METHODS FOR STRENGTHENING THE DESIGN AND ANALYSIS OF CLINICAL TRIALS TO SHOW NON-INFERIORITY OF A NEW TREATMENT TO A REFERENCE TREATMENT FOR A BINARY RESPONSE VARIABLE

Rebekkah S. Dann

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctorate of Public Health in the School of Public Health (Biostatistics)

Chapel Hill
2006

Approved by:

Gary Koch

Amy Herring

Lisa LaVange

Jeanenne Nelson

John Preisser

ABSTRACT

Rebekkah S. Dann: Methods for Stregthening the Design and Analysis of Clinical Trials to Show Non-inferiority of a New Treatment to a Reference Treatment for a Binary Response Variable
(Under the direction of Gary G. Koch)

Non-inferiority clinical trials are increasingly becoming more prominent in research and development of new pharamaceuticals. The objective of such trials is to show that the amount by which a new treatment is worse than an active control is below a specified amount. Methodology specifically for the design and analysis of these trials is essential for the assurance of quality trials that are statistically defensible in the scientific community as well as in a regulatory setting, where traditionally focus has been on superiority.

Standard methodology must be reviewed and assessed as to its appropriateness for addressing the non-inferiority hypothesis. Categorical data analysis for a dichotomous primary endpoint may include analysis of a risk ratio or a risk difference which compares the test and active control treatments. The effect of sample size allocation and other parameters of interest on the performance of these methods will be assessed. In addition, appropriate sample size formulas will be developed and evaluated to aid in trial planning.

In some non-inferiority trials, it is possible to include a placebo arm as well as an active control arm which allows non-inferiority to be assessed relative to the percentage of the difference between the control and placebo arms that the test treatment preserves over placebo. Methodology for this assessment is also of interest along with appropriate sample size calculations. This setting also presents an area of research for discussion of the one versus two trials paradigm.

Extensions to the methodology for the risk ratio and risk difference are assessed when stratification is necessary, specifically for large subgroups such as gender. Methods for stratification are an important component, and additionally the effects of stratification in a non-inferiority setting need evaluation.

Review, development, and assessment of this methodology for categorical data as specifically focused on the non-inferiority setting is an important addition to the current statistical practice. This research is a cohesive presentation for each of the measures of interest through assessment of methodology and its relation to appropriate design components such as sample size calculation. The importance of helping statisticians understand and implement methods in these areas is of most concern.

ACKNOWLEDGEMENTS

This dissertation is the product of many years of work, spurred on by those I have been privileged to know. My family has been a constant source of encouragement to pursue education and finish this endeavor. My friends are so important to me and have walked alongside of me through the highs and lows of the past few years. Lastly, my academic mentors have been instrumental in guiding me through from high school to completing graduate school, with much support and affirmation. Dr. Koch has provided assistance in teaching me the practice of biostatistics as well as many opportunities to launch my career. Thank you to all who have been a part of this process.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

INTRODUCTION

Non-inferiority clinical trials are increasingly becoming more prominent in the research and development of new pharamaceuticals. Methodology specifically for the design and analysis of these trials is an essential component for the assurance of quality trials that are statistically defensible in the scientific community as well as in a regulatory setting, where traditionally the focus has been on superiority.

The main goal of non-inferiority trials is to show that the new experimental medication (test treatment) is not unacceptably worse than the current standard of care (active control treatment) by a specified amount, but the test treatment may have other desirable aspects such as a better safety profile or properties which make patient compliance better. This is a reversal from the goals of a superiority trial which generally includes the test treatment in comparison to a placebo where the goal is to show that this new treatment is more effective than placebo. In certain disease areas such as infections, the use of a placebo control arm is unethical due to widespread use of the active control for treatment of the disease.

There are specific guidance documents which discuss the issues surrounding the design and implementation of non-inferiority trials. The ICH-E10 guidance[1] on the "Choice of Control Group and Related Issues in Clinical Trials" provides the rationale for use of an active-control treatment in a non-inferiority setting. In addition, the trial must address assay sensitivity through historical evidence of efficacy, and the conduct of the trial must make efforts to increase compliance and minimize dropouts, since poor trial conduct can bias these trials toward non-inferiority. Additionally, the choice of a margin for testing the inferiority hypothesis must be established by clinical and statistical judgment[2].These issues are all very important aspects in the design of a non-inferiority trial. However, the current discussion will assume that these issues are appropriately addressed and the focus will include statistical issues related to sample size calculation, sample size allocation, and analysis in non-inferiority trials. These issues are essential for statisticians who need to know how to better design and analyze these trials, with specific emphasis on methods related to dichotomous categorical data.

Standard methodology must be reviewed and assessed as to its appropriateness for addressing the non-inferiority hypothesis. Categorical data analysis for a dichotomous primary endpoint may include analysis of a risk ratio or a risk difference which compares the test and active control treatments. This assessment of non-inferiority is performed by computing a confidence interval and determining if the applicable limit is below (or similarly above) the pre-specified non-inferiority margin. A test statistic can also be used for this assessment where rejection of the null hypothesis of inferiority would require a p-value less than the pre-specified alpha level. Methods for computing either the confidence interval or the corresponding test statistic will be assessed according to their performance with respect to type I error and power through simulations for relevant scenarios. The effect of sample size allocation on the performance of these methods will also be assessed. In addition, appropriate sample size formulas will be developed to aid in trial planning.

In some non-inferiority trials it is possible to include a placebo arm as well as an active control arm. This placebo arm can address issues related to assay sensitivity and appropriate trial conduct. This also allows non-inferiority to be assessed relative to the placebo arm, using the percentage of effect that the test treatment has over placebo compared to the effect that the control treatment preserves over placebo. Methodology for this assessment is also of interest along with appropriate sample size calculations. This setting also presents an area of research for discussion of the one versus two trials paradigm. Frequently, regulatory agencies require two confirmatory trials. However, if these trials are run in an identical manner with similar protocols, it may be beneficial to run one large trial. The implications of these scenarios are assessed related to type I error control and the resulting power for rejecting the null hypothesis of inferiority.

Extensions to the methodology for the risk ratio and risk difference are assessed when stratification is necessary, specifically for large subgroups such as gender. Methods for stratification are an important component, and additionally the effects of stratification in a non-inferiority setting.

Review, development, and assessment of this methodology for dichotomous data specifically focused on the non-inferiority setting is an important addition to the current statistical practice. This research is a cohesive presentation for each of the measures of interest through assessment of methodology and its relation to appropriate design components such as sample size calculation. The importance of helping statisticians understand and implement methods in these areas is of most concern.

References

1. U. S. Department of Health and Human Services, Food and Drug Administration. Guidance for Industry, E10 Choice of Control Group and Related Issues in Clinical Trials. Rockville, MD **2001**.

2. Hung, H. M. J., Wang, S. J., Tsong Y., Lawrence, J., and O'Neill, R. T. Some Fundamental Issues with Non-Inferiority Testing in Active Controlled Trials. Statistics in Medicine **2003**, 22, 213-225.

Chapter 1

Review and Evaluation of Methods for Computing Confidence Intervals for the Ratio of Two Proportions and Considerations for Non-inferiority Clinical Trials

I. Introduction

Ratios of proportions are often called risk ratios in a clinical trials setting. These ratios are used to compare two independent groups, usually on two different treatments. A non-inferiority clinical trial can compare an active control group to a group taking a new treatment for an efficacy outcome (or a placebo group to a group taking a new treatment for a safety outcome). The goal is to show that the new treatment is not unacceptably worse than the active control (or placebo) treatment[1]. The new treatment may have other beneficial aspects such as a reduction in severity of side effects, easier use, or lower cost.

Assessing non-inferiority is often done through a confidence interval for the risk ratio of the two groups[1], particularly if control failure rates are small (e.g., ≤ 0.20) or control success rates are large (e.g., ≥ 0.80). If failure rates are very small (e.g., < 0.05) then the odds ratio can be conservatively used to approximate the risk ratio (when defined so as to have the larger expected rate in the numerator and the smaller expected rate in the denominator). For situations where failure rates are larger (e.g., > 0.20), then the difference in rates is typically emphasized[2,3,4]. In some cases, if the new treatment group has a risk that is not more than twice that of the control group for a failure outcome through an upper confidence limit of 2 or less, then the new treatment will be judged non-inferior. This non-inferiority limit can be set at a variety of pre-determined levels[1] denoted θ_0. Accordingly, a corresponding test of non-inferiority has the null hypothesis as H_O: $\theta = \pi_T / \pi_C \geq \theta_0$ and the alternative hypothesis as H_A: $\theta = \pi_T / \pi_C < \theta_0$ where $\theta = \pi_T / \pi_C$ is the population risk ratio for the test group versus the control group with π_T as the population proportion of events in the test group and π_C as the population proportion of events in the control group, and $\theta_0=2$ was the previously mentioned example.

There are many methods in existence for computing a confidence interval for a risk ratio. Several of the methods for forming confidence intervals for ratios of two independent binomial proportions will be reviewed and evaluated for their statistical performance. These methods include use of a Taylor Series expansion to estimate variance, solutions to a quadratic equation, and maximum likelihood methods. Simulations were used to identify the better methods for controlling the type I error rate while maintaining power. Applications of these findings include sample size calculations which arise in randomized clinical trials conducted to show non-inferiority.

II. Methods

A. Taylor Series Expansion Methods

The literature contains many methods for forming confidence intervals for risk ratios. The first group of these uses a variance formed through a Taylor Series expansion. The following method seen in (1.1), hereafter called the Taylor Series method, is the simplest in this group discussed by Katz, Baptista, Azen and Pike[5] and used by SAS in the FREQ procedure[6] and by EquivTest[7] to form a 100(1-2α)% confidence interval for a risk ratio:

$$\exp\left\{\log_e\left(\frac{y_T/n_T}{y_C/n_C}\right)\pm z_\alpha\left[\frac{1}{y_T}+\frac{1}{y_C}-\frac{1}{n_T}-\frac{1}{n_C}\right]^{1/2}\right\} \tag{1.1}$$

where y_T is the number of events and n_T is the total sample size in the treatment group, y_C is the number of events and n_C is the total sample size in the control group, and z_α is the 100(1-α) percentile from a standard normal distribution.

In 1988, Gart and Nam[8] revised this original method so that the confidence interval would be defined if y_T or y_C were equal to zero. The formula seen in (1.2) is this modified confidence interval used by StatXact[9] for risk ratios.

$$\exp\left\{\log_e\left(\frac{(y_T+0.5)/(n_T+0.5)}{(y_C+0.5)/(n_C+0.5)}\right)\pm z_\alpha\left[\frac{1}{y_T+0.5}+\frac{1}{y_C+0.5}-\frac{1}{n_T+0.5}-\frac{1}{n_C+0.5}\right]^{1/2}\right\} \tag{1.2}$$

This Modified Taylor Series method adds a half to the event count for each group as well as the total sample size for each group.

The last method in this group of Taylor Series expansion methods is adapted from a confidence interval for a single binomial proportion proposed by Agresti and Coull[10]. For a confidence interval for a single binomial proportion, Agresti and Coull suggested adding half the squared z-value (at the corresponding alpha level) to each outcome for each group to produce a more conservative interval. This strategy was adapted for a test of non-inferiority where the null hypothesis is not one of equality. The additional $2z_\alpha^2$ counts must be distributed to each group according to the null hypothesis (θ_0) and the allocation of sample size to each group ($R=n_T/n_C$) as seen in (1.3)

$$\exp\left\{\log_e\left(\frac{(y_T+\gamma_T)/(n_T+\gamma_{Tot,T})}{(y_C+\gamma_C)/(n_C+\gamma_{Tot,C})}\right)\pm z_\alpha\left[\frac{1}{y_T+\gamma_T}+\frac{1}{y_C+\gamma_C}-\frac{1}{n_T+\gamma_{Tot,T}}-\frac{1}{n_C+\gamma_{Tot,C}}\right]^{1/2}\right\}$$

(1.3)

where $\gamma_{\text{Tot,T}}=2z_\alpha^2*\frac{R}{R+1}$, $\gamma_{\text{Tot,C}}=2z_\alpha^2*\frac{1}{R+1}$, $\gamma_T=\gamma_{\text{Tot,T}}*\frac{\theta_0}{\theta_0+1}$, and

$\gamma_C=\gamma_{\text{Tot,C}}*\frac{1}{\theta_0+1}$.

For example, at an α=0.025 level an additional $2z_\alpha^2=2(1.96)^2\approx 8$ counts must be added. For a setting with twice as many patients allocated to the test group than the control group, R=2, and a null hypothesis of θ_0=2, there are a total of γ_T=3.56 events added to the test group, γ_C=0.89 events added to the control group with a total of $\gamma_{\text{Tot,T}}$=5.33 added to the overall number of patients in the test group and $\gamma_{\text{Tot,C}}$=2.67 patients added to the control group.

The Taylor Series Adjusted Alpha method was added so as to correct inflation of type I error by the Taylor Series method seen in initial simulations. This method is the Taylor Series method with an alpha level that is 0.0025 less than the alpha level for the α=0.025 scenario. For example, this method would use an alpha level of 0.025 – 0.0025 = 0.0225 when α=0.025 was specified. The choice of 0.0025 was motivated by findings from the simulations for the specific scenarios presented where this modification was needed to offset the small inflation in type I error of the Taylor Series method. This 0.0025 adjustment of the alpha level is dependent on the application at hand and simulations can be used to determine the appropriate adjustment for any scenario. This adjustment to the alpha level is a way to address studies with finite samples rather than infinite (or very large) samples by increasing the z-criterion for significance slightly (i.e., for alpha=0.025 the z-criterion would increase from 1.96 to 2.00).

B. Solution to Quadratic Equation Methods

The next group of methods is slightly more complicated because the confidence limits are the solutions to a quadratic equation. After algebraic manipulations, a quadratic form of the equations provided below are then solved for θ. The upper and lower confidence limits are the smaller and larger of the two solutions, respectively. However, these methods may produce complex-valued results (when square roots of negative numbers are involved).

Fieller[11] first presented the most basic of these methods in 1944 as seen in (1.4), hereafter called the Quadratic method where $\hat{p}_T = y_T / n_T$ and $\hat{p}_C = y_C / n_C$.

$$\frac{(\hat{p}_T - \theta\,\hat{p}_C)^2}{\left\{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \theta^2\,\frac{\hat{p}_C(1-\hat{p}_C)}{n_C}\right\}} = z_\alpha^2 \tag{1.4}$$

The second of this group of methods in (1.5) was proposed by Bailey in 1987[12] which is a modification of the Quadratic method to produce limits with more desirable properties as will be discussed in more detail in the literature review section.

$$\frac{(\hat{p}_T{}^{1/3} - \theta^{1/3}\,\hat{p}_C{}^{1/3})^2}{\frac{1}{9}\left\{\frac{\hat{p}_T{}^{-1/3}(1-\hat{p}_T)}{n_T} + \theta^{2/3}\,\frac{\hat{p}_C{}^{-1/3}(1-\hat{p}_C)}{n_C}\right\}} = z_\alpha^2 \tag{1.5}$$

The last of this group of methods was proposed by Farrington and Manning in 1990[4] with three possible variations on the equation in (1.6).

$$\frac{(\hat{p}_T - \theta\,\hat{p}_C)^2}{\left\{\dfrac{\tilde{\pi}_T(1-\tilde{\pi}_T)}{n_T} + \theta^2\,\dfrac{\tilde{\pi}_C(1-\tilde{\pi}_C)}{n_C}\right\}} = z_\alpha^2 \tag{1.6}$$

Each variation suggests computing $\tilde{\pi}_C$ and $\tilde{\pi}_T$ in a different manner. The first of these, F-M 1, uses the observed values and sets $\tilde{\pi}_C = \hat{p}_C$ and $\tilde{\pi}_T = \hat{p}_T$. The second variation, F-M 2, uses fixed marginal totals to compute $\tilde{\pi}_T = \dfrac{\theta_0\left(n_T\hat{p}_T + n_C\hat{p}_C\right)}{(\theta_0 n_T + n_C)}$ and $\tilde{\pi}_C = \dfrac{(n_T\hat{p}_T + n_C\hat{p}_C)}{(\theta_0 n_T + n_C)}$.

The third variation, F-M 3, uses maximum likelihood estimation under the null hypothesis to obtain $\tilde{\pi}_C$ and $\tilde{\pi}_T$ with details found in Farrington and Manning's paper[4] and solutions for $\tilde{\pi}_T$ and $\tilde{\pi}_C$ below:

$$\tilde{\pi}_T = \frac{-b-\sqrt{b^2-4ac}}{2a} \text{ and } \tilde{\pi}_C = \tilde{\pi}_T / R$$

where $a = 1 + \dfrac{n_C}{n_T}$, $b = -\left\{\theta_0\left(1+\dfrac{n_C}{n_T}\hat{p}_C\right) + \dfrac{n_C}{n_T} + \hat{p}_T\right\}$, and $c = \theta_0\left(\hat{p}_T + \dfrac{n_C}{n_T}\hat{p}_C\right)$.

In addition, Gart and Nam[8] summarize an interval attributed to Noether where the equation in (1.7) is solved for θ to yield upper and lower confidence limits, θ_L and θ_U.

$$\frac{\left(\hat{p}_T / \hat{p}_C - \theta\right)^2}{\left\{\dfrac{\theta^2(1-\hat{p}_C)}{n_C\hat{p}_C} + \dfrac{\theta(1-\theta\hat{p}_C)}{n_T(\theta\hat{p}_C)}\right\}} = z_\alpha^2 \tag{1.7}$$

C. Maximum Likelihood Methods

The third group of confidence interval methods includes those that use maximum likelihood estimators for the proportion of events in the treatment and control groups based on the joint distribution of the events as the product of two independent binomial distributions for the treatment and control groups. The first of these methods calculates a deviance statistic as seen in (1.8)

$$\text{Deviance} = 2 * [\log \mathrm{L}(\hat{\pi}_T, \hat{\pi}_C) - \log \mathrm{L}(\theta_0 \hat{\pi}^*, \hat{\pi}^*)] \qquad (1.8)$$

where $\hat{\pi}_T$ and $\hat{\pi}_C$ are the maximum likelihood estimators of π_T and π_C under the alternative hypothesis and $(\theta_0 \hat{\pi}^*)$ and $\hat{\pi}^*$ are the corresponding maximum likelihood estimators under the null hypothesis $\theta = \theta_0$.

The second of these maximum likelihood methods is based on a Pearson statistic in the form of [(observed – expected)^2 / expected] in (1.9).

$$\frac{\{y_T - \theta_0 n_T \hat{\pi}^*\}^2}{\theta_0 n_T \hat{\pi}^*} + \frac{\{(n_T - y_T) - n_T(1 - \theta_0 \hat{\pi}^*)\}^2}{n_T(1 - \theta_0 \hat{\pi}^*)} + \frac{\{y_C - n_C \hat{\pi}^*\}^2}{n_C \hat{\pi}^*} + \frac{\{(n_C - y_C) - n_C(1 - \hat{\pi}^*)\}^2}{n_C(1 - \hat{\pi}^*)}$$

(1.9)

using $\theta_0 \hat{\pi}^*$ and $\hat{\pi}^*$, the maximum likelihood estimators of π_T and π_C under the null hypothesis $\theta = \theta_0$. Koopman[13] proposed this method in 1984, and StatXact[9] is a software package that provides these confidence intervals.

In addition, Bedrick[14] discusses a set of methods termed the power divergence methods seen in (1.10) where various values of λ can be used, with this discussion focusing on λ=-0.5, 0.5, 0.67, 1.0, and 1.25.

$$I^{\lambda} = \frac{2}{\lambda(\lambda+1)}\left\{ n_T\left[\hat{p}_T\left(\frac{\hat{p}_T}{\tilde{p}_T}\right)^{\lambda} + (1-\hat{p}_T)\left(\frac{1-\hat{p}_T}{1-\tilde{p}_T}\right)^{\lambda} - 1\right] + n_C\left[\hat{p}_C\left(\frac{\hat{p}_C}{\tilde{p}_C}\right)^{\lambda} + (1-\hat{p}_C)\left(\frac{1-\hat{p}_C}{1-\tilde{p}_C}\right)^{\lambda} - 1\right]\right\}$$

(1.10)

The Deviance, Pearson, and Power Divergence methods produce test statistics for which p-values can be obtained using the chi-square distribution under one degree of freedom. The appropriate confidence limits can be found through an iterative process. The hypothesized ratio θ of π_T to π_C is modified until the desired p-value (e.g., 0.05 or 0.025) is obtained. This process identifies the largest θ_0 that would not be rejected as H_0: $\theta \geq \theta_0$. The ratio that produces the desired p-value is then the upper confidence limit. This iterative process requires changing the maximum likelihood estimator pertaining to the null hypothesis as θ_0 changes. This group of methods is more complicated than the others due to the iterative nature of finding the confidence intervals as all hypotheses not rejected, thus requiring intensive computer resources.

A summary of available software resources for the computation of the methods described can be found in Table 1.1.

III. Review of Literature

Different combinations of the methods described above have been compared in the literature. In 1978, Katz et al.[5] compared the Taylor Series method and the Quadratic method using simulations and calculating coverage probabilities. Katz et al. suggested that the Quadratic method could be erratic and may not produce confidence limits at all; the Taylor Series method was recommended for use instead of the Quadratic method.

Again in 1984, Koopman[13] used simulations and coverage probabilities to compare the Taylor Series method and the Pearson maximum likelihood method. Findings suggested that the Pearson method maintains a coverage probability closer to the (1 - 2α) level, and in addition, the one-sided probabilities of exceeding the upper limit or being lower than the lower limit are much closer to α. Therefore, Koopman recommended use of the Pearson method.

In 1987, Bailey[12] extended the Quadratic method to produce Bailey's method, which should reduce the skewness of the confidence interval as well as maintain the nominal coverage probability better than the Quadratic method. This new method is also compared to the Taylor series method and the Pearson method. Bailey concluded that his method results in confidence limits that are closer to the nominal level than the Taylor Series method. In addition, Bailey's method more often maintains the nominal coverage probability better than the Pearson method.

Gart and Nam[8] produced a comprehensive comparison of the methods presented previous to 1988. They indicated that the Quadratic method and Bailey's method tend to produce confidence limits that are either above or below the nominal coverage probability, whereas the Modified Taylor Series method and the Pearson method achieve coverage probabilities close to the nominal level, with the Pearson method slightly better. They also discuss a skewness-corrected score method, which is iterative in nature, that is slightly better than the Power Divergence method (λ=0.5) of Bedrick[14].

In 1990, Farrington and Manning[4] presented results on the three variations of quadratic methods for producing confidence limits for risk ratios. Their recommendation was the third of these methods, F-M 3, based on maximum likelihood estimation for the proportions.

IV. Confidence Limit Comparisons

An initial comparison of the methods includes computing the upper confidence limits for selected cases. At a one-sided alpha level of 0.025, the upper confidence limits are presented for each of the methods producing confidence limits and p-values for the methods producing a test statistic (with confidence limits computed through an iterative process). The methods are grouped by the three method types: the Taylor Series variance expansion methods (Table 1.2), the quadratic methods (Table 1.3), and the maximum likelihood methods (Table 1.4).

Within the Taylor Series variance expansion methods, the Taylor Series method and the Taylor Series Adjusted Alpha method produce higher confidence limits for the 1:2 allocation whereas the Adapted Agresti method has higher confidence limits for the allocations that place more sample size in the test treatment for the 3:2, 2:1, and 3:1 allocations. The Quadratic method and Farrington-Manning method 1 produce very similar upper confidence limits due to their similarity in computation. Noether's method produces higher confidence limits for all sample size allocations. Farrington-Manning methods 2 and 3 also produce similar upper confidence limits for the selected cases presented. The Deviance and Pearson methods produce similar p-values. The group of Power Divergence methods yields decreasing p-values for increasing choices of λ.

V. Simulations

Data were generated from known distributions to compare the behavior of the methods with respect to power and type I error. Scenarios included varying the following parameters:

1. π_C, the population proportion of events in the control group: 0.10, 0.15, 0.20, 0.25
2. $\theta = \pi_T/\pi_C$, the population risk ratio: 0.667, 0.800, 1.000, 1.250, 1.500, 2.000, 2.500
3. π_T, the population proportion of events in the test group: $\pi_T = \theta\pi_C$
4. θ_0, the null hypothesis risk ratio: 1.5, 2.0, 2.5
5. α, the one-sided alpha level: 0.005, 0.025, 0.050
6. n_C, the sample size in the test group is calculated to have 85% power to contradict the null hypothesis θ_0, given a risk ratio of 1 for test versus control with $n_T = Rn_C$:

$$n_C = \frac{\left(z_\alpha + z_\beta\right)^2 \left\{\frac{1}{R\pi_T} + \frac{1}{\pi_C}\right\}}{\left\{\ln\left(1/\theta_0\right)\right\}^2}$$

7. Sample size allocation for test:control as 1:2, 1:1, 3:2, 2:1, 3:1

For each combination of the parameters, 100,000 simulations were generated using a random sample from the two binomial distributions of $y_T \sim \text{bin}(n_T, \pi_T)$ and $y_C \sim \text{bin}(n_C, \pi_C)$. For each combination of y_T and y_C, upper confidence limits or test statistics with corresponding p-values for all methods were calculated. If y_T or y_C were equal to zero or the method failed to produce a valid result, then the exact confidence limit for the odds ratio was the default. This modification using the odds ratio is conservative because it employs exact methodology and because the odds ratio exceeds the risk ratio when both exceed one. As a note, if y_C=0 then the upper confidence limit for the odds ratio is essentially infinite, and so it was set to 100 and the null hypothesis of inferiority was not rejected. This modification, where the upper confidence limit was set to 100, is also necessary in cases where the group of quadratic methods lead to square roots of negative numbers (i.e., complex solutions) or where the Deviance or Pearson methods fail to produce interpretable results because of computational singularities. No modifications were necessary for the Modified Taylor Series or the Adapted Agresti methods.

For each method, an indicator variable was created for each simulation that takes the value of 1 if the upper confidence limit produced was less than θ_0 and 0 otherwise or similarly if the p-value was less than alpha the indicator takes the value 1 and 0 otherwise. This indicator was then averaged across all 100,000 simulations to produce a probability. For $\theta < \theta_0$, this probability is the power for the test of non-inferiority and can be written in the following manner: power = pr(reject H_O: $\theta=\pi_T/\pi_C \geq \theta_0$ | H_A: $\theta < \theta_0$ true). For $\theta = \theta_0$, this probability is the type I error rate for the test of non-inferiority, and can be written in the following manner: α = type I error = pr(reject H_O: $\theta=\pi_T/\pi_C \geq \theta_0$ | H_O: $\theta \geq \theta_0$ true).

A summary of the type I error of the methods generated from the 100,000 simulations is displayed in Figure 1.1 for the Taylor Series methods, Figure 1.12 for the quadratic methods, and Figure 1.3 for the maximum likelihood methods. Farrington-Manning method 1 is dropped from summaries due to its similarities to the Quadratic method. Displays include only the α=0.025 level with similar patterns seen for the other alpha levels.

The performance of the methods with respect to the type I error varies in relation to the sample size allocation of treatment to control. All of the Taylor series expansion methods have approximately nominal type I error rates for the 1:2 allocation. However, as more sample size is placed in the test group, the type I error rates become inflated higher than the nominal level. The Adapted Agresti method yields type I error rates closest to the nominal level, but this method still shows inflation for the 3:2, 2:1, and 3:1 allocations.

Out of the group of quadratic methods, the Quadratic method and Noether's method have type I error rates that are consistently below the nominal level for all allocations. However, Bailey's, F-M 2, and F-M 3 have appropriate type I error rates for the 1:2 and 1:1 allocations with higher than nominal type I error rates for the 3:2, 2:1, and 3:1 scenarios.

The group of maximum likelihood methods perform similarly for the 1:2 allocation, with type I errors approximately nominal or just slightly higher than nominal. The Deviance method performs adequately for all sample size allocation scenarios with type I errors close to the nominal level. The Pearson method has slightly inflated type I errors for all other scenarios. The group of power divergence methods yields higher type I errors as λ increases with λ=-0.5 yielding lower than nominal type I errors and λ=1.25 yielding higher than nominal type I errors.

Figure 1.4 provides a graphical summary of the methods with better type I error performance including the Taylor series method, Adapted Agresti method, Bailey's method, and Deviance method. Discussions of power will be limited to these methods for scenarios where the simulated type I error is appropriately controlled.

The Deviance method seems to perform appropriately for all sample size allocation scenarios, with slightly higher type I errors for the 1:2 allocation. Figure 1.5 compares the Taylor Series power to the Deviance power for the 1:2 allocation, for the null hypothesis θ_0=2. These methods tend to perform similarly in this setting. Figure 1.6 is a comparison of the Adapted Agresti and Deviance simulated powers for the allocations including 1:2 and

1:1. The Deviance method produces similar or slightly higher simulated powers in these scenarios. Figure 1.7 displays Bailey’s method compared to the Deviance method for the 1:1 allocation setting, also showing similar simulated powers between the two methods.

These findings suggest that in the 1:2 or 1:1 allocation settings, the simpler Taylor Series or Adapted Agresti methods perform similarly to the computer intensive Deviance method with respect to power. However, the Deviance method may be the preferred method for allocations with more sample size in the test group in order to maintain the nominal type I error level.

