 0.05.

Summary

The three tables illustrate the explicit links between the equivalence limit and key parameters. The following concludes some tips for equivalence limit determination settings:

1. An increasing measurement precision (small sample size or large sample variation) or a decreasing type of error leads to greater equivalence limits.

2. The equivalence limit must be greater than the reference group's measurement precision $s_R/\sqrt{n}$, so that the test assesses equivalence based solely on the magnitude of the sample mean difference.

3. The value of an equivalent limit should also be advised by the statisticians and subject matter experts together as the related experimental experiences often defines the appropriate action.

4. Typically, the type I and II errors are set to 0.05 ($\alpha = \beta = 0.05$) when a reasonable equivalence limit is determined.

3.5 Sample size determination

Determining sample size is always an important stage in experimental design. By rewriting the equivalence limit formula, a sample size for the equivalence test is computed as follows:

$$\theta - \delta = s^*[t_{1-\alpha/2,2n-2} + t_{1-\beta/2,2n-2}]\sqrt{\frac{2}{n}} \approx \sigma_R[z_{1-\alpha/2} + z_{1-\beta/2}]\sqrt{\frac{2}{n}} \tag{3.11}$$

Because the equivalence limit and preset mean difference are highly related to the reference mean, their values can be rewritten as $\theta = p_\theta \mu_R$ and $\delta = p_\delta \mu_R$. The coefficient of variation ($CV = \sigma_R/\mu_R$), a relative measure of variability in relation to its mean, is undertaken. Large CV indicates that the standard deviation is relatively larger than the mean. Hence, the sample size formula with new-defined parameters can be simplified as follows:

$$\begin{aligned} (p_\theta - p_\delta)\mu_R &= s_R[z_{1-\alpha/2} + z_{1-\beta/2}]\sqrt{\frac{2}{n}} \\ n &= 2\left[\frac{CV_R(z_{1-\alpha/2} + z_{1-\beta/2})}{(p_\theta - p_\delta)}\right]^2 + 1 \end{aligned} \tag{3.12}$$

Table 3.7 determines the sample size required for a particular equivalence limit, a preset mean difference, and a sample standard deviation. As either one of the preset mean difference and CV increases or the equivalence limit decreases, the required sample size rises.

δ_{pc}	CV_R	θ_{pc} =0.05	θ_{pc} =0.1	θ_{pc} =0.15	θ_{pc} =0.2	θ_{pc} =0.25	θ_{pc} =0.3
0	0.05	32	9	5	3	3	2
0	0.1	124	32	15	9	6	5
0	0.15	278	71	32	19	13	9
0	0.2	493	124	56	32	21	15
0	0.25	770	194	87	50	32	23
0	0.3	1108	278	124	71	46	32
0.02	0.05	87	14	6	4	3	2
0.02	0.1	343	50	20	11	7	5
0.02	0.15	770	110	42	23	15	10
0.02	0.2	1367	194	74	39	25	17
0.02	0.25	2136	302	115	61	38	26
0.02	0.3	3075	434	165	87	54	37
0.04	0.05	770	23	8	5	3	3
0.04	0.1	3075	87	27	14	8	6
0.04	0.15	6916	194	59	29	17	12
0.04	0.2	12294	343	103	50	29	20
0.04	0.25	19209	535	160	77	45	30
0.04	0.3	27660	770	230	110	64	42
0.06	0.05	770	50	11	5	4	3
0.06	0.1	3075	194	39	17	10	7
0.06	0.15	6916	434	87	37	21	14
0.06	0.2	12294	770	153	64	36	23
0.06	0.25	19209	1202	239	99	55	35
0.06	0.3	27660	1730	343	143	78	50
0.08	0.05	87	194	17	7	4	3
0.08	0.1	343	770	64	23	12	8
0.08	0.15	770	1730	143	50	25	16
0.08	0.2	1367	3075	252	87	44	27
0.08	0.25	2136	4803	393	135	68	41
0.08	0.3	3075	6916	566	194	97	59
0.1	0.05	32	Inf	32	9	5	3
0.1	0.1	124	Inf	124	32	15	9
0.1	0.15	278	Inf	278	71	32	19
0.1	0.2	493	Inf	493	124	56	32
0.1	0.25	770	Inf	770	194	87	50
0.1	0.3	1108	Inf	1108	278	124	71

Table 3.7: Table of sample size n for various percentage of equivalence limits p_θ, percentage of preset mean difference p_δ and CV_R.

For example, suppose paired samples are generated from their normal distributions and their CV of reference group is computed as 0.25. The equivalence limit θ and the preset mean difference are set as 20% and 0% of the reference mean, respectively. So, the sample size for each group should be at least 32 to assess equivalence with sufficient confidence from Table 3.7. It is notable that when p_θ equals to p_δ, the denominator of the sample size formula equals 0, which results in an Inf value.

3.6 Power comparison for equivalence tests

This section aims to compare the probability of two-sample t-test does not reject the null hypothesis of equality and powers of three equivalence tests (probabilities of correctly rejecting the null hypothesis, which is, concluding the equivalence between two groups in equivalence tests of power analysis, TOST, and Hauck-Anderson method) as parameters (mean difference, parameters, distributions, and data kinds) vary while taking 20% of the reference mean as the half width of the equivalence range θ_{fix}. It summarizes the findings for tests that have stable power to assess equivalence when the t-test concludes the preset mean difference has an insignificant difference compared to the population mean difference. The simulation analysis is conducted based on normally and log-normally distributed data separately.

3.6.1 Normally distributed data

Initial settings

Consider two groups with the same sample size ($n = n_R = n_T$) are randomly generated from following:

$$\begin{cases} y_{1_R}, ..., y_{n_R} \sim N(20, 2), \\ y_{1_T}, ..., y_{n_T} \sim N(\mu_T, 2). \end{cases} \tag{3.13}$$

The simulation assumes the test mean is greater than the reference mean. In order to bring the true mean difference within the equivalence region while ensuring the mean difference remains greater than zero, the test mean is set to be slightly greater than the reference group. The test mean is determined by a percentage change from the reference mean as $\mu_T = \mu_R * (1 + p)$.

Testing procedure

Limentani's formula (3.10) indicates that an estimated equivalence limit consists of two terms: a predetermined mean difference δ and the width of an equivalence range θ_{fix}. By taking δ as the desired center of the equivalence range, a t-test for assessing a significant difference between an actual and a preset mean difference is conducted with the following hypothesis:

$$\begin{cases} H'_0 : \mu_T - \mu_R = \delta, \\ H'_A : \mu_T - \mu_R \neq \delta. \end{cases} \tag{3.14}$$

When the p-value of the t-test is greater than the significance level, it implies that the sample data are inadequate to reject the null hypothesis that $\mu_D = \delta$. This conclusion could be a consequence of either high measurement precision or the insignificant magnitude of the difference between μ_D and δ. Then three equivalence tests are conducted as a post-hoc examination. They determine whether sample data has adequate power to draw the reliable inference that the CI of the mean difference falls into the equivalence range $[\theta_1, \theta_2]$ centered at δ. Their hypotheses are shown as follows:

$$\begin{cases} H_0 : \mu_T - \mu_R < \theta_1 \text{ or } \mu_T - \mu_R > \theta_2, \\ H_A : \theta_1 \leq \mu_T - \mu_R \leq \theta_2. \end{cases} \tag{3.15}$$

Simulation setting

After specifying the hypotheses, three parameters are considered to compare the three tests' power performance. They are the sample size for each group $n = n_T = n_R$, the reference group's standard deviation s_R, and a percentage change p of the reference mean as population mean difference μ_D.

The sample sizes for this simulation study are adjusted to 5, 10, 20, 30, and 40. The assigned values for the standard deviation are 0.1, 0.3, 0.5, 0.7, 0.9, 1, 2, 3, 4, 5. A sequence of percentage changes (p) such as 0, 0.05, 0.10, and 0.15 of the reference mean represents the population mean difference $\mu_D = \mu_R - \mu_T = p * \mu_R$. We consider $\delta = \mu_D$, so the probability of concluding an insignificant difference between the preset and actual mean difference is proved to be stable at $(1 - \alpha)100\%$.

The width of the equivalence limit θ_{fix} is set to 20% of the reference mean. The type I and II errors for TOST and Hauck-Anderson methods are fixed at 0.05 to reach an overall type I error value of around 0.025. Because the power analysis is part of the t-test, its error rates are 0.05. There are $4 * 5 * 10 = 200$ parameter combinations, and 1,000 pairs of samples are generated for each parameter setting.

After assigning a series of suitable values for parameters of interest, a t-test is conducted to compare the significant difference between the actual and preset mean difference for each paired sample. When the p-value is greater than the significance level, three equivalence tests are conducted. Their average powers of determining equivalence between two populations under various parameters are summarized in the figures below.

T-test

Figure 3.6 shows an average probability that the test does not reject the null hypothesis that $\mu_D = \delta$ versus sample size, along with an increasing mean difference. Each line represents the probability trend of a t-test with a specific standard deviation value.

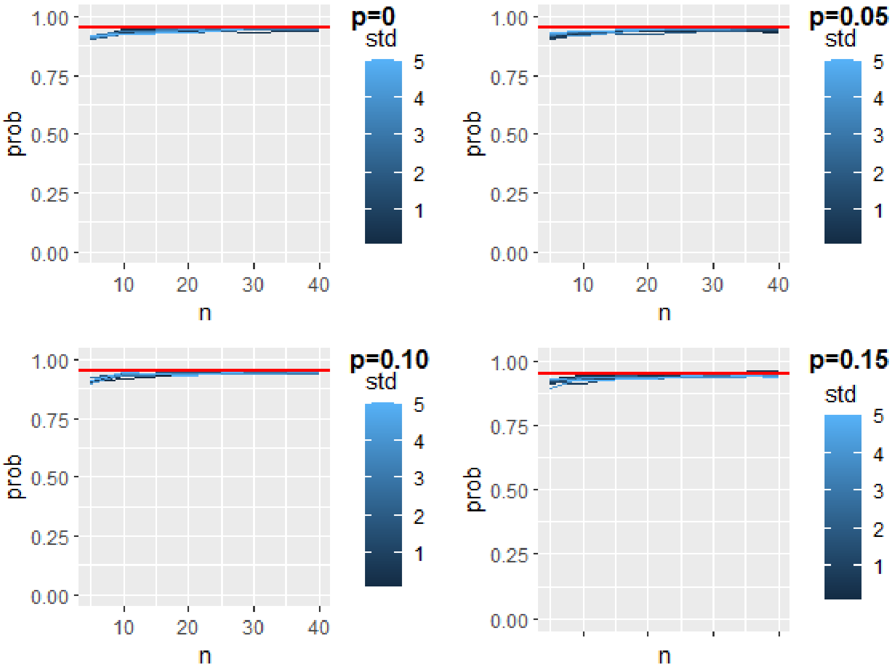


Figure 3.6: Probability of two-sample t-test does not reject the null hypothesis that $\delta = \mu_D$ as parameters sample size ($n_T = n_R =$ 5, 10, 20, 30, 40) and standard deviation ($\sigma_T = \sigma_R =$ 0.1, 0.3, 0.5, 0.7, 0.9, 1, 2, 3, 4, 5) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

As expected, the plot demonstrates that the t-test has a stable probability of an insignificant difference between μ and δ when they are actually equal. It also indicates one of its shortcomings: the t-test does not penalize samples with fewer data points. When a paired sample has more than 1 unit mean difference and the sample group only contains 5 data points, the t-test still has around a 95% chance of not rejecting the null hypothesis of equality.

Equivalence test - power analysis

For those paired samples with a p-value greater than the significance level, equivalence tests are conducted. Figure 3.7 first shows the average power trend of power analysis under various parameters. Given specific equivalence limits and additional constraints on sample parameters, the measured power reflects an observed probability that sample data distinguish if the CI of the mean difference between two groups is within the predetermined equivalence range $[\theta_1, \theta_2]$ is true. It provides evidence against the possibility that they are equivalent due to the high precision of the sampling measurement as well.

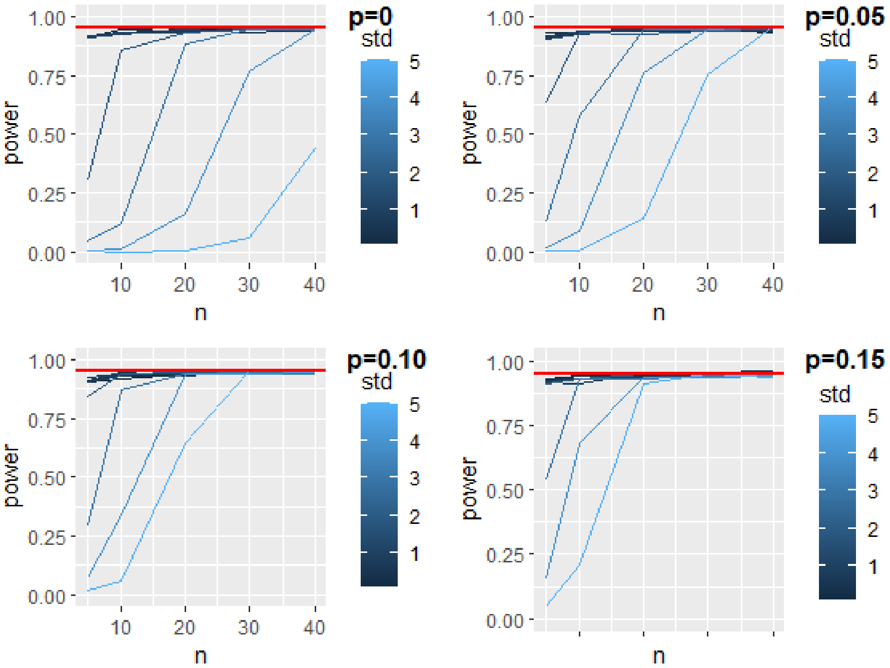


Figure 3.7: Power of power analysis as parameters sample size ($n_T = n_R =$ 5, 10, 20, 30, 40) and standard deviation ($\sigma_T = \sigma_R =$ 0.1, 0.3, 0.5, 0.7, 0.9, 1, 2, 3, 4, 5) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

Figure 3.7 shows how the power analysis penalizes the case of high measurement precision and narrow equivalence ranges. When the sample has a large measurement precision, a test lacks the power to conclude equivalence. For example, the Table 3.4 indicates that a sample with a size of 15 and an equivalence limit of 4 requires the sample standard deviation to be smaller than 3.5 to achieve sufficient power to assess equivalence. So, the line charts clearly show that for samples with more than 15 data points and a standard deviation lower than 3.5, their average powers reach 95%.

We expect that the power trend will remain stable as the population mean difference changes as $\delta = \mu_D$. However, the power trend increases significantly as the population mean difference increasese where the other parameters remain stable.

Equivalence test - TOST

Figures 3.8 and 3.9 illustrate the power performance of the TOST and the Hauck-Anderson test as a function of various parameter values. The results exhibit some comparable trends to the power analysis. Their powers remain constant when the parameters satisfy the conditions from the Table**??**. Both tests are restricted by small equivalence limits, small sample sizes, and great variances. Unlike the power analysis, their power trends are stable as the population mean difference changes.

