MEDIRES
Symbol of Quality Research

Clinical Trials and Bioavailability Research
info@mediresonline.org

ISSN: 2836-5836 RESEARCH ARTICLE

A Reference Scaled Difference in Means Approach to Equivalence Test for Binary Clinical Outcomes

Chao Wang[1*], Yixin Ren[1], Meiyu Shen[1] and Yi Tsong[1]

[1]*Office of Biostatistics, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, United States*

***Corresponding Author**: Chao Wang, Office of Biostatistics, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, United States.

Received Date: 16 February 2024; **Accepted Date:** 11 March 2024; **Published date:** 22 March 2024

Citation: Chao Wang, Yixin Ren, Meiyu Shen, and Yi Tsong. (2024). A Reference Scaled Difference in Means Approach to Equivalence Test for Binary Clinical Outcomes. Clinical Trials and Bioavailability Research. 3(1); DOI: 10.58489/2836-5836/012

Abstract

A proposed generic drug product needs to be compared with some reference product through equivalence test to support marketing approval. One important type of treatment outcome is the binary response indicating whether a favorable outcome is achieved. The binary response rates of the test and reference treatments are often compared via their difference with some margin. While the difference can usually be estimated in a clinical study, it is frequently difficult to determine a proper margin. Existing approaches suggest a fixed margin or a margin as a function of the reference response rate, such as step-wise constant margin, piece-wise smooth margin, etc. The issues with existing margin choices were discussed recently and a variable margin was proposed for non-inferiority test, which is a constant multiple of the standard deviation of reference response rate. In this paper, we extended the discussion to equivalence test, which can be formulated as two one-sided tests. Our discussion revealed some unique features of the equivalence test for binary responses with a variable margin. For instance, this approach may improve power control when the reference product has a high response rate. On the other hand, when both sample size and margin multiplier is small, the rejection rate of the equivalence test is nearly zero.

Keywords: *equivalence test, binary response, variable margin, reference-scaled test, Wald test.*

Introduction

In clinical trials, one important type of treatment outcome is binary response indicating whether a subject has a favorable outcome or unfavorable outcome, often conveniently denoted by 1 and 0 respectively. Let p_T and p_R be the binary response rates of the test and reference products respectively. For a new test drug proposed as a generic to a marketed reference product, one of the critical requirements for regulatory approval is to establish the equivalence between p_T and p_R, which is often assessed via some form of equivalence test.

A common form of equivalence test is focused on the difference between the two response rates and is formulated as below,

$$H_0: |p_T - p_R| \geq \delta \text{ versus } H_a: |p_T - p_R| < \delta \quad (1)$$

where $\delta > 0$ is referred to as the equivalence margin.

The equivalence test can also be reformulated as the following two one-sided tests (TOST),

$$H_{10}: p_T - p_R \leq -\delta \text{ versus } H_{1a}: p_T - p_R > \delta \quad (2)$$

and

$$H_{20}: p_T - p_R \geq \delta \text{ versus } H_{2a}: p_T - p_R < \delta. \quad (3)$$

The hypothesis test in Eq. (2) is often presented as a non-inferiority (NI) test. An FDA guidance for industry recommends that the NI test be performed with a test size of 2.5% when it is used to establish effectiveness (FDA, 2016). For the case of equivalence test, it is generally accepted that both null hypotheses H_{10} and H_{20} be rejected at the type I error rate of 5% to achieve an overall 5% test size.

Regulatory authorities may predetermine a fixed constant margin δ for a particular drug product or a category of drug products. An FDA guidance for industry (FDA, 2003) recommends that bioequivalence be established with clinical endpoints for drugs with low absorption in blood system. A clinical trial of generic drug bioequivalence assessment typically consists of three arms, i.e., placebo (B), test (T), and reference (R) treatments. The investigator needs to compare test with placebo

for efficacy, using the following hypotheses,

$H_0: p_T - p_B \leq 0$ versus $H_a: p_T - p_B > 0$,

and to compare reference with placebo to show validation in the study population with the following hypotheses,

$H_0: p_R - p_B \leq 0$ versus $H_a: p_R - p_B > 0$.

Both null hypotheses should be rejected with 2.5% type I error rate before testing for bioequivalence with the hypotheses in Eq (2) and (3) with $\delta = 0.2$.

This fixed margin can lead to some difficulties. The variance of the sample response rate is a function of the response rate. Therefore, when the response rate is close to 0 or 1, the variance is smaller, and the equivalence test enjoys a higher power than the case when the response rate is close to 0.5. In fact, the sample size required with a fixed margin can increase up to a magnitude of 2.7 when the true response rates vary from 0.1 or 0.9 to 0.5, assuming equal response rates and equal sample sizes for test and reference treatments (see Eq. 4.2.4 (Chow, Wang, and Shao, Chow et al.)). In addition, Yuan et al. (2018) discussed the power and sample size determination in bioequivalence test of binary endpoints in generic drug applications with a fixed margin of δ = 20%, using the same sample size, for the same size of difference, $p_T - p_R$, and showed that the power of bioequivalence test depends heavily on the reference response rate. It is of interest to drug developers to keep sample sizes small to reduce cost and also to regulatory authorities as not to increase unnecessary burden to drug developers. It may be impractical to calculate sample size needed by assuming the worst-case scenario of both response rates equal to 0.5.

In a recent discussion (Ren et al., 2019) for NI test, several common margin choices were compared, such as fixed margin (Tsong, 2007; FDA, 2016), variable margins which are functions of p_R, i.e., $\delta = \delta(p_R)$, including step-wise constant margin (FDA, 1992; Röhmel, 1998), and smooth margins (Röhmel, 2001). It was also noted that the reference response rate in the smooth margin was considered to be deterministic by Röhmel (2001). In addition, there was also a comprehensive discussion for NI test with a variable margin (Zhang, 2006).

Step-wise constant margins have also been used in recent studies. To design a proper immunogenicity study of an insulin biosimilar drug product, a constant margin determined by a step function with a maximum sample size of 500 patients was used in Wang (2018), which is reproduced in Table 1.

Table 1: A step function to determine margin, reproduced from Wang (2018).