VI. Sample Size Calculations

An immediate application of these results arises in the design of non-inferiority clinical trials. The Taylor Series method provides a fairly straightforward form from which to obtain sample size calculations. A conservative form of the variance is seen in (1.11).

$$\begin{aligned}&\operatorname{var}\left\{\log_e \frac{y_T / n_T}{y_C / n_C}\right\} \approx \left\{\frac{1}{y_T}+\frac{1}{y_C}-\frac{1}{n_T}-\frac{1}{n_C}\right\} = \\ &\quad \left\{\frac{1}{n_C\pi_C}\left(\frac{1}{\theta R}+1\right)-\frac{1}{n_T}-\frac{1}{n_C}\right\} < \left\{\frac{1}{n_C\pi_C}\left(\frac{1}{\theta R}+1\right)\right\} = v^*\end{aligned} \tag{1.11}$$

Motivation for obtaining a sample size formula begins with formulation of a z-statistic in (1.12) where θ_0 is the value of θ under the null hypothesis.

$$\mathrm{z} = \frac{\log_{\mathrm{e}} \theta - \log_{\mathrm{e}} \theta_0}{\sqrt{\mathrm{v}^*}} \tag{1.12}$$

The equation in (1.13) results from squaring equation (1.12) and writing z in terms of the type I and type II errors which produces equation (1.14) after algebraic manipulations.

$$(z_\alpha + z_\beta)^2 = \frac{\left(\log_e \frac{\theta}{\theta_0}\right)^2}{v^*} \tag{1.13}$$

$$\frac{(z_\alpha + z_\beta)^2 \left\{ \frac{1}{n_C \pi_C} \left(\frac{1}{\theta R} + 1 \right) \right\}}{\left\{ \log_e \left(\frac{\theta}{\theta_0} \right) \right\}^2} = 1 \tag{1.14}$$

This form is then solved for the sample size, n_C, and can be written as in (1.15)

$$n_C = \frac{(z_\alpha + z_\beta)^2 \left\{ \frac{1}{\pi_C} \left(\frac{1}{\theta R} + 1 \right) \right\}}{\left\{ \log_e \left(\frac{\theta}{\theta_0} \right) \right\}^2} \tag{1.15}$$

which depends only on a pre-specified one-sided type I error (α), power (1-β), event rate in the control group (π_C), the sample size allocation ($R=n_T/n_C$), and a hypothesized ratio of events in the treatment versus the control group (θ) with θ_0, the null hypothesis, specified. This formula is useful in practice due to ease of computation.

To evaluate whether formula (1.15) produces sample sizes that maintain the pre-specified power, results were compared to those obtained from simulations. These results were based on 100,000 simulations. The sample size formula (1.15) was written in terms of power as seen in equation (1.16)

$$z_\beta = \frac{\sqrt{n_C} \log_e \left(\frac{\theta}{\theta_0} \right)}{\sqrt{\left\{ \frac{1}{\pi_C} \left(\frac{1}{\theta R} + 1 \right) \right\}}} - z_\alpha \tag{1.16}$$

where power = $\Phi(z_\beta)$ and $\Phi(.)$ is the standard normal probability. In addition, this sample size formula and power calculation in (1.15) and (1.16) can be modified for the Taylor Series Adjusted Alpha method which controls type I error better than the Taylor Series method (although in allocations with more subjects in the test group, the type I error is still above the nominal level). This adjustment uses an alpha level of 0.0025 lower than that specified. For example, at a specified α=0.025 the critical value would be calculated at 0.025-0.0025=0.0225.

In addition Farrington and Manning[4] present sample size formula (1.17) and power formula (1.18) based on their methods.

$$n_T = \frac{\left\{z_\alpha\sqrt{\tilde{\pi}_T(1-\tilde{\pi}_T)+R\theta_0^2\tilde{\pi}_C(1-\tilde{\pi}_C)}+z_\beta\sqrt{\pi_T(1-\pi_T)+R\theta_0^2\pi_C(1-\pi_C)}\right\}^2}{\left(\pi_T-\theta_0\pi_C\right)^2} \qquad (1.17)$$

$$z_\beta = \frac{\left\{\sqrt{n_T}(\pi_T-\theta_0\pi_C)-z_\alpha\sqrt{\tilde{\pi}_T(1-\tilde{\pi}_T)+R\theta_0^2\tilde{\pi}_C(1-\tilde{\pi}_C)}\right\}}{\sqrt{\pi_T(1-\pi_T)+R\theta_0^2\pi_C(1-\pi_C)}} \qquad (1.18)$$

where $\tilde{\pi}_T$ and $\tilde{\pi}_C$ are specified differently for each of the three methods. The first method presented by Farrington and Manning use $\tilde{\pi}_T = \pi_T$ and $\tilde{\pi}_C = \pi_C$. The second of these methods uses the following values:

$$\tilde{\pi}_T = \frac{\theta_0\left(n_T\pi_T + n_C\pi_C\right)}{(\theta_0 n_T + n_C)} \text{ and } \tilde{\pi}_C = \frac{(n_T\pi_T + n_C\pi_C)}{(\theta_0 n_T + n_C)}$$

Farrington-Manning method 3 replaces $\tilde{\pi}_T$ and $\tilde{\pi}_C$ using the following equations:

$$\tilde{\pi}_T = \frac{-b-\sqrt{b^2-4ac}}{2a} \text{ and } \tilde{\pi}_C = \tilde{\pi}_T / R$$

where $a = 1+\frac{n_C}{n_T}$, $b = -\left\{\theta_0\left(1+\frac{n_C}{n_T}\pi_C\right)+\frac{n_C}{n_T}+\pi_T\right\}$, and $c = \theta_0\left(\pi_T+\frac{n_C}{n_T}\pi_C\right)$.

For combinations of α, π_C, θ, R, and n_C generated in the 100,000 simulations, a power based on the sample size formula was calculated using formula (1.16) and the Taylor Series Adjusted Alpha formula. Power was also calculated using the sample size formulas presented by Farrington and Manning for F-M method 1, F-M method 2, and F-M method 3. This calculated power was then compared to the power obtained from the simulations for each method.

Figures 1.8 – 1.14 graphically display the comparison between the calculated and simulated power for the Taylor Series, Taylor Series Adjusted Alpha, F-M 1, F-M 2, and F-M 3 methods, at an alpha level of 0.025 for a null hypothesis θ_0=2. For most cases, the simulated power is similar to or larger than the calculated power; therefore the sample size formulas are somewhat conservative which is beneficial when determining sample size for clinical trials.

The Deviance method does not have a corresponding sample size formula, therefore the simulated power from this method is compared to the calculated Taylor series power in Figure 1.3f and the calculated F-M 3 power in Figure 1.3g. The simulated Deviance power is both larger and smaller than the Taylor Series calculated power for specific scenarios. However, the calculated F-M 3 power seems to agree consistently with the Deviance power for scenarios with power higher than 0.80. For power values lower than 0.80, the F-M 3 calculated power yields slightly higher values. However, when planning a trial it is usually necessary to have at least 0.80 power and in these cases the F-M 3 calculations would be appropriate.

VII. Conclusions

The properties and behavior of many different methods for computing confidence intervals for risk ratios have been reviewed. The performance of the methods tends to vary according to the sample size allocation. The Deviance method seems to consistently perform at the nominal type I error level for most settings, and specifically for allocations with equal sample sizes or more sample size in the test group compared to the control group. The Taylor series method and the Adapted Agresti method tend to perform fairly well for the 1:2 and 1:1 sample size allocation settings while being easier to implement.

Problems due to small event rates were avoided due to use of methods for exact odds ratios. Due to this modification, the performance of the methods may be slightly different than that presented previously in the literature.

The straightforward sample size formula for the Taylor Series method makes it attractive for use in designing non-inferiority clinical trials, but may be appropriate for scenarios with allocations of 1:2 or 1:1. The sample size formula presented by Farrington and Manning based on method 3 is useful in trial design for the other allocation settings and agrees with the Deviance method for trials designed with fairly high power.

References

1. Hung, H. M. J., Wang, S. J., Tsong Y., Lawrence, J., and O'Neill, R. T. Some Fundamental Issues with Non-Inferiority Testing in Active Controlled Trials. Statistics in Medicine **2003**, 22, 213-225.

2. Blackwelder, W. C. "Proving the Null Hypothesis" in Clinical Trials. Controlled Clinical Trials **1982**, 3: 345-353.

3. *nQuery Advisor Version 5.0 User's Guide*. Statistical Solutions Ltd.: Cork, Ireland, 2002.

4. Farrington, C. P., and Manning, G. Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-zero Risk Difference or Non-unity Relative Risk. Statistics in Medicine **1990**, 9, 1447-1454.

5. Katz, D., Baptista, J., Azen, S. P., and Pike, M. C. Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies. Biometrics **1978**, 34, 469-474.

6. *SAS Online Doc®, Version 8*. SAS Institute, Inc: Cary, NC, 1999.

7. EquivTest 1.0. *Software for the statistical analysis of equivalence and bioavailability studies*. Statistical Solutions Ltd.: Cork, Ireland, 2000.

8. Gart, J. J., and Nam, J. Approximate Interval Estimation of the Ratio of Binomial Parameters: A Review and Corrections for Skewness. Biometrics **1988**, 44, 323-338.

9. *StatXact4 for Windows User Manual*. CYTEL Software Corporation: Cambridge, MA, 1990, 435-452.

10. Agresti, A. and Coull, B. A. Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions. The American Statistician **1998**, 52, 119-126.

11. Fieller, E. C. A Fundamental Formula in the Statistics of Biological Assay and Some Applications. Quarterly Journal of Pharmacy and Pharmacology **1944**, 17, 117-123.

12. Bailey, B. J. R. Confidence Limits to the Risk Ratio. Biometrics **1987**, 43, 201-205.

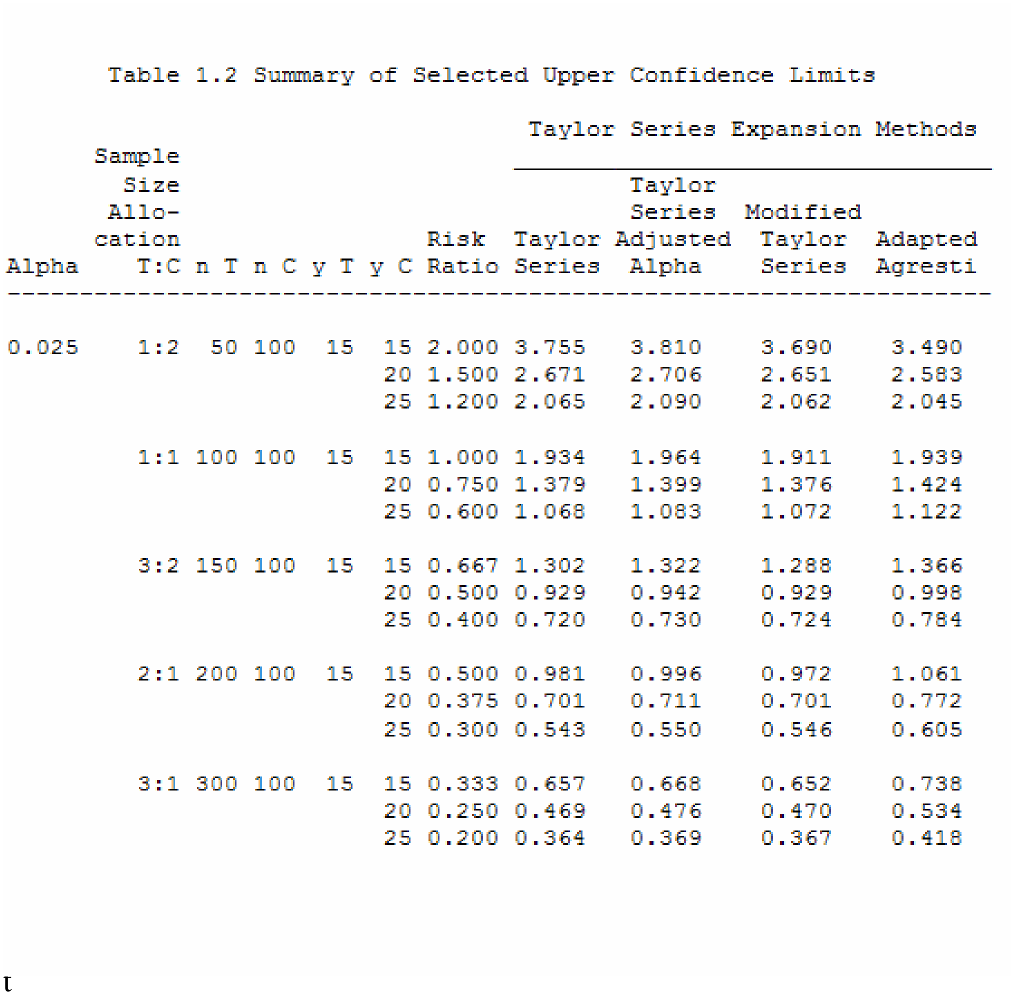
13. Koopman, P. A. R. Confidence Intervals for the Ratio of Two Binomial Proportions. Biometrics **1984**, 40, 513-517.

14. Bedrick, E. J. A Family of Confidence Intervals for the Ratio of Two Binomial Proportions. Biometrics **1987**, 43, 993-998.

Table 1.1 Software Resources for Methods

Method	Software Resources
Taylor Series	SAS using PROC FREQ[6] EquivTest[7]
Taylor Series Adjusted Alpha	SAS using PROC FREQ[6] †
Modified Taylor Series	SAS using PROC FREQ[6] ‡ StatXact[9]
Adapted Agresti	SAS using PROC FREQ[6] ‡
Quadratic	No resources available
Farrington-Manning 1	No resources available
Farrington-Manning 2	No resources available
Farrington-Manning 3	No resources available
Bailey	No resources available
Noether	No resources available
Deviance	SAS using PROC GENMOD[6] €
Pearson	SAS using PROC GENMOD[6] €
Power Divergence	No resources available

† The alpha level can be modified to produce this interval
‡ Event counts can be modified to produce this interval
€ Additional programming is required to use this computer resource



```
       Table 1.2 Summary of Selected Upper Confidence Limits

                                                    Taylor Series Expansion Methods
      Sample                                      _________________________________
       Size                                               Taylor
      Allo-                                               Series  Modified
      cation                        Risk  Taylor Adjusted  Taylor  Adapted
Alpha    T:C n T n C y T y C Ratio Series  Alpha    Series  Agresti
--------------------------------------------------------------------

0.025    1:2  50 100  15  15 2.000 3.755  3.810    3.690    3.490
                          20 1.500 2.671  2.706    2.651    2.583
                          25 1.200 2.065  2.090    2.062    2.045

         1:1 100 100  15  15 1.000 1.934  1.964    1.911    1.939
                          20 0.750 1.379  1.399    1.376    1.424
                          25 0.600 1.068  1.083    1.072    1.122

         3:2 150 100  15  15 0.667 1.302  1.322    1.288    1.366
                          20 0.500 0.929  0.942    0.929    0.998
                          25 0.400 0.720  0.730    0.724    0.784

         2:1 200 100  15  15 0.500 0.981  0.996    0.972    1.061
                          20 0.375 0.701  0.711    0.701    0.772
                          25 0.300 0.543  0.550    0.546    0.605

         3:1 300 100  15  15 0.333 0.657  0.668    0.652    0.738
                          20 0.250 0.469  0.476    0.470    0.534
                          25 0.200 0.364  0.369    0.367    0.418
```

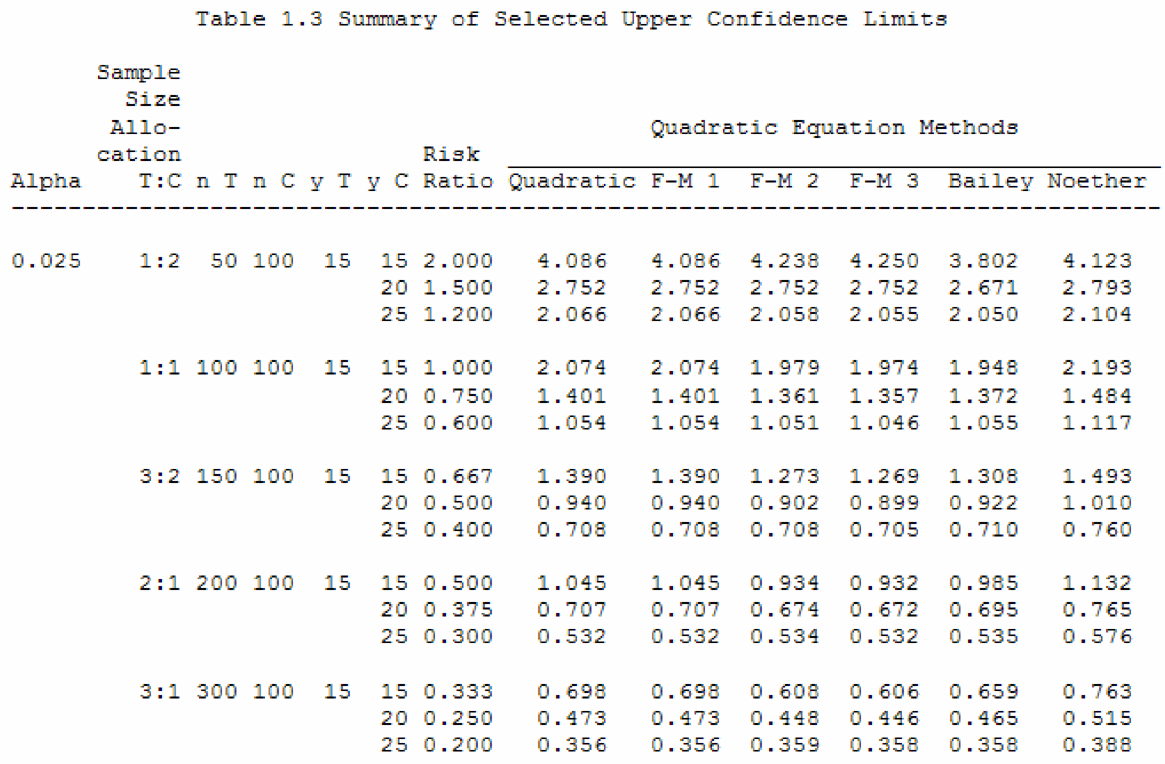


Table 1.3 Summary of Selected Upper Confidence Limits

Alpha	Sample Size Allocation T:C	n T	n C	y T	y C	Risk Ratio	Quadratic Equation Methods					
							Quadratic	F-M 1	F-M 2	F-M 3	Bailey	Noether
0.025	1:2	50	100	15	15	2.000	4.086	4.086	4.238	4.250	3.802	4.123
					20	1.500	2.752	2.752	2.752	2.752	2.671	2.793
					25	1.200	2.066	2.066	2.058	2.055	2.050	2.104
	1:1	100	100	15	15	1.000	2.074	2.074	1.979	1.974	1.948	2.193
					20	0.750	1.401	1.401	1.361	1.357	1.372	1.484
					25	0.600	1.054	1.054	1.051	1.046	1.055	1.117
	3:2	150	100	15	15	0.667	1.390	1.390	1.273	1.269	1.308	1.493
					20	0.500	0.940	0.940	0.902	0.899	0.922	1.010
					25	0.400	0.708	0.708	0.708	0.705	0.710	0.760
	2:1	200	100	15	15	0.500	1.045	1.045	0.934	0.932	0.985	1.132
					20	0.375	0.707	0.707	0.674	0.672	0.695	0.765
					25	0.300	0.532	0.532	0.534	0.532	0.535	0.576
	3:1	300	100	15	15	0.333	0.698	0.698	0.608	0.606	0.659	0.763
					20	0.250	0.473	0.473	0.448	0.446	0.465	0.515
					25	0.200	0.356	0.356	0.359	0.358	0.358	0.388

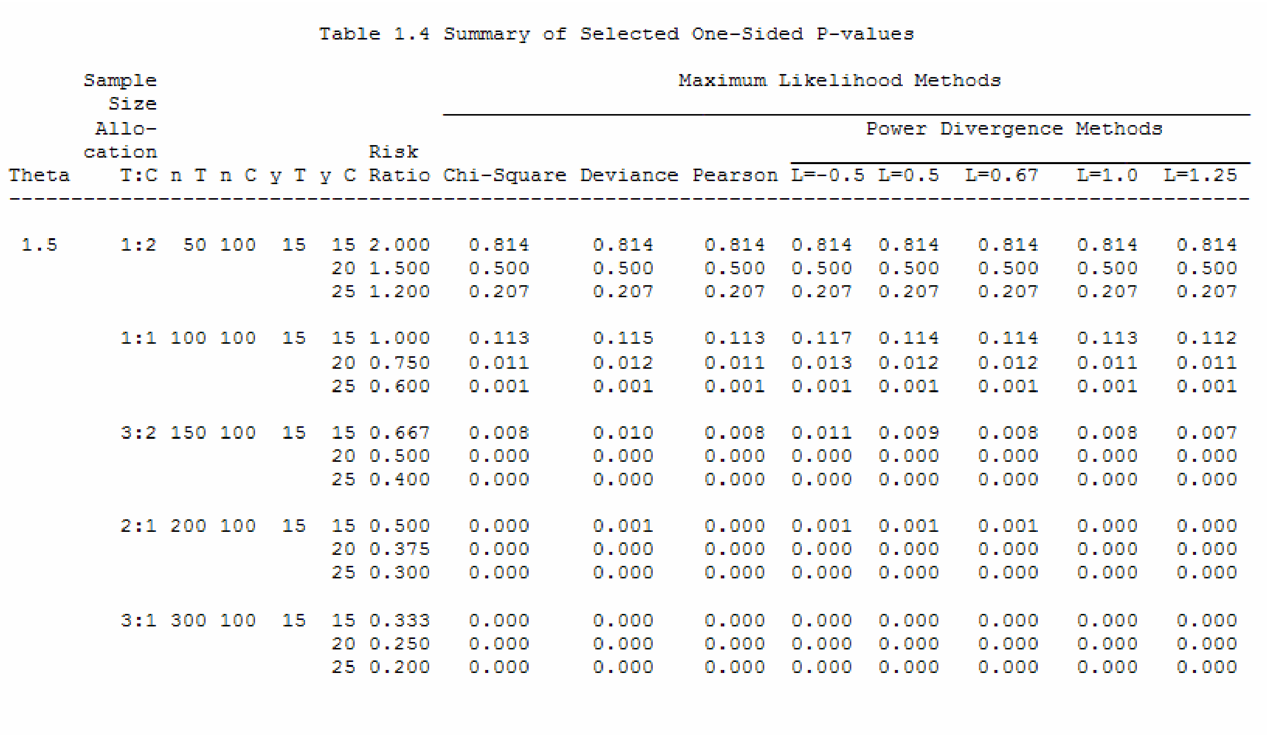


Table 1.4 Summary of Selected One-Sided P-values

							Maximum Likelihood Methods							
										Power Divergence Methods				
Theta	Sample Size Allo-cation T:C	n T	n C	y T	y C	Risk Ratio	Chi-Square	Deviance	Pearson	L=-0.5	L=0.5	L=0.67	L=1.0	L=1.25
1.5	1:2	50	100	15	15	2.000	0.814	0.814	0.814	0.814	0.814	0.814	0.814	0.814
					20	1.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
					25	1.200	0.207	0.207	0.207	0.207	0.207	0.207	0.207	0.207
	1:1	100	100	15	15	1.000	0.113	0.115	0.113	0.117	0.114	0.114	0.113	0.112
					20	0.750	0.011	0.012	0.011	0.013	0.012	0.012	0.011	0.011
					25	0.600	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	3:2	150	100	15	15	0.667	0.008	0.010	0.008	0.011	0.009	0.008	0.008	0.007
					20	0.500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
					25	0.400	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2:1	200	100	15	15	0.500	0.000	0.001	0.000	0.001	0.001	0.001	0.000	0.000
					20	0.375	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
					25	0.300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	3:1	300	100	15	15	0.333	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
					20	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
					25	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 1.1 Summary of Simulated Type I Error

Taylor Series Expansion Methods
Alpha=0.025

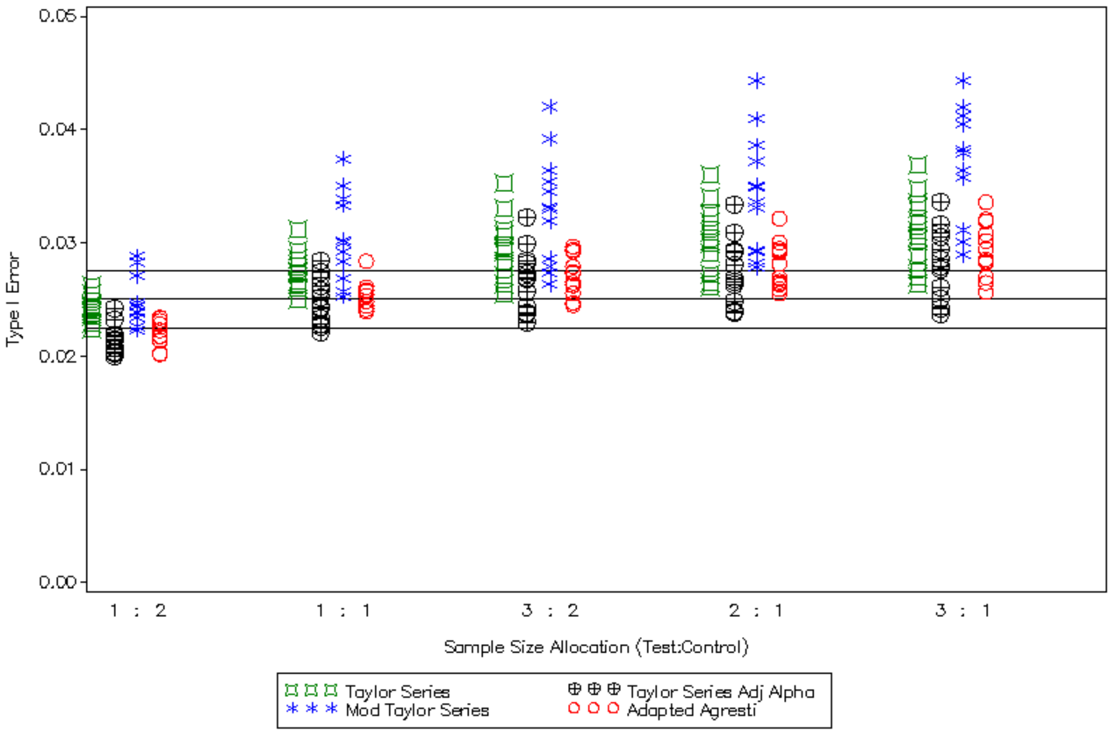


Figure 1.2 Summary of Simulated Type I Error

Quadratic Equation Methods
Alpha=0.025

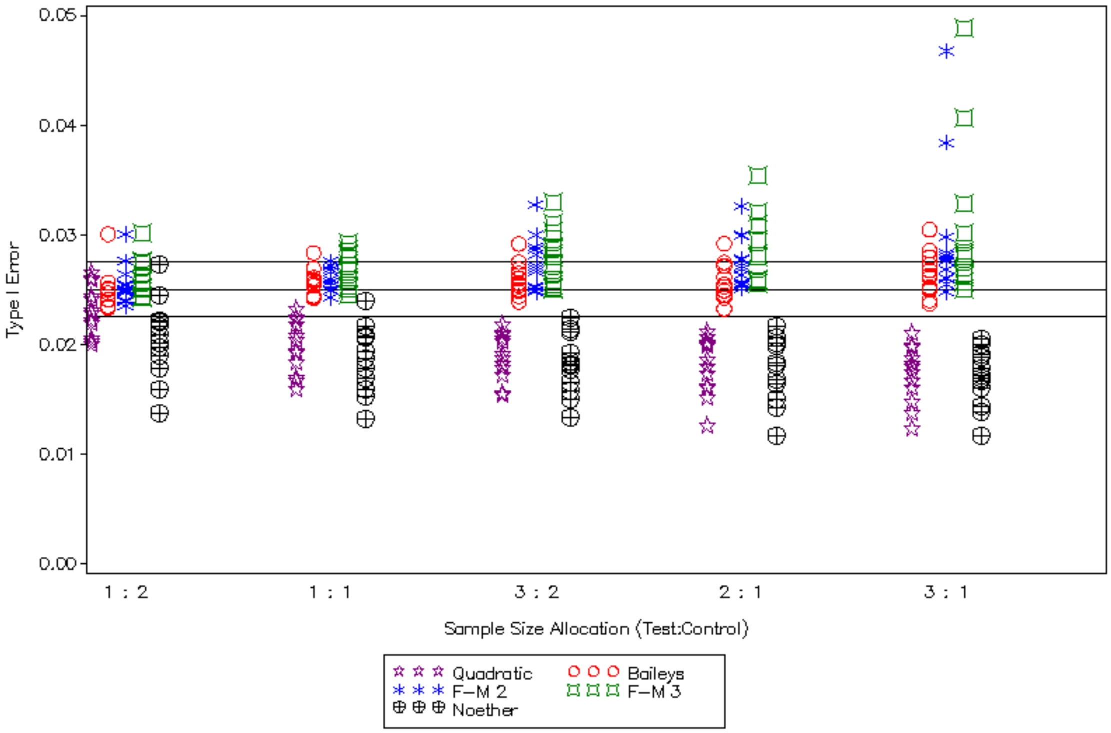


Figure 1.3 Summary of Simulated Type I Error
Maximum Likelihood Methods
Alpha=0.025

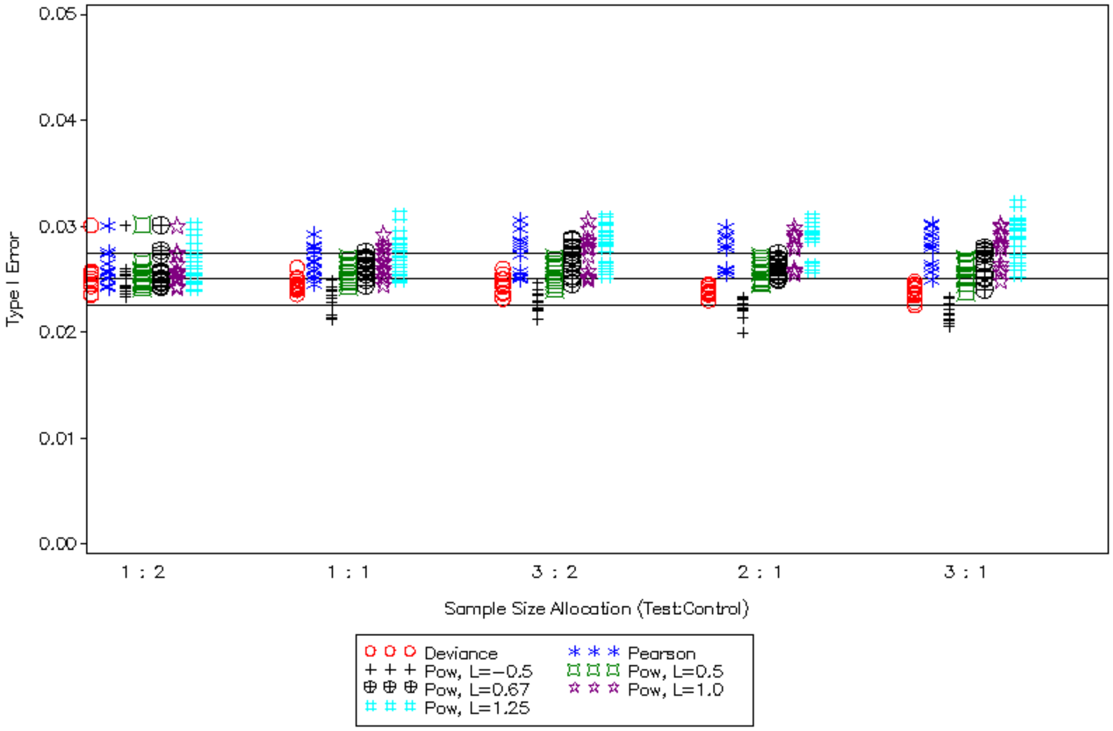


Figure 1.4 Summary of Simulated Type I Error
Overview of Better Methods
Alpha=0.025

