A Statistical Method for Analyzing Risk Difference in Trials with a Three-Level Paired Design

Xinjun Wang

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2017

Committee:

Xiao-Hua (Andrew) Zhou

Patrick Heagerty

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

A Statistical Method for Analyzing Risk Difference
in Trials with a Three-Level Paired Design

Xinjun Wang

Chair of the Supervisory Committee:
Professor Xiao-Hua (Andrew) Zhou
Biostatistics

This thesis is motivated from an animal trial for a wearable external cardiac defibrillator. Each pig in the trial will be treated with two devices (test vs. control) after induced to experience ventricular fibrillation multiple times. A set of different defibrillation waveforms will be tested and are randomized in blocks. The efficacy of two waveforms, defined as the probability of shock success, will be compared in absolute difference. We propose a statistical method, applying t-test on a discrete variable transformed from the original binary outcome, to study the (average) risk difference in trials with a similar three-level paired design, which can properly adjust for the random effects for pig (subject) and block, as well as for the random slope of treatment for pig, which is also known as the "*Subject – Varying Treatment Effect*". We develop two naive data generating procedures for the simulation study, and test the performance of both procedures. The results of a pilot study are used to inform design parameters used to simulate data. The performance of our proposed method is evaluated through a set of simulation studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to express my great gratitude to Dr. Andrew Zhou and Dr. Patrick Heagerty for their endless support, guidance, and patience during my master's thesis. This experience would be very beneficial to my future career.

Second, I would like to acknowledge our collaborators, including Tyson Taylor and Fred Chapman at Physio-Control, and Professor Gregory P. Walcott at the University of Alabama at Birmingham, for sharing and explaining their interesting study to us, and providing us the data of their pilot study for analysis.

Finally, I would like to thank all people in our department for providing a terrific academic environment during my master's study.

DEDICATION

To my family

Chapter 1. INTRODUCTION

1.1 MOTIVATION FROM AN ANIMAL TRIAL

This master's thesis was motivated from an animal trial I have been working on with my thesis advisor Dr. Andrew Zhou. In this research, our primary goal is to help Kestra Medical Technologies (Kirkland, WA) calculate the sample size of pigs required to demonstrate non-inferiority of the efficacy of their newly designed external cardiac defibrillation waveform, compared to the waveform produced by an external cardiac defibrillator that is currently approved for patient used in the United States. Results from a pilot study will be used to inform design parameters used to simulate data and to compute power for candidate analysis strategies.

In the proposed animal study, different defibrillator waveforms will be used, which are defined by their device type and by the different simulated transthoracic impedance (TTI) values. Different TTI values are achieved by adding resistors in series with the defibrillator and the pig. In our study, the TTI values may not be associated with efficacy because the therapy is attenuated differently for waveforms at each TTI. Thus, we will do separate analysis for each TTI value. In each study episode, a pig will be first induced to experience ventricular fibrillation (VF), and then a randomly selected defibrillation waveform will be delivered. If the shock is unsuccessful, up to three rescue shocks will be delivered by a LIFEPAK (LP12) as soon as possible. The protocol ends for that animal if the rescue shocks are unsuccessful. Defibrillation waveforms used for each episode are block randomized, i.e., each block contains all different episodes in randomized order. Because of confidentiality, an overview of the study protocol for each animal is not shown.

In general, this prospective block randomized study will test the efficacy of all different defibrillation waveforms used to treat episodes of electrically-induced VF in a closed-chest porcine model of short duration cardiac arrest. The efficacy of each defibrillation waveform, which is defined as the probability of 1st shock defibrillation success, will be compared between the test and the control waveforms at each TTI value. Non-inferiority can be demonstrated if the difference in efficacy between two devices (test - control) at all different TTI values is greater than -10%, the absolute risk difference according to ANSI/AAMI Standard (ANSI/AAMI Standard DF80:2003. Medical electrical equipment – Part 2-4).

There are two major challenges to statistically analysis of data from this study. First, the direct measurement of the study has a binary outcome, shock success or failure, but as we are interested in studying the difference in probabilities (also known as "risk difference") instead of an odd ratio, we prefer using a linear model compared to a logit model. Second, the study design is a three-level paired design, where the study episode, block and subject (pig) are corresponding level 1 to level 3 structure, respectively. Thus, we need to properly adjust for or eliminate both within block correlation and within subject correlation.

1.2 The Analysis of Binary Outcomes

When the clinical outcome of interests is a binary variable such as "success" or "failure", the most common statistical method used to evaluate the relationship between predictors and the outcome would be logistic regression (Cox, 1958). In logistic regression, we model the odds of the response variable as a log-linear function of candidate predictor variables. Logistic regression can also be viewed as a special case of generalized linear regression with a logit link function.

The followings are advantages of using logistic regression over linear models for binary outcomes. First, the logit function is always bounded between 0 and 1; Second, when the

outcome is rare, the odds ratio is a good approximation of risk ratio (Greenland & Thomas, 1982); Third, the S-shaped relationship between the probability of getting a disease given predictors, P(D|X), and predictors, X, specified by logistic regression is what one might expect for probabilities. However, when the outcome is not rare, odds ratios and risk ratios may not be close. The understanding of odds ratios is often difficult for non-statisticians, and most of the time, they interpret an odds ratio as a risk ratio, which will lead to misleading a conclusion. Nevertheless, there are a variety of alternative methods to analyze a binary outcome.

Relative risk regression (Zhang J, 1998), also called log-binomial regression, is used to estimate a relative risk (risk ratio) instead of an odds ratio. As a result, the interpretation of relative risk regression is generally more straight forward. Unlike logistic regression, the relationship between P(D|X) and X is exponential in relative risk regression, which means that we could get P(D|X) > 1. Thus, computational challenges might exist because of the constraint that probabilities must be bounded between 0 and 1. Algorithms for finding maximum likelihood estimates might fail to converge. The other popular alternative is Poisson regression, which is a general regression method when the outcome is count data. However, we can also use Poisson regression for a binary response, where the outcome is either 0 or 1. The Poisson regression and the relative risk regression are very similar in that they are both generalized linear regression with a log link function, and for each combination of independent variables, the distribution of the dependent variable is assumed to be binomial distribution. However, the difference is that Poisson model treats this distribution as Poisson distribution while the relative risk regression correctly treats it as binomial distribution. Researchers have shown that for very high prevalence and with moderate sample size, the Poisson regression gives less biased estimates of risk ratio than the relative risk regression. However, when there are moderate prevalence and moderate

sample size, the relative risk regression yields less biased estimates than the Poisson regression. In addition, the relative risk regression generally has slightly higher power and smaller standard error than the Poisson regression (Petersen & Deddens, 2008).

Linear regression is commonly used to evaluate the relationship between predictors and a continuous outcome. However, when the outcome is binary, linear regression could still be used if the researchers are interested in a risk difference and the sample size is sufficiently large (Lumley, Diehr, Emerson, & Chen, 2002). The greatest advantage of linear regression is its interpretability. In a linear model with a binary outcome, the coefficient of a predictor simply represents the corresponding risk difference. When there is only one predictor in the linear regression and it is a binary variable, then it is the same as a t-test. Equivalently, a t-test could also be used in the case of binary outcome if the sample size is not small.

Furthermore, in some situations a research may be instead interested in classification approaches, and in this case there are a lot of potential discriminant analysis methods such as linear discriminant analysis (LDA) could be used (Fisher, 1936). In this manuscript, the motivating example is not focused on classification performance but rather statistical inference comparing two experimental conditions and therefore regression methods are most appropriate for generating confirmatory inference.

All of the methods mentioned above are potential valid strategies when dealing with binary outcomes. The choice of which one to use in practice should always depend on the scientific questions.

1.3 PROPOSED STATISTICAL MODEL AND TEST

1.3.1 *The Use of a Linear Model*

For each TTI value, we can assume the following linear model with non-parametric error distribution, where Y is the direct measurement of success (represented by 1) or failure (represented by 0) of a specific shock:

$$Y_{ijk} = (\alpha + \beta X_{ijk}) + b_i X_{ijk} + (e_i + e_{ij} + e_{ijk}) \quad \text{- Model (1)}$$

where,

i: Subject (pig) number: from 1 to n_I

j: Block number: from 1 to n_J

k: 1 for test device; 0 for control device

X_{ijk}: treatment device, where X_{ij1} = 1 for new device; X_{ij0} = 0 for control device

α – baseline successful rate for control device

β – average effect size (risk difference), *parameter of interest*

$b_i \sim (0, \sigma_b^2)$ – random slope of treatment for subject

$e_i \sim (0, \sigma_1^2)$ – random intercept for subject

$e_{ij} \sim (0, \sigma_2^2)$ – random intercept for block

$e_{ijk} \sim (0, \sigma_3^2)$ – residual error

Here e_i and e_{ij} are used to specify the within-cluster correlation for subject and block effects, respectively. Here, b_i is used to indicate the *Subject (pig) – Varying Treatment Effect* (*SVTE*), which means that we allow the difference in efficacy between two devices to vary across different subjects. If the *SVTE* is not included the model, we would make an assumption that such a difference in efficacy stays constant across different subjects, which may not be accurate

in practice. The effect size β is the parameter of interest, and our goal is to estimate β. Note that β here is the average effect size since it varies across different subjects.

Here, we do not assume any distribution for b_i, e_i, e_{ij}, and e_{ijk}, and we only assume that those variables are all independent, each with mean 0 and finite variance.

1.3.2 *An Unbiased Estimator for the Average Effect Size β*

Although we include three random effects in *Model (1)*, due to the paired design of the study we could easily eliminate the two random intercepts, e_i and e_{ij}, by using the derived variable that compares the outcomes across the two treatment conditions: Y_{ij1} - Y_{ij0}.

Let $V_{ij} = Y_{ij1} - Y_{ij0}$, then

$V_{ij} = (\alpha + \beta) + b_i + (e_i + e_{ij} + e_{ij1}) - (\alpha + e_i + e_{ij} + e_{ij0})$

$= \beta + b_i + e^*$, where $e^* = e_{ij1} - e_{ij0}$

Here, the resulting V_{ij}'s (n = $n_I * n_J$ observations) are not independent due to b_i. If *SVTE* does not need to be included in the model, i.e., b_i is not in the model, we can easily come up with a method using $\overline{V_{IJ}} = \frac{1}{n_I * n_J}\sum_{ij} V_{ij}$ as an unbiased estimator for β. However, in our case we want to find a method that is relative robust to different circumstances, so we need to appropriately adjust for the effect of b_i by looking for another estimator. A direct method would be to further transform the direct outcome Y_{ijk} by averaging them across all the different blocks for each device. By doing so, we would have a smaller sample size (n = n_I observations after transformation) compared to using V_{ij}'s directly, but we are able to adjust for the effect of b_i.

Let $W_i = Y_{i.1} - Y_{i.0} = \frac{1}{n_J}\sum_j Y_{ij1} - \frac{1}{n_J}\sum_j Y_{ij0}$. Then,

$$W_i = \frac{1}{n_J}\sum_j((\alpha + \beta) + b_i + (e_i + e_{ij} + e_{ij1}) - (\alpha + (e_i + e_{ij} + e_{ij0}))$$

$= \beta + b_i + \frac{1}{n_J}\sum_j (e_{ij1} - e_{ij0})$

$= \beta + e_i^*$, *where* $e_i^* \sim (0, \sigma_*^2)$ *- Model (2)*

and $\sigma_*^2 = \begin{cases} \frac{2}{n_J}\sigma_3^2 + \sigma_b^2, & assuming\ comman\ variance\ for\ e_{ij1} \text{and}\ e_{ij0} \\ \frac{1}{n_J}\left(\sigma_{ij1}^2 + \sigma_{ij0}^2\right) + \sigma_b^2, & assuming\ unequal\ variances\ for\ e_{ij1} \text{and}\ e_{ij0} \end{cases}$

Thus, we have E[W_i] = β, and Var(W_i) = σ_*^2.

As a result, $\overline{W_i} = \frac{1}{n_I}\sum_i W_i$ is an unbiased estimator for β. Furthermore, W_i 's are independent, and can be easily computed from the direct outcome Y_{ijk}.

1.3.3 *The Use of a t-test*

By the Central Limit Theorem, we have

$\sqrt{n_I}\,(\overline{W_i} - \beta) \xrightarrow[d]{} N\,(0, \sigma_*^2)$.

Thus, we have $T_i = \frac{\overline{W_i} - \beta}{s\,/\,\sqrt{n_I}} \sim t_{(n_I - 1)}$, where s^2 is the sample variance of W_i, and we can then apply a t-test on W_i's for hypothesis testing (non-inferiority) as follows:

For each of three different TTI values,

$\begin{cases} H_0\text{: } \beta \leq \text{NI margin} \\ H_1\text{: } \beta > \text{NI margin} \end{cases}$

We reject the null hypothesis if the one-sided test is significant at α = 0.025 level, i.e., we reject H_0 when $t_{obs} = \frac{\overline{W_i} - NI\ margin}{s\,/\,\sqrt{n_I}} > t_{0.025,(n_I - 1)}$.