ADA Rate of Reference Product (%)	Margin (%)
5	5.70
10	7.90
15	9.30
20	10.50
25	11.30
30	12.00
35	12.50
40	12.80
45	13.00
50	13.10
55	13.00
60	12.80

In this paper, we extend the smooth variable margin originally proposed for NI test (Ren et al., 2019) to the two-sided equivalence test. Although there are some similarities between the equivalence test and NI test, we found some features unseen for the case of the one-sided NI test. For instance, the rejection rate of the equivalence test is nearly zero when both sample size and margin multiplier is small, regardless the values of p_T and p_R.

The paper is organized as follows. A formal discussion of the problem and test statistics are presented in Section 2. Section 3 reports simulation studies. Section 4 concludes the paper with additional discussion. More technical details are deferred to the Appendix.

Materials And Methods

FDA recommends to use the reference scaled average bioequivalence test for products with large variability (FDA, 2001, 2011; Tothfalusi and Endrenyi, 2016). The test can be stated as below,

$H_0: (\mu_T - \mu_R)/\sigma_R \leq -\delta \; or \; (\mu_T - \mu_R)/\sigma_R \geq \delta$ versus $H_a: -\delta < (\mu_T - \mu_R)/\sigma_R < \delta$ *(4)*

where μ_T *and* μ_R are the means of test and reference respectively; σ_R is the standard deviation of reference product; δ is the pre-specified equivalence margin. This hypothesis can also be presented as follow,

$H_0: \mu_T - \mu_R \leq -\delta\sigma_R \; or \; \mu_T - \mu_R \geq \delta\sigma_R$ versus $H_a: -\sigma\delta_R < \mu_T - \mu_R < \sigma\sigma_R$. *(5)*

For binary data, the test can be presented as follow,

$H_0: p_T - p_R \le -\delta(p_R)$ or $p_T - p_R \ge \delta(p_R)$ versus $H_a: -\delta(p_R) < p_T - p_R < \delta(p_R)$ (6)

where p_T *and* p_R are the response rates of test and reference respectively; $\delta(p_R)$ is the equivalence margin represented by $k\sqrt{p_R(1-p_R)}$.

Test statistics

Let $(X_{I,i})_{i=1}^{n_I}$ be the observed responses for product $I = T, R$. Throughout, it is assumed that $X_{I,i} \sim_{iid} Bernoulli(p_I)$ and the two samples $(X_{T,i})_{i=1}^{n_T}$ and $(X_{R,i})_{i=1}^{n_R}$ are mutually independent. Let $\hat{p_I} = \sum_i X_{I,i}/n_I$ be the sample estimate of the response rate for p_I $I = T, R$.

In the following discussion, we use the margin function originally proposed by (Ren et al., 2019), which is a multiple of the standard deviation of sample estimate for p_R,

$\delta(p_R) = k\sqrt{p_R(1-p_R)}$ *(7)*

where $k > 0$ is referred to as the margin multiplier, the selection of which will be discussed later.

We consider the following Wald type test statistics for equivalence test,

$$T_1 = \frac{\hat{p}_T - \hat{p}_R + \delta(\hat{p}_R)}{\sqrt{\mathrm{var}(\hat{p}_T - \hat{p}_R + \delta(\hat{p}_R))}},$$

$$T_2 = \frac{\hat{p}_T - \hat{p}_R - \delta(\hat{p}_R)}{\sqrt{\mathrm{var}(\hat{p}_T - \hat{p}_R - \delta(\hat{p}_R))}}.$$

Let

$f_1(x) = x - k\sqrt{x(1-x)}$, $f_2(x) = x + k\sqrt{x(1-x)}$,

then T_1 and T_2 can be written as follows,

$$T_1 = \frac{\hat{p}_T - f_1(\hat{p}_R)}{\sqrt{\mathrm{var}(\hat{p}_T - f_1(\hat{p}_R))}}, \quad (8)$$

$$T_2 = \frac{\hat{p}_T - f_2(\hat{p}_R)}{\sqrt{\mathrm{var}(\hat{p}_T - f_2(\hat{p}_R))}}. \quad (9)$$

Three methods for calculating the variances, $\mathrm{var}\left(\hat{p}_T - f_j(\hat{p}_R)\right), j = 1,2$, were discussed and compared based on asymptotic normal approximation previously (Ren et al., 2019). For the purpose of benchmarking the performance, the same methods are used here.

The first method ignores the margin variability, which gives a naive version of the variance and is frequently used in practice,

$$\nu_1(p_T, p_R) = \frac{p_T(1-p_T)}{n_T} + \frac{p_R(1-p_R)}{n_R}. \quad (10)$$

The second and third methods take into account the margin variability but differ in the response rates at which they are evaluated. Since $\sqrt{n_R}(\hat{p}_R - p_R) \to_d N(0, p_R(1-p_R))$, by the delta method, it follows that

$$\sqrt{n_R}(f_j(\hat{p}_R) - f_j(p_R)) \to_d N\left(0, \left(\sqrt{p_R(1-p_R)} + (-1)^j k\left(\frac{1}{2} - p_R\right)\right)^2\right).$$

Therefore, a more accurate version of the variance for $\hat{p}_T - f_j(\hat{p}_R)$ is

$$\nu_{j,2}(p_T, p_R) = \frac{p_T(1-p_T)}{n_T} + \frac{\left(\sqrt{p_R(1-p_R)} + (-1)^j k\left(\frac{1}{2} - p_R\right)\right)^2}{n_R}. \quad (11)$$

For a null hypothesis H_{i0}, i = 1, 2, let $\check{p}_{i,T}$ and $\check{p}_{i,R}$ be the restricted maximum likelihood estimates of p_T and p_R respectively, restricted at the boundary of H_{i0}, i.e., $p_T - p_R = (-1)^i\delta(p_R)$. Then the three test statistics are defined below.