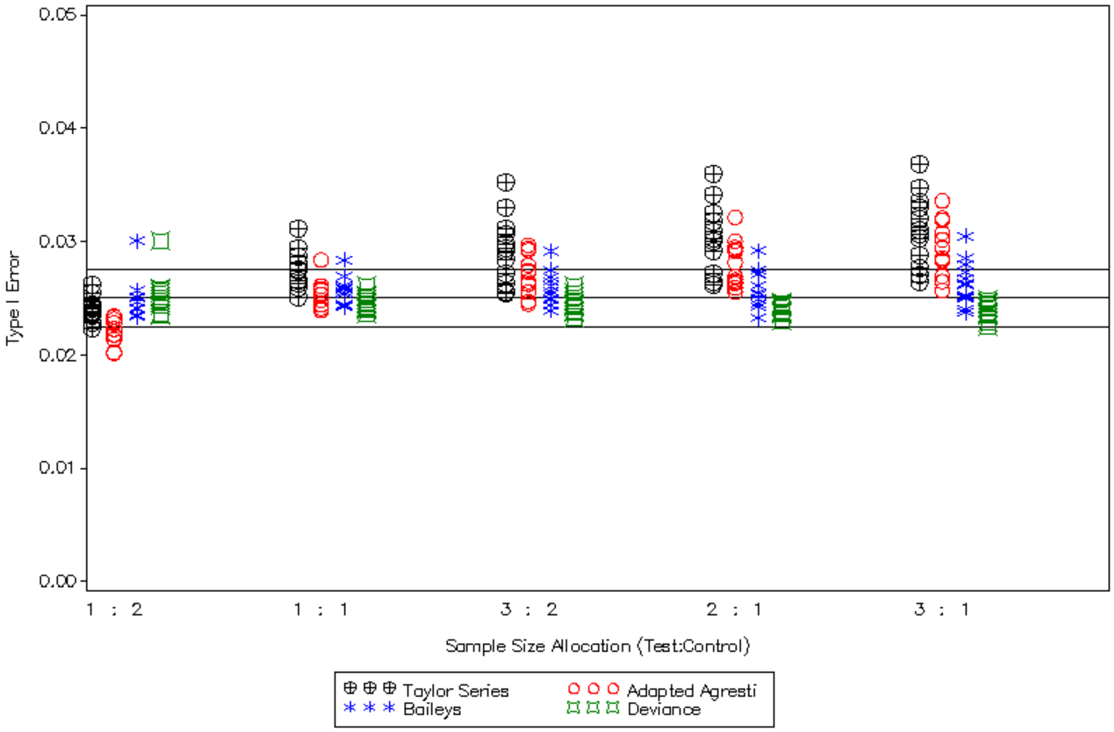


Figure 1.5 Comparison of Simulated Power

Deviance and Taylor Series Methods
By Population Risk Ratio
Alpha=0.025, Theta:H0=2, Allocation 1:2

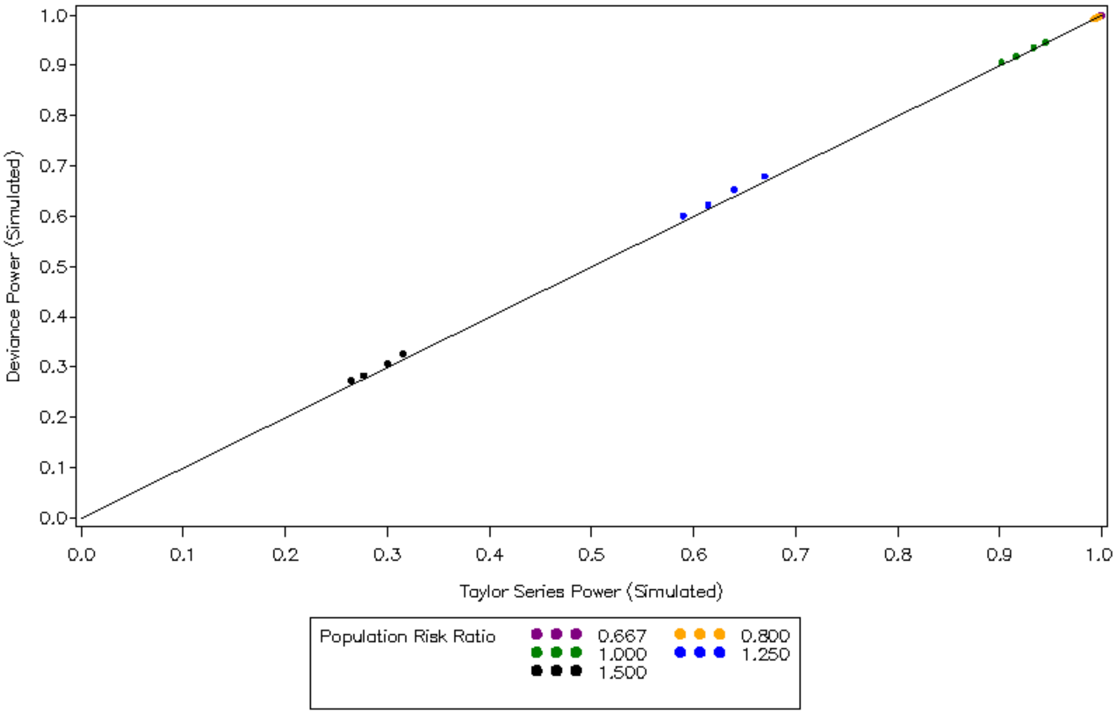


Figure 1.6 Comparison of Simulated Power

Deviance and Adapted Agresti Methods
By Population Risk Ratio
Alpha=0.025, Theta:H0=2, Allocation 1:2, 1:1

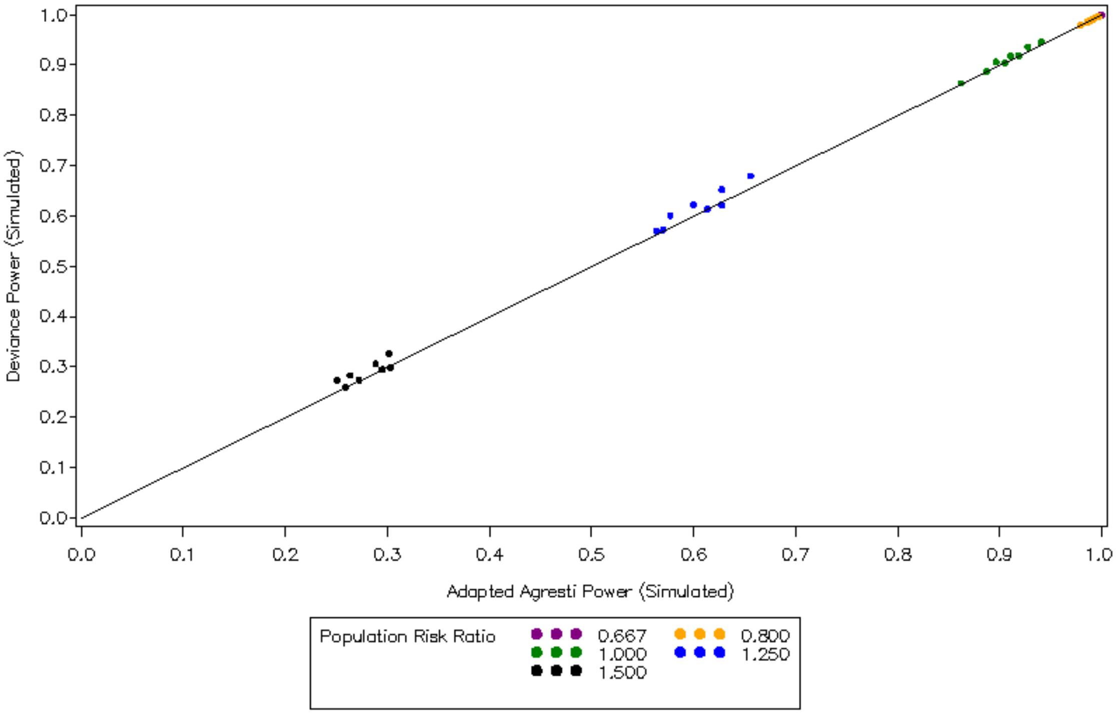


Figure 1.7 Comparison of Simulated Power

Deviance and Bailey Methods
By Population Risk Ratio
Alpha=0.025, Theta:H0=2, Allocation 1:1

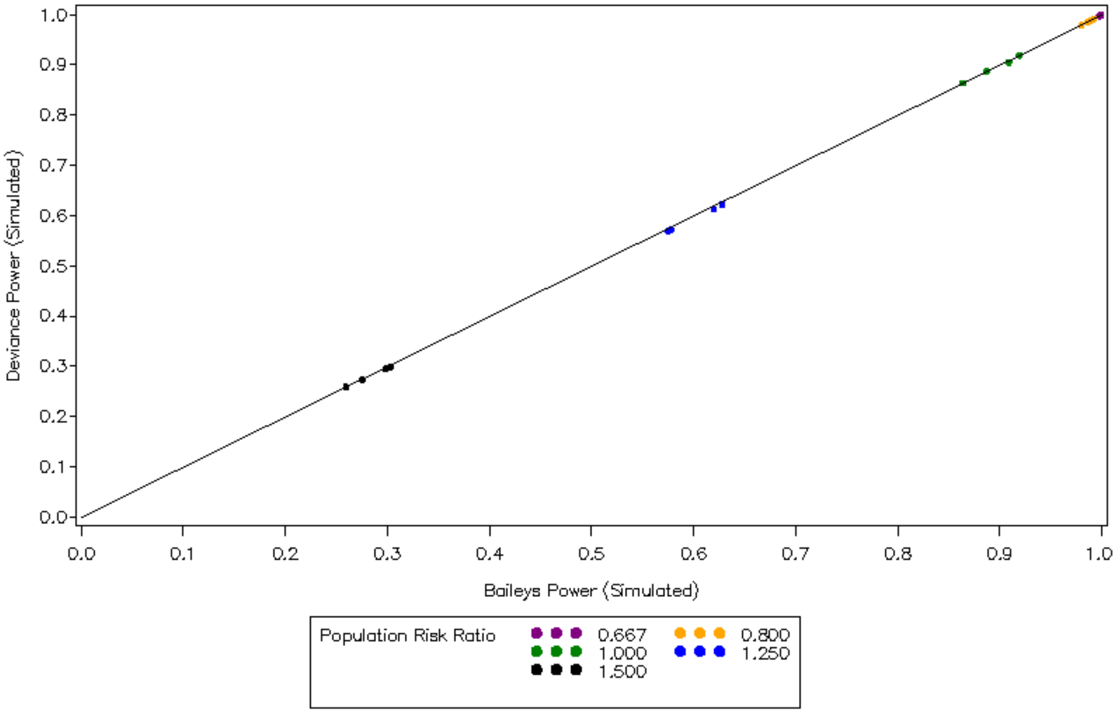


Figure 1.8 Comparison of Simulated and Calculated Power

Taylor Series Method
By Population Risk Ratio
Alpha=0.025, Theta:H0=2

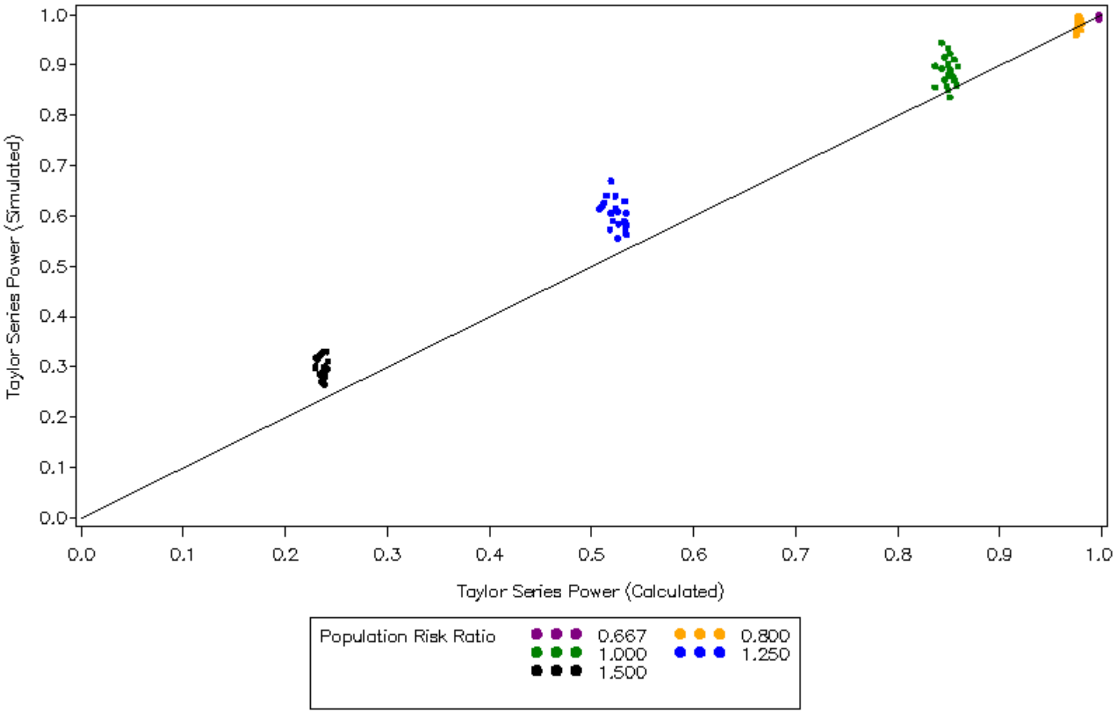


Figure 1.9 Comparison of Simulated and Calculated Power
Taylor Series Adjusted Alpha Method
By Population Risk Ratio
Alpha=0.025, Theta:H0=2

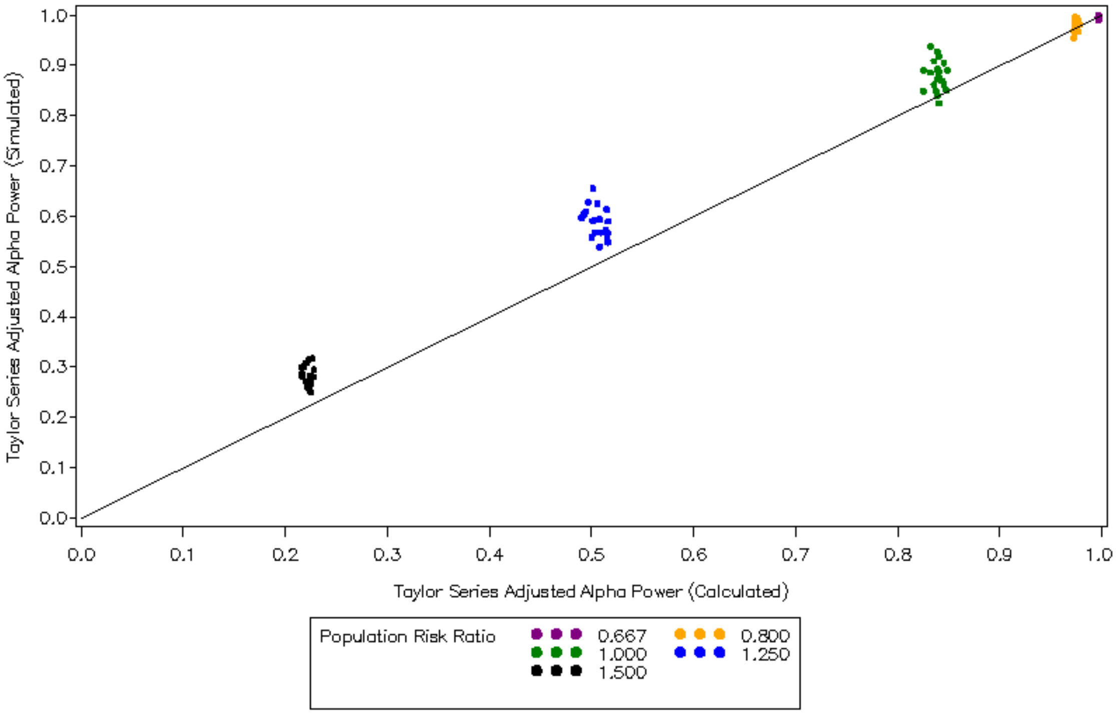


Figure 1.10 Comparison of Simulated and Calculated Power
Farrington–Manning 1 Method
By Population Risk Ratio
Alpha=0.025, Theta:H0=2

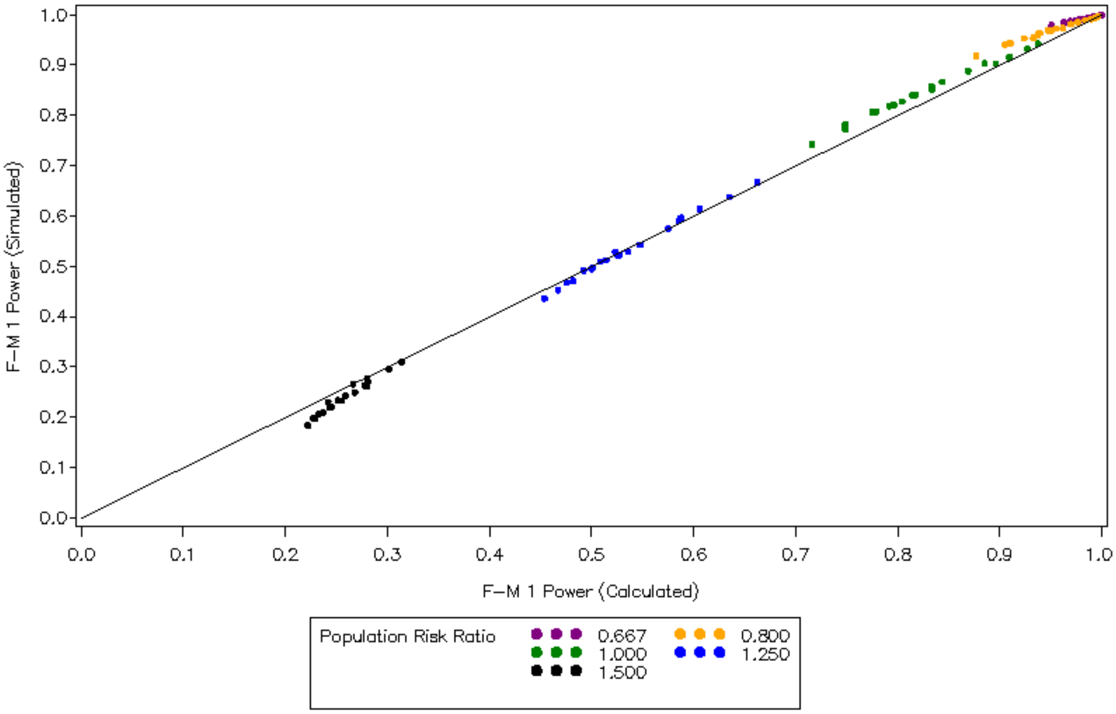


Figure 1.11 Comparison of Simulated and Calculated Power

Farrington–Manning 2 Method
By Population Risk Ratio
Alpha=0.025, Theta:H0=2

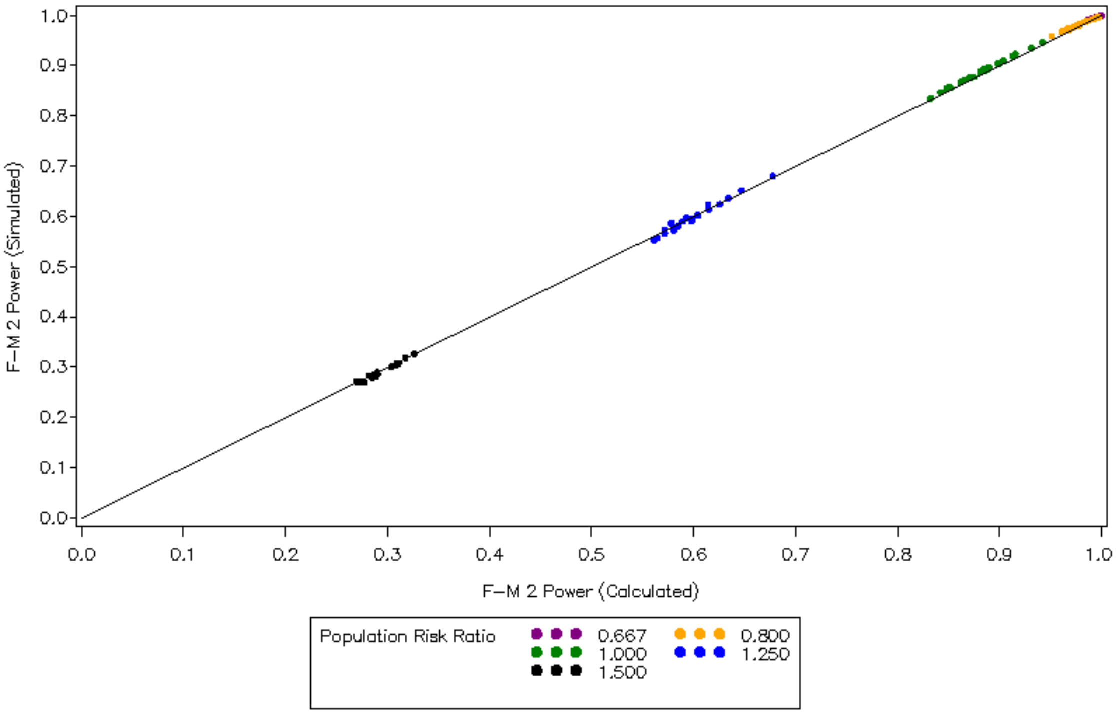


Figure 1.12 Comparison of Simulated and Calculated Power

Farrington–Manning 3 Method
By Population Risk Ratio
Alpha=0.025, Theta:H0=2

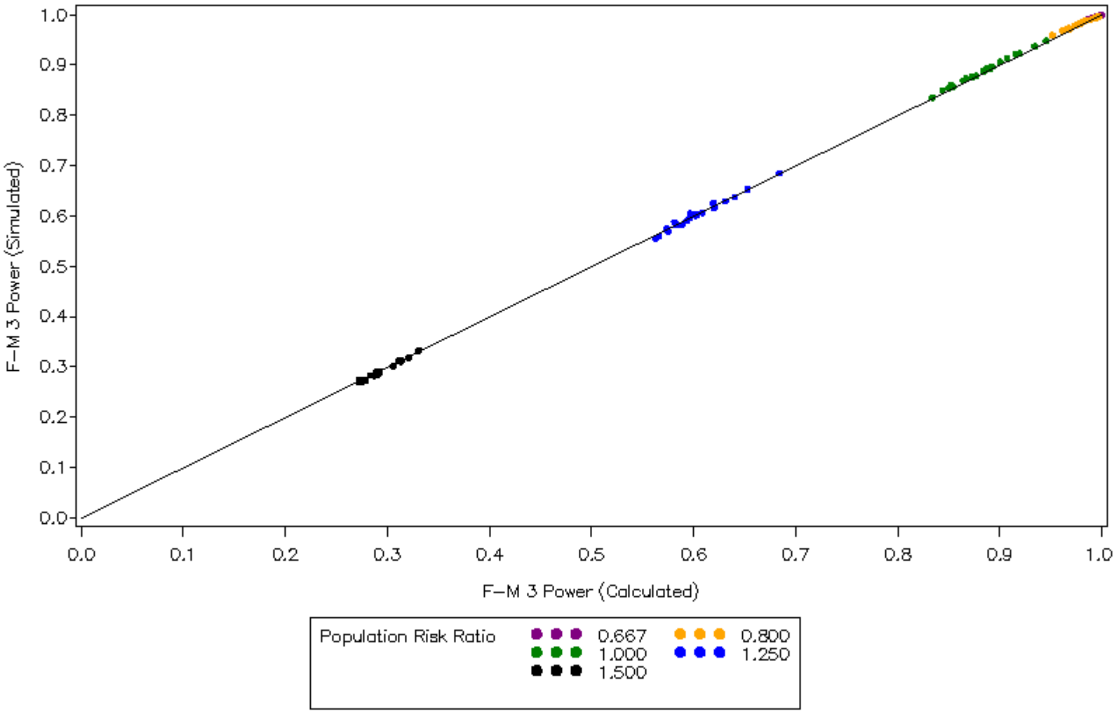


Figure 1.13 Comparison of Simulated and Calculated Power
Deviance Simulated Method & Taylor Series Calculated Method
By Population Risk Ratio
Alpha=0.025, Theta:H0=2

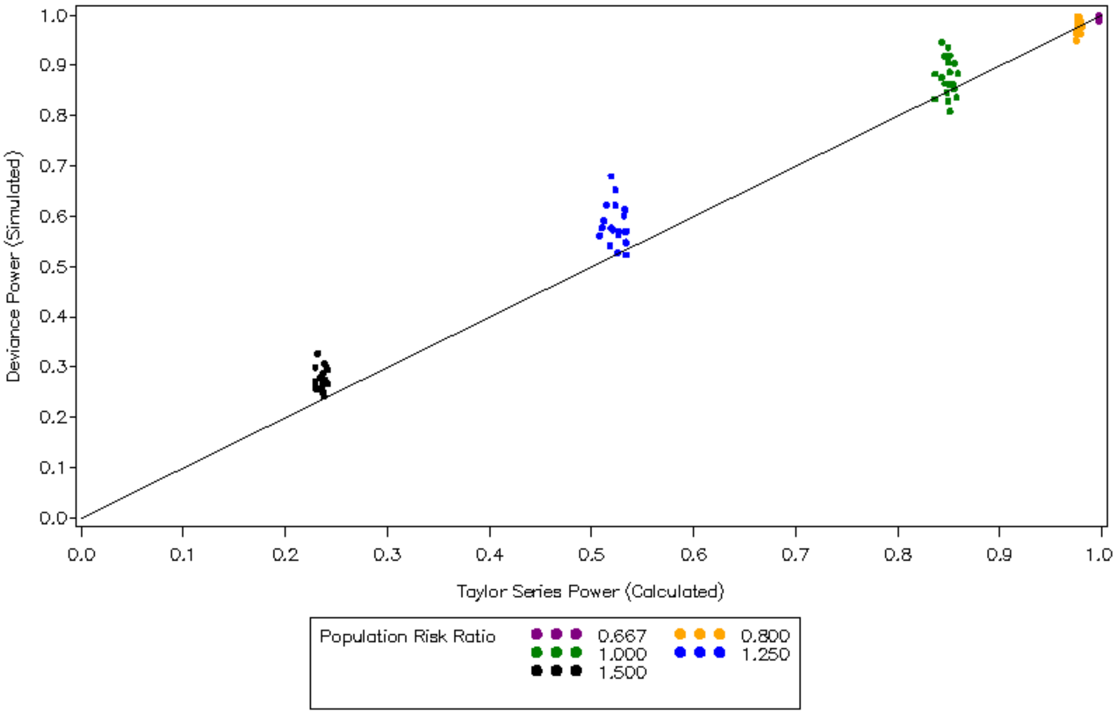
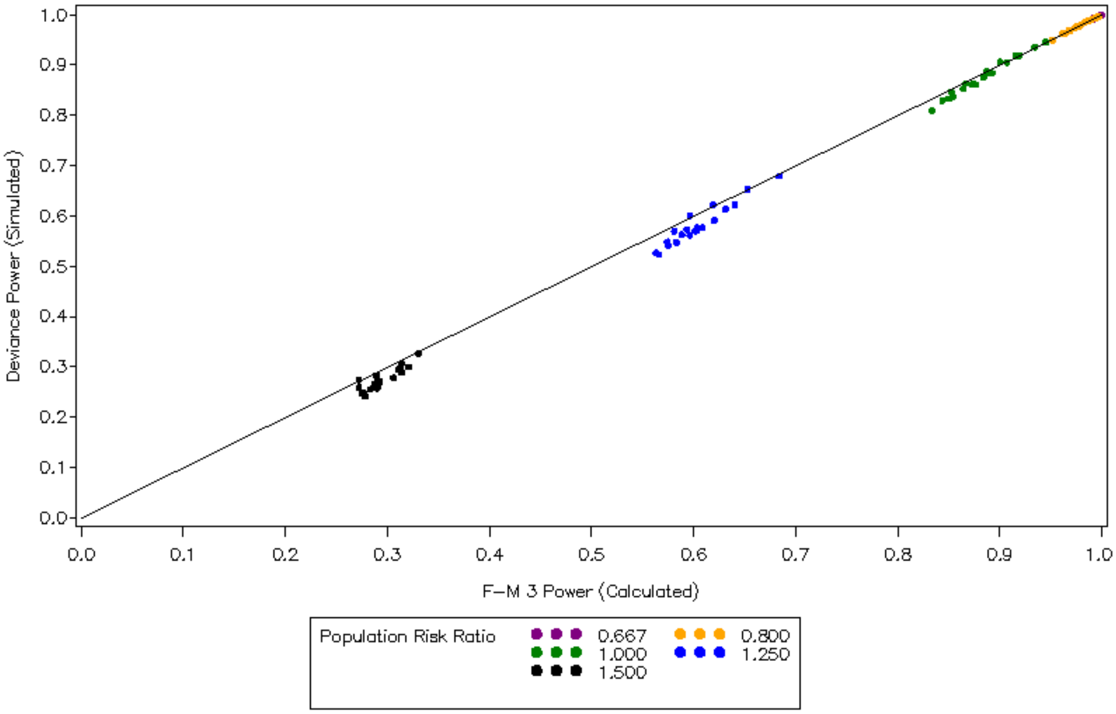


Figure 1.14 Comparison of Simulated and Calculated Power
Deviance Simulated Method & F-M 3 Calculated Method
By Population Risk Ratio
Alpha=0.025, Theta:H0=2



Chapter 2

Review of Methods for One-Sided Testing of the Difference between Proportions and Sample Size Considerations

I. Introduction

Proportions are used in many clinical trials to describe the distributions of dichotomous response variables with independent binomial distributions for treatments under study. Comparisons between treatment groups are often made through one-sided confidence intervals on the difference in the two treatment group proportions. In a non-inferiority setting, the goal is to show that the investigational treatment (test) group is no worse than an active control group by a predetermined non-inferiority margin. For proportions pertaining to favorable response, the lower confidence bound on the difference between the test and control groups must usually be larger than this margin in order to conclude that the test treatment is not inferior to the active control with respect to efficacy. Such a lower confidence bound must exceed zero to demonstrate superiority.