Here, W_i is not continuous However, by definition, $W_i \in \{-1, -1+\frac{1}{n_J}, -1+\frac{2}{n_J}, \ldots, 1\}$, which still provides a relatively large pool of values. For example, when the number of block in the study (n_J) is 8, $W_i \in \{-1, -\frac{7}{8}, -\frac{6}{8}, \ldots, \frac{6}{8}, \frac{7}{8}, 1\}$), and has 17 different values.

As a result, applying a t-test on W_i's should be valid for hypothesis testing when sample size is large. We will evaluate the performance of t-test in a simulation study when the sample size is not too large in *Chapter 3*.

1.4 PILOT STUDY

A pilot study was conducted and the data was collected in order to inform subsequent studies. In the pilot study, 10 pigs were used (n_I = 10), and the number of block was set at 8 (n_J = 8). The three TTI values used were 25, 50, and 100 **Ω**. These TTI values were not exactly the same as those to be used in the future design, in which only 50 **Ω** will be used again. Thus, the data from shocks with TTI = 50 **Ω** in the pilot study will be used to provide point estimates that can be used in the simulation studies.

Based on *Model (2)*, we can directly calculate $\overline{W_i}$ and s^2, the sample variance of W_i, from the pilot data. As a result, $\overline{W_i}$ = 0.10, and s = 0.23. In addition, we can estimate all the parameters in *Model (1)* using a linear mixed model to account for multilevel structure. Although the standard error for $\hat{\beta}$ may not be valid due to heteroscedasticity, the point estimates for α and β using the best linear unbiased estimator (BLUE), as well as the random effects using the best linear unbiased predictor (BLUP) should still be unbiased and thus valid. As a result, $\hat{\alpha}$ = 0.36, $\hat{\beta}$ = 0.11, $\widehat{\sigma_b^2}$ = 0.01, $\widehat{\sigma_1^2}$ = 0.05, $\widehat{\sigma_2^2}$ < 0.001, $\widehat{\sigma_3^2}$ = 0.18.

As mentioned above, the TTI values may not be associated with waveform efficacy. From the results of the pilot study, we observed that $\hat{\beta}$ = 0.05 for TTI = 25 **Ω**, and $\hat{\beta}$ = 0.16 for TTI = 100 **Ω**. Thus, the investigators thought that the larger TTI value could be associated with a larger difference in efficacy between the two devices (the test device with a greater efficacy) because the difference in waveform morphology increases at higher impedance values. Although

it is not of our primary interest to study such an association, what is interesting to us is it is highly likely that the sample size required to demonstrate non-inferiority for TTI = 50 Ω waveforms would be large enough for comparison of the waveforms at the other higher impedances in the future design.

Chapter 2. DATA GENERATING PROCEDURE

Although applying a t-test on W_i's was shown to be a valid method in *Section 1.3.3* for a large sample size, the actual test performance still needed to be evaluated in a finite sample size using Monte-Carlo simulation. The challenge of simulation in this case is how to simulate data with a three-level correlated structure and a binary outcome like the one used in this study. In this chapter, we proposed two data generating procedures: 1) generating the direct binary outcome Y_{ijk} based on *Model (1)* as shown in *Section 2.1*; 2) generating W_i, the transformation of Y_{ijk}, based on *Model (2)* as shown in *Section 2.2*.

Generating Y_{ijk} is ideal, as it keeps the original three-level data structure, but we need to control many nuisance parameters, such as the variances of different random effects. On the other hand, generating W_i is less complicated because we need to control only one nuisance parameter, the residual variance σ_*^2, but we can no longer keep the original data structure. As a result, we did two separate simulation studies to evaluate the test performance in *Chapter 3*, each using a different data generating procedure. Generating Y_{ijk} is used for our primary analysis, whereas generating W_i is our secondary method used for sensitivity analysis.

Each of the two data generating procedures we proposed is naïve, thus there could be potential bias between the estimate of a parameter and its given value used to generate data. For

example, if we simulate a set of $W_i's$ with given mean β, the sample mean of a simulated dataset could be consistently larger or smaller than β. As a result, in this chapter we evaluated the performance and potential limitations of each data generating procedure in advance to the simulation study assessing the test performance.

In general, we first simulated 10,000 datasets for each parameter setting, and then we decided if the data generating procedure was adequate for that parameter setting by comparing the mean estimates of each parameter with its corresponding given value. By doing so, we could find the range of the parameter of interests when the data generating procedure had relative good performance (with no or extremely small bias of the estimated value from its given value), which could be used to further guide our simulation study evaluating the test performance.

In addition, we set the number of block n_J at 8, the same as used in the pilot study. The range of parameters being tested were also based on the pilot study, which is common in real life studies.

2.1 SIMULATE DATA ON Y_{ijk}

2.1.1 *Data Generating Procedure*

The proposed data generating procedure is a modified version of the procedure used in Deke's Brief. In their article, the procedure is used to generate a binary outcome for two treatment groups, which has a baseline risk *p* and an effect size *impact*. They first generate a continuous latent variable y_i^{latent}, and then dichotomize it into a binary variable z_i using the cutoff at $F_{y_i^{latent}}^{-1}(p)$. For the control group, the final binary outcome is just z_i. For the treatment group, an adjustment is further applied to introduce the effect size (Deke, 2014). The detailed data generating procedure is shown in *Appendix A1*.

Our modifications include: 1) random intercepts for subject and block are introduced in addition to residual error to generate the continuous latent variable y_i^{latent}; 2) to introduce the random slope the average effect size is generated from a Normal distribution instead of being a fixed number. Although we didn't assume a certain distribution for random effects b_i, e_i, e_{ij} and residual error e_{ijk}, for simplicity we generated those effects based on a Normal distribution.

2.1.2 *Assessing the Performance of the Data Generating Procedure*

The performance of this data generating procedure was evaluated by fitting a linear mixed model. Although the standard errors for the estimators obtained from a linear mixed model may not be valid due to heteroscedasticity, the unbiased estimates for α and β, as well as the variance for random effects are valid.

Based on the results of the pilot study, the variance of random intercept for a block effect was extremely small (<1e-5), thus σ_2 was set at 1e-5. Due to the binary transformation, the estimates of those variance (or standard deviation) parameters, including σ_1, σ_2, and σ_3, could have potential bias from their given values. One approach to handle this issue is to set σ_1 as a tuning parameter ($\widetilde{\sigma_1}$, range from 0 to 1), and let $\sigma_3 = \sqrt{1 - \sigma_1^2 - \sigma_2^2}$, where σ_2 = 1e-5. Through a simulation study, we obtained a general association between the tuning parameter $\widetilde{\sigma_1}$ and the mean estimates of those variance parameters.

We specified a set of parameters for the simulation study, including α, β, σ_b^2, σ_1^2, σ_2^2 and σ_3^2. Based on the results of the pilot study,

1) α was set at 0.1, 0.3, 0.35 and 0.5;
2) β was set at 0, 0.05 and 0.1;
3) σ_b was set at 0, 0.1 and 0.2;

4) n_I was set at 10, 30 and 50;

5) $\widetilde{\sigma_1}$, was set at 0.1, 0.3, 0.5, 0.7 and 0.9.

2.1.3 *The Performance of the Data Generating Procedure*

From the simulation results, we observed the followings:

i. In general, the tuning parameter $\widetilde{\sigma_1}$ worked very well. Increasing $\widetilde{\sigma_1}$ led to an increased mean $\widehat{\sigma_1}$ and a decreased mean $\widehat{\sigma_3}$, indicating a simultaneous increase in random effects on pig and a decrease of residual error variation. The mean $\widehat{\sigma_2}$ was increased as well, although it was set fixed. However, it was still well controlled at less than 0.03 under all the circumstances in the simulation study and could indicate a low block effect.

ii. When α was set to lower than 0.3, for example when α = 0.1 was set in the simulation study, the performance of this data generating procedure was relatively bad due to the inflation of mean $\hat{\beta}$, and it is summarized in Table 2.1. For instances,

   (a) When σ_b was set at 0.2 (indicating a strong *SVTE*), and β set at 0 (indicating no treatment effect), the mean $\hat{\beta}$ was greater than 0.05.

   (b) When σ_b was set at 0.2, and β set at 0.1, the mean $\hat{\beta}$ was greater than 0.12.

Based on the simulation results, when α was set to greater than 0.3, the absolute difference between the mean $\hat{\beta}$ and β was lowered to around 0.01, which could guarantee the results of the following simulation study evaluating the test performance is relatively reliable with a small sample size. However, if the sample size is large, we would expect a "falsely inflated type I error rate", because the true β in this case is not 0 even when we set it to be 0 in the data generating procedure. When α was 0.5, the absolute difference between the mean $\hat{\beta}$ and

β was further reduced to 0.001 or less so that we should expect our type I error rate roughly at constant.

Table 2.1 Sample results showing limitations of data generating procedure (Y_{ijk})

	σ_b = 0.2	Given α= 0.1, β = 0		Given α= 0.35, β = 0		Given α= 0.5, β = 0	
	tuning parameter ($\widetilde{\sigma_1}$)	Mean $\hat{\alpha}$	Mean $\hat{\beta}$	Mean $\hat{\alpha}$	Mean $\hat{\beta}$	Mean $\hat{\alpha}$	Mean $\hat{\beta}$
n_I = 10	0.1	0.1	0.057	0.35	0.012	0.499	0.003
	0.3	0.101	0.055	0.351	0.013	0.5	-0.002
	0.5	0.1	0.055	0.349	0.01	0.501	0.001
	0.7	0.1	0.057	0.35	0.014	0.5	0
	0.9	0.101	0.056	0.35	0.009	0.5	-0.001
n_I= 30	0.1	0.1	0.054	0.35	0.012	0.5	0
	0.3	0.099	0.056	0.35	0.01	0.499	0.002
	0.5	0.101	0.054	0.35	0.011	0.5	-0.001
	0.7	0.1	0.055	0.351	0.008	0.499	0.004
	0.9	0.1	0.058	0.35	0.008	0.5	0
n_I = 50	0.1	0.1	0.056	0.351	0.009	0.499	0.001
	0.3	0.1	0.055	0.349	0.014	0.5	-0.001
	0.5	0.1	0.054	0.35	0.011	0.5	0.001
	0.7	0.101	0.055	0.351	0.007	0.499	0.001
	0.9	0.1	0.056	0.35	0.009	0.5	-0.001
n_I = 100	0.1	0.1	0.055	0.35	0.011	0.5	0.002
	0.3	0.1	0.055	0.35	0.012	0.5	0.001
	0.5	0.1	0.054	0.35	0.011	0.5	0.001
	0.7	0.1	0.056	0.35	0.011	0.5	0
	0.9	0.1	0.055	0.35	0.011	0.5	0

2.2 SIMULATE DATA ON W_i

2.2.1 *Data Generating Procedure*

By definition, W_i is a discrete random variable with the support {-1, $-\frac{7}{8}$, $-\frac{6}{8}$, …, $\frac{6}{8}$, $\frac{7}{8}$, 1 }. Although we do not assume a certain distribution for W_i, we could first generate a continuous

variable from a common distribution, and then choose a set of cut points to categorize the latent continuous variable into a discrete random variable.

Let $w_1, w_2, \ldots, w_{17}$ be -1, $-\frac{7}{8}, \ldots, 1$, respectively.

1) Determine the 16 cut points, denoted by c_i's, where i = 1, 2, …, 16. For a given common distribution *D* (for example, Normal distribution) with mean 0 and variance σ_*^2, we determine the $c_i's$ by solving $\Phi_D(w_{j+1}) - \Phi_D(c_i) = \Phi_D(c_i) - \Phi_D(w_j)$ for each pair of i and j, where Φ_D is the cumulative distribution function of *D*, and c_i is the cut point between w_j and w_{j+1}. By doing so, if we generate a random variable *d*'s ~ *D* (0, σ_*^2), we would have $P(d = w_j) = P(d = w_{j+1})$ for $d \in (w_j, w_{j+1})$.

2) Generate a latent continuous variable *Y* of size n_I from distribution *D*', a shifted distribution of *D* with mean β, i.e., *Y* ~ *D* (β, σ_*^2).

3) Categorize *Y* into *W* using the set of cut points $c_i's$ determined in step 1). The resulting W_i's should approximately have mean β and variance σ_*^2.

2.2.2 *Assessing the Performance of the Data Generating Procedure*

Four common distributions were selected as the distribution *D's* as described in *Section 2.2.1*, including Normal distribution, Laplace distribution and two t-distributions each with degree of freedom of 5 and 20.