1. Plugging $\hat{p}_T$ and $\hat{p}_R$ into v_1, we have

$$T_{i,MWO} = \frac{\hat{p}_T - f_i(\hat{p}_R)}{\sqrt{\nu_1(\hat{p}_T, \hat{p}_R)}}. \quad (12)$$

2. Plugging $\check{p}_{i,T}$ and $\check{p}_{i,R}$ to v_1, we have

$$T_{i,RWO} = \frac{\hat{p}_T - f_i(\hat{p}_R)}{\sqrt{\nu_1(\check{p}_{i,T}, \check{p}_{i,R})}}. \quad (13)$$

3. Plugging $\check{p}_{i,T}$ and $\check{p}_{i,R}$ to $v_{i,2}$, we have

$$T_{i,RW} = \frac{\hat{p}_T - f_i(\hat{p}_R)}{\sqrt{\nu_{i,2}(\check{p}_{i,T}, \check{p}_{i,R})}}. \quad (14)$$

The distributions of the three test statistics $T_{i,I}$ for $I = MWO, RWO, RW$ at the finite boundary of H_{i0} can be approximated by the standard normal distribution. Let $Z_{1-\alpha}$ denote the $(1-\alpha)$-quantile of the standard normal distribution, then the null hypothesis H_{10} is rejected if $T_{1,I} > Z_{1-\alpha}$ and H_{20} is rejected if $T_{2,I} < -Z_{1-\alpha}$. The null hypothesis of equivalence test H_0 is rejected if and only if both H_{10} and H_{20} are rejected.

Power function

Both $T_{1,MWO}$ and $T_{1,RWO}$ were shown to control type I error considerably inferior to $T_{1,RW}$ for NI test (Ren et al., 2019). This is also true for T_{MWO} and T_{RMO} compared with T_{RW} for equivalence test, which can be seen in the simulation studies reported later in this paper. Thus, only the power function of T_{RW} is considered here.

Assume that there exist some $\bar{p}_{i,T}$ and $\bar{p}_{i,R}$ such that $\check{p}_{i,T} \to \bar{p}_{i,T}$ and $\check{p}_{i,R} \to \bar{p}_{i,R}$ in probability. Further

assume that $p_T - p_R = k_0\sqrt{p_R(1-p_R)}$. Then the approximate power functions can be given as below (more details can be found in the Appendix),

$$P_{RW} \approx P\left(Z_1 > \sqrt{\frac{\nu_{1,2}(\bar{p}_{1,T},\bar{p}_{1,R})}{\nu_{1,2}(p_T,p_R)}}\left(Z_{1-\alpha} - \frac{p_T - f_1(p_R)}{\sqrt{\nu_{1,2}(\bar{p}_{1,T},\bar{p}_{1,R})}}\right) \&\right.$$
$$\left. Z_2 < \sqrt{\frac{\nu_{2,2}(\bar{p}_{2,T},\bar{p}_{2,R})}{\nu_{2,2}(p_T,p_R)}}\left(-Z_{1-\alpha} - \frac{p_T - f_2(p_R)}{\sqrt{\nu_{2,2}(\bar{p}_{2,T},\bar{p}_{2,R})}}\right)\right), \quad (15)$$

where

$$Z_j = \frac{\hat{p}_T - f_j(\hat{p}_R) - (p_T - f_j(p_R))}{\sqrt{\nu_{j,2}(p_T,p_R)}},$$

$j = 1,2$, *and* (Z_1, Z_2) *is a random vector* with an asymptotic bivariate normal distribution, each following the standard normal distribution with asymptotic correlation given by

$$\frac{\frac{1}{n_T}p_T(1-p_T) + \frac{1}{n_R}(p_R(1-p_R) - k^2(0.5-p_R)^2)}{\sqrt{\nu_{1,2}(p_T,p_R)\nu_{2,2}(p_T,p_R)}}.$$

Then the power function can be given approximately by the probability of a bivariate normal distribution over a region defined by Eq. (15), which can be calculated numerically by the pmvnorm function in the R package mvtnorm (Genz et al., 2018; Genz and Bretz, 2009).

The selection of the margin multiplier *k*

Two methods can be used to determine k based on the sample size of the study (Ren et al., 2019). The first method is intended to be used for large sample studies, and the second for small sample studies. First, we note that for a constant k, when $p_R < \frac{k^2}{1+k^2}$ then $p_R - k\sqrt{p_R(1-p_R)} < 0$, and thus H_{10} is always rejected since $p_T \geq 0$ is a probability. Similarly, if $p_R > \frac{1}{1+k^2}$, then $p_R + k\sqrt{p_R(1-p_R)} > 1$, and thus H_{20} is always rejected.

For studies with large sample sizes, k may be selected so that the variable margin is similar to margins used in previous studies. As an example, for the biosimilar immunogenicity trial of insulin product in Wang (2018), which has at least 250 per arm, we match the step-function margin with our variable margin at p_R = 0.5,

$$0.131 = k\max_{p_R}\sqrt{p_R(1-p_R)} = k\sqrt{0.5 \times (1-0.5)}. \quad (16)$$

Solving Eq. (16) for k, then we have k = 0.262 and margin function $\delta(p_R) = 0.262\sqrt{p_R(1-p_R)}$. In fact, the margin function agrees with the step function at the thresholds.

For small sample studies, setting k = 0.262 may render the test of little power. Simulation studies reported below showed that when $n_T = n_R = 50$ the equivalence test cannot reject H_0 for any combination of p_T and p_R. However, there are practical situations where it is impractical to obtain relevant large sample sizes, while sufficient power is still needed. In this regard, one may choose k for a test statistic by setting the power function in Eq. (15) to be $1 - \beta$ and solve for k for some presumed p_T and p_R. The solved k is implicitly also a function of p_T, p_R and the sample sizes. To facilitate the discussion, we consider the case when both test and reference response rates are the same, i.e., $p_T = p_R = p$, so that the equivalence test has the maximum power, and the sample sizes $n_T = n_R = n$.

For the power function in Eq (15), the main difficulty lies in the unknown $\bar{p}_T$ and $\bar{p}_R$ and their dependency with k. Because no explicit formulas of $\bar{p}_T$ and $\bar{p}_R$ are available, we estimate it by bootstrap for any given k. Then k is solved by the R function uniroot in the stats package (R Core Team, 2017).

Assuming α = 0.05, β = 0.1, i.e., 90% power, and n = 50, 100, 150, the calculated margin multipliers for T_{RW} are illustrated in Fig. 1. Except when p is very small, the margin multiplier k seems to be symmetric at $p = 0.5$ and an increasing function with respect to $|p - 0.5|$. Note that when $n = 50$, very large k is required to obtain the prespecified power. A large k can result in trivial lower bound $p_R - kp_R(1-p_R)$ or upper bound $p_R + kp_R(1-p_R)$ and this is more often as k increases. The induced margins are shown in Fig. 2. Interestingly, the induced margin curves are rather flat for a given sample size except when p_R is close to 0 or 1. For $n = 50$, the margin ranges from 0.315 to 0.341, excluding the margins for p_R = 0.05 and p_R = 0.95. The margin ranges from 0.18 to 0.236 for *n* = 100 and from 0.122 to 0.192 for *n* = 150.