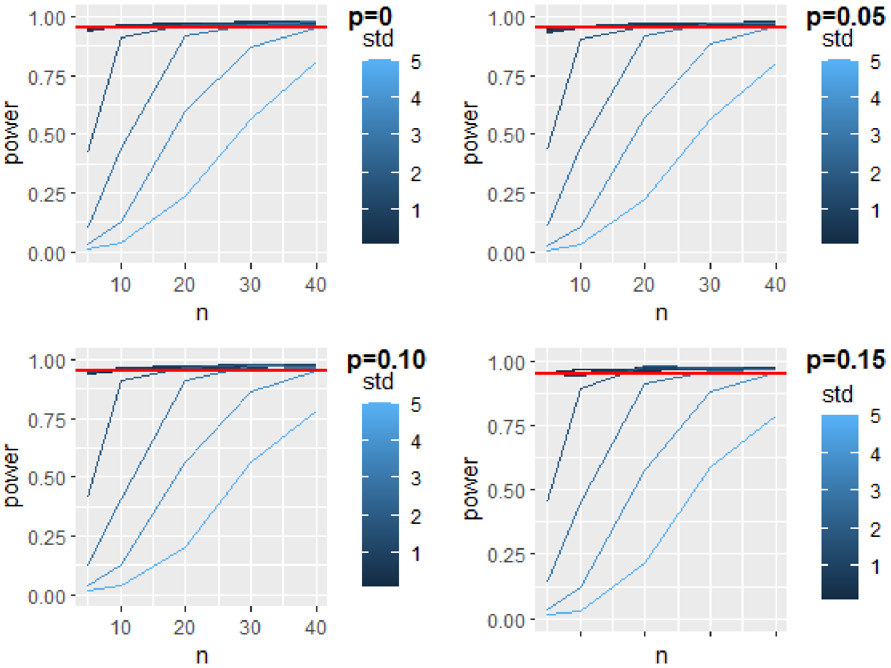


Figure 3.8: Power of TOST as the parameters sample size ($n_T = n_R = 5$, 10, 20, 30, 40) and standard deviation ($\sigma_T = \sigma_R$ =0.1, 0.3, 0.5, 0.7, 0.9, 1, 2, 3, 4, 5) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

Equivalence Test - Hauck-Anderson test

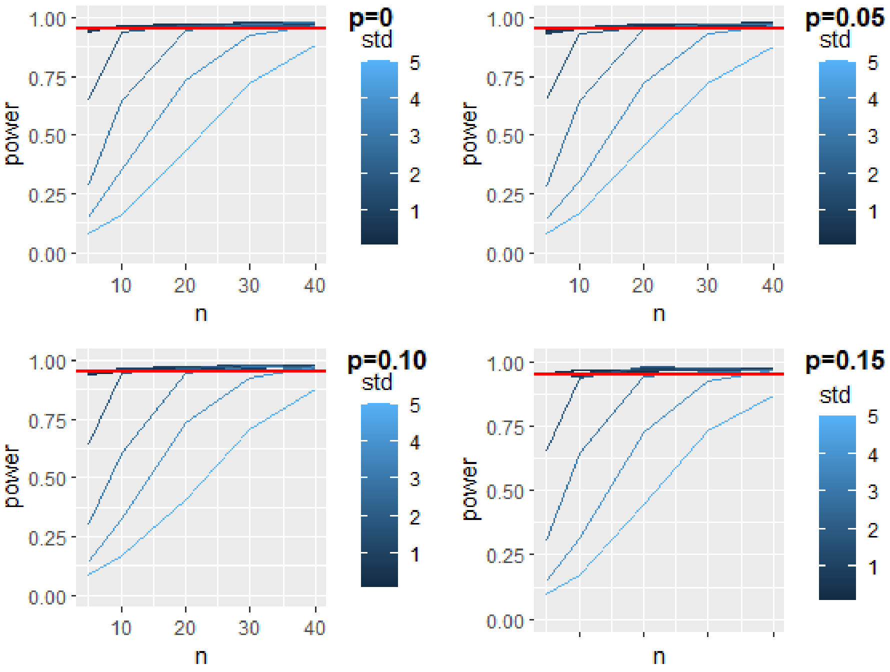


Figure 3.9: Power of Hauck-Anderson method as the parameters sample size ($n_T = n_R = 5$, 10, 20, 30, 40) and standard deviation ($\sigma_T = \sigma_R = 0.1$, 0.3, 0.5, 0.7, 0.9, 1, 2, 3, 4, 5) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

3.6.2 Log-normally distributed data

Most chemicals are log-normally distributed, so the comparison of powers for equivalence tests on the log-scaled data is run as well.

We consider the reference group to be derived from a normal distribution with a mean equal to its log mean and a standard deviation equal to its log standard deviation. We also assume the test group has the same distribution and standard deviation of the logs as its reference group. Moreover, the mean of its test group differs from the mean of its reference group by a specified percentage change, $p = \frac{M_T - M_R}{M_R}$, where M_R and M_T are the means of the logs for reference and test groups. Thus, the population mean for test groups equals $M_T = (1 + p) * M_R$.

By configuring equivalent parameter settings and setting $\sigma_R^2 = \sigma_T^2$ to be the logs of their standard deviation and $n_R = n_T$, data are generated as follows:

$$\begin{cases} \log(y_{1_R}), ..., \log(y_{n_R}) \sim N(5, \sigma_R), \\ \log(y_{1_T}), ..., \log(y_{n_T}) \sim N(M_T, \sigma_T). \end{cases} \tag{3.16}$$

The sample size is set as $n = n_T = n_R = 5$, 10, 20, 30, 40, and the percentage changes of the reference mean are set as 0, 0.05, 0.10, and 0.15. The standard deviation of each group is set at 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. $4 * 5 * 10 = 200$ parameter permutations are conducted, and 1,000 simulations are generated for each parameter setting.

The Appendix A.2 and Appendix A.3 contain figures and summary tables of the power result for each equivalence test.

Tests on the log-normally distributed data show similar results as in the normally distributed case. When $\delta = \mu_D$, the t-test easily demonstrates an insignificant difference between δ and μ_D, even when the sample size is extremely small. So, it cannot offer a conclusive result when samples' measurement precision is inadequate. Power analysis is too strict to conclude in-equivalence, especially when the population mean difference is around 0 and is sensitive to varying mean differences. TOST and the Huack-Anderson method are effective at penalizing simulations with a small sample size, considerable variance, and small equivalence limits. With specified parameter settings based on Limentani's formula, both tests generate stable power to assess equivalence.

3.6.3 Summary for power comparison

The main features of different testing procedures and equivalence limits through simulation analysis are shown as follows:

- When data distribution and specific population parameters are known, taking 20% of the reference mean as an equivalence limit is appropriate for equivalence tests.
- Taking a suitable preset mean difference as a part of the equivalence limit gives equivalence tests a reliable foundation to determine the equivalence, so it is important to employ the t-test as a prior examination for preset and actual mean differences.
- The equivalence limit determination should not only be based on the formula but also be constrained by practical limitations, as concluded in the previous section.
- The power trend of power analysis is varied with various population mean differences. When the difference is close to 0, the power analysis lacks the power on the samples with large measurement precision.
- Both the TOST and Hauck-Anderson methods for normally distributed and log-normally distributed data have comparable power trends to draw conclusions, and they are good at penalizing cases with large measurement precision and small equivalence limits.

Chapter 4

Application of equivalence tests in clinical trials with application to tobacco data

Presently, HC does not have full data from these WHO studies, instead only sparse chemical data collected from different tobacco products' emissions and constituents using HC's measurement method is available. The cigarettes are randomly collected various not only products, but manufacturers to reduce variability. So, this section conducts a simulation-based comparison of equivalence tests for the level of analytical equivalence in clinical trials with their application to tobacco data using official HC measurement methods. It takes various design constraints for the equivalence tests (TOST, Hauck-Anderson test, power analysis) into account and provides useful suggestions for experimental design.

4.1 Introduction to reference data

Chemical data collected from three different cigarette products manufactured by three different companies using the HC reference techniques is taken as the reference group. These chemicals are T-111B (nitrosamines, both under mainstream ISO and extreme smoking circumstances), T-304 (humectants as tobacco's constituents), and T-115 (nicotine and carbon monoxide under mainstream intense smoking conditions).

Data distribution

The distributions of the HC's data on the original and log-transformed scales are graphed in Appendix B. A quantile-quantile (Q-Q) plot was used to determine if the data on the original or log scales is normally distributed.

Appendix B.1 displays the distribution of nitrosamines (NAB, NAT, NNK, and NNN) in conditions of mainstream strong smoking; B.2. displays the distribution of nitrosamines (NAB, NAT, NNK, and NNN) under mainstream ISO smoking settings; B.3. displays the distribution of humectants (glycerol, propylene glycol, and trienthylene glycol), which are components in a cigarette; and B.4. describes the distribution of nicotine and carbon monoxide (CO) under mainstream, intense smoking conditions.

According to these distributions, only CO is normally distributed. All 128 observations of humectant glycerol are beyond the limits of detection (LOD) and quantification (LOQ), whereas triethylene glycol had 110/128 observations below the LOD and propylene glycol had 81/128 observations below the LOD and 3/128 observations below the LOQ. Because their distributions contain more than half of the constants, these data were not used for further simulation analysis. Other chemicals, such as Nicotine, nitrosamines, etc., are fairly log-normally distributed.

Descriptive summary with normal-scaled data

Table 4.1 provides summary statistics of tobacco's emissions and constituent levels on the original scale. The table also includes population size, mean, standard deviations, and coefficient of variation (CV).

Condition	Chemical	Units	N	Mean	Std	CV
Intense	CO	mg/cig	1037	27.08	2.64	0.10
Intense	NAB	ng/cig	140	17.41	9.17	0.53
Intense	NAT	ng/cig	140	118.50	83.05	0.70
Intense	Nicotine	mg/cig	1037	2.10	0.46	0.22
Intense	NNK	ng/cig	140	80.11	61.50	0.77
Intense	NNN	ng/cig	140	111.50	95.73	0.86
ISO	NAB	ng/cig	140	7.06	3.82	0.54
ISO	NAT	ng/cig	140	49.57	34.70	0.70
ISO	NNK	ng/cig	140	31.89	24.34	0.76
ISO	NNN	ng/cig	140	42.34	37.23	0.88

Table 4.1: Table of descriptive statistics for original-scale data on chemicals from tobacco products.

Descriptive summary with logged-scaled data

Because the data have a lognormal distribution, Table 4.2 presents their log scale summary statistics (CO is excluded from the table since the data is normally distributed). The columns of mean, SD, and CV are computed based on their log values. Their original means (Geometric Mean, GM) and standard deviations (Geometric Standard Deviation, GSD) are computed using the formulas $E(X) = e^{\mu+\frac{\sigma^2}{2}}$ and $var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$. It is clear that GM and GSD values are fairly close to the mean and variance of the original-scale data.

Condition	chemical	units	N	log_mean	log_std	log_CV	GM	GSD
Intense	CO	mg/cig	1037	-	-	-	-	-
Intense	NAT	ng/cig	140	4.54	0.69	0.15	118.87	92.82
Intense	NNK	ng/cig	140	4.17	0.61	0.15	77.95	52.34
Intense	NNN	ng/cig	140	4.43	0.71	0.16	107.99	87.43
Intense	NAB	ng/cig	140	2.72	0.53	0.19	17.47	9.95
ISO	NAT	ng/cig	140	3.63	0.81	0.22	52.36	50.42
ISO	NNK	ng/cig	140	3.2	0.75	0.23	32.5	28.24
ISO	NNN	ng/cig	140	3.44	0.8	0.23	42.95	40.67
ISO	NAB	ng/cig	140	1.82	0.54	0.3	7.14	4.15
Intense	Nicotine	mg/cig	1037	0.72	0.23	0.32	2.11	0.49

Table 4.2: Table of descriptive statistics for log-scale data on chemicals from tobacco products.

4.2 Equivalence tests with sample size determination

When the data distribution is determined from the sampling histogram, the half width of the equivalence limits θ_{fix} and the preset mean difference δ should be determined by experts and statisticians together. In this case, we take θ_{fix} as 20% of the reference mean and $\delta = 0$. Then the sample size for each chemical is calculated based on the Table 3.7. It should be noted that calculated size only provides the minimum boundary, and it is always good to have more sample data.

After an appropriate equivalence limit and sample size for each tobacco chemical are determined, a t-test is conducted as a prior examination to compare the actual and preset mean differences. Three equivalence tests (Power analysis, TOST, and the Hauck-Anderson method) with assigned parameter values are used when the p-value is greater than the significance level. The power for each equivalence test represents the probability that sample data is capable of showing the CI of population mean difference is within an acceptable equivalence range, while the sample does not have sufficient ability to infer a significant difference between the preset δ and the actual population mean difference.

Figures 4.1, 4.2 and 4.3 visualize power trends for determining equivalence between each chemical data set from two different methodologies versus the number of simulated pairs. Each line represents the average power of concluding a insignificant mean difference or equivalence between two groups. The calculated sample size for each simulated pair is also shown in each figure.

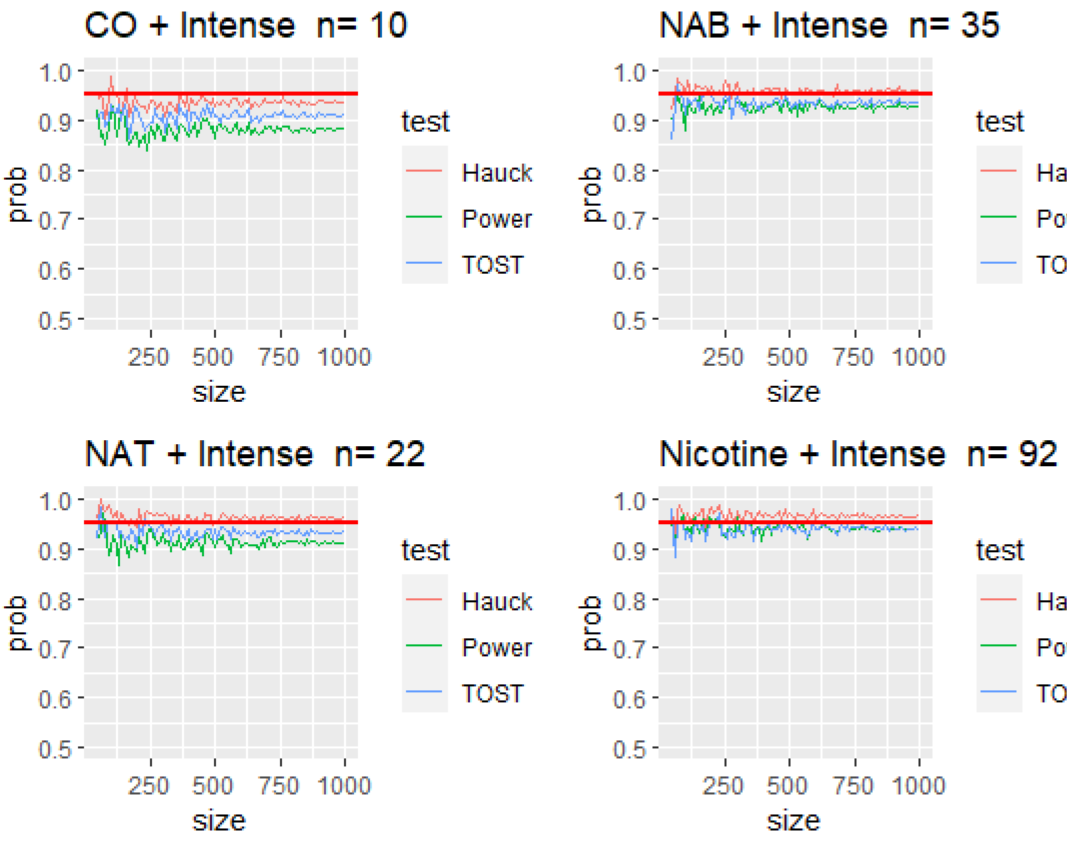
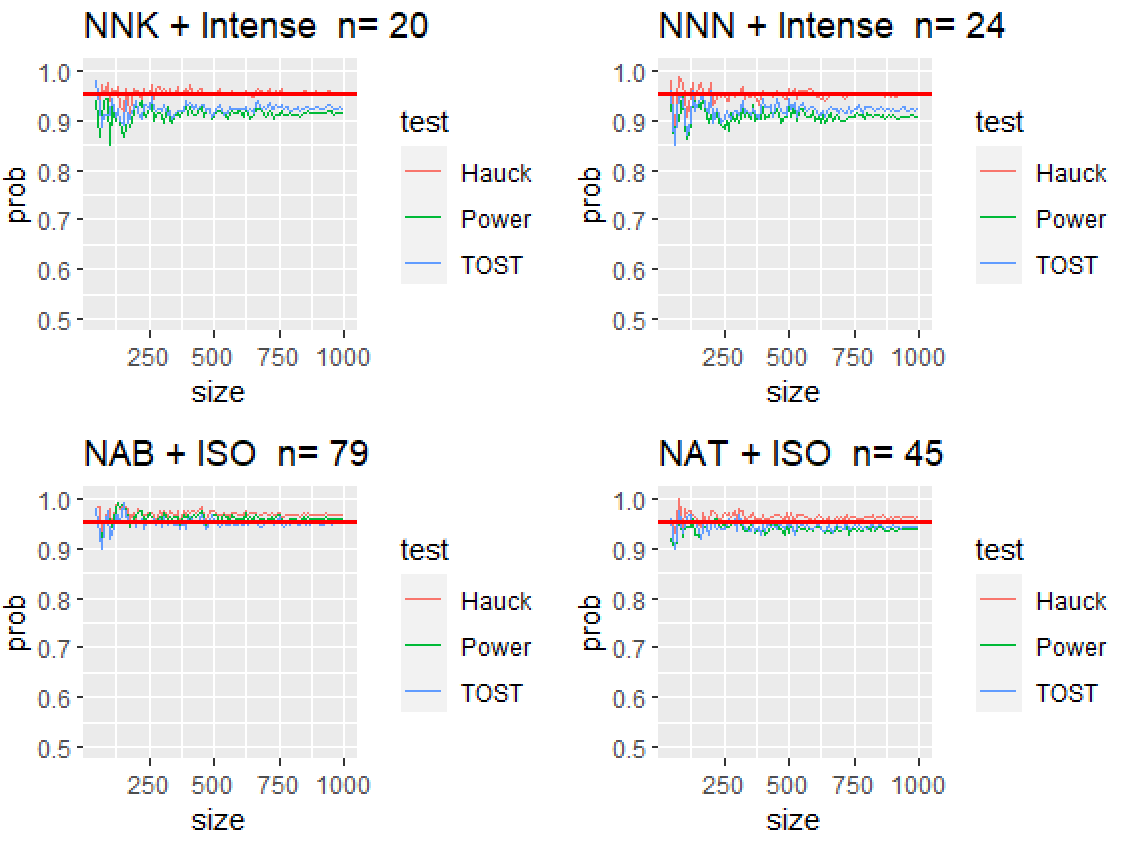


