# A COMPARISON OF TWO SCALING TECHNIQUES TO REDUCE UNCERTAINTY IN PREDICTIVE MODELS

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Austin Luke Todd

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2020

Fargo, North Dakota

North Dakota State University

Graduate School

**Title**

# A COMPARISON OF TWO SCALING TECHNIQUES TO REDUCE UNCERTAINTY IN PREDICTIVE MODELS.

**By**

Austin Luke Todd

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

**MASTER OF SCIENCE**

SUPERVISORY COMMITTEE:

Rhonda Magel

Chair

Curt Doetkott

Gursimran Walia

Approved:

04/23/2020 — Date

Rhonda Magel — Department Chair

## ABSTRACT

This research examines the use of two scaling techniques to accurately transfer information from small-scale data to large-scale predictions in a handful of nonlinear functions. The two techniques are (1) using random draws from distributions that represent smaller time scales and (2) using a single draw from a distribution representing the mean over all time represented by the model. This research used simulation to create the underlying distributions for the variable and parameters of the chosen functions which were then scaled accordingly. Once scaled, the variable and parameters were plugged into our chosen functions to give an output value. Using simulation, output distributions were created for each combination of scaling technique, underlying distribution, variable bounds, and parameter bounds. These distributions were then compared using a variety of statistical tests, measures, and graphical plots.

## ACKNOWLEDGEMENTS

I would like to thank Curt Doetkott for advising me in developing this thesis and for his guidance of me in my learning during my time at North Dakota State University. He helped facilitate decision making at every step of this process. I would also like to thank Dr. Magel for her critiques and advice in preparing my thesis. I would also like to thank her for advising me throughout my graduate and undergraduate years at NDSU, she has been a key component in my development as a statistician. Thanks, are also in order for Dr. Gursimran Walia, who was on my thesis committee and who's unique perspective helped me make improvements to my thesis.

I would also like to thank my parents for their never-ending support and guidance. Thanks, are also in order for my friends at Neptune and Company for bringing this topic to me and trusting me to complete this analysis on my own. Finally, I would like to thank all the professors and fellow graduate students in the Department of Statistics.

## TABLE OF CONTENTS

ABSTRACT ..... iii

ACKNOWLEDGEMENTS ..... iv

LIST OF FIGURES ..... vii

LIST OF SYMBOLS ..... viii

LIST OF APPENDIX FIGURES ..... ix

CHAPTER 1. INTRODUCTION/BACKGROUND ..... 1

CHAPTER 2. MOTIVATING EXAMPLE ..... 4

2.1. Linear Function. ..... 4

2.1.1. Technique 1 (drawing a new value at each time step) ..... 4

2.1.2. Technique 2 (drawing one value and applying to all time steps) ..... 5

2.2. Nonlinear Function. ..... 7

2.2.1. Technique 1: Annual Sampling ..... 7

2.2.2. Technique 2 (drawing one value and applying to all time steps) ..... 8

CHAPTER 3. METHODOLOGY ..... 11

3.1. Function Selection. ..... 11

3.2. Parameter/ Distribution Selection. ..... 12

3.3. Simulation Procedure. ..... 13

CHAPTER 4. RESULTS ..... 16

4.1. Example Output. ..... 16

4.2. Parameters Bounded Between 0-1 vs Bounded Between 1-10. ..... 19

4.3. Semi-Linear vs Nonlinear Functions. ..... 20

4.4. Normal Distribution. ..... 22

4.5. Lognormal Distribution. ..... 24

4.6. Variable Bounded 0-1 vs Bounded 1-10. ..... 26

4.7. Number of Parameters. 27
CHAPTER 5. CONCLUSION 29
REFERENCES 31
APPENDIX A. EXAMPLE DISTRIBUTION PLOTS 32
APPENDIX B. EXAMPLE Q-Q PLOTS 58
APPENDIX C. EXAMPLE FIGURES OF DESCRIPTIVE STATISTICS 85
APPENDIX D. R CODE 93
D.1. Example R Program 93
D.2. Example R Output Program 95

## LIST OF FIGURES


Figure Page

1. Flowchart of half of all combinations ..... 13
2. Distributions of technique 1 vs technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 16
3. Q-Q plot of technique 1 vs technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 17
4. Descriptive statistics for Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 18
5. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1 (left). Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 0-1 (right) ..... 20
6. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * xb$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 20
7. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 0-1 for semi-linear function (left) and a nonlinear function (right). ..... 21
8. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ..... 23
9. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ..... 24
10. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 25
11. Q-Q plot of technique 1 vs technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 26
12. Variance ratios across number of parameters. ..... 27
13. Mean ratios across number of parameters. ..... 28

## LIST OF SYMBOLS

μ .....(mu) A Greek letter used to represent the population mean.

σ .....(sigma) A Greek letter used to represent the population standard deviation.

## LIST OF APPENDIX FIGURES


Figure — Page

A1. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 32

A2. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 33

A3. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 34

A4. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 35

A5. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 36

A6. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ... 37

A7. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ... 38

A8. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ... 39

A9. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ... 40

A10. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ... 41

A11. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 42

A12. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 43

A13. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 44

A14. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 45

A15. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ... 46

A16. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 47

A17. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 48

A18. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 49

A19. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 50

A20. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 51

A21. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 52

A22. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 53

A23. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 54

A24. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 55

A25. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 56

A26. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 57

B1. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 58

B2. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 59

B3. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 60

B4. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 61

B5. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 62

B6. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. 63

B7. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. 64

B8. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. 65

B9. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. 66

B10. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. 67

B11. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10. 68

B12. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. 69

B13. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. 70

B14. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. 71

B15. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. 72

B16. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. 73

B17. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. 74

B18. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. 75

B19. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. 76

B20. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. 77

B21. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1. 78

B22. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 79

B23. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 80

B24. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 81

B25. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 82

B26. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 83

B27. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 84

C1. Descriptive statistics for Technique 1 vs Technique 2 for y = 1-(1xa) using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 85

C2. Descriptive statistics for Technique 1 vs Technique 2 for y = a*xb using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 85

C3. Descriptive statistics for Technique 1 vs Technique 2 for y = (a*b*x)/(1 + b*x) using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 85

C4. Descriptive statistics for Technique 1 vs Technique 2 for y = 1/(a + b*x + c*x2) using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 86

C5. Descriptive statistics for Technique 1 vs Technique 2 for y = a*exp (-expb-c*x) using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 86

C6. Descriptive statistics for Technique 1 vs Technique 2 for y = a + b*exp (-c*x-d2) using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 86

C7. Descriptive statistics for Technique 1 vs Technique 2 for y = 1/(x + a) using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ..... 86

C8. Descriptive statistics for Technique 1 vs Technique 2 for y = log (a + b*x) using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ..... 87

C9. Descriptive statistics for Technique 1 vs Technique 2 for y = a*(1-exp-b*x)γ using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ..... 87

C10. Descriptive statistics for Technique 1 vs Technique 2 for y = a*exp-b*x + γ*exp (-δ*x) using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ..... 87

C11. Descriptive statistics for Technique 1 vs Technique 2 for $y = a/(1 + \exp(b\text{-}\gamma^{*}x))1/\delta$ using normal distributions, parameters bounded 1-10, and variable bounded 1-10. ..... 87

C12. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(x + a)$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 88

C13. Descriptive statistics for Technique 1 vs Technique 2 for $y = \log(a + b^{*}x)$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 88

C14. Descriptive statistics for Technique 1 vs Technique 2 for $y = a^{*}(1\text{-}\exp\text{-}b^{*}x)\gamma$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 88

C15. Descriptive statistics for Technique 1 vs Technique 2 for $y = a^{*}\exp\text{-}b^{*}x + \gamma^{*}\exp(-\delta^{*}x)$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 89

C16. Descriptive statistics for Technique 1 vs Technique 2 for $y = a/(1 + \exp(b\text{-}\gamma^{*}x))1/\delta$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 89

C17. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(x + a)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 89

C18. Descriptive statistics for Technique 1 vs Technique 2 for $y = \log(a + b^{*}x)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 90

C19. Descriptive statistics for Technique 1 vs Technique 2 for $y = a^{*}(1\text{-}\exp\text{-}b^{*}x)\gamma$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 90

C20. Descriptive statistics for Technique 1 vs Technique 2 for $y = a^{*}\exp\text{-}b^{*}x + \gamma^{*}\exp(-\delta^{*}x)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 90

C21. Descriptive statistics for Technique 1 vs Technique 2 for $y = a/(1 + \exp(b\text{-}\gamma^{*}x))1/\delta$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1. ..... 90

C22. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1\text{-}(1xa)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 91

C23. Descriptive statistics for Technique 1 vs Technique 2 for $y = a^{*}xb$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 91

C24. Descriptive statistics for Technique 1 vs Technique 2 for $y = (a^{*}b^{*}x)/(1 + b^{*}x)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 91

C25. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(a + b*x + \gamma*x2)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 91

C26. Descriptive statistics for Technique 1 vs Technique 2 for $y = a*\exp(-\exp b - \gamma*x)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 92

C27. Descriptive statistics for Technique 1 vs Technique 2 for $y = a + b*\exp(-\gamma*(x-\delta)2)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10. ..... 92

# CHAPTER 1. INTRODUCTION/BACKGROUND

Sample size has always been something that statisticians wrestle with. Larger samples allow researchers to make more precise estimates, but they come at an increased cost. In predictive modeling, researchers use sampled data to build input distributions in which they can make draws from and apply those values to a model that covers larger temporal and/or spatial scales. Any time we want to predict future values we are required to use past data and one of the issues of predictive modeling is that uncertainty can be overestimated if we do not correct for it. This uncertainty occurs because we are using small samples of data to represent much larger spatial or temporal domains. We can use our data to model hundreds and thousands of years into the future, or we can use samples from localized areas to make predictions about the entire region but if we don't scale our data the uncertainty associated with our samples will be applied throughout the model and will result in a greatly overestimated variance in our input distributions.

Scaling is a technique that predictive models use in which measurements are collected and either downscaled or upscaled before they are applied to a model or function. Downscaling occurs in almost all climate models where the data collected by researchers is on a large scale, usually in grids of 200 km to 300 km on the Earth's surface, and they must use downscaling techniques to make predictions and inference at smaller, local scales. Upscaling is essentially the opposite of downscaling. In upscaling the researcher is collecting data on a small scale (temporal or spatial) and uses upscaling to make prediction or inference at a larger scale. Consider a researcher who want to model how much rainfall a certain state receives in a chosen time period. The researcher could collect samples from multiple locations in the state over that time period, calculate averages for each sample and build a new distribution of those averages. This

distribution of averages would be a better representation of the entire state than sampling a single value from any of the original samples.

For the purposes of this paper only two specific types of upscaling will be investigated. The two types of upscaling investigated in this paper are essentially implemented as a form of averaging. This scaling question is predominantly motivated from predictive models used for probabilistic risk assessment, which are used to address complex human health risk assessment from exposure to contaminant chemicals. An example of this type of modeling would be fate and transport modeling. These models simulate the movement and chemical alteration of radionuclides as they move through the subsurface to help determine potential risk to humans. These models are usually set up so that random numbers are drawn to represent the beginning of time and are projected throughout time (and/or space) for the duration/extent of the model. Each realization corresponds to a unique, randomly selected, value from each of the input distributions. Each input distribution is created using available data.

The two upscaling techniques that will be compared in this paper are (1) using random draws from distributions that represent smaller time scales and (2) using a single draw from a distribution representing the mean over all time represented by the model. The main difference between these two techniques is that for technique 1, we apply the chosen function to our data before averaging occurs but in technique 2, we average our data first and then apply the function. For this research, technique 1 will be treated as the gold standard that we wish to replicate with technique 2. Although the capability exists to draw new random numbers at each time step it adds more computational complexity to problems that are already computationally intensive, therefore we want to see if technique 2 accurately mimics technique 1. If we can show that technique 2 is a reasonable alternative to technique 1 it will greatly reduce the computational

complexity of the model. For very simple functional forms we can show mathematically whether these two methods will yield similar results. The purpose of this paper is to examine how well technique 2 mimics technique 1 for more complex functional forms that we cannot mathematically compare, but first we will show how these two techniques compare for the simplest of functional forms.

# CHAPTER 2. MOTIVATING EXAMPLE

## 2.1. Linear Function

For this example, we will consider the simplest type of function, a linear response function. Suppose I wish to model how many scores will occur in NFL football games over the next 100 games. For this example, it is assumed that the number of scores that occurs is a simple linear function that consists of a variable X and a constant c. Let X be the random variable representing the distribution of possessions in a single game, and let c be some constant that represents what proportion of those possessions that result in a score. For simplicity, we will also assume that each game is independent and has no impact on other games. Y has the functional form:

$$y = f(x) = cx$$

Thus, Y is the random variable representing the number of scoring events, not points, that occur in a game. We are ultimately interested in the distribution of the total number of scoring events for the next 100 games.

### 2.1.1. Technique 1 (drawing a new value at each time step)

Technique 1 involves drawing a distinct realization from an underlying input distribution at each time step(game), so there are 100 realizations drawn across the 100-game period. We can obtain the mean and variance of the distribution of interest mathematically by considering the random variable $O_1$ that represents the sum of the number of scores over the 100-game time period.

$$O_1 = \sum_{i=1}^{100} Y_i = \sum_{i=1}^{100} cX_i = c\sum_{i=1}^{100} X_i$$

The expectation of $O_1$ is:

$$E[O_1] = E\left[c\sum_{i=1}^{100} X_i\right] = 100cE[X]$$

And since the $X_i$'s are assumed independent and identically distributed, the variance of $O_1$ is:

$$Var(\sum_{i=1}^{100} cX_i) = c^2[\sum_{i=1}^{100} Var(X_i)] = 100c^2Var(X_i)$$

Suppose that the average number of possessions per game is 25 but there is some uncertainty in this amount based on the data that has been collected, so the input distribution for the number of possessions per game is represented by a normal distribution:

$$X \sim N(\mu_1 = 25, \sigma_1 = 4)$$

Assume that c = 0.35, or that roughly thirty five percent of possessions result in a score. So, the E[X] = 25 possessions per game, n = 100 games, and c = 0.35. Therefore, the expectation of $O_1$ is E[$O_1$] = 100*0.35*25 = 875. So, I can expect to see around eight hundred and seventy-five scores over the next 100 games. The variability of $O_1$ is:

$$Var\left(\sum_{i=1}^{100} cX_i\right) = c^2\left[\sum_{i=1}^{100} Var(X_i)\right] = 100c^2Var(X_i) = 100(0.35)^2(4)^2 = 196$$

### 2.1.2. Technique 2 (drawing one value and applying to all time steps)

Technique 2 considers the input distribution as a random variable representing the average value of the input across the entire 100-game time period. This mimics the process of selecting one single value at the beginning of the 100-game time period and using this same value for each game of the entire time period.

The input distribution must be scaled appropriately to represent the effect of summing the number of scores across the 100-game time span. X must be temporally scaled so that it

represents the total number of scores in 100 games, not just a single game. If we denote the scaled distribution as $X_{100}$ , then the number of scores that occur in 100 games can be denoted as:

$$O_{100} = f_{lin}(X_{100}) = 100 * c * X_{100}$$

That is, a random value from the scaled distribution is chosen and multiplied by 100 to represent using that value at each of the 100-time steps, where $X_{100}$ is the distribution of the average of X. While this model appears to imply that there are the same number of possessions each game, the purpose is to create a distribution that represents the sum of the number of possessions across the 100 games. This sum can be represented as:

$$O_{100} = 100 * c * O_{100} = 100 * c * \frac{\sum_{i=1}^{100} X_i}{100} = c * \sum_{i=1}^{100} X_i$$

This will result in an expected value of:

$$E[O_{100}] = E\left[c \sum_{i=1}^{100} X_{100}\right] = 100 * c * E[X_{100}] = 100 * c * E[X]$$

The distribution of X is the same as in technique 1, however, the distribution of interest is of $O_{100}$, which is:

$$X_{100} \sim N(\mu_{100} = \mu_1 = 25, \sigma_{100} = \frac{\sigma_1}{\sqrt{100}} = 0.4)$$

Based on the underlying distribution, the expected value is:

$$E[O_{100}] = 100 * c * E[X_{100}] = 100 * 0.35 * 25 = 875$$

And the variance is:

$$Var(O_{100}) = 100 * c^2 * Var(X_{100}) = 100 * 0.35^2 * 4^2 = 196$$

If scaling is not performed for models of this nature, then the uncertainty will be overstated, and the resulting distributional variance will be too large. For this linear case, the distribution of the total number of scores that occur over the 100-game time period is the same for both technique 1 and 2. This example demonstrates that if our function is linear and the simulation involves drawing a number at the beginning of time and applying that same value at each time step, then the distribution of the average can be used, which corresponds to technique 2. Under these specific conditions, technique 2 mimics technique 1 and would be preferable due to the simplicity of this technique.

## 2.2. Nonlinear Function

### 2.2.1. Technique 1: Annual Sampling

Now we will consider a more complex case. For case two we will consider a nonlinear function. For the sake of the example we will assume that the number of scores in each game follows a quadratic function. The motivation for this function could be that as more possessions occur, the offense gets better at adjusting to the defensive schemes and can therefore score more easily. Whether or not this is realistic is not important for this demonstration. We are only interested in determining the effect of scaling in these situations. In this case the functional form of the response, y, is:

$$y = f(x) = cX^2$$

Under technique 1, we draw a distinct realization from our underlying annual input for each time step. In this Case,

$$O_1 = \sum_{i=1}^{100} Y_i = \sum_{i=1}^{100} cX^2{}_i = c \sum_{i=1}^{100} X_i{}^2$$

In this example, an analytical solution can be obtained when applying the function prior to addressing expectation and variance. The distribution of a squared normal is a chi-square with expectation,

$$E(X^2) = \sigma^2 + \mu^2$$

Applying the nonlinear function to that expectation provides,

$$E(O_1) = 100 * c * (\sigma^2 + \mu^2)$$

Similar calculations can be performed for the variance,

$$Var(O_1) = 2 * 100 * c^2 * \sigma^2 * (\sigma^2 + 2\mu^2)$$

Using the same example as before with $X \sim N(\mu_1 = 25, \sigma_1 = 4)$, and c = 0.35 we get,

$$E[O_1] = 100 * 0.35 * (4^2 + 25^2) = 22435$$

$$Var(O_1) = 2 * 100 * 0.35^2 * 4^2 * (4^2 + (2 * 25^2)) = 496272$$

**2.2.2. Technique 2 (drawing one value and applying to all time steps)**

Now we consider the input distribution as a random variable representing the average across the 100 game time period, which is the same equation that we used in Case 1,

$$X_{100} \sim N(\mu_{100} = \mu_1 = 25, \sigma_{100} = \frac{\sigma_1}{\sqrt{100}} = 0.4)$$

For Technique 2 we are interested in the distribution of the sum of all the scores over our 100-game time period. We obtain the distribution of interest by applying the nonlinear function to our random variable.

$$O = f_{nl}(X_{100}) = c{X_{100}}^2$$

From this we can find the expectation of Y as,

$$E[O_{100}] = E\left[c\left(\sum_{i=1}^{100} {X_{100}}^2\right)\right] = cE\left[\left(\sum_{i=1}^{100} {X_{100}}^2\right)\right] = 100 * c * (\mu_{100}^2 + \sigma_{100}^2)$$

We can also find the variance of Y as,

$$Var(O_{100}) = Var\left(c\left(\sum_{i=1}^{100} X_{100}\right)^2\right) = c^2 * 100^2 * Var(X_{100}^2) =$$

$$c^2 * 100^2 * [E((X_{100}^2)^2) - [E(X_{100}^2)]^2] = c^2 * 100^2 * [E(X_{100}^4) - [E(X_{100}^2)]^2] =$$

$$c^2 * 100^2 * [(\mu_{100}^4 + 6\mu^2\sigma^2 + 3\sigma^4) - (\mu^2 + \sigma^2)^2]$$

Now we use our input distribution for $O_{100}$ and apply those parameters to the previous two equations to give us:

$$E[O_{100}] = 100 * 0.35 * (25^2 + .4^2) = 21880.6$$

$$Var(O_{100}) = 0.35^2 * 100^2 * [(25^4 + 6 * 25^2 * 0.4^2 + 3 * 0.4^4) - (25^2 + 0.4^2)^2] =$$

$$490062.72$$

For the nonlinear function, Technique 1 and 2 give us different results for both the mean and variance. Technique 1 had a mean of 22435 and a variance of 496272, Technique 2 had a mean

of 21880.6 and a variance of 490062.72. For this nonlinear equation it appears that taking f(E(x)) or applying the function to the distribution of the average of the data, does not give us the same result as E(f(x)), or taking the average after applying the function to the data. Technique 1 is our gold standard and Technique 2 does not give us the same result as was the result of the linear case. Although they are not the same, our values for the mean and variance are still quite similar in this example. This might be an acceptable difference depending on the situation.

Now that we've set up our problem of interest, our next step is to choose a myriad of nonlinear functions, with a varying number of parameters that are too complex to compare mathematically and use simulation to see how well Technique 2 mimics Technique 1.

# CHAPTER 3. METHODOLOGY

## 3.1. Function Selection

There were 14 functions chosen to be used in testing the two scaling techniques against each other. These 14 functions were selected from David A. Ratkowsky's book "Handbook of Nonlinear Regression Models.". The functions were chosen based on the shape of the function, the number of parameters in the function, and whether they are well known, named, functions. In his book, Ratkowsky describes functions with one to six parameters with most of them containing one to four parameters. Each set of functions is also separated into semi-linear and nonlinear functions. The semi-linear functions tend to have shapes that are closer to a linear shape while the nonlinear functions tend to deviate strongly from a linear shape. Two to four functions were chosen from each class of parameters: 1, 2, 3, and 4 with roughly half from each class categorized as semi-linear and the other half categorized as nonlinear. This separation allowed us to analyze how the functions compared with respect to the category as well as the scaling technique used. Here is the list of functions used in this analysis, those denoted with * means they are classified as semi-linear:

1. $y = 1 - (\frac{1}{x^a})$*
2. $y = 1/(x + a)$
3. $y = a * x^b$ (Freundlich Model)*
4. $y = \log(a + b * x)$
5. $y = (a * b * x)/(1 + b * x)$ (Langmuir Model)
6. $y = 1/(a + b * x + \gamma * x^2)$ (Holliday Model)*
7. $y = a * (1 - \exp(-b * x))^{\gamma}$ (Chapman-Richards Model)
8. $y = a * \exp(-\exp(b - \gamma * x))$ (Gompertz Model)*

9. $y = a * \exp(-b * x) + \gamma * \exp(-\delta * x)$ (Classical Sum of Exponentials)
10. $y = a/(1 + \exp(b - \gamma * x))^{1/\delta}$ (Richards Model)
11. $y = a + b * \exp(-\gamma * (x - \delta)^2)$ (Bragg Equation)*

### 3.2. Parameter/ Distribution Selection

Within each of the chosen functions there are between 1 and 4 parameters, with one variable. Each of these parameters and the variable has an underlying distribution that we use to draw the parameter value from. The distributions of interest are the normal distribution and the lognormal. These distributions were chosen because they are some of the most common distributions and they represent two distinct types of distributions (symmetric and skewed). For each distribution we also chose the distributional parameters such as the mean and standard deviation for the normal distributions and the meanlog and sdlog for the lognormal distributions. These decide the spread and the shape of the underlying distributions. There are two distinct classes of distributional parameters chosen: bounded between 0 and 1 and bounded between 1 and 10. These two classes were chosen because, generally, functions behave more linearly when their parameters are bounded between 0 and 1, and less linearly when their parameters are larger than 1. We also chose to bound our variable between 0 -1 and 1-10 because in practice many functions do not have the same bounds for the variable and the parameters. When considering all possible combinations (2 types of functions, 2 types of distributions, and 2 types of parameter bounds, 2 types of variable bounds), there were 16 total outputs to analyze with each output containing 5 (semi-linear) or 6 (nonlinear) functions. Figure 1 is a flowchart that represents all possible combinations under the semi-linear functions. This represents half of all combinations.

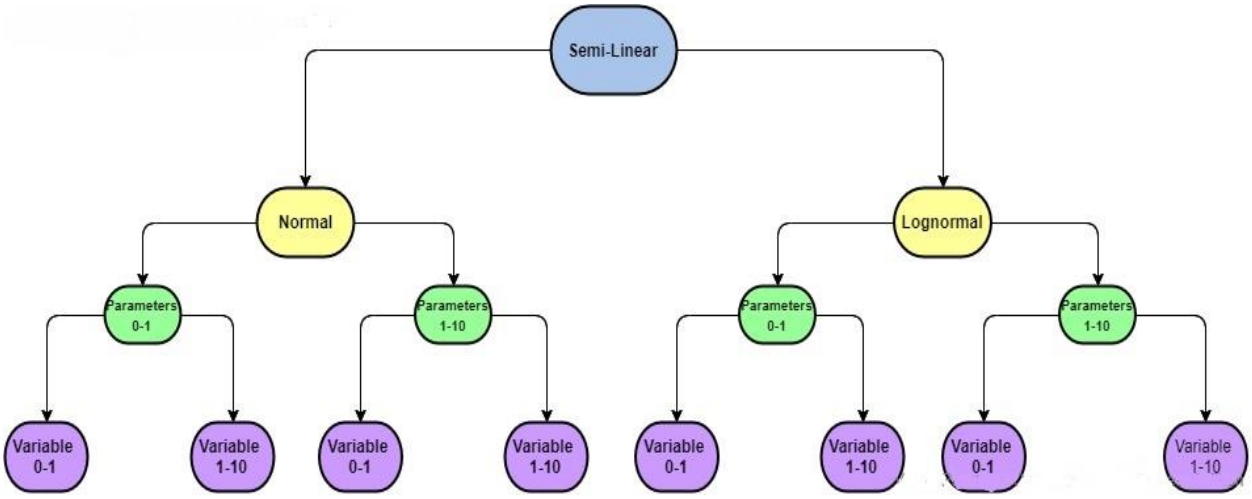


Figure 1. Flowchart of half of all combinations

### 3.3. Simulation Procedure

Each simulation program was created in R and an example of the code used is listed in appendix D.1. There were four programs in total, one was made for each number of parameters. Each program allows for the input of an underlying distribution and the corresponding distributional information for each parameter as well as the functional form to be tested. The simulation procedure used in this work followed three general steps: The first step was to simulate data using the parameter estimates and the chosen underlying distributions. The second step was to apply the sampled data to the chosen function, creating a new distribution of outputs for each sampling technique. The third step involved creating plots, summary statistics and running tests to compare the two output distributions. The following paragraphs will explain these steps in greater detail.

The data simulation for the first scaling technique was done using the rnorm, and rlnorm statements in R depending on the chosen distribution. In order to simulate data for the second scaling technique, the central limit theorem was used in conjunction with rnorm to simulate a scaled distribution of the average of those distributions originally chosen. In order to reduce the

number of combinations, each simulated data set had a sample size of 100. Once the two unique data sets were generated, they were then run through our chosen R program. After this we were left with two distributions of data, one distribution created using scaling Technique 1 and the other distribution created using scaling Technique 2. For this simulation, each run generated one single data point so in running 10,000 simulations we created a resulting distribution of 10,000 data points for each technique. This was repeated for each combination of function, distribution, parameter bounding, and variable bounding. The next step was to compare these two techniques based on those resulting distributions.

The goal of this work was to determine how well Technique 2 mimics Technique 1, or to determine how similar the two resulting distribution were. Determining similarity is somewhat ambiguous in this case, so in order to make that comparison a variety of measures were chosen. These measures included descriptive statistics and graphical summaries. The descriptive statistics used were mean, median, variance, and quantiles. We also calculated the reduction in means, medians, and variances from one technique to the other. The graphics used to compare the two distributions were overlaying histograms and quantile-quantile (QQ) plots. Both types of plots were created in R using the ggplot2 and the EnvStats packages. To complement the graphical comparisons, we planned on using the Anderson Darling K sample and Kolmorgorov-Smirnov two sample tests. The Anderson Darling K sample test is a nonparametric statistical procedure that tests the hypothesis that the populations from which two or more groups of data were drawn are identical. The Kolmogorov-Smirnov two sample test is used to test that two data samples come from the same distribution. In the end, we decided against using these distributional tests as well as any tests to compare the means and variances because our sample sizes of 10,000 gave us too much power and we were detecting even the most miniscule

differences between the two techniques. Using the statistical and graphical summaries together, we were able to compare the distributions resulting from the two techniques on a variety of levels.

# CHAPTER 4. RESULTS

## 4.1. Example Output

Here is an example of what the output looked like for one combination of function type, distribution, parameter bounding, and variable bounding. For this case we chose the two-parameter function $y = a * x^b$ which is considered a semi-linear function. This example output also had underlying distributions that were normal for all the parameters, and those distributions were bounded between 0-1 with the variable bounded 1-10.

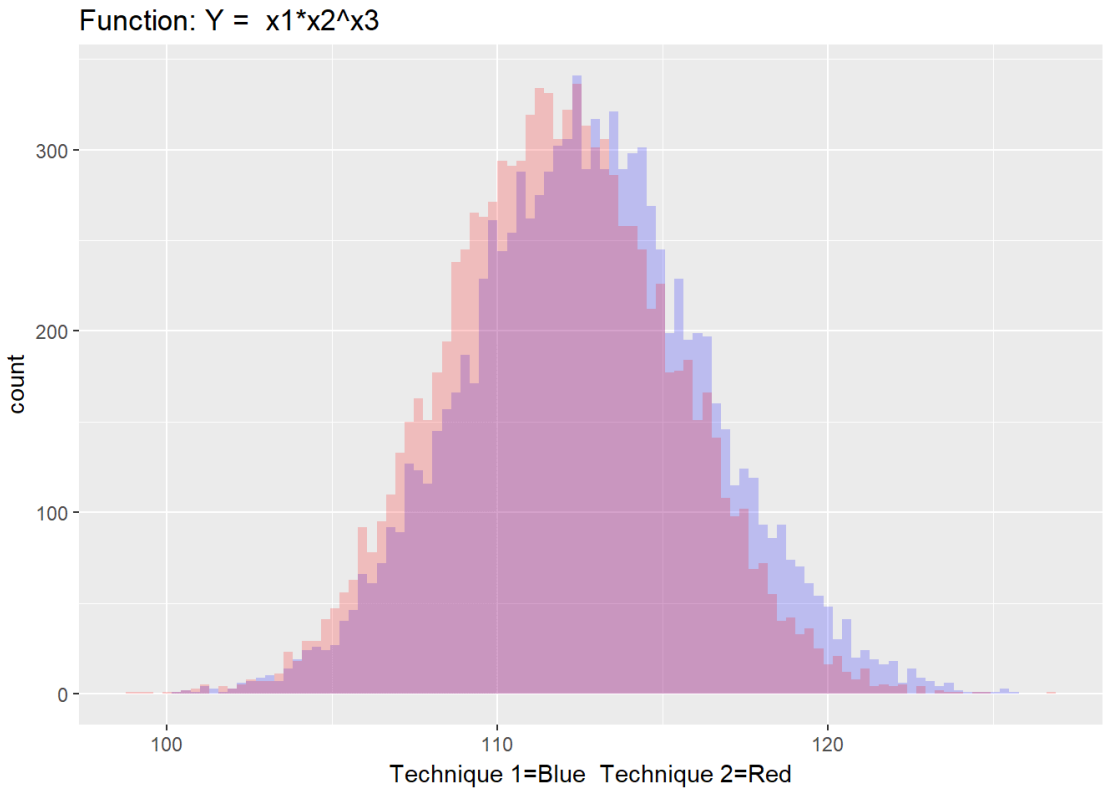


Figure 2. Distributions of Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 1-10.