Since the cut points were selected based on a centered distribution *D* (mean at 0), it is expected that there could be some bias when β is not 0. Also, the categorizing step, step 3) in *Section 2.2.1*, may not preserve the value of σ_*^2.

In the pilot study, the estimated $\hat{\beta}$ and $\widehat{\sigma_*}$ were 0.1 and 0.23, respectively. Thus, β = 0, 0.1 and 0.2, and σ_* = 0.15, 0.2 and 0.3 were used as different parameter settings in the simulation

study to test the performance of this data generation procedure. Under each circumstance, 10,000 datasets were simulated, and the mean estimates of β and σ_*^2 were computed.

2.2.3 *The Performance of the Data Generating Procedure*

The results are summarized in Table 2.2. It was observed that under all four distributions the mean $\hat{\beta}$ was generally inflated compared to the given β when β is set not equal to 0, and such a bias got larger when the given β is larger. This observed bias is expected as we mentioned above.

However, such a bias would be reduced with larger σ_*. For example, when β was set at 0.2 and σ_* set larger than 0.2, the absolute difference between the mean $\hat{\beta}$ and β would be controlled to less than 0.01. In addition, the mean $\widehat{\sigma_*}$ is close to the corresponding given σ_*, although it is not of primary interest. As a result, although our data generating procedure had some limitations, it should work well when we set β to smaller than 0.2 and σ_* greater than 0.2, which covers the estimated parameters from the pilot study as well as many cases in real life.

Table 2.2 Performance of data generating procedure (W_i) comparing the mean estimates of β and σ_* with their corresponding given values under different circumstances

(1) Based on Normal distribution

		Given σ_*					
	Given β	0.15		0.2		0.3	
		Mean $\hat{\beta}$	Mean $\widehat{\sigma_*}$	Mean $\hat{\beta}$	Mean $\widehat{\sigma_*}$	Mean $\hat{\beta}$	Mean $\widehat{\sigma_*}$
n_I = 10	0	-0.0003	0.1623	-0.0007	0.2074	-0.0007	0.3006
	0.1	0.1079	0.1621	0.1042	0.2071	0.1012	0.3005
	0.2	0.2155	0.1611	0.209	0.2066	0.2031	0.2998
n_I = 25	0	-0.0002	0.1651	-0.0003	0.2109	-0.0003	0.3057
	0.1	0.1081	0.1648	0.1045	0.2107	0.1017	0.3056
	0.2	0.2158	0.1638	0.2093	0.2103	0.2035	0.3048

	0	0.0001	0.1659	0	0.2121	0.0001	0.3075
n_I = 50	0.1	0.1083	0.1656	0.1048	0.2119	0.1021	0.3073
	0.2	0.2161	0.1646	0.2096	0.2115	0.204	0.3065
	0	0	0.1663	-0.0001	0.2124	-0.0001	0.308
n_I = 100	0.1	0.1083	0.1659	0.1047	0.2123	0.102	0.3078
	0.2	0.216	0.1649	0.2094	0.212	0.2038	0.307

(2) Based on Laplace distribution

	Given β	Given σ_* 0.15		0.2		0.3	
		Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$	Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$	Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$
	0	0.0001	0.1591	-0.0002	0.2022	0	0.2879
n_I = 10	0.1	0.1111	0.1547	0.1073	0.1999	0.1028	0.2867
	0.2	0.2158	0.1506	0.211	0.1968	0.204	0.284
	0	0.0004	0.1635	0.0004	0.2078	0.0007	0.2955
n_I = 25	0.1	0.1115	0.1592	0.1079	0.2059	0.1035	0.2942
	0.2	0.2162	0.1553	0.2114	0.2028	0.2047	0.2916
	0	-0.0002	0.1651	-0.0002	0.2099	-0.0003	0.2983
n_I = 50	0.1	0.1111	0.161	0.1072	0.208	0.1025	0.297
	0.2	0.2157	0.1572	0.2108	0.205	0.2037	0.2942
	0	-0.0001	0.166	-0.0001	0.2112	0	0.2998
n_I = 100	0.1	0.1112	0.162	0.1072	0.2093	0.1028	0.2987
	0.2	0.2158	0.1583	0.2109	0.2063	0.204	0.296

(3) Based on t-distribution with df = 5

	Given β	Given σ_* 0.15		0.2		0.3	
		Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$	Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$	Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$
	0	0	0.1581	-0.0001	0.2003	-0.0002	0.2858
n_I = 10	0.1	0.1094	0.1561	0.105	0.1993	0.1015	0.2853
	0.2	0.2157	0.152	0.209	0.1967	0.203	0.2835
	0	0.0001	0.1621	0.0002	0.2053	0.0002	0.2925
n_I = 25	0.1	0.1096	0.1602	0.1054	0.2043	0.1019	0.2919
	0.2	0.216	0.1559	0.2093	0.2018	0.2032	0.29
	0	0	0.1637	-0.0002	0.2075	-0.0002	0.2952

n_I = 50	0.1	0.1095	0.1619	0.1051	0.2064	0.1015	0.2946
	0.2	0.2158	0.1577	0.209	0.2039	0.2028	0.2926
	0	0	0.1644	0	0.2084	0	0.2964
n_I = 100	0.1	0.1096	0.1627	0.1052	0.2074	0.1017	0.2958
	0.2	0.2159	0.1585	0.2092	0.2049	0.2031	0.2938

(4) Based on t-distribution with df = 20

	Given β	Given σ_* 0.15		0.2		0.3	
		Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$	Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$	Mean $\hat{\beta}$	Mean $\hat{\sigma}_*$
	0	-0.0002	0.1612	-0.0003	0.2063	0	0.2978
n_I = 10	0.1	0.1083	0.1605	0.105	0.2061	0.1019	0.2974
	0.2	0.2157	0.1589	0.2096	0.2046	0.2036	0.2963
	0	0.0001	0.1642	0.0001	0.21	0.0003	0.3032
n_I = 25	0.1	0.1085	0.1636	0.1052	0.2097	0.1022	0.3028
	0.2	0.2159	0.1617	0.2099	0.2085	0.2038	0.3017
	0	0	0.1656	0	0.2117	0	0.3056
n_I = 50	0.1	0.1083	0.1649	0.105	0.2114	0.1019	0.3053
	0.2	0.2158	0.163	0.2096	0.2102	0.2035	0.3039
	0	-0.0001	0.166	-0.0001	0.2122	-0.0002	0.3064
n_I = 100	0.1	0.1083	0.1654	0.1049	0.2119	0.1017	0.306
	0.2	0.2157	0.1635	0.2094	0.2107	0.2033	0.3047

Chapter 3. SIMULATION STUDY

After assessing the performance of the two data generating procedures in *Chapter 2*, we were able to apply some formal simulation studies to evaluate the performance of the t-test. As mentioned before, we did two separate simulation studies based on different data generating procedures. As for our primary analysis, we generated data directly on Y_{ijk}, as shown in *Section 3.1*. As for our secondary analysis, we generated data on W_i, the transformation of Y_{ijk}, as shown in *Section 3.2*.

In general, we simulated 10,000 datasets for each parameter setting using each data generating method. The test performance was evaluated by the corresponding type I error rate and power of the test.

We set the number of blocks n_J at 8, the same as used in the pilot study. The range of parameters being tested were decided based on the results in *Chapter 2*.

3.1 SIMULATE DATA ON Y_{ijk}

3.1.1 *Parameter Setting for Simulation Study*

Based on *Model (1)*, six parameters, α, β, σ_b, σ_1, σ_2, and σ_3, were to be set in the simulation study. However, as shown in *Section 2.1*, we could fix σ_2 at 1e-5, and use the tuning parameter $\widetilde{\sigma_1}$ to change the estimates of σ_1 and σ_3.

According to the results in *Section 2.1.3*, the performance of the data generating procedure was generally good when α was set greater than 0.3. Thus, we set α = 0.3, 0.4, and 0.5, β = 0, 0.05, 0.1 and 0.15, $\widetilde{\sigma_1}$ = 0.5, 0.7 and 0.9 in our simulation study. In addition, we set σ_b = 0 to indicate no or a small *SVTE*, and 0.2 to indicate a strong *SVTE*. The sample size n_I was set at 10, 20, 25, 30, 50 and 100. The simulation results are summarized in the following section.

3.1.2 *Simulation Results*

Type I error rate and power were both evaluated for the performance of the t-test. The results are shown in Figure 3.1 through Figure 3.9. Among those, Figure 3.1 and Figure 3.2 showed type I error rate and power of the test, respectively, when α = 0.3 and σ_b = 0 (indicating no or small *SVTE*); Figure 3.3 and Figure 3.4 showed the test performance when α = 0.3 and σ_b = 0.2 (indicating strong *SVTE*); Figure 3.5 and Figure 3.6 showed the test performance when α = 0.4

and σ_b = 0; Figure 3.7 and Figure 3.8 showed the test performance when α = 0.4 and σ_b = 0.2. The results for α = 0.5 are shown in *Appendix A2*.

Type I error rate was generally constant at 0.05 when there was no or a small *SVTE*, but around 0.04 when there was a strong *SVTE*. Thus, the test was valid for a finite sample size and sometimes could be slightly conservative. However, we could observe a scenario of "falsely inflated type I error rate" due to the limitation of data generating procedure as shown in *Section 2.1.3*, that under some certain circumstances the true β would slightly vary from 0 when it was set to be 0. For instance, such a scenario is shown in Figure 3.3 when α = 0.3 and with a strong *SVTE*, as the mean $\hat{\beta}$ of the simulated data was not 0 but in fact greater than 0.01. But, such a scenario was not observed under other circumstances when the mean $\hat{\beta}$ of simulated data was extremely close to 0 (around 0.001). When there was no or little *SVTE* in the model, and the effect size to be detected was moderate (β = 0.15), the power of the test was relatively large at a small sample size (n_I = 20). However, when a stronger *SVTE* was included in the model, the power was reduced. In addition, compared to the effects of effect size (β) and *SVTE* (σ_b), the effects of baseline successful rate (α) and random effects of subject and block (σ_1 and σ_2, respectively) had much smaller influence on the power of the test.

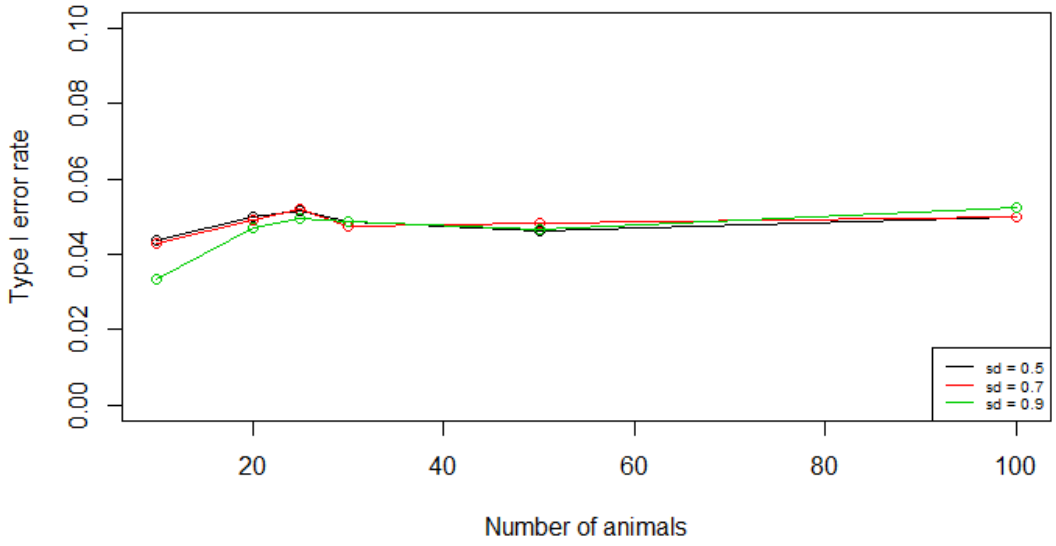

Figure 3.1 Type I error rate of t-test through simulation (Y_{ijk}) without SVTE when α = 0.3