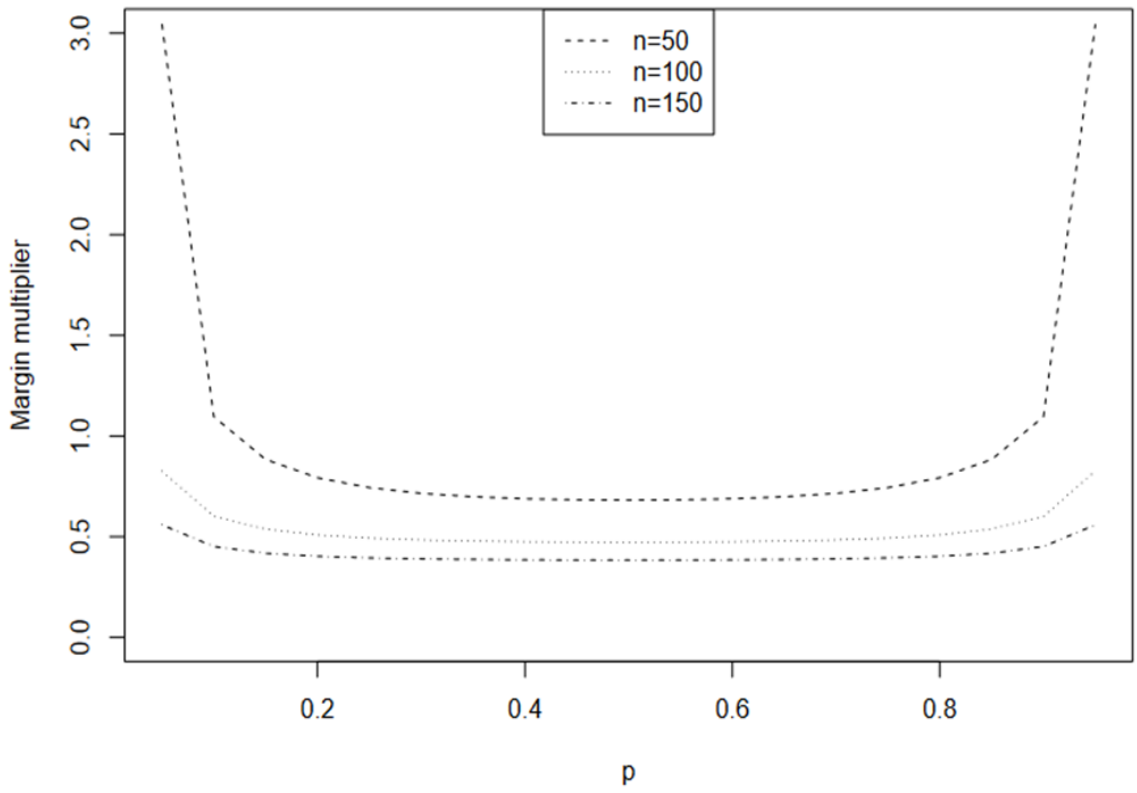

Fig 1: Plot of margin multiplier as a function of p to obtain a constant power of 90% and type I error 5% with $p_T = p_R = p$ and $n_T = n_R = 50$.

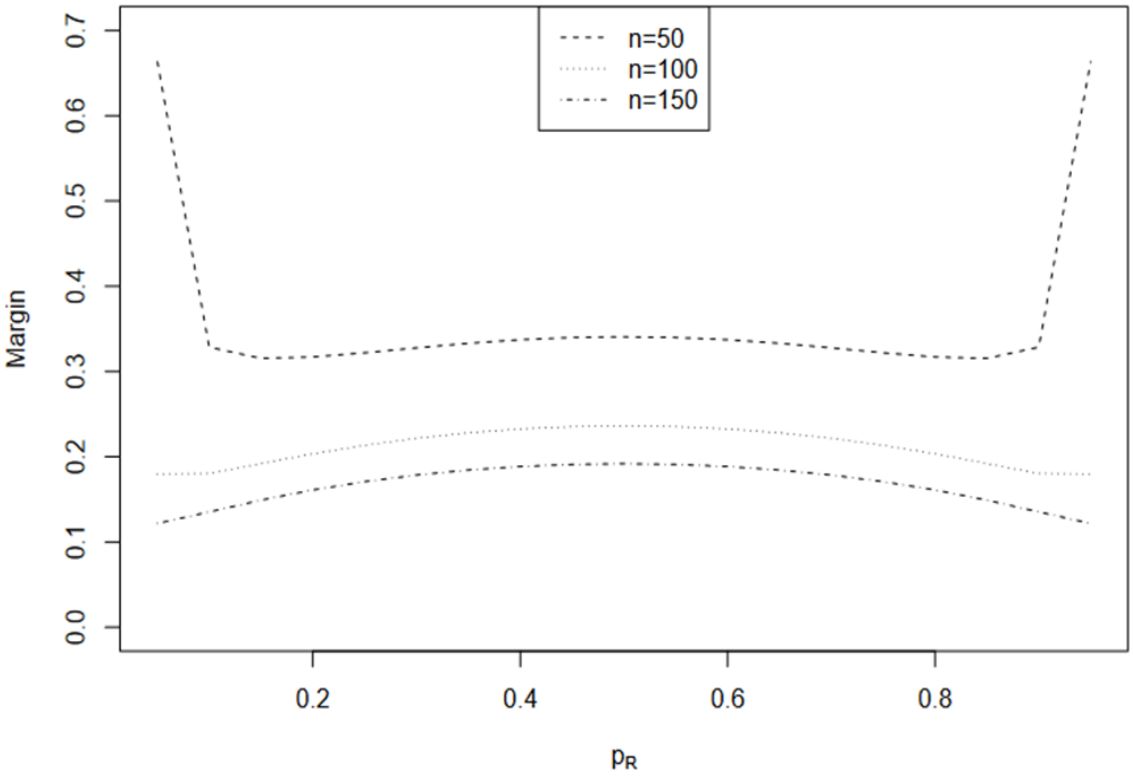

Fig 2: Plot of the variable margin to obtain a constant power of 90% and type I error 5% with $p_T = p_R = p$ and $n_T = n_R = 50, 100, 150$.