There are many methods in the statistical literature for computing the confidence interval for the difference between two independent binomial proportions. However, each method has both advantages and disadvantages to its use. It is important to understand in which settings these methods are more useful and appropriate. Scenarios include a wide range of proportions of favorable response of 0.60 and higher.

II. Review of Methods in the Literature

Historically, the most well known method for computing the confidence interval for a difference between two independent binomial proportions is the Wald method as based on a normal approximation. This method as seen in (2.1) is easy for computation and understanding, and so it is presented in most basic statistics textbooks and implemented in standard statistical software packages.

$$\{\hat{p}_T - \hat{p}_C\} \pm z_\alpha \sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}} \qquad (2.1)$$

In (2.1), $\hat{p}_T = y_T/n_T$ is the observed proportion for favorable response in the test group with y_T representing the total number of such outcomes in a total sample size of n_T and $\hat{p}_C = y_C/n_C$ is the observed proportion for favorable response in the control group with y_C representing the total number of such outcomes in a total sample size of n_C. Also, z_α is the (1-α) quantile of the standard normal distribution. This method has traditionally been shown to have poor performance for even moderate sample sizes with respect to excessive inflation of the type I error rate when consideration includes the entire confidence interval using both the upper and lower bounds[1,2]. However, Roebruck and Kühn[3] found this method to perform adequately for sample sizes large enough to yield power of at least 0.70 where the sample size allocation for test:control is 3:2 for the one-sided limit as a one-sided test. Li and Chuang-Stein[4] also found this method to perform well in an equal allocation setting when event rates were moderate enough so as to provide expected cell frequencies of at least 15 for all cells.

For alleviation of some of these issues, a continuity corrected version of the Wald method is suggested as seen in (2.2).

$$\{\hat{p}_T - \hat{p}_C\} \pm \left\{ z_\alpha \sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}} + \frac{1}{2}\left[\frac{1}{n_T} + \frac{1}{n_C}\right] \right\} \tag{2.2}$$

This method should resolve the inflation of the type I error rate, while perhaps being overly conservative by reducing the power to test the hypothesis that the difference between treatment groups does not equal a predetermined value[2]. In addition, these methods are also presented using an unbiased estimate of the variance with (n_T-1) and (n_C-1) used in the denominators for $\hat{p}_T(1-\hat{p}_T)$ and $\hat{p}_C(1-\hat{p}_C)$, but these methods are rarely found in standard statistical software packages.

Agresti and Caffo[1] developed an adjustment to the Wald confidence interval to produce results that maintain the nominal type I error of a statistical test, analogous to a confidence interval including a predetermined value, while still being simple to calculate. This method as seen in (2.3) uses adjusted proportions when computing the confidence interval.

$$\{\tilde{p}_T - \tilde{p}_C\} \pm z_\alpha \sqrt{\frac{\tilde{p}_T(1-\tilde{p}_T)}{n_T + 2} + \frac{\tilde{p}_C(1-\tilde{p}_C)}{n_C + 2}} \tag{2.3}$$

In (2.3), $\tilde{p}_T = (y_T + 1)/(n_T + 2)$and $\tilde{p}_C = (y_C + 1)/(n_C + 2)$. These proportions are calculated by adding one success and one failure to each group and thereby two successes and two failures in total. Zhou et. al.[5] also agree that Agresti and Caffo’s method performs well at the 0.95 two-sided confidence level (for which it was designed), but the performance is unknown at other levels. In addition, Zhou et. al. point out that although the method is easy to implement, the theoretical properties of this method are still unknown.

Newcombe[2] provides a method based on the Wilson score method for a single proportion that is more complicated to compute than the Wald or Agresti and Caffo methods, but suggests that it has better coverage properties. This Newcombe hybrid score interval solves $|p_T - \hat{p}_T| = z_\alpha \sqrt{p_T(1-p_T)/n_T}$ for p_T resulting in two solutions, l_T and u_T. Similarly, the equation $|p_C - \hat{p}_C| = z_\alpha \sqrt{p_C(1-p_C)/n_C}$ is solved for p_C yielding solutions l_C and u_C. In addition, a continuity corrected version is proposed where the two solutions for p_T solve the following equation $|p_T - \hat{p}_T| - \frac{1}{2n_T} = z_\alpha \sqrt{p_T(1-p_T)/n_T}$ and the two solutions for p_C solve the following equation $|p_C - \hat{p}_C| - \frac{1}{2n_C} = z_\alpha \sqrt{p_C(1-p_C)/n_C}$. The lower and upper bounds of the interval are then computed as in (2.4) for both the Newcombe hybrid score and the continuity corrected version, using the solutions previously obtained.

$$\{\hat{p}_T - \hat{p}_C\} - \sqrt{(\hat{p}_T - l_T)^2 + (u_C - \hat{p}_C)^2}, \{\hat{p}_T - \hat{p}_C\} + \sqrt{(u_T - \hat{p}_T)^2 + (\hat{p}_C - l_C)^2} \tag{2.4}$$

Newcombe recommends this method over the Wald methods and the other methods reviewed in his paper[2] because of its performance with respect to coverage in the setting which involves both upper and lower confidence limits. In addition, Agresti and Caffo[1] suggest that this method is an appropriate method with the limitation of being more complicated to implement. Agresti and Caffo[1] also suggest that the Newcombe hybrid score method is an appropriate method except when proportions are close to 0 or 1. Zhou et. al.[5] use the Newcombe hybrid score as one of the best known methods in their paper, but suggest that its use be limited because the theoretical properties are not known.

Instead, Zhou et. al.[5] propose two new methods which are theoretically defensible and perform similarly to the Newcombe hybrid score and the Agresti and Caffo methods. These methods make adjustments to address limitations of using a normal approximation to estimate an interval when sample sizes may not be large or when the distribution of the data is skewed. The first proposed interval uses an Edgeworth expansion of the Wald statistic and corrects the interval using the error term to adjust for the skewed nature of the distribution. This Edgeworth expansion interval takes the form in (2.5).

$$\begin{gathered}
\{\hat{p}_T - \hat{p}_C\} - \left\{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}\right\}^{1/2} \left\{z_{1-\alpha} - (n_T + n_C)^{-1/2}\hat{Q}(z_{1-\alpha/2})\right\}, \\
\{\hat{p}_T - \hat{p}_C\} - \left\{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}\right\}^{1/2} \left\{z_{\alpha} - (n_T + n_C)^{-1/2}\hat{Q}(z_{\alpha/2})\right\}
\end{gathered} \tag{2.5}$$

The components of the interval in (2.5) are defined in (2.6).

$$\begin{gathered}
\hat{\mathrm{Q}}(\mathrm{t}) = \hat{\sigma}^{-1}(\hat{\mathrm{a}} + \hat{\mathrm{b}}\mathrm{t}^2), \quad \hat{\mathrm{a}} = \frac{\hat{\delta}}{6\hat{\sigma}^2}, \quad \hat{\mathrm{b}} = \frac{(\mathrm{n_T} + \mathrm{n_C})(1 - 2\hat{\mathrm{p}}_\mathrm{T})}{2\mathrm{n_T}} - \frac{\hat{\delta}}{6\hat{\sigma}^2}, \\
\hat{\sigma} = \left\{\frac{\mathrm{n_T} + \mathrm{n_C}}{\mathrm{n_T}}\hat{\mathrm{p}}_\mathrm{T}\left(1 - \hat{\mathrm{p}}_\mathrm{T}\right) + \frac{\mathrm{n_T} + \mathrm{n_C}}{\mathrm{n_C}}\hat{\mathrm{p}}_\mathrm{C}\left(1 - \hat{\mathrm{p}}_\mathrm{C}\right)\right\}^{1/2} \\
\hat{\delta} = \left(\frac{\mathrm{n_T} + \mathrm{n_C}}{\mathrm{n_T}}\right)^2 \hat{\mathrm{p}}_\mathrm{T}(1 - \hat{\mathrm{p}}_\mathrm{T})(1 - 2\hat{\mathrm{p}}_\mathrm{T}) - \left(\frac{\mathrm{n_T} + \mathrm{n_C}}{\mathrm{n_T}}\right)^2 \hat{\mathrm{p}}_\mathrm{C}(1 - \hat{\mathrm{p}}_\mathrm{C})(1 - 2\hat{\mathrm{p}}_\mathrm{C})
\end{gathered} \tag{2.6}$$

In addition, Zhou et. al.[5] propose an additional method to address skewness by using a Transformation approach as seen in the interval in (2.7).

$$\begin{gathered}
\{\hat{p}_T - \hat{p}_C\} - \left\{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}\right\}^{1/2} \left\{z_{1-\alpha} - (n_T + n_C)^{-1/2}\hat{g}^{-1}(z_{1-\alpha/2})\right\}, \\
\{\hat{p}_T - \hat{p}_C\} - \left\{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C}\right\}^{1/2} \left\{z_{\alpha} - (n_T + n_C)^{-1/2}\hat{g}^{-1}(z_{\alpha/2})\right\}
\end{gathered} \tag{2.7}$$

In (2.7), $\hat{g}^{-1}(t) = (n_T + n_C)^{1/2}(\hat{b}\hat{\sigma})^{-1}\left\{\left[1 + 3(\hat{b}\hat{\sigma})(n_T + n_C)^{-1/2}t - (n_T + n_C)^{-1}\hat{a}\hat{\sigma}\right]^{1/3} - 1\right\}$ and the additional components are as defined in the Edgeworth expansion method described in (2.6). Zhou et. el. suggest that the Edgeworth expansion method has poor coverage when proportions are near 0 or 1, but otherwise it is their recommended method for computing a confidence interval on the difference in proportions. They also suggest that this method has slightly better coverage properties than the Newcombe hybrid score method or the Agresti and Caffo method. When proportions are near 0 or 1, they recommend the Transformation method on the basis of similar coverage properties to the best known methods of Newcombe and Agresti and Caffo.

There are other methods presented by Newcombe[2] in his review of methods used to compute confidence intervals for the difference between proportions. These include a method attributed to Beal and Haldane as seen in the interval in (2.8).

$$\theta^* \pm w \tag{2.8}$$

$$with\ \ \theta^* = \frac{(\hat{p}_T - \hat{p}_C) + z_\alpha v(1-2\tilde{\psi})}{1 + z_\alpha^2 u}, \tilde{\psi} = \frac{1}{2}(\hat{p}_T + \hat{p}_C), u = \frac{1}{4}\left(\frac{1}{n_T} + \frac{1}{n_C}\right),$$

$$v = \frac{1}{4}\left(\frac{1}{n_T} - \frac{1}{n_C}\right) \text{ and } \ w = \frac{z_\alpha}{1 + z_\alpha^2 u}[u\{4\tilde{\psi}(1-\tilde{\psi}) - (\hat{p}_T - \hat{p}_C)^2\} + 2v(1-2\tilde{\psi})(\hat{p}_T - \hat{p}_C) + 4z_\alpha^2 u^2(1-\tilde{\psi})\tilde{\psi} + z_\alpha^2 v^2(1-2\tilde{\psi})^2]^{1/2}$$

In addition, an adjustment to the Beal's Haldane method is presented and identified as the Beal's Jeffreys-Perks interval and is similar to the Beal's Haldane method but with $\tilde{\psi} = \frac{1}{2}\left(\frac{y_T + 0.5}{n_T + 1} + \frac{y_C + 0.5}{n_C + 1}\right)$. These methods were developed in an attempt to fix the problems associated with the use of the Wald and Wald continuity corrected intervals. Newcombe[2] suggests that they provide improvements but still do not have properties which surpass those of the Newcombe hybrid score interval.

Dunnett and Gent[6] discuss methods to compute p-values instead of confidence intervals for a test of a predetermined margin of non-inferiority for the difference between proportions. These methods include modifications of the Wald and Wald continuity corrected intervals that use modified proportions in the computation of the variance estimate with these adjustments represented by $p`_T = \dfrac{y_T + y_C + n_C\Delta_0}{n_T + n_C}$ and $p`_C = \dfrac{y_T + y_C - n_T\Delta_0}{n_T + n_C}$ which are constrained by the null hypothesis (Δ_0) and fixed marginal totals. The test statistic (termed the Wald Adjusted method) is written in the form of a confidence interval as seen in (2.9).

$$\{\hat{p}_T - \hat{p}_C\} \pm z_\alpha \sqrt{\frac{p`_T(1-p`_T)}{n_T} + \frac{p`_C(1-p`_C)}{n_C}} \tag{2.9}$$

The Wald adjusted continuity corrected interval simply adds $\dfrac{1}{2}\left[\dfrac{1}{\mathrm{n_T}} + \dfrac{1}{\mathrm{n_C}}\right]$ to the right of (2.9).

Dunnett and Gent[6] describe test statistics based on a Chi-square distribution which can be used to produce a test of the null hypothesis of inferiority. These statistics have been modified from those presented so that they follow a standard normal distribution. The Chi-squared statistic is seen in (2.10) with the continuity corrected form seen in (2.11) which uses the adjusted proportions as in the Wald adjusted intervals.

$$\{\mathrm{y_T - n_T p`_T}\}\left\{\frac{1}{\mathrm{n_T p`_T}} + \frac{1}{\mathrm{y_T + y_C - n_T p`_T}} + \frac{1}{\mathrm{n_T - n_T p`_T}} + \frac{1}{\mathrm{n_C - y_T - y_C + n_T p`_T}}\right\}^{1/2} \tag{2.10}$$

$$\left\{\left|y_T - n_T p\text{`}_T\right| - \frac{1}{2}\right\}\left\{\frac{1}{n_T p\text{`}_T} + \frac{1}{y_T + y_C - n_T p\text{`}_T} + \frac{1}{n_T - n_T p\text{`}_T} + \frac{1}{n_C - y_T - y_C + n_T p\text{`}_T}\right\}^{1/2} \quad (2.11)$$

Dunnett and Gent suggest that the continuity corrected Chi-square test statistic is the most preferred method out of those presented in their paper.

Farrington and Manning[7] also propose a series of methods for the difference in proportions. Their methods all follow the general from see in (2.12) but with varying estimates for $\tilde{\pi}_T$ and $\tilde{\pi}_C$.

$$\left\{\hat{p}_T - \hat{p}_C\right\} \pm z_\alpha \sqrt{\frac{\tilde{\pi}_T(1-\tilde{\pi}_T)}{n_T} + \frac{\tilde{\pi}_C(1-\tilde{\pi}_C)}{n_C}} \quad (2.12)$$

The first of these suggested by Farrington and Manning uses the observed proportions $\hat{p}_T$ and $\hat{p}_C$ for $\tilde{\pi}_T$ and $\tilde{\pi}_C$ which results in an interval that is identical to the Wald interval. The second of these methods uses estimates of $p\text{`}_T$ and $p\text{`}_C$ which yields an identical interval to the Wald adjusted interval. Finally, the third method proposed by Farrington and Manning (F-M 3) uses estimates of $\tilde{\pi}_T$ and $\tilde{\pi}_C$ which are maximum likelihood estimates under the null hypothesis of inferiority at Δ_0, and Farrington and Manning discuss their computation as closed form solutions seen in (2.13).

$$\tilde{\pi}_T = 2u\cos(w) - \frac{b}{3a} \text{ and } \tilde{\pi}_C = \tilde{\pi}_T - \Delta_0 \quad (2.13)$$

$$a = 1 + \frac{n_C}{n_T}, b = -\left\{1 + \frac{n_C}{n_T} + \hat{p}_T + \frac{n_C}{n_T}\hat{p}_C + \Delta_0\left(\frac{n_C}{n_T} + 2\right)\right\},$$

where $c = \Delta_0^{\ 2} + \Delta_0\left(2\hat{p}_T + \frac{n_C}{n_T} + 1\right) + \hat{p}_T + \frac{n_C}{n_T}\hat{p}_C, d = -\hat{p}_T\Delta_0(1+\Delta_0),$

$$u = sign(v)\left\{\frac{b^2}{(3a)^2} - \frac{c}{3a}\right\}^{1/2}, v = \frac{b^3}{(3a)^3} - \frac{bc}{6a^2} + \frac{d}{2a}, w = \frac{1}{3}\left\{\Pi + \cos^{-1}\left(\frac{v}{u^3}\right)\right\}$$

Software, such as SAS[8], can be used to compute the maximum likelihood estimates $\tilde{\pi}_T$ and $\tilde{\pi}_C$, and these values can then be placed in (2.12) to produce F-M 3 confidence limits. In this regard for implementation in SAS through PROC GENMOD, a procedure used to fit general linear models, there would be specification of a binomial distribution with an identity link. The model statement fits only the intercept and includes an offset term where the offset for the control group is zero and for the test group is set equal to the specified non-inferiority margin Δ_0. Farrington and Manning recommend this last method because of closer to nominal coverage probabilities than the Wald or Wald Adjusted intervals.

Falk and Koch[9] suggest an additional method which attempts to improve on the Wald interval using a more appropriate unbiased estimator of the variance as seen in (2.14).

$$\{\hat{p}_T - \hat{p}_C\} \pm \left\{ z_\alpha \sqrt{\hat{\text{var}}_{\Delta_0}} + \frac{1}{2}\left[\frac{1}{n_T} + \frac{1}{n_C}\right]\right\} \qquad (2.14)$$

with $\hat{\text{var}}_{\Delta_0} = C_t \dfrac{\hat{\pi}_T(1-\hat{\pi}_T)}{n_T} + C_C \dfrac{\hat{\pi}_C(1-\hat{\pi}_C)}{n_C}$

$$\hat{\pi}_T = \overline{P} + \rho\Delta_0, \hat{\pi}_C = \overline{P} - (1-\rho)\Delta_0, \rho = \frac{n_C}{n_T + n_C}, \overline{P} = (1-\rho)\hat{p}_T + \rho\hat{p}_C,$$

where
$$C_C = 1 + \frac{\rho^2\left(\dfrac{1}{n_T} + \dfrac{1}{n_C}\right)}{\left\{1 - \dfrac{(1-\rho)^2}{n_T} - \dfrac{\rho^2}{n_C}\right\}}, C_T = 1 + \frac{(1-\rho)^2\left(\dfrac{1}{n_T} + \dfrac{1}{n_C}\right)}{\left\{1 - \dfrac{(1-\rho)^2}{n_T} - \dfrac{\rho^2}{n_C}\right\}}$$

Brown and Li[10] include in their list of methods the Yule's method as seen in (2.15) which is similar to the Wald interval but uses an average estimate of the proportion $\overline{p} = (y_T + y_C)/(n_T + n_C)$ in the variance, which Brown and Li suggest estimates the variance better when $\pi_T - \pi_C = 0$.

$$\{\hat{p}_T - \hat{p}_C\} \pm z_\alpha \sqrt{\left(\frac{1}{n_T} + \frac{1}{n_C}\right)\bar{p}(1-\bar{p})} \tag{2.15}$$

Brown and Li[10] present a Modified Yule's interval as seen in (2.16) which modifies the estimate of the proportion $\breve{p} = (n_C\hat{p}_T + n_T\hat{p}_C)/(n_T + n_C)$ used in the variance that should perform better when the sample sizes are not equal.

$$\{\hat{p}_T - \hat{p}_C\} \pm z_\alpha \sqrt{\left(\frac{1}{n_T} + \frac{1}{n_C}\right)\breve{p}(1-\breve{p})} \tag{2.16}$$

The Yule's method and Modified Yule's method are equivalent when $n_T = n_C$.

Another modification of the Wald method is based on Bayesian methodology, using a modification from a single proportion based on a prior distribution, and attributed as the Jeffrey's interval which adds one to each group, with half being attributed to an event where $p_T^* = (y_T + 0.5)/(n_T + 1)$ and $p_C^* = (y_C + 0.5)/(n_C + 1)$ with the interval calculated as in (2.17).

$$\{p_T^* - p_C^*\} \pm z_\alpha \sqrt{\frac{p_T^*(1-p_T^*)}{n_T} + \frac{p_C^*(1-p_C^*)}{n_C}} \tag{2.17}$$

A refinement of this interval is the Approximate Jeffrey's interval which adjusts the denominator of the variance estimate as seen in (2.18).

$$\{p_T^* - p_C^*\} \pm z_\alpha \sqrt{\frac{p_T^*(1-p_T^*)}{n_T + 2} + \frac{p_C^*(1-p_C^*)}{n_C + 2}} \tag{2.18}$$

Brown and Li[10] develop the new Recentered interval which they suggest performs well in relation to coverage probabilities of the interval. This interval uses an estimate $\breve{p} = (n_C\hat{p}_T + n_T\hat{p}_C)/(n_T + n_C)$ as seen in the Modified Yules interval for the variance but forces a truncated estimate $\tilde{\mathrm{p}}$ to meet the conditions in (2.19) so that $\tilde{p}$ is never estimated out of the appropriate range.

$$\tilde{p} = \begin{cases} \Delta_0 n_C /(n_T + n_C) & \text{if } \breve{\mathrm{p}} < \Delta_0 n_C /(n_T + n_C) \\ \breve{p} & \text{if } \Delta_0 n_C /(n_T + n_C) \le \breve{p} \le 1 - \Delta_0 n_T /(n_T + n_C) \\ 1 - \Delta_0 n_T /(n_T + n_C) & \text{if } \breve{\mathrm{p}} > 1 \text{ - } \Delta_0 n_T /(n_T + n_C) \end{cases} \tag{2.19}$$

The Recentered interval is seen in (2.20) where κ is the 1-α quantile of the t-distribution with $(\mathrm{n_T} + \mathrm{n_C} - 2)$ degrees of freedom.

$$\frac{\hat{\mathrm{p}}_\mathrm{T} - \hat{\mathrm{p}}_\mathrm{C}}{1 + \kappa^2 /(\mathrm{n_T} + \mathrm{n_C})} \pm \frac{\kappa \sqrt{\left\{1 + \kappa^2 /(\mathrm{n_T} + \mathrm{n_C})\right\}\left(\frac{1}{\mathrm{n_T}} + \frac{1}{\mathrm{n_C}}\right)\tilde{\mathrm{p}}(1 - \tilde{\mathrm{p}}) - \frac{(\hat{\mathrm{p}}_\mathrm{T} - \hat{\mathrm{p}}_\mathrm{C})^2}{(\mathrm{n_T} + \mathrm{n_C})}}}{1 + \kappa^2 /(\mathrm{n_T} + \mathrm{n_C})} \tag{2.20}$$

Pan[11] also presents a new interval for the difference in proportions which he suggests is an improvement on the Wald and Agresti and Caffo intervals. Pan's interval seen in (2.21) is similar to the Agresti and Caffo interval but uses a critical value from the t-distribution with degrees of freedom as specified in (2.22) instead of using a critical value from a standard normal distribution to account for use of asymptotic methodology for finite samples.

$$\{\tilde{p}_T - \tilde{p}_C\} \pm t_{1-\alpha,df} \sqrt{\frac{\tilde{p}_T(1 - \tilde{p}_T)}{n_T + 2} + \frac{\tilde{p}_C(1 - \tilde{p}_C)}{n_C + 2}} \tag{2.21}$$

$$\mathrm{df} \approx \frac{2\left\{\frac{\tilde{\mathrm{p}}_\mathrm{T}(1 - \tilde{\mathrm{p}}_\mathrm{T})}{\mathrm{n_T} + 2} + \frac{\tilde{\mathrm{p}}_\mathrm{C}(1 - \tilde{\mathrm{p}}_\mathrm{C})}{\mathrm{n_C} + 2}\right\}^2}{\Omega(\tilde{\mathrm{p}}_\mathrm{T}, \mathrm{n_T} + 2) + \Omega(\tilde{\mathrm{p}}_\mathrm{C}, \mathrm{n_C} + 2)} \tag{2.22}$$

$$\text{where } \Omega(\mathrm{p}, \mathrm{n}) = \frac{(\mathrm{p} \text{ - } \mathrm{p}^2)}{\mathrm{n}^3} + \left[\mathrm{p} + (6\mathrm{n} - 7)\mathrm{p}^2 + 4(\mathrm{n} - 1)(\mathrm{n} - 3)\mathrm{p}^2 - 2(\mathrm{n} - 1)(2\mathrm{n} - 3)\mathrm{p}^3\right]/\,\mathrm{n}^5 \\ - 2\left[\mathrm{p} + (2\mathrm{n} - 3)\mathrm{p}^2 - 2(\mathrm{n} - 1)\mathrm{p}^3\right]/\,\mathrm{n}^4$$

Other methods produce tests for significance of the non-inferiority hypothesis based on estimators obtained from maximum likelihood methods for the proportion of events in the treatment and control groups based on the joint distribution of the events as the product of two independent binomial distributions for the treatment and control groups. The first method is based on the Deviance statistic in (2.23) where $\hat{\pi}_T$ and $\hat{\pi}_C$ are the maximum likelihood estimators under the alternative hypothesis H_A: $(\pi_T - \pi_C) > \Delta_0$, and $\left(\hat{\pi}^* - \Delta_0\right)$ and $\hat{\pi}^*$ are the corresponding maximum likelihood estimators under the null hypothesis H_0: $(\pi_T - \pi_C) \leq \Delta_0$.

$$2\left\{\log L\left(\hat{\pi}_T, \hat{\pi}_C\right) - \log L\left(\hat{\pi}^* - \Delta_0, \hat{\pi}^*\right)\right\} \tag{2.23}$$

The second of these methods is based on a Pearson statistic in the form of [(observed – expected)^2 / expected] as seen in (2.24) using $\left(\hat{\pi}^* - \Delta_0\right)$ and $\hat{\pi}^*$, the maximum likelihood estimators of π_T and π_C under the null hypothesis.

$$\begin{aligned}&\frac{\left\{y_T - n_T\left(\hat{\pi}^* - \Delta_0\right)\right\}^2}{n_T\left(\hat{\pi}^* - \Delta_0\right)} + \frac{\left\{\left(n_T - y_T\right) - n_T\left(1 - \hat{\pi}^* + \Delta_0\right)\right\}^2}{n_T\left(1 - \hat{\pi}^* + \Delta_0\right)} + \\ &\quad \frac{\left\{y_C - n_C\hat{\pi}^*\right\}^2}{n_C\hat{\pi}^*} + \frac{\left\{\left(n_C - y_C\right) - n_C(1 - \hat{\pi}^*)\right\}^2}{n_C(1 - \hat{\pi}^*)}\end{aligned} \tag{2.24}$$

Additionally, a method produced using proportions restricted under the null hypothesis using weighted least squares for estimation instead of maximum likelihood estimation ($\hat{p}_{T,WLS}$ and $\hat{p}_{C,WLS}$) yields a test statistic as seen in (2.25) where

$$v_T = \frac{\hat{p}_{T,WLS}(1 - \hat{p}_{T,WLS})}{(n_T - 1)} \text{ and } v_C = \frac{\hat{p}_{C,WLS}(1 - \hat{p}_{C,WLS})}{(n_C - 1)}.$$

$$Q_{WLS} = \frac{\left(\hat{p}_{T,WLS} - \hat{p}_{C,WLS} - \Delta_0\right)^2}{v_T + v_C} \tag{2.25}$$

These methods produce test statistics for which p-values can be obtained by using the chi-square distribution under one degree of freedom.

In addition to the methods described above, there are other methods which are iterative in nature and require intensive computational resources. These include Gart's method[6], the Score test[10], the Real Jeffrey's interval[10], the Approximate unconditional exact test[12], and the methods by Mee[2], Miettinen & Nurminen[2], the Profile likelihood method[2], the Profile likelihood method based on exact tail areas[2], and the Profile likelihood method based on 'mid-p' tail areas[2].

The lower limit of the confidence interval for the difference in proportions exceeding the non-inferiority margin (Δ_0) is used as the counterpart to a test statistic for H_0: $(\pi_T - \pi_C) \leq \Delta_0$ versus H_A: $(\pi_T - \pi_C) > \Delta_0$ as the alternative hypothesis for non-inferiority. Also, $\Delta_0 = 0$ corresponds to a one-sided assessment of superiority. Due to the one-sided nature of the non-inferiority hypothesis, the alpha level of interest is one-sided. Discussion and results will focus on the lower confidence limit through its provision of a one-sided test. Results may be entirely different when assessment includes the upper confidence limit and similarly the two-sided test, but they are outside the scope of this discussion.

For some hypothetical illustrations, Table 2.1 displays selected one-sided lower confidence limits from the confidence limit methods. Scenarios are provided for the one-sided 0.975 confidence level. Table 2.2 summarizes methods which consistently yield the largest lower confidence limits including Falk & Koch, Beal's-Haldane, Transformation, and Wald Adjusted methods. These methods seem to be less conservative than the other methods.

Table 2.1 summarizes methods which result in the smallest lower confidence limits including Wald Adjusted Continuity Corrected, Wald Continuity Corrected, Newcombe Hybrid Score Continuity Corrected, Edgeworth Expansion, Recentered, and Pan methods. These methods seem to be the most conservative methods.

Table 2.1 also summarizes methods that have small lower confidence limits for sample size allocations of 1:2 or 1:1 and have increasingly larger lower confidence limits for allocations which place more subjects in the test group (3:2, 2:1, 3:1). These methods include Beal's Jeffreys-Perks, Yules, Modified Yules, Jeffrey's, and Approximate Jeffrey's.

The methods summarized in Table 2.1 produce moderate lower confidence limits including Wald, Agresti & Caffo, Newcombe Hybrid Score, and F-M 3 methods. Table 2.2 summarizes the one-sided p-values for those methods producing test statistics (with confidence limits available through an iterative process). These methods include Chi-Square, Chi-Square Continuity Corrected, Weighted Least Squares, Deviance, and Pearson.