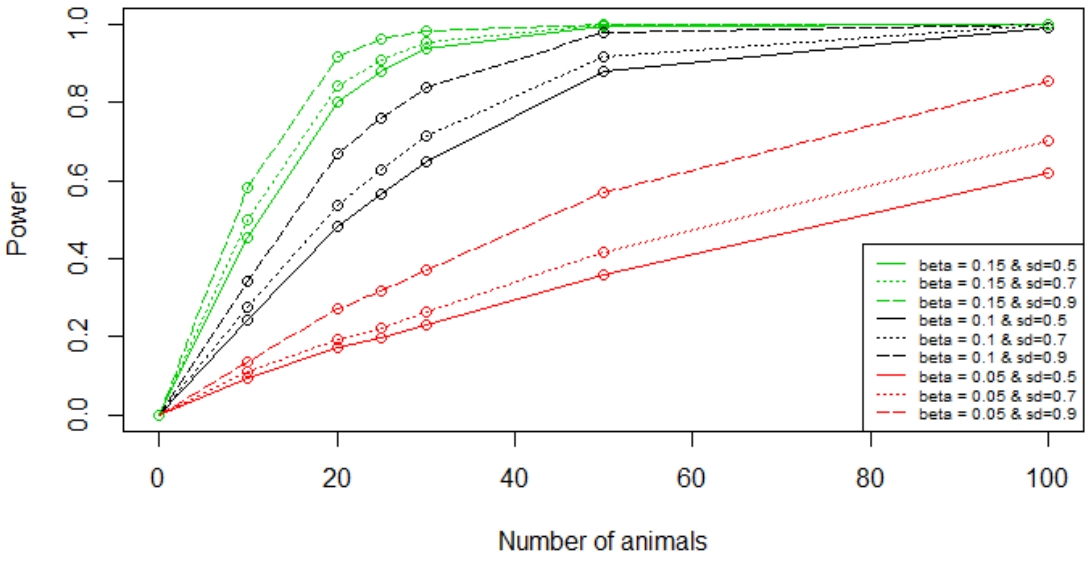

Figure 3.2 Power of t-test through simulation (Y_{ijk}) without SVTE when α = 0.3

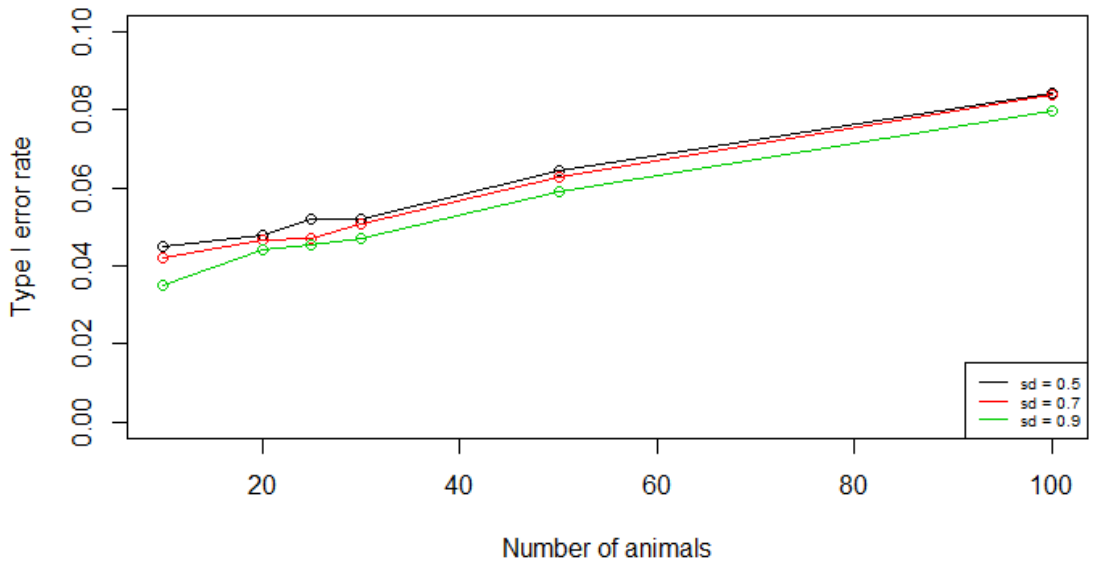

Figure 3.3 Type I error rate of t-test through simulation (Y_{ijk}) with SVTE when α = 0.3

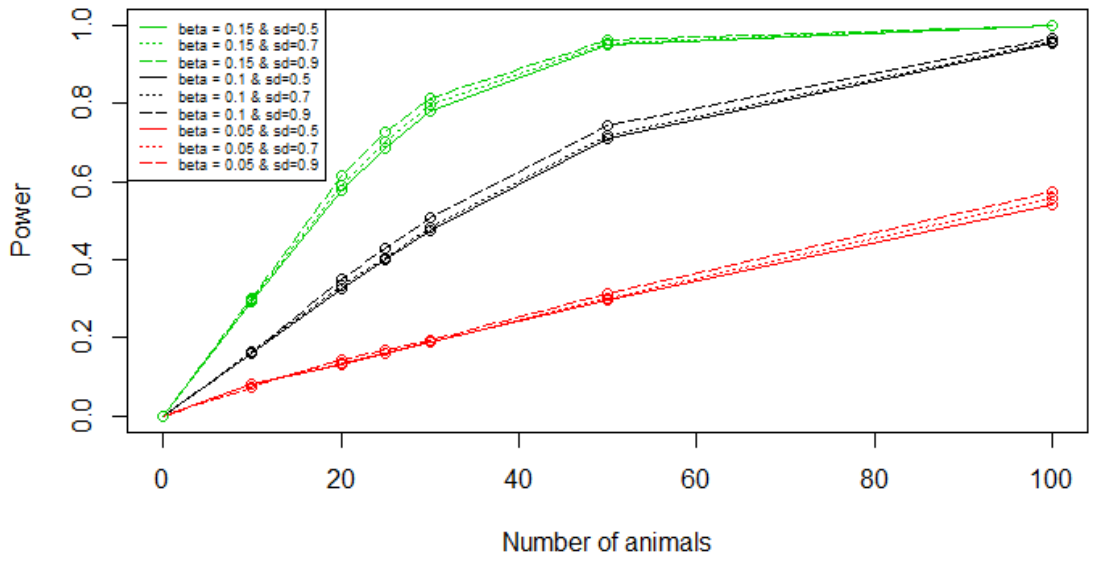

Figure 3.4 Power of t-test through simulation (Y_{ijk}) with SVTE when α = 0.3

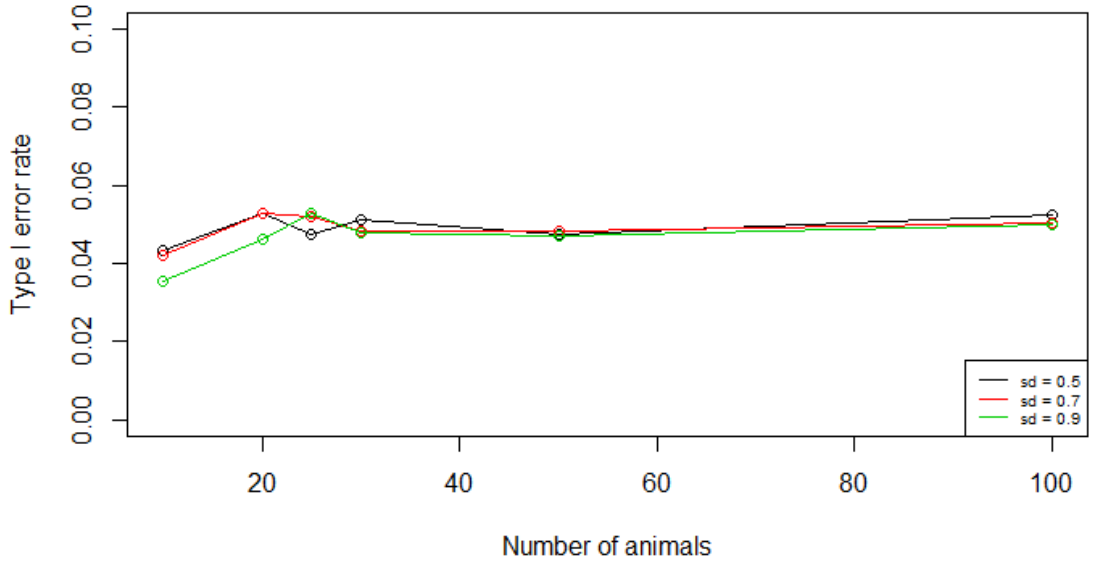

Figure 3.5 Type I error rate of t-test through simulation (Y_{ijk}) without SVTE when α = 0.4

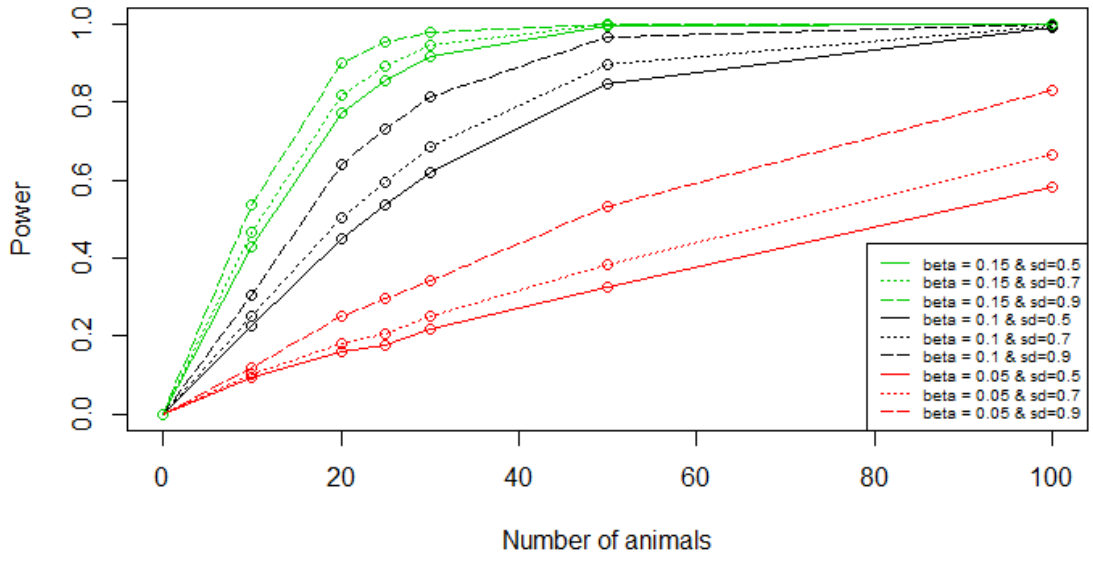

Figure 3.6 Power of t-test through simulation (Y_{ijk}) without SVTE when α = 0.4

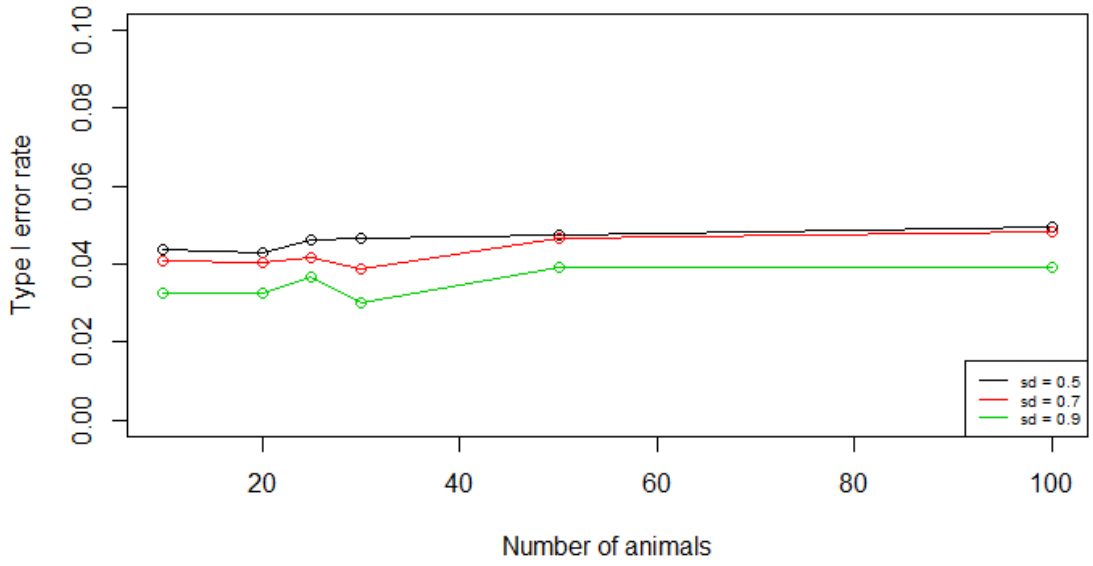

Figure 3.7 Type I error rate of t-test through simulation (Y_{ijk}) with SVTE when α = 0.4

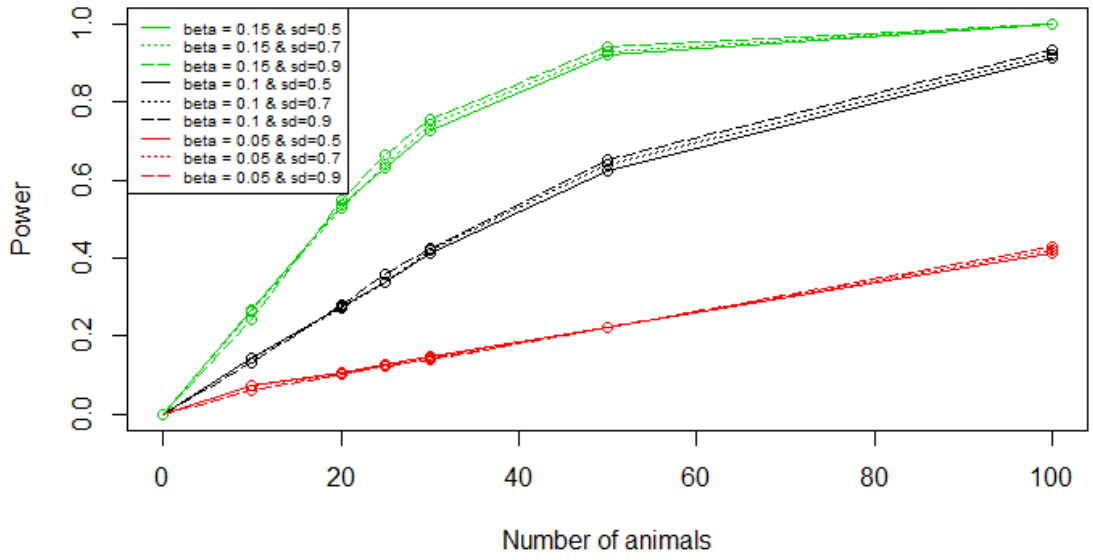

Figure 3.8 Power of t-test through simulation (Y_{ijk}) with SVTE when α = 0.4

3.2 SIMULATE DATA ON W_i

3.2.1 *Parameter Setting for Simulation Study*

Based on *Model (2)*, two parameters, β and σ_*, were to be set in the simulation study. Again, four different distributions were used in the data generating procedure, including a Normal distribution, a Laplace distribution and two t-distributions each with 5 and 20 degrees of freedom.