Results

Empirical type I error

Here we report a simulation study of the empirical type I error of the three test statistics in large sample studies with $k = 0.262$. We assume equal sample size for both arms, and consider a variety of sample

sizes, $n_T = n_R = 50, 100, \cdots, 500$ and the true reference probability $p_R = 0.1, 0.2, \cdots, 0.9$. For each p_R, the margin and true p_T are given by $\delta(p_R) = kp_R(1-p_R)$ and $p_T = p_R - \delta(p_R)$ respectively. For given sample size and true response rates, the test and reference data are simulated$^{\delta}$ by $X_{I,i} \sim Bernoulli(p_I)$ for $i = 1, \cdots, n_I$, and $I = T, R$. Three equivalence tests based on test statistics T_{RWO}, T_{MWO}, and T_{RW} are performed for each simulated sample at a nominal level of $\alpha = 0.05$. The simulation study is replicated for 10^6 times.

The performances of the three tests are similar when sample sizes changes, so only empirical rejection rates (ERR) (type I errors) for T_{RWO}, T_{MWO}, and T_{RW} and theoretical rejection rates (TRR) for T_{RW} are reported in Figure 3 for the case when $n_T = n_R = 50, 100, 150, 200, 250, 500$. When sample size is 50, all tests have almost zero rejection rate, regardless the true reference probability. This is due to the small sample size and the small k value. When sample size is 100, the rejection rates are around 0.03. Note that the theoretical rejection rate for T_{RW} is also close to 0.03. The rejection rates for sample sizes at least 150 are similar across different sample sizes, showing that both T_{RWO} and T_{MWO} have seriously bias in type I error, compared with T_{RW} for which both the empirical type I error and theoretical approximations are much closer to the nominal 0.05 level.

The evident trends in ERR for T_{RWO} and T_{MWO} are mainly due to the variance estimates ignoring the variability in the margin. The uprising trend cross the 0.05 nominal size at $p_R = 0.5$ may be explained below. First note that the simulation study is set at the lower boundary $p_T - p_R = -\delta$. Thus, the sample estimate $\hat{p}_I$ is close to the RMLE $\check{p}_{i,I}$, both of which are consistent estimators of p_I for $I = T, R$. This is the reason for the comparable rejection rates of T_{RWO} and T_{MWO}. Also, $H_{2,0}$ is almost always rejected. So the type I error of the equivalence test is close to the rejection rate of $H_{1,0}$. For this selected k and any $\check{p}_{i,R} < 1/2$, $v_1(\check{p}_{i,T}, \check{p}_{i,R}) > v_{1,2}(\check{p}_{i,T}, \check{p}_{i,R})$, implying $T_{1,RWO} < T_{1,RW}$ and thus $T_{1,RWO}$ rejects $H_{1,0}$ less frequently than $T_{1,RW}$.

Additional simulation study conducted with $p_T = p_R + \delta(p_R)$ unreported here shows a mirrored pattern for T_{MWO} and T_{RWO}, of which downward trends were observed for type I error.

Empirical power of T_{RW}

Due to the inferior performance of T_{MWO} and T_{RWO} in controlling type I error, only T_{RW} is considered in the power study. The simulation settings for power function of large sample studies are similar to those in Section 3.1. For any $p_T \in [p_R - \delta, p_R + \delta]$, its deviation from H_0 can be measured by $\Delta = (p_T - p_R)/\delta \in [-1, 1]$, interpreted as the signed distance between the test probability p_T and p_R standardized by δ. The reference probability p_R is from 0.1 to 0.9 with increment of 0.1. For each p_R, we consider a series of p_T's with the corresponding Δ from −1 to 1 with an increasement of 0.25. In particular, $\Delta = \pm 1$ implies that the test response rate is at the H_0 boundary and $\Delta = 0$ implies that the test and reference response rates are identical.

The simulation results are illustrated in Fig. 4. The power for $n = 50$ is virtually 0 for all combinations of true test and reference probability. This is likely to be due to the small k value and sample size. For a fixed p_R, the power in general increases as sample size increases and p_T is closer to p_R. However, the power curves seem to be asymmetric at $\Delta = 0$, which is due to the variance components of the margin variability and unseen for equivalence test with a fixed margin. The only case for a symmetric power curve is when $p_R = 0.5$. For $p \neq 0.5$, the power curves for $p_R = p$ appears to be the power curves for $p_R = 1 - p$ flipped at $\Delta = 0$.

Verification of margin multiplier k for a prespecified power

Here we verify the calculation of margin multiplier k for T_{RW}. The data are simulated with $p_T = p_R = p$ and $n = n_T = n_R = 50, 100, 150$ with $p = 0.05, 0.1, \cdots, 0.90$. For each n and each p, two sets of simulations are performed. In the first set, the test is performed with the k calculated with a target of 90% power and true p with given n. In the second set, the test is performed with the previously computed k which yields 90% power for $p_T = p_R = 0.5$,