III. Simulations for Non-inferiority

Simulations were used to study the properties of these methods in a non-inferiority setting. Scenarios included varying the following parameters:

1. π_C, the population proportion of events in the control group: 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95
2. $\Delta = \pi_T - \pi_C$, the population risk difference: $-\Delta_0$, $-\Delta_0/2$, -0.01, 0. 0.01, 0.025, 0.05
3. π_T, the population proportion of events in the test group: $\pi_T = \pi_C + \Delta$
4. Δ_0, the null hypothesis risk difference: -0.10, -0.075, -0.05
5. α, the one-sided alpha level: 0.005, 0.025, 0.05
6. n_C, the sample size in the control group is calculated to have 85% power to contradict the null hypothesis Δ_0 for equality of test and control groups, $n_T = Rn_C$

$$n_C = \frac{\left(z_\alpha + z_\beta\right)^2 \left\{(R+1)\pi_C(1-\pi_C)\right\}}{R\Delta_0^2}$$

7. Sample size allocation for test:control as 1:2, 1:1, 3:2, 2:1, 3:1

For each of the 100,000 replications performed, a sample from the specified binomial distribution was drawn separately for the test group and for the control group. All of the confidence limits and p-values were computed for the same replication and a conclusion of non-inferiority or not was determined according to whether or not the one-sided lower confidence limit exceeded the specified non-inferiority margin or similarly if the p-value exceeded the alpha level. The average of zero or one indicator variables for demonstration of non-inferiority or not resulted in a simulated power for the methods when the specified difference in proportions is better than the non-inferiority margin and a type I error rate when the specified difference in proportions is equal to (or poorer than) the non-inferiority margin. If the number of failures in either group was equal to zero or the method failed to produce a logical confidence limit, then the Agresti and Caffo method was used in place of the methods. In practice, an alternative (exact) method might be used for small event rates greater than zero, but the performance of the methods with this minimal modification is adequate for this discussion without additional modifications.

A summary of the type I error can be found in Figures 2.1 – 2.4 for α=0.025. Similar results are seen for other alpha levels. Patterns observed for the selected scenarios from Tables 2.1-2.4 are similar for the simulated type I error rates. Methods including the Chi-Square, Chi-Square Continuity Corrected, Falk & Koch, Beal's-Haldane, Transformation, and Wald Adjusted consistently yield type I errors above the nominal level as seen in Figure 2.1. The opposite is observed in Figure 2.2 for the Wald Adjusted Continuity Corrected, Wald Continuity Corrected, Newcombe Hybrid Score Continuity Corrected, Edgeworth Expansion, Recentered, and Pan methods with type I error levels consistently below the nominal level. Figure 2.3 summarizes methods which produce higher than nominal type I errors for sample size allocations of 1:2 and 1:1, but with approximately nominal type I error levels for the 3:2, 2:1, and 3:1 allocations. These methods include Beal's Jeffreys-Perks, Yules, Modified Yules, Jeffrey's, Approximate Jeffrey's, and Weighted Least Squares.

The methods summarized in Figure 2.4 will be the focus of further discussion. These methods generally yield approximately nominal type I errors, at least for certain sample size allocations. The Wald method has higher than nominal type I error rates for the 1:2 and 1:1 allocations, but nominal values for 3:2, 2:1, and 3:1 allocations. The Agresti & Caffo method generally produces nominal type I error levels, with values becoming closer to nominal as more sample size is placed in the test group (3:2, 2:1, 3:1). The F-M 3method consistently produces nominal type I error rates, but increasing as more sample size is placed in the test group. Similar patterns are also seen with the Newcombe Hybrid Score and Pearson methods. The Deviance method generally has nominal rates but has higher type I error rates for the 1:2 and 1:1 allocations, with lower rates for allocations with more sample size in the test group.

These methods are also summarized by the non-inferiority margin in Figure 2.5. The performance of these methods seems to be unaffected by the choice of margin. Figure 2.6 summarizes the type I error by values of π_C, with type I error values becoming more variable with larger values of π_C. The most plausible reason for this is that the expected cell frequencies become smaller as π_C increases. Therefore, the asymptotic assumption may not be appropriate for larger π_C depending on the sample sizes in each treatment group.

Figures 2.7 – 2.10 summarize the simulated power for varying scenarios of $\Delta=\pi_T - \pi_C$, the population risk difference. As would be expected, if a trial was planned for equality of the treatment groups, and this was misspecified as in Figure 2.7 with $\Delta=\Delta_0/2$, the power is drastically reduced and falls below 0.40 for most cases. However, if the equality assumption is valid as in Figure 2.8 with $\Delta=0$, then the simulated power hovers around the planned power of 0.85 with the 1:2 allocation yielding lower simulated power and the allocations with more sample size in the test group yielding power higher than 0.85. If the test group is better than the control group as seen in Figure 2.9 for $\Delta=0.01$ and Figure 2.10 for $\Delta=0.025$, the simulated power is higher than 0.85.

The most appropriate discussion of power is focused on situations where the methods perform close to the nominal level for type I error. The Newcombe Hybrid Score, F-M 3, and Pearson methods tend to yield close to nominal type I error levels for the 1:2 and 1:1 sample size allocation settings. Figures 2.11 – 2.13 compare each of these methods with respect to simulated power for the setting where $\Delta=0$ and $\alpha=0.025$. These methods have similar simulated powers, with the Newcombe Hybrid Score method yielding slightly higher power than the Pearson and F-M 3 methods.

The Wald, Agresti & Caffo, and Deviance methods tend to perform at the nominal type I error level for sample size allocations of 3:2, 2:1, and 3:1. Figures 2.14 – 2.16 compare the simulated power of these methods. Both the Deviance and Agresti & Caffo methods tend to have higher simulated power than the Wald method in these settings. The Deviance method yields slightly higher power for the 3:1 scenario, with the Agresti & Caffo method being slightly higher in the 3:2 and 2:1 settings.

IV. Sample Size Considerations

In addition to appropriate methods for analyses when the difference between proportions is the measure of interest for treatment comparisons in a one-sided non-inferiority setting, it is important to have corresponding sample size formulas in the planning stages of the trial. Sample size calculation based on the Wald method is a popular and straightforward way to plan for patient recruitment in non-inferiority trials for the difference in proportions. This sample size calculation (with respect to $\Delta_0<0$) is shown in (2.26) for the test group, with the sample size in the control group defined as $n_C=n_T/R$ where $R=n_T/n_C$.

$$n_T = \frac{\left\{z_{1-\alpha} + z_{1-\beta}\right\}^2 \left\{\pi_T\left(1-\pi_T\right) + R\pi_C\left(1-\pi_C\right)\right\}}{\left\{\pi_T - \pi_C - \Delta_0\right\}^2} \tag{2.26}$$

This sample size formula, through algebraic manipulation, can be written to produce power for specified sample sizes in the test and control groups as in (2.27)

$$z_{1-\beta} = \frac{\sqrt{n_T}\left\{\pi_T - \pi_C - \Delta_0\right\}}{\sqrt{\pi_T\left(1-\pi_T\right) + R\pi_C\left(1-\pi_C\right)}} - z_{1-\alpha} \tag{2.27}$$

where power is obtained as the probability (1-β) from $z_{1-\beta}$ as the (1-β) quantile of the standard normal distribution.

Additionally, Farrington and Manning[7] provide the sample size formula in (2.28) that is analogous to their methods, with the appropriate proportions substituted for $\tilde{\pi}_T$ and $\tilde{\pi}_C$ where for F-M 3 the values from solving the maximum likelihood equations under the null hypothesis are used. As a note, (2.28) reduces to (2.26) if $\tilde{\pi}_T = \pi_T$ and $\tilde{\pi}_C = \pi_C$.

$$n_T = \frac{\left\{z_{1-\alpha}\sqrt{\tilde{\pi}_T(1-\tilde{\pi}_T) + R\tilde{\pi}_C(1-\tilde{\pi}_C)} + z_{1-\beta}\sqrt{\pi_T\left(1-\pi_T\right) + R\pi_C\left(1-\pi_C\right)}\right\}^2}{\left\{\pi_T - \pi_C - \Delta_0\right\}^2} \tag{2.28}$$

The corresponding power calculation is seen in (2.29).

$$z_{1-\beta} = \frac{\sqrt{n_T}\,(\pi_T - \pi_C - \Delta_0) - z_{1-\alpha}\sqrt{\tilde{\pi}_T(1-\tilde{\pi}_T) + R\tilde{\pi}_C(1-\tilde{\pi}_C)}}{\sqrt{\pi_T(1-\pi_T) + R\pi_C(1-\pi_C)}} \tag{2.29}$$

The calculated power from the F-M 3 sample size formula is compared to the simulated power from the F-M 3 method in Figure 2.17 for the 1:2 and 1:1 sample size allocation settings. The simulated power is consistently equal to or greater than the calculated power. Similar patterns are seen when the Newcombe Hybrid Score and Pearson simulated powers are compared to the F-M 3 calculated power as seen in Figures 2.18 and 2.19 for the 1:2 and 1:1 allocations. Table 2.3 includes a summary of selected scenarios for the 1:2 and 1:1 sample size allocations with the simulated power for the F-M 3, Newcombe Hybrid Score, and Pearson methods along with the calculated power based on the F-M 3 method.

The sample size formula based on the Wald method is useful in the 3:2, 2:1, and 3:1 allocation settings. The Wald, Deviance, and Agresti & Caffo simulated powers seen in Figures 2.20, 2.21, and 2.22 are higher than the calculated power. However, in some cases, especially as more sample size is placed in the test group (3:1 allocation), the simulated power may be higher than the calculated power by over 0.05. Having a conservative sample size formula is beneficial in trial design, however with limited resources too much conservatism may be costly. Therefore, in these situations the sample size calculations could be reduced slightly to match more closely with the simulated final power so as to conserve resources if necessary. Table 2.4 summarizes selected scenarios for the 3:2, 2:1, and 3:1 sample size allocations with the simulated power for the Wald, Agresti & Caffo, and Pearson methods as well as the calculated power based on the Wald method.

V. Discussion

There are many methods available in the literature for a non-inferiority setting focused on the difference between two proportions. Research involving these methods shows that the performance related to maintaining the nominal type I error rate depends on the sample size allocation of interest. For the 1:2 and 1:1 settings for test:control, the F-M 3, Newcombe Hybrid Score, and Pearson methods perform appropriately. Use of the sample size calculation based on the F-M 3 method allows for appropriate planning of a non-inferiority trial with slightly conservative sample sizes calculations.

In the scenarios with allocations of 3:2, 2:1, and 3:1 with more sample size allocated to the test group, the Wald, Agresti & Caffo, and Deviance methods are appropriate. In addition, the Wald sample size calculation can be used in trial design with the caviat that as more sample size is placed in the test group this formula may become fairly conservative.

References

1. Agresti A., and Caffo B. Simple and Effective Confidence Intervals for Proportions and Difference of Proportions Results from Adding Two Successes and Two Failures. The American Statistician **2000**, 54(4): 280-288.

2. Newcombe G. Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. Statistics in Medicine **1998**, 17: 873-890.

3. Roebruck P, Kühn A. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine* 1995; **14**: 1583-1594.

4. Li Z, Chuang-Stein C. A note on comparing two binomial proportions in confirmatory noninferiority trials. *Drug Information Journal* 2006; **40**: 203-208.

5. Zhou X., Tsao M., and Qin G. New Intervals for the Difference Between Two Independent Binomial Proportions. Journal of Statistical Planning and Inference **2004**, 123: 97-115.

6. Dunnett C. W., and Gent M. Significance Testing to Establish Equivalence Between Treatments, with Special Reference to Data in the Form of 2 x 2 Tables. Biometrics **1977**, 33(4): 593-602.

7. Farrington, C. P., and Manning G. Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-zero Risk Difference or Non-unity Relative Risk. Statistics in Medicine **1990**, 9: 1447-1454.

8. SAS®, Version 8.02. SAS Institute Inc.: Cary, NC, 1999.

9. Falk R. W., and Koch G. G. Testing a Specified Difference Between Proportions. Biometrics **1998**, 54(4): 1602-1614.

10. Brown L., and Li X. Confidence Intervals for Two Sample Binomial Distribution. Journal of Statistical Planning and Inference **2005**, 130: 359-375.

11. Pan W. Approximate Confidence Intervals for One Proportion and Difference of Two Proportions. Computational Statistics & Data Analysis **2002**, 40: 143-157.

12. Kang S.-H., and Chen J. J. An Approximate Unconditional Test of Non-inferiority Between Two Proportions. Statistics in Medicine **2000**, 19: 2089-2100.

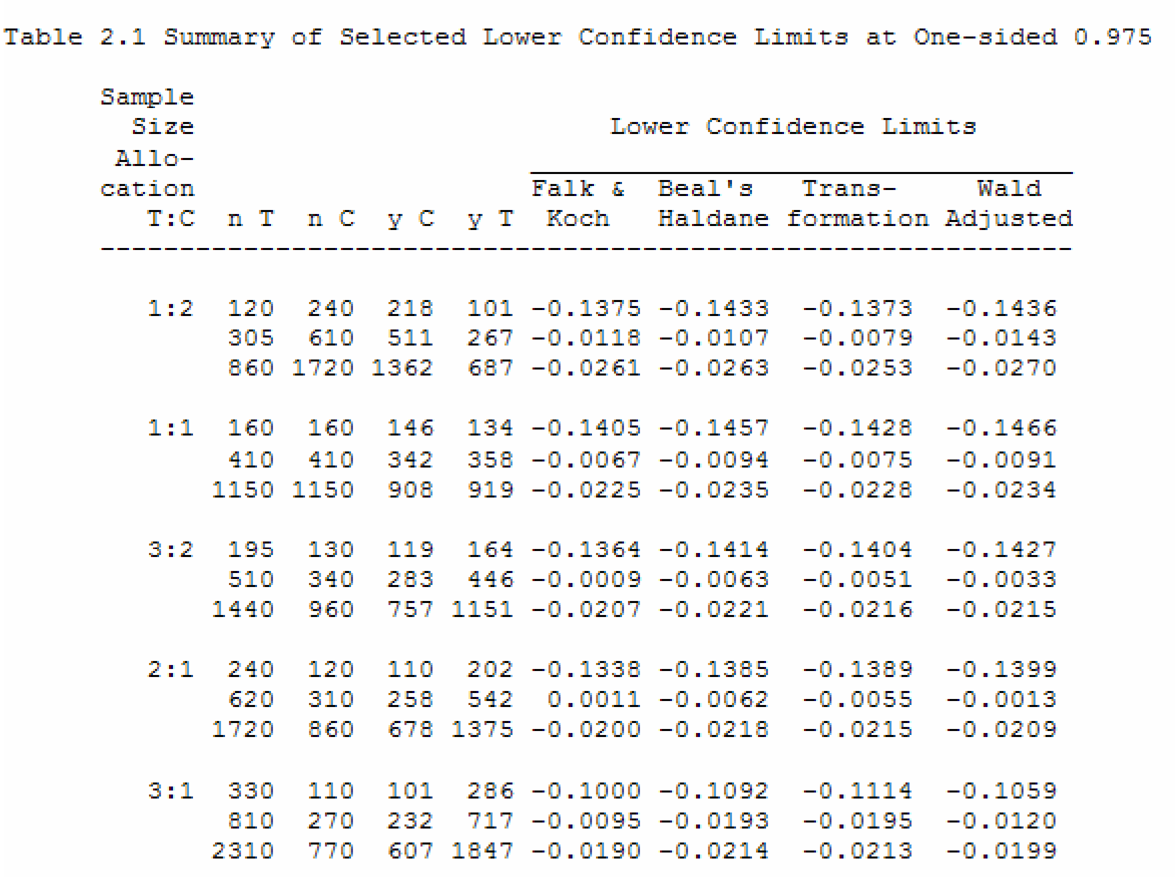


Table 2.1 Summary of Selected Lower Confidence Limits at One-sided 0.975

Sample Size Allocation					Lower Confidence Limits			
T:C	n T	n C	y C	y T	Falk & Koch	Beal's Haldane	Transformation	Wald Adjusted
1:2	120	240	218	101	-0.1375	-0.1433	-0.1373	-0.1436
	305	610	511	267	-0.0118	-0.0107	-0.0079	-0.0143
	860	1720	1362	687	-0.0261	-0.0263	-0.0253	-0.0270
1:1	160	160	146	134	-0.1405	-0.1457	-0.1428	-0.1466
	410	410	342	358	-0.0067	-0.0094	-0.0075	-0.0091
	1150	1150	908	919	-0.0225	-0.0235	-0.0228	-0.0234
3:2	195	130	119	164	-0.1364	-0.1414	-0.1404	-0.1427
	510	340	283	446	-0.0009	-0.0063	-0.0051	-0.0033
	1440	960	757	1151	-0.0207	-0.0221	-0.0216	-0.0215
2:1	240	120	110	202	-0.1338	-0.1385	-0.1389	-0.1399
	620	310	258	542	0.0011	-0.0062	-0.0055	-0.0013
	1720	860	678	1375	-0.0200	-0.0218	-0.0215	-0.0209
3:1	330	110	101	286	-0.1000	-0.1092	-0.1114	-0.1059
	810	270	232	717	-0.0095	-0.0193	-0.0195	-0.0120
	2310	770	607	1847	-0.0190	-0.0214	-0.0213	-0.0199

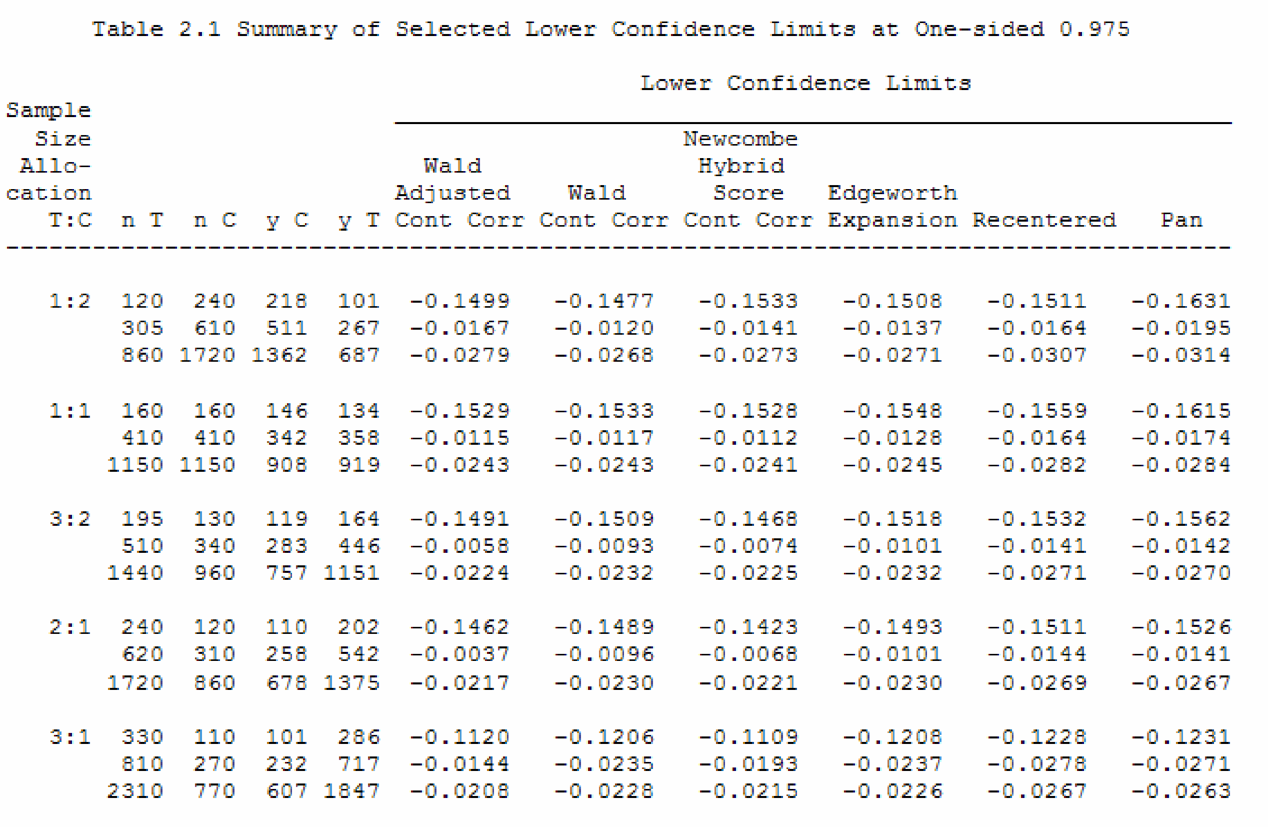


Table 2.1 Summary of Selected Lower Confidence Limits at One-sided 0.975

					Lower Confidence Limits					
Sample Size Allo-cation T:C	n T	n C	y C	y T	Wald Adjusted Cont Corr	Wald Cont Corr	Newcombe Hybrid Score Cont Corr	Edgeworth Expansion	Recentered	Pan
1:2	120	240	218	101	-0.1499	-0.1477	-0.1533	-0.1508	-0.1511	-0.1631
	305	610	511	267	-0.0167	-0.0120	-0.0141	-0.0137	-0.0164	-0.0195
	860	1720	1362	687	-0.0279	-0.0268	-0.0273	-0.0271	-0.0307	-0.0314
1:1	160	160	146	134	-0.1529	-0.1533	-0.1528	-0.1548	-0.1559	-0.1615
	410	410	342	358	-0.0115	-0.0117	-0.0112	-0.0128	-0.0164	-0.0174
	1150	1150	908	919	-0.0243	-0.0243	-0.0241	-0.0245	-0.0282	-0.0284
3:2	195	130	119	164	-0.1491	-0.1509	-0.1468	-0.1518	-0.1532	-0.1562
	510	340	283	446	-0.0058	-0.0093	-0.0074	-0.0101	-0.0141	-0.0142
	1440	960	757	1151	-0.0224	-0.0232	-0.0225	-0.0232	-0.0271	-0.0270
2:1	240	120	110	202	-0.1462	-0.1489	-0.1423	-0.1493	-0.1511	-0.1526
	620	310	258	542	-0.0037	-0.0096	-0.0068	-0.0101	-0.0144	-0.0141
	1720	860	678	1375	-0.0217	-0.0230	-0.0221	-0.0230	-0.0269	-0.0267
3:1	330	110	101	286	-0.1120	-0.1206	-0.1109	-0.1208	-0.1228	-0.1231
	810	270	232	717	-0.0144	-0.0235	-0.0193	-0.0237	-0.0278	-0.0271
	2310	770	607	1847	-0.0208	-0.0228	-0.0215	-0.0226	-0.0267	-0.0263

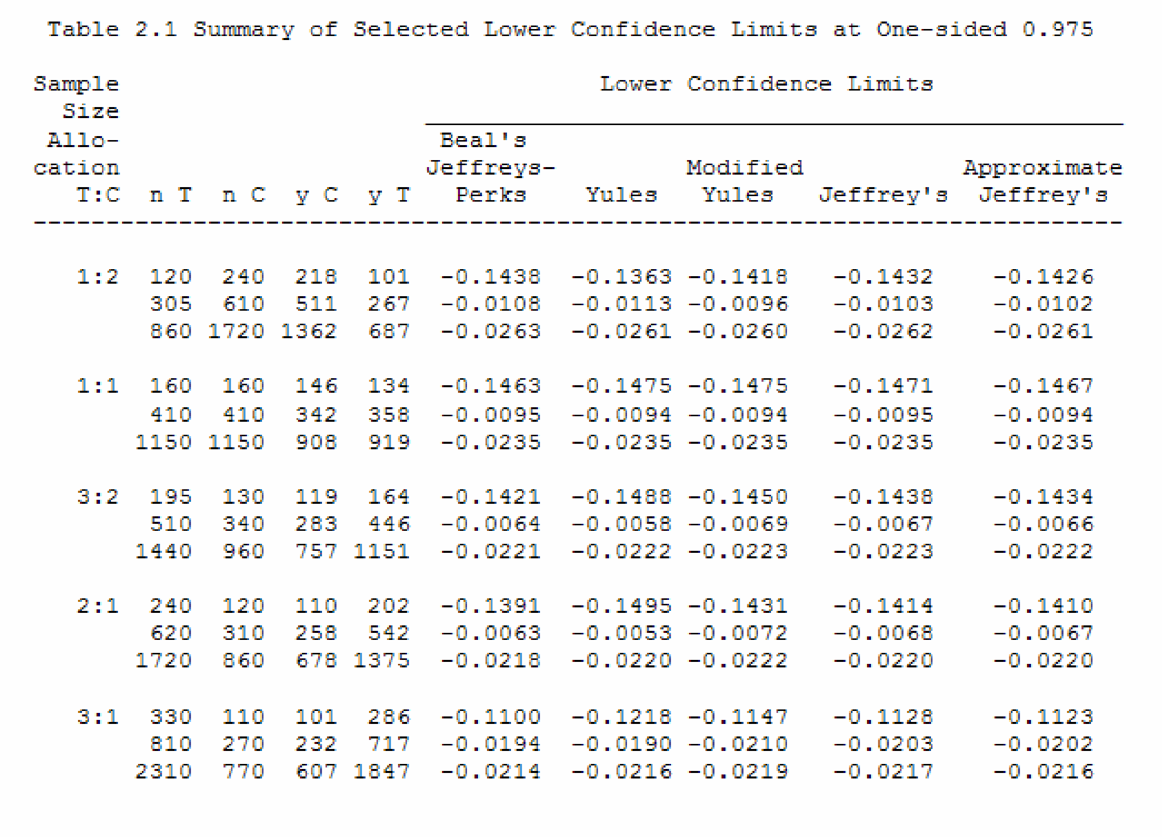


Table 2.1 Summary of Selected Lower Confidence Limits at One-sided 0.975

Sample Size Allocation T:C	n T	n C	y C	y T	Lower Confidence Limits Beal's Jeffreys-Perks	Yules	Modified Yules	Jeffrey's	Approximate Jeffrey's
1:2	120	240	218	101	-0.1438	-0.1363	-0.1418	-0.1432	-0.1426
	305	610	511	267	-0.0108	-0.0113	-0.0096	-0.0103	-0.0102
	860	1720	1362	687	-0.0263	-0.0261	-0.0260	-0.0262	-0.0261
1:1	160	160	146	134	-0.1463	-0.1475	-0.1475	-0.1471	-0.1467
	410	410	342	358	-0.0095	-0.0094	-0.0094	-0.0095	-0.0094
	1150	1150	908	919	-0.0235	-0.0235	-0.0235	-0.0235	-0.0235
3:2	195	130	119	164	-0.1421	-0.1488	-0.1450	-0.1438	-0.1434
	510	340	283	446	-0.0064	-0.0058	-0.0069	-0.0067	-0.0066
	1440	960	757	1151	-0.0221	-0.0222	-0.0223	-0.0223	-0.0222
2:1	240	120	110	202	-0.1391	-0.1495	-0.1431	-0.1414	-0.1410
	620	310	258	542	-0.0063	-0.0053	-0.0072	-0.0068	-0.0067
	1720	860	678	1375	-0.0218	-0.0220	-0.0222	-0.0220	-0.0220
3:1	330	110	101	286	-0.1100	-0.1218	-0.1147	-0.1128	-0.1123
	810	270	232	717	-0.0194	-0.0190	-0.0210	-0.0203	-0.0202
	2310	770	607	1847	-0.0214	-0.0216	-0.0219	-0.0217	-0.0216