According to the results in *Section 2.2.3*, the performance of the data generating procedure was good when β was set to less than 0.2, and σ_* set to greater than 0.2. Thus, we set β = 0, 0.05, 0.1 and 0.15, σ_* = 0.2, 0.23 (results from the pilot study) and 0.3 in our simulation study. The sample size n_I was set at 10, 20, 25, 30, 50 and 100. The simulation results are summarized in the following Section.

3.2.2 *Simulation Results*

Type I error rate and power were both evaluated for the t-test. The results are shown in Figure 3.9 through Figure 3.16. Among those, Figure 3.9 and Figure 3.10 showed type I error rate and power of the test, respectively, when data were generated based on a Normal distribution; Figure 3.11 and Figure 3.12 showed the test performance when data were generated based on a Laplace distribution; Figure 3.13 and Figure 3.14 showed the test performance when data were generated based on a t-distribution with 5 degrees of freedom; Figure 3.15 and Figure 3.16 showed the test performance when data were generated based on a t-distribution with 20 degrees of freedom.

Type I error rate was generally constant at 0.05 under each of the tested circumstances, so the test was valid for a finite sample size as expected. The power of the test was relatively large

at a small sample size (n_I = 20) when the effect size to be detected was moderate (β = 0.15), which was consistent with the results in Section 3.1.2. However, if the effect size was extremely small (β = 0.05), the test was not powerful unless with a huge sample size. The test performance was similar among four different given distributions, which made the test potentially robust to different distributions of residual error.

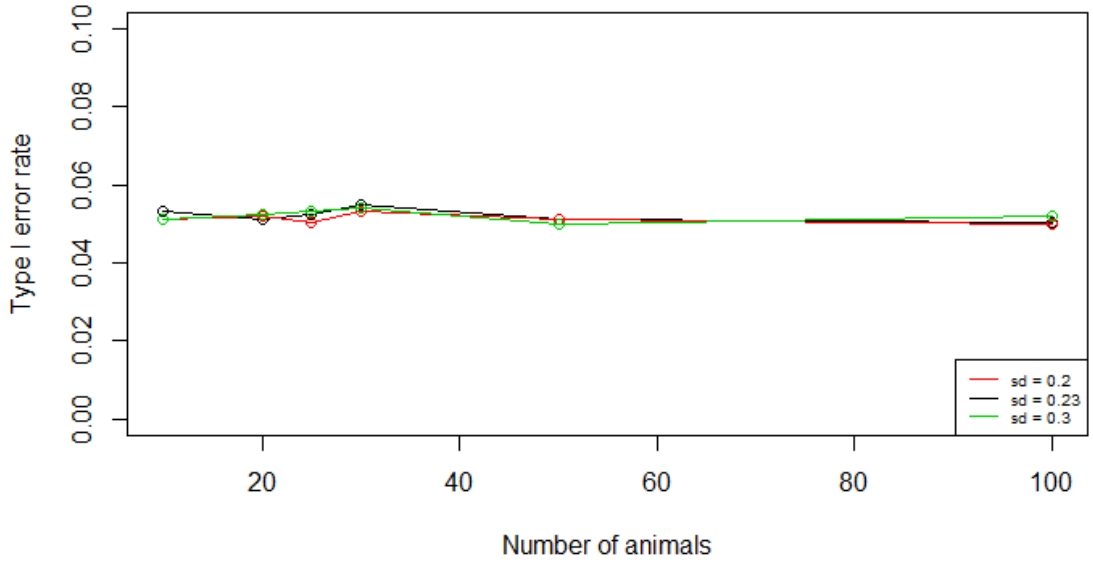

Figure 3.9 Type I error rate of t-test through simulation (W_i) based on a Normal distribution

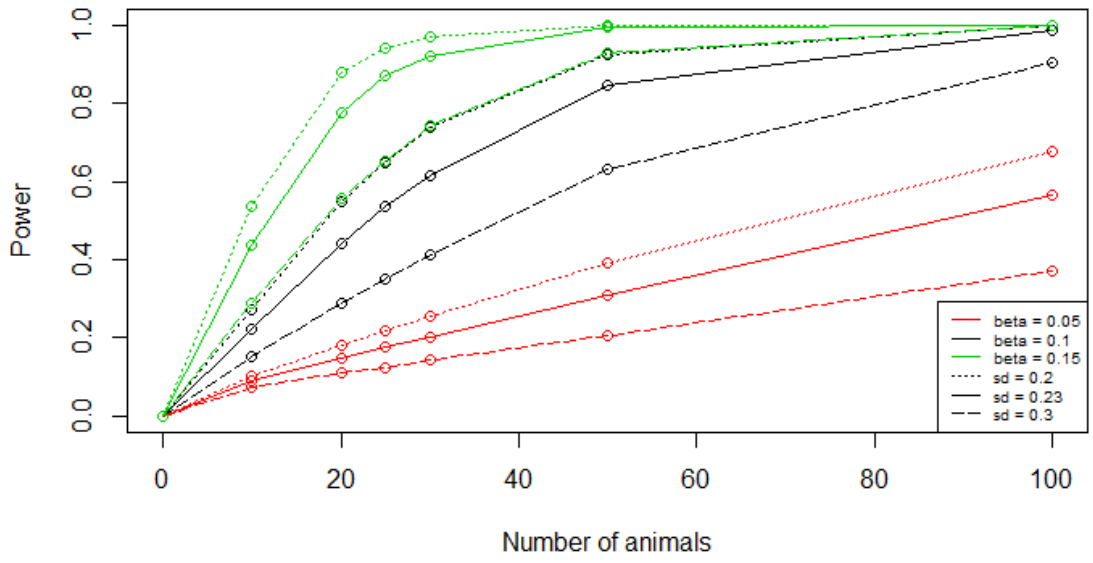

Figure 3.10 Power of t-test through simulation (W_i) based on a Normal distribution

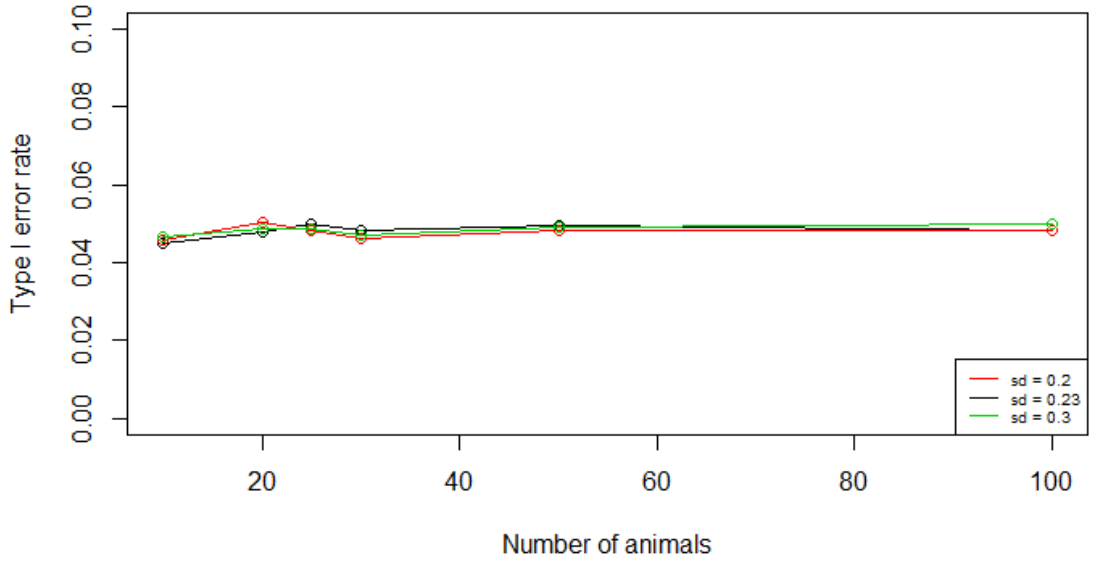

Figure 3.11 Type I error rate of t-test through simulation (W_i) based on a Laplace distribution

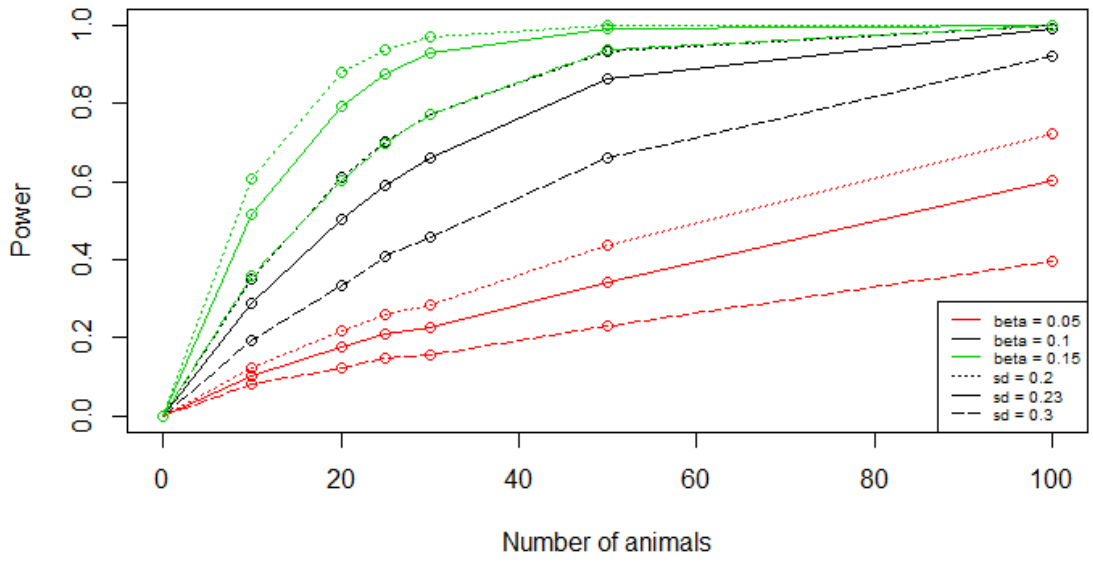

Figure 3.12 Power of t-test through simulation (W_i) based on a Laplace distribution