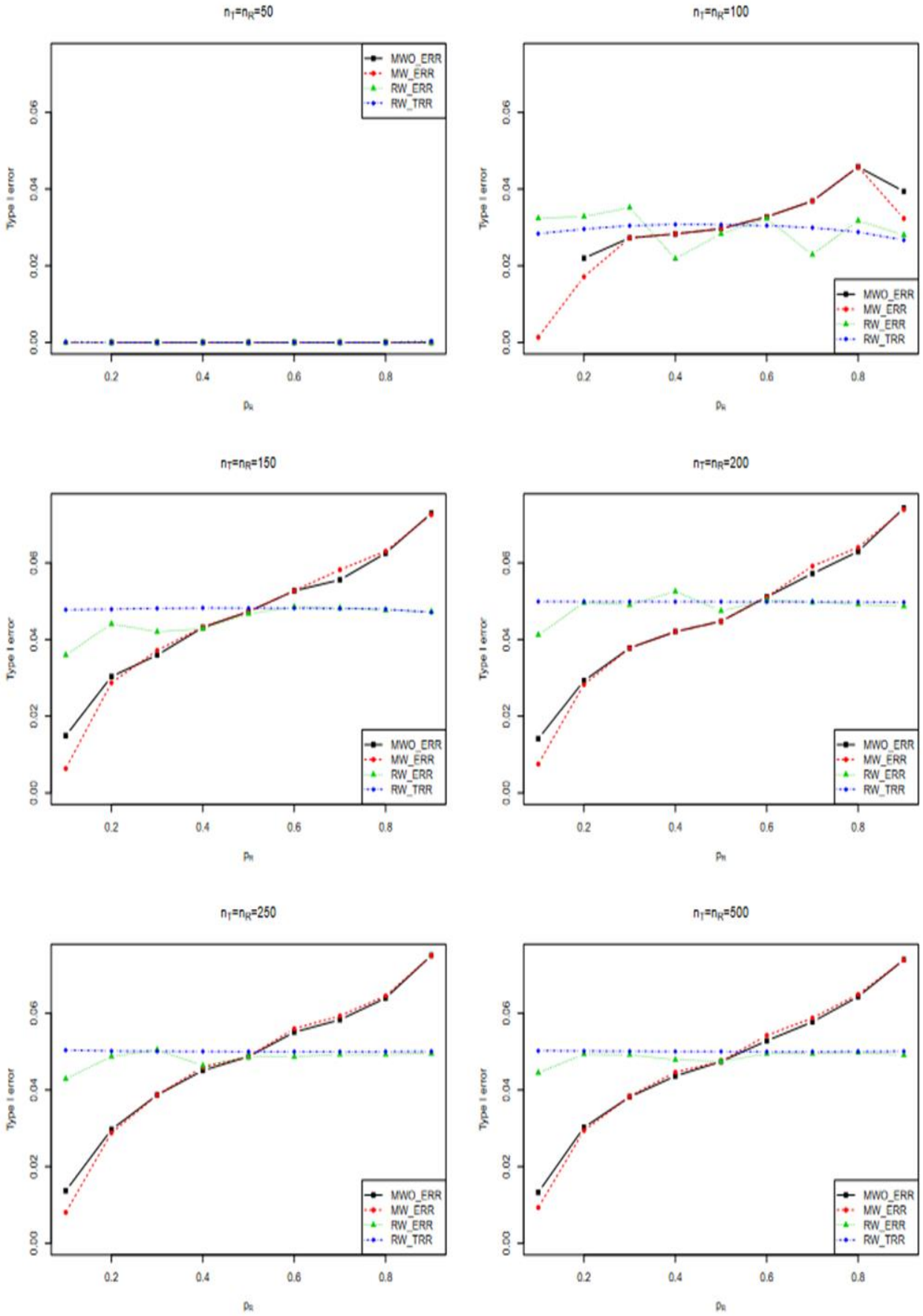

Fig 3: The empirical and theoretical rejection rates for equivalence tests.

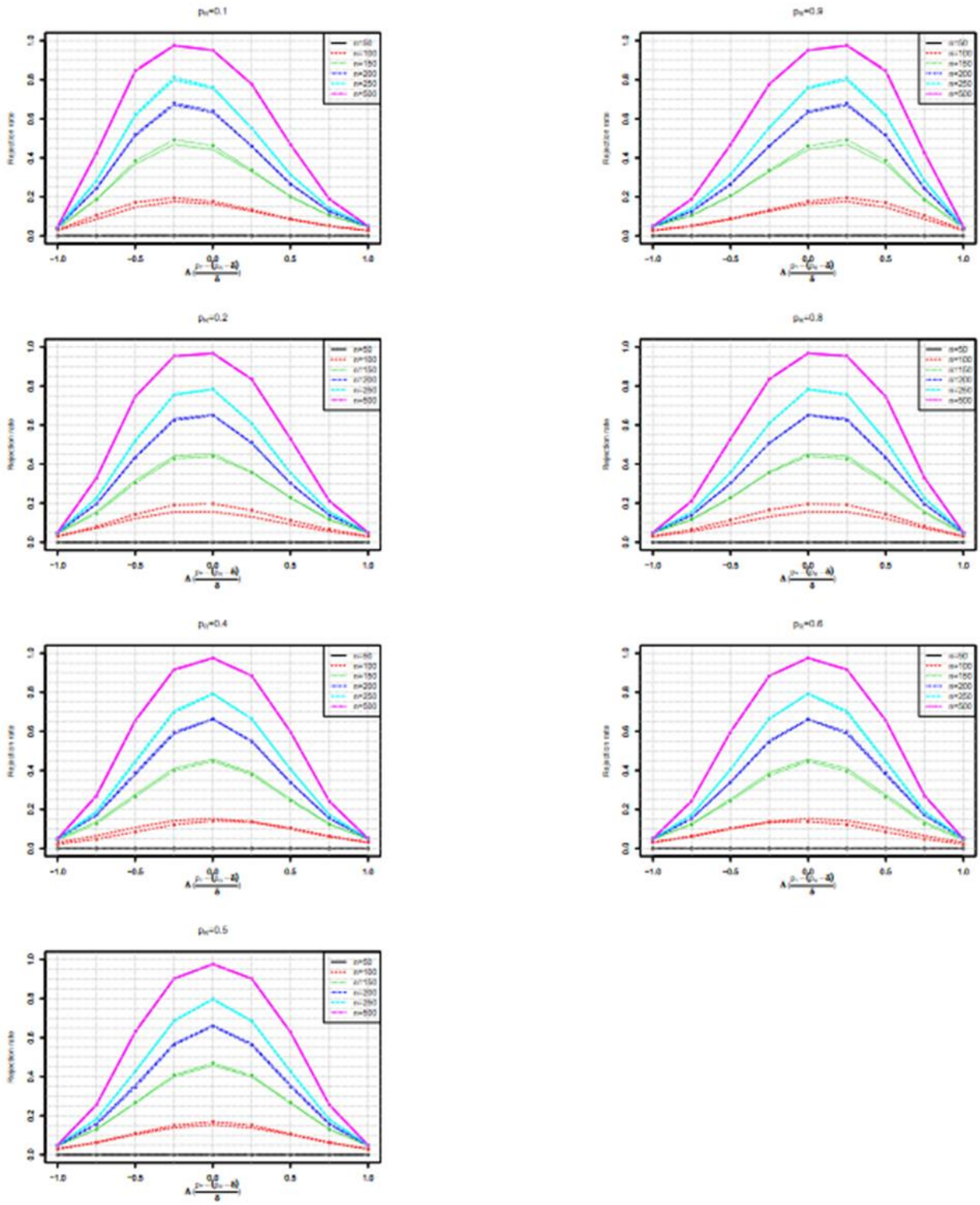

Fig 4: The empirical and theoretical power functions for T_{RW} with $k = 0.262$. The empirical rejection rates are illustrated by lines with solid circles and the theoretical rejection rates by lines only. The results for different sample sizes are shown in different lines. regardless the true $p = p_T = p_R$. The empirical powers are based on 105 replications and illustrated in Fig. 5. For the tests computed with the k calculated with true p, the empirical powers very close to 0.9. When a fixed k is used in the test, the empirical power decreases as p deviates from 0.5, which is more evident for $n = 50$ but less so for $n = 100$ and $n = 150$, due to the fact that the k values are more stable for relatively large sample sizes (cf. Fig. 1).