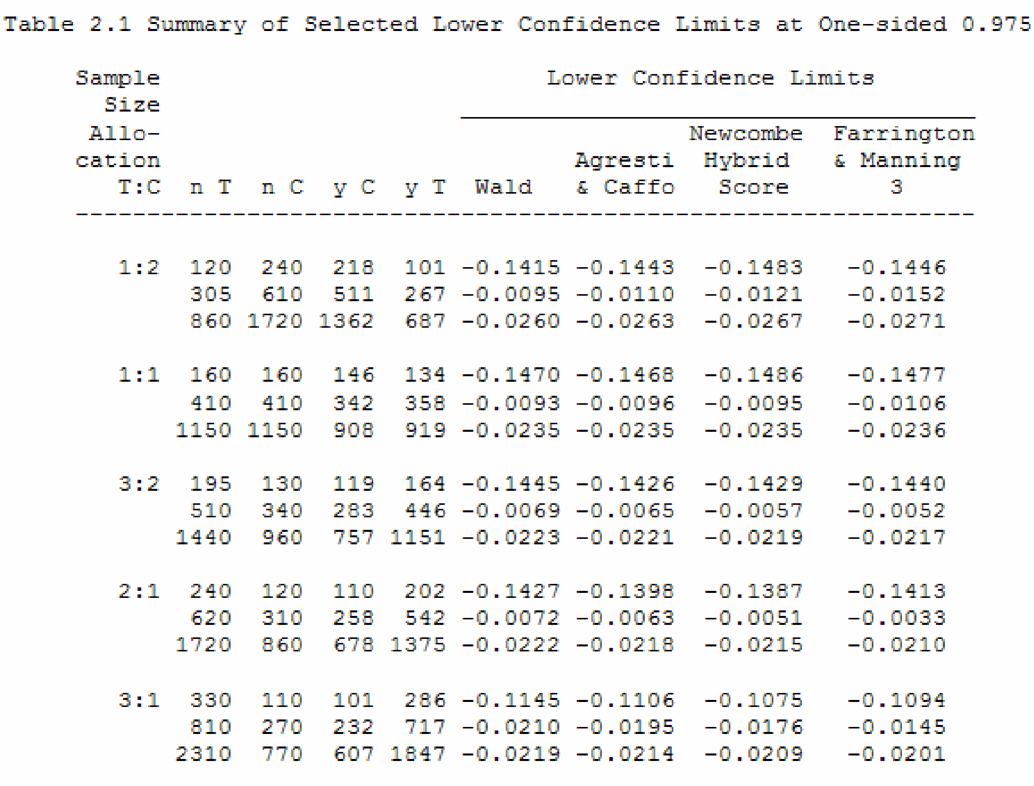
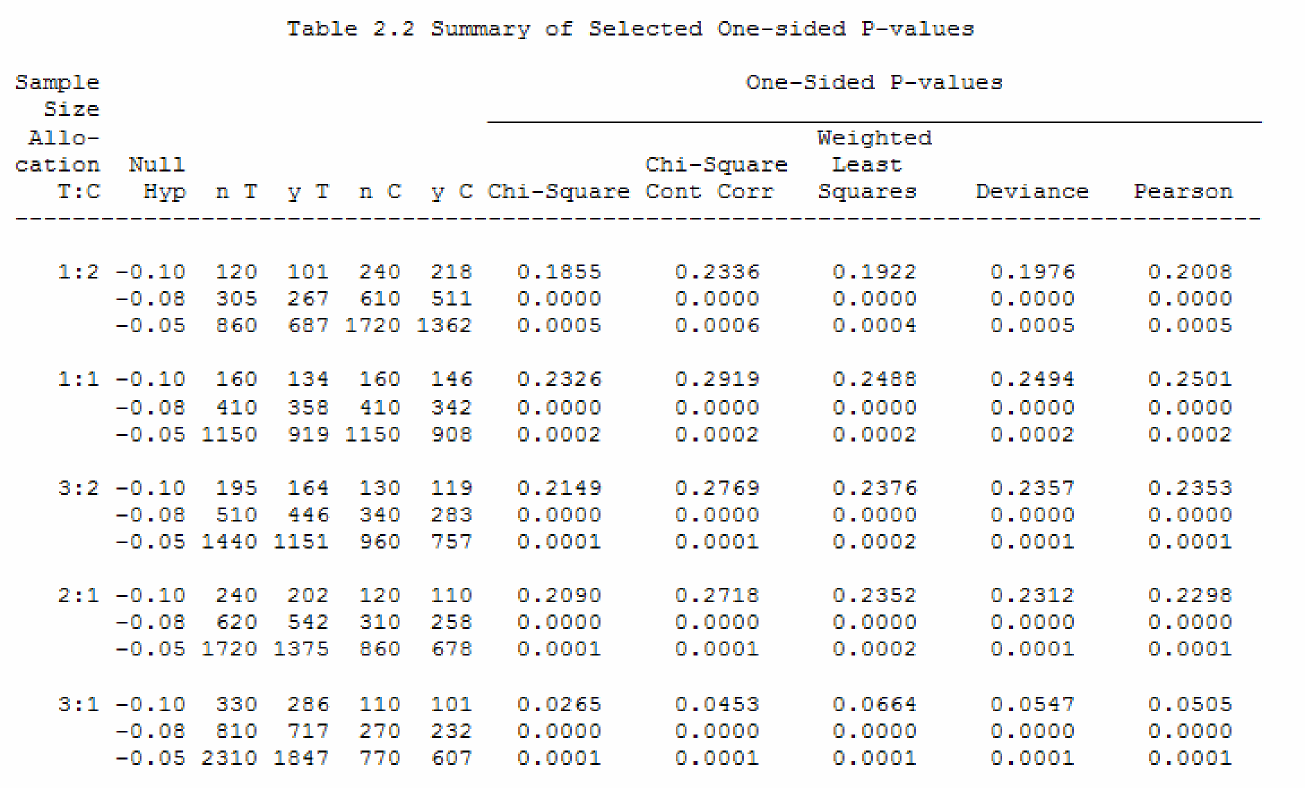


Table 2.1 Summary of Selected Lower Confidence Limits at One-sided 0.975

Sample Size Allo- cation T:C	n T	n C	y C	y T	Lower Confidence Limits			
					Wald	Agresti & Caffo	Newcombe Hybrid Score	Farrington & Manning 3
1:2	120	240	218	101	-0.1415	-0.1443	-0.1483	-0.1446
	305	610	511	267	-0.0095	-0.0110	-0.0121	-0.0152
	860	1720	1362	687	-0.0260	-0.0263	-0.0267	-0.0271
1:1	160	160	146	134	-0.1470	-0.1468	-0.1486	-0.1477
	410	410	342	358	-0.0093	-0.0096	-0.0095	-0.0106
	1150	1150	908	919	-0.0235	-0.0235	-0.0235	-0.0236
3:2	195	130	119	164	-0.1445	-0.1426	-0.1429	-0.1440
	510	340	283	446	-0.0069	-0.0065	-0.0057	-0.0052
	1440	960	757	1151	-0.0223	-0.0221	-0.0219	-0.0217
2:1	240	120	110	202	-0.1427	-0.1398	-0.1387	-0.1413
	620	310	258	542	-0.0072	-0.0063	-0.0051	-0.0033
	1720	860	678	1375	-0.0222	-0.0218	-0.0215	-0.0210
3:1	330	110	101	286	-0.1145	-0.1106	-0.1075	-0.1094
	810	270	232	717	-0.0210	-0.0195	-0.0176	-0.0145
	2310	770	607	1847	-0.0219	-0.0214	-0.0209	-0.0201



```
                  Table 2.2 Summary of Selected One-sided P-values

Sample                                                        One-Sided P-values
 Size                                   ______________________________________________________
Allo-                                                      Weighted
cation  Null                                    Chi-Square   Least
  T:C    Hyp   n T   y T   n C   y C Chi-Square Cont Corr   Squares     Deviance     Pearson
-----------------------------------------------------------------------------------------------

  1:2 -0.10   120   101   240   218    0.1855      0.2336      0.1922      0.1976      0.2008
      -0.08   305   267   610   511    0.0000      0.0000      0.0000      0.0000      0.0000
      -0.05   860   687  1720  1362    0.0005      0.0006      0.0004      0.0005      0.0005

  1:1 -0.10   160   134   160   146    0.2326      0.2919      0.2488      0.2494      0.2501
      -0.08   410   358   410   342    0.0000      0.0000      0.0000      0.0000      0.0000
      -0.05  1150   919  1150   908    0.0002      0.0002      0.0002      0.0002      0.0002

  3:2 -0.10   195   164   130   119    0.2149      0.2769      0.2376      0.2357      0.2353
      -0.08   510   446   340   283    0.0000      0.0000      0.0000      0.0000      0.0000
      -0.05  1440  1151   960   757    0.0001      0.0001      0.0002      0.0001      0.0001

  2:1 -0.10   240   202   120   110    0.2090      0.2718      0.2352      0.2312      0.2298
      -0.08   620   542   310   258    0.0000      0.0000      0.0000      0.0000      0.0000
      -0.05  1720  1375   860   678    0.0001      0.0001      0.0002      0.0001      0.0001

  3:1 -0.10   330   286   110   101    0.0265      0.0453      0.0664      0.0547      0.0505
      -0.08   810   717   270   232    0.0000      0.0000      0.0000      0.0000      0.0000
      -0.05  2310  1847   770   607    0.0001      0.0001      0.0001      0.0001      0.0001
```

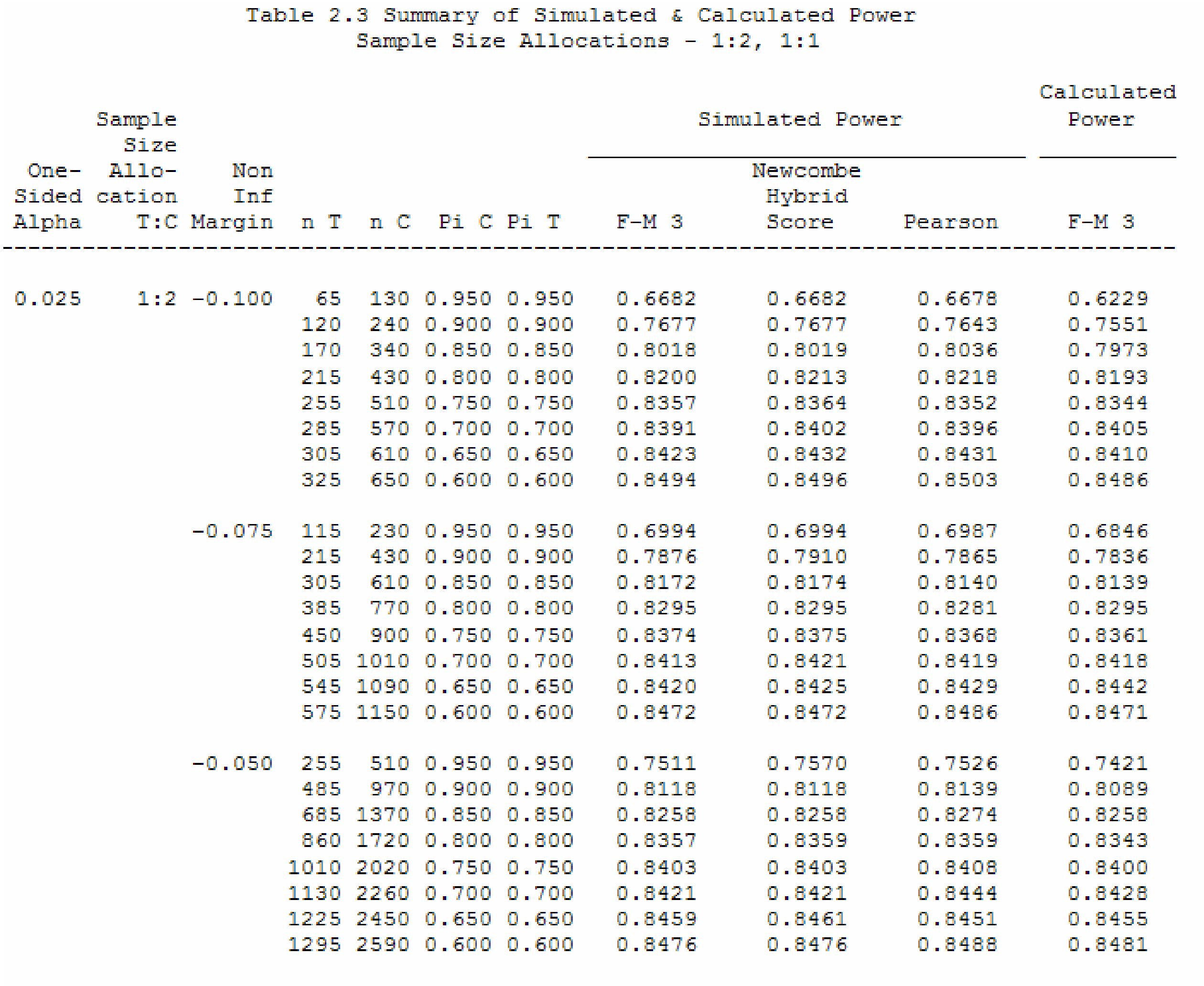


Table 2.3 Summary of Simulated & Calculated Power
Sample Size Allocations - 1:2, 1:1

							Simulated Power			Calculated Power
One-Sided Alpha	Sample Size Allocation T:C	Non Inf Margin	n T	n C	Pi C	Pi T	F-M 3	Newcombe Hybrid Score	Pearson	F-M 3
0.025	1:2	-0.100	65	130	0.950	0.950	0.6682	0.6682	0.6678	0.6229
			120	240	0.900	0.900	0.7677	0.7677	0.7643	0.7551
			170	340	0.850	0.850	0.8018	0.8019	0.8036	0.7973
			215	430	0.800	0.800	0.8200	0.8213	0.8218	0.8193
			255	510	0.750	0.750	0.8357	0.8364	0.8352	0.8344
			285	570	0.700	0.700	0.8391	0.8402	0.8396	0.8405
			305	610	0.650	0.650	0.8423	0.8432	0.8431	0.8410
			325	650	0.600	0.600	0.8494	0.8496	0.8503	0.8486
		-0.075	115	230	0.950	0.950	0.6994	0.6994	0.6987	0.6846
			215	430	0.900	0.900	0.7876	0.7910	0.7865	0.7836
			305	610	0.850	0.850	0.8172	0.8174	0.8140	0.8139
			385	770	0.800	0.800	0.8295	0.8295	0.8281	0.8295
			450	900	0.750	0.750	0.8374	0.8375	0.8368	0.8361
			505	1010	0.700	0.700	0.8413	0.8421	0.8419	0.8418
			545	1090	0.650	0.650	0.8420	0.8425	0.8429	0.8442
			575	1150	0.600	0.600	0.8472	0.8472	0.8486	0.8471
		-0.050	255	510	0.950	0.950	0.7511	0.7570	0.7526	0.7421
			485	970	0.900	0.900	0.8118	0.8118	0.8139	0.8089
			685	1370	0.850	0.850	0.8258	0.8258	0.8274	0.8258
			860	1720	0.800	0.800	0.8357	0.8359	0.8359	0.8343
			1010	2020	0.750	0.750	0.8403	0.8403	0.8408	0.8400
			1130	2260	0.700	0.700	0.8421	0.8421	0.8444	0.8428
			1225	2450	0.650	0.650	0.8459	0.8461	0.8451	0.8455
			1295	2590	0.600	0.600	0.8476	0.8476	0.8488	0.8481

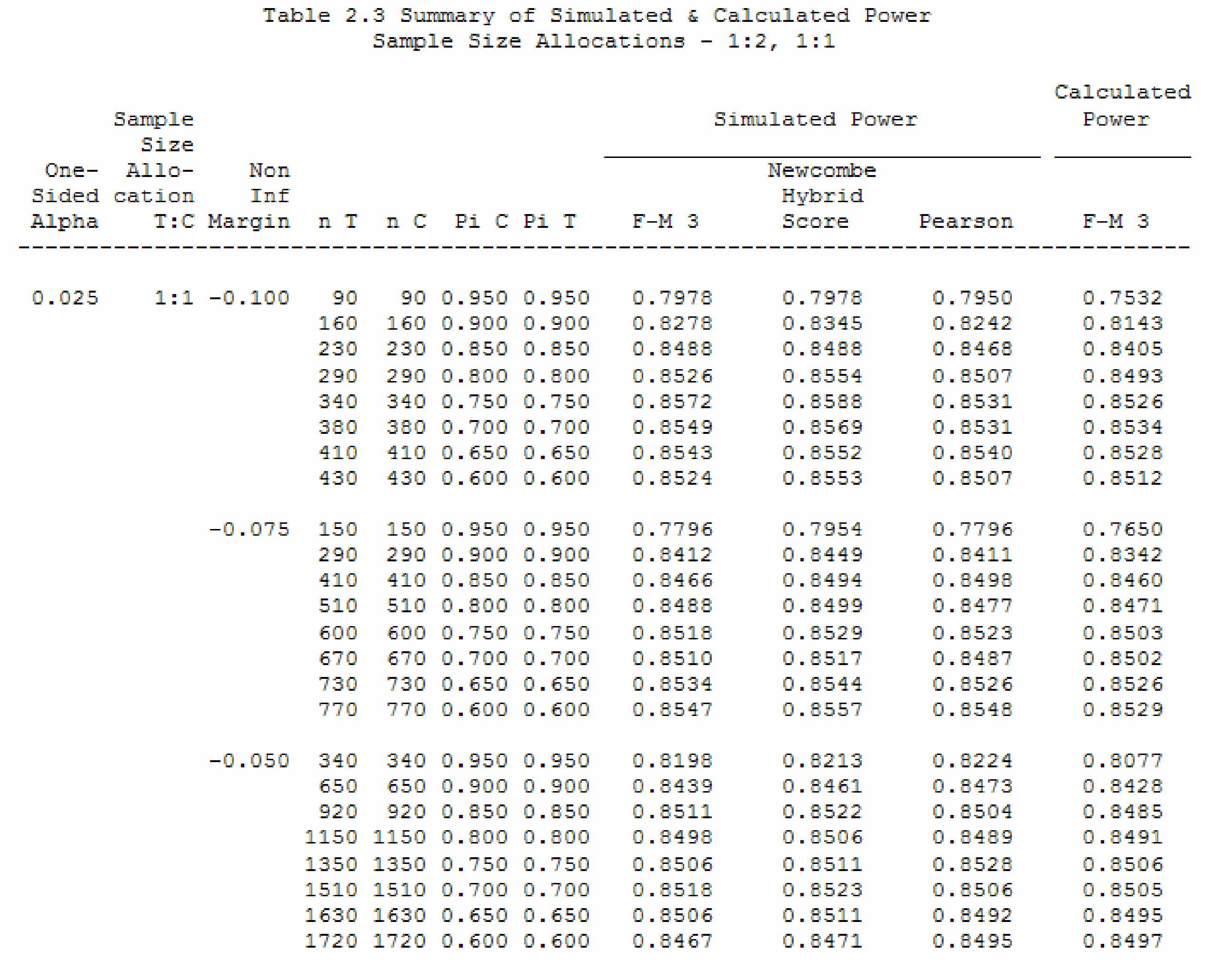


Table 2.3 Summary of Simulated & Calculated Power
Sample Size Allocations - 1:2, 1:1

							Simulated Power			Calculated Power
	Sample Size									
One-Sided Alpha	Allocation T:C	Non Inf Margin	n T	n C	Pi C	Pi T	F-M 3	Newcombe Hybrid Score	Pearson	F-M 3
0.025	1:1	-0.100	90	90	0.950	0.950	0.7978	0.7978	0.7950	0.7532
			160	160	0.900	0.900	0.8278	0.8345	0.8242	0.8143
			230	230	0.850	0.850	0.8488	0.8488	0.8468	0.8405
			290	290	0.800	0.800	0.8526	0.8554	0.8507	0.8493
			340	340	0.750	0.750	0.8572	0.8588	0.8531	0.8526
			380	380	0.700	0.700	0.8549	0.8569	0.8531	0.8534
			410	410	0.650	0.650	0.8543	0.8552	0.8540	0.8528
			430	430	0.600	0.600	0.8524	0.8553	0.8507	0.8512
		-0.075	150	150	0.950	0.950	0.7796	0.7954	0.7796	0.7650
			290	290	0.900	0.900	0.8412	0.8449	0.8411	0.8342
			410	410	0.850	0.850	0.8466	0.8494	0.8498	0.8460
			510	510	0.800	0.800	0.8488	0.8499	0.8477	0.8471
			600	600	0.750	0.750	0.8518	0.8529	0.8523	0.8503
			670	670	0.700	0.700	0.8510	0.8517	0.8487	0.8502
			730	730	0.650	0.650	0.8534	0.8544	0.8526	0.8526
			770	770	0.600	0.600	0.8547	0.8557	0.8548	0.8529
		-0.050	340	340	0.950	0.950	0.8198	0.8213	0.8224	0.8077
			650	650	0.900	0.900	0.8439	0.8461	0.8473	0.8428
			920	920	0.850	0.850	0.8511	0.8522	0.8504	0.8485
			1150	1150	0.800	0.800	0.8498	0.8506	0.8489	0.8491
			1350	1350	0.750	0.750	0.8506	0.8511	0.8528	0.8506
			1510	1510	0.700	0.700	0.8518	0.8523	0.8506	0.8505
			1630	1630	0.650	0.650	0.8506	0.8511	0.8492	0.8495
			1720	1720	0.600	0.600	0.8467	0.8471	0.8495	0.8497

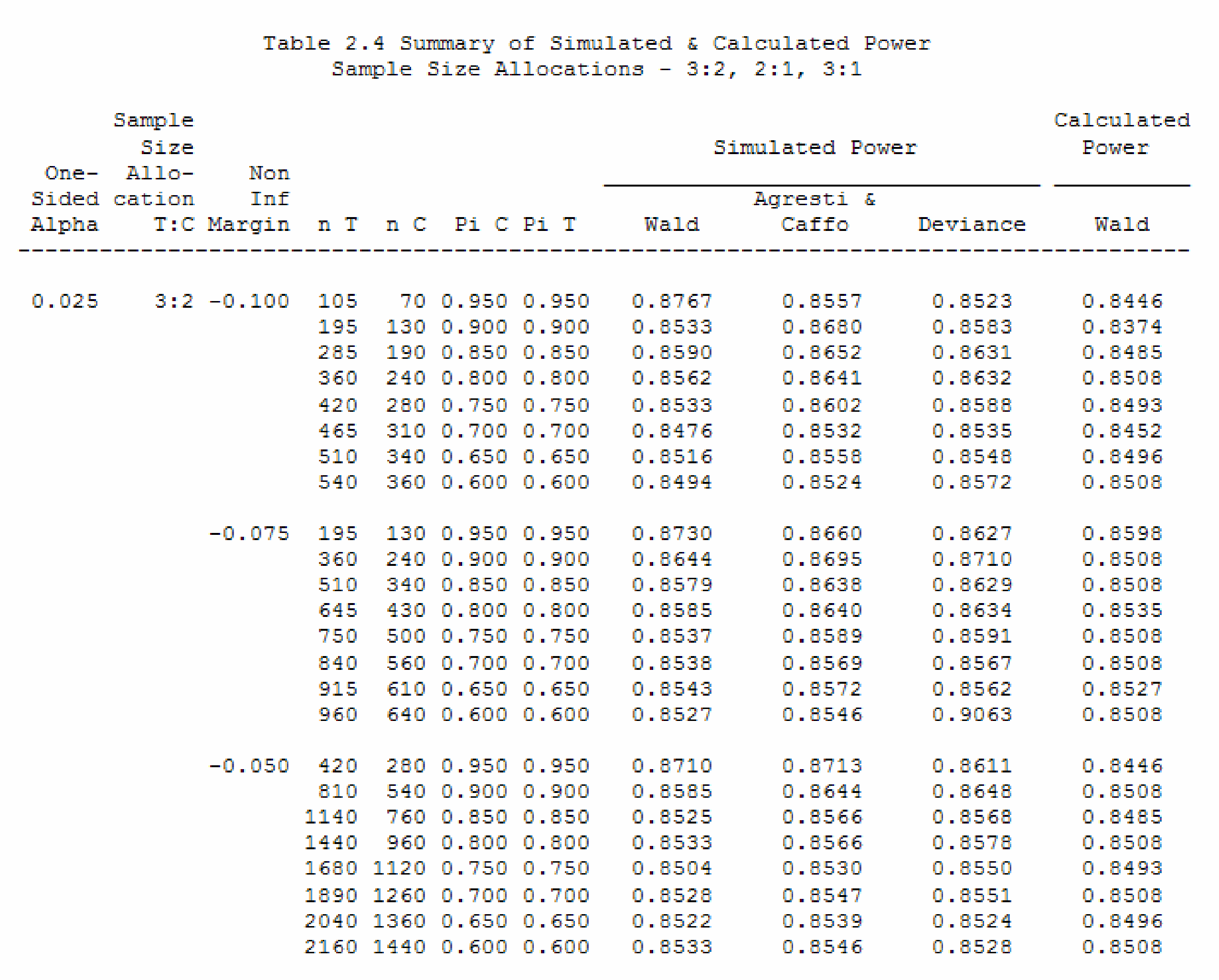


Table 2.4 Summary of Simulated & Calculated Power
Sample Size Allocations - 3:2, 2:1, 3:1

	Sample						Simulated Power			Calculated
	Size									Power
One-	Allo-	Non								
Sided	cation	Inf						Agresti &		
Alpha	T:C	Margin	n T	n C	Pi C	Pi T	Wald	Caffo	Deviance	Wald
0.025	3:2	-0.100	105	70	0.950	0.950	0.8767	0.8557	0.8523	0.8446
			195	130	0.900	0.900	0.8533	0.8680	0.8583	0.8374
			285	190	0.850	0.850	0.8590	0.8652	0.8631	0.8485
			360	240	0.800	0.800	0.8562	0.8641	0.8632	0.8508
			420	280	0.750	0.750	0.8533	0.8602	0.8588	0.8493
			465	310	0.700	0.700	0.8476	0.8532	0.8535	0.8452
			510	340	0.650	0.650	0.8516	0.8558	0.8548	0.8496
			540	360	0.600	0.600	0.8494	0.8524	0.8572	0.8508
		-0.075	195	130	0.950	0.950	0.8730	0.8660	0.8627	0.8598
			360	240	0.900	0.900	0.8644	0.8695	0.8710	0.8508
			510	340	0.850	0.850	0.8579	0.8638	0.8629	0.8508
			645	430	0.800	0.800	0.8585	0.8640	0.8634	0.8535
			750	500	0.750	0.750	0.8537	0.8589	0.8591	0.8508
			840	560	0.700	0.700	0.8538	0.8569	0.8567	0.8508
			915	610	0.650	0.650	0.8543	0.8572	0.8562	0.8527
			960	640	0.600	0.600	0.8527	0.8546	0.9063	0.8508
		-0.050	420	280	0.950	0.950	0.8710	0.8713	0.8611	0.8446
			810	540	0.900	0.900	0.8585	0.8644	0.8648	0.8508
			1140	760	0.850	0.850	0.8525	0.8566	0.8568	0.8485
			1440	960	0.800	0.800	0.8533	0.8566	0.8578	0.8508
			1680	1120	0.750	0.750	0.8504	0.8530	0.8550	0.8493
			1890	1260	0.700	0.700	0.8528	0.8547	0.8551	0.8508
			2040	1360	0.650	0.650	0.8522	0.8539	0.8524	0.8496
			2160	1440	0.600	0.600	0.8533	0.8546	0.8528	0.8508

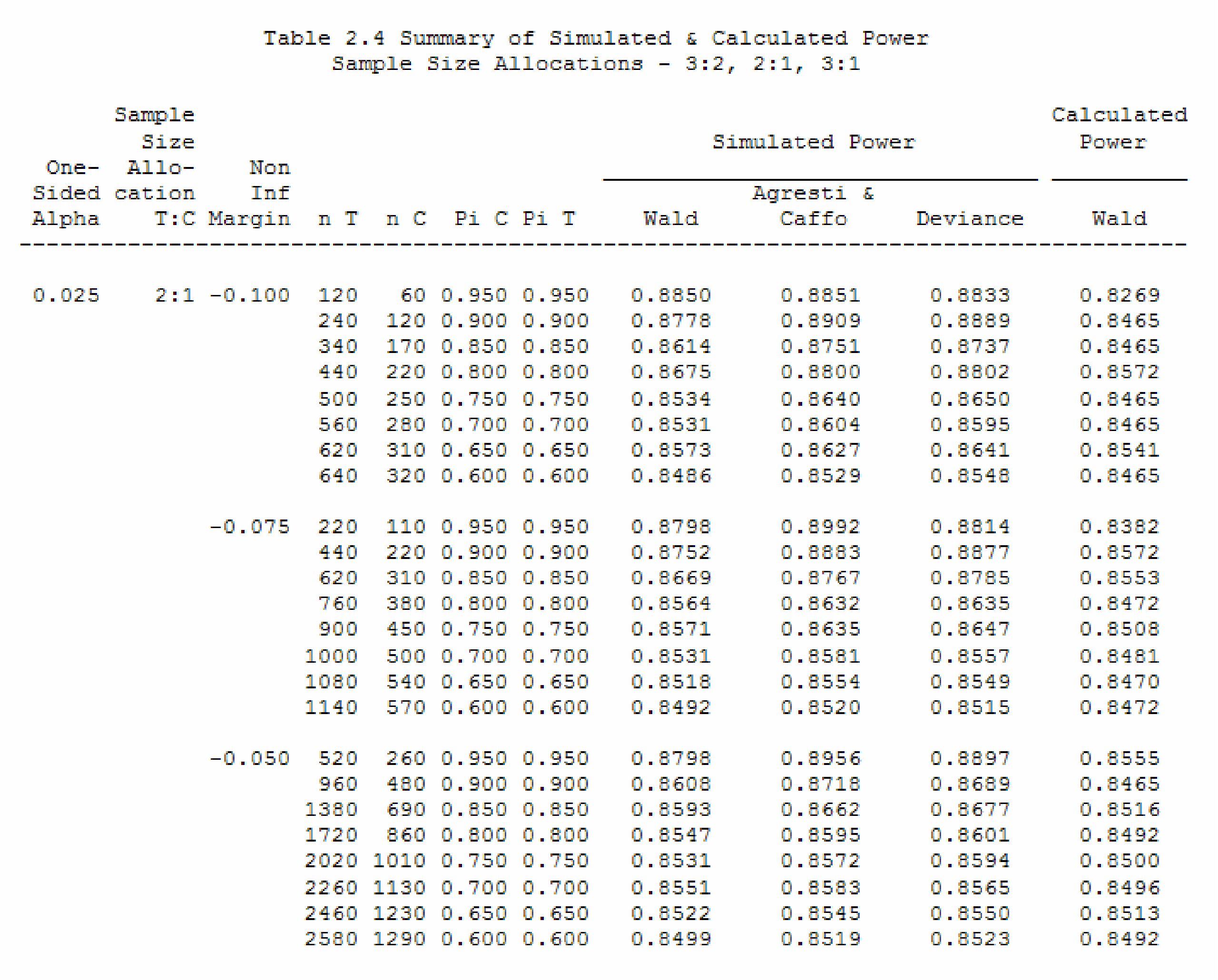


Table 2.4 Summary of Simulated & Calculated Power
Sample Size Allocations - 3:2, 2:1, 3:1

	Sample Size						Simulated Power			Calculated Power
One-Sided Alpha	Allo-cation T:C	Non Inf Margin	n T	n C	Pi C	Pi T	Wald	Agresti & Caffo	Deviance	Wald
0.025	2:1	-0.100	120	60	0.950	0.950	0.8850	0.8851	0.8833	0.8269
			240	120	0.900	0.900	0.8778	0.8909	0.8889	0.8465
			340	170	0.850	0.850	0.8614	0.8751	0.8737	0.8465
			440	220	0.800	0.800	0.8675	0.8800	0.8802	0.8572
			500	250	0.750	0.750	0.8534	0.8640	0.8650	0.8465
			560	280	0.700	0.700	0.8531	0.8604	0.8595	0.8465
			620	310	0.650	0.650	0.8573	0.8627	0.8641	0.8541
			640	320	0.600	0.600	0.8486	0.8529	0.8548	0.8465
		-0.075	220	110	0.950	0.950	0.8798	0.8992	0.8814	0.8382
			440	220	0.900	0.900	0.8752	0.8883	0.8877	0.8572
			620	310	0.850	0.850	0.8669	0.8767	0.8785	0.8553
			760	380	0.800	0.800	0.8564	0.8632	0.8635	0.8472
			900	450	0.750	0.750	0.8571	0.8635	0.8647	0.8508
			1000	500	0.700	0.700	0.8531	0.8581	0.8557	0.8481
			1080	540	0.650	0.650	0.8518	0.8554	0.8549	0.8470
			1140	570	0.600	0.600	0.8492	0.8520	0.8515	0.8472
		-0.050	520	260	0.950	0.950	0.8798	0.8956	0.8897	0.8555
			960	480	0.900	0.900	0.8608	0.8718	0.8689	0.8465
			1380	690	0.850	0.850	0.8593	0.8662	0.8677	0.8516
			1720	860	0.800	0.800	0.8547	0.8595	0.8601	0.8492
			2020	1010	0.750	0.750	0.8531	0.8572	0.8594	0.8500
			2260	1130	0.700	0.700	0.8551	0.8583	0.8565	0.8496
			2460	1230	0.650	0.650	0.8522	0.8545	0.8550	0.8513
			2580	1290	0.600	0.600	0.8499	0.8519	0.8523	0.8492