This is one of the two plots generated by each R program, this plot overlays the distributions for the two techniques to give us an idea of how well Technique 2 mimics our gold standard, Technique 1. We can see that for this combination of function, distribution, and parameters, Technique 2 mimics Technique 1 closely with a slight shift to the left.

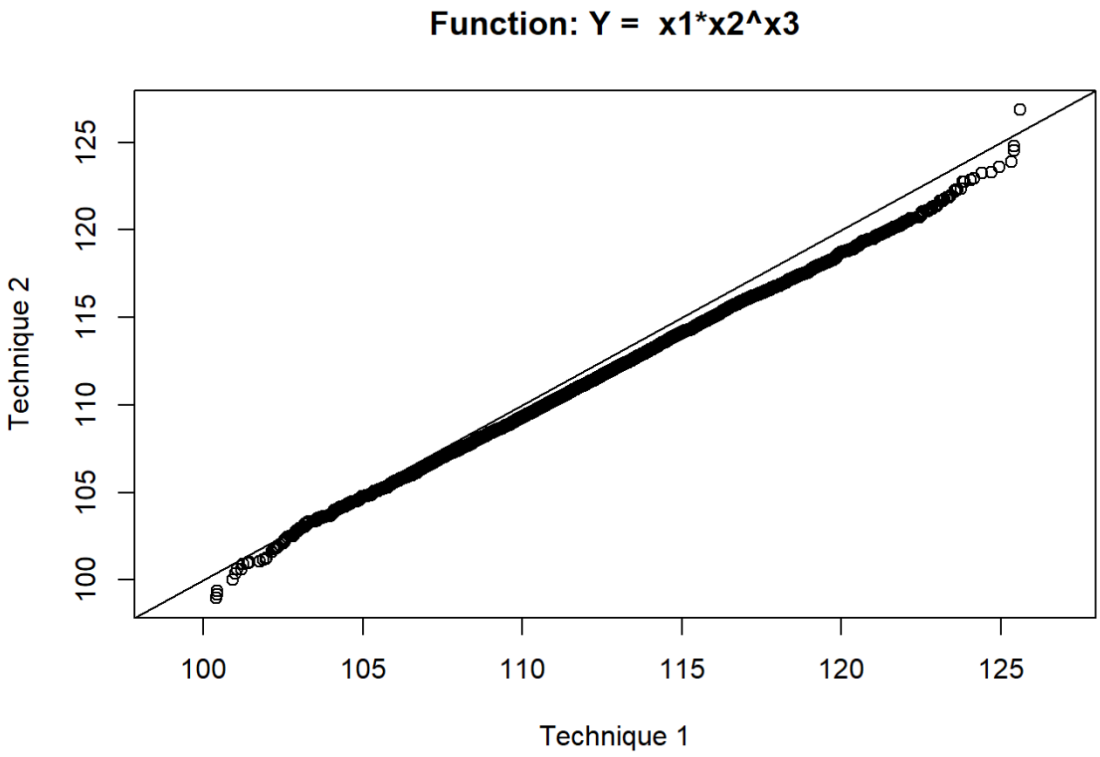


Figure 3. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 1-10.

The second plot given by each R program is a quantile-quantile (Q-Q) plot. This plot allows us to compare the distributions by plotting their quantiles against each other. The more similar the two distributions are, the more points will fall along the diagonal. For this plot we can see that most points fall close to the diagonal, but many are slightly under. This indicates that we have two similar distributions with one shifted slightly, which supports what we saw in the first plot.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 112.6571 | 112.575 | 13.0875 | 106.8217 | 118.7592 |
| Y2 | 111.86 | 111.811 | 12.0873 | 106.2865 | 117.7529 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9929 | 0.9932 | 0.9236 |

Figure 4. Descriptive statistics for Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 1-10.

This figure is an example of what kind of descriptive statistics we got for each function within each combination of type of function, underlying distribution, parameter bounding, and variable bounding. From this figure we can see the basic descriptive statistics for each of the two scaling techniques as well as the ratios of the estimates of the variance, mean, and median of Technique 2 compared to Technique 1. From the descriptive statistics we can see that the two techniques give us similar means, medians, and variances, with Technique 2 giving a slightly smaller range of values. Using the ratios row, we can confirm that the mean and median are almost identical, but the variance of Technique 2 is approximately 7.6% less than Technique 1. Considering these results, we can say that for this semi-linear function with an underlying normal distribution, parameters bounded between 0-1, and variable bounded between 1-10, the two scaling techniques give us similar, but not the same result.

Due to the high number of output statistics created we focused on the ratios of the estimates for much of our analysis. The ratio values given in the figures were used to make data sets for each comparison of interest. We had 40-48 total functions for each comparison which gave us 40-48 ratios of means and variances to use in our comparisons. Using the ratio of the means and variances we were able to calculate averages, medians, and standard deviations as well as the 5% and 95% quantiles for each comparison. Next, we looked at how the techniques

compared across all combinations of function type, distribution, parameter bounding, and variable bounding to see if there were any patterns or commonalities.

### 4.2. Parameters Bounded Between 0-1 vs Bounded Between 1-10

The first comparison of techniques we consider is parameter bounding between 0-1 vs parameter bounding between 1-10. For this comparison we would expect that, in general, when parameters are bounded between 0-1 our resulting functions would be more linear than when the parameters are bounded between 1-10. This would give us more similar distributions for Technique 1 and Technique 2. Our results appear to support our assumptions. For the parameter bounding of 0-1, the quantiles of the ratios of variances were (0.5764, 1.0460) with an average of 0.9040 and a median of 0.9699. The quantiles of the ratios of means were (0.7742, 1.0606) with an average of 0.9698 and a median of 1.0003. This is compared to quantiles of the ratios of variances of (0, 1.175) with an average of 1.0424 and a median of 0.9126. With quantiles of the ratios of means of (0, 1.1130), an average of 0.8077, and a median of 0.9998 for the 1-10 bounding. We found that only one or two functions performed considerably worse for the 0-1 bounded functions but there were more than 10 that performed considerably worse for the 1-10 bounded functions, this is more apparent in the standard deviations of the ratios of variances. The standard deviation for the 0-1 bounds group was 0.1731 with but the standard deviation of the 1-10 bounds group was 2.18. This is partly due to the fact that for certain equations, when certain parameters become larger than 1, functions such as $y = a * x^b$ end up giving us some extremely large values for Technique 1 but Technique 2 averages out those outliers and gives a much lower mean and variance. When our parameters remain below 1, we do not run into that same problem. Figure 5 shows us how different the results can be for the same function when we

choose different bounds and figure 6 gives us the summary statistics of the bounded 1-10 function.

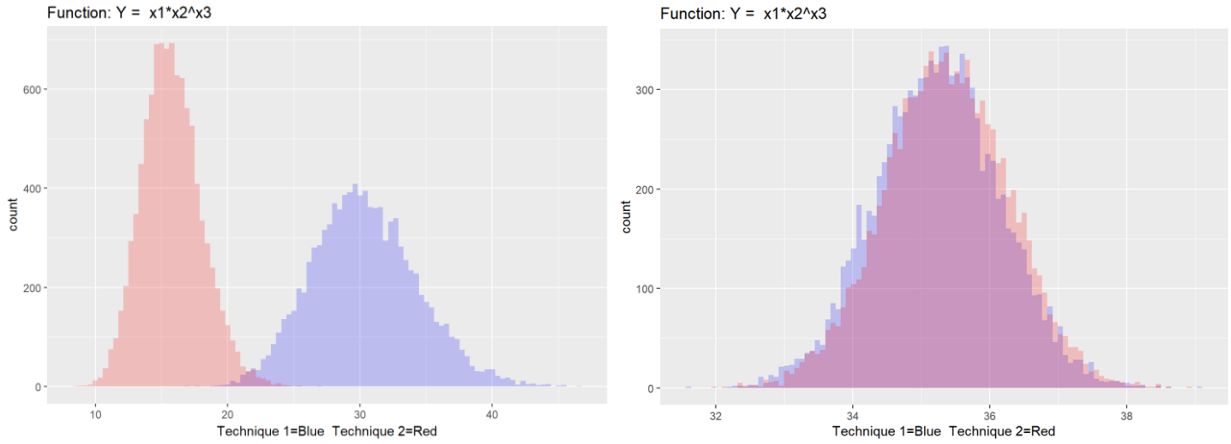


Figure 5. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1 (left). Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 0-1 (right).

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 30.4065 | 30.2332 | 15.6355 | 24.1451 | 37.1968 |
| Y2 | 15.74 | 15.6048 | 5.0911 | 12.2422 | 19.6691 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 0.5177 | 0.5161 | 0.3256 |

Figure 6. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * x^b$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

### 4.3. Semi-Linear vs Nonlinear Functions

The second comparison was the semi-linear functions vs the nonlinear functions for our two techniques. This is one of the two comparisons where different functions are used in each of the groups we are comparing. The reasoning behind this comparison is that we expect that equations who are closer to a linear form should result in more similar distributions than those who are more nonlinear, as we saw in the motivating example. Surprisingly, the nonlinear functions seem to perform just as well, if not better than their semi-linear counterparts. The

quantiles of the ratios of variances for the nonlinear equations are (0.5, 1.0895) with an average of .8710, a median of 0.9610, and a standard deviation of 0.2256. The quantiles of the ratios of means are (0.6972, 1.0524) with an average of 0.9225, a median of 1.004, and a standard deviation of 0.2246. This is compared to quantiles of the ratios of variances of (0, 1.1606) with an average of 1.0585, a median of 0.9285, and a standard deviation of 2.0769. The quantiles of the ratios of means were (0.0024, 1.0914) with an average of 0.8607, a median of 0.9994, and a standard deviation of 0.2650 for the semi-linear equations. The main difference between these two categories was that the nonlinear functions had a lower standard deviation of both variance and mean ratios which suggests that Technique 2 mimics Technique 1 better more consistently for the nonlinear functions.

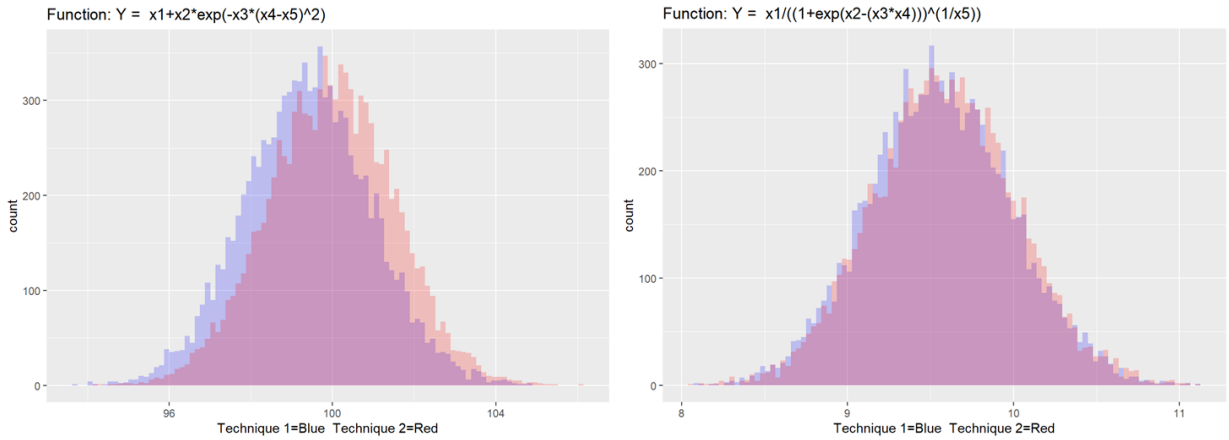


Figure 7. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 0-1, and variable bounded 0-1 for semi-linear function (left) and a nonlinear function (right).

### 4.4. Normal Distribution

Due to our normal and lognormal distributions have different parameters we did not compare across distributions but rather saw how the two techniques compared within each distributional category. Using an underlying normal distribution for our generated data we found that Technique 1 and 2 yielded similar results in most of our combinations. We had quantiles of the ratios of variances of (0, 1.0328) with an average of 1.0733, a median of 0.9245, and a standard deviation 2.1443. Having an average that close to 1 even when considering the outlier ratios means that in most cases these two techniques are giving us very similar spreads of distributions. The same goes for the ratios of means. We had quantiles of (0.0065, 1.0503) with an average of 0.8904, a median of 0.9998, and a standard deviation of 0.3133. Of the 44 combinations of functions, over half had ratios of variances in the range (.8, 1.2) and around 80% had ratios of the means in the same range. Figure 8 and 9 show an example function where the two techniques perform almost identically using underlying normal distributions.

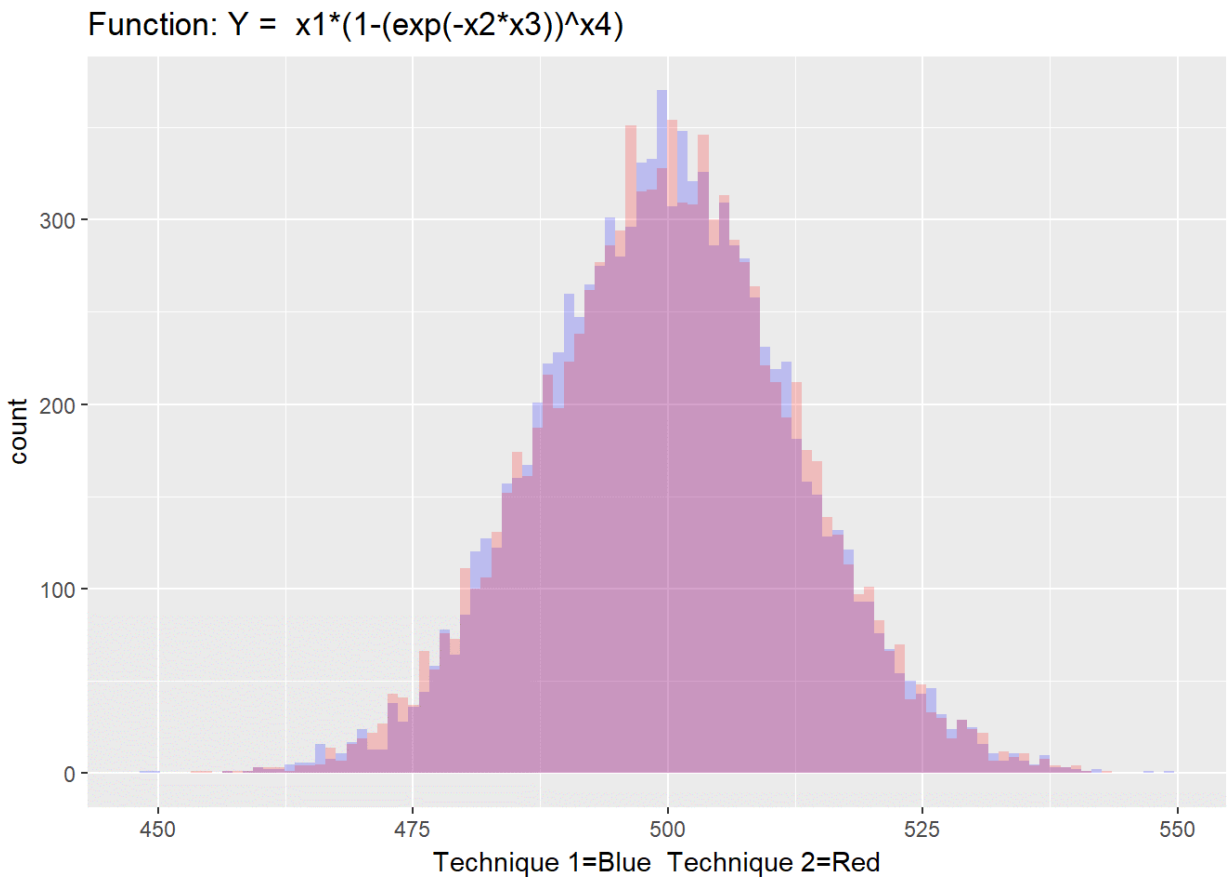


Figure 8. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

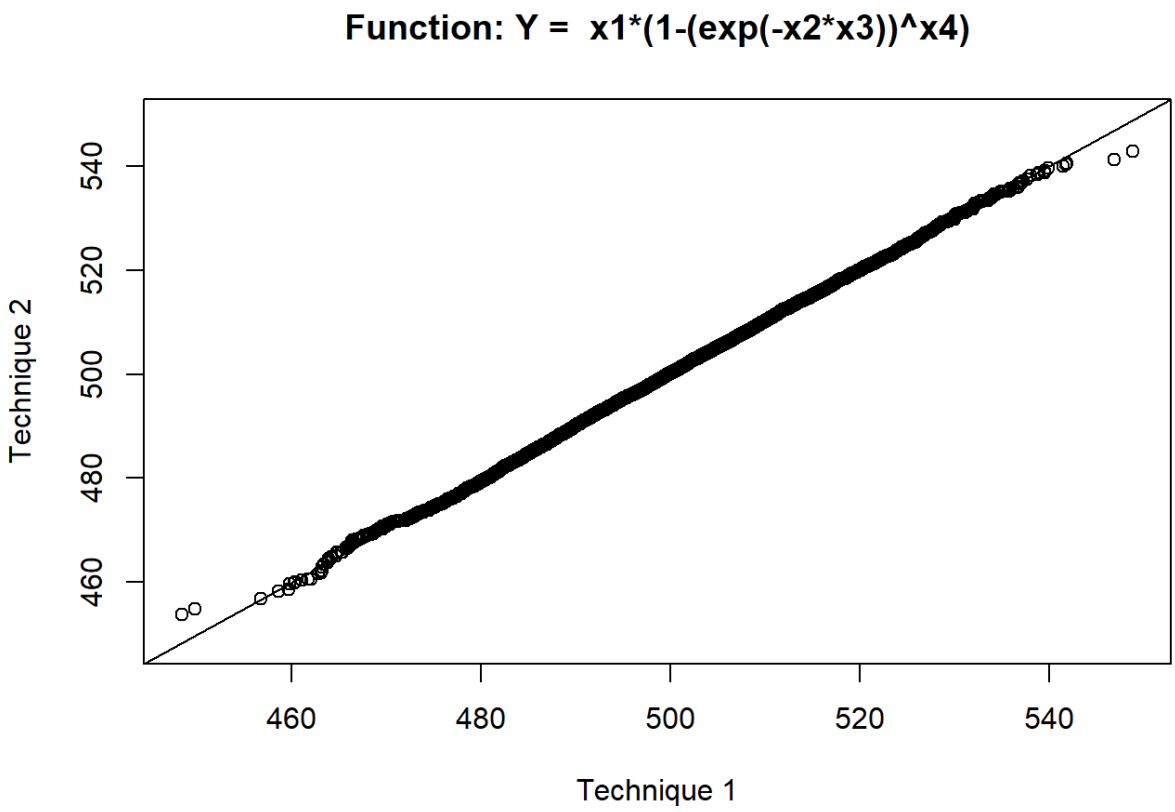


Figure 9. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

### 4.5. Lognormal Distribution

Using an underlying lognormal distribution for our generated data we also found that Technique 1 and Technique 2 yielded similar results in most of our equations. We found that the quantiles of the ratios of variances were (0.0012, 1.1661) with an average of 0.8731, a median of 0.9685, and a standard deviation of 0.4160. We found that the ratios of the means fared even better than our variances with quantiles of (0.0488, 1.0955) an average of 0.8871, a median of 1.0002, and a standard deviation of 0.3085. Of the 44 functions almost 75% had ratios of variances between (.8, 1.2) and around 80% had ratios of means in the same range. Figure 10 and 11 give an example of how well even complex functions performed with underlying lognormal distributions.

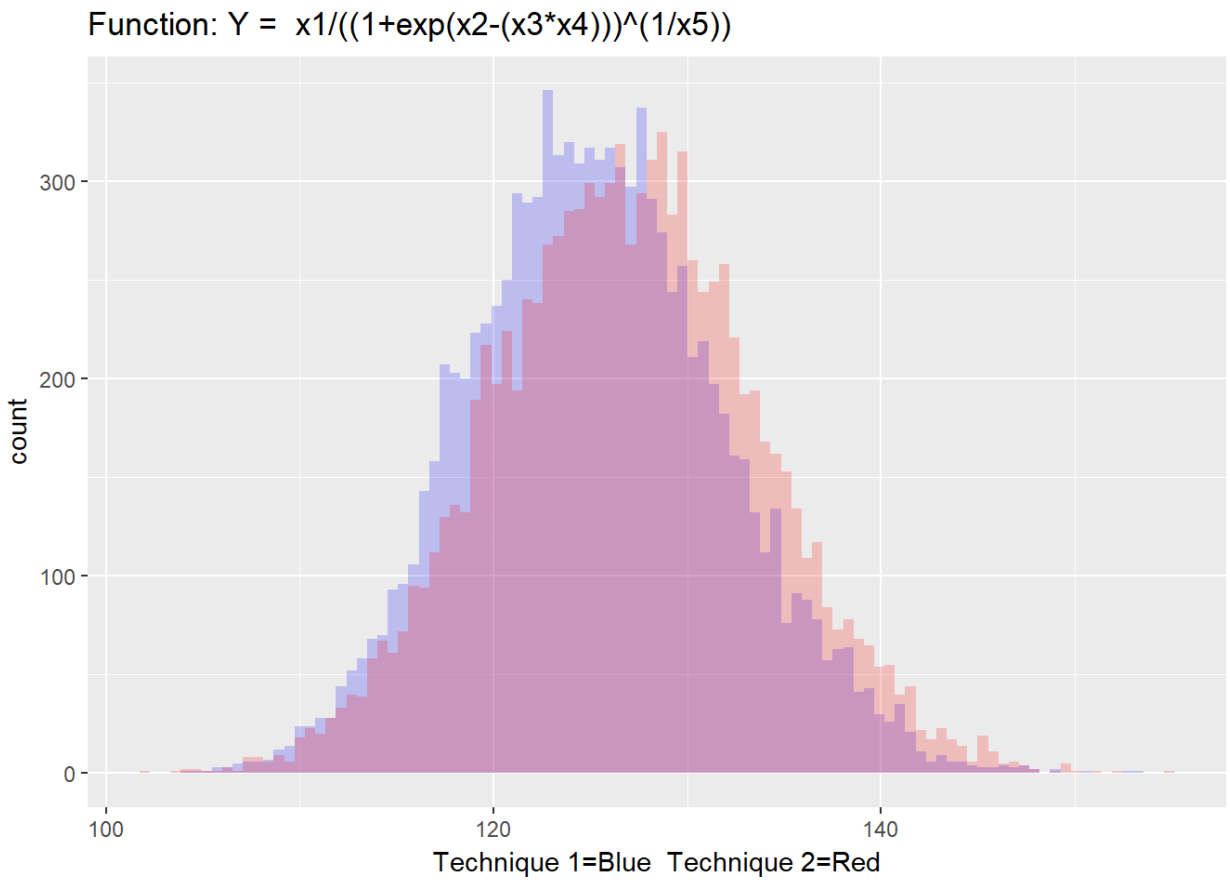


Figure 10. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

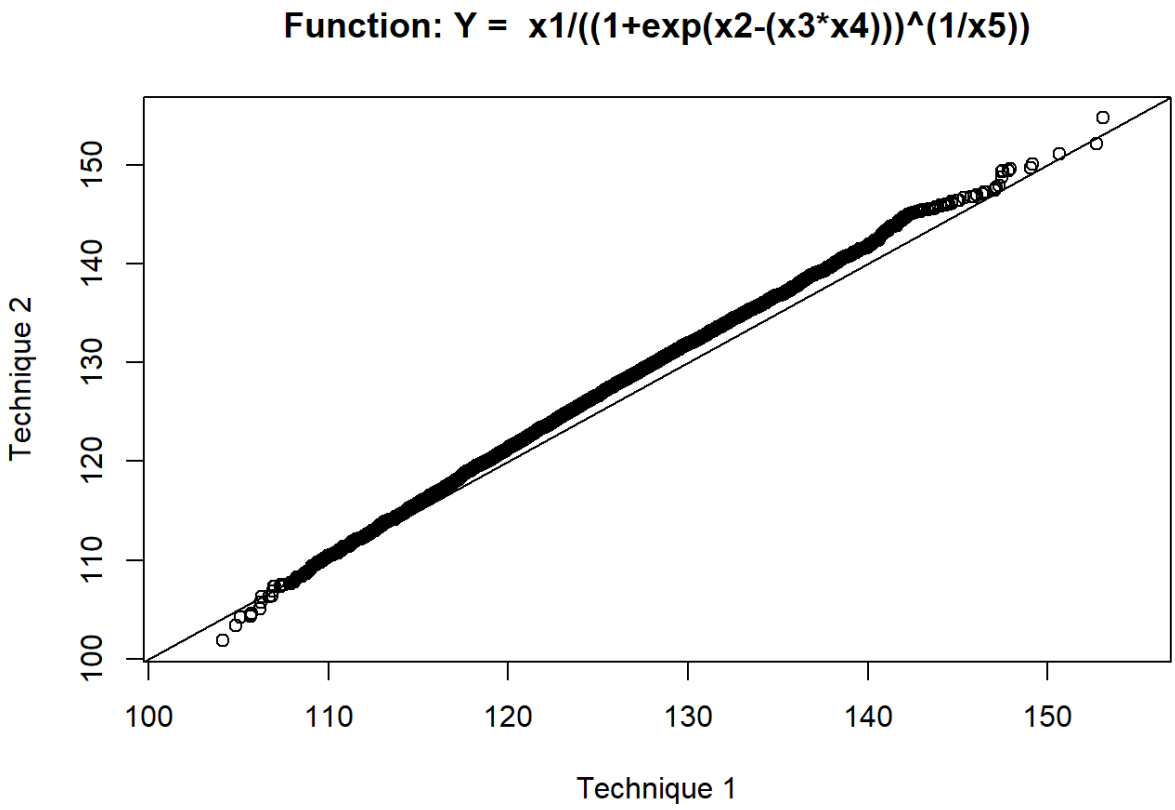


Figure 11. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

### 4.6. Variable Bounded 0-1 vs Bounded 1-10

The fourth comparison was bounding the variable between 0-1 vs bounding the variable between 1-10. We chose to do this comparison because in practice variables can take on many values, just like the parameters, and we wanted to see if there was a significant difference in our two scaling techniques depending on the scale of our variable. We found that for the variable bounded from 0-1 group there were quantiles of the ratios of variances of (0.0001, 1.0229) with a mean of 0.8136, a median of 0.9753, and a standard deviation of 0.3291. Quantiles of the ratios of means were (0.0012, 1.0241) with a mean of 0.8610, a median of 0.9970, and a standard deviation of 0.3037. These are compared to quantiles of the ratios of variances of (0, 1.1630)

with a mean of 1.1328, a median of 0.9380, and a standard deviation of 2.1520. Quantiles of the ratios of means were (0.0488, 1.0955) with a mean of 0.9165, a median of 1.0007, and a standard deviation of 0.3154. These numbers suggest a similar conclusion as our parameter bounds section, the median variance ratio is closer to 1 and the standard deviation of our variance ratios is considerably smaller when we bound our variable below 1. This suggest that, for our functions, Technique 2 mimics Technique 1 more consistently when we bound our variable below 1.

### 4.7. Number of Parameters

One other comparison worth noting is the comparison across number of parameters in our functions. When looking at figure 12 we can see that the variance ratio for all number of parameters is similar. The four parameter functions appear to have the closest results for Technique 1 and 2 with the two and three parameter functions performing slightly worse. The one parameter functions have the worse average variance ratio. One thing to note is that we removed a severe outlier in the four-parameter category to get values that were more representative of the whole dataset.

| Parameters | Avg Var Ratio | Med Var Ratio | Var 5%ile | Var 95%ile | SD Var |
|---|---|---|---|---|---|
| 1 | 0.609 | 0.609 | 0 | 1.02 | 0.363 |
| 2 | 0.867 | 1.00 | 0.0305 | 1.13 | 0.348 |
| 3 | 0.799 | 0.919 | 0.0452 | 1.04 | 0.324 |
| 4 | 0.937 | 1.01 | 0.0613 | 1.10 | 0.464 |

Figure 12. Variance ratios across number of parameters.

When we look at figure 13, we can see that Technique 2 does a better job of mimicking the means than it does the variances. There is less variation between the averages, medians, and standard deviations of the number of parameters. The two, three, and four parameter functions

averages are nearly 90% with the one parameter functions doing slightly worse at nearly 85% but their medians are all great with each class having better than 90%.

| Parameters | Avg Mean Ratio | Med Mean Ratio | Mean 5%ile | Mean 95%ile | SD Mean |
|---|---|---|---|---|---|
| 1 | 0.844 | 0.940 | 0 | 1.07 | 0.334 |
| 2 | 0.874 | 1.01 | 0.117 | 1.06 | 0.304 |
| 3 | 0.887 | 1.00 | 0.121 | 1.09 | 0.288 |
| 4 | 0.892 | 0.997 | 0.114 | 1.06 | 0.310 |

Figure 13. Mean ratios across number of parameters.

# CHAPTER 5. CONCLUSION

The purpose of this research was to determine how well Technique 2 could mimic Technique 1 under a variety of different circumstances. When considering the results outlined in the last chapter some conclusions can be made. When we consider all the results together, we found that over 65% of all ratios of variances were between (0.8, 1.2) with a mean of 0.9706, a median of 0.9668 and a standard deviation of 1.539. The means fared even better across all results, we had over 75% of all ratios of means between (0.8, 1.2) with an average of 0.8853, a median of 0.9999, and a standard deviation of 0.309. This is a good indicator that in most cases these two techniques give us similar, although not exact, distributions. This also tells us that Technique 2 does a better job of mimicking the center of the distribution than it does the spread. Conclusions can also be made about the comparisons detailed in the results chapter.

In 5 out of the 6 comparisons made in chapter 4 we found that the distributions and proportional means and variances of Technique 1 and Technique 2 were relatively similar. The one case in which we found considerable differences between our two techniques was for the comparison of bounded below one parameter values vs the bounded above 1 parameter values. When we allowed our parameter values to go above 1 this caused many of the functions to become highly nonlinear and Technique 2 did worse at mimicking Technique 1. For example, the equation $y = a * \exp(-b * x) + \gamma * \exp(-\delta * x)$ had ratios of variances of 1.0036 when the parameters were bounded below one, but that number dropped to approximately 0 when the parameters could go above 1. This is just one example but there were many cases where this occurred.

The acceptable degree of difference between our two techniques would obviously depend on the specific scenario but in many of our analyses we found that the ratios of means and

variances of the two distributions were close to 1 but rarely exactly 1. This suggests that it would be acceptable to use Technique 2 as an alternative to Technique 1 if the researcher is comfortable with trading some accuracy for reducing the computational complexity of their model. Although, if you have nonlinear equations where the all parameters can go above 1, we cannot recommend using Technique 2 to replace Technique 1. Another interesting finding is that complexity of function also does not seem to have an adverse effect on how well Technique 2 can mimic Technique 1 for these one variable functions. The 2, 3, and 4 parameter functions performed similarly well with the one parameter functions performing slightly worse. This is likely because there were less 1 parameter functions which resulted in smaller samples of ratios for analysis.