Figure 4.1: Power of equivalence tests with a fixed equivalence limit and a computed sample size versus the number of simulated pairs for chemicals a) CO under intense condition; b) NAB under intense condition; c) NAT under intense condition and d) nicotine under intense condition.



NNK + Intense n= 20

NNN + Intense n= 24

NAB + ISO n= 79

NAT + ISO n= 45

Figure 4.2: Power of equivalence tests with a fixed equivalence limit and a computed sample size versus the number of simulated pairs for chemicals a) NNK under intense condition; b) NNN under intense condition; c) NAB under ISO condition and d) NAT under ISO condition.

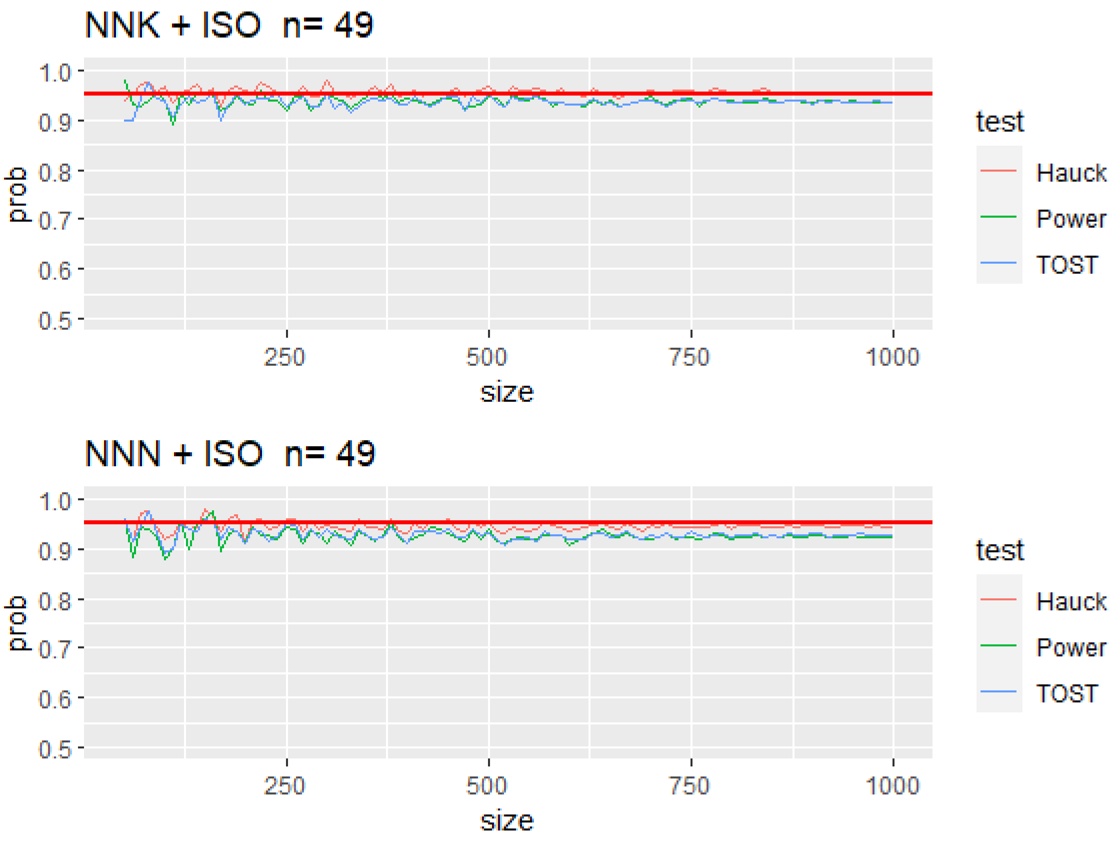


Figure 4.3: Power of equivalence tests with a fixed equivalence limit and a computed sample size versus the number of simulated pairs for chemicals a) NNK under ISO condition; b) NNN under ISO condition.

As expected, the power trends of power analysis, TOST, and the Hauck-Anderson method are stable at 95% to conclude the equivalence between two groups, especially when the desired sample size for each group is greater than 30. When the sample size is less than 30, it is clear that the three equivalence tests have lower rates to conclude equivalence,especially the power analysis. It strongly indicates that they are strict with a small sample. Moreover, power trends for most chemicals vary significantly when the number of simulated pairs is less than 250.

Chapter 5

Conclusions

This report first illustrates that the traditional two-sample t-test is not a reliable way to assess equivalence between two groups. One reason is that it only determines a significant mean difference. But, failure to reject the null hypothesis does not imply that the means of two groups are equal or similar. A higher measurement precision leads to a greater probability of failing to reject the null hypothesis of inequality.

So, three equivalence approaches are examined in this report: power analysis, TOST, and the Hauck-Anderson method. Each assumes that the two groups have a significant mean difference and then uses sample data to assess equivalence.

The equivalence limit θ and sample size are important parameters for tests; summary tables based on Limentani's formula present their trade-offs with other key parameters. By taking the t-test as a prior examination for the preset mean difference δ, the simulation is conducted to compare the power trends of three equivalence tests for normally and log-normally distributed data across a variety of parameter values. The figures show that they all eventually yield comparable testing capabilities of $(1-\alpha)100\%$ on concluding the equivalence with varying actual mean differences, while power analysis is more sensitive in terms of measurement precision and actual mean differences.

The simulation-based comparison of equivalence tests for the level of analytical equivalence in clinical trials with their application to tobacco data using official HC measurement methods is presented in the last section. The final figures show that the three tests have comparable power and they always penalize cases with a small sample size. Some restrictions on parameter settings and testing procedures are shown as follows:

- There are trade-offs between the equivalence limit and sample size with other parameters, so one can be determined by the summary table when the others are known. Some restrictions on equivalence limit and sample size determination are summarized in this report.

- An acceptable mean difference δ should be determined based on sample statistics and the subject expert's practical experiences, and a t-test should be conducted for the significant difference between the preset and actual mean differences.

- When the t-test shows not enough evidence to reject the initial assumption of equality, an equivalence test with pre-defined parameters is undertaken. Its p-value represents the probability that the sample data has adequate power to suggest a fitting preset mean difference and that its value is within an acceptable equivalence range.

- Equivalence tests are strict on the measurement precision and their equivalence limits can also be suggested by the experimental experiences, so they are more suitable to assess equivalence compared to the t-test.

This report conducts a simulation-based experimental design to determine equivalence with the application of HC's official measurement method for tobacco chemical data. It only considers the case where the measured values using two methods under the specific data distributions have the same population means and variances are independent of each other, and their means are the same. So, future work on this project would include exploring other data types. It is also important to figure out the difference among the three equivalence tests from the perspective of formalizing their power calculations, etc.

Bibliography

[1] ISO 2077. Cigarettes — routine analytical cigarette smoking machine — definitions and standard conditions with an intense smoking regime. 2018.

[2] Lakens Daniël. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. 2017.

[3] Center for Drug Evaluation and Research. Bioavailability studies submitted in ndas or inds – general considerations. 2022.

[4] W. Hauck, W and S. Anderson. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. 1984.

[5] M. C.and Ye F. Bergquist M. L. Limentani, G. B.and Ringo and E. O. MCSorley. Beyond the t-test: Statistical equivalence testing. 2005.

[6] D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. 1987.

[7] Tobacco Control Directorate (TCD). Regulations amending the tobacco reporting regulations. 2022.

[8] W. J. Westlake. Response to t.b.l. kirkwood:bioequivlaence testing – a need to rethink. 1981.

Appendix A

The validation of power for equivalence test

A.1 Tables of powers for equivalence tests with normally-distributed data

A.1.1 0% change between reference and test methods for various theta $(\mu_T - \mu_R = 0)$

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	0.003	-0.162	0.168	0.92	0.92	0.949	0.949
5	0.3	0.011	-0.47	0.493	0.91	0.91	0.946	0.946
5	0.5	0.005	-0.835	0.846	0.912	0.912	0.941	0.941
5	0.7	-0.012	-1.169	1.144	0.921	0.921	0.954	0.954
5	0.9	-0.025	-1.473	1.424	0.918	0.918	0.95	0.95
5	1	-0.023	-1.671	1.625	0.911	0.91	0.94	0.943
5	2	0.003	-3.255	3.261	0.907	0.303	0.422	0.649
5	3	0.023	-4.927	4.974	0.916	0.043	0.104	0.282
5	4	0.073	-6.591	6.737	0.915	0.002	0.03	0.149
5	5	0.038	-8.141	8.218	0.921	0.002	0.012	0.078
10	0.1	0.003	-0.103	0.109	0.935	0.935	0.967	0.967
10	0.3	-0.002	-0.32	0.315	0.943	0.943	0.967	0.967
10	0.5	0.016	-0.521	0.553	0.936	0.936	0.966	0.966
10	0.7	0.008	-0.745	0.762	0.93	0.93	0.953	0.953
10	0.9	0.002	-0.953	0.958	0.933	0.933	0.96	0.96
10	1	0.019	-1.049	1.086	0.923	0.923	0.962	0.962
10	2	-0.001	-2.114	2.111	0.926	0.857	0.91	0.938
10	3	-0.026	-3.216	3.164	0.937	0.124	0.441	0.644
10	4	0.018	-4.224	4.26	0.924	0.011	0.13	0.352
10	5	-0.005	-5.324	5.315	0.926	0	0.036	0.162
20	0.1	0.001	-0.072	0.073	0.943	0.943	0.97	0.97
20	0.3	-0.005	-0.224	0.215	0.95	0.95	0.974	0.974
20	0.5	-0.001	-0.366	0.363	0.943	0.943	0.977	0.977
20	0.7	0.002	-0.505	0.509	0.941	0.941	0.969	0.969
20	0.9	0.024	-0.635	0.683	0.936	0.936	0.964	0.964
20	1	0.003	-0.723	0.73	0.944	0.944	0.968	0.968
20	2	0.012	-1.455	1.479	0.93	0.93	0.966	0.966
20	3	0.018	-2.159	2.194	0.945	0.888	0.916	0.947
20	4	-0.006	-2.915	2.903	0.944	0.163	0.6	0.735
20	5	-0.005	-3.646	3.636	0.933	0.006	0.234	0.438
30	0.1	0	-0.059	0.059	0.945	0.945	0.97	0.97
30	0.3	0.005	-0.172	0.181	0.948	0.948	0.971	0.971
30	0.5	-0.008	-0.302	0.287	0.931	0.931	0.958	0.958
30	0.7	0.003	-0.41	0.416	0.945	0.945	0.974	0.974
30	0.9	0.002	-0.527	0.532	0.945	0.945	0.979	0.979
30	1	0	-0.59	0.591	0.961	0.961	0.979	0.979
30	2	-0.015	-1.194	1.164	0.949	0.949	0.97	0.97
30	3	-0.013	-1.774	1.748	0.943	0.943	0.97	0.972
30	4	-0.008	-2.374	2.359	0.957	0.769	0.872	0.928
30	5	0.028	-2.911	2.968	0.955	0.06	0.567	0.724
40	0.1	0	-0.051	0.051	0.952	0.952	0.977	0.977
40	0.3	-0.001	-0.153	0.15	0.94	0.94	0.968	0.968
40	0.5	-0.004	-0.257	0.25	0.948	0.948	0.971	0.971
40	0.7	0	-0.357	0.357	0.939	0.939	0.959	0.959
40	0.9	0.008	-0.449	0.464	0.957	0.957	0.978	0.978
40	1	0.001	-0.508	0.51	0.948	0.948	0.975	0.975
40	2	-0.001	-1.012	1.011	0.943	0.943	0.969	0.969
40	3	-0.006	-1.535	1.523	0.945	0.945	0.978	0.978
40	4	0.028	-2.002	2.058	0.952	0.949	0.956	0.967
40	5	0.01	-2.53	2.549	0.956	0.445	0.808	0.884

A.1.2 5% change between reference and test methods for varying values of theta ($\mu_T - \mu_R = 1$)