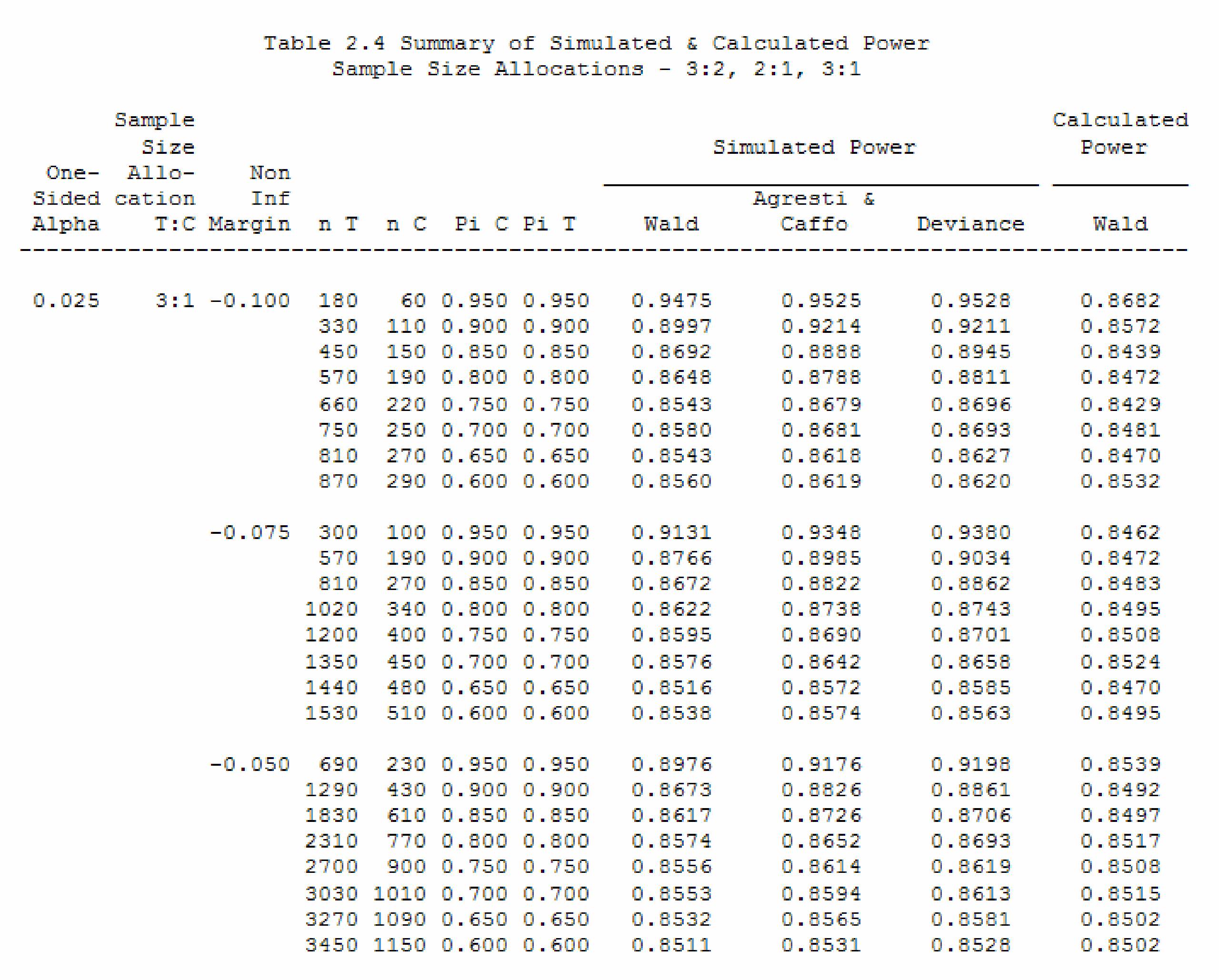


Table 2.4 Summary of Simulated & Calculated Power
Sample Size Allocations - 3:2, 2:1, 3:1

							Simulated Power			Calculated Power
One-Sided Alpha	Sample Size Allocation T:C	Non Inf Margin	n T	n C	Pi C	Pi T	Wald	Agresti & Caffo	Deviance	Wald
0.025	3:1	-0.100	180	60	0.950	0.950	0.9475	0.9525	0.9528	0.8682
			330	110	0.900	0.900	0.8997	0.9214	0.9211	0.8572
			450	150	0.850	0.850	0.8692	0.8888	0.8945	0.8439
			570	190	0.800	0.800	0.8648	0.8788	0.8811	0.8472
			660	220	0.750	0.750	0.8543	0.8679	0.8696	0.8429
			750	250	0.700	0.700	0.8580	0.8681	0.8693	0.8481
			810	270	0.650	0.650	0.8543	0.8618	0.8627	0.8470
			870	290	0.600	0.600	0.8560	0.8619	0.8620	0.8532
		-0.075	300	100	0.950	0.950	0.9131	0.9348	0.9380	0.8462
			570	190	0.900	0.900	0.8766	0.8985	0.9034	0.8472
			810	270	0.850	0.850	0.8672	0.8822	0.8862	0.8483
			1020	340	0.800	0.800	0.8622	0.8738	0.8743	0.8495
			1200	400	0.750	0.750	0.8595	0.8690	0.8701	0.8508
			1350	450	0.700	0.700	0.8576	0.8642	0.8658	0.8524
			1440	480	0.650	0.650	0.8516	0.8572	0.8585	0.8470
			1530	510	0.600	0.600	0.8538	0.8574	0.8563	0.8495
		-0.050	690	230	0.950	0.950	0.8976	0.9176	0.9198	0.8539
			1290	430	0.900	0.900	0.8673	0.8826	0.8861	0.8492
			1830	610	0.850	0.850	0.8617	0.8726	0.8706	0.8497
			2310	770	0.800	0.800	0.8574	0.8652	0.8693	0.8517
			2700	900	0.750	0.750	0.8556	0.8614	0.8619	0.8508
			3030	1010	0.700	0.700	0.8553	0.8594	0.8613	0.8515
			3270	1090	0.650	0.650	0.8532	0.8565	0.8581	0.8502
			3450	1150	0.600	0.600	0.8511	0.8531	0.8528	0.8502

Figure 2.1 Summary of Simulated Type I Error
By Sample Size Allocation
Alpha=0.025

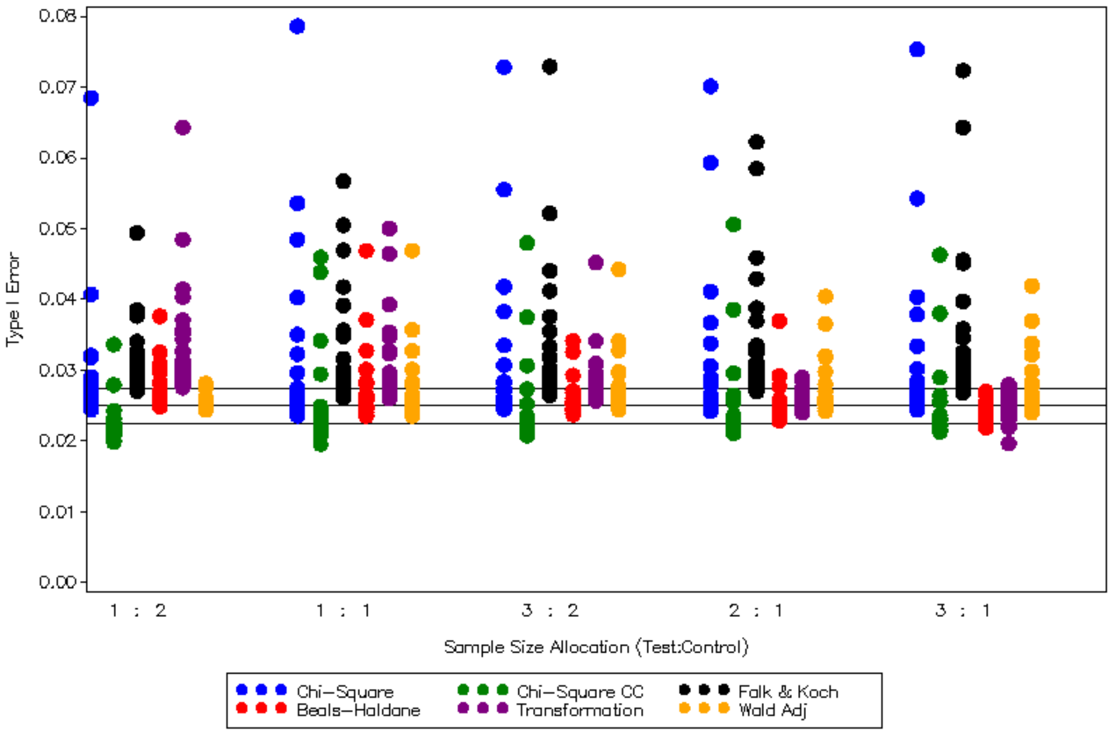


Figure 2.2 Summary of Simulated Type I Error
By Sample Size Allocation
Alpha=0.025

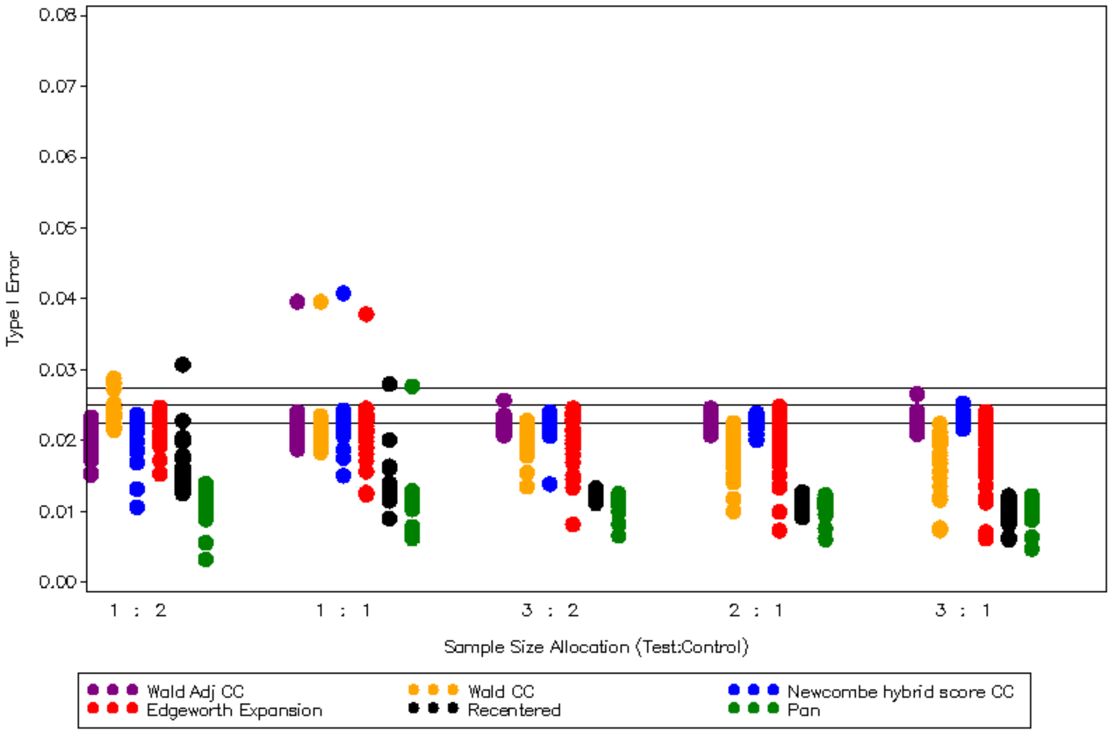


Figure 2.3 Summary of Simulated Type I Error
By Sample Size Allocation
Alpha=0.025

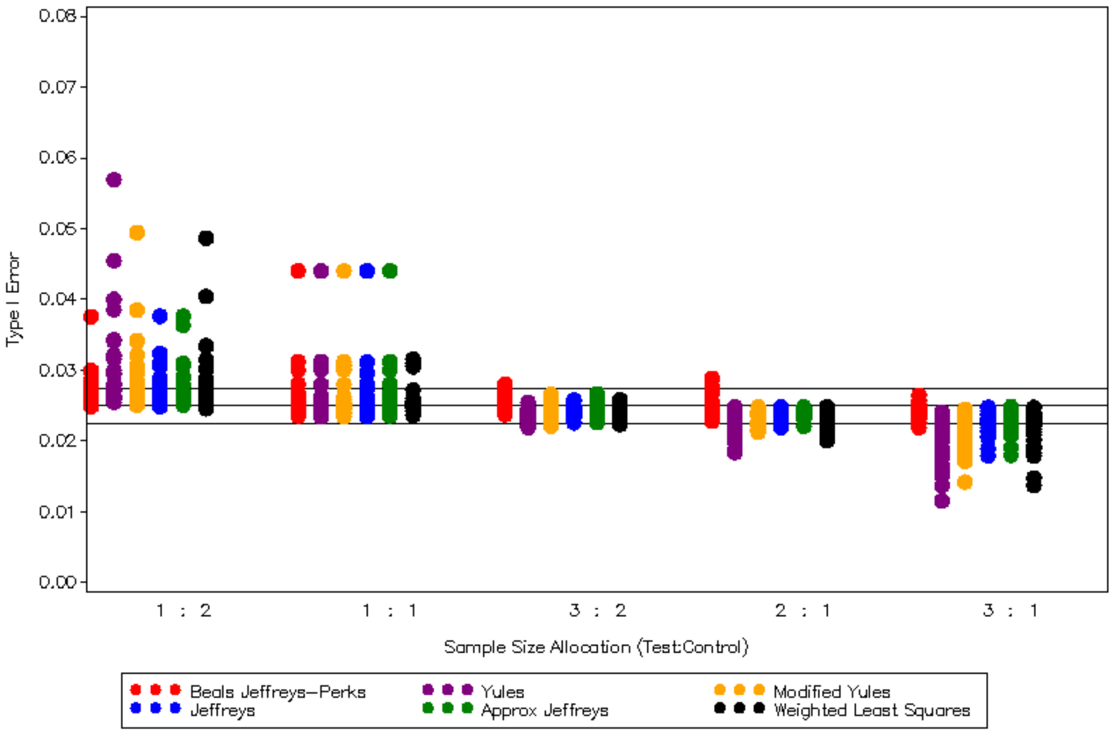


Figure 2.4 Summary of Simulated Type I Error
By Sample Size Allocation
Alpha=0.025

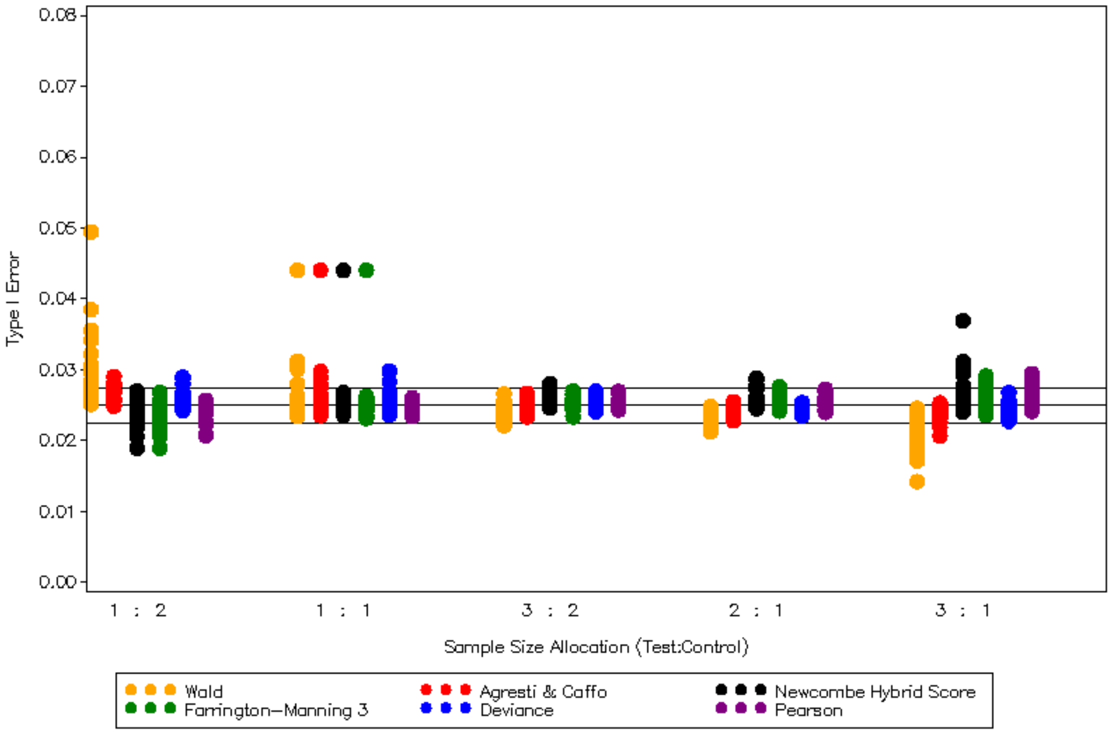


Figure 2.5 Summary of Simulated Type I Error

By Non–inferiority Margin
Alpha=0.025

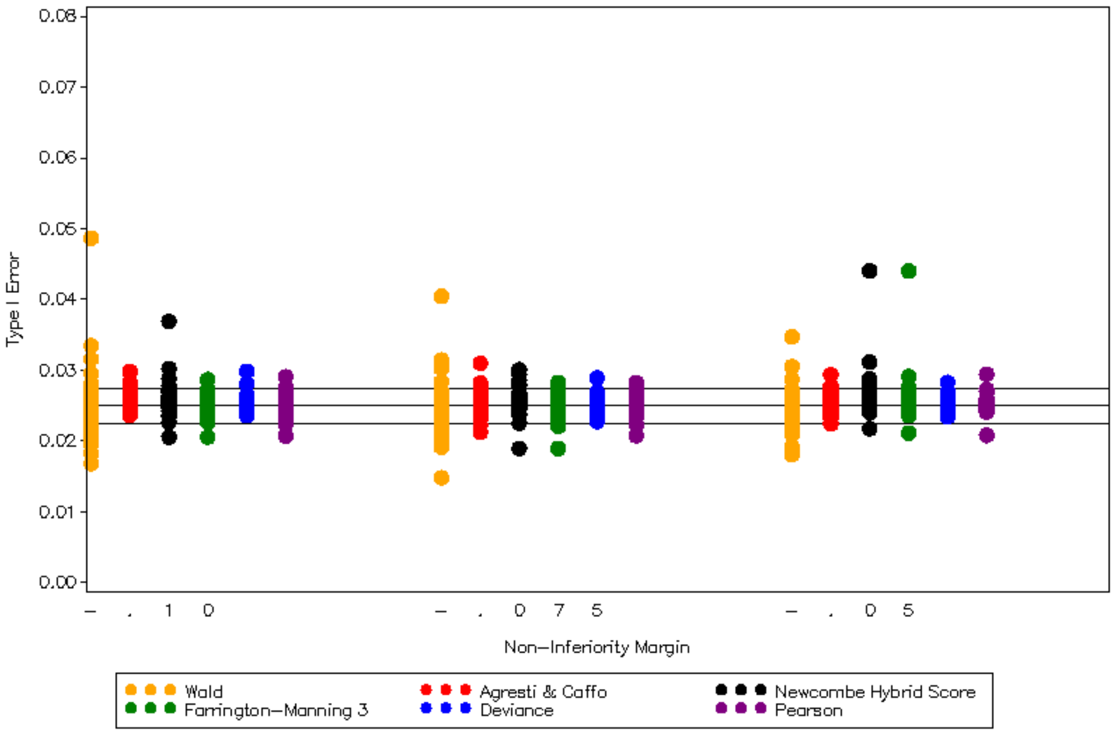


Figure 2.6 Summary of Simulated Type I Error

By Pi C
Alpha=0.025

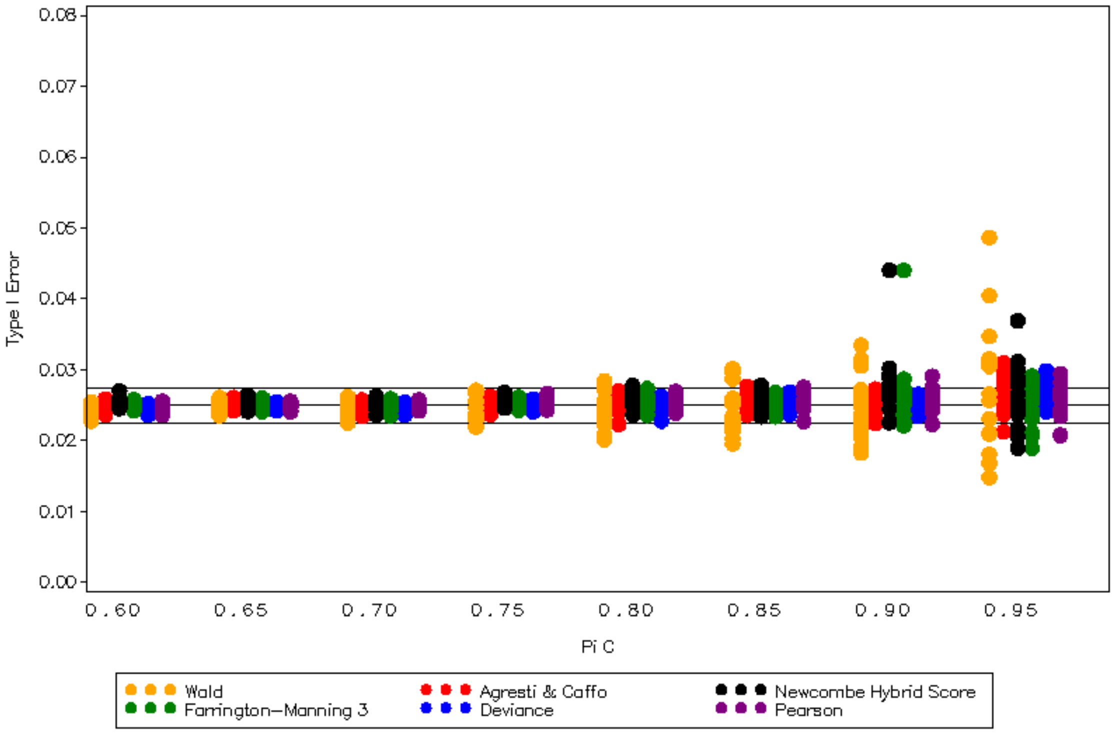


Figure 2.7 Summary of Simulated Power
By Sample Size Allocation
Alpha=0.025, Delta=−Non−Inf Margin/2

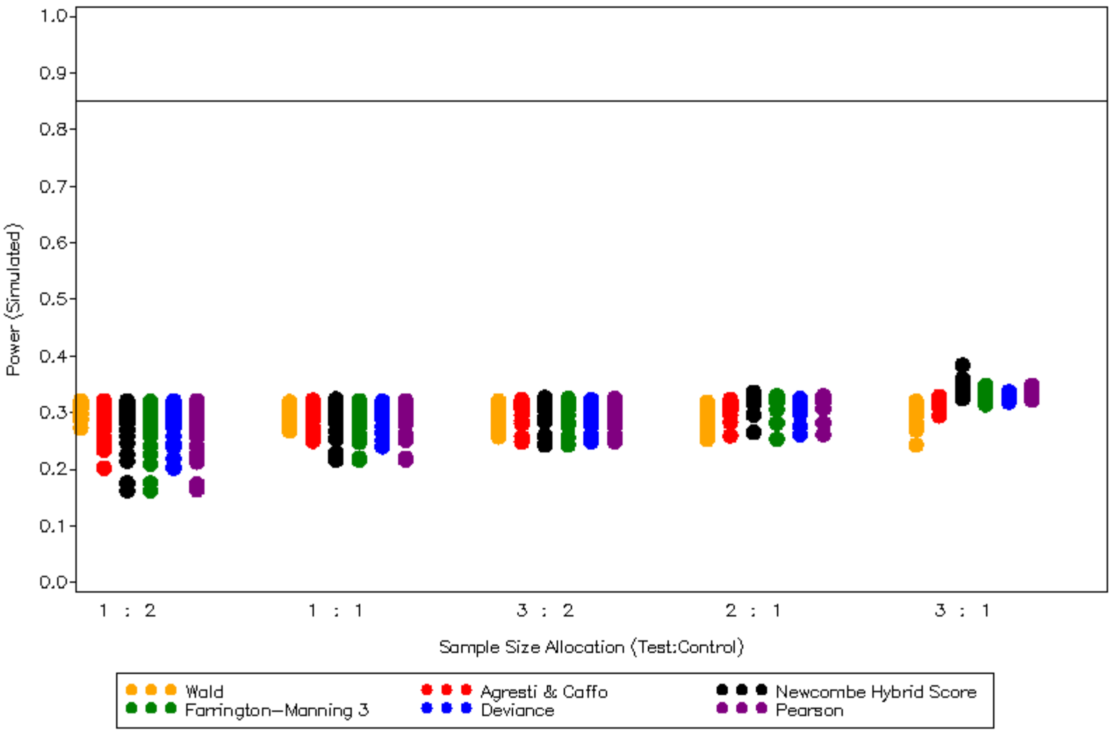


Figure 2.8 Summary of Simulated Power
By Sample Size Allocation
Alpha=0.025, Delta=0

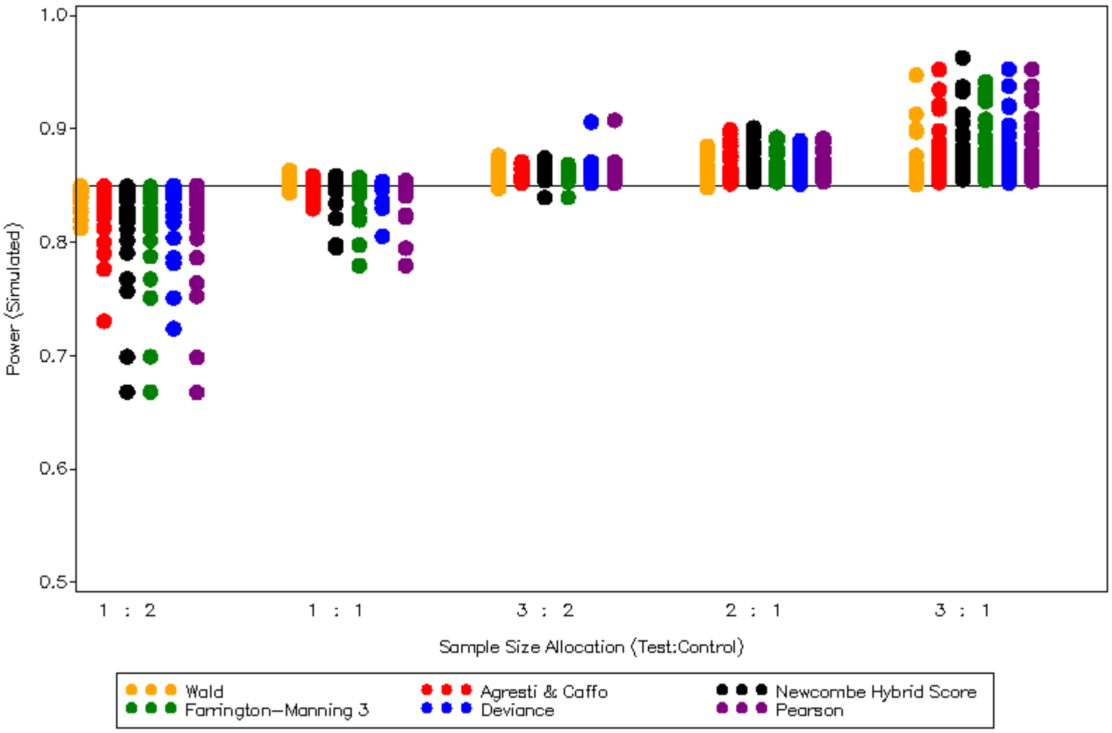


Figure 2.9 Summary of Simulated Power
By Sample Size Allocation
Alpha=0.025, Delta=0.01

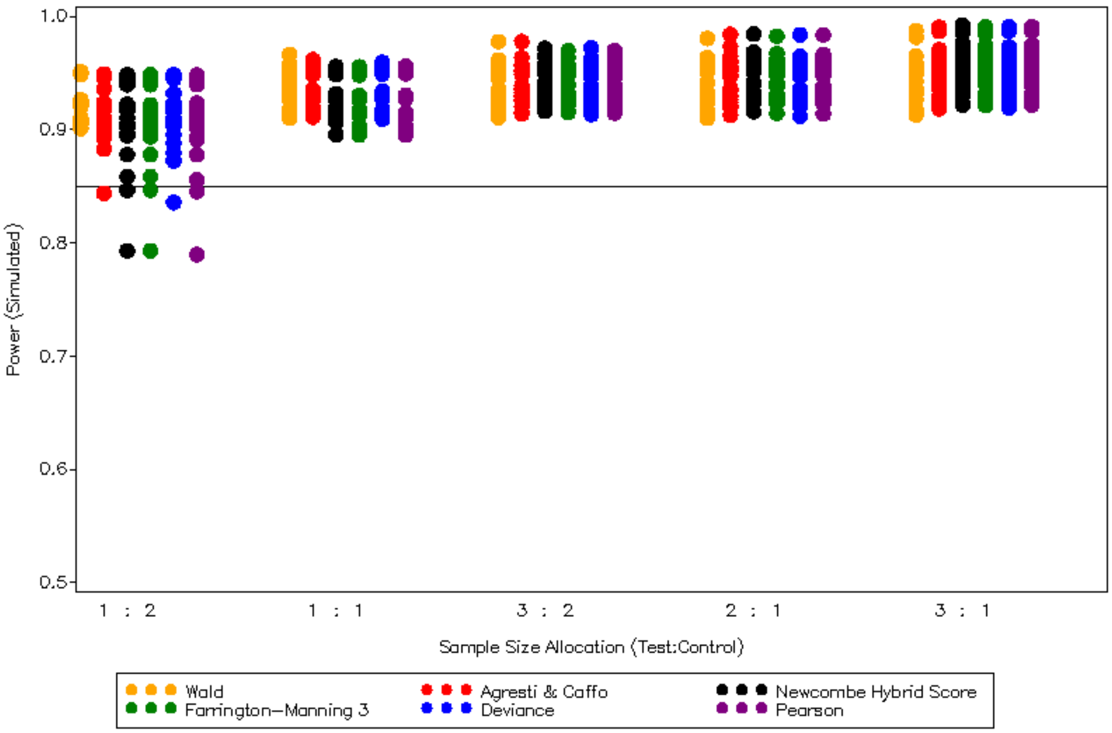


Figure 2.10 Summary of Simulated Power
By Sample Size Allocation
Alpha=0.025, Delta=0.025