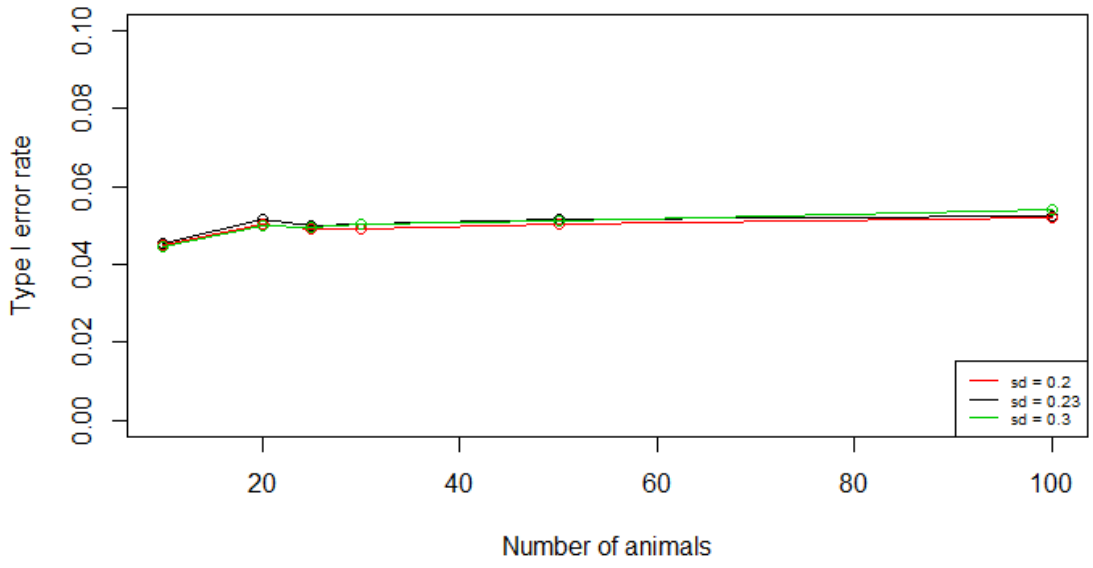

Figure 3.13 Type I error rate of t-test through simulation (W_i) based on a t-distribution (df = 5)

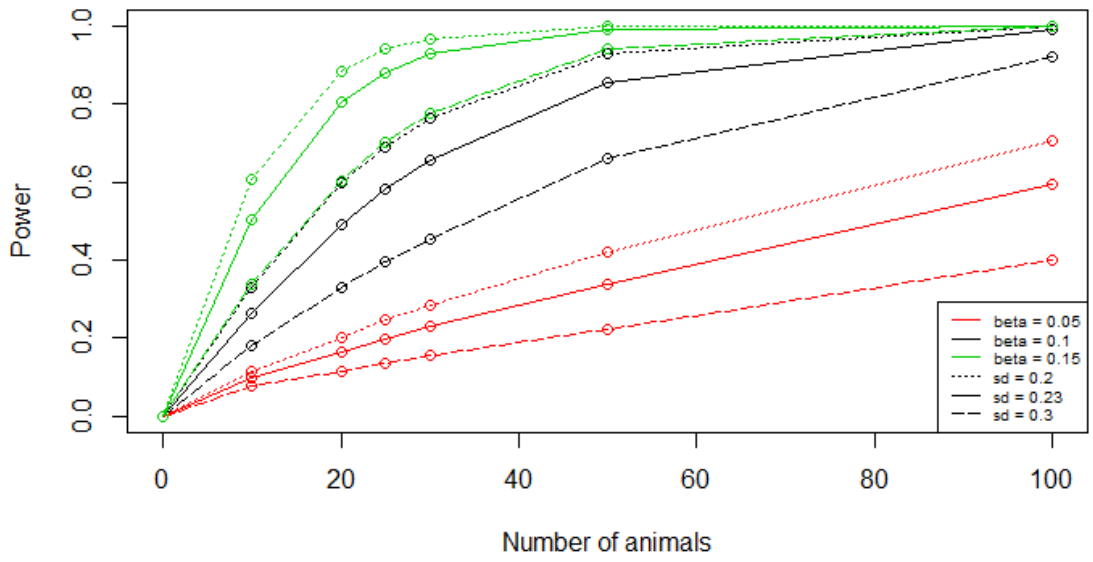

Figure 3.14 Power of t-test through simulation (W_i) based on a t-distribution (df = 5)

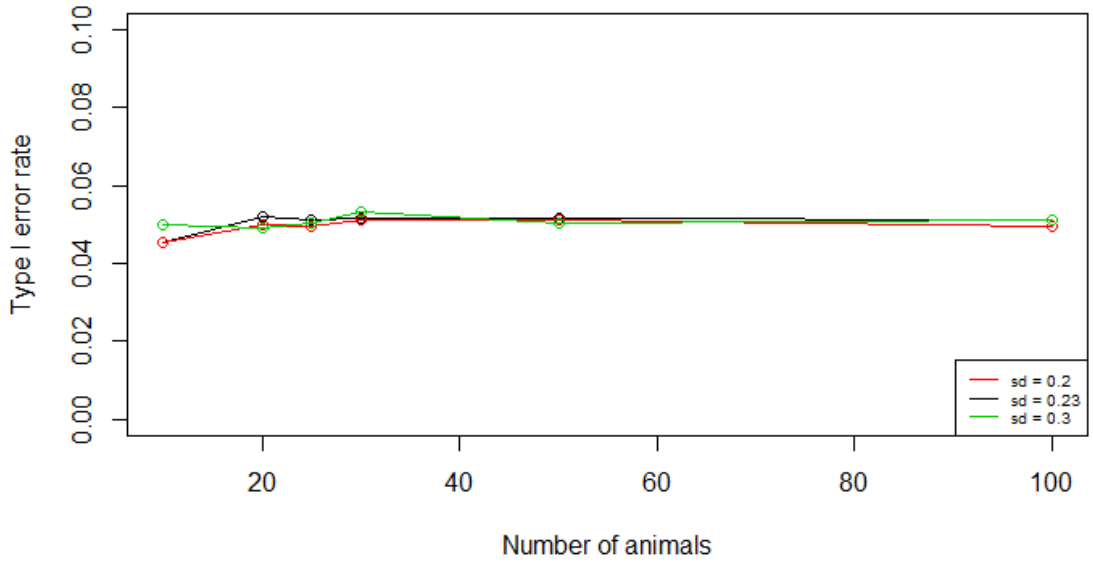

Figure 3.15 Type I error rate of t-test through simulation (W_i) based on a t-distribution (df = 20)

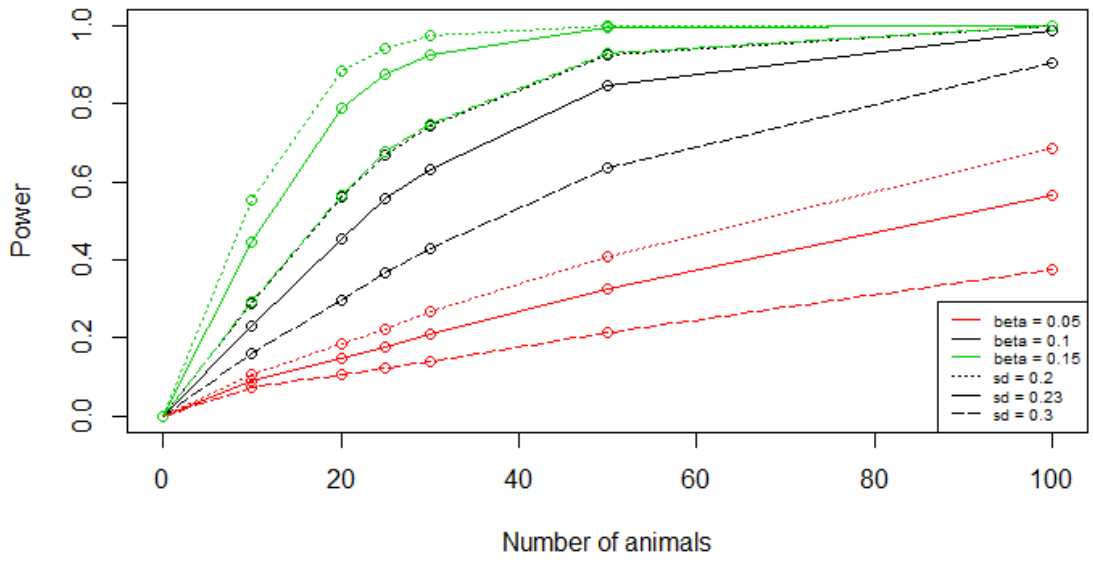

Figure 3.16 Power of t-test through simulation (W_i) based on a t-distribution (df = 20)

Chapter 4. DISCUSSION AND LIMITATIONS

Based on the simulation results in *Chapter 3*, it is reasonable to conclude that applying t-test on W_i, the transformation of the original binary outcome Y_{ijk}, as described in *Section 1.3.3*, is a valid and reliable statistical method to study the risk difference in a three-level paired design with a binary outcome, as introduced in *Section 1.1*.

The two data generating methods, simulating data on Y_{ijk} based on *Model (1)*, and W_i based on *Model (2)* are different approaches to evaluate the test performance, and each of them has its own advantages. Simulating data on Y_{ijk} is a complicated procedure (having more nuisance parameters), but it allows us to study the effects of baseline risk and *SVTE* on the test performance. On the other hand, simulating data on W_i is an easier procedure, but we are unable to study the effects of baseline risk or *SVTE* on the test performance. However, the results of

those two methods are expected to be consistent, as *Model (2)* was derived from *Model (1)*. In fact, they are indeed consistent in our study, which provides some strong evidence leading to our conclusion.

The power of the test to detect a risk difference of 0.15 is around 80% for a sample size of 20 subjects when there is a small *SVTE*. However, when *SVTE* was strong, the power would drop to 60%. Thus, our proposed method is more efficient to deal with cases with a small *SVTE*, for example in the current study, as suggested by the pilot study.

Our proposed method is very useful in non-inferiority trials when the test drug (or device) is in fact better than the active control, as a smaller sample size is required to demonstrate non-inferiority compared to demonstrating superiority in this case (CPMP, 2001). For example, the current study is designed as a non-inferiority trial with a non-inferiority margin of -10%. The sample size determination based on the estimates from the pilot study can be achieved through simulation. Again, two simulation procedures were applied, and the results are shown in *Appendix A3*. Based on the results, around 10 subjects are required to demonstrate non-inferiority if the true average effect size is 0.15, and around 25 subjects are required if the true average effect size is only 0.05, even with a relatively strong *SVTE*. Although no closed form is derived for sample size calculation using our proposed method, simulation is a useful alternative method and easy to implement.

This study has some limitations, which are from the two data generating procedures. As mentioned in *Section 2.1.3* and *Section 2.2.3*, both procedures have their limitations. In the first procedure used to simulate Y_{ijk}, the data was generated well only when α was set to greater than 0.3. In the second procedure used to simulate W_i, the data was generated well only when β was set to smaller than 0.2. Although both methods are good enough to cover a lot of circumstances

in real life, for example the pilot study, and such limitations shouldn't be expected to affect our conclusion, we still could have trouble when doing sample size calculation through simulation. Thus, further study on modifying the data generating procedure would be helpful.

BIBLIOGRAPHY

Cox, D. (1958). The regression analysis of binary sequences (with discussion). *J Roy Stat Soc B,* 20:215-242.

CPMP. (2001). Points to consider on switching between superiority and non-inferiority. *Br J Clin Pharmacol*, 52(3): 223–228.

Deke, J. (2014). *Using the Linear Probability Model to Estimate Impacts.* Mathematica Policy Research.

Fisher, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7 (2): 179-188.

Greenland, S., & Thomas, D. (1982). On the need for the rare disease assumption in case-control studies. *Am. J. Epidemiol*, 116(3): 547-53.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). THE IMPORTANCE OF THE NORMALITY ASSUMPTION IN LARGE PUBLIC HEALTH DATA SETS. *Annual Review of Public Health*, Vol.231(1), p.151-169.

Petersen, M. R., & Deddens, J. A. (2008). A comparison of two methods for estimating prevalence ratios. *BMC Medical Research Methodology*, 8: 9.

Zhang J, Y. K. (1998). What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA,* 280:1690-1691.