This thesis shows that there are many things one must consider when deciding which of these two scaling techniques is most appropriate. It is important that the researcher understand how different functions, distributions, parameter and variable bounding can affect the distribution of interest and that each scenario may have completely different outcomes. It may be advisable for a researcher to conduct their own simulations, when possible, using their own functions, distributions, and parameters to determine if these two techniques give them results that are acceptably close. They can then make an informed decision on how to adequately scale their data. This exploratory analysis only scratches the surface of how to compare these two scaling techniques and is meant to be used as a starting point.

## REFERENCES

Neptune and Company Inc. (2018, October). *Scaling Considerations for Performance Assessments*. Retrieved from Energy.gov: https://www.energy.gov/sites/prod/files/2019/01/f58/4-Black.pdf

Ohio Environmental Protection Agency Division of Drinking and Ground Waters. (1995). Ground Water Flow and Fate and Transport Modeling. *Technical Guidance Manual for Hydrogeologic Investigations and Ground Water Monitoring*.

Ratkowsky, D. A. (1990). *Handbook of Nonlinear Regression Models.* New York: Marcel Dekker.

Sylwia Trzaska, E. S. (2014). *A Review of Downscaling Methods for Climate Change Projections.* United States Agency of International Development.

## APPENDIX A. EXAMPLE DISTRIBUTION PLOTS

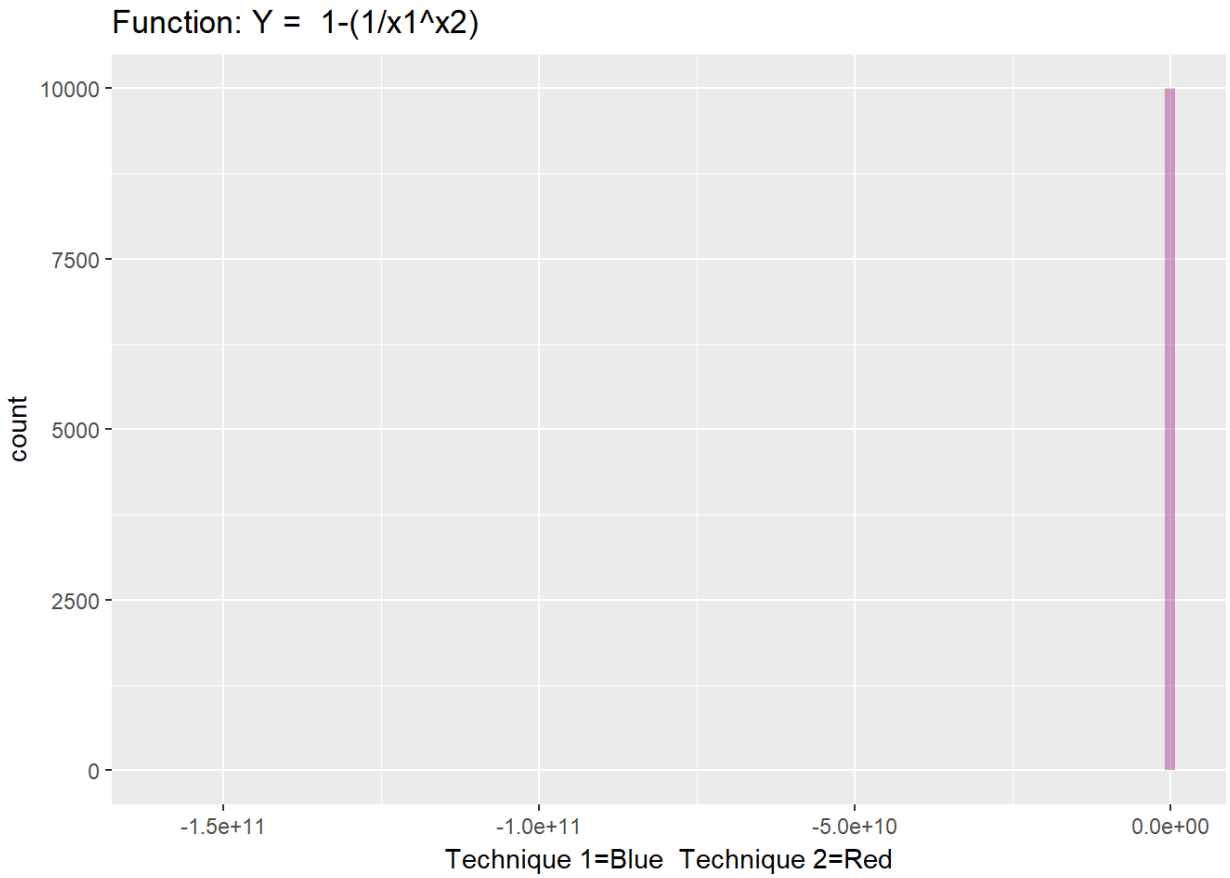


Figure A1. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

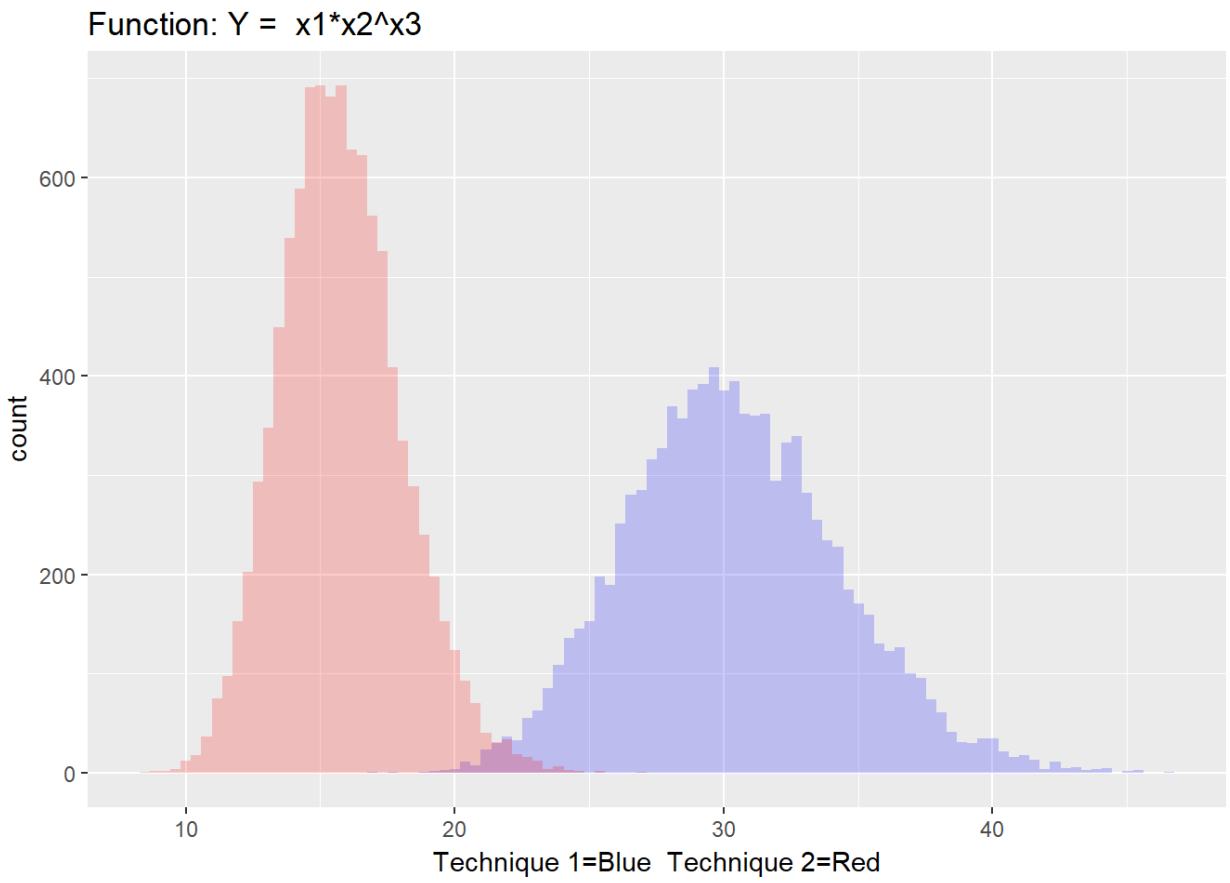


Figure A2. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

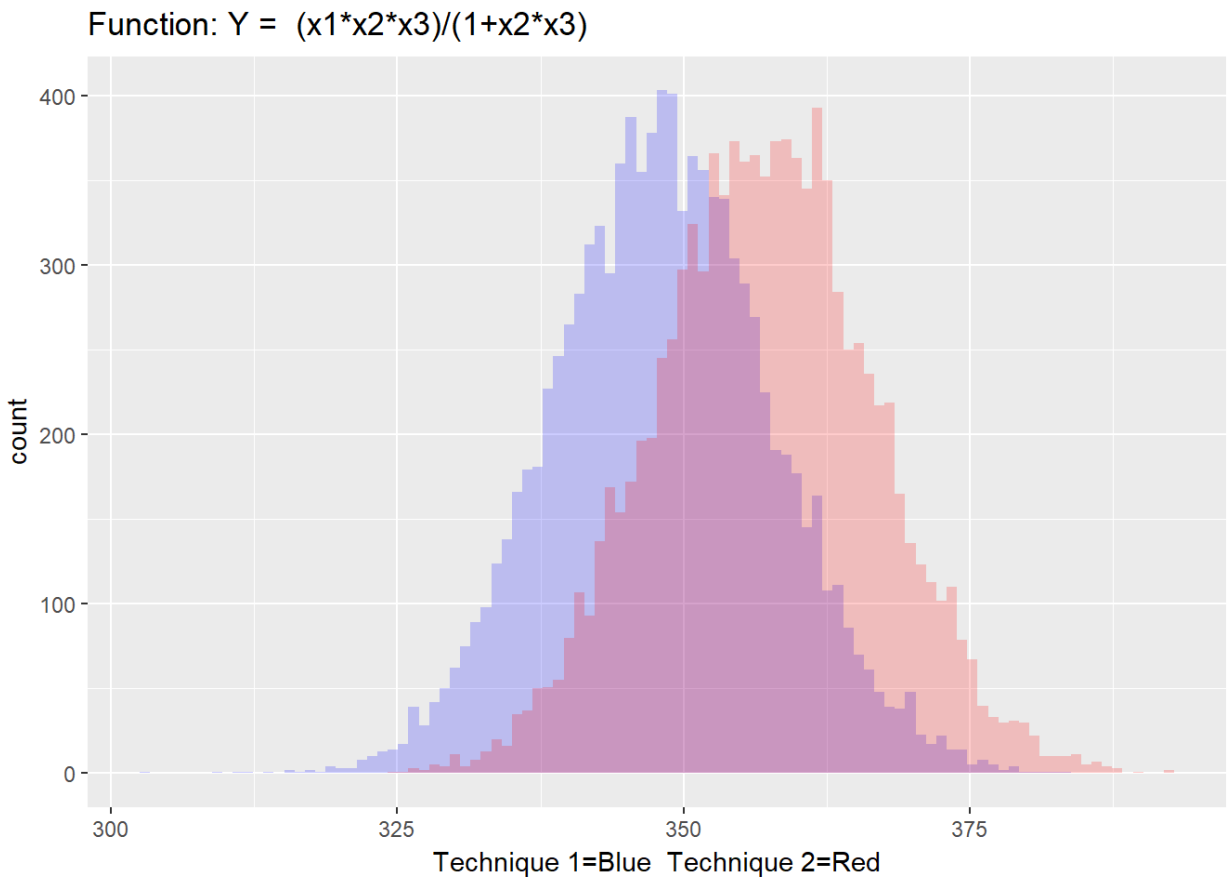


Figure A3. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

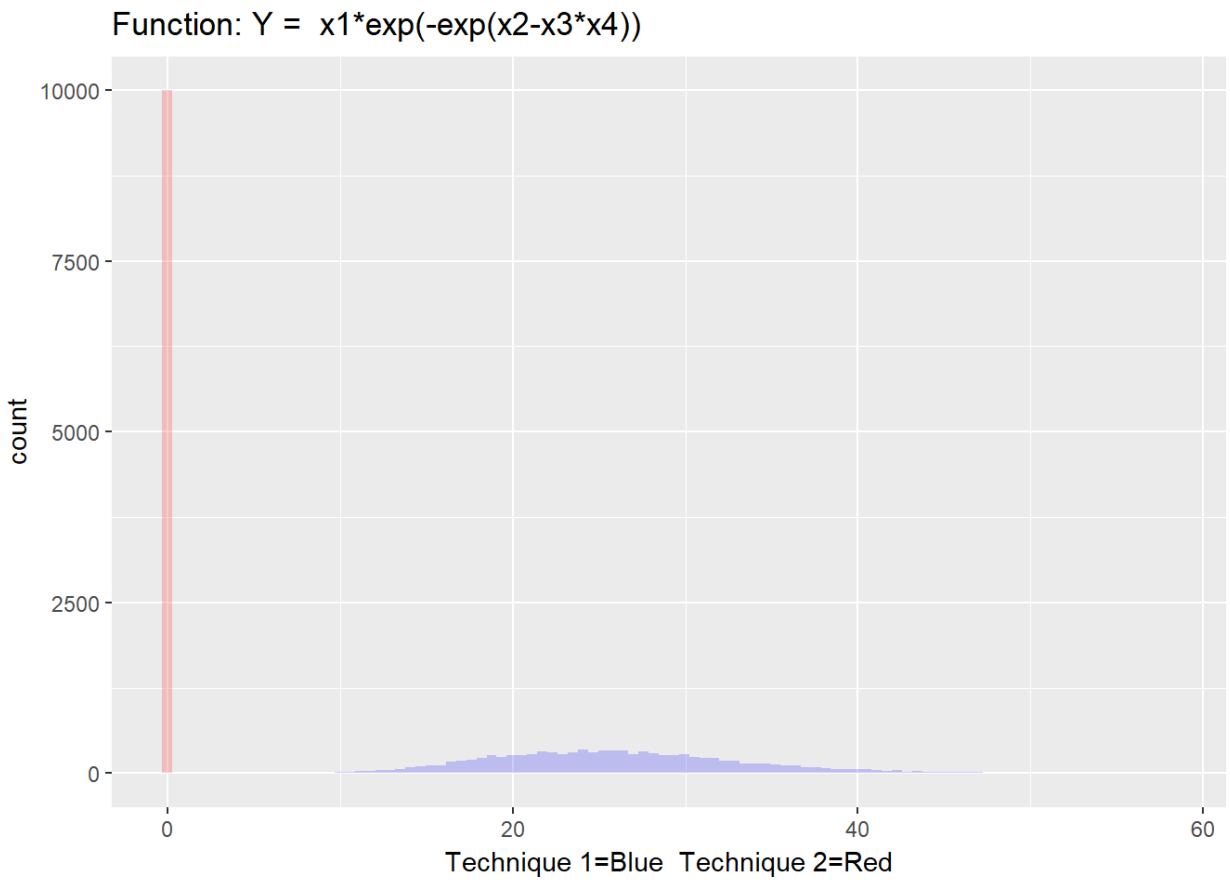


Figure A4. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

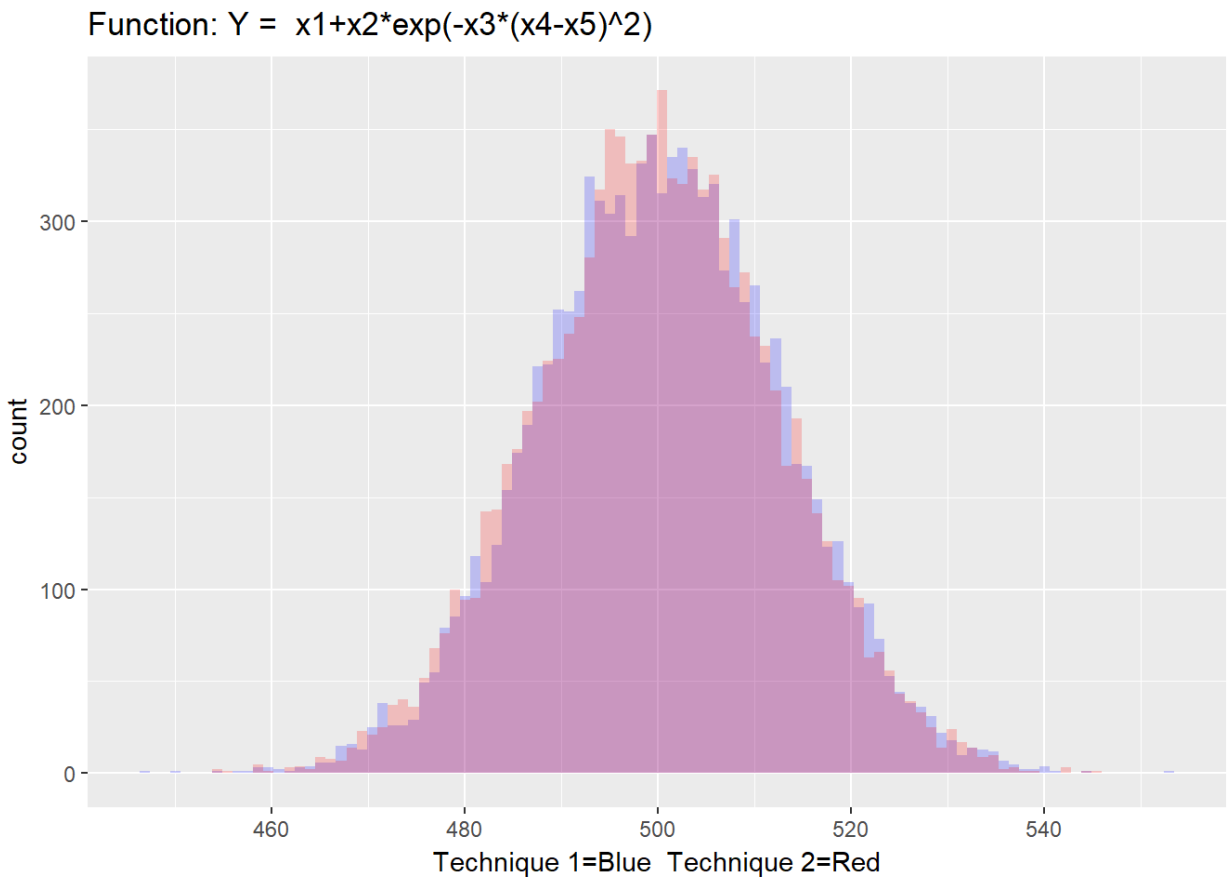


Figure A5. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

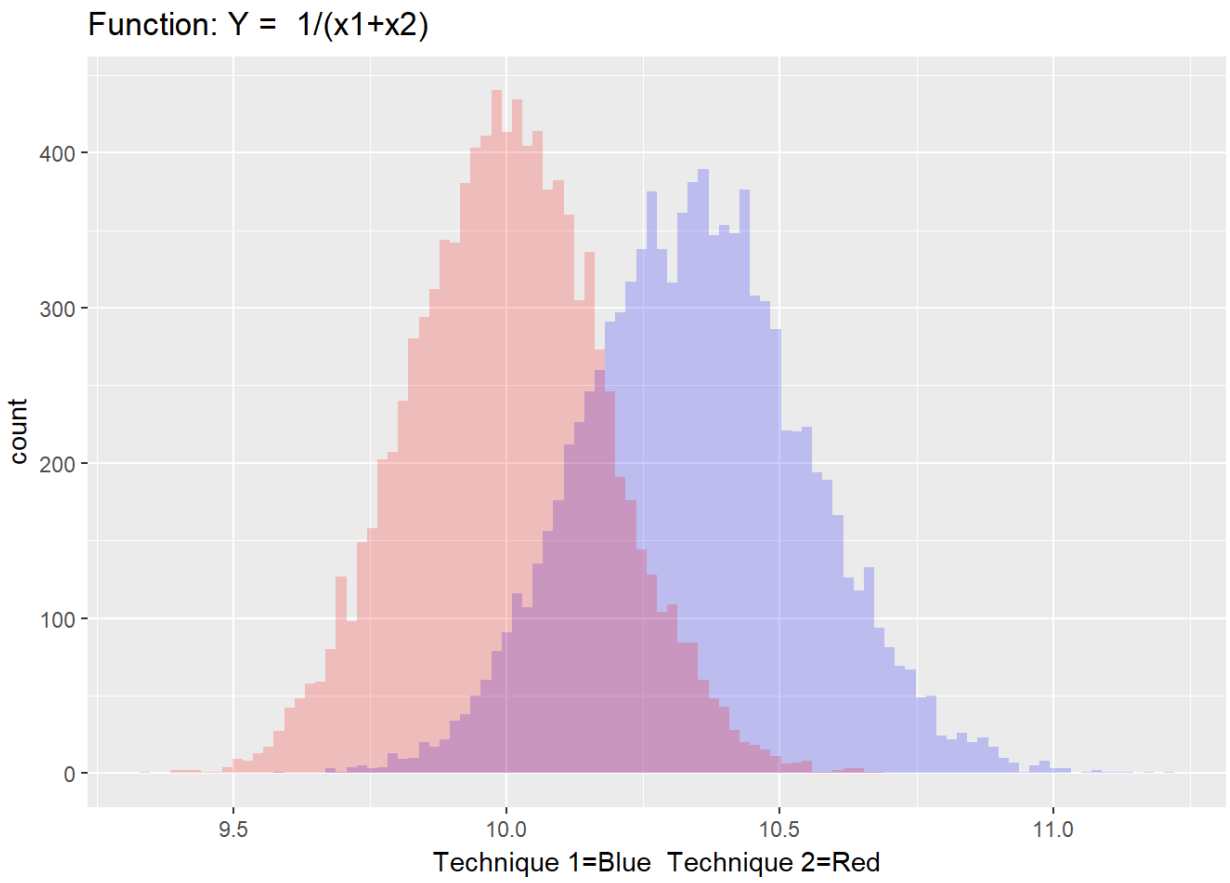


Figure A6. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

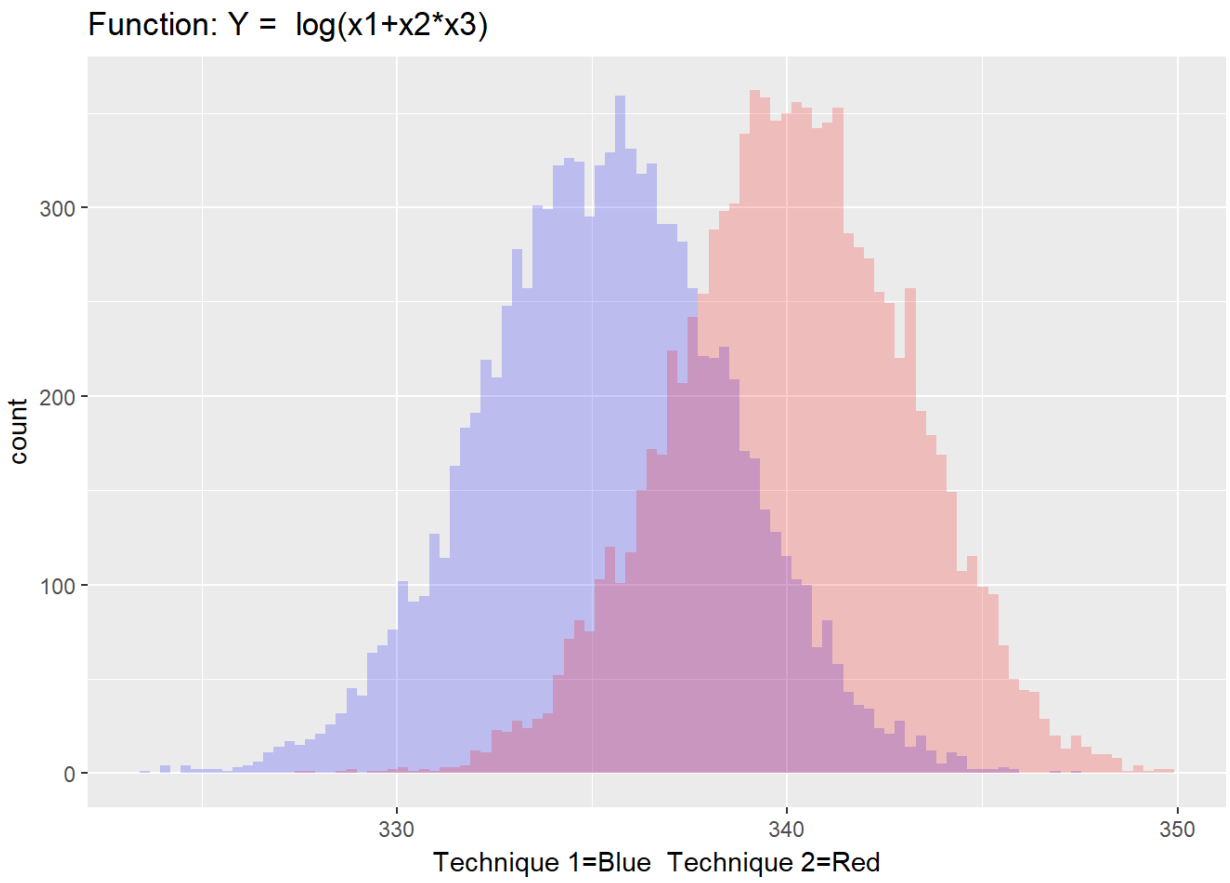


Figure A7. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

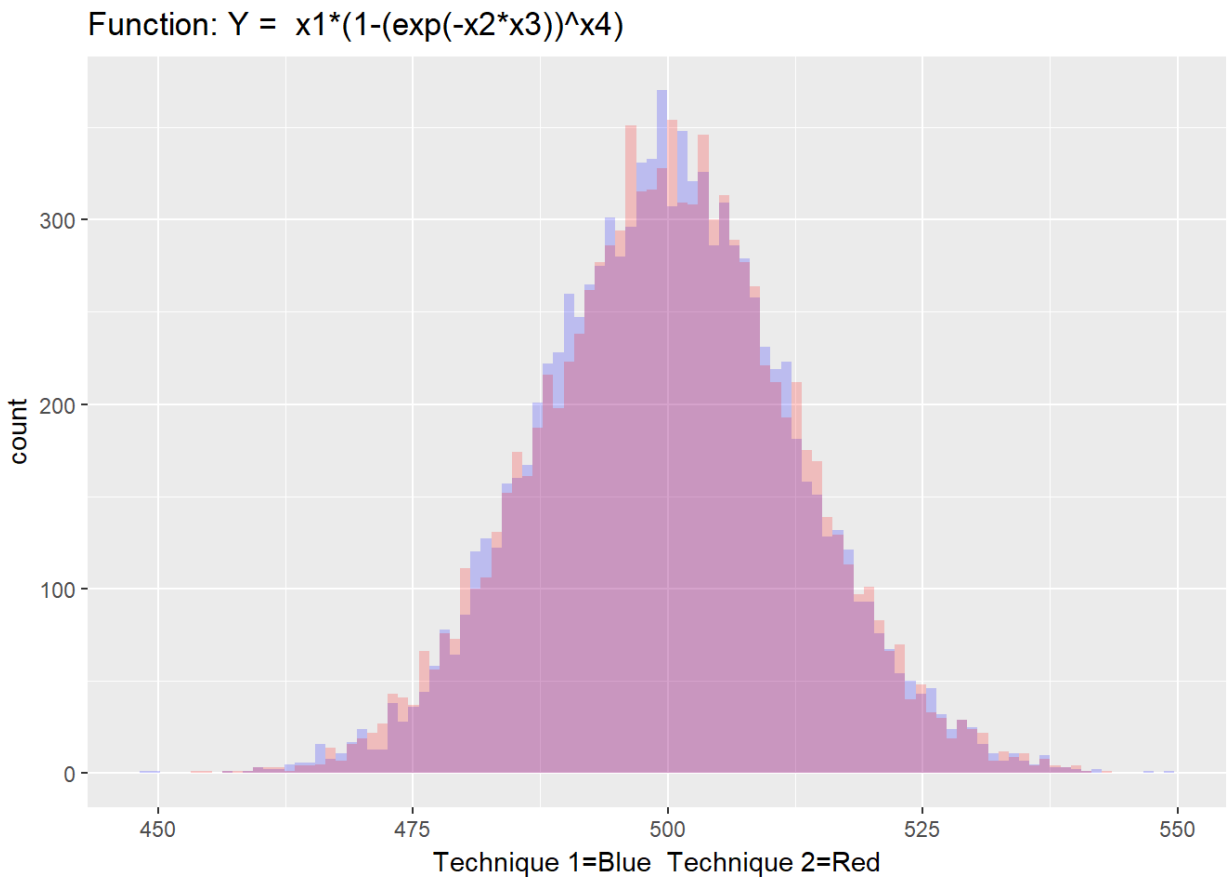


Figure A8. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

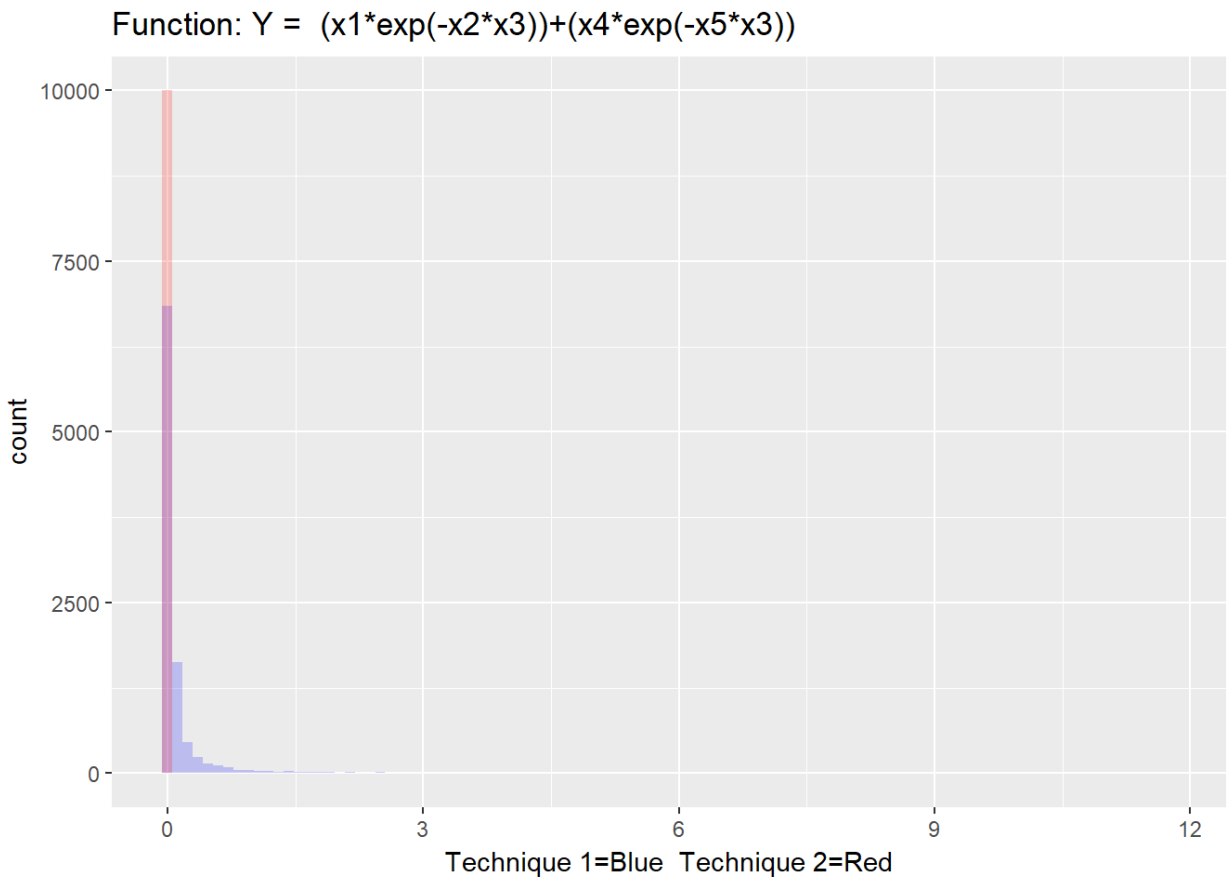


Figure A9. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

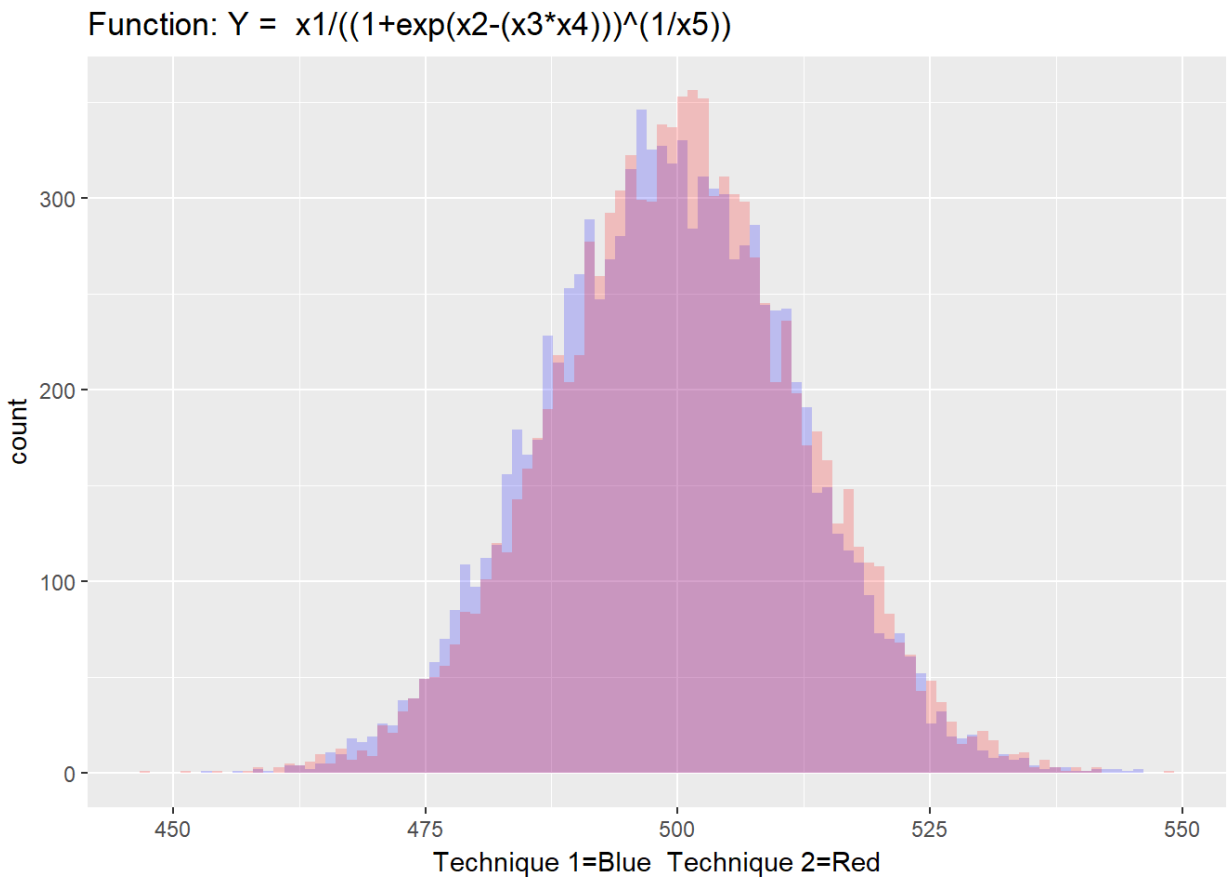


Figure A10. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

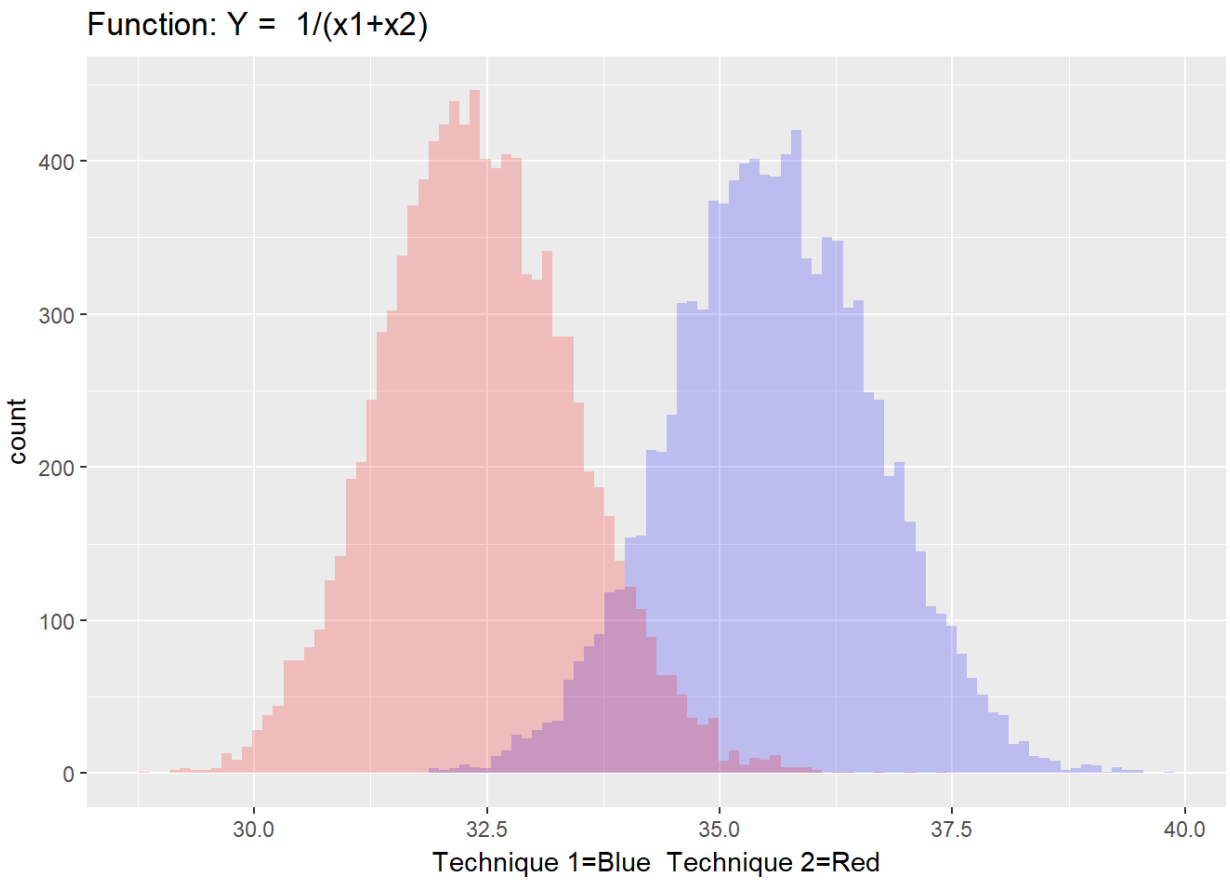


Figure A11. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

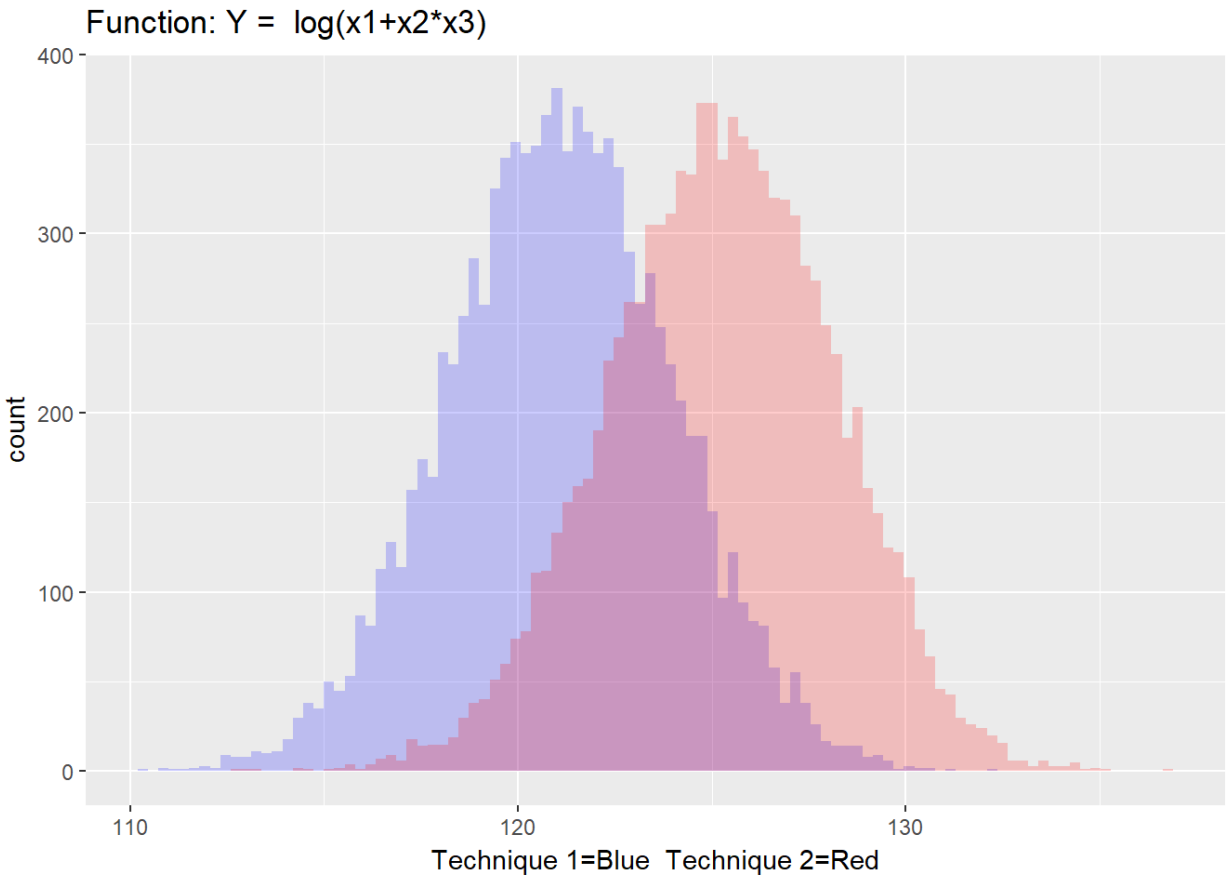


Figure A12. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

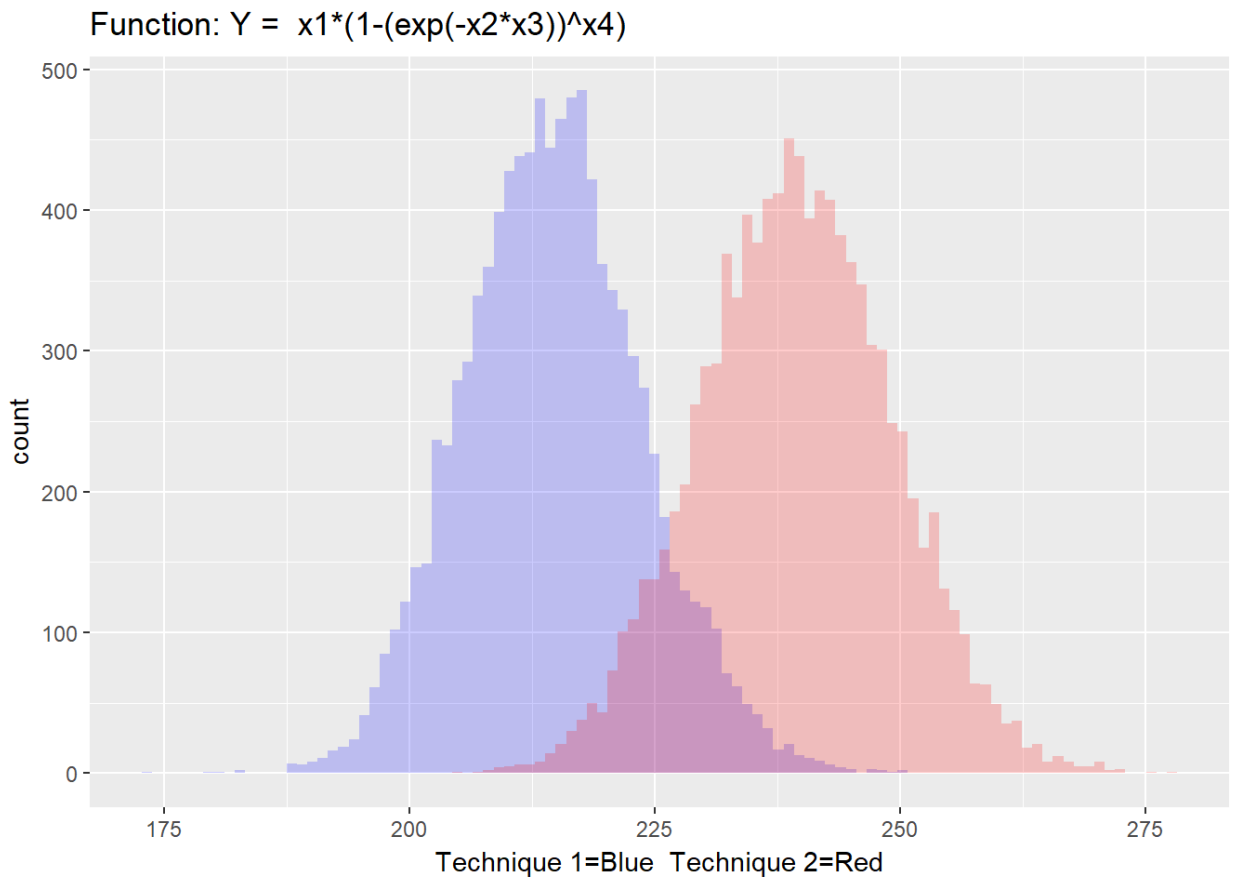


Figure A13. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