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	1	0.837	1.163	0.931	0.931	0.952	0.952
5	0.3	0.999	0.507	1.492	0.909	0.909	0.947	0.947
5	0.5	1.006	0.209	1.802	0.919	0.919	0.941	0.941
5	0.7	0.993	-0.155	2.141	0.906	0.906	0.933	0.933
5	0.9	0.993	-0.512	2.497	0.911	0.911	0.937	0.937
5	1	0.987	-0.642	2.616	0.905	0.905	0.93	0.936
5	2	0.97	-2.314	4.255	0.925	0.634	0.432	0.656
5	3	0.966	-4.078	6.011	0.92	0.125	0.109	0.281
5	4	1.023	-5.61	7.656	0.926	0.021	0.028	0.142
5	5	0.948	-7.383	9.28	0.93	0.003	0.006	0.079
10	0.1	1.002	0.896	1.108	0.932	0.932	0.957	0.957
10	0.3	1.001	0.682	1.319	0.927	0.927	0.955	0.955
10	0.5	0.99	0.455	1.524	0.933	0.933	0.957	0.957
10	0.7	1	0.249	1.751	0.926	0.926	0.958	0.958
10	0.9	1	0.053	1.947	0.938	0.938	0.961	0.961
10	1	1	-0.057	2.057	0.926	0.926	0.95	0.95
10	2	0.997	-1.152	3.145	0.936	0.935	0.905	0.934
10	3	1.018	-2.169	4.204	0.938	0.58	0.446	0.646
10	4	0.979	-3.278	5.236	0.921	0.084	0.107	0.308
10	5	0.955	-4.3	6.209	0.938	0.007	0.034	0.166
20	0.1	0.999	0.927	1.072	0.943	0.943	0.968	0.968
20	0.3	0.997	0.779	1.215	0.946	0.946	0.975	0.975
20	0.5	1.009	0.646	1.371	0.938	0.938	0.969	0.969
20	0.7	1.008	0.495	1.522	0.946	0.946	0.974	0.974
20	0.9	0.994	0.34	1.648	0.948	0.948	0.973	0.973
20	1	0.988	0.255	1.721	0.96	0.96	0.977	0.977
20	2	1.002	-0.459	2.463	0.927	0.927	0.96	0.96
20	3	0.994	-1.173	3.161	0.948	0.948	0.916	0.952
20	4	0.963	-1.96	3.887	0.944	0.765	0.569	0.723
20	5	0.896	-2.749	4.541	0.945	0.145	0.221	0.457
30	0.1	1.001	0.941	1.06	0.945	0.945	0.968	0.968
30	0.3	0.999	0.822	1.176	0.94	0.94	0.97	0.97
30	0.5	0.996	0.7	1.293	0.948	0.948	0.968	0.968
30	0.7	0.994	0.582	1.407	0.953	0.953	0.972	0.972
30	0.9	1.012	0.478	1.546	0.946	0.946	0.971	0.971
30	1	1.001	0.409	1.593	0.951	0.951	0.974	0.974
30	2	0.998	-0.174	2.171	0.94	0.94	0.96	0.96
30	3	0.967	-0.802	2.736	0.944	0.944	0.974	0.974
30	4	0.945	-1.407	3.298	0.949	0.949	0.882	0.932
30	5	0.996	-1.967	3.958	0.962	0.755	0.564	0.718
40	0.1	0.999	0.948	1.049	0.931	0.931	0.966	0.966
40	0.3	1.002	0.85	1.154	0.949	0.949	0.969	0.969
40	0.5	0.999	0.744	1.254	0.937	0.937	0.964	0.964
40	0.7	1	0.644	1.356	0.953	0.953	0.981	0.981
40	0.9	1.001	0.543	1.459	0.949	0.949	0.968	0.968
40	1	1.005	0.494	1.515	0.955	0.955	0.976	0.976
40	2	1.019	0.005	2.033	0.943	0.943	0.963	0.963
40	3	1.012	-0.516	2.539	0.954	0.954	0.976	0.976
40	4	1.032	-0.998	3.063	0.961	0.961	0.958	0.97
40	5	1.03	-1.502	3.563	0.958	0.958	0.806	0.88

A.1.3 10% change between reference and test methods for varying values of theta ($\mu_T - \mu_R = 2$)

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	1.998	1.833	2.162	0.908	0.908	0.938	0.938
5	0.3	1.988	1.499	2.478	0.913	0.913	0.95	0.95
5	0.5	1.997	1.187	2.808	0.914	0.914	0.944	0.944
5	0.7	1.986	0.847	3.124	0.915	0.915	0.949	0.949
5	0.9	2.002	0.521	3.483	0.908	0.908	0.94	0.945
5	1	2.01	0.372	3.648	0.924	0.924	0.944	0.949
5	2	1.949	-1.32	5.218	0.926	0.846	0.415	0.642
5	3	1.946	-2.944	6.836	0.925	0.294	0.123	0.301
5	4	1.928	-4.542	8.398	0.898	0.076	0.036	0.141
5	5	2.003	-6.093	10.099	0.912	0.016	0.02	0.085
10	0.1	2.001	1.894	2.107	0.921	0.921	0.951	0.951
10	0.3	2.005	1.685	2.325	0.933	0.933	0.954	0.954
10	0.5	2.002	1.467	2.538	0.942	0.942	0.97	0.97
10	0.7	2.008	1.252	2.764	0.947	0.947	0.963	0.963
10	0.9	1.983	1.02	2.947	0.934	0.934	0.967	0.967
10	1	1.991	0.919	3.063	0.942	0.942	0.967	0.967
10	2	1.993	-0.126	4.112	0.942	0.942	0.909	0.949
10	3	2.021	-1.199	5.242	0.935	0.868	0.411	0.604
10	4	2.108	-2.103	6.318	0.94	0.342	0.126	0.325
10	5	1.964	-3.398	7.325	0.946	0.063	0.041	0.168
20	0.1	2	1.928	2.073	0.942	0.942	0.965	0.965
20	0.3	2.005	1.787	2.223	0.94	0.94	0.969	0.969
20	0.5	2.005	1.637	2.373	0.938	0.938	0.969	0.969
20	0.7	1.994	1.482	2.506	0.936	0.936	0.962	0.962
20	0.9	2.01	1.35	2.669	0.955	0.955	0.974	0.974
20	1	1.989	1.265	2.713	0.94	0.94	0.972	0.972
20	2	1.988	0.53	3.445	0.947	0.947	0.966	0.966
20	3	2.01	-0.182	4.201	0.937	0.937	0.913	0.946
20	4	1.996	-0.92	4.912	0.942	0.94	0.566	0.735
20	5	2.03	-1.611	5.671	0.932	0.649	0.202	0.409
30	0.1	2	1.941	2.059	0.943	0.943	0.976	0.976
30	0.3	2.003	1.827	2.178	0.947	0.947	0.966	0.966
30	0.5	1.999	1.704	2.294	0.947	0.947	0.974	0.974
30	0.7	1.998	1.587	2.408	0.95	0.95	0.975	0.975
30	0.9	2.005	1.479	2.532	0.939	0.939	0.966	0.966
30	1	2.002	1.412	2.591	0.948	0.948	0.978	0.978
30	2	1.99	0.803	3.177	0.953	0.953	0.982	0.982
30	3	1.976	0.207	3.744	0.944	0.944	0.971	0.973
30	4	1.996	-0.347	4.34	0.938	0.938	0.866	0.927
30	5	2.024	-0.917	4.965	0.955	0.955	0.562	0.71
40	0.1	2	1.949	2.051	0.949	0.949	0.974	0.974
40	0.3	1.997	1.844	2.149	0.952	0.952	0.979	0.979
40	0.5	1.999	1.745	2.253	0.95	0.95	0.968	0.968
40	0.7	1.991	1.636	2.347	0.943	0.943	0.971	0.971
40	0.9	2.003	1.547	2.459	0.952	0.952	0.98	0.98
40	1	1.998	1.49	2.507	0.957	0.957	0.981	0.981
40	2	1.995	0.977	3.013	0.948	0.948	0.972	0.972
40	3	2.007	0.494	3.52	0.938	0.938	0.965	0.965
40	4	1.994	-0.035	4.024	0.944	0.944	0.955	0.971
40	5	2.004	-0.553	4.56	0.96	0.96	0.785	0.876

A.1.4 15% change between reference and test methods for varying values of theta $(\mu_T - \mu_R = 3)$

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	3.001	2.838	3.165	0.93	0.93	0.955	0.955
5	0.3	2.997	2.507	3.487	0.922	0.922	0.952	0.952
5	0.5	3.017	2.207	3.826	0.925	0.925	0.956	0.956
5	0.7	2.999	1.851	4.147	0.912	0.912	0.949	0.949
5	0.9	3.003	1.535	4.471	0.916	0.916	0.953	0.955
5	1	3.011	1.37	4.652	0.919	0.919	0.949	0.954
5	2	2.974	-0.318	6.265	0.924	0.911	0.45	0.651
5	3	3.076	-1.782	7.934	0.924	0.541	0.138	0.303
5	4	3.036	-3.513	9.586	0.895	0.154	0.036	0.147
5	5	2.916	-5.201	11.033	0.928	0.05	0.019	0.097
10	0.1	3.003	2.897	3.109	0.94	0.94	0.966	0.966
10	0.3	3.002	2.682	3.322	0.946	0.946	0.971	0.971
10	0.5	2.996	2.468	3.524	0.93	0.93	0.957	0.957
10	0.7	2.989	2.244	3.733	0.93	0.93	0.956	0.956
10	0.9	3.001	2.048	3.954	0.928	0.928	0.956	0.956
10	1	3.014	1.956	4.073	0.911	0.911	0.943	0.943
10	2	2.981	0.873	5.088	0.932	0.932	0.893	0.936
10	3	3.036	-0.135	6.208	0.935	0.932	0.452	0.647
10	4	2.936	-1.298	7.17	0.936	0.681	0.121	0.313
10	5	2.85	-2.352	8.051	0.924	0.212	0.028	0.17
20	0.1	3	2.927	3.073	0.936	0.936	0.966	0.966
20	0.3	3.005	2.788	3.221	0.934	0.934	0.961	0.961
20	0.5	2.997	2.632	3.362	0.943	0.943	0.972	0.972
20	0.7	2.992	2.485	3.5	0.937	0.937	0.967	0.967
20	0.9	3.002	2.344	3.661	0.938	0.938	0.968	0.968
20	1	2.995	2.263	3.728	0.955	0.955	0.974	0.974
20	2	2.981	1.521	4.442	0.95	0.95	0.979	0.979
20	3	3.03	0.87	5.19	0.932	0.932	0.911	0.942
20	4	2.964	0.03	5.898	0.937	0.937	0.574	0.725
20	5	2.943	-0.686	6.572	0.931	0.912	0.216	0.446
30	0.1	3	2.941	3.059	0.945	0.945	0.968	0.968
30	0.3	2.998	2.822	3.175	0.951	0.951	0.976	0.976
30	0.5	3.004	2.707	3.3	0.953	0.953	0.974	0.974
30	0.7	3.005	2.591	3.419	0.937	0.937	0.97	0.97
30	0.9	3.003	2.471	3.534	0.947	0.947	0.976	0.976
30	1	2.997	2.407	3.588	0.945	0.945	0.968	0.968
30	2	2.986	1.799	4.173	0.95	0.95	0.975	0.975
30	3	2.993	1.218	4.768	0.94	0.94	0.963	0.964
30	4	3.006	0.661	5.352	0.938	0.938	0.879	0.927
30	5	2.963	0.048	5.879	0.95	0.95	0.586	0.73
40	0.1	2.999	2.949	3.05	0.945	0.945	0.977	0.977
40	0.3	2.998	2.845	3.151	0.945	0.945	0.971	0.971
40	0.5	2.994	2.738	3.25	0.959	0.959	0.977	0.977
40	0.7	3	2.645	3.356	0.94	0.94	0.969	0.969
40	0.9	3.001	2.544	3.458	0.942	0.942	0.966	0.966
40	1	3.005	2.496	3.514	0.953	0.953	0.972	0.972
40	2	2.999	1.982	4.016	0.943	0.943	0.97	0.97
40	3	2.983	1.449	4.517	0.948	0.948	0.971	0.971
40	4	2.94	0.912	4.967	0.941	0.941	0.955	0.971
40	5	3.011	0.474	5.548	0.937	0.937	0.787	0.868

A.2 Tables of powers for equivalence tests with log-normally distributed data

A.2.1 0% change between reference and test methods for varying values of theta $(M_T - M_R = 0)$

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	0.003	-0.162	0.168	0.92	0.92	0.949	0.949
5	0.2	0.008	-0.313	0.329	0.91	0.91	0.946	0.946
5	0.3	0.003	-0.501	0.507	0.912	0.878	0.87	0.926
5	0.4	-0.007	-0.668	0.654	0.921	0.607	0.668	0.821
5	0.5	-0.014	-0.818	0.791	0.918	0.338	0.444	0.652
5	0.6	-0.014	-1.003	0.975	0.911	0.126	0.248	0.468
5	0.7	0.001	-1.139	1.141	0.907	0.041	0.138	0.326
5	0.8	0.006	-1.314	1.326	0.916	0.021	0.082	0.247
5	0.9	0.016	-1.483	1.516	0.915	0.003	0.047	0.183
5	1	0.008	-1.628	1.644	0.921	0.008	0.028	0.13
10	0.1	0.003	-0.103	0.109	0.935	0.935	0.967	0.967
10	0.2	-0.001	-0.213	0.21	0.943	0.943	0.967	0.967
10	0.3	0.01	-0.313	0.332	0.936	0.936	0.966	0.966
10	0.4	0.005	-0.426	0.435	0.93	0.929	0.947	0.951
10	0.5	0.001	-0.53	0.532	0.933	0.874	0.895	0.935
10	0.6	0.011	-0.629	0.652	0.923	0.549	0.761	0.862
10	0.7	0	-0.74	0.739	0.926	0.254	0.531	0.716
10	0.8	-0.007	-0.858	0.844	0.937	0.071	0.35	0.58
10	0.9	0.004	-0.95	0.958	0.924	0.03	0.226	0.448
10	1	-0.001	-1.065	1.063	0.926	0.008	0.122	0.314
20	0.1	0.001	-0.072	0.073	0.943	0.943	0.97	0.97
20	0.2	-0.003	-0.15	0.143	0.95	0.95	0.974	0.974
20	0.3	-0.001	-0.219	0.218	0.943	0.943	0.977	0.977
20	0.4	0.001	-0.289	0.291	0.941	0.941	0.969	0.969
20	0.5	0.013	-0.353	0.379	0.936	0.936	0.964	0.964
20	0.6	0.002	-0.434	0.438	0.944	0.944	0.967	0.968
20	0.7	0.004	-0.509	0.518	0.93	0.912	0.931	0.955
20	0.8	0.005	-0.576	0.585	0.945	0.779	0.871	0.925
20	0.9	-0.001	-0.656	0.653	0.944	0.416	0.734	0.827
20	1	-0.001	-0.729	0.727	0.933	0.139	0.564	0.736
30	0.1	0	-0.059	0.059	0.945	0.945	0.97	0.97
30	0.2	0.003	-0.114	0.12	0.948	0.948	0.971	0.971
30	0.3	-0.005	-0.181	0.172	0.931	0.931	0.958	0.958
30	0.4	0.001	-0.235	0.237	0.945	0.945	0.974	0.974
30	0.5	0.001	-0.293	0.295	0.945	0.945	0.979	0.979
30	0.6	0	-0.354	0.354	0.961	0.961	0.979	0.979
30	0.7	-0.005	-0.418	0.407	0.949	0.949	0.97	0.97
30	0.8	-0.003	-0.473	0.466	0.943	0.943	0.969	0.971
30	0.9	-0.002	-0.534	0.531	0.957	0.946	0.936	0.963
30	1	0.006	-0.582	0.594	0.955	0.761	0.886	0.939
40	0.1	0	-0.051	0.051	0.952	0.952	0.977	0.977
40	0.2	-0.001	-0.102	0.1	0.94	0.94	0.968	0.968
40	0.3	-0.002	-0.154	0.15	0.948	0.948	0.971	0.971
40	0.4	0	-0.204	0.204	0.939	0.939	0.959	0.959
40	0.5	0.004	-0.249	0.258	0.957	0.957	0.978	0.978
40	0.6	0.001	-0.305	0.306	0.948	0.948	0.975	0.975
40	0.7	0	-0.354	0.354	0.943	0.943	0.969	0.969
40	0.8	-0.002	-0.409	0.406	0.945	0.945	0.978	0.978
40	0.9	0.006	-0.451	0.463	0.952	0.952	0.973	0.975
40	1	0.002	-0.506	0.51	0.956	0.955	0.961	0.974