APPENDIX A

A1. *DATA GENERATING PROCEDURE IN DEKE'S BRIEF*

$$
\begin{aligned}
& e \sim N(0,1) \\
(1)\quad & x \sim N(0,1) \\
& y_i^{latent} = -x_i + e_i
\end{aligned}
$$

$$(2)\quad T \sim Bernouli(0.5)$$

$$(3)\quad z_i = \begin{cases} 1 \text{ if } y_i^{latent} < F^{-1}_{y^{latent}}(p) \\ 0 \text{ if } y_i^{latent} \geq F^{-1}_{y^{latent}}(p) \end{cases}$$

$$(4)\quad y_i = \begin{cases} z_i \text{ if } T_i = 0 \\ 1 \text{ if } T_i = 1,\ impact > 0, \text{ and } z_i = 1 \\ Bernouli\left(impact/(1-p)\right) \text{ if } T_i = 1,\ impact > 0, \text{ and } z_i = 0 \\ 0 \text{ if } T_i = 1,\ impact < 0, \text{ and } z_i = 0 \\ Bernouli\left((p + impact)/p\right) \text{ if } T_i = 1,\ impact < 0, \text{ and } z_i = 1 \end{cases}$$

Figure A.1 Data generating procedure in Deke's Brief

A2. *SIMULATION RESULTS* (Y_{ijk}) *FOR* $\alpha = 0.5$

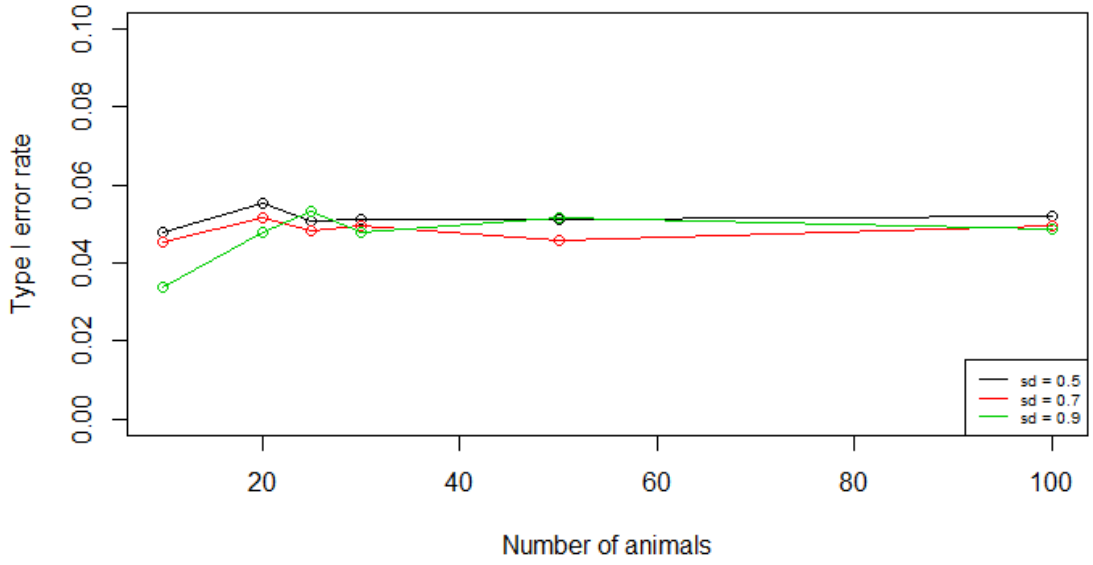

Figure A.2 Type I error rate of t-test through simulation (Y_{ijk}) without SVTE when α = 0.5

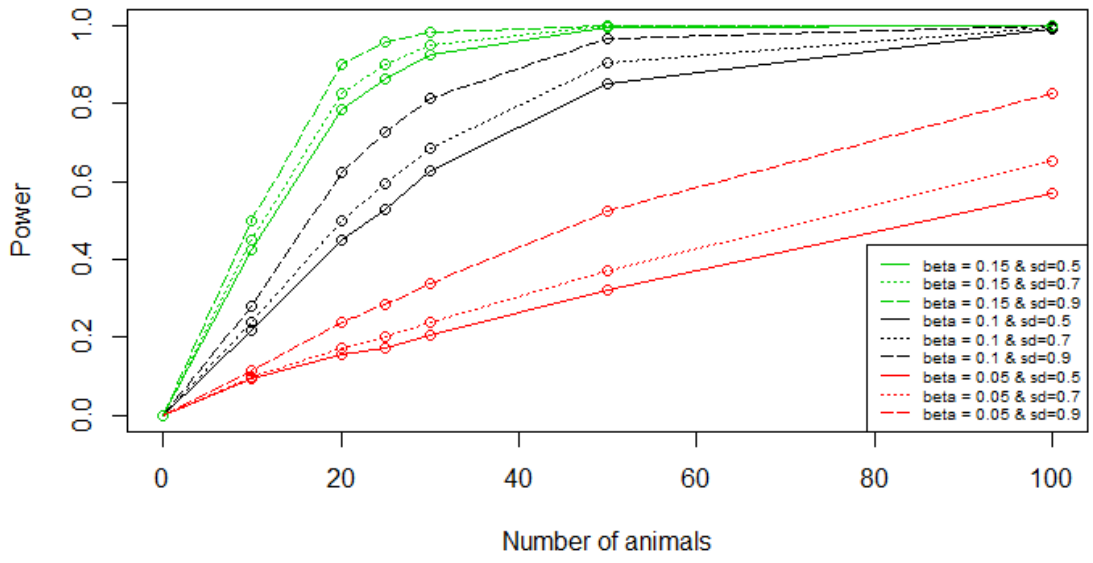

Figure A.3 Power of t-test through simulation (Y_{ijk}) without SVTE when α = 0.5

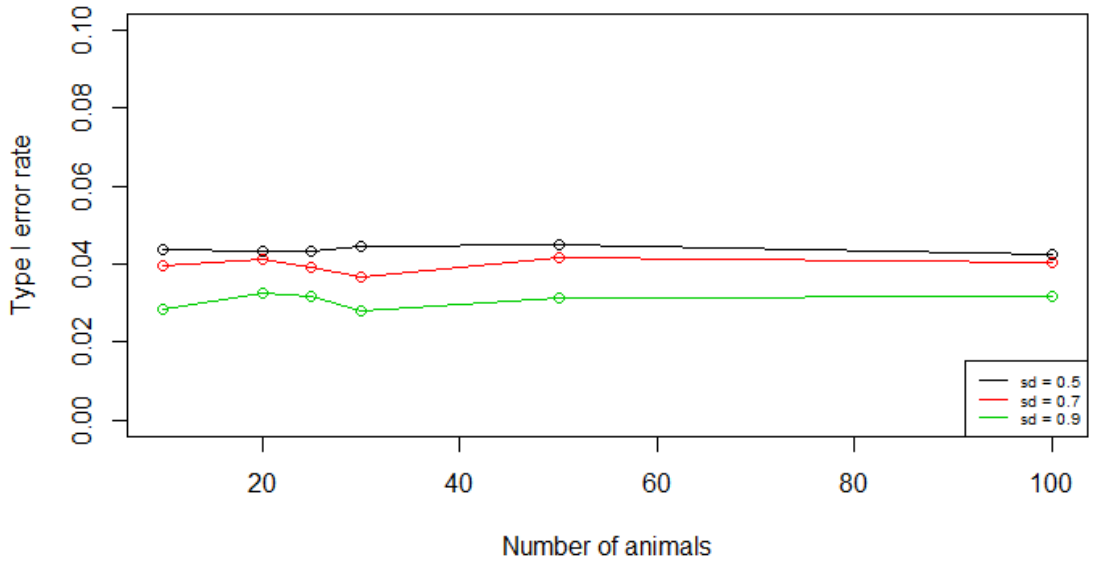

Figure A.4 Type I Error Rate of t-test through Simulation (Y_{ijk}) with SVTE when α = 0.5

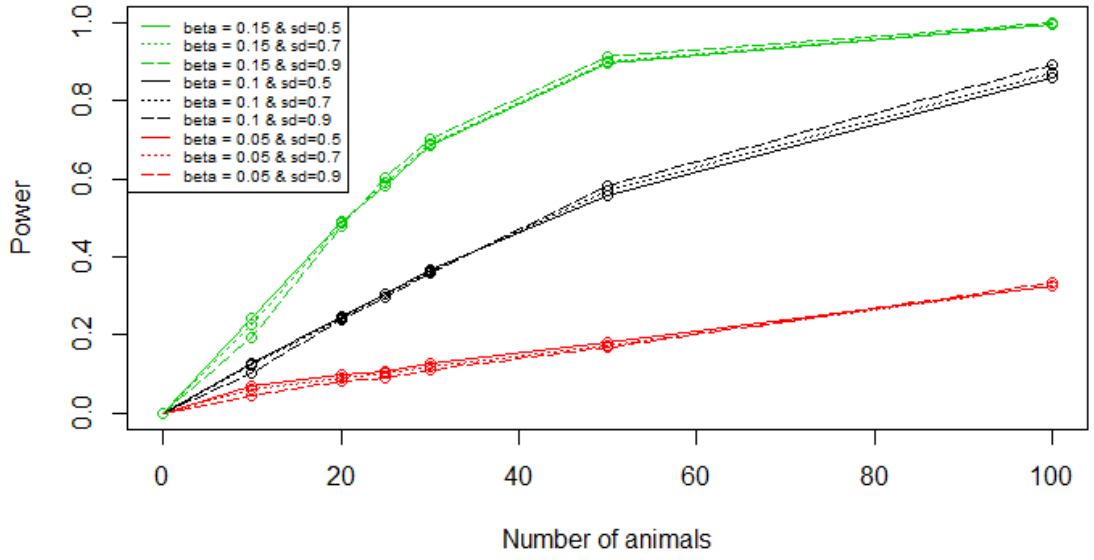

Figure A.5 Power of t-test through Simulation (Y_{ijk}) with SVTE when α = 0.5

A3. *Sample Size Determination for Non-inferiority Design*

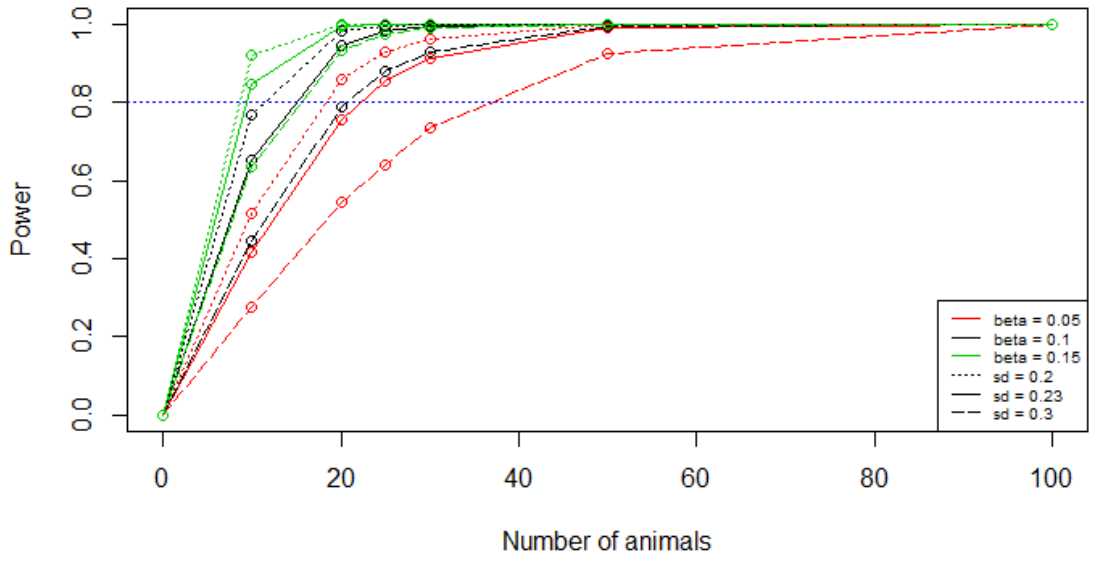

Figure A.6 Sample size calculation through simulation (W_i) based on Normal Distribution

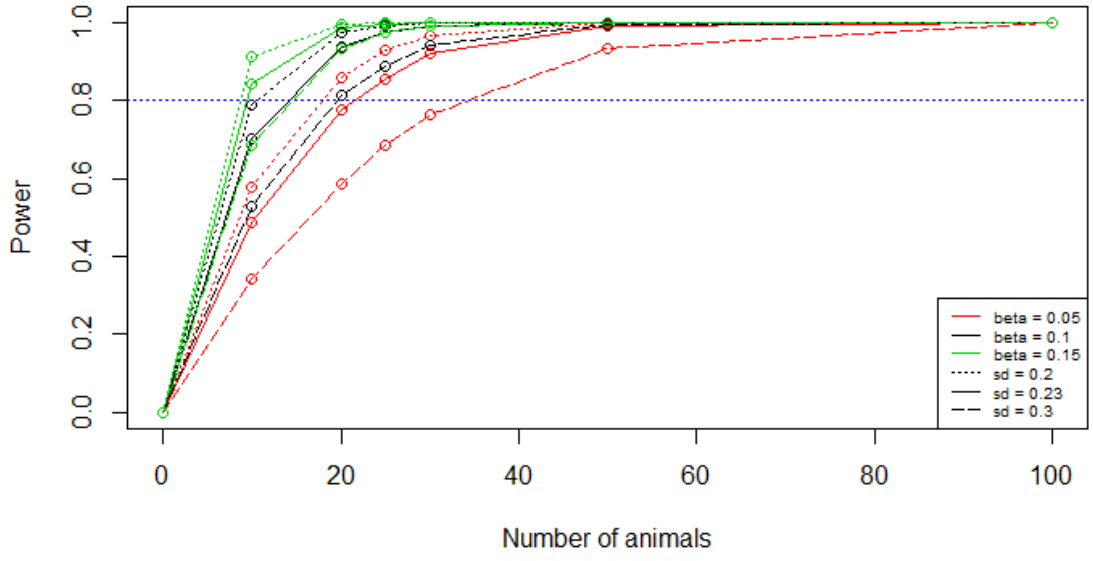

Figure A.7 Sample size calculation through simulation (W_i) based on Laplace Distribution