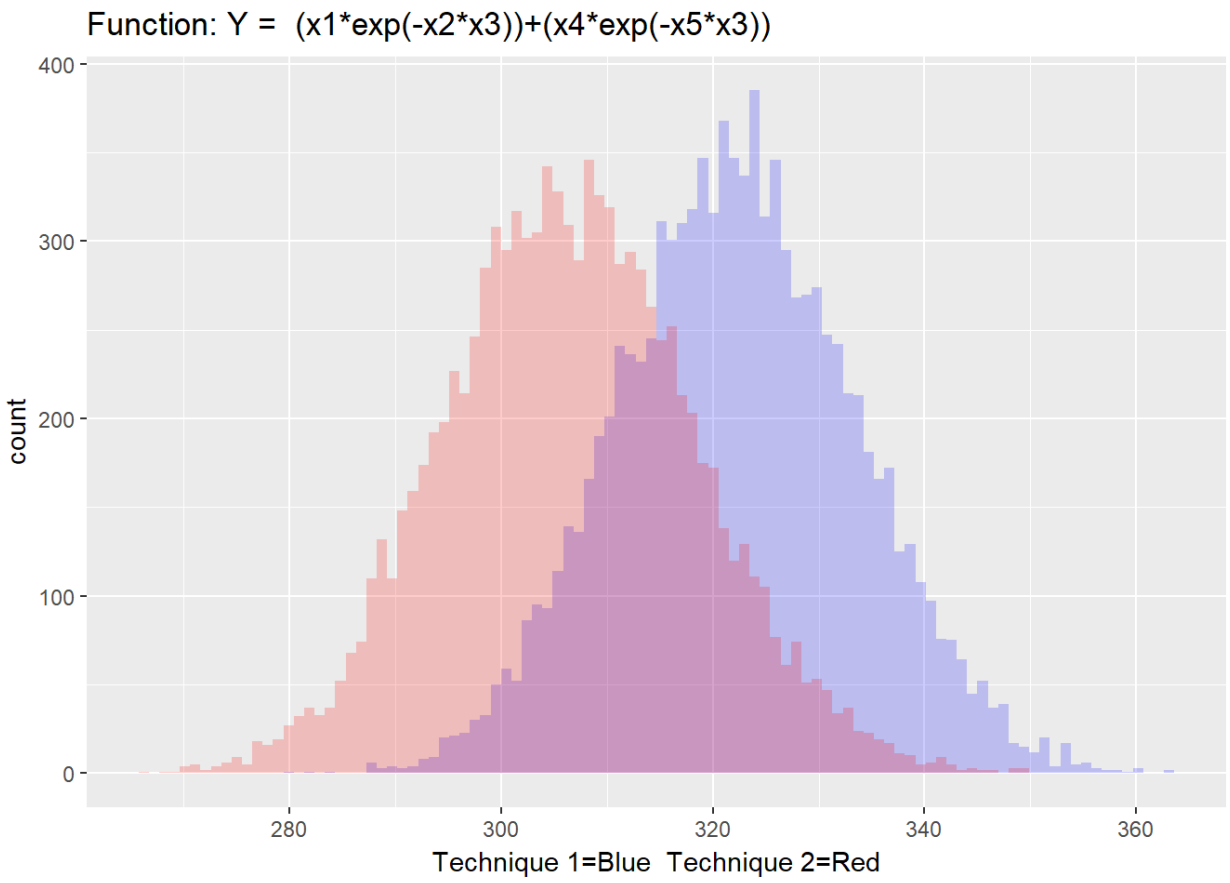


Figure A14. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

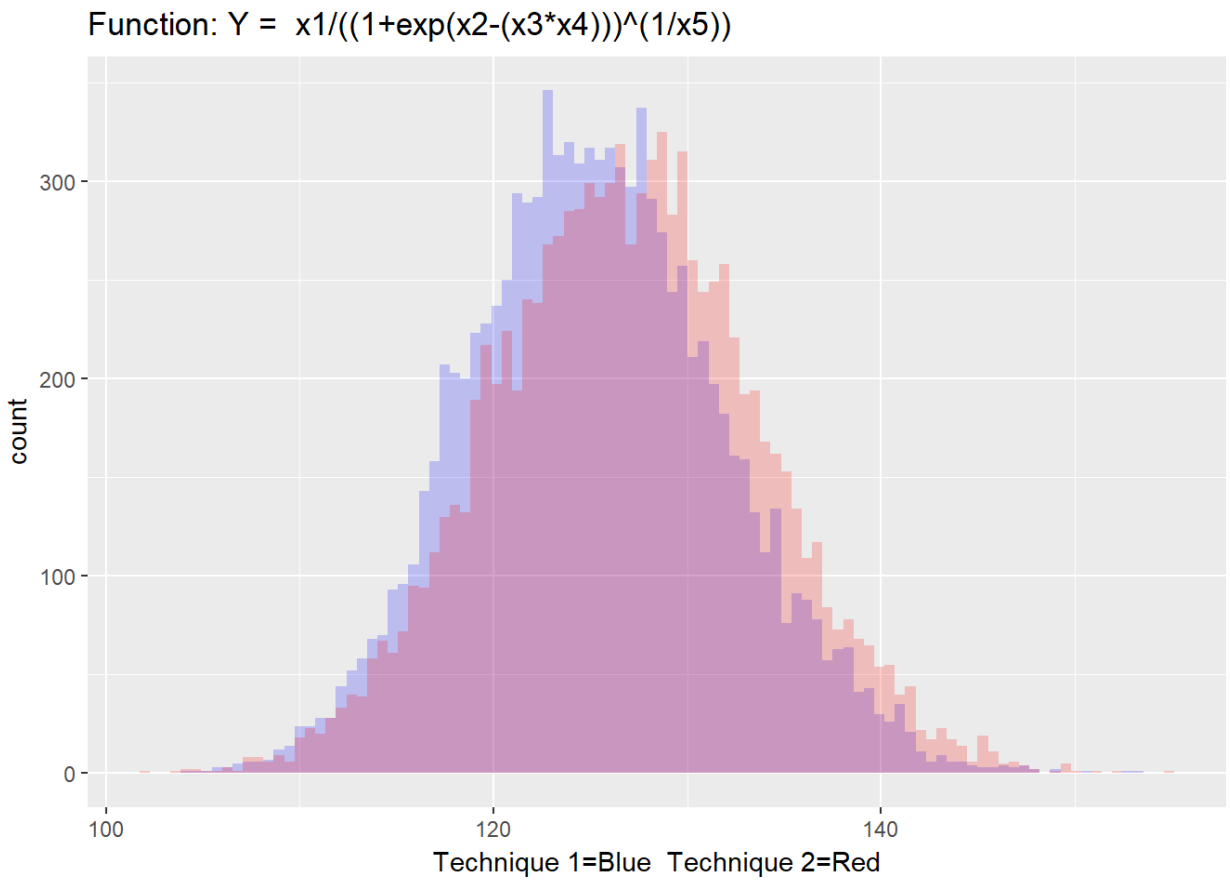


Figure A15. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

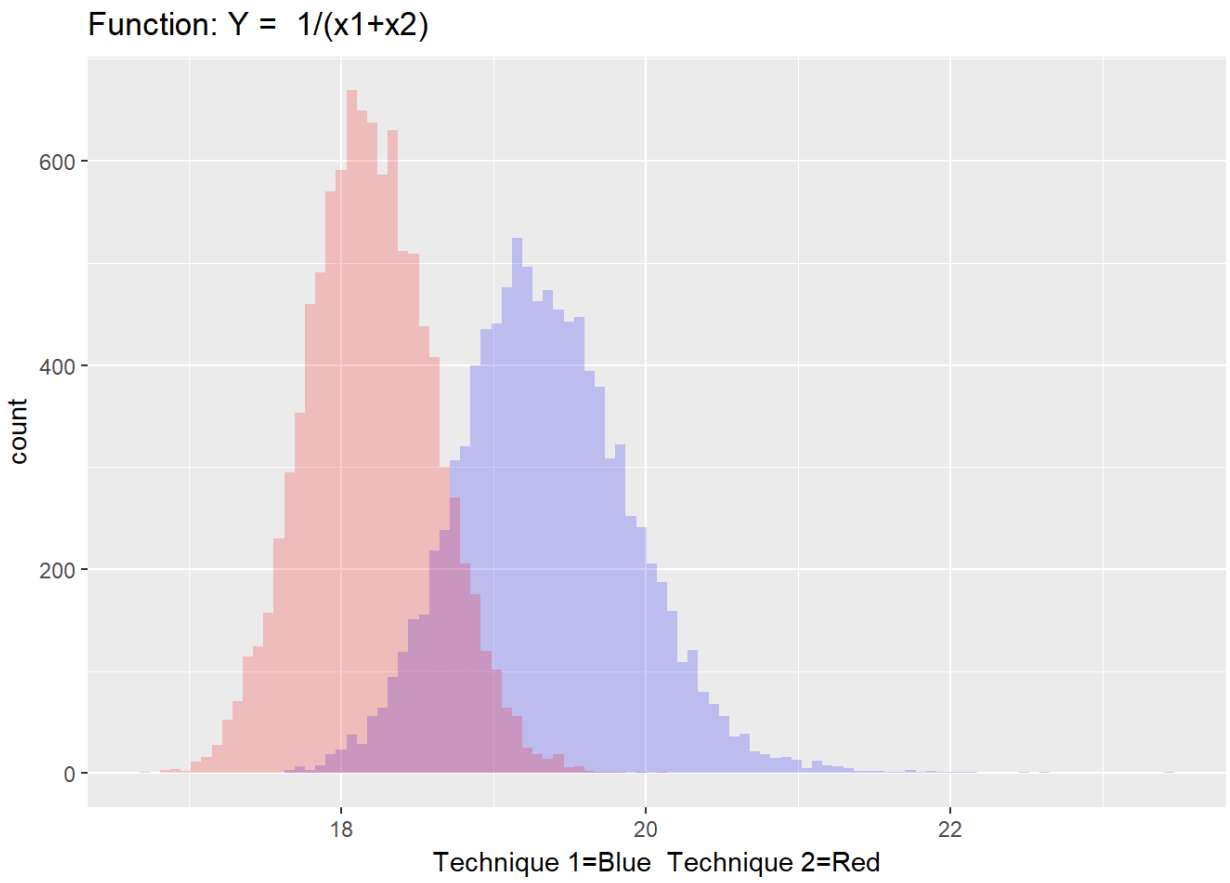


Figure A16. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

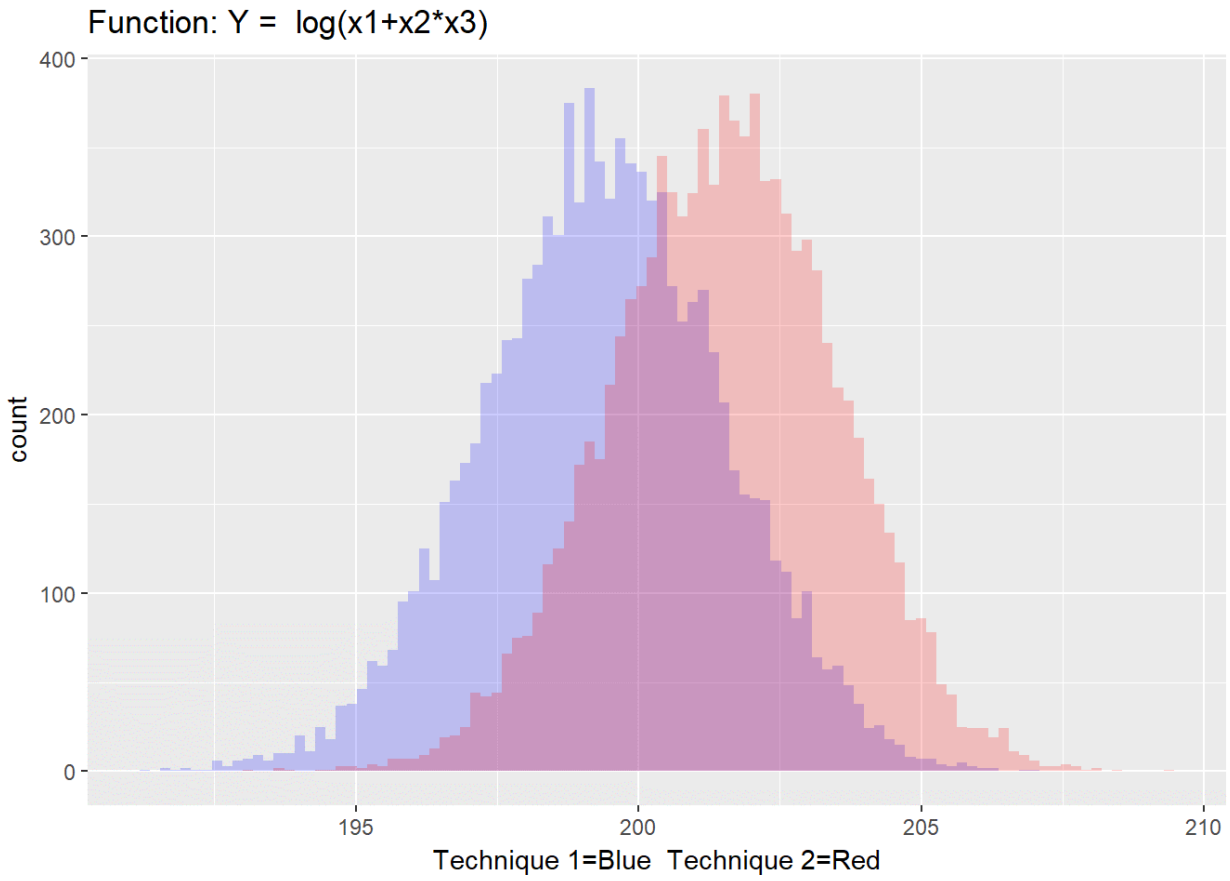


Figure A17. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

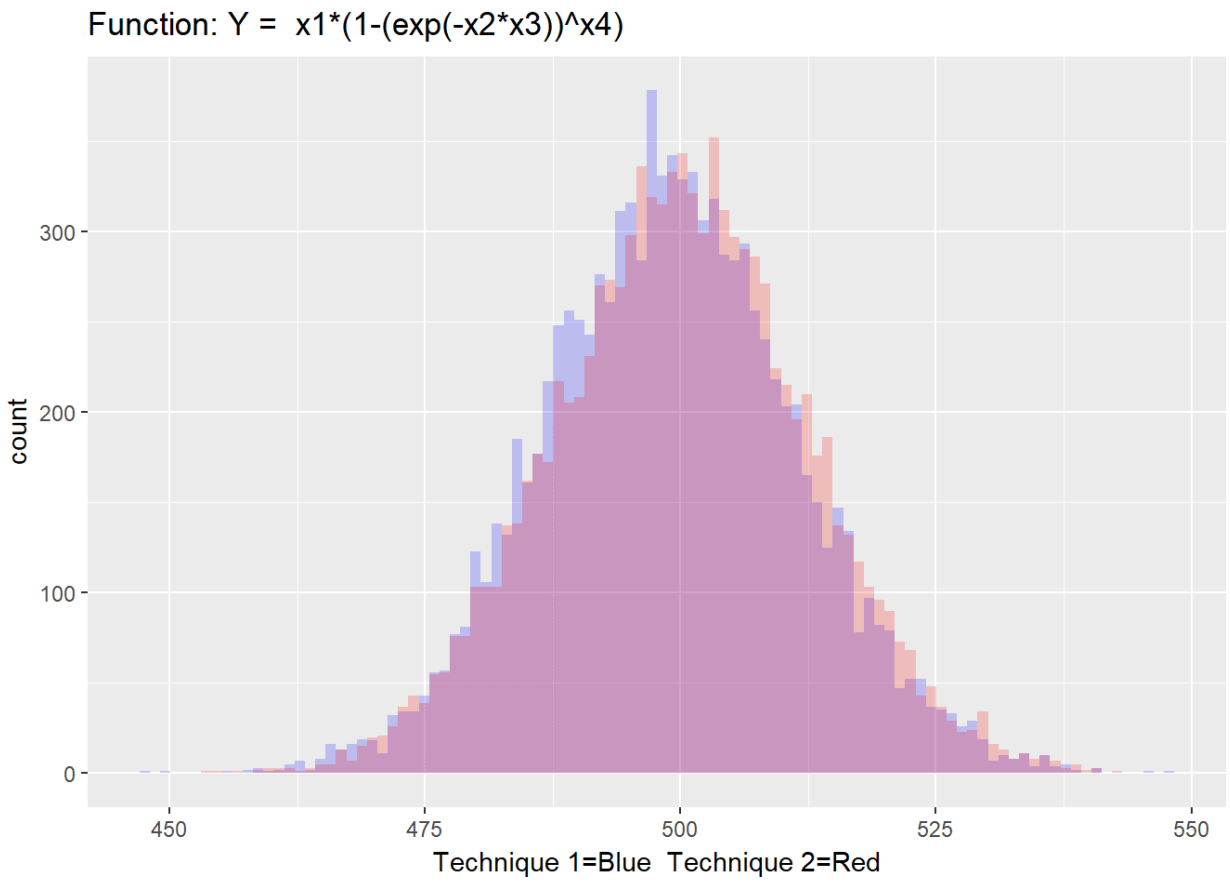


Figure A18. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

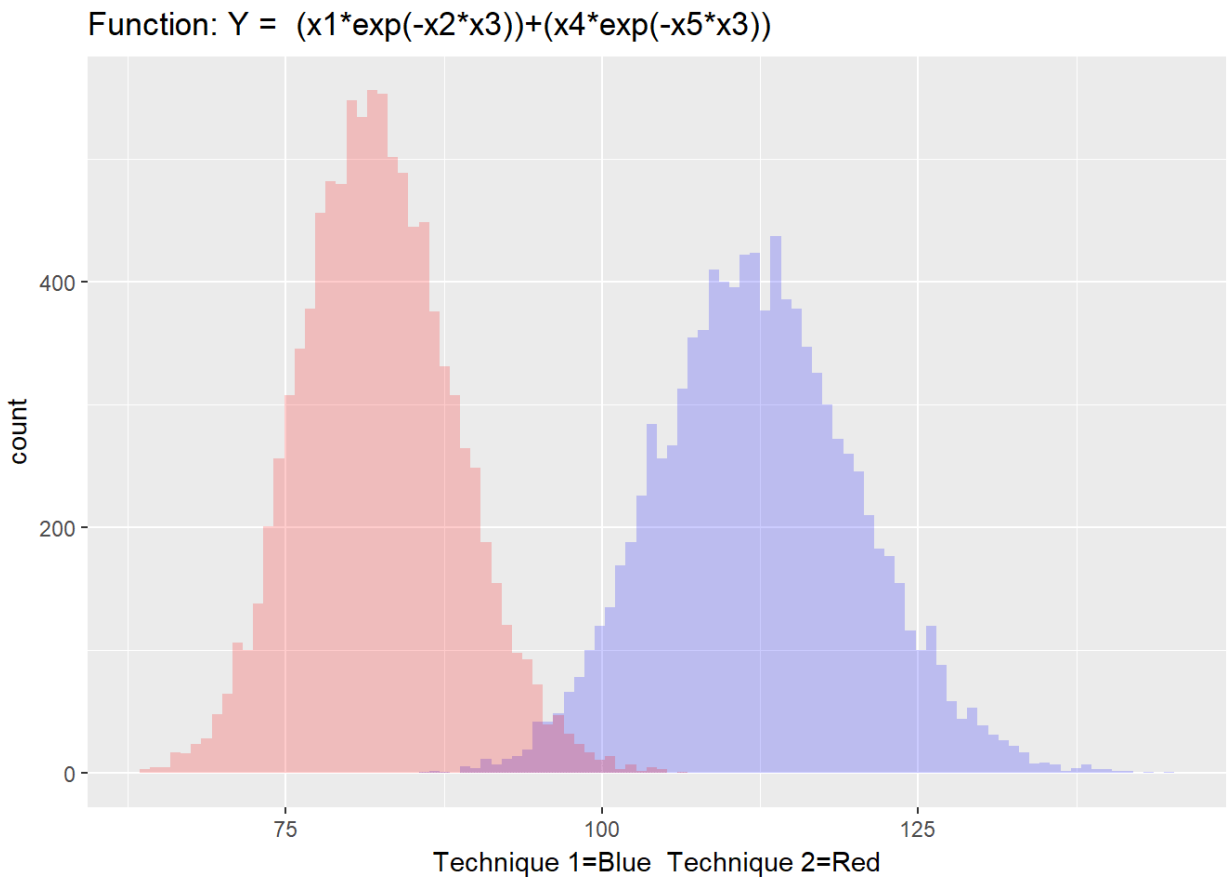


Figure A19. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

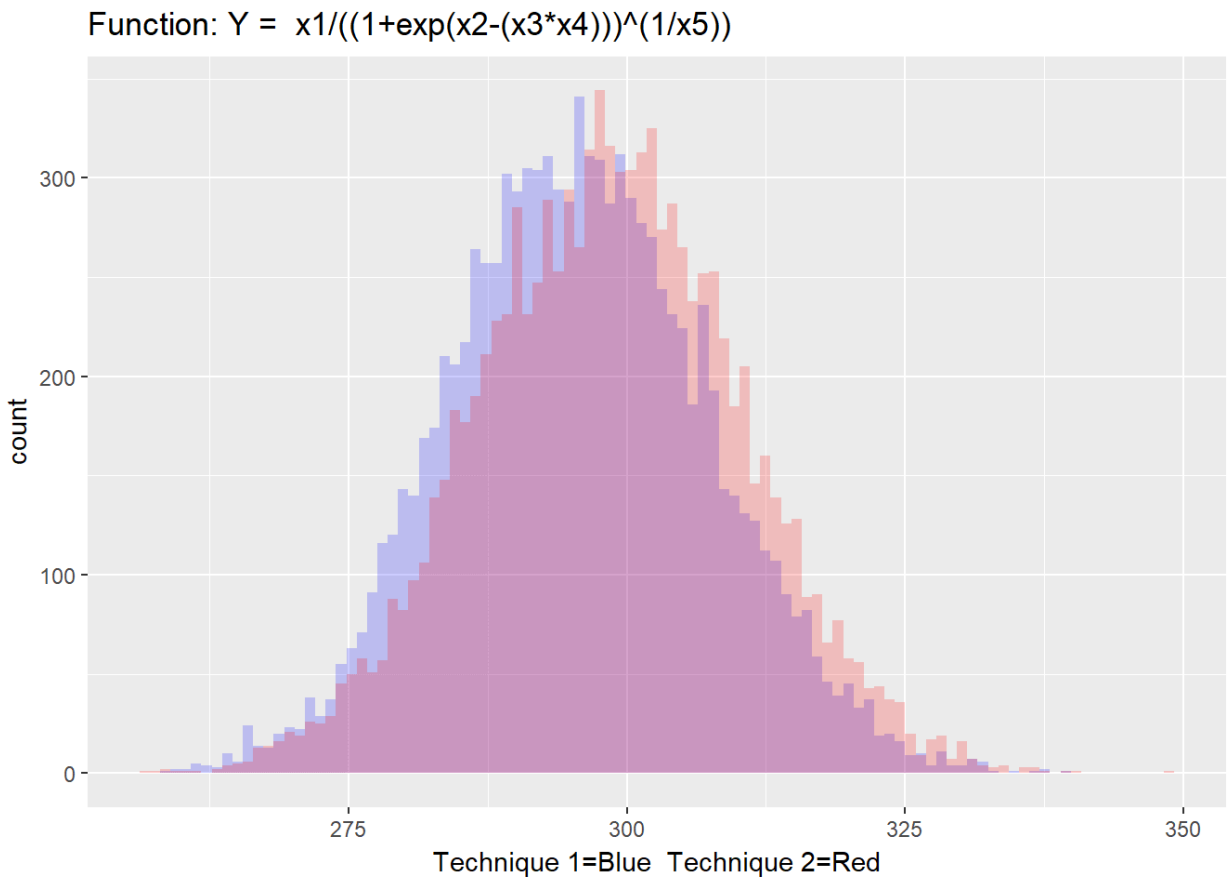


Figure A20. Distributions for Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

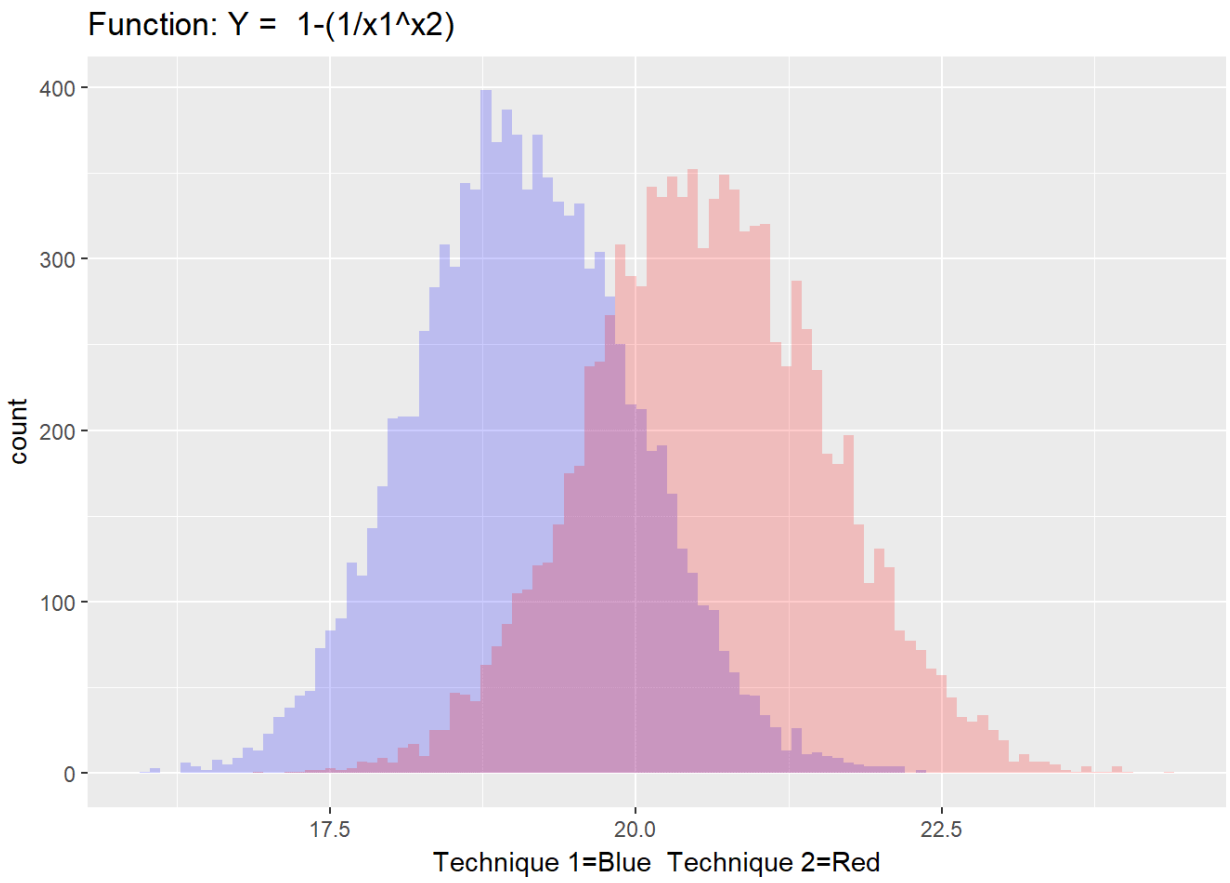


Figure A21. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

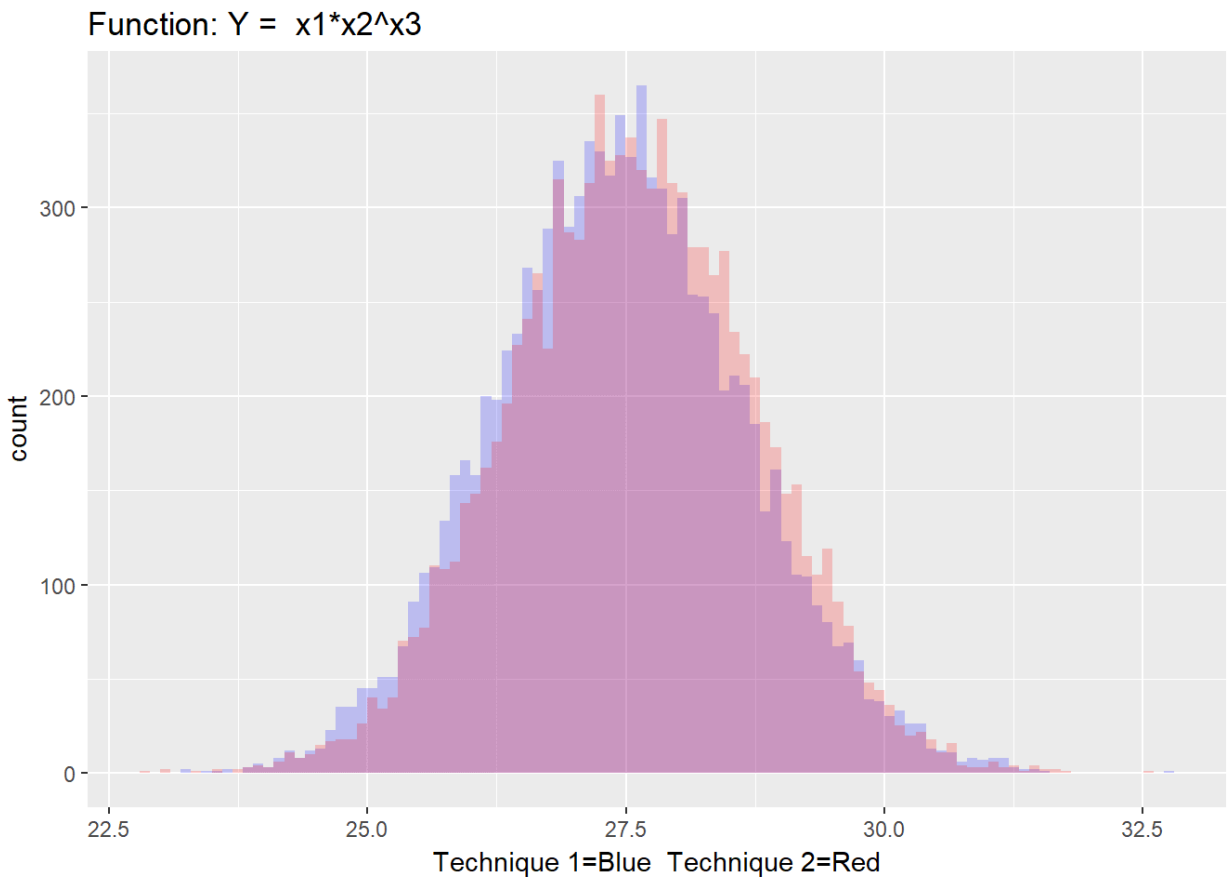


Figure A22. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

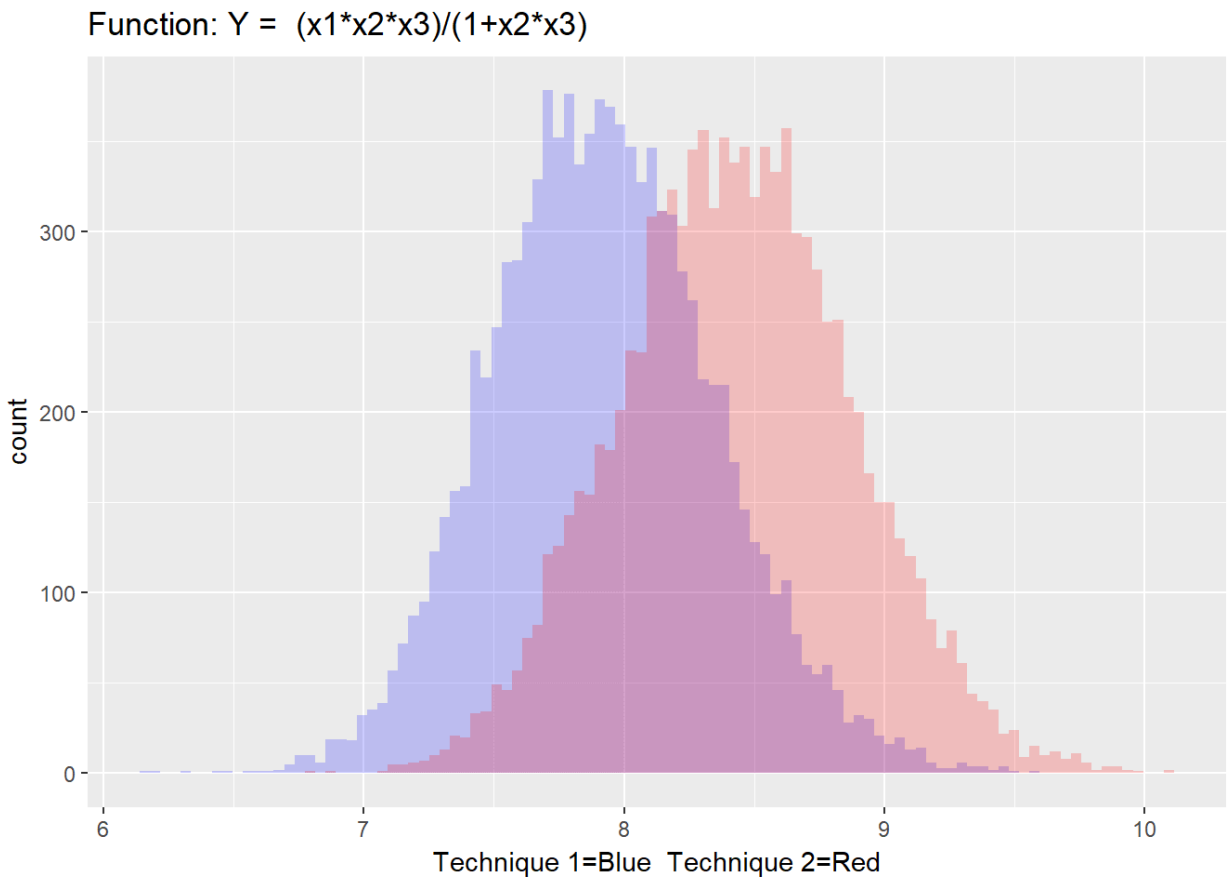


Figure A23. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

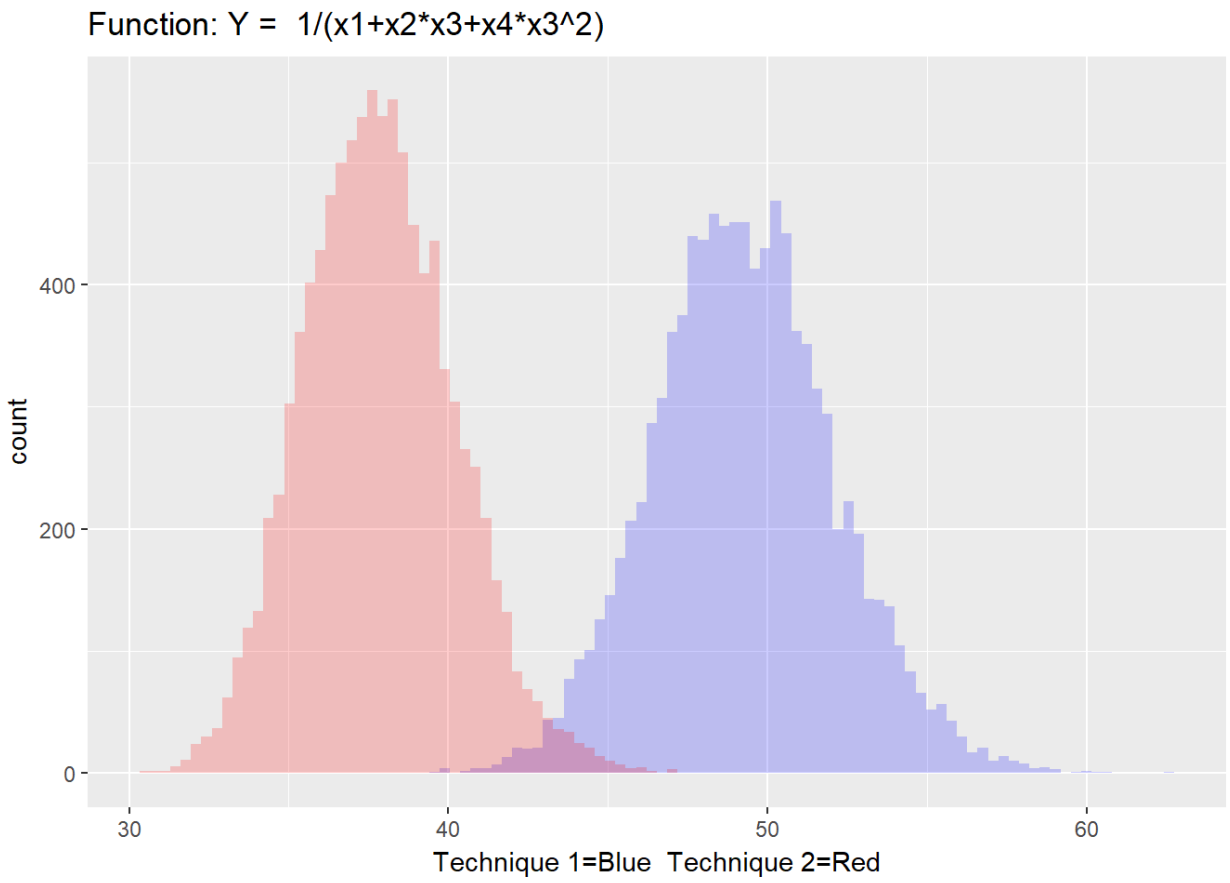


Figure A24. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

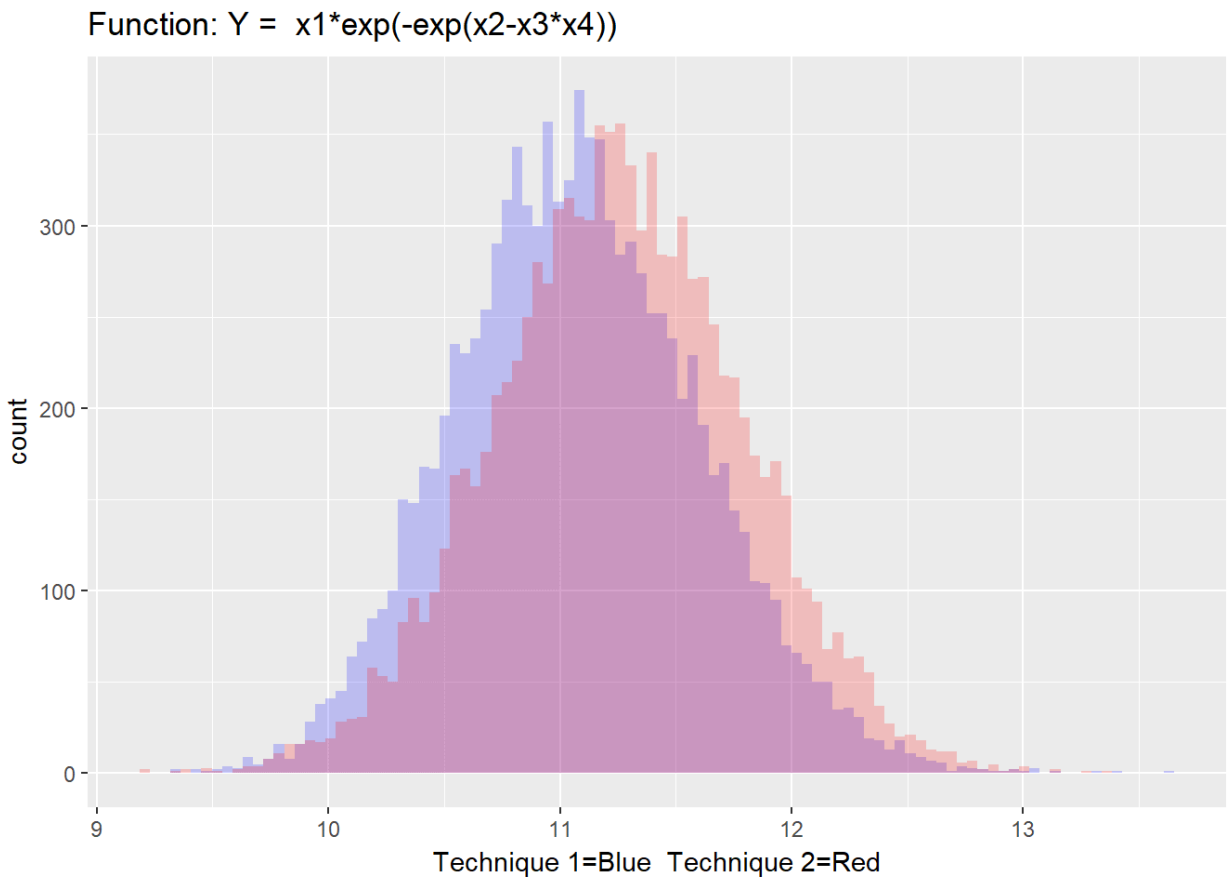


Figure A25. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

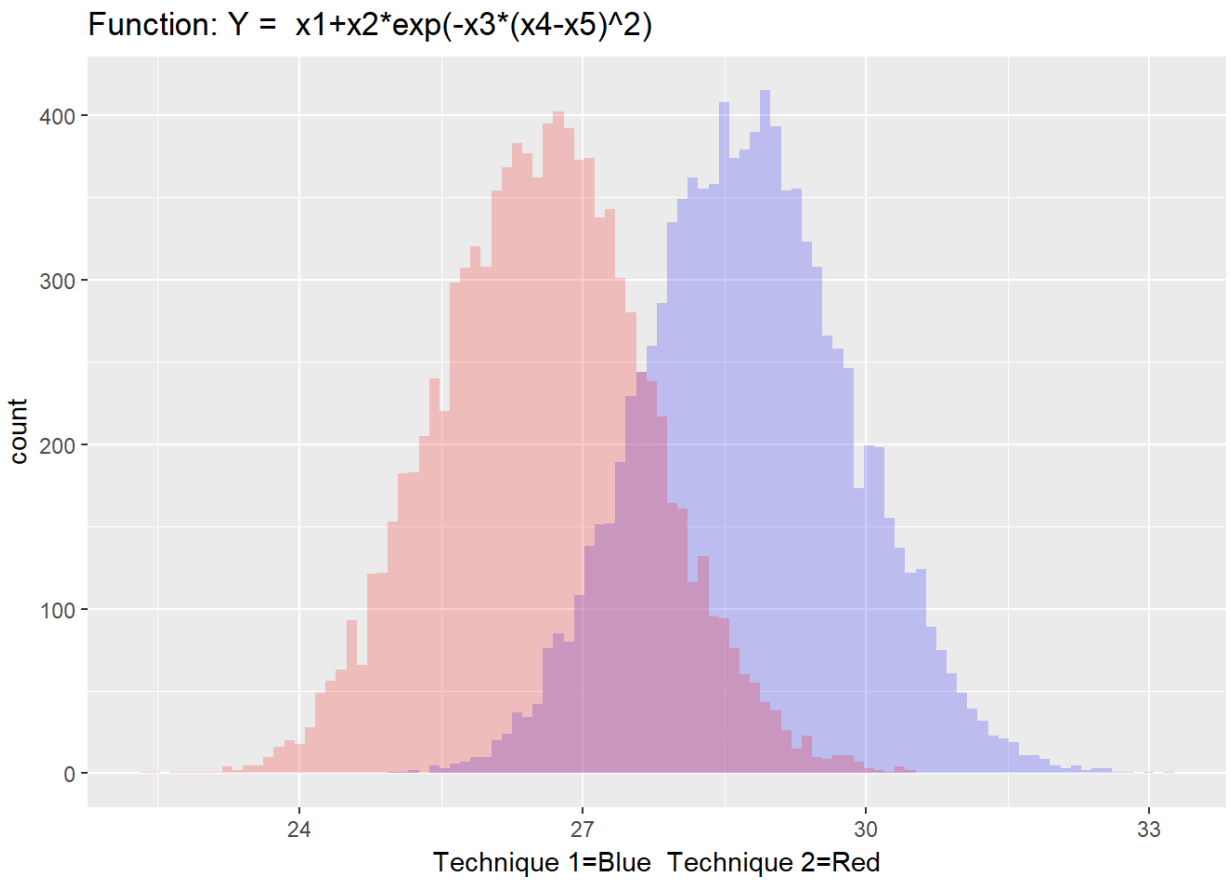


Figure A26. Distributions for Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