A.2.2 5% change between reference and test methods for varying values of theta $(M_T - M_R = 0.25)$

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	0.251	0.086	0.415	0.92	0.92	0.953	0.953
5	0.2	0.253	-0.077	0.583	0.91	0.91	0.935	0.935
5	0.3	0.243	-0.252	0.739	0.924	0.923	0.909	0.944
5	0.4	0.251	-0.414	0.916	0.925	0.865	0.652	0.827
5	0.5	0.272	-0.555	1.099	0.924	0.635	0.416	0.657
5	0.6	0.241	-0.721	1.204	0.905	0.367	0.27	0.485
5	0.7	0.227	-0.94	1.395	0.925	0.197	0.123	0.315
5	0.8	0.269	-1.048	1.586	0.91	0.094	0.084	0.233
5	0.9	0.237	-1.19	1.664	0.92	0.055	0.068	0.194
5	1	0.245	-1.395	1.885	0.912	0.021	0.022	0.142
10	0.1	0.252	0.146	0.358	0.932	0.932	0.965	0.965
10	0.2	0.251	0.038	0.463	0.937	0.937	0.961	0.961
10	0.3	0.248	-0.075	0.571	0.941	0.941	0.969	0.969
10	0.4	0.249	-0.177	0.674	0.932	0.932	0.954	0.959
10	0.5	0.245	-0.286	0.776	0.94	0.939	0.903	0.94
10	0.6	0.246	-0.392	0.884	0.939	0.893	0.764	0.873
10	0.7	0.255	-0.489	1	0.932	0.712	0.516	0.703
10	0.8	0.266	-0.58	1.112	0.936	0.426	0.358	0.563
10	0.9	0.253	-0.715	1.222	0.938	0.199	0.199	0.409
10	1	0.254	-0.811	1.319	0.943	0.07	0.127	0.328
20	0.1	0.251	0.179	0.324	0.942	0.942	0.968	0.968
20	0.2	0.248	0.101	0.394	0.929	0.929	0.961	0.961
20	0.3	0.248	0.03	0.467	0.947	0.947	0.97	0.97
20	0.4	0.252	-0.039	0.544	0.944	0.944	0.97	0.97
20	0.5	0.246	-0.124	0.615	0.955	0.955	0.98	0.98
20	0.6	0.259	-0.179	0.698	0.947	0.947	0.968	0.972
20	0.7	0.264	-0.245	0.773	0.93	0.93	0.934	0.952
20	0.8	0.246	-0.337	0.828	0.953	0.953	0.884	0.942
20	0.9	0.24	-0.414	0.894	0.951	0.926	0.737	0.848
20	1	0.236	-0.494	0.967	0.944	0.753	0.568	0.717
30	0.1	0.249	0.19	0.308	0.96	0.96	0.973	0.973
30	0.2	0.248	0.13	0.366	0.946	0.946	0.971	0.971
30	0.3	0.249	0.072	0.426	0.942	0.942	0.966	0.966
30	0.4	0.251	0.016	0.486	0.94	0.94	0.972	0.972
30	0.5	0.254	-0.041	0.549	0.937	0.937	0.961	0.961
30	0.6	0.246	-0.106	0.598	0.948	0.948	0.968	0.968
30	0.7	0.25	-0.164	0.663	0.955	0.955	0.978	0.978
30	0.8	0.246	-0.225	0.718	0.938	0.938	0.96	0.962
30	0.9	0.252	-0.278	0.782	0.958	0.958	0.94	0.965
30	1	0.263	-0.328	0.855	0.953	0.953	0.867	0.922
40	0.1	0.25	0.199	0.301	0.949	0.949	0.971	0.971
40	0.2	0.25	0.148	0.352	0.939	0.939	0.964	0.964
40	0.3	0.25	0.097	0.403	0.939	0.939	0.966	0.966
40	0.4	0.255	0.052	0.458	0.949	0.949	0.972	0.972
40	0.5	0.251	-0.003	0.505	0.946	0.946	0.968	0.968
40	0.6	0.245	-0.058	0.548	0.937	0.937	0.972	0.972
40	0.7	0.251	-0.103	0.605	0.944	0.944	0.962	0.962
40	0.8	0.247	-0.156	0.65	0.945	0.945	0.969	0.969
40	0.9	0.247	-0.212	0.706	0.945	0.945	0.973	0.973
40	1	0.254	-0.252	0.76	0.946	0.946	0.959	0.968

A.2.3 10% change between reference and test methods for varying values of theta ($M_T - M_R = 0.5$)

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	0.504	0.34	0.668	0.917	0.917	0.943	0.943
5	0.2	0.5	0.174	0.826	0.927	0.927	0.95	0.95
5	0.3	0.49	-0.001	0.982	0.913	0.913	0.883	0.93
5	0.4	0.499	-0.168	1.167	0.911	0.905	0.658	0.819
5	0.5	0.503	-0.328	1.334	0.915	0.817	0.41	0.65
5	0.6	0.511	-0.477	1.499	0.92	0.628	0.249	0.474
5	0.7	0.509	-0.649	1.667	0.923	0.394	0.138	0.335
5	0.8	0.487	-0.821	1.795	0.928	0.244	0.073	0.257
5	0.9	0.511	-0.979	2.001	0.929	0.132	0.063	0.207
5	1	0.555	-1.068	2.178	0.914	0.079	0.043	0.154
10	0.1	0.501	0.395	0.607	0.945	0.945	0.974	0.974
10	0.2	0.494	0.283	0.705	0.935	0.935	0.957	0.957
10	0.3	0.507	0.189	0.826	0.931	0.931	0.959	0.959
10	0.4	0.504	0.077	0.932	0.92	0.92	0.953	0.957
10	0.5	0.498	-0.041	1.036	0.936	0.936	0.909	0.944
10	0.6	0.518	-0.116	1.152	0.94	0.939	0.749	0.846
10	0.7	0.505	-0.241	1.251	0.946	0.916	0.538	0.702
10	0.8	0.494	-0.351	1.339	0.93	0.789	0.342	0.552
10	0.9	0.485	-0.473	1.443	0.93	0.563	0.199	0.41
10	1	0.501	-0.558	1.56	0.928	0.338	0.12	0.324
20	0.1	0.5	0.428	0.573	0.943	0.943	0.969	0.969
20	0.2	0.502	0.356	0.647	0.943	0.943	0.966	0.966
20	0.3	0.498	0.281	0.716	0.955	0.955	0.972	0.972
20	0.4	0.494	0.204	0.785	0.93	0.93	0.965	0.965
20	0.5	0.497	0.132	0.862	0.947	0.947	0.969	0.971
20	0.6	0.507	0.073	0.941	0.943	0.943	0.964	0.965
20	0.7	0.503	-0.008	1.014	0.96	0.96	0.954	0.971
20	0.8	0.506	-0.077	1.088	0.933	0.933	0.863	0.922
20	0.9	0.499	-0.159	1.156	0.94	0.94	0.746	0.857
20	1	0.499	-0.231	1.228	0.96	0.959	0.598	0.753
30	0.1	0.5	0.441	0.559	0.957	0.957	0.978	0.978
30	0.2	0.498	0.38	0.616	0.939	0.939	0.965	0.965
30	0.3	0.502	0.325	0.679	0.961	0.961	0.976	0.976
30	0.4	0.498	0.26	0.736	0.947	0.947	0.968	0.968
30	0.5	0.5	0.205	0.794	0.949	0.949	0.971	0.971
30	0.6	0.502	0.148	0.855	0.958	0.958	0.976	0.976
30	0.7	0.496	0.084	0.909	0.948	0.948	0.974	0.975
30	0.8	0.503	0.029	0.977	0.94	0.94	0.958	0.965
30	0.9	0.507	-0.021	1.035	0.935	0.935	0.932	0.95
30	1	0.506	-0.084	1.095	0.949	0.949	0.885	0.94
40	0.1	0.499	0.449	0.55	0.948	0.948	0.968	0.968
40	0.2	0.498	0.396	0.6	0.952	0.952	0.98	0.98
40	0.3	0.501	0.348	0.653	0.958	0.958	0.974	0.974
40	0.4	0.501	0.297	0.704	0.936	0.936	0.962	0.962
40	0.5	0.498	0.243	0.752	0.946	0.946	0.971	0.971
40	0.6	0.499	0.195	0.804	0.946	0.946	0.966	0.966
40	0.7	0.495	0.139	0.851	0.948	0.948	0.973	0.973
40	0.8	0.506	0.098	0.914	0.94	0.94	0.966	0.966
40	0.9	0.491	0.031	0.951	0.943	0.943	0.966	0.969
40	1	0.503	-0.002	1.007	0.939	0.939	0.946	0.963

A.2.4 15% change between reference and test methods for varying values of theta ($M_T - M_R = 0.75$)

Sample Statistics					Equivalence Testing Result			
n	std	MeanDiff	LCI	UCI	TTEST	Power	TOST	Hauck
5	0.1	0.745	0.582	0.908	0.923	0.923	0.954	0.954
5	0.2	0.75	0.427	1.074	0.93	0.93	0.961	0.961
5	0.3	0.738	0.239	1.237	0.933	0.933	0.905	0.946
5	0.4	0.754	0.079	1.429	0.919	0.919	0.653	0.816
5	0.5	0.749	-0.061	1.559	0.924	0.904	0.447	0.658
5	0.6	0.749	-0.218	1.715	0.9	0.799	0.258	0.448
5	0.7	0.76	-0.412	1.931	0.924	0.604	0.145	0.329
5	0.8	0.797	-0.517	2.11	0.916	0.42	0.085	0.251
5	0.9	0.759	-0.726	2.243	0.926	0.301	0.048	0.19
5	1	0.733	-0.875	2.341	0.918	0.191	0.036	0.145
10	0.1	0.751	0.645	0.858	0.942	0.942	0.968	0.968
10	0.2	0.753	0.542	0.963	0.945	0.945	0.964	0.964
10	0.3	0.754	0.434	1.073	0.948	0.948	0.966	0.966
10	0.4	0.749	0.33	1.169	0.929	0.929	0.951	0.955
10	0.5	0.748	0.209	1.288	0.939	0.939	0.89	0.938
10	0.6	0.765	0.123	1.408	0.93	0.93	0.737	0.85
10	0.7	0.754	0.001	1.506	0.927	0.926	0.508	0.704
10	0.8	0.736	-0.11	1.582	0.935	0.921	0.351	0.577
10	0.9	0.762	-0.197	1.721	0.946	0.855	0.215	0.433
10	1	0.758	-0.304	1.819	0.929	0.668	0.121	0.326
20	0.1	0.748	0.676	0.821	0.943	0.943	0.964	0.964
20	0.2	0.75	0.603	0.897	0.954	0.954	0.98	0.98
20	0.3	0.747	0.527	0.968	0.929	0.929	0.955	0.955
20	0.4	0.754	0.462	1.046	0.942	0.942	0.974	0.974
20	0.5	0.749	0.386	1.113	0.952	0.952	0.974	0.974
20	0.6	0.736	0.304	1.168	0.935	0.935	0.961	0.962
20	0.7	0.756	0.249	1.262	0.938	0.938	0.939	0.953
20	0.8	0.755	0.172	1.338	0.948	0.948	0.858	0.922
20	0.9	0.759	0.101	1.416	0.942	0.942	0.734	0.838
20	1	0.742	0.024	1.46	0.937	0.937	0.583	0.728
30	0.1	0.751	0.691	0.81	0.946	0.946	0.974	0.974
30	0.2	0.75	0.632	0.868	0.936	0.936	0.965	0.965
30	0.3	0.753	0.576	0.931	0.953	0.953	0.968	0.968
30	0.4	0.75	0.515	0.985	0.938	0.938	0.972	0.972
30	0.5	0.754	0.458	1.05	0.945	0.945	0.975	0.975
30	0.6	0.758	0.404	1.112	0.938	0.938	0.969	0.969
30	0.7	0.754	0.339	1.17	0.944	0.944	0.968	0.968
30	0.8	0.756	0.285	1.227	0.936	0.936	0.953	0.96
30	0.9	0.746	0.218	1.274	0.951	0.951	0.937	0.96
30	1	0.74	0.148	1.332	0.944	0.944	0.87	0.922
40	0.1	0.748	0.697	0.799	0.944	0.944	0.965	0.965
40	0.2	0.747	0.645	0.849	0.947	0.947	0.978	0.978
40	0.3	0.748	0.595	0.9	0.954	0.954	0.977	0.977
40	0.4	0.753	0.551	0.956	0.939	0.939	0.971	0.971
40	0.5	0.756	0.503	1.008	0.945	0.945	0.973	0.973
40	0.6	0.751	0.447	1.056	0.938	0.938	0.966	0.966
40	0.7	0.753	0.397	1.109	0.945	0.945	0.971	0.971
40	0.8	0.749	0.343	1.155	0.945	0.945	0.97	0.97
40	0.9	0.731	0.274	1.188	0.96	0.96	0.973	0.974
40	1	0.748	0.243	1.253	0.952	0.952	0.97	0.975

A.3 Visual comparison on each equivalence tests for log-normally distributed data

A.3.1 T-test

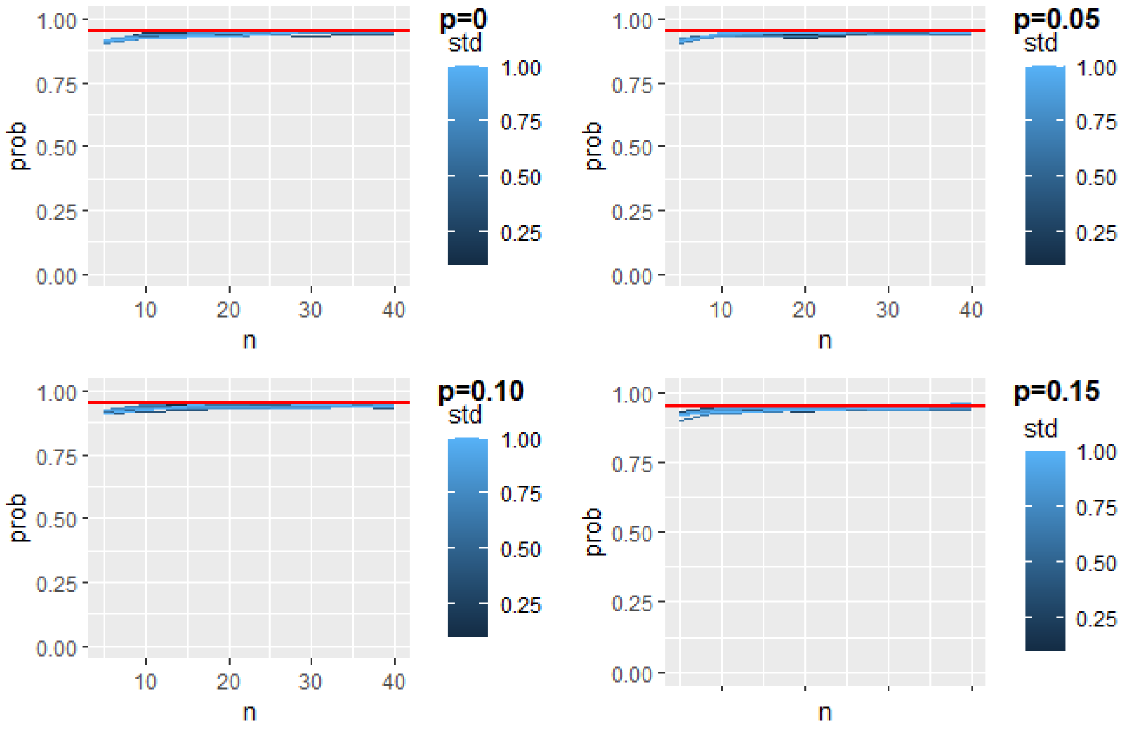


Figure A.1: Probability of two-sample t-test does not reject the null hypothesis that $\delta = \mu_D$ as the parameters sample size ($n_T = n_R = 5,\ 10,\ 20,\ 30,\ 40$) and standard deviation ($\sigma_T = \sigma_R = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9,\ 1$) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