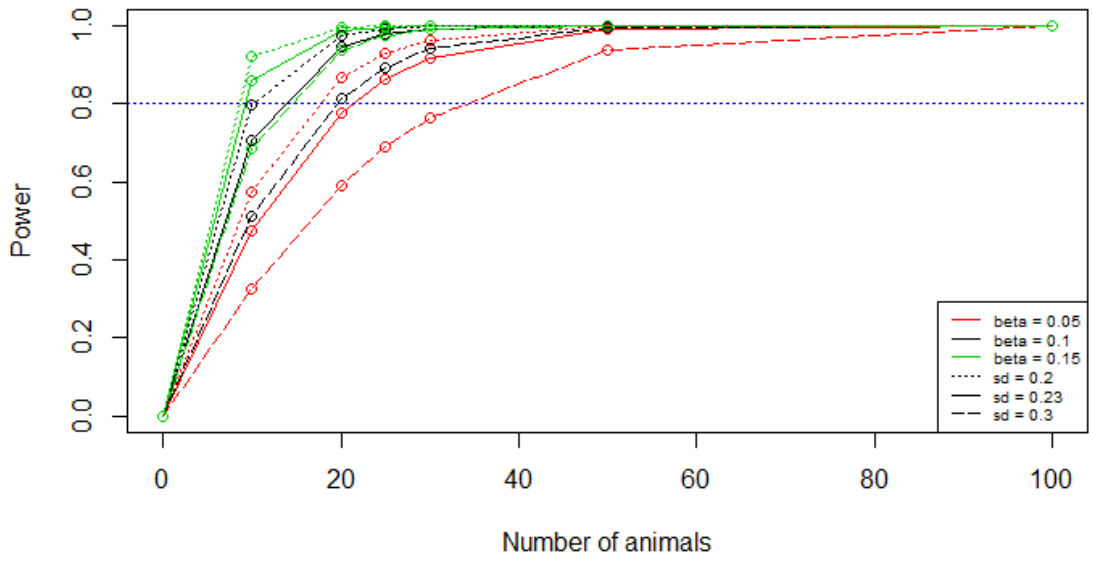

Figure A.8 Sample size calculation through simulation (W_i) based on t-distribution (df = 5)

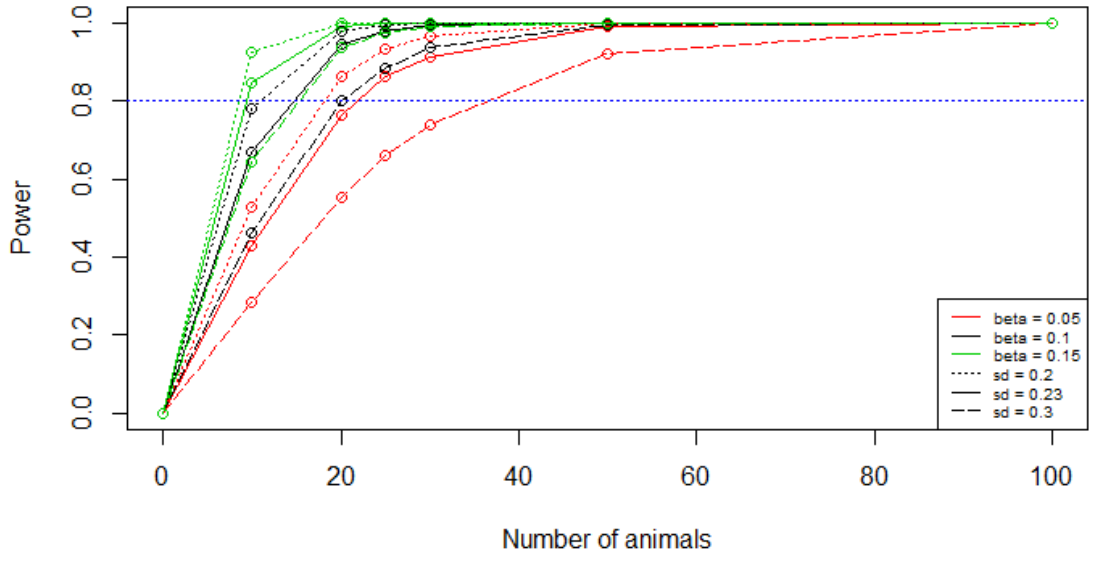

Figure A.9 Sample size calculation through simulation (W_i) based on t-distribution (df = 20)

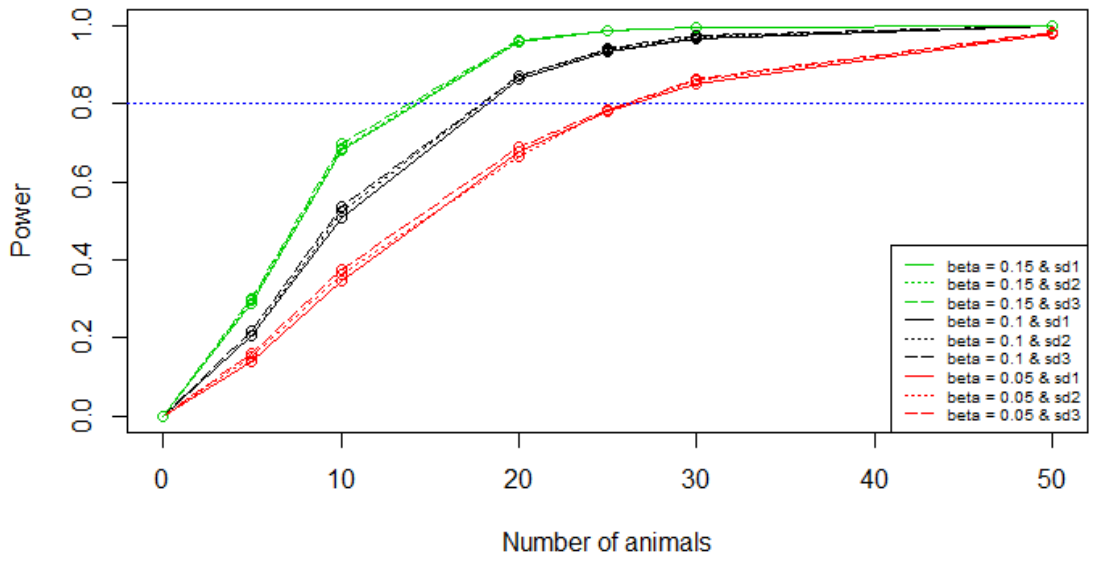

Figure A.10 Sample size calculation through simulation (Y_{ijk}) with SVTE when α = 0.3

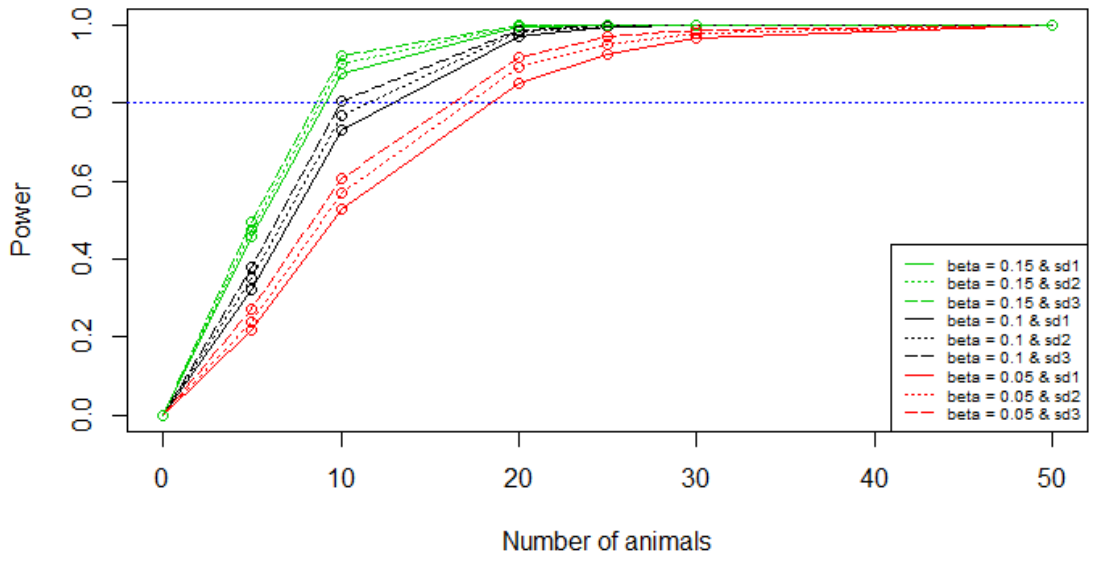

Figure A.11 Sample size calculation through simulation (Y_{ijk}) without SVTE when α = 0.3

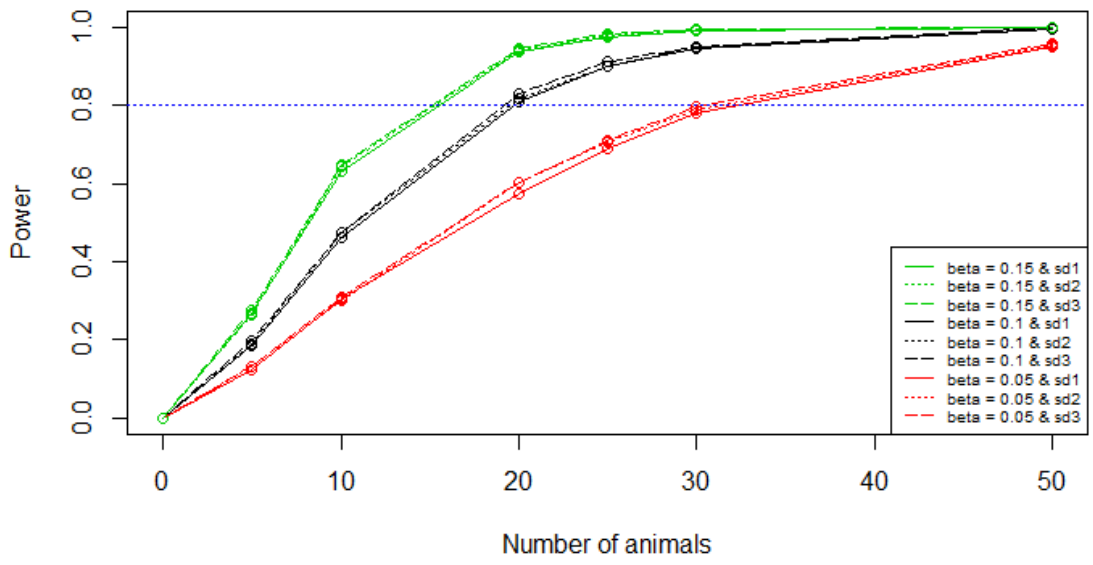

Figure A.12 Sample size calculation through simulation (Y_{ijk}) with SVTE when α = 0.4

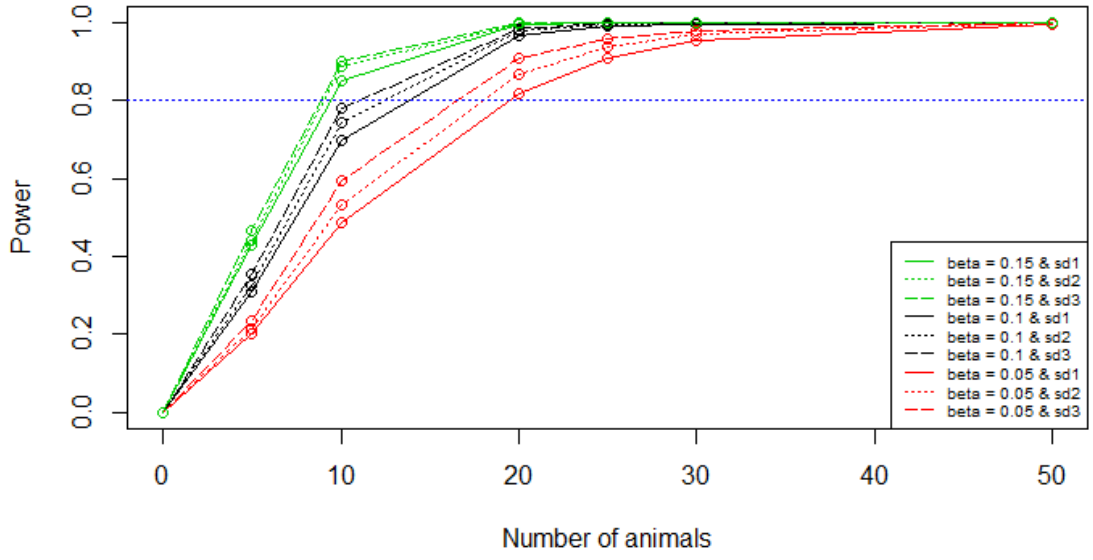

Figure A.13 Sample size calculation through simulation (Y_{ijk}) without SVTE when α = 0.4