## APPENDIX B. EXAMPLE Q-Q PLOTS

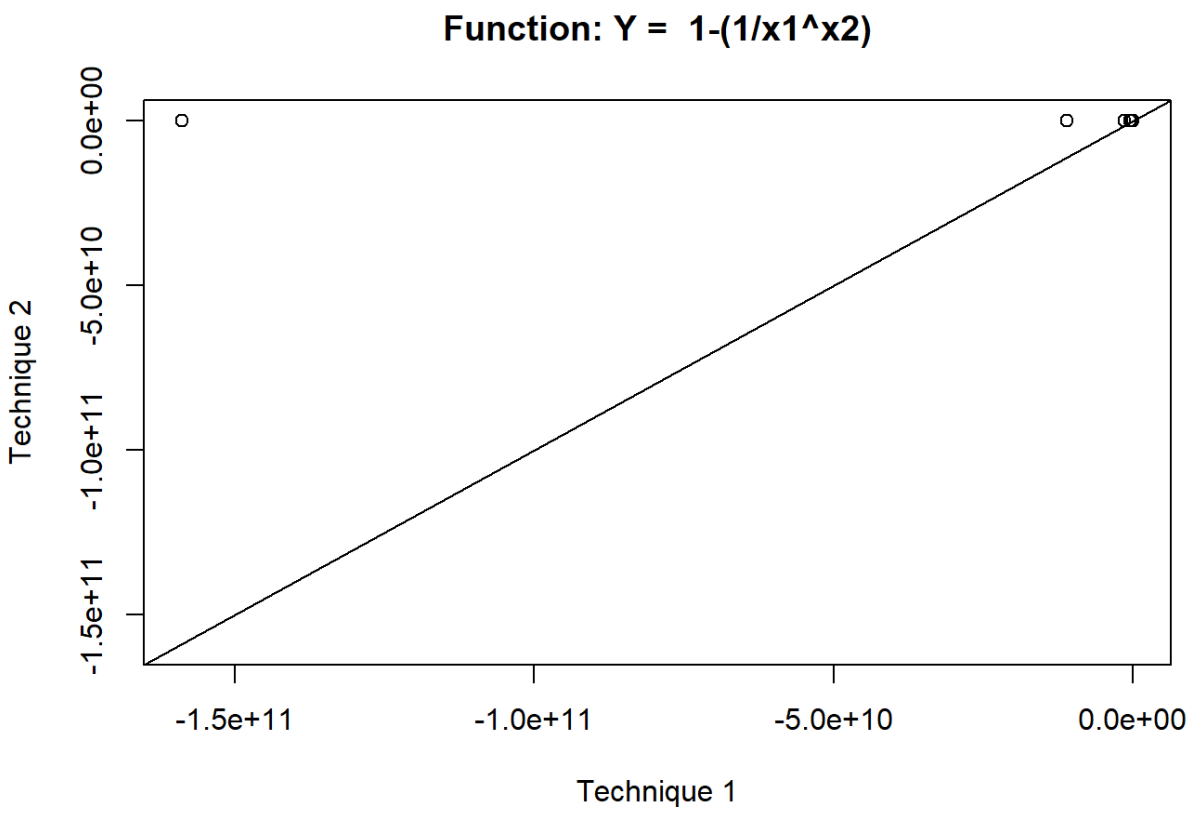


Figure B1. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

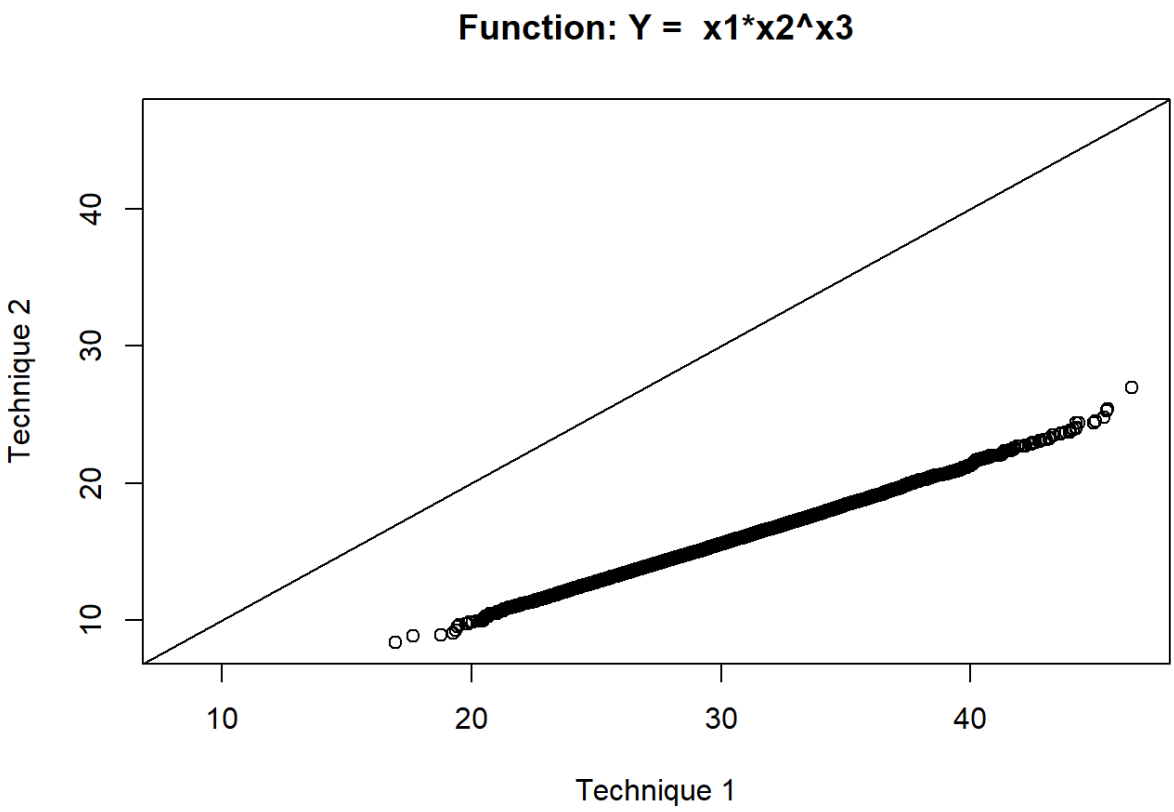


Figure B2. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

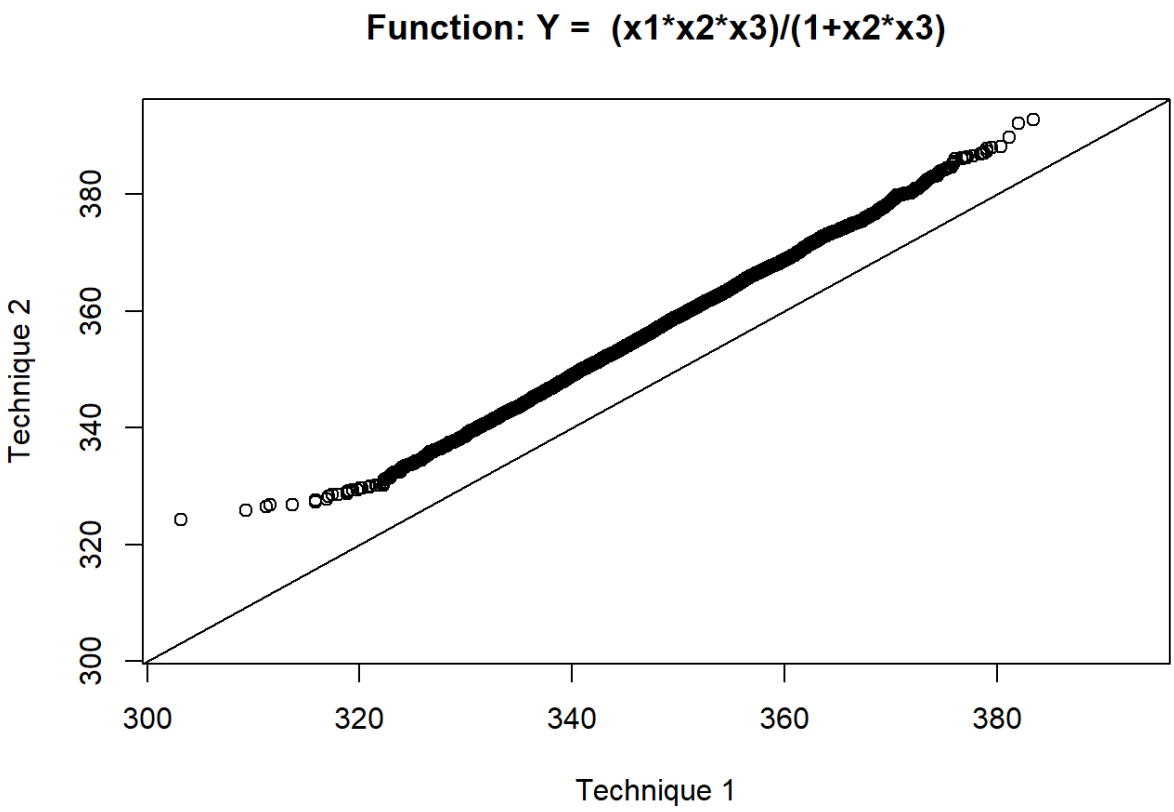


Figure B3. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

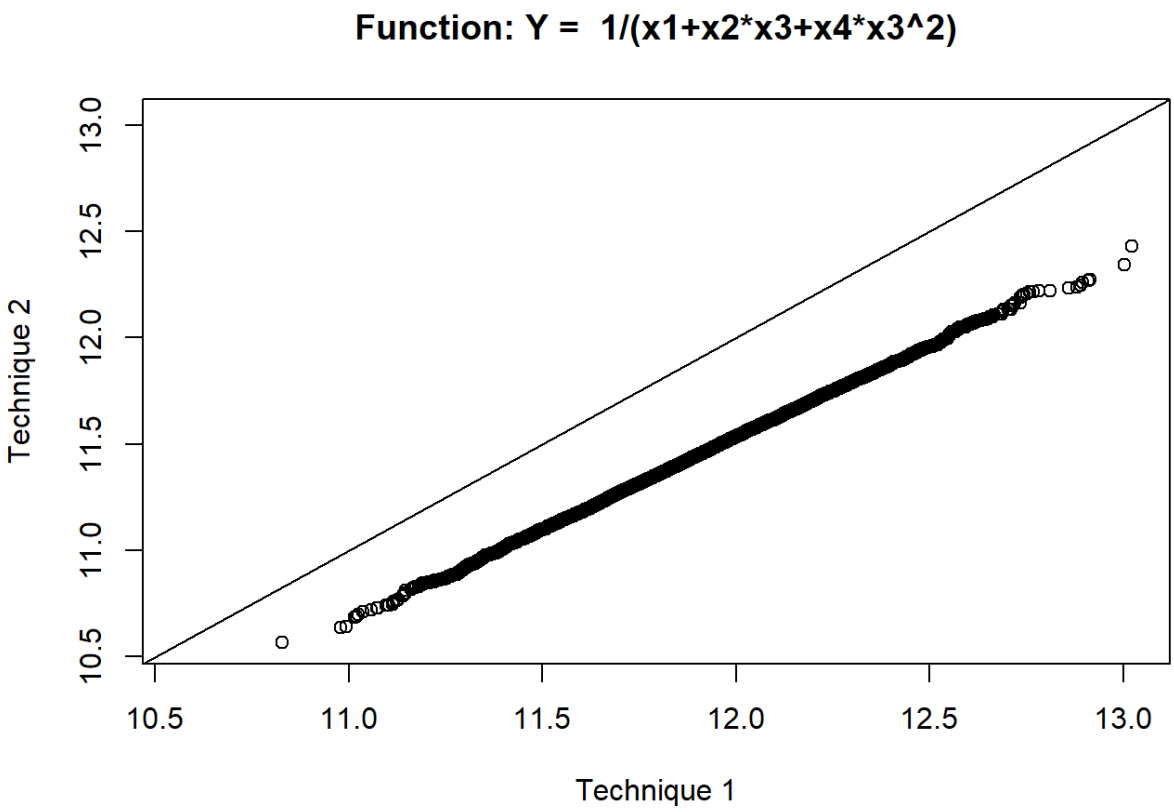


Figure B4. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

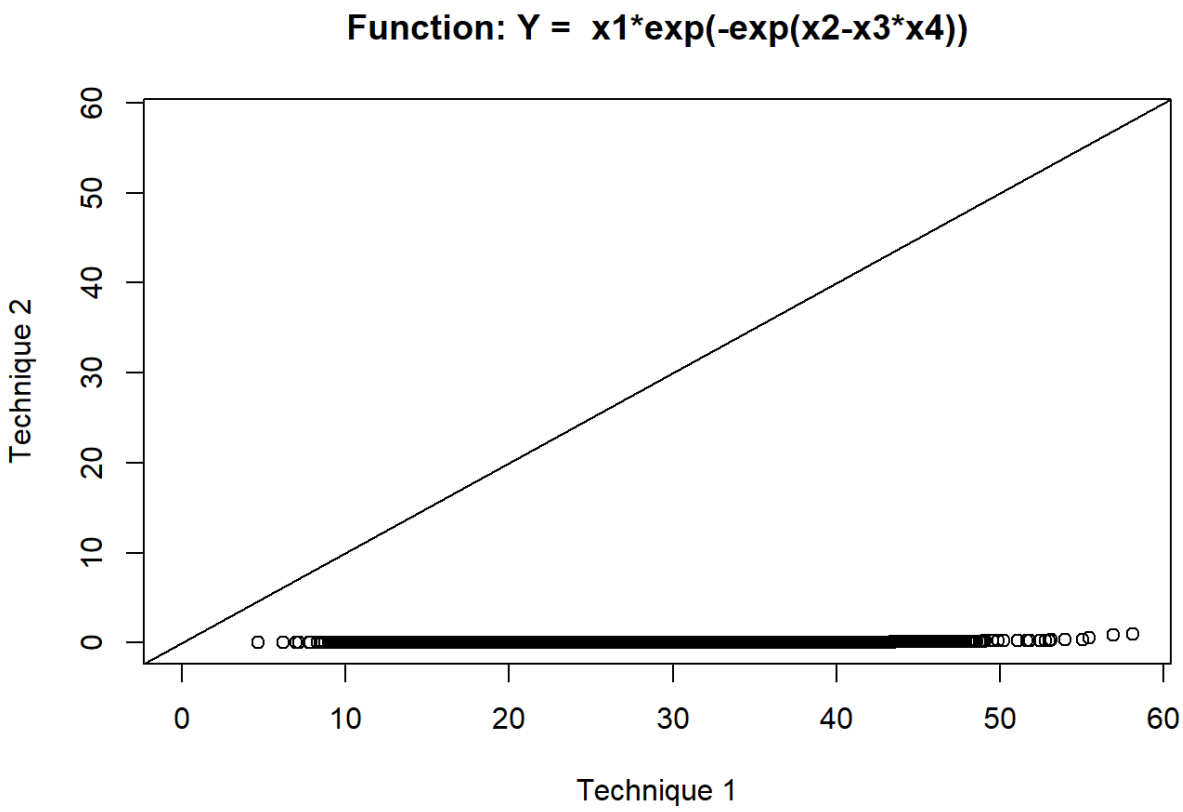


Figure B5. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

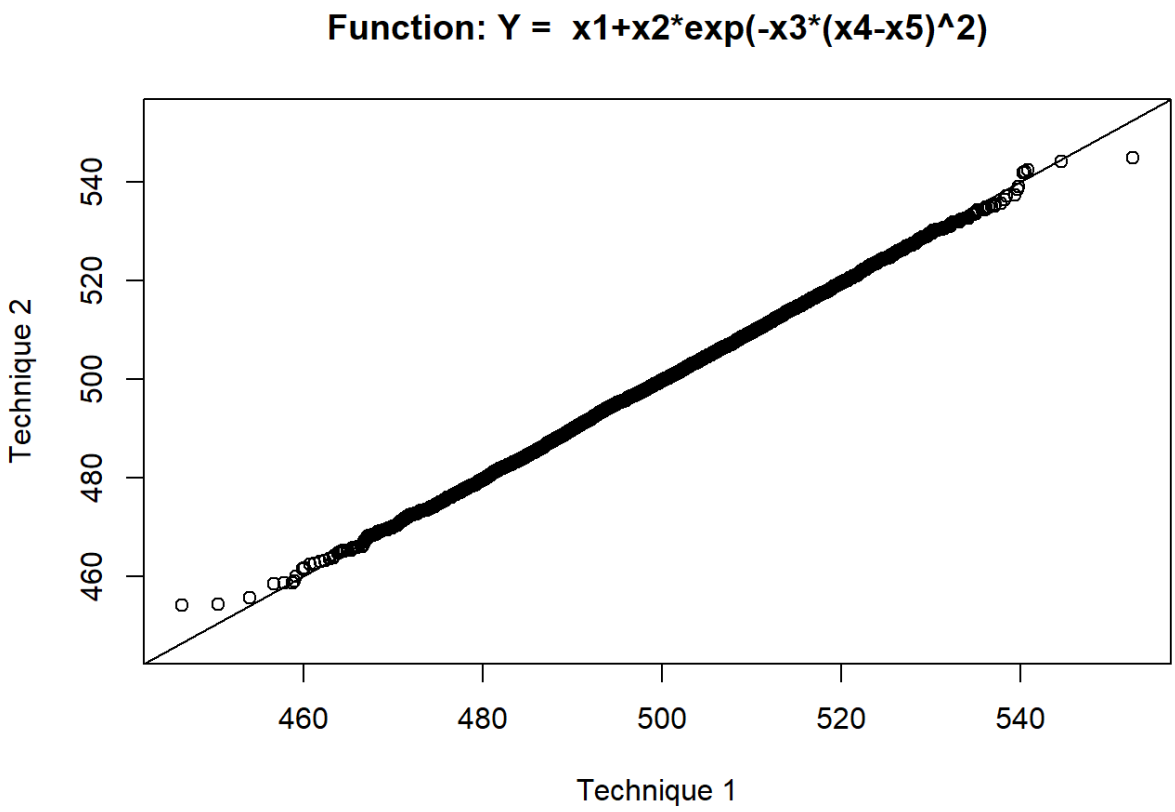


Figure B6. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

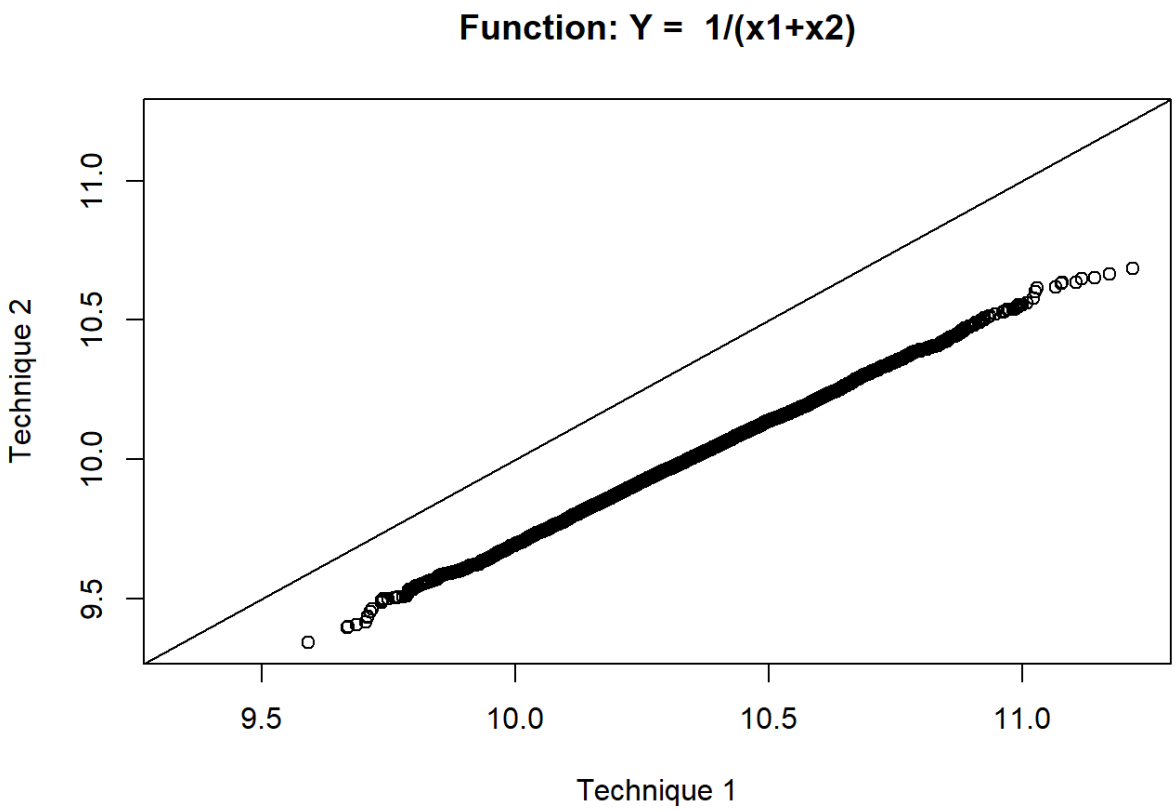


Figure B7. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

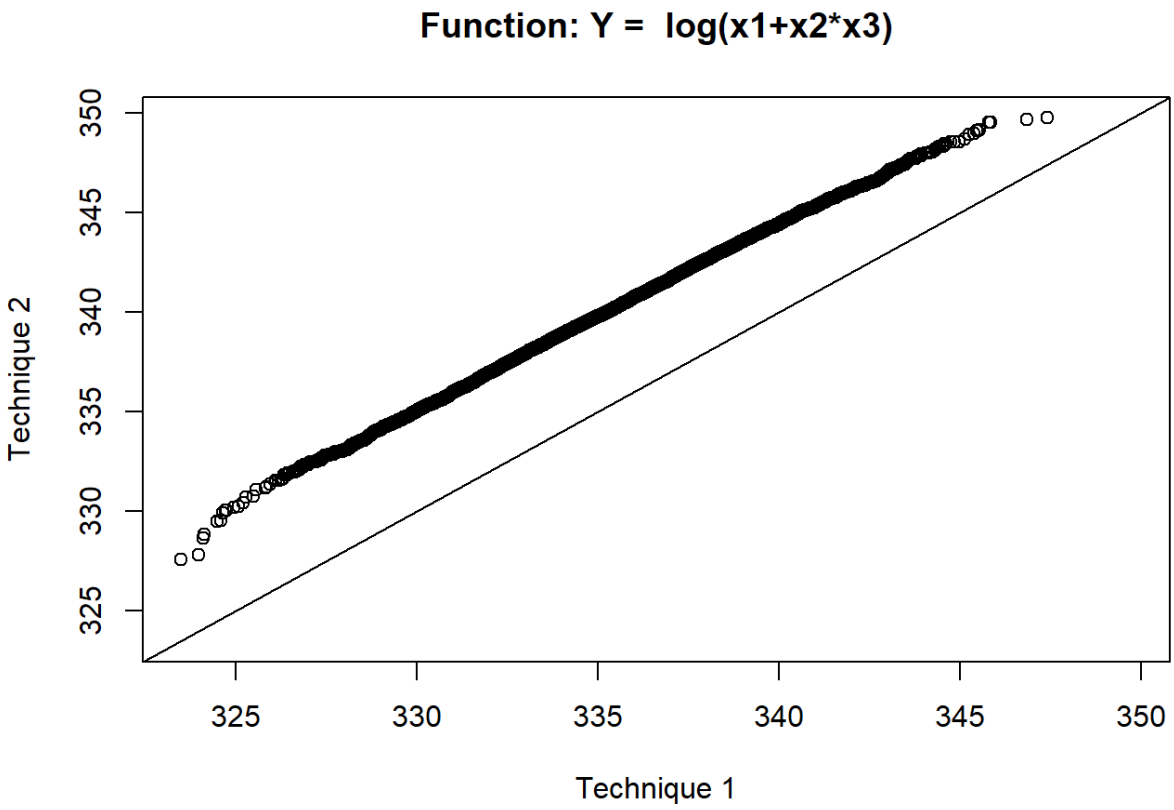


Figure B8. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

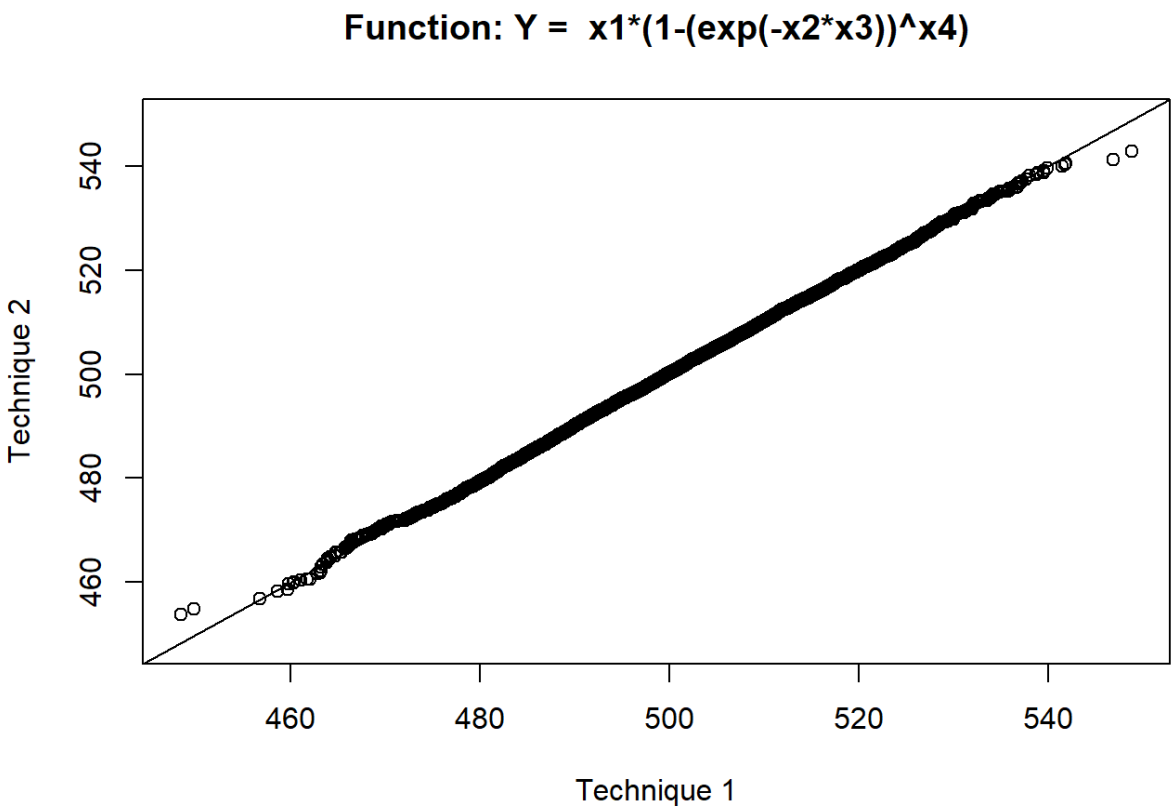


Figure B9. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

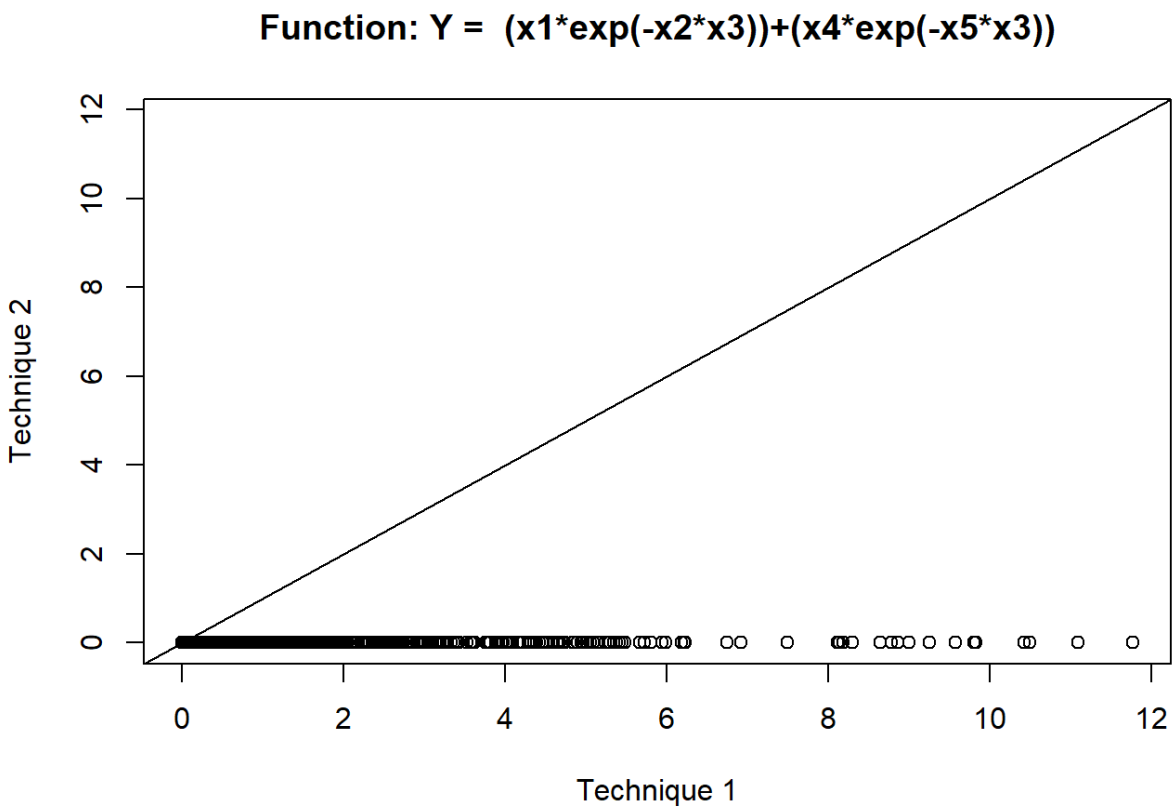


Figure B10. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

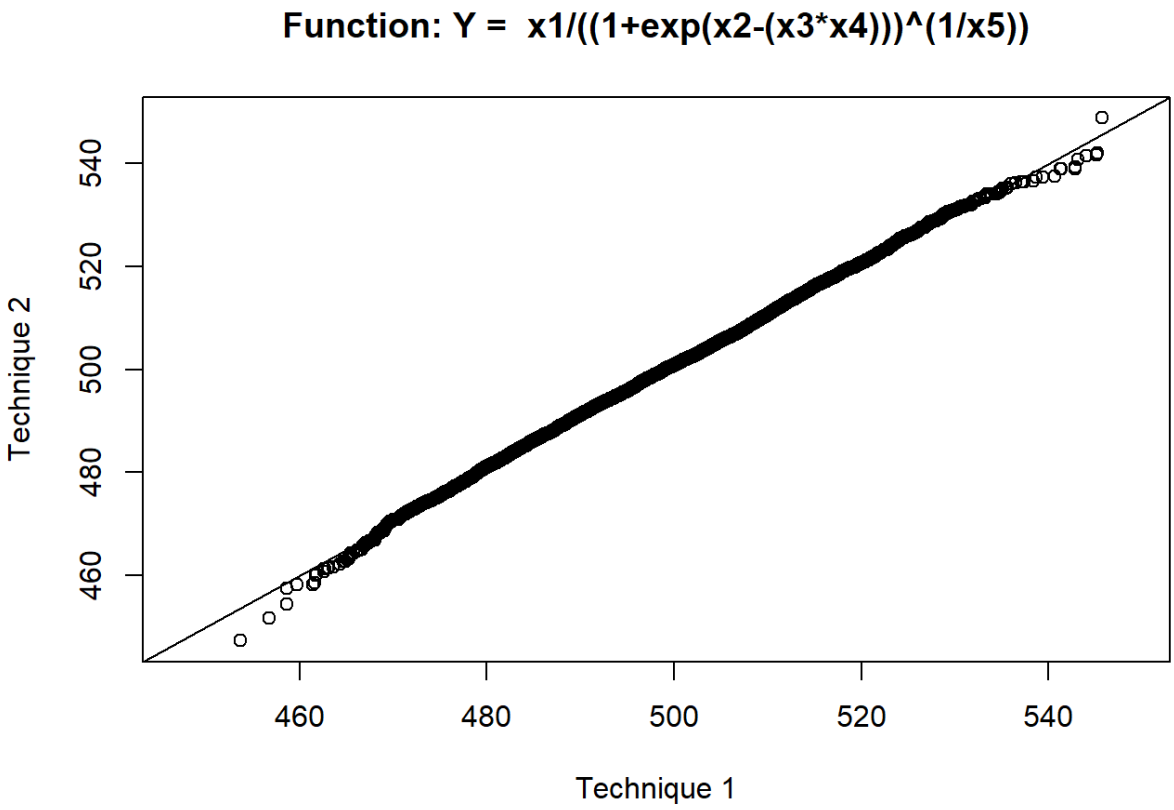