A.3.2 Power analysis

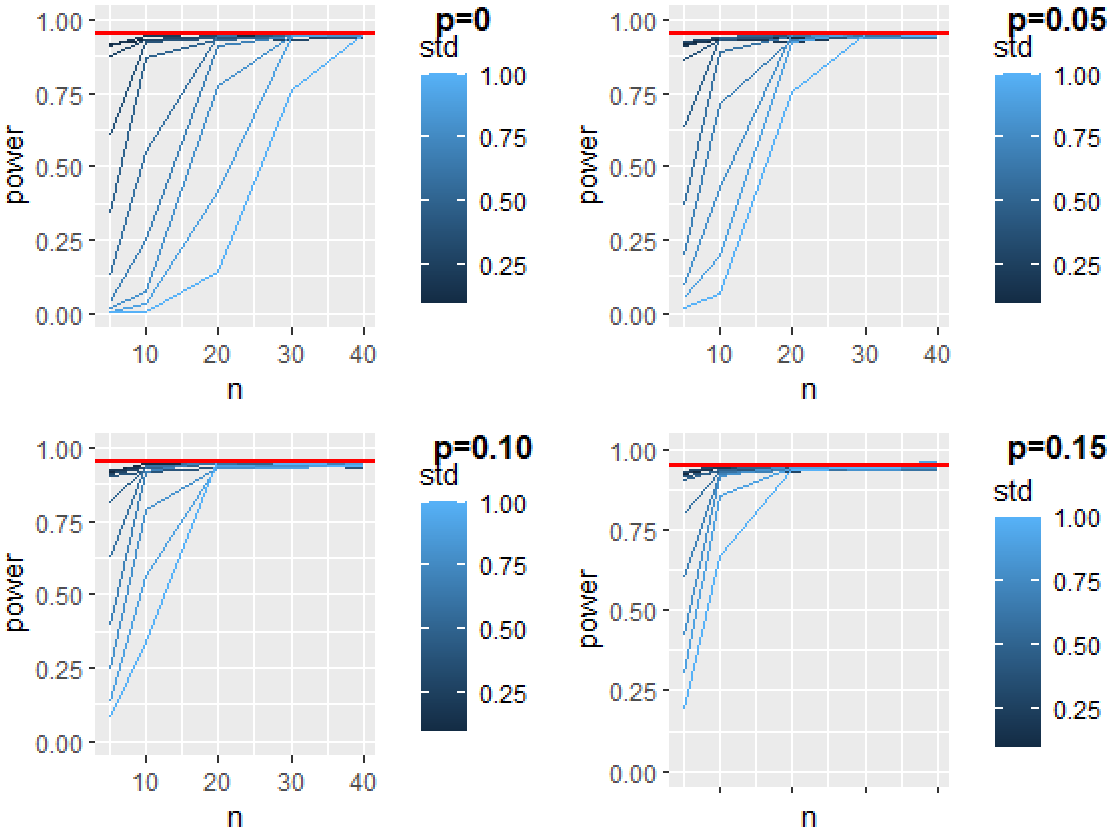


Figure A.2: Power of power analysis as the parameters sample size ($n_T = n_R = 5$, 10, 20, 30, 40) and standard deviation ($\sigma_T = \sigma_R =$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

A.3.3 TOST

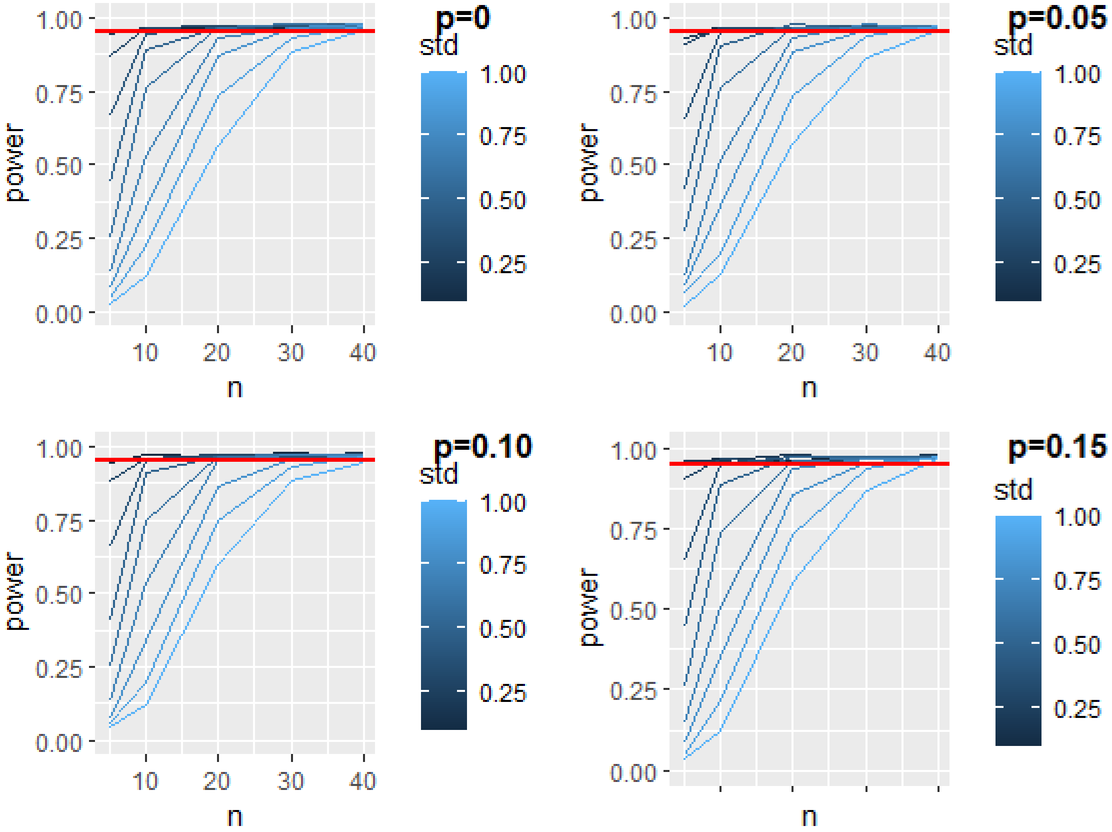


Figure A.3: Power of TOST as the parameters sample size ($n_T = n_R = 5$, 10, 20, 30, 40) and standard deviation ($\sigma_T = \sigma_R = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

A.3.4 Hauck-Anderson test

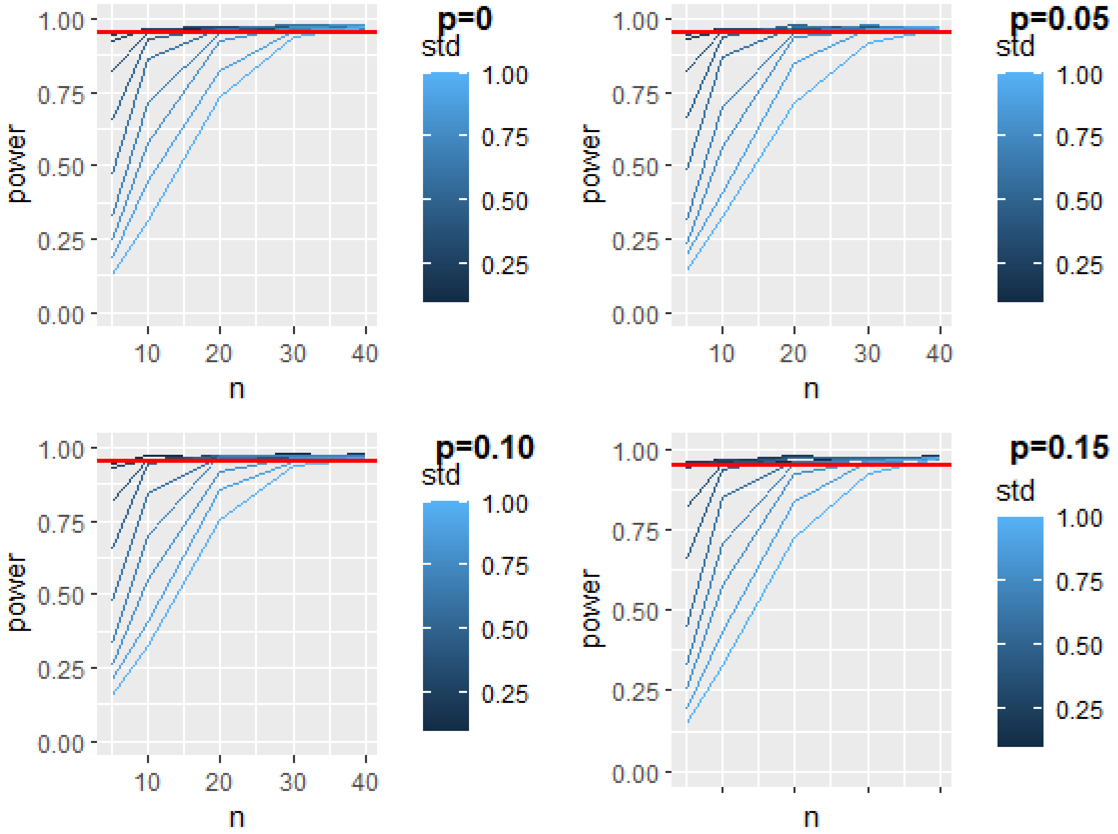


Figure A.4: Power of Hauck-Anderson method as the parameters sample size ($n_T = n_R =$ 5, 10, 20, 30, 40) and standard deviation ($\sigma_T = \sigma_R =$ 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) vary, under different percentage changes of reference group (i.$p = 0, \mu_D = 0$; ii.$p = 0.05, \mu_D = 1$; iii. $p = 0.10, \mu_T - \mu_R = 2$; iv. $p = 0.15, \mu_T - \mu_R = 3$).

Appendix B

Distributions of chemicals

B.1 Distributions of T-111B / SOP03: NAB, NAT, NNK and NNN from various cigarettes under intense smoking conditions

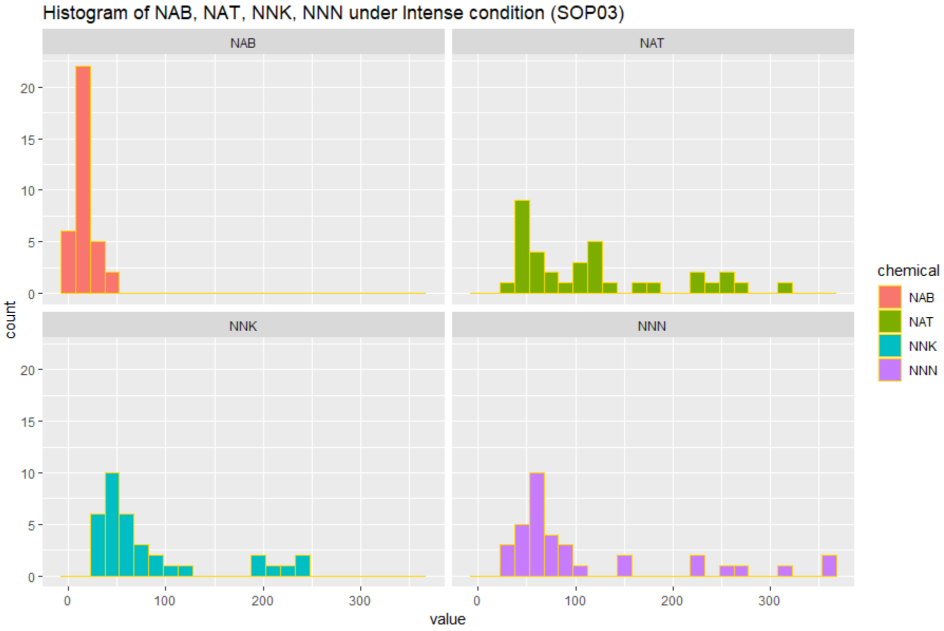


Figure B.1: (a) Distributions of normal-scale chemicals

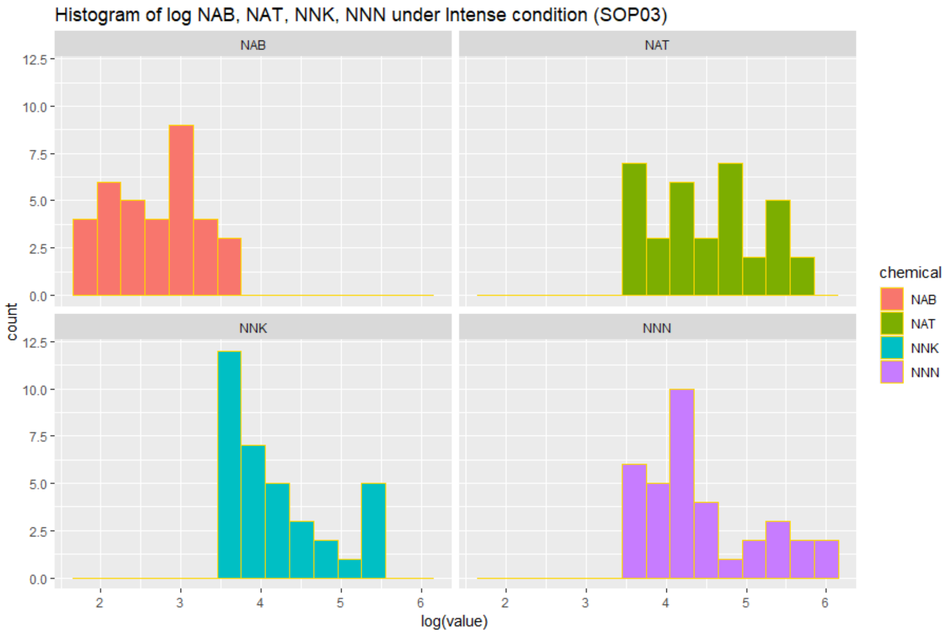


Figure B.2: (b) Distributions of log-scale chemicals

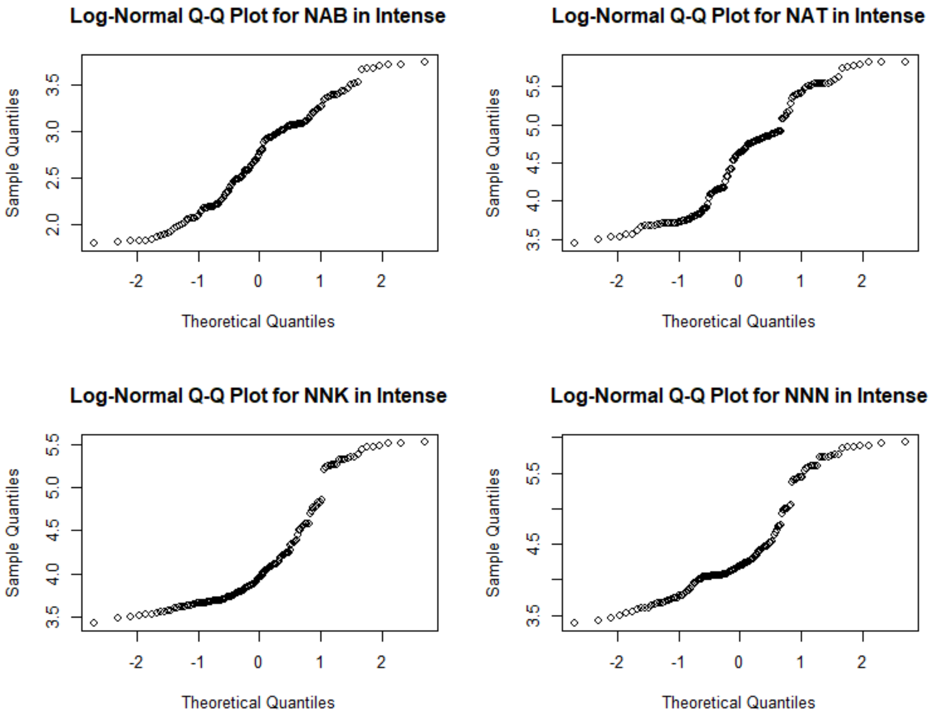


Figure B.3: (c) QQ-plot of log-scale chemicals

B.2 Distributions of T-111B / SOP03: NAB, NAT, NNK and NNN from various cigarettes under ISO smoking conditions