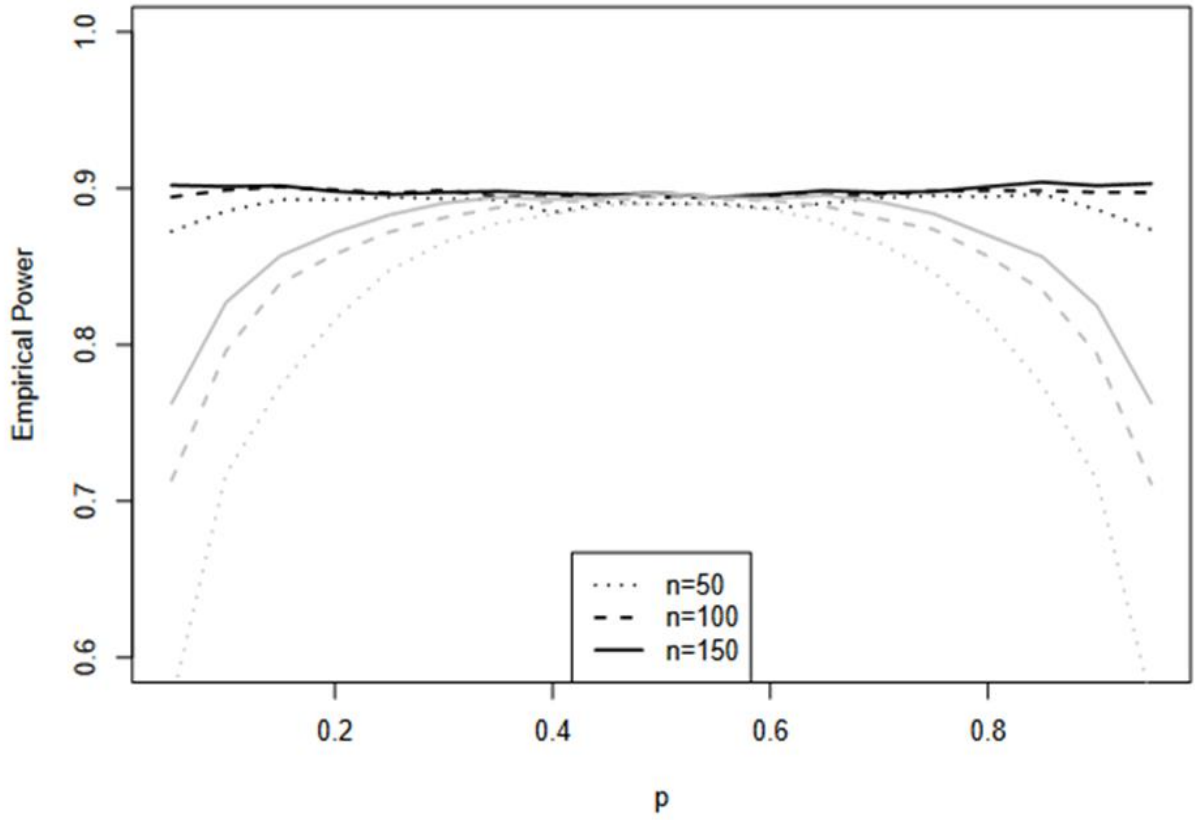

Fig 5: Plot of empirical rejection rates (power) of T_{RW} for true response rate p with the corresponding k targeting power 90% (black lines) and the empirical power for different p tested with k fixed at the value which yields 90% power for $p = 0.5$ (grey lines).

Discussion

In this paper we extended the discussion of comparison of binary test and reference responses with a variable margin in the form of reference scaled difference in means from NI test (Ren et al., 2019) to equivalence test. Three test statistics were studied and both theoretical discussion and simulation studies showed that the variability in the margin should be taken into account to construct proper test statistics so T_{RW} is recommended. The type I error of T_{RW} is closer to the nominal size for p_R close to 1 than p_R close to 0. Our calculation of the margin multiplier showed that when a fixed power is desired, the margin does not vary much as the reference response rate changes. Simulation studies showed that small margin multiplier may result in zero rejection rates so sufficiently large sample size should be maintained to achieve proper control of type I error and desired power as the weakness of two one-sided tests.