Figure B11. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

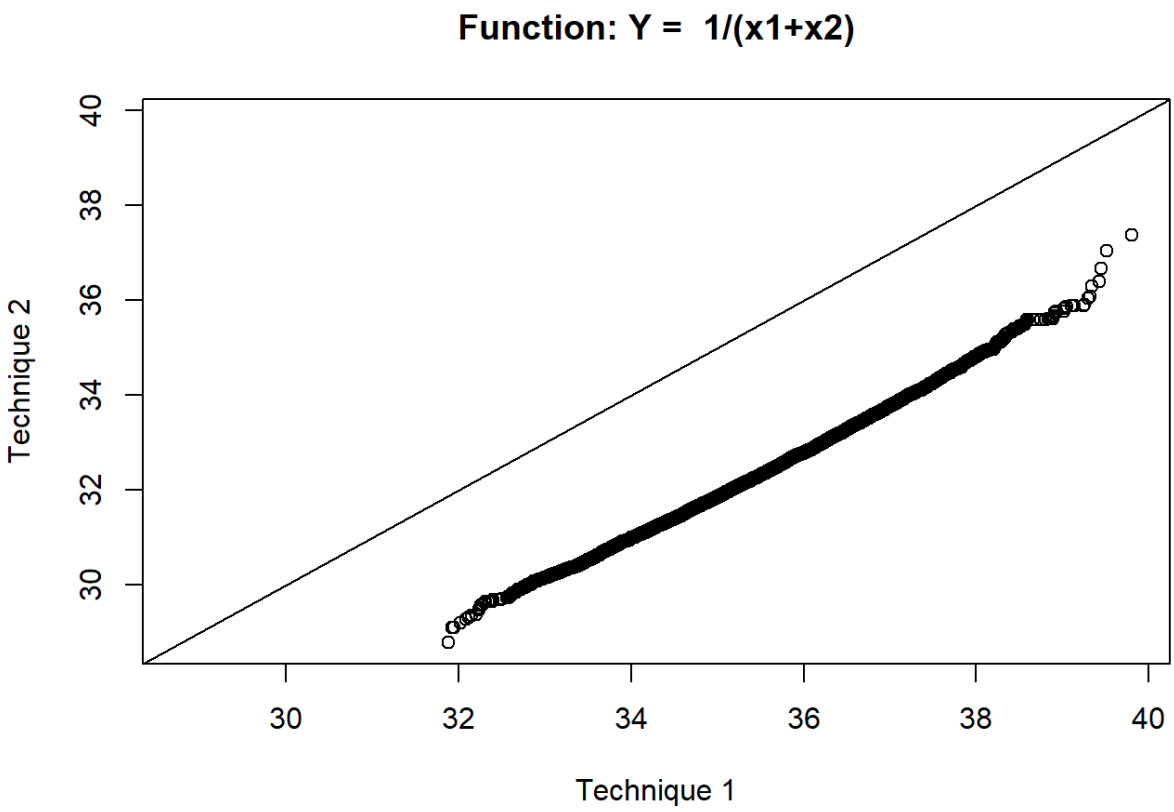


Figure B12. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

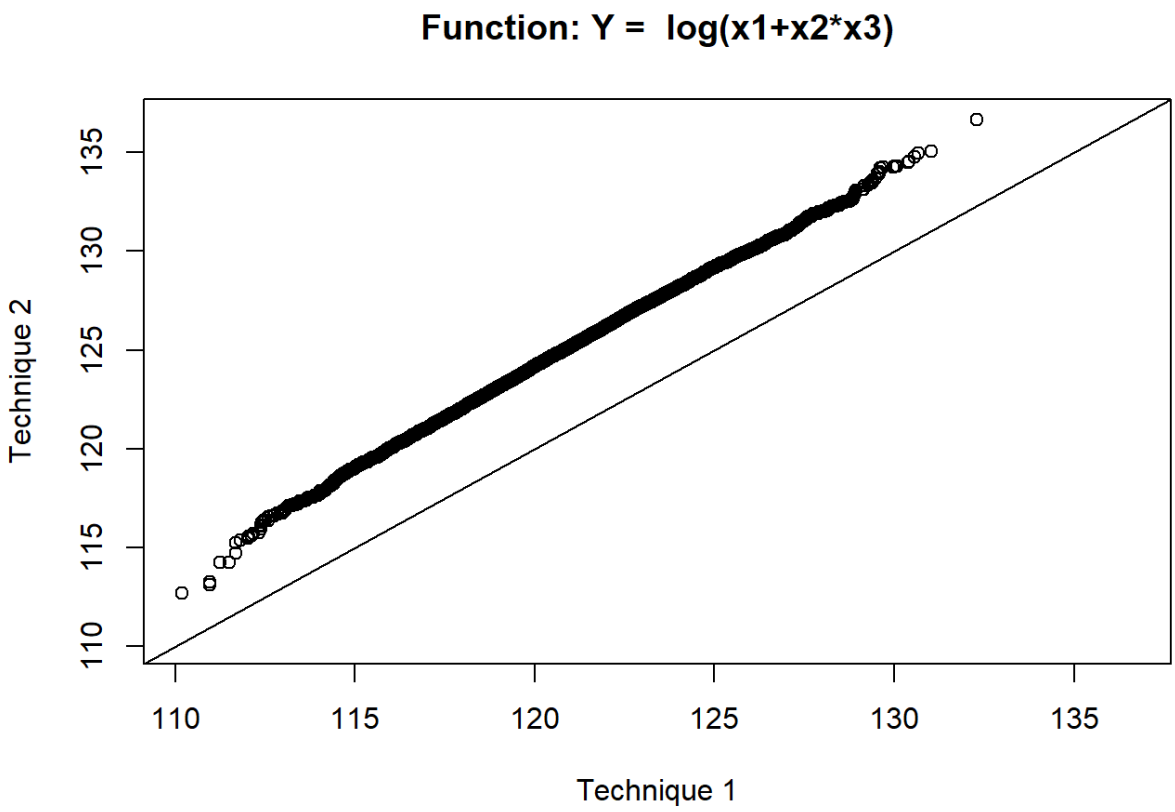


Figure B13. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

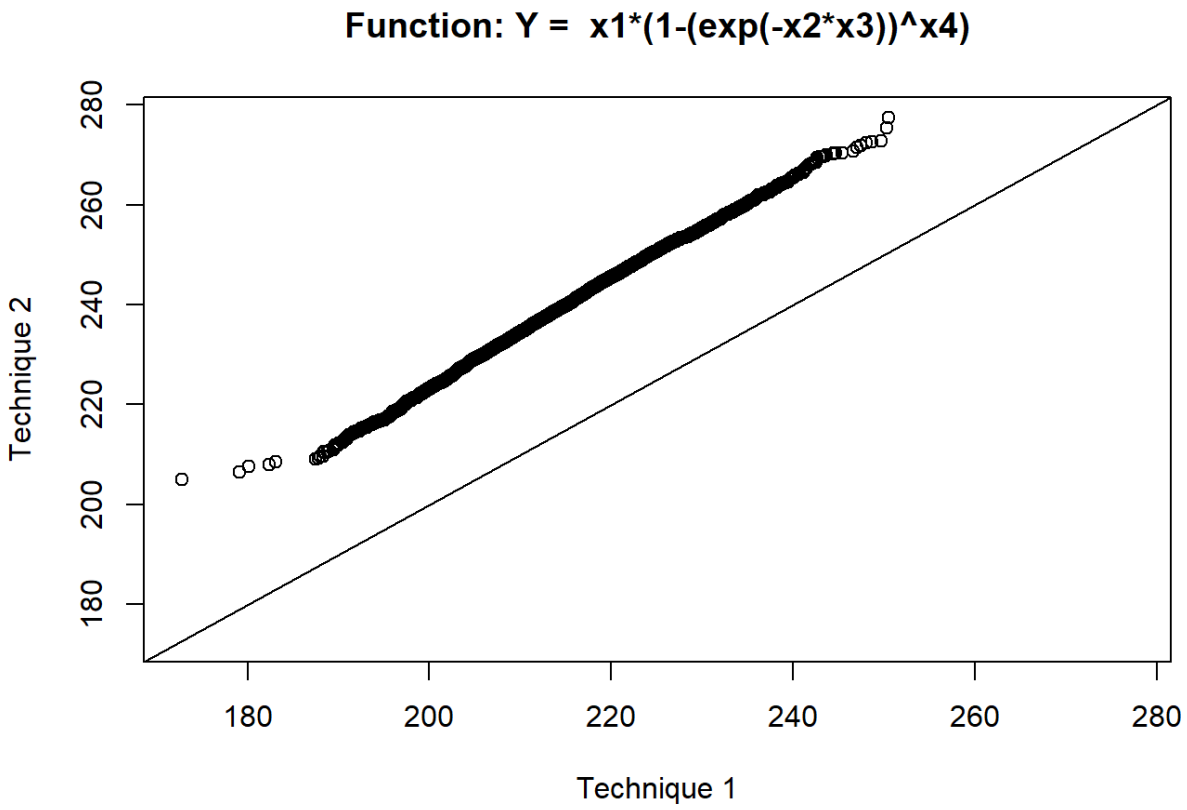


Figure B14. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

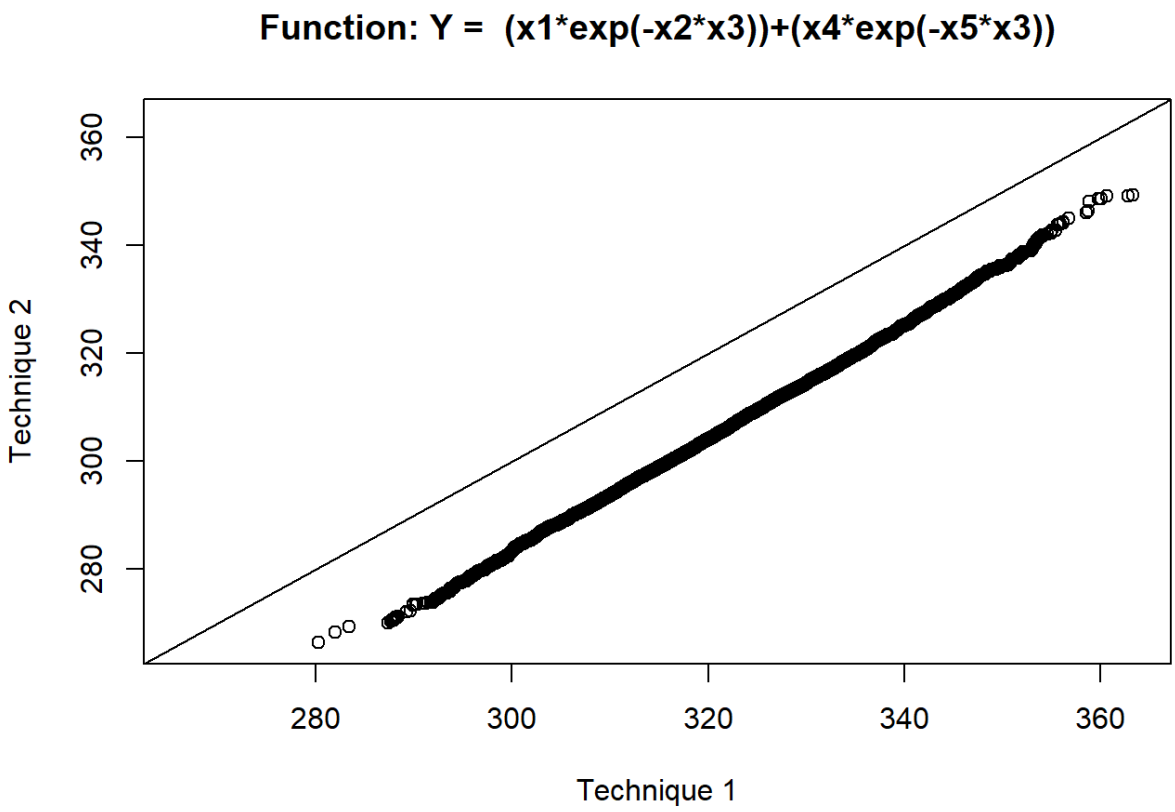


Figure B15. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

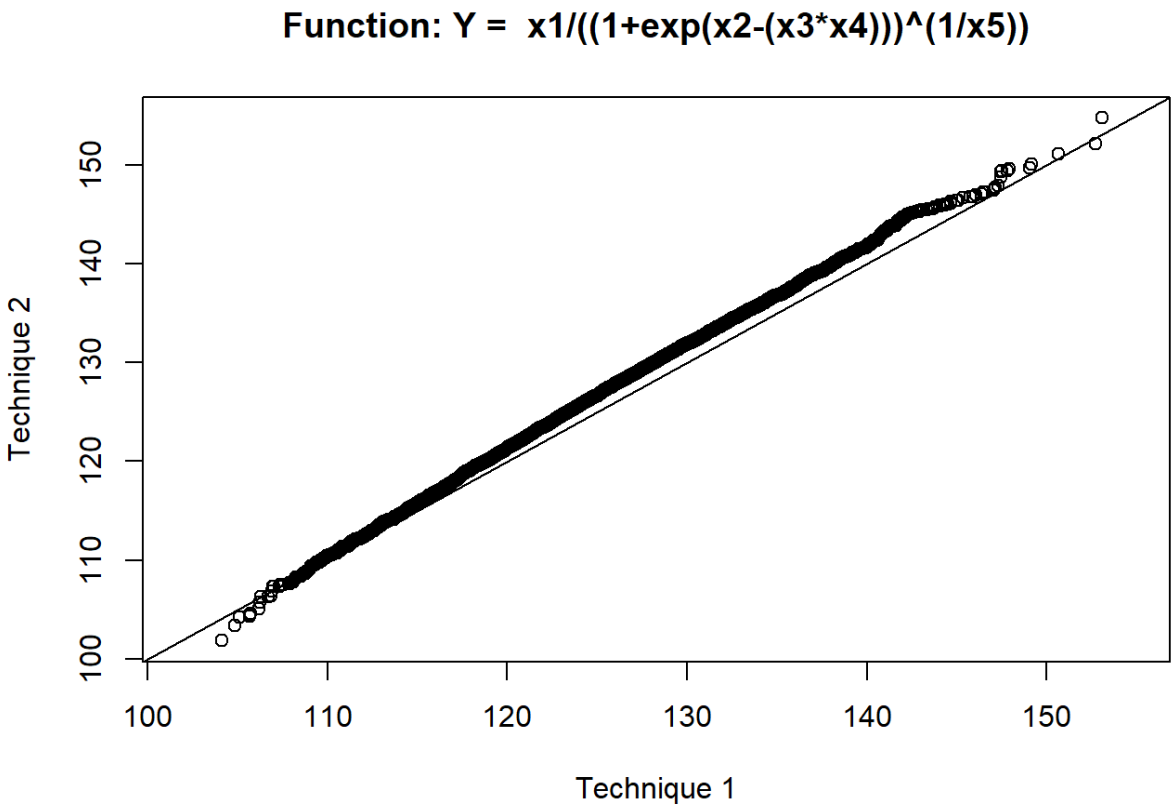


Figure B16. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

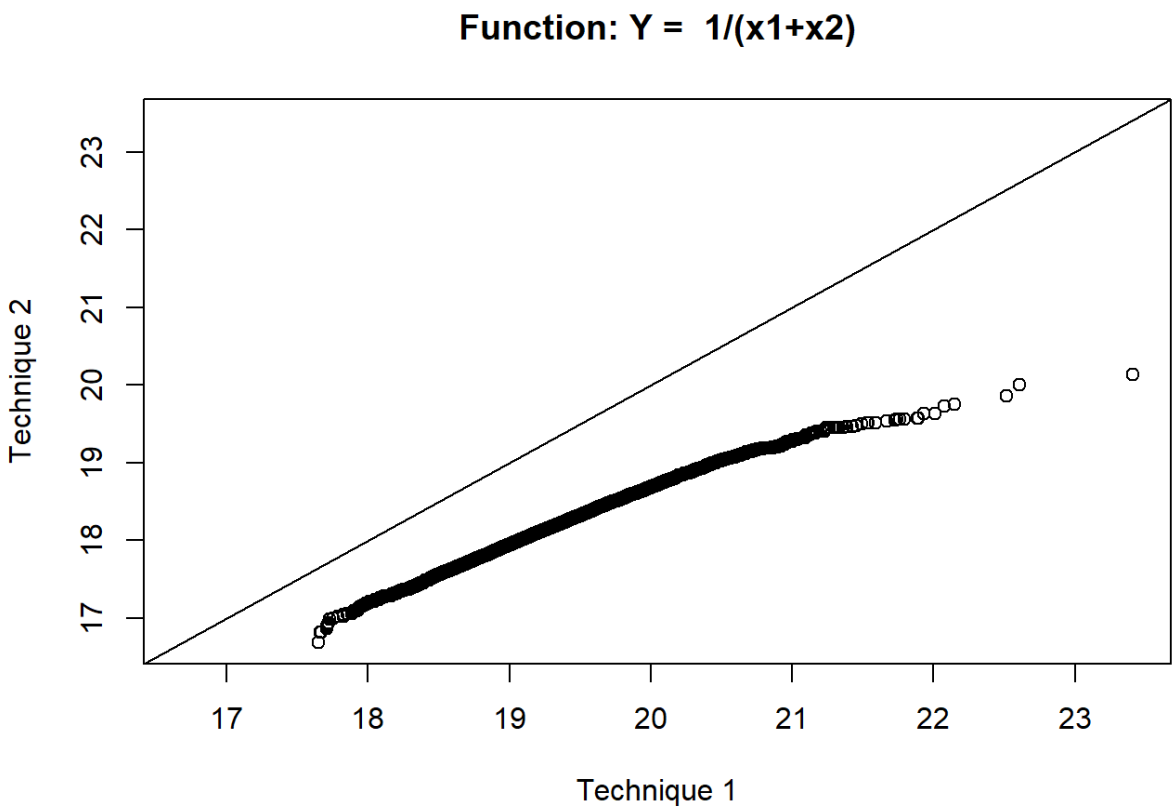


Figure B17. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

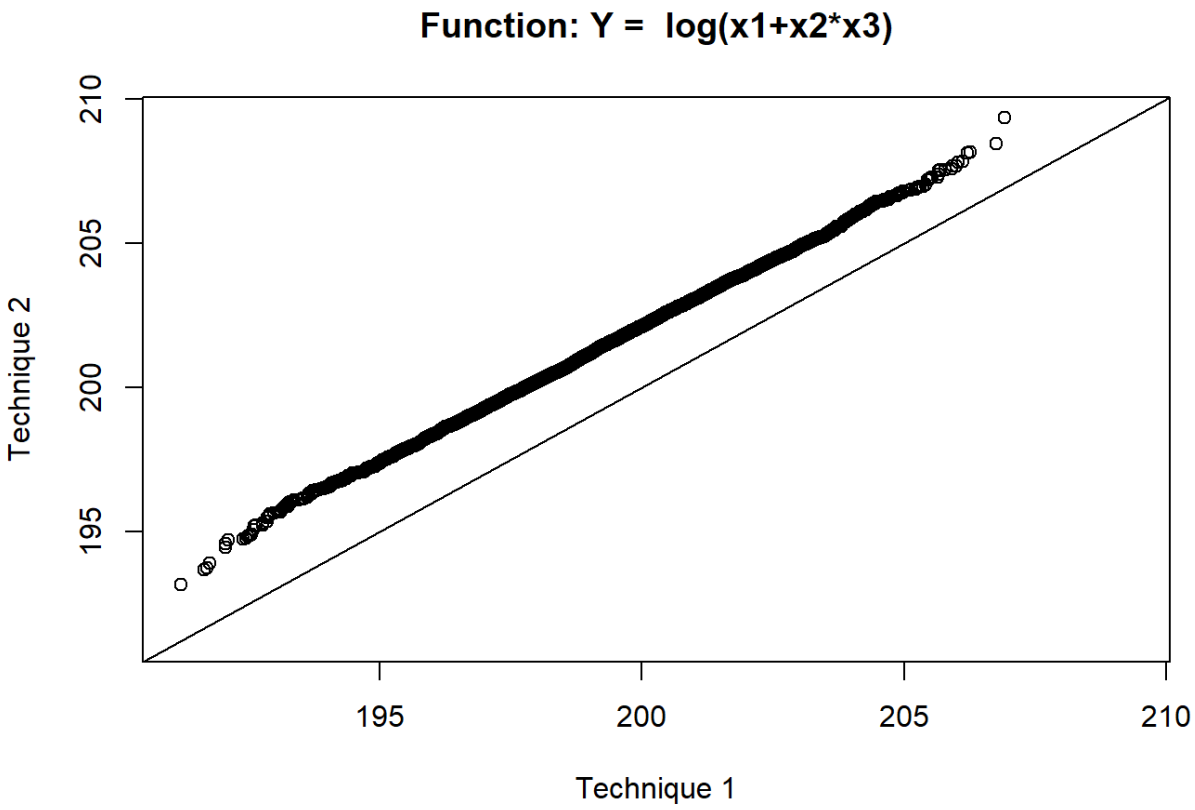


Figure B18. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

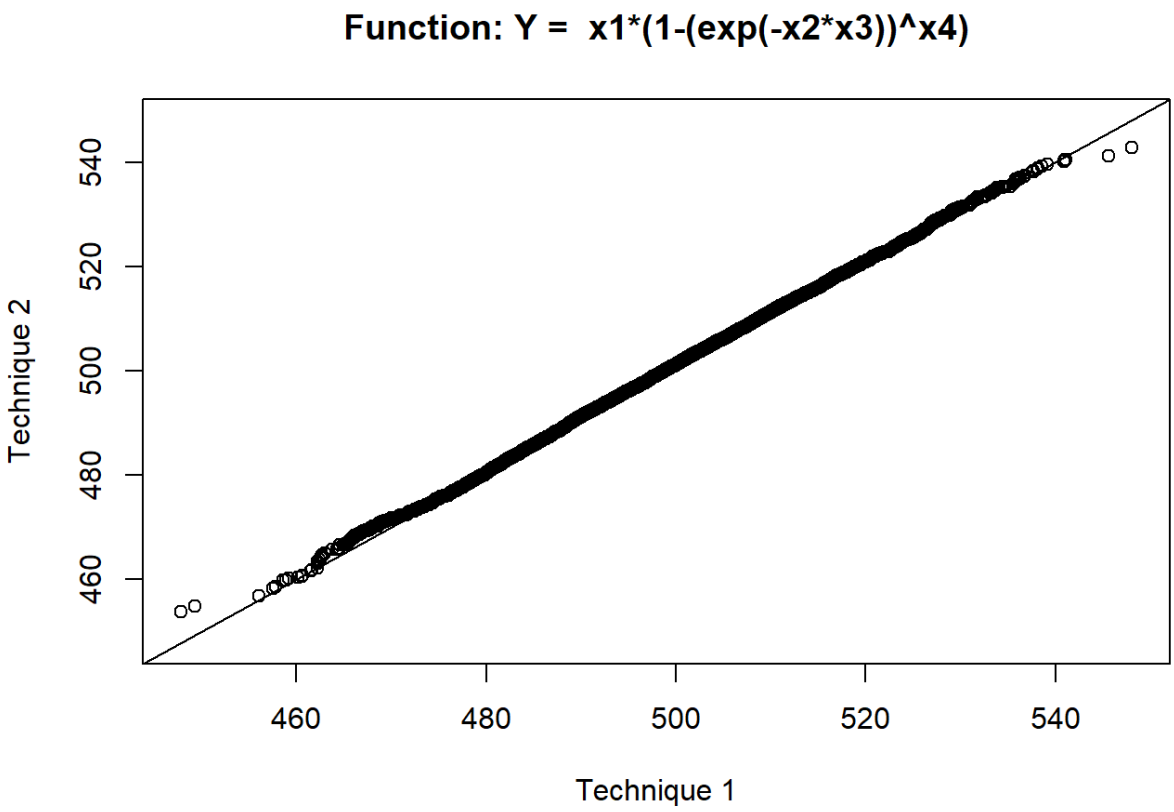


Figure B19. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

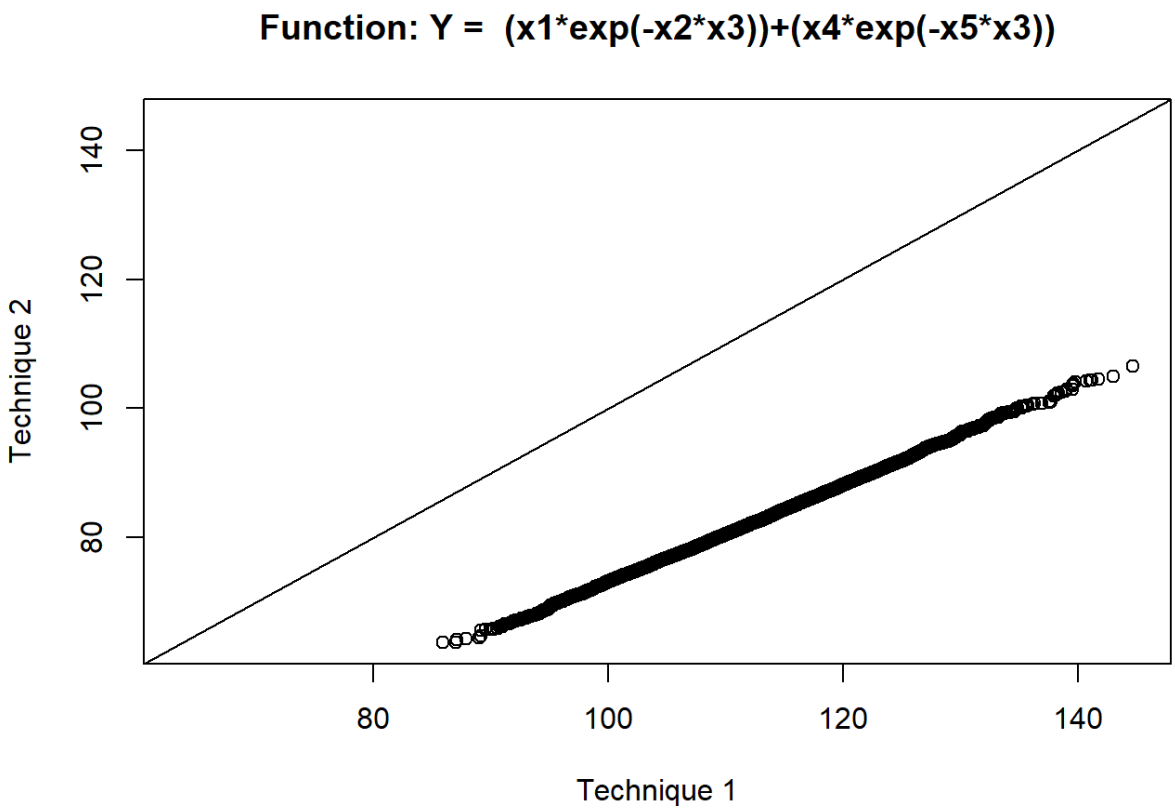


Figure B20. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

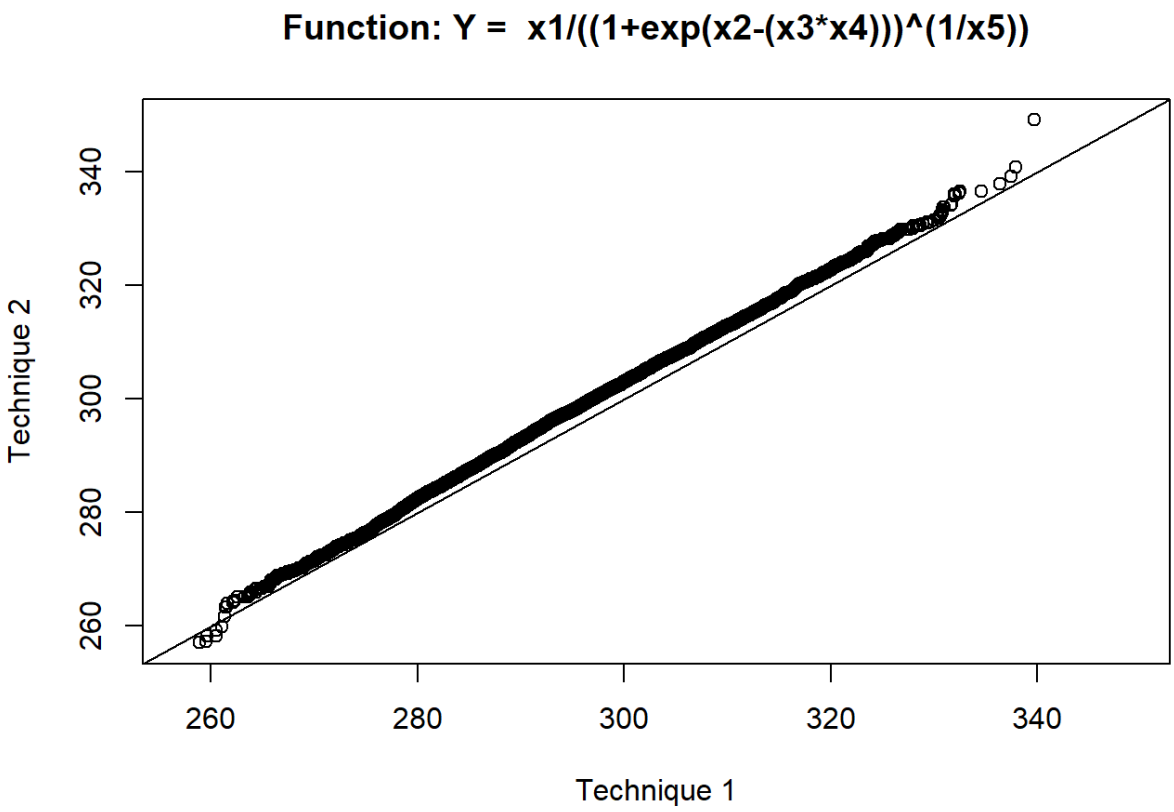


Figure B21. Q-Q plot of Technique 1 vs Technique 2 using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