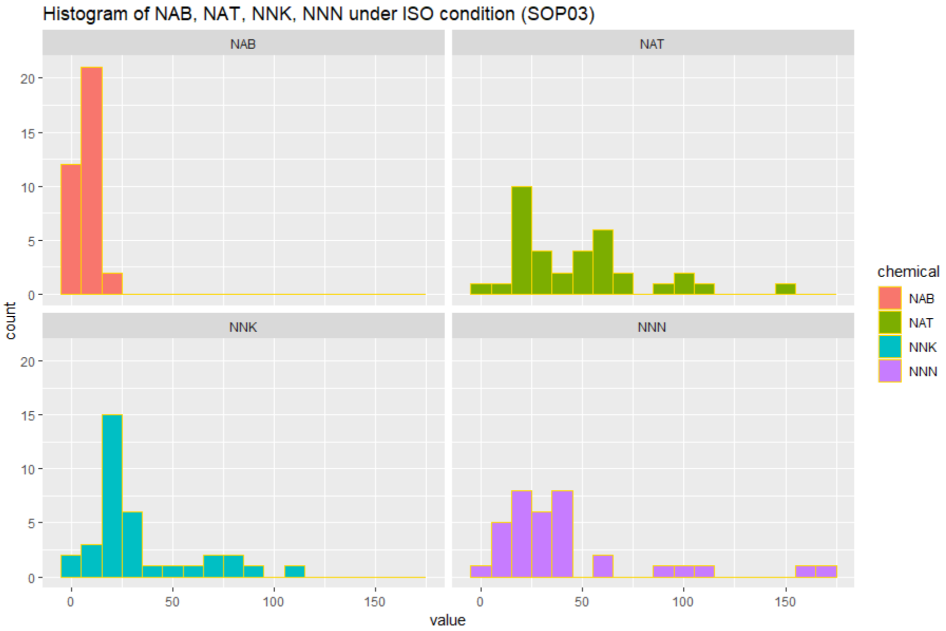


Figure B.4: (a) Distributions of normal-scale chemicals

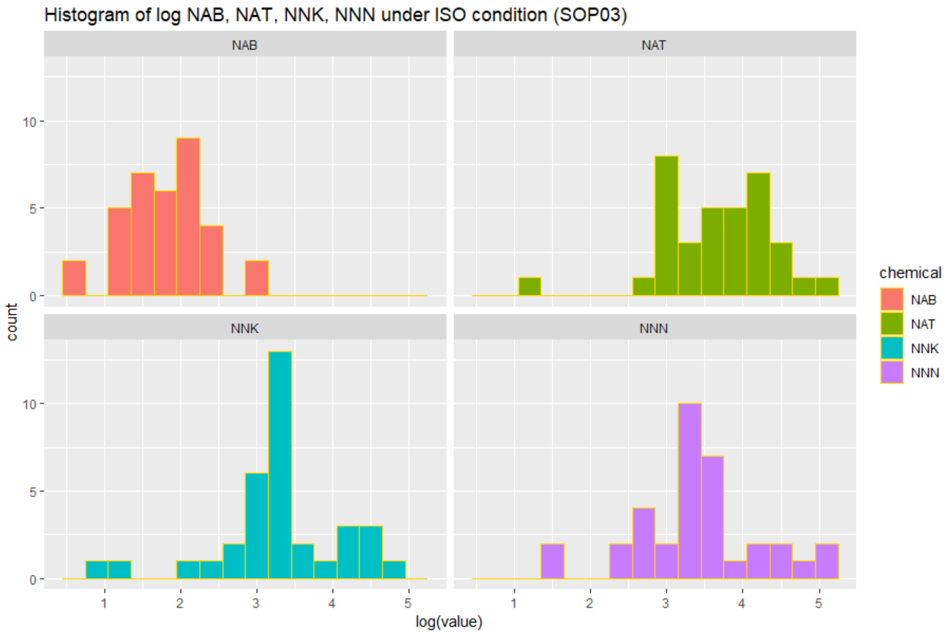


Figure B.5: (b) Distributions of log-scale chemicals

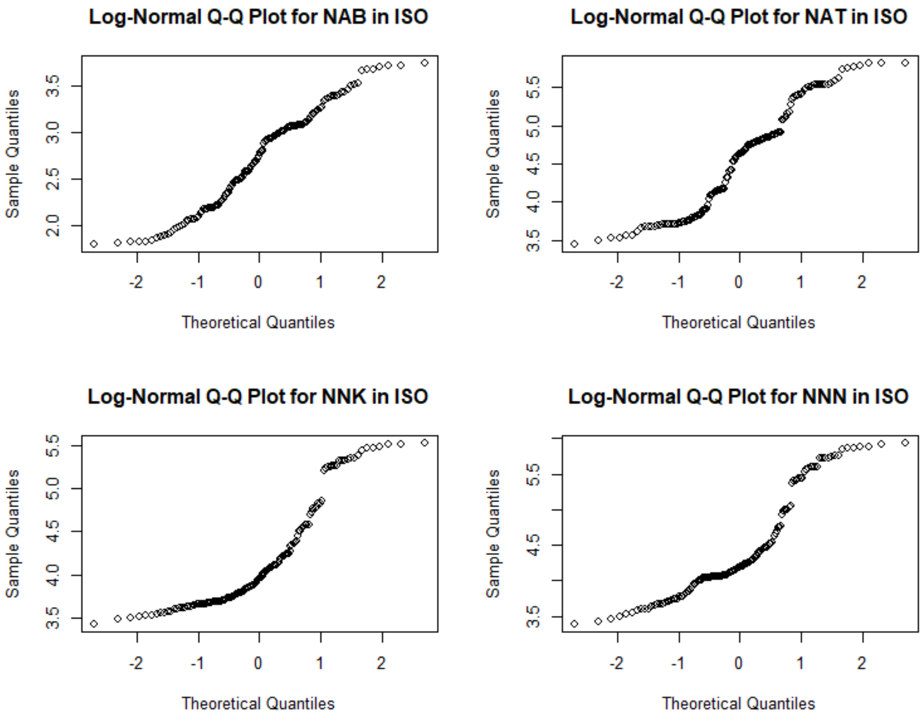


Figure B.6: (c) QQ-plot of log-scale chemicals

B.3 Distributions of T-304 / SOP06: glcerol, porpylene glycol, trienthylene glycol from various cigarettes (constituents)

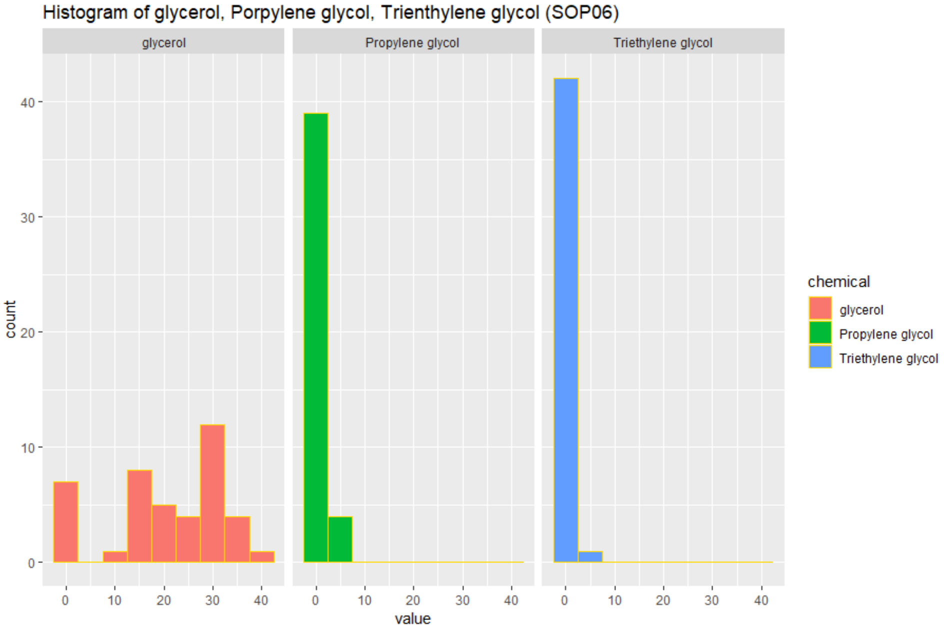


Figure B.7: (a) Distributions of normal-scale chemicals

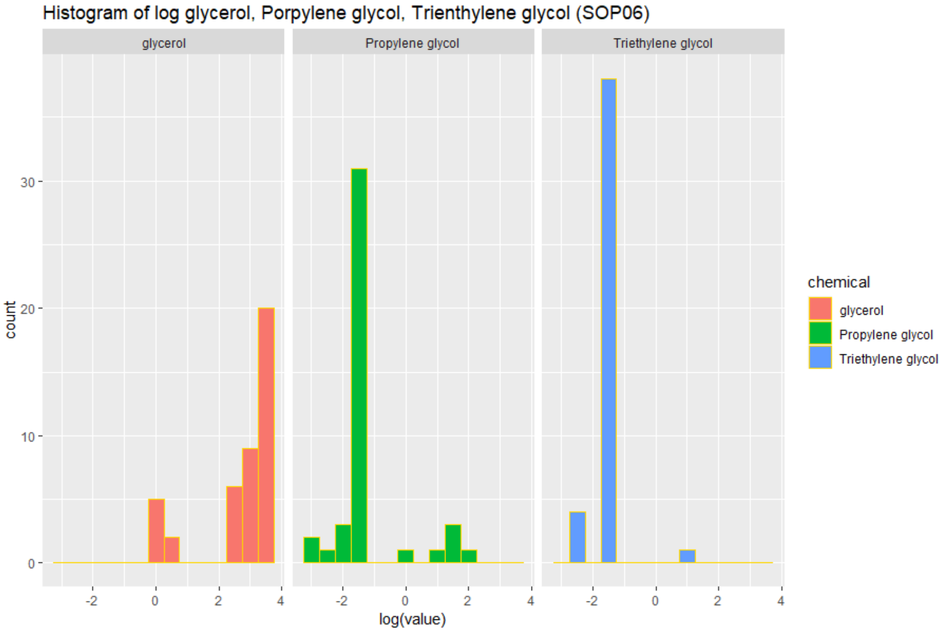


Figure B.8: (b) Distributions of log-scale chemicals

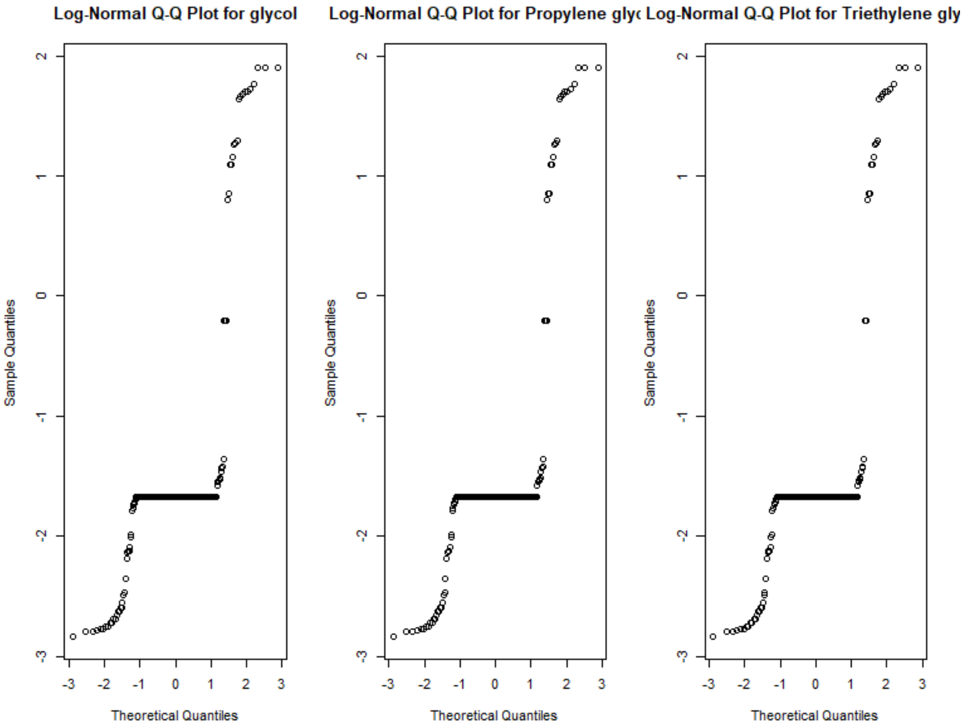


Figure B.9: (c) QQ-plot of log-scale chemicals

B.4 Distributions of T-115 / SOP10: CO and Nicotine from various cigarettes under intense smoking conditions