Appendix

For T_{RW}, the power function is

$$
\begin{aligned}
&P(T_{1,RW} > Z_{1-\alpha} \,\&\, T_{2,RW} < -Z_{1-\alpha}) \\
=&P\left(\frac{\hat{p}_T - \hat{p}_R + k\sqrt{\hat{p}_R(1-\hat{p}_R)}}{\sqrt{\nu_{1,2}(\check{p}_{1,T},\check{p}_{1,R})}} > Z_{1-\alpha} \,\&\, \frac{\hat{p}_T - \hat{p}_R - k\sqrt{\hat{p}_R(1-\hat{p}_R)}}{\sqrt{\nu_{2,2}(\check{p}_{2,T},\check{p}_{2,R})}} < -Z_{1-\alpha}\right) \\
=&P\left(\frac{\hat{p}_T - f_1(\hat{p}_R) - (p_T - f_1(p_R))}{\sqrt{\nu_{1,2}(\check{p}_{1,T},\check{p}_{1,R})}} > Z_{1-\alpha} - \frac{p_T - f_1(p_R)}{\sqrt{\nu_{1,2}(\check{p}_{1,T},\check{p}_{1,R})}} \,\&\right. \\
&\left.\frac{\hat{p}_T - f_2(\hat{p}_R) - (p_T - f_2(p_R))}{\sqrt{\nu_{2,2}(\check{p}_{2,T},\check{p}_{2,R})}} < -Z_{1-\alpha} - \frac{p_T - f_2(p_R)}{\sqrt{\nu_{2,2}(\check{p}_{2,T},\check{p}_{2,R})}}\right) \\
=&P\left(\frac{\hat{p}_T - f_1(\hat{p}_R) - (p_T - f_1(p_R))}{\sqrt{\nu_{1,2}(p_T,p_R)}} > \sqrt{\frac{\nu_{1,2}(\check{p}_{1,T},\check{p}_{1,R})}{\nu_{1,2}(p_T,p_R)}}\left(Z_{1-\alpha} - \frac{p_T - f_1(p_R)}{\sqrt{\nu_{1,2}(\check{p}_{1,T},\check{p}_{1,R})}}\right)\right. \\
&\left.\& \frac{\hat{p}_T - f_2(\hat{p}_R) - (p_T - f_2(p_R))}{\sqrt{\nu_{2,2}(p_T,p_R)}} < \sqrt{\frac{\nu_{2,2}(\check{p}_{2,T},\check{p}_{2,R})}{\nu_{2,2}(p_T,p_R)}}\left(-Z_{1-\alpha} - \frac{p_T - f_2(p_R)}{\sqrt{\nu_{2,2}(\check{p}_{2,T},\check{p}_{2,R})}}\right)\right) \\
:=&P\left(Z_1 > \sqrt{\frac{\nu_{1,2}(\check{p}_{1,T},\check{p}_{1,R})}{\nu_{1,2}(p_T,p_R)}}\left(Z_{1-\alpha} - \frac{p_T - f_1(p_R)}{\sqrt{\nu_{1,2}(\check{p}_{1,T},\check{p}_{1,R})}}\right)\right. \\
&\left.\& Z_2 < \sqrt{\frac{\nu_{2,2}(\check{p}_{2,T},\check{p}_{2,R})}{\nu_{2,2}(p_T,p_R)}}\left(-Z_{1-\alpha} - \frac{p_T - f_2(p_R)}{\sqrt{\nu_{2,2}(\check{p}_{2,T},\check{p}_{2,R})}}\right)\right) \\
\approx&P\left(Z_1 > \sqrt{\frac{\nu_{1,2}(\bar{p}_{1,T},\bar{p}_{1,R})}{\nu_{1,2}(p_T,p_R)}}\left(Z_{1-\alpha} - \frac{p_T - f_1(p_R)}{\sqrt{\nu_{1,2}(\bar{p}_{1,T},\bar{p}_{1,R})}}\right)\right. \\
&\left.\& Z_2 < \sqrt{\frac{\nu_{2,2}(\bar{p}_{2,T},\bar{p}_{2,R})}{\nu_{2,2}(p_T,p_R)}}\left(-Z_{1-\alpha} - \frac{p_T - f_2(p_R)}{\sqrt{\nu_{2,2}(\bar{p}_{2,T},\bar{p}_{2,R})}}\right)\right)
\end{aligned}
$$

There is no deterministic order between the random variables Z_1 and Z_2. Both have asymptotic standard normal distribution with asymptotic correlation given by

$$
\frac{\frac{1}{n_T}p_T(1-p_T) + \frac{1}{n_R}\left(p_R(1-p_R) - k^2(0.5-p_R)^2\right)}{\sqrt{\nu_{1,2}(p_T,p_R)\nu_{2,2}(p_T,p_R)}}
$$

Then the power function can be given approximated by the probability in the last display, which is a probability of a bivariate normal distribution over a region, can be calculated by the function pmvnorm of the R package mvtnorm (Genz et al., 2018; Genz and Bretz, 2009).

References

1. Chow, S.-C., H. Wang, and J. Shao. Sample Size Calculations in Clinical Research (2nd ed.). Chapman and Hall/CRC.
2. FDA (1992). Points to consider. Clinical development and labeling of anti-infective drug products. Silver Spring, Maryland: U.S. Department of Health and Human Services, FDA, CDER.
3. FDA (2001). FDA Guidance for Industry: Statistical approaches to establishing bioequivalence. Silver Spring, Maryland: U.S. Department of Health and Human Services, FDA, CDER.
4. FDA (2003). FDA Guidance for Industry: Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action. Silver Spring, Maryland: U.S. Department of Health and Human Services, FDA, CDER.
5. FDA (2011). FDA Guidance for Industry:

Bioequivalence recommendations for progesterone oral capsules. Silver Spring, Maryland: U.S. Department of Health and Human Services, FDA, CDER.

6. FDA (2016). Guidance for Industry. Non-inferiority clinical trials to establish effectiveness. Silver Spring, Maryland: U.S. Department of Health and Human Services, FDA, CDER.
7. Genz, A., & Bretz, F. (2009). Computation of multivariate normal and t probabilities (Vol. 195). Springer Science & Business Media.
8. Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2018). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-8.
9. R Core Team (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
10. Ren, Y., Wang, C., Shen, M., & Tsong, Y. (2019). Non-inferiority tests for binary endpoints with variable margins. Journal of Biopharmaceutical Statistics, 29(5), 822-833.
11. Röhmel, J. (1998). Therapeutic equivalence investigations: statistical considerations. Statistics in medicine, 17(15-16), 1703-1714.
12. Röhmel, J. (2001). Statistical considerations of FDA and CPMP rules for the investigation of new anti-bacterial products. Statistics in Medicine, 20(17-18), 2561-2571.
13. Tothfalusi, L., & Endrenyi, L. (2016). An exact procedure for the evaluation of reference-scaled average bioequivalence. The AAPS journal, 18, 476-489.
14. Tsong, Y. (2007). The utility of active-controlled noninferiority/equivalence trials in drug development. International Journal of Pharmaceutical Medicine, 21, 225-233.
15. Wang, Y. (2018). Considerations in designing comparative immunogenicity study. Presentation at 2018 ICSA Applied Stat. Symposium.
16. Yuan, M., J. Luan, and H. Sun (2018). New proposal for equivalence testing on binary endpoint in clinical bioequivalence studies. Presentation at 2018 ICSA Applied Statistics Symposium.
17. Zhang, Z. (2006). Non-inferiority testing with a variable margin. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 48(6), 948-965.