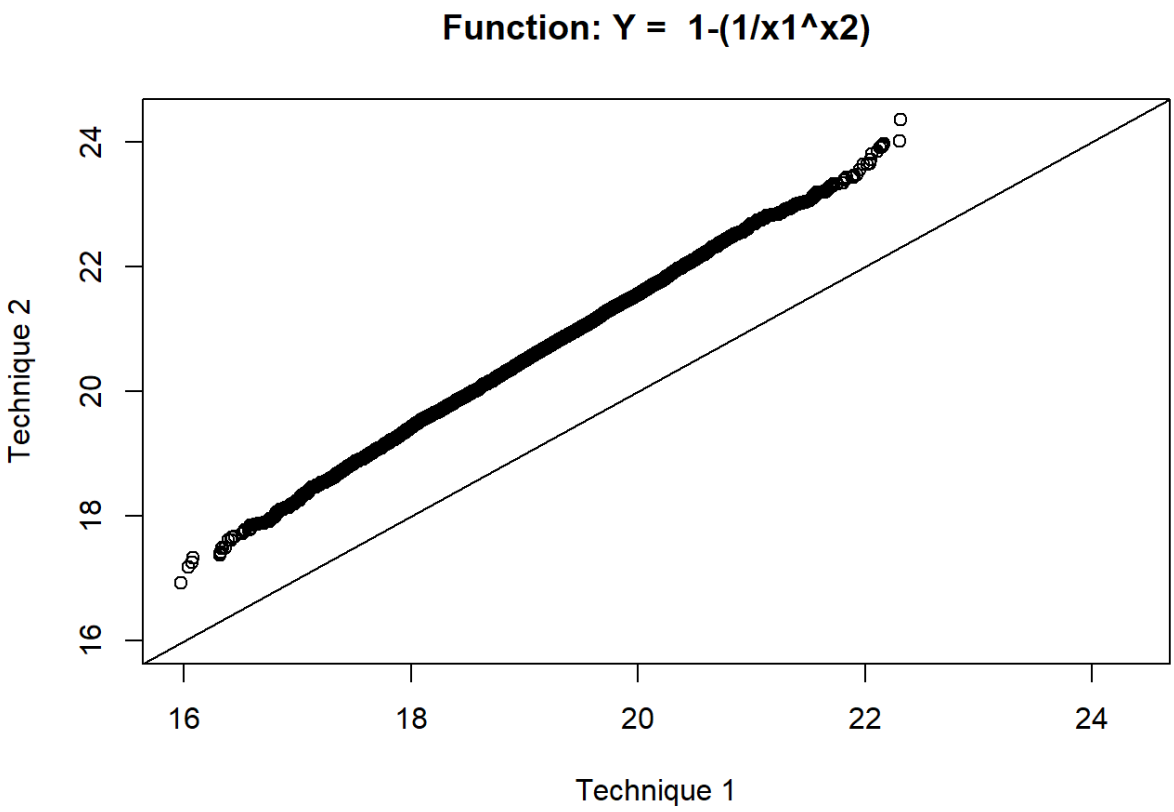


Figure B22. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

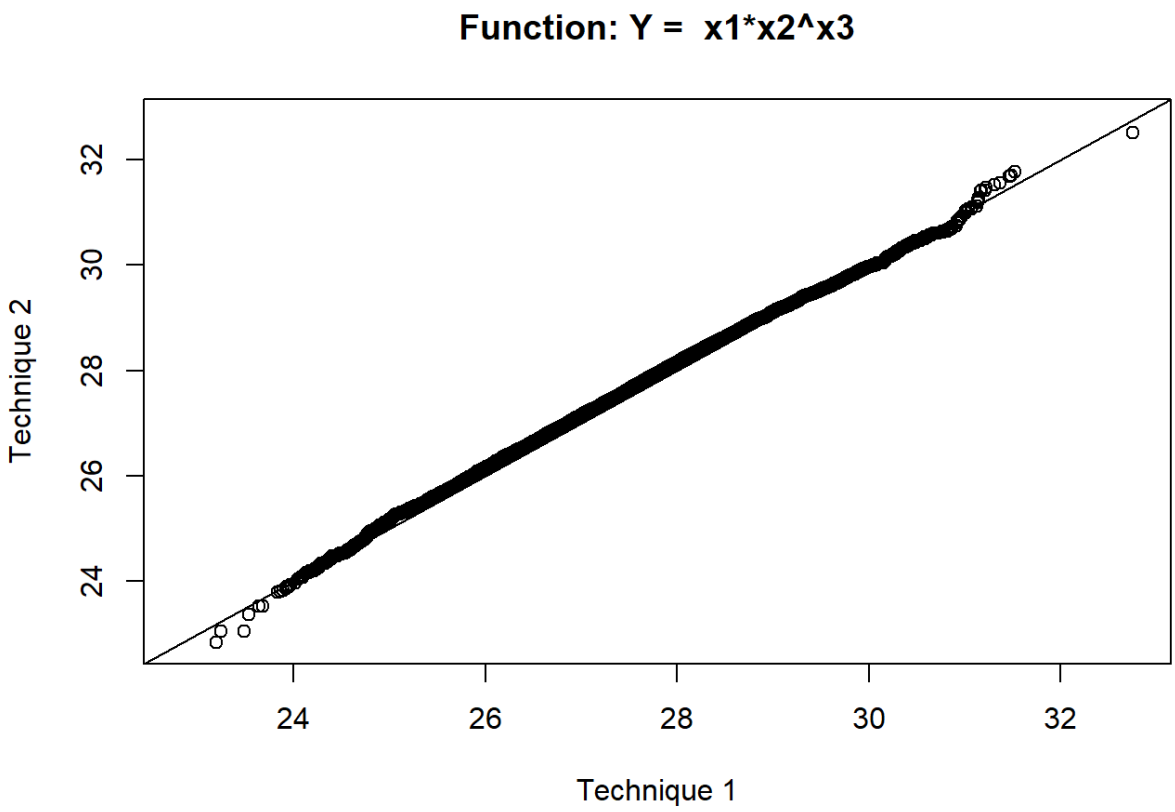


Figure B23. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

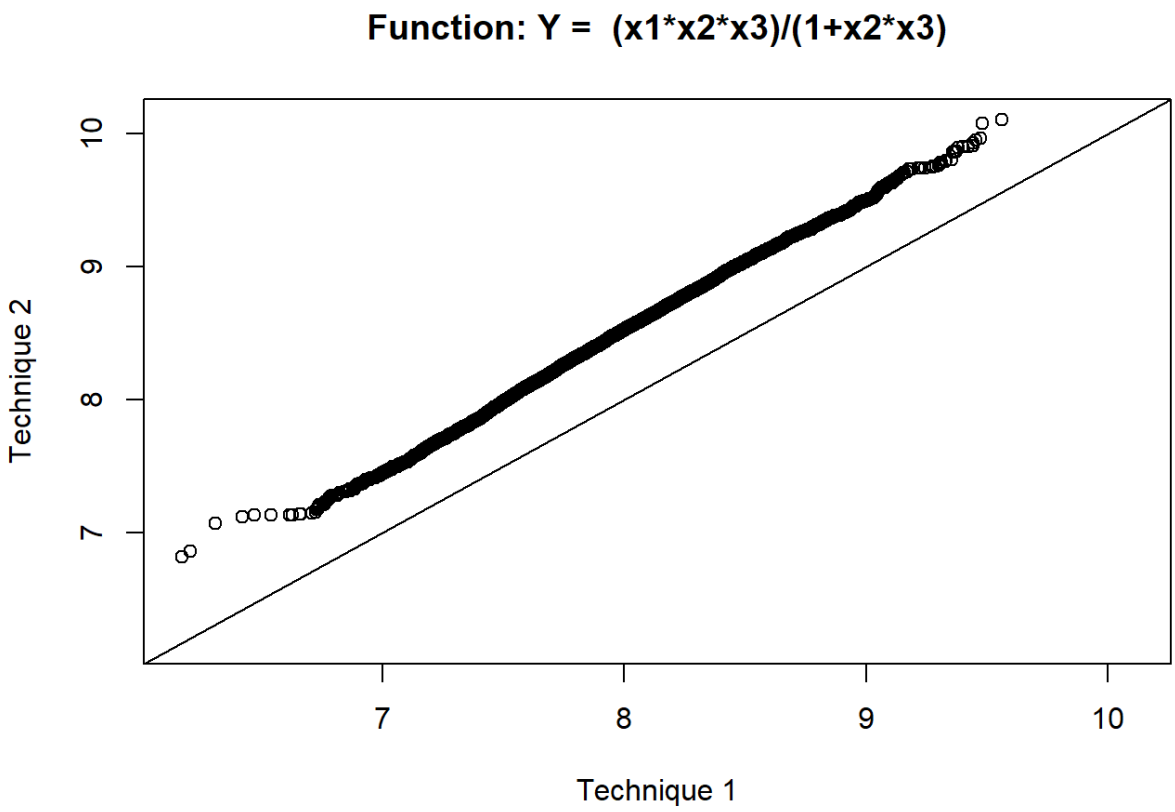


Figure B24. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

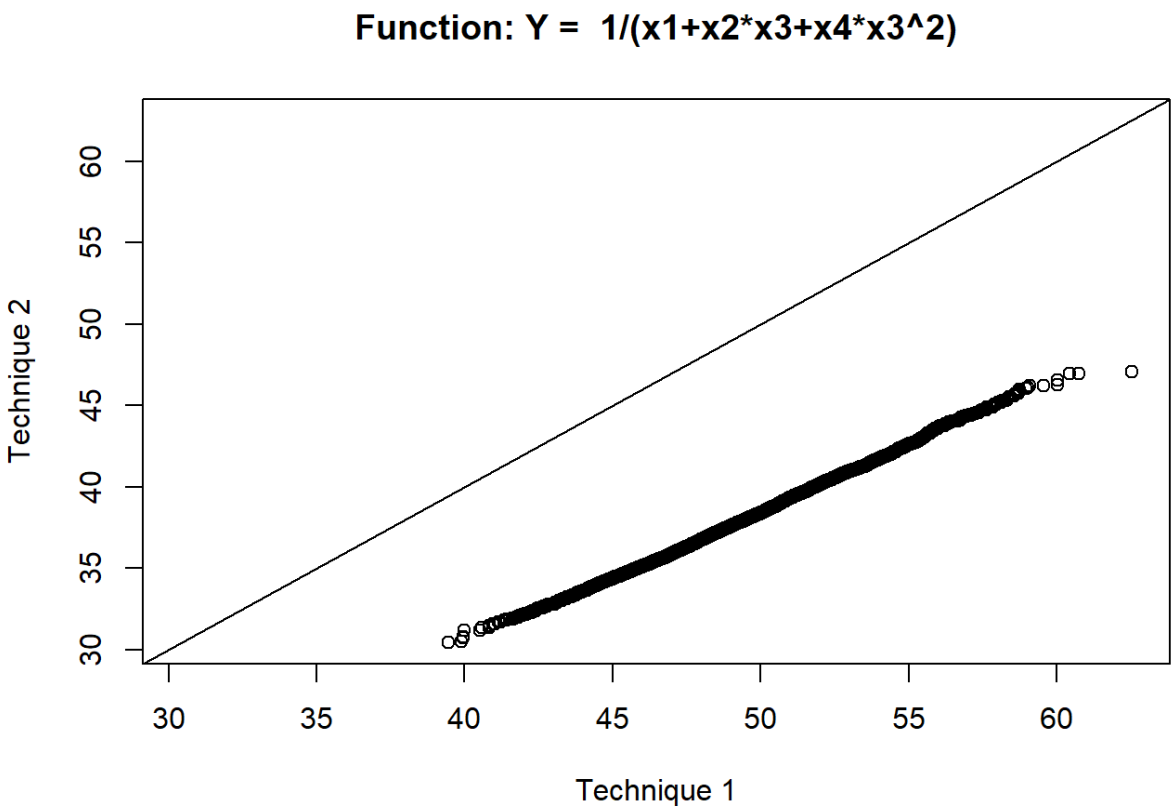


Figure B25. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

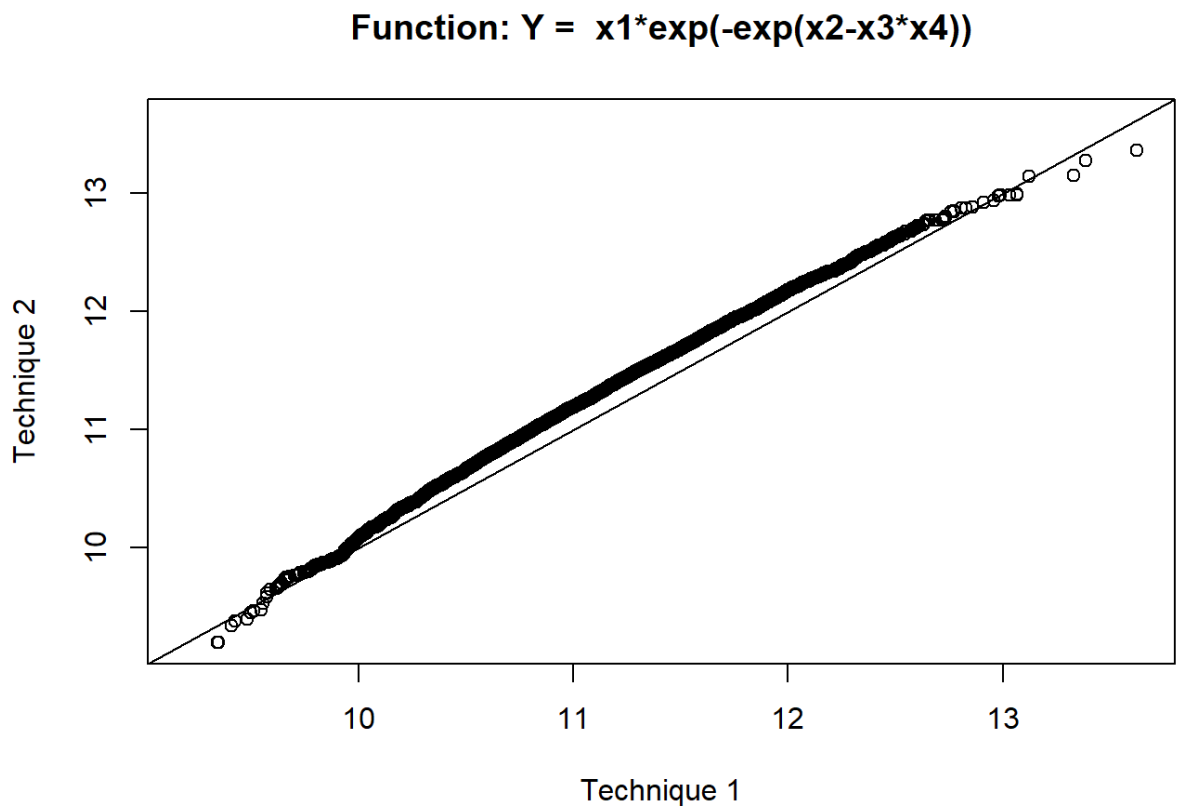


Figure B26. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

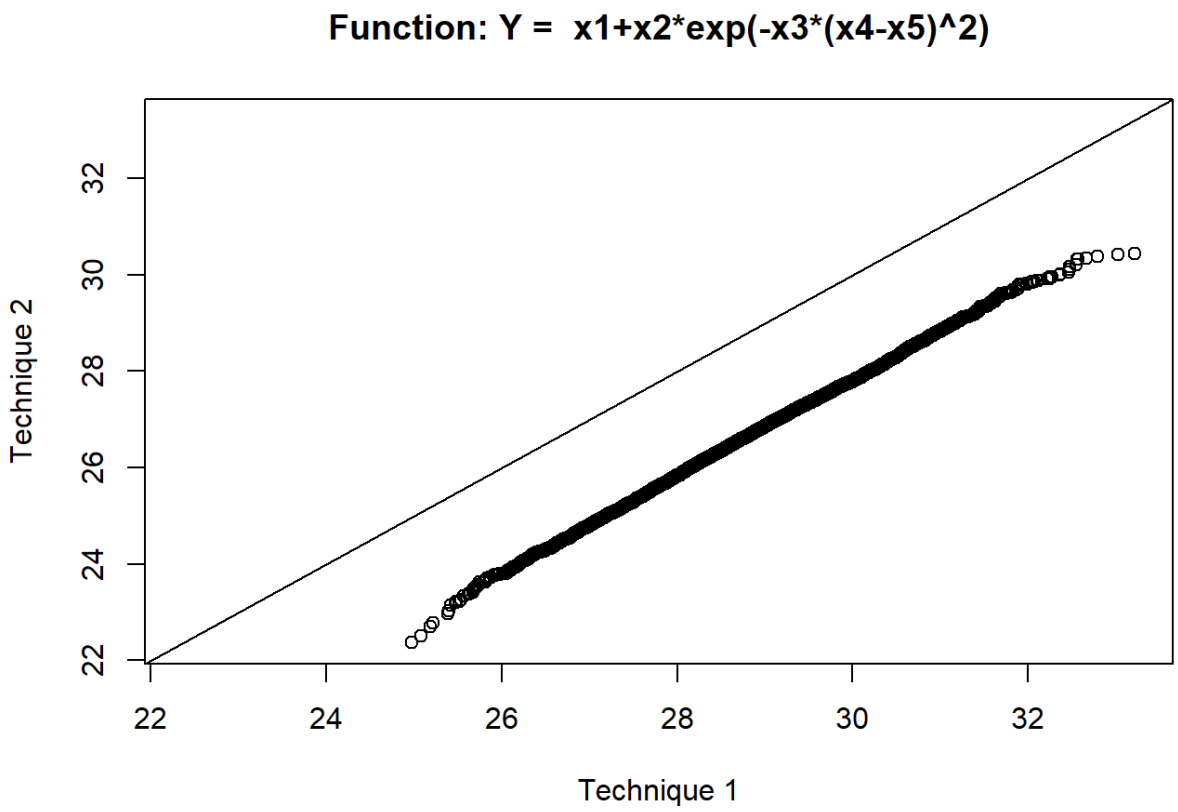


Figure B27. Q-Q plot of Technique 1 vs Technique 2 using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

## APPENDIX C. EXAMPLE FIGURES OF DESCRIPTIVE STATISTICS

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | -1.745642e+07 | -1.774988e+04 | 2.536801e+18 | -1.579893e+05 | -8.165115e+03 |
| Y2 | -3.140590e+03 | -3.093812e+03 | 2.156114e+05 | -3.965653e+03 | -2.456363e+03 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0.0002 | 0.1743 | 0.0000 |

Figure C1. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1 - (\frac{1}{x^a})$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 30.373 | 30.166 | 15.7501 | 24.1856 | 37.2086 | |
| Y2 | 15.78 | 15.6475 | 5.0458 | 12.3338 | 19.675 | 0.3204 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0.5196 | 0.5187 | 0.3204 |

Figure C2. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * x^b$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 348.2015 | 348.1346 | 91.767 | 332.5437 | 364.0274 | |
| Y2 | 357.08 | 357.1385 | 91.5099 | 341.3246 | 373.0147 | 0.9972 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0255 | 1.0259 | 0.9972 |

Figure C3. Descriptive statistics for Technique 1 vs Technique 2 for $y = (a * b * x)/(1 + b * x)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 11.8829 | 11.8753 | 0.0746 | 11.4355 | 12.3443 | |
| Y2 | 11.43 | 11.4241 | 0.0563 | 11.0428 | 11.8262 | 0.7542 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9619 | 0.9620 | 0.7542 |

Figure C4. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(a + b * x + c * x^2)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 26.0175 | 25.5416 | 52.4167 | 15.1347 | 38.8448 | |
| Y2 | 0.01 | 0.0025 | 5e-04 | 1e-04 | 0.0367 | 0 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 3e-04 | 1e-04 | 0e+00 |

Figure C5. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * \exp(-\exp(b - c * x))$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 500.2624 | 500.2915 | 158.6085 | 479.6506 | 520.9925 | |
| Y2 | 499.85 | 499.8812 | 155.6149 | 479.1397 | 520.4322 | 0.9811 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9992 | 0.9992 | 0.9811 |

Figure C6. Descriptive statistics for Technique 1 vs Technique 2 for $y = a + b * \exp(-c * (x - d)^2)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 10.3469 | 10.3437 | 0.0421 | 10.0160 | 10.6890 |
| Y2 | 10.0000 | 10.0020 | 0.0319 | 9.7063 | 10.3028 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9667 | 0.9670 | 0.7581 |

Figure C7. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(x + a)$ using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 335.3584 | 335.391 | 10.1276 | 330.0848 | 340.555 | |
| Y2 | 340.1 | 340.119 | 8.8948 | 335.1052 | 344.964 | 0.8783 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0141 | 1.0141 | 0.8783 |

Figure C8. Descriptive statistics for Technique 1 vs Technique 2 for $y = \log(a + b * x)$ using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 499.958 | 499.8705 | 154.181 | 479.9865 | 520.6856 | |
| Y2 | 500.01 | 500.089 | 156.3849 | 479.3314 | 520.6649 | 1.0143 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0001 | 1.0004 | 1.0143 |

Figure C9. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * (1 - \exp(-b * x))^{\gamma}$ using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 0.1814 | 0.0251 | 0.4285 | 0.0021 | 0.7474 | |
| Y2 | 0 | 0 | 0 | 0 | 0 | 0 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0 | 0 | 0 |

Figure C10. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * \exp(-b * x) + \gamma * \exp(-\delta * x)$ using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 499.1568 | 499.1033 | 155.7355 | 478.6181 | 519.6421 | |
| Y2 | 500.03 | 500.1523 | 154.5366 | 479.4341 | 520.4074 | 0.9923 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0017 | 1.0021 | 0.9923 |

Figure C11. Descriptive statistics for Technique 1 vs Technique 2 for $y = a/(1 + \exp(b - \gamma * x))^{1/\delta}$ using normal distributions, parameters bounded 1-10, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 35.5710 | 35.5622 | 1.2030 | 33.7649 | 37.3986 |
| Y2 | 32.4100 | 32.3721 | 1.0630 | 30.7780 | 34.1620 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9112 | 0.9103 | 0.8836 |

Figure C12. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(x + a)$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 121.0855 | 121.0932 | 8.394 | 116.3221 | 125.8704 | |
| Y2 | 125.23 | 125.2678 | 8.6075 | 120.3621 | 129.9307 | 1.0254 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0342 | 1.0345 | 1.0254 |

Figure C13. Descriptive statistics for Technique 1 vs Technique 2 for $y = \log(a + b * x)$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 214.6447 | 214.4729 | 82.5049 | 200.0796 | 230.248 | |
| Y2 | 239.42 | 239.3373 | 95.8198 | 223.2429 | 255.5542 | 1.1614 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 1.1154 | 1.1159 | 1.1614 |

Figure C14. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * (1 - \exp(-b * x))^{\gamma}$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 322.2771 | 322.1907 | 127.1983 | 303.8133 | 341.0166 | |
| Y2 | 306.56 | 306.3007 | 140.12 | 287.7495 | 326.4237 | 1.1016 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9512 | 0.9507 | 1.1016 |

Figure C15. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * \exp(-b * x) + \gamma * \exp(-\delta * x)$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 125.2771 | 125.0991 | 43.3852 | 114.8561 | 136.5598 | |
| Y2 | 126.83 | 126.766 | 47.8745 | 115.5532 | 138.435 | 1.1035 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1 |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0124 | 1.0133 | 1.1035 |

Figure C16. Descriptive statistics for Technique 1 vs Technique 2 for $y = a/(1 + \exp(b - \gamma * x))^{1/\delta}$ using lognormal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 19.3441 | 19.3145 | 0.3227 | 18.4619 | 20.3040 |
| Y2 | 18.2000 | 18.1821 | 0.1734 | 17.5281 | 18.8944 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9406 | 0.9414 | 0.5374 |

Figure C17. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(x + a)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 199.3674 | 199.3781 | 4.5153 | 195.8441 | 202.8716 | |
| Y2 | 201.5 | 201.5445 | 3.9981 | 198.2011 | 204.7183 | 0.8855 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0107 | 1.0109 | 0.8855 |

Figure C18. Descriptive statistics for Technique 1 vs Technique 2 for $y = \log(a + b * x)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 498.9555 | 498.869 | 154.2526 | 478.9887 | 519.6279 | |
| Y2 | 500.01 | 500.0877 | 156.3835 | 479.3303 | 520.6624 | 1.0138 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0021 | 1.0024 | 1.0138 |

Figure C19. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * (1 - \exp(-b * x))^{\gamma}$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile | Var(y2)/Var(y1) |
|---|---|---|---|---|---|---|
| Y1 | 112.4006 | 112.1369 | 63.0587 | 99.7495 | 125.8901 | |
| Y2 | 82.33 | 82.0907 | 36.3981 | 72.9084 | 92.6358 | 0.5772 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 0.7325 | 0.7321 | 0.5772 |

Figure C20. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * \exp(-b * x) + \gamma * \exp(-\delta * x)$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 295.6816 | 295.5475 | 134.6267 | 277.2963 | 315.2081 |
| Y2 | 298.46 | 298.4673 | 140.0524 | 279.1176 | 317.9752 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0094 | 1.0099 | 1.0403 |

Figure C21. Descriptive statistics for Technique 1 vs Technique 2 for $y = a/(1 + \exp(b - \gamma * x))^{1/\delta}$ using normal distributions, parameters bounded 1-10, and variable bounded 0-1.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 19.0914 | 19.0680 | 0.8091 | 17.6365 | 20.5861 |
| Y2 | 20.5800 | 20.5671 | 0.9468 | 18.9963 | 22.2101 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0782 | 1.0786 | 1.1702 |

Figure C22. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1 - (\frac{1}{x^a})$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 27.4325 | 27.4231 | 1.4604 | 25.4803 | 29.4685 |
| Y2 | 27.56 | 27.5607 | 1.4132 | 25.6162 | 29.5007 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0047 | 1.0050 | 0.9677 |

Figure C23. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * x^b$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 7.9219 | 7.9092 | 0.1836 | 7.2455 | 8.6471 |
| Y2 | 8.43 | 8.4261 | 0.202 | 7.7004 | 9.1742 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0643 | 1.0653 | 1.1001 |

Figure C24. Descriptive statistics for Technique 1 vs Technique 2 for $y = (a * b * x)/(1 + b * x)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 49.2829 | 49.2036 | 8.0714 | 44.6979 | 54.0869 |
| Y2 | 37.84 | 37.7617 | 5.5318 | 34.169 | 41.7886 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 0.7679 | 0.7675 | 0.6854 |

Figure C25. Descriptive statistics for Technique 1 vs Technique 2 for $y = 1/(a + b * x + \gamma * x^2)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 11.0763 | 11.0623 | 0.2787 | 10.2372 | 11.973 |
| Y2 | 11.25 | 11.2519 | 0.2905 | 10.3696 | 12.1452 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 1.0161 | 1.0171 | 1.0424 |

Figure C26. Descriptive statistics for Technique 1 vs Technique 2 for $y = a * \exp(-\exp(b - \gamma * x))$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

| | Mean | Median | Variance | 5% Quantile | 95% Quantile |
|---|---|---|---|---|---|
| Y1 | 28.7528 | 28.7353 | 1.2207 | 26.9758 | 30.5915 |
| Y2 | 26.59 | 26.5953 | 1.2305 | 24.7739 | 28.4243 |

| | Mean(y2)/Mean(y1) | Median(y2)/Median(y2) | Var(y2)/Var(y1) |
|---|---|---|---|
| Ratios of T2 to T1 | 0.9246 | 0.9255 | 1.0080 |

Figure C27. Descriptive statistics for Technique 1 vs Technique 2 for $y = a + b * \exp(-\gamma * (x - \delta)^2)$ using lognormal distributions, parameters bounded 0-1, and variable bounded 1-10.

# APPENDIX D. R CODE

## D.1. Example R Program

```
Any2fn=function(Seed,Steps,Simulations,Distribution1,Parametera1,Parameterb1,Distribution2,Parametera2,Parameterb2,func){
  #Set seed based on input
  set.seed(Seed)

  #Determine function from input
  eval(parse(text = paste('fn = function(x1,x2) { return(' , func , ')}', sep='')))

  #Create dataframe to save distributional and parameter inputs
  params_df<-data.frame(Distribution = c(Distribution1,Distribution2),
                        Parametera = c(Parametera1,Parametera2),
                        Parameterb = c(Parameterb1,Parameterb2))

  #Set number of parameters
  nparam = 2

  #Create empty matrixes and vectors to fill with simulated data
  input = matrix(rep(NA,Steps*Simulations*nparam),nrow = nparam, ncol = Steps*Simulations)
  inputmean = matrix(rep(NA,Simulations*nparam),nrow = nparam, ncol = Simulations)
  ExpectedVal = as.vector(rep(NA,nparam))
  Var = as.vector(rep(NA,nparam))

  #Set f = parsed function
  f=fn

  #Run this loop for each parameter
  for(i in 1:nparam){
  #Set of if else statements that determine which distribution to draw from for each variable
  if(params_df$Distribution[i]=="normal"){
    #Draws for technique 1
    input[i,] = abs(rnorm(Steps*Simulations,params_df$Parametera[i], params_df$Parameterb[i]))
    #Draws for technique 2
    inputmean[i,]= rnorm(Simulations, params_df$Parametera[i], sd=(params_df$Parameterb[i]/sqrt(Steps)))
  }
  else if(params_df$Distribution[i]=="lognormal"){
    #Convert lognormal parameters into normal approximations
    ExpectedVal[i] = exp(params_df$Parametera[i]+((params_df$Parameterb[i]^2)/2))
```

```
    Var[i] = (exp(params_df$Parameterb[i]^2)-1)*exp(2*params_df$Parametera[i]+(params_df$Parameterb[i]^2))
    #Draws for technique 1
    input[i,] = rlnorm(Steps*Simulations,params_df$Parametera[i],params_df$Parameterb[i])
    #Draws for technique 2 using converted lognormal parameters
    inputmean[i,]= rnorm(Simulations,mean=ExpectedVal[i], sd=sqrt(Var[i]/Steps))
  }
  }



  #Function applied to Variable and Parameters
  output1 = f(input[1,],input[2,])
  output2= f(inputmean[1,],inputmean[2,])


  #Create Matrices for Variable
  outmat1 = matrix(output1, nrow=Simulations, ncol=Steps)


  #Scale Y's to have an equal number of data points
  y1= rowSums(outmat1)
  y2= output2*Steps


  #Convert data into dataframes
  y1dat=data.frame(y1)
  y2dat=data.frame(y2)


  #Combine dataframes for graphics
  ycomb = cbind(y1,y2)
  ydat = data.frame(ycomb)


  #Plot histograms for each outcome
  Compared = ggplot(ydat) + geom_histogram(aes(x=y1),fill = "blue", alpha = .2, bins = 100)+
                            geom_histogram(aes(x=y2),fill = "red", alpha = .2, bins = 100)+
                            labs(title = paste("Function: Y = ",func),x = "Technique 1=Blue  Technique 2=Red")
  grid.arrange(Compared)


  #Create qqplot to compare distributions
  qqPlot(y1,y2,xlab = "Technique1", ylab = "Technique2", main = paste("Function: Y = ",func), add.line = TRUE,
qq.line.type = "0-1", equal.axes = TRUE)


  #Create table of descriptive statistics
  out_table = matrix(c(round(mean(y1),4),round(median(y1),4),round(var(y1),4),round(as.numeric(quantile(y1,.05,
na.rm =  TRUE)),4),round(as.numeric(quantile(y1,.95,na.rm = TRUE)),4),round(mean(y2,na.rm = TRUE),2),round(medi
an(y2),4),round(var(y2,na.rm = TRUE),4),round(as.numeric(quantile(y2,.05,na.rm = TRUE)),4),round(as.numeric(qua
ntile(y2,.95,na.rm = TRUE)),4)), ncol = 5, byrow = TRUE)
```

```
  colnames(out_table) = c("Mean", "Median", "Variance", "5% Quantile","95% Quantile")

  rownames(out_table) = c("Y1","Y2")

  print(as.table(out_table))


  #Create table of proportional differences

  out_table2 = matrix(c(round(mean(y2)/mean(y1),4),round(median(y2)/median(y1),4),round(var(y2)/var(y1),4)),nco
l = 3,byrow = TRUE)

  colnames(out_table2) = c("Mean(y2)/Mean(y1)","Median(y2)/Median(y2)","Var(y2)/Var(y1")

  rownames(out_table2) = c("Prop. diff.")

  print(as.table(out_table2))


}
```

## D.2. Example R Output Program

```
#######Semi-linear#######

#Bound a between 0,1


#2 parameters

"y = 1-(1/x^a)"

Any2fn(100,100,10000,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"1-(1/x1^x2)")


#3 parameters

"y = a*x^b"

Any3fn(200,100,10000,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"x1*x2^x3")

"y = (a*b*x)/(1+b*x)"

Any3fn(300,100,10000,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"(x1*x2*x3)/(1+x2*x3)")


#4 paramters

"y = 1/(a+b*x+c*x^2)"

Any4fn(400,100,10000,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"1/(x1+x2*
x3+x4*x3^2)")

"y = a*exp(-exp(b-c*x))"

Any4fn(500,100,10000,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"x1*exp(-e
xp(x2-x3*x4))")


#5 parameters

"y = a+b*exp(-c*(x-d)^2)"

Any5fn(600,100,10000,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal",-1.6,.4,"lognormal
",-1.6,.4,"x1+x2*exp(-x3*(x4-x5)^2)")
```