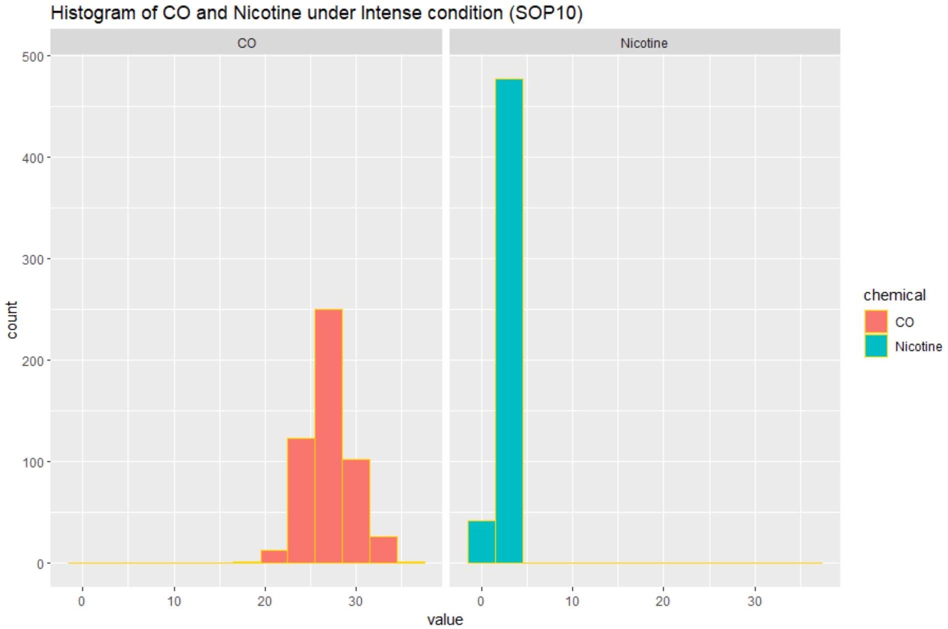


Figure B.10: (a) Distributions of normal-scale chemicals

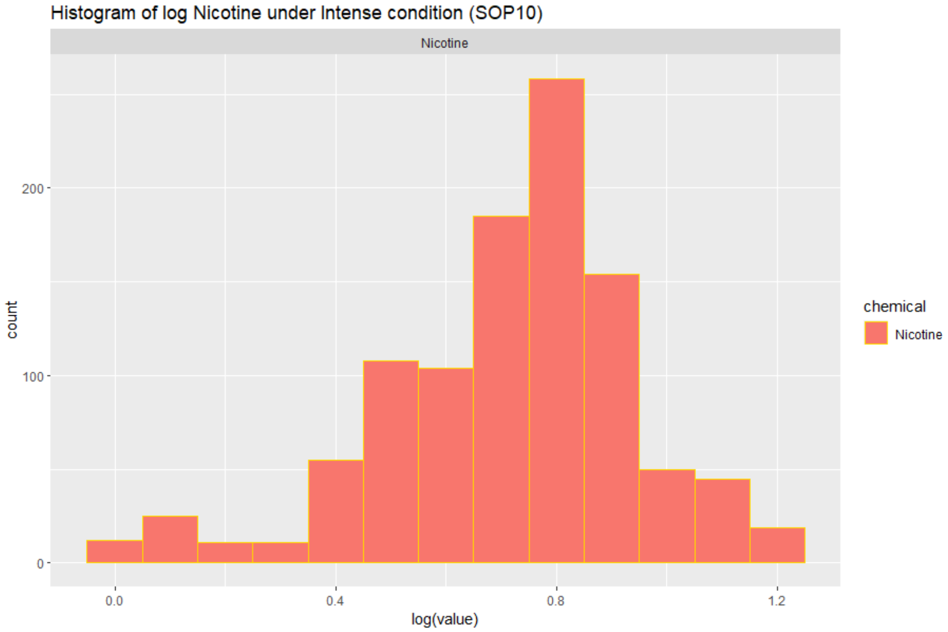


Figure B.11: (b) Distributions of log-scale chemicals

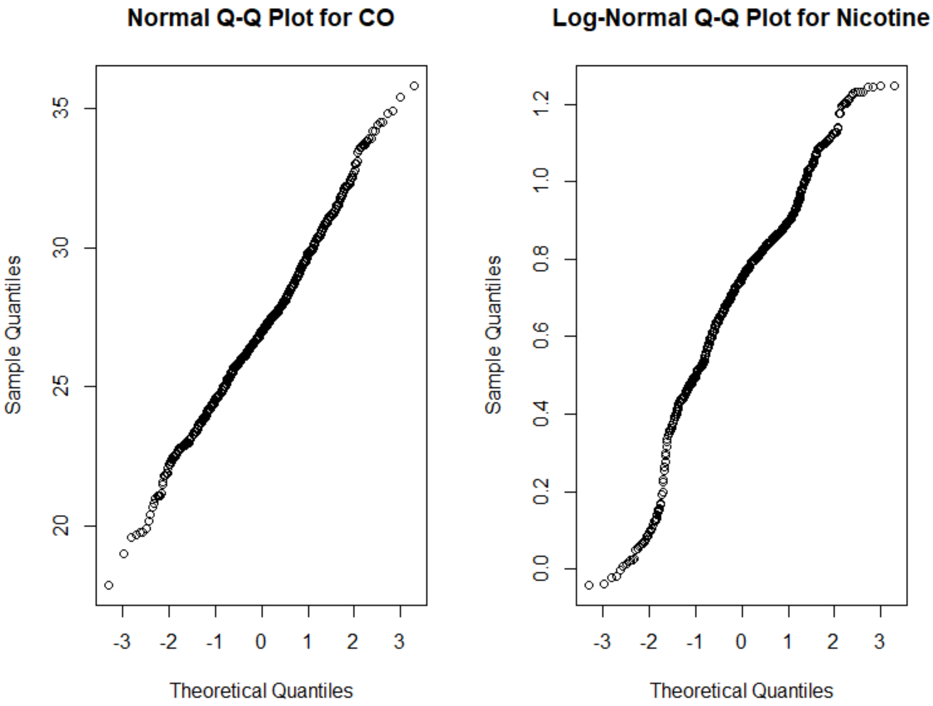


Figure B.12: (c) QQ-plot of log-scale chemicals

Appendix C

R codes

C.1 R codes for the Table 3.4 - 3.6 about equivalence limit determination with various parameters

```
library(openxlsx)
delta = 0
n = c(5,7,9,11,13,15,17,20,25,30,35,40)
wb <- createWorkbook()

## with various std values
alpha = 0.05
beta = 0.05
gamma=0.20
std_list = c(seq(0.5,0.9,0.1), seq(1,5,0.5))

theta_vec <- c()
theta <- data.frame(matrix(ncol = length(n), nrow = length(std_list)))
colnames(theta) <- paste0('n=',n,sep='')
rownames(theta) <- paste0('std=',std_list,sep='')

for (i in 1:length(n)){
  m = n[i]
  for (j in 1:length(std_list)){
    std = std_list[j]
    sst=std*sqrt((m-1)/qchisq(p=gamma, df=m-1, lower.tail=FALSE))
    theta_value = round(delta+sst*(qt(p=1-alpha, df=2*m-2) +
        qt(p=1-beta/2, df=2*m-2))*sqrt(2/m),2)
    theta_vec = append(theta_vec, theta_value)
  }
  theta[,i] = theta_vec
  theta_vec = c()}

addWorksheet(wb, sheetName = 'std')
writeData(wb, sheet = 'std', theta, rowNames=TRUE)
```

```
## with various alpha values
std = 3
n = c(5,7,9,11,13,15,17,20,25,30,35,40)
alpha = c(0.01, 0.05, 0.10, 0.15, 0.20)

theta_vec <- c()
theta <- data.frame(matrix(ncol = length(n), nrow = length(alpha)))
colnames(theta) <- paste0(’n=’,n,sep=’’)
rownames(theta) <- paste0(’alpha=’,alpha,sep=’’)

for (i in 1:length(n)){
  m = n[i]
  for (j in 1:length(alpha)){
    a = alpha[j]
    sst=std*sqrt((m-1)/qchisq(p=gamma, df=m-1))
    theta_value = round(delta+sst*(qt(p=1-a, df=2*m-2) +
        qt(p=1-beta/2, df=2*m-2))*sqrt(2/m),2)
    theta_vec = append(theta_vec, theta_value)
  }
  theta[,i] = theta_vec
  theta_vec = c()}

addWorksheet(wb, sheetName = ’alpha’)
writeData(wb, sheet = ’alpha’, theta, rowNames=TRUE)

## with various beta values
std = 3
alpha = 0.05
beta = c(0.01, 0.05, 0.10, 0.15, 0.20)

theta_vec <- c()
theta <- data.frame(matrix(ncol = length(n), nrow = length(beta)))
colnames(theta) <- paste0(’n=’,n,sep=’’)
rownames(theta) <- paste0(’beta=’,beta,sep=’’)

for (i in 1:length(n)){
  m = n[i]
  for (j in 1:length(beta)){
    b = beta[j]
    sst=std*sqrt((m-1)/qchisq(p=gamma, df=m-1))
    theta_value = round(delta+sst*(qt(p=1-alpha, df=2*m-2) +
        qt(p=1-b/2, df=2*m-2))*sqrt(2/m),2)
    theta_vec = append(theta_vec, theta_value)
  }
  theta[,i] = theta_vec
  theta_vec = c()}

addWorksheet(wb, sheetName = ’beta’)
writeData(wb, sheet = ’beta’, theta, rowNames=TRUE)

saveWorkbook(wb, "␣")
```

C.2 R codes for the Table 3.7 for sample size determination with various parameters

```
## read libraries into R
library(readxl)
library(writexl)

set.seed(1231) ##to get the exactly same sampling data on each round

##### initial setting
alpha=0.05
beta=0.05
gamma=0.20
size = 1000
cv_vec = seq(0.05, 0.4, 0.05)
delta_vec = seq(0, 0.1, 0.02)
theta_vec = seq(0.05, 0.3, 0.05)

n_vec <- c()
m_vec <- c()
final<- data.frame(CV=rep(cv_vec, length(delta_vec)))

for (t in 1:length(theta_vec)){
  theta = theta_vec[t]
  for (k in 1:length(delta_vec)){
    p = delta_vec[k]
    for (j in 1:length(cv_vec)){
      cv = cv_vec[j]
      cv1 = as.character(cv)

      n = ceiling(2*(cv*(qnorm(1-alpha)+qnorm(1-beta/2))/(theta-p))^2)+1
      n_vec = append(n_vec, n)
    }
    m_vec = append(m_vec,n_vec)
    n_vec = c()}
    final = cbind(final,m_vec)
    colnames(final)[t+1]=paste(’theta_pc_’,theta_vec[t])
    m_vec = c()}
final <- data.frame(delta_pc=rep(delta_vec, each=length(cv_vec)),final)

write_xlsx(final,"␣")
```

C.3 R codes for the Figure 3.6 - 3.9 about power comparison of equivalence tests for normally distributed data

```
## read libraries into R
library(writexl)
library(openxlsx)
library(dplyr)
library(tidyr)
library(ggplot2)
library(ggpubr)

set.seed(1231) ##to get the exactly same sampling data on each round

##### initial setting
alpha = 0.025
beta = 0.025
alpha1 = 0.05
beta1 = 0.05
size = 1000
n_vec = c(5, 10, 20, 30, 40)
std_vec = c(seq(0.1, 0.9, 0.2), seq(1, 5, 1))
pc_vec = c(0, 0.05, 0.1, 0.15)

refmean = 20 # reference mean
thetafix = 0.2*refmean # FDA theta suggestion

df <- NULL
sum_list <- list()
wb <- createWorkbook()
chem_summary <- data.frame(n=numeric(),
                           std=numeric(),
                           MeanDiff=numeric(),
                           LCI=numeric(),
                           UCI=numeric())

for (k in 1:length(pc_vec)){
  p = pc_vec[k]
  delta = refmean*p

  for (j in 1:length(n_vec)){
    m = n_vec[j]
    m1 = as.character(1:m)

    for (t in 1:length(std_vec)){
      std = std_vec[t]
      std1 = as.character(std)

      refchem = as.data.frame(matrix(rnorm(n=m*size, mean=refmean,
        sd=std), nrow=size, ncol=m))
      testchem = as.data.frame(matrix(rnorm(n=m*size, mean=refmean*(1+p),
        sd=std), nrow=size, ncol=m))
```

```r
colnames(refchem) = c(paste('r', m1, sep=''))
colnames(testchem) = c(paste('t', m1, sep=''))

refchem$ref_freq = rep(m,size)
refchem$ref_mean = apply(refchem[,1:m], 1, mean)
refchem$ref_std = apply(refchem[,1:m], 1, sd)
refchem$ref_var = refchem$ref_std^2

testchem$test_freq = rep(m,size)
testchem$test_mean = apply(testchem[,1:m], 1, mean)
testchem$test_std = apply(testchem[,1:m], 1, sd)
testchem$test_var = testchem$test_std^2
chemn = cbind(refchem,testchem)

Testing <- as.data.frame(chemn) %>%
  mutate(mean_diff = test_mean-ref_mean,
         tvalue = (mean_diff-delta)/sqrt(ref_var*(2/m)),
         pvalue = 2*(1-pt(abs(tvalue),2*m-2)),
         TTEST_result = ifelse((pvalue<=alpha), "UnEqual", "Equal"),

         theta = delta+thetafix,
         negtheta = delta-thetafix,
         t = qt(p=1-alpha/2, df=2*m-2),
         LCI = mean_diff-t*sqrt(ref_var*(2/m)),
         UCI = mean_diff+t*sqrt(ref_var*(2/m)),
         TOST_result = ifelse(TTEST_result=="Equal",
          ifelse(negtheta<=LCI & UCI<=theta, "Equiv",
          "UnEquiv"), NA),

         sigma = (theta/(sqrt(2/m)))/(qnorm(p=1-alpha1/2)+
          qnorm(1-beta1))*(((theta>=0)*1)-((theta<0)*1)),
         TTEST_result1 = ifelse((pvalue<=alpha1),
          "UnEqual", "Equal"),
         Power_result = ifelse(TTEST_result1=="Equal",
          ifelse(sqrt((ref_var+test_var)/2)<sigma, "Equiv",
          "UnEquiv"), NA),

         Hauck_delta = thetafix/(sqrt(2*ref_var/m)),
         Hauck_Tvalue = (mean_diff-delta)/(sqrt(2*ref_var/m)),
         Hauck_pvalue = pt(abs(Hauck_Tvalue)-Hauck_delta, 2*m-2)-
          pt(-abs(Hauck_Tvalue)-Hauck_delta, 2*m-2),
         Hauck_result = ifelse(TTEST_result=="Equal",
          ifelse(Hauck_pvalue<=alpha, "Equiv", "UnEquiv"), NA))

chem_summary[t, 'n'] = m
chem_summary[t, 'std'] = std
chem_summary[t, 'MeanDiff'] = round(mean(Testing$mean_diff),3)
chem_summary[t, 'LCI'] = round(mean(Testing$LCI),3)
chem_summary[t, 'UCI'] = round(mean(Testing$UCI),3)
chem_summary[t, 'TTEST'] = round(length(which(Testing$TTEST_result1==
  'Equal'))/size, 3)
chem_summary[t, 'Power'] = round(length(which(Testing$Power_result==
  'Equiv'))/size, 3)
```

```
      chem_summary[t, 'TOST'] = round(length(which(Testing$TOST_result==
'Equiv'))/size, 3)
      chem_summary[t, 'Hauck'] = round(length(which(Testing$Hauck_result==
        'Equiv'))/size,3)}
    df[[j]] = chem_summary}
  finaltest = rbind(df[[1]], df[[2]], df[[3]], df[[4]], df[[5]])
  sum_list[[k]] <- finaltest
  addWorksheet(wb, sheetName = paste0('p=',p,'mean␣diff=',round(delta,2),
    sep=''))
  writeData(wb, sheet = paste0('p=',p,'mean␣diff=',round(delta,2), sep=''),
    finaltest)}

saveWorkbook(wb, "␣")

long_list <- list()
for (i in 1:4){
  long_list[[i]] <- gather(sum_list[[i]], test, power, TTEST:Hauck,
    factor_key=TRUE)}

list <- c('TTEST', 'Power', 'TOST', 'Hauck')
for (i in 1:length(list)){
  p00 <- long_list[[1]] %>%
    filter(test==list[i]) %>%
    ggplot(aes(x=n, y=power, group=std, color=std))+
    coord_cartesian(ylim = c(0, 1))+
    geom_line()+
    geom_hline(yintercept = 0.95, col='red', cex=1)
  p05 <- long_list[[2]] %>%
    filter(test==list[i]) %>%
    ggplot(aes(x=n, y=power, group=std, color=std))+
    coord_cartesian(ylim = c(0, 1))+
    geom_line()+
    geom_hline(yintercept = 0.95, col='red', cex=1)
  p10 <- long_list[[3]] %>%
    filter(test==list[i]) %>%
    ggplot(aes(x=n, y=power, group=std, color=std))+
    coord_cartesian(ylim = c(0, 1))+
    geom_line()+
    geom_hline(yintercept = 0.95, col='red', cex=1)
  p15 <- long_list[[4]] %>%
    filter(test==list[i]) %>%
    ggplot(aes(x=n, y=power, group=std, color=std)) +
    coord_cartesian(ylim = c(0, 1)) +
    geom_line()+
    geom_hline(yintercept = 0.95, col='red', cex=1)
  print(ggarrange(p00, p05, p10, p15 + rremove("x.text"),
                  labels = c("␣␣p=0", "p=0.05", "p=0.10", "p=0.15"),
                  label.x = 0.7, label.y = 1,
                  font.label = list(size = 12),
                  ncol = 2, nrow = 2))}
```