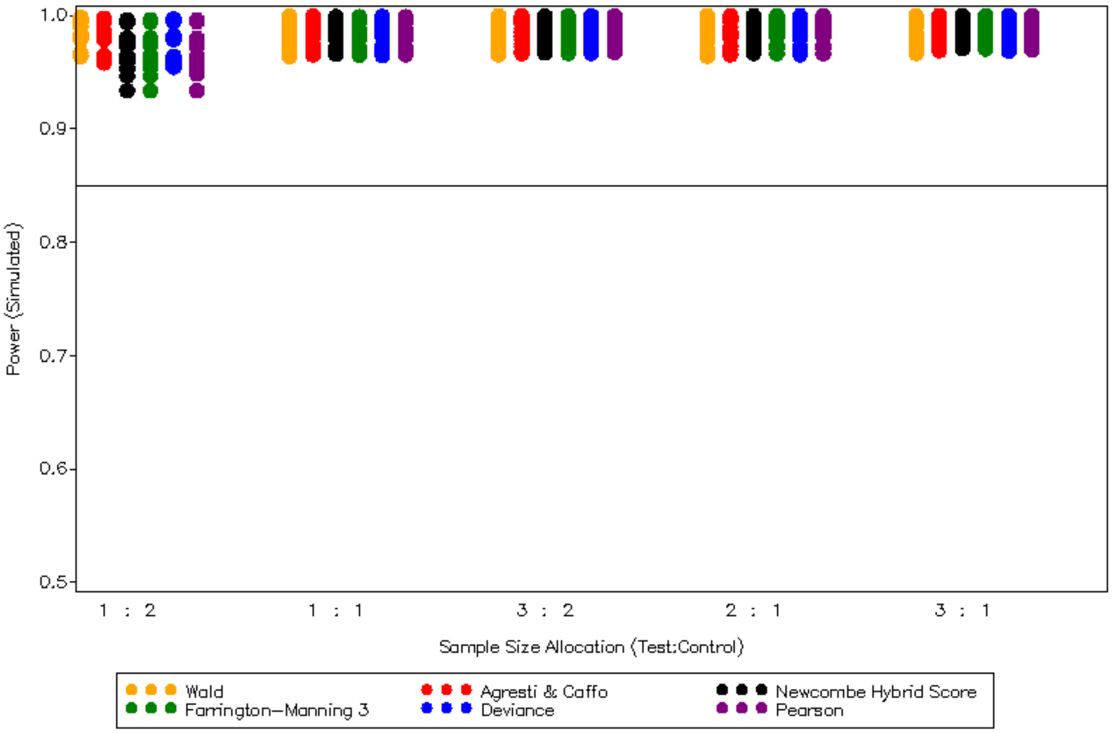


Figure 2.11 Comparison of Simulated Power

Newcombe Hybrid Score and Pearson Methods
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 1:2, 1:1

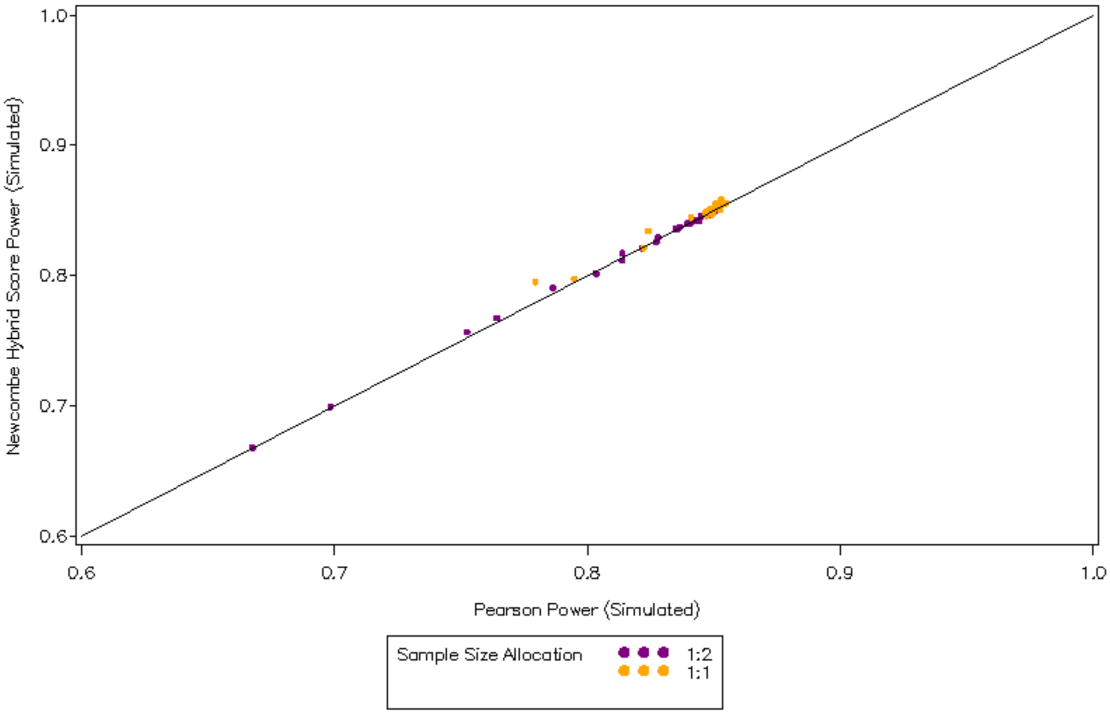


Figure 2.12 Comparison of Simulated Power

Farrington–Manning 3 and Pearson Methods
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 1:2, 1:1

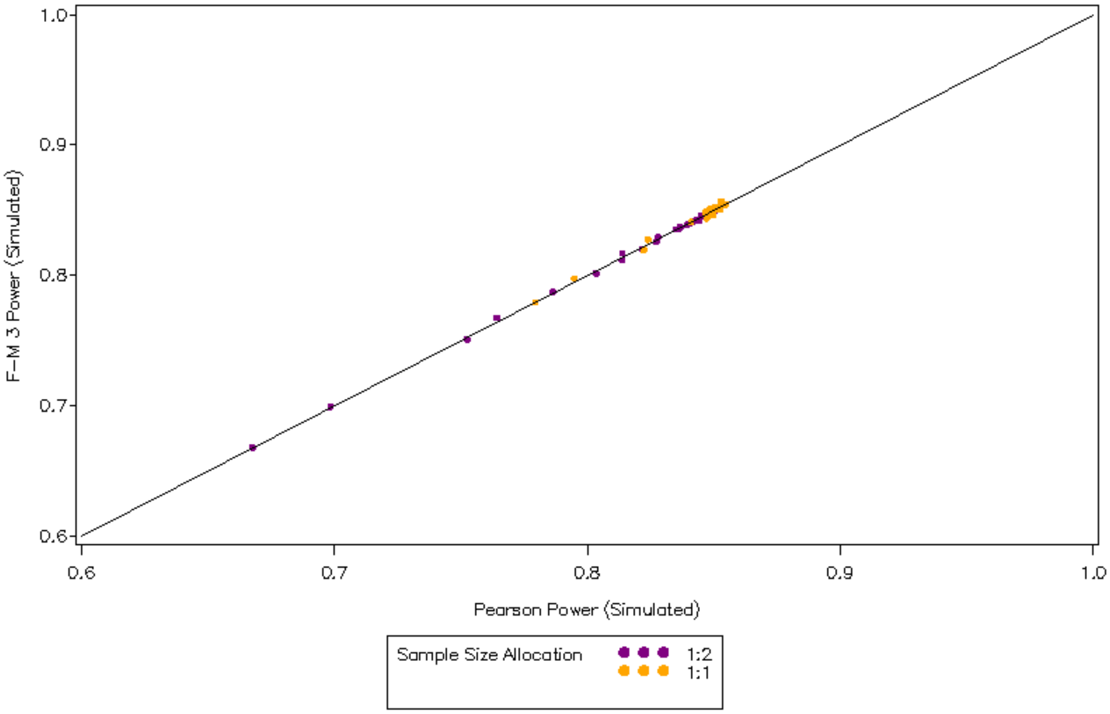


Figure 2.13 Comparison of Simulated Power

Farrington–Manning 3 and Newcombe Hybrid Score Methods
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 1:2, 1:1

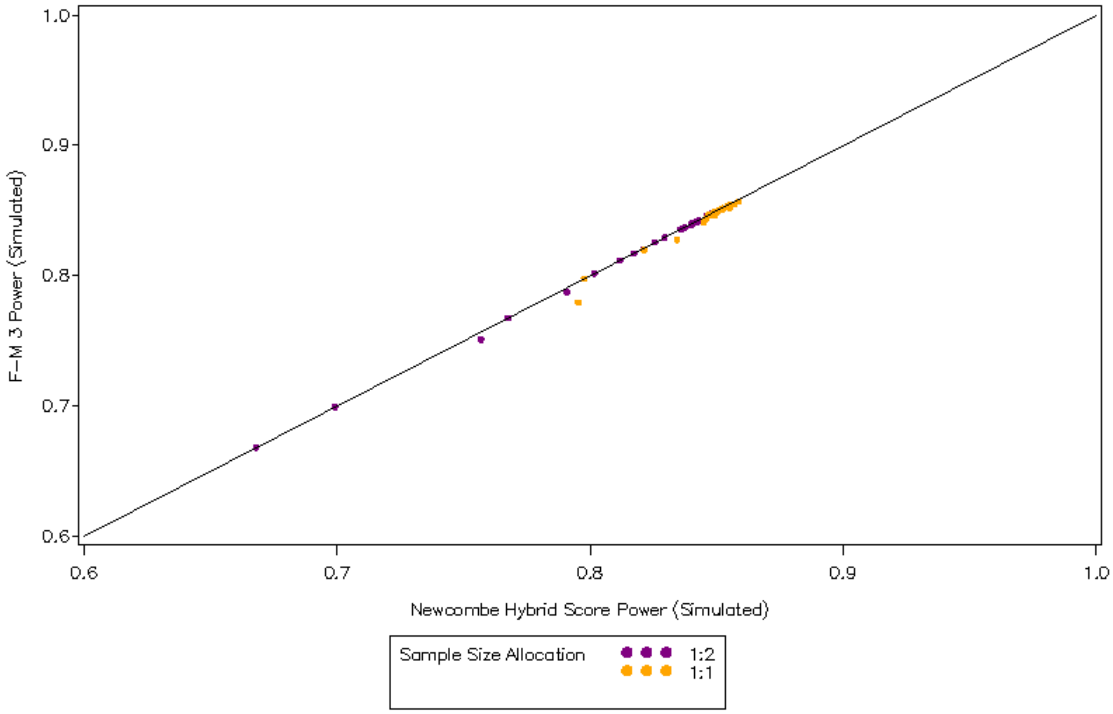


Figure 2.14 Comparison of Simulated Power

Agresti & Caffo and Deviance Methods
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 3:2, 2:1, 3:1

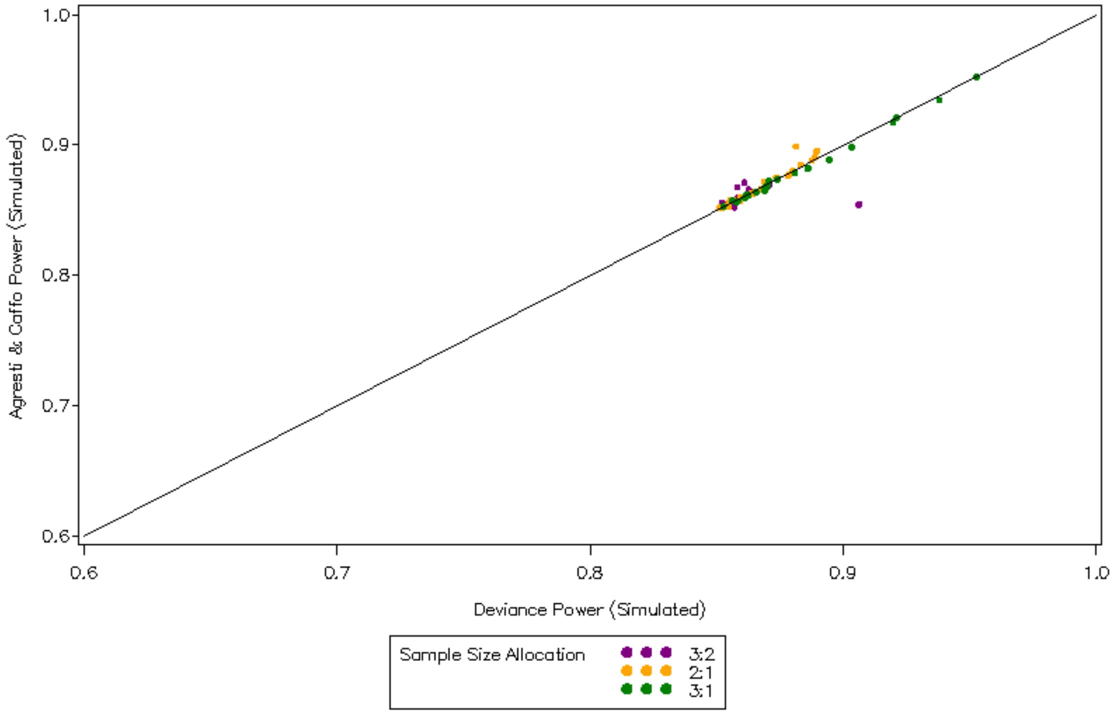


Figure 2.15 Comparison of Simulated Power
Wald and Deviance Methods
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 3:2, 2:1, 3:1

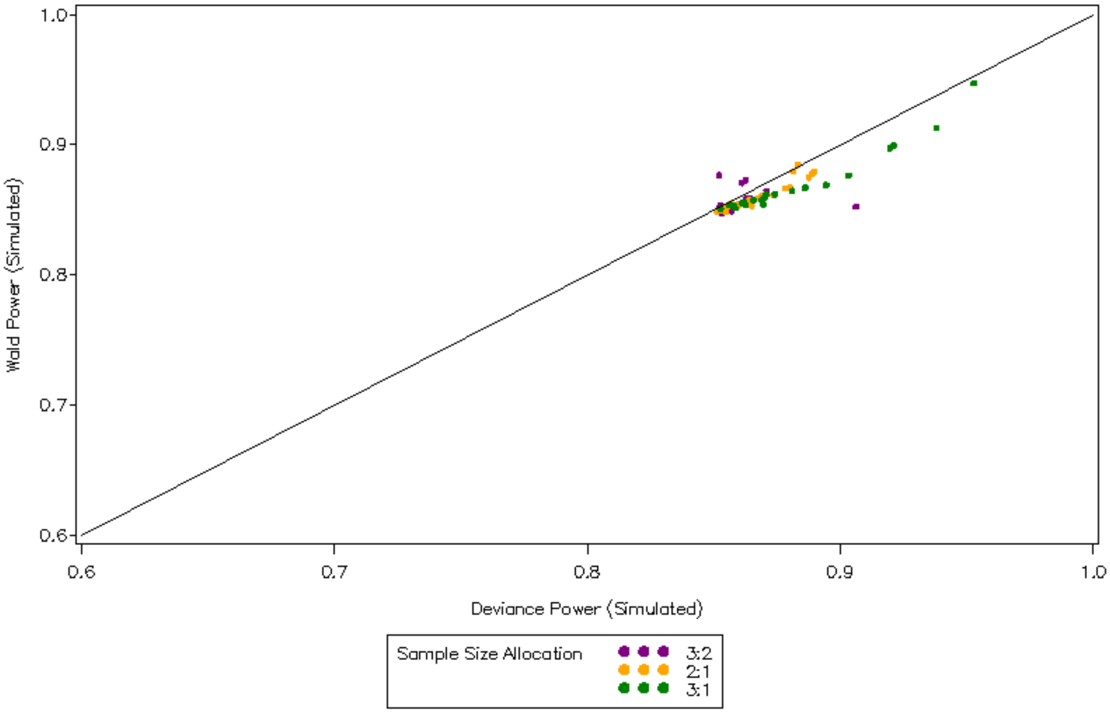


Figure 2.16 Comparison of Simulated Power
Wald and Agresti & Caffo Methods
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 3:2, 2:1, 3:1

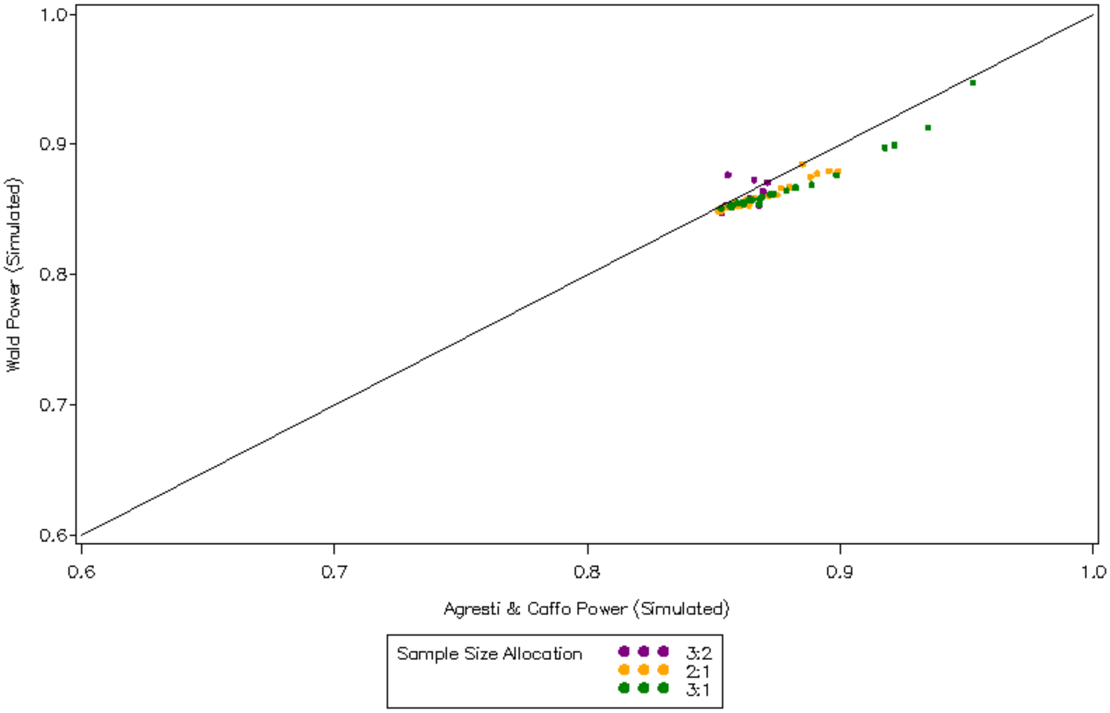


Figure 2.17 Comparison of Simulated and Calculated Power

F–M 3 Method
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 1:2, 1:1

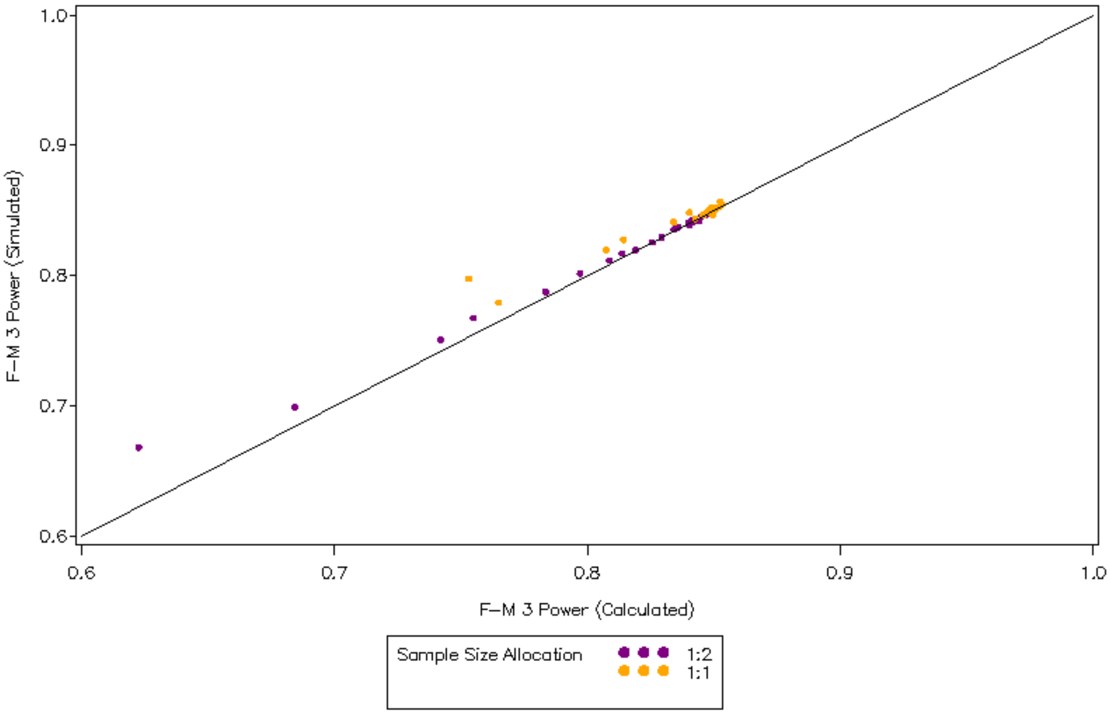


Figure 2.18 Comparison of Simulated and Calculated Power

Newcombe Hybrid Score Simulated Method & F–M 3 Calculated Method
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 1:2, 1:1

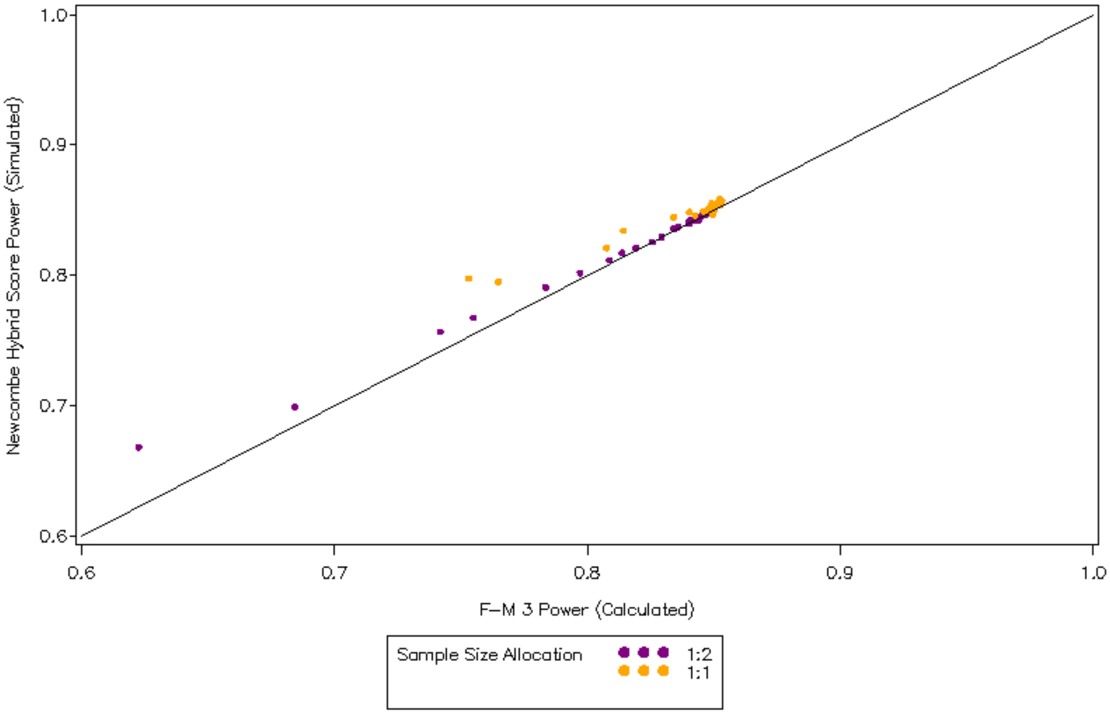


Figure 2.19 Comparison of Simulated and Calculated Power
Pearson Simulated Method & F–M 3 Calculated Method
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 1:2, 1:1

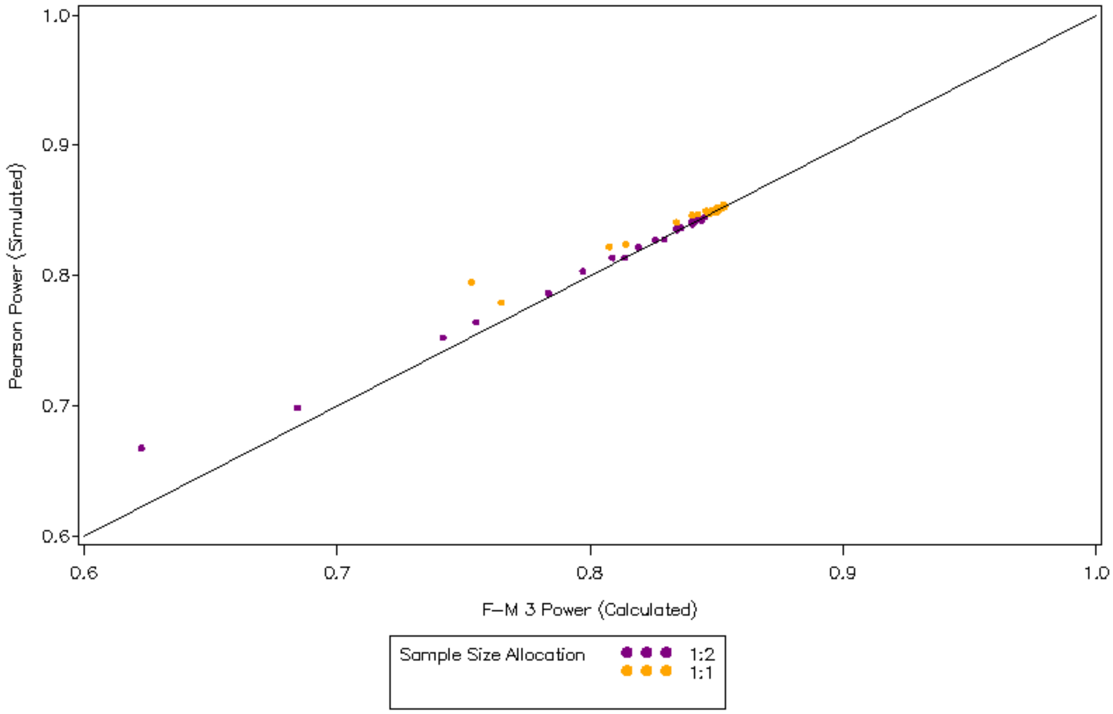


Figure 2.20 Comparison of Simulated and Calculated Power
Wald Method
By Sample Size Allocation
Alpha=0.025, Delta=0, Allocation 3:2, 2:1, 3:1

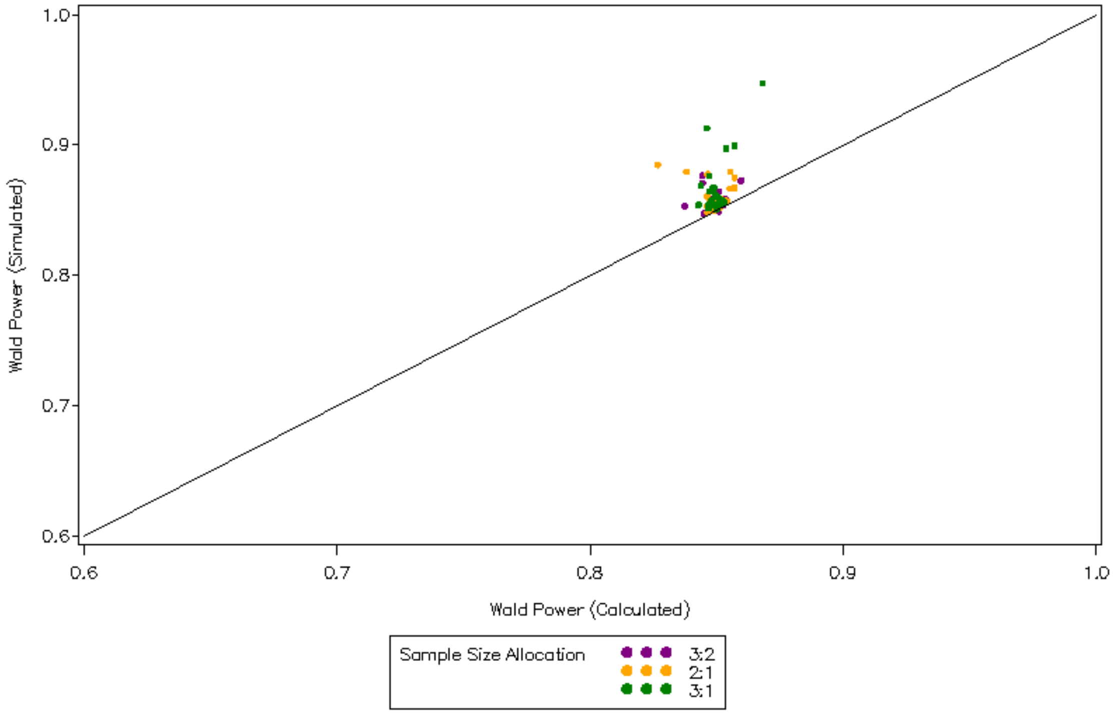


Figure 2.21 Comparison of Simulated and Calculated Power

Deviance Simulated Method & Wald Calculated Method

By Sample Size Allocation

Alpha=0.025, Delta=0, Allocation 3:2, 2:1, 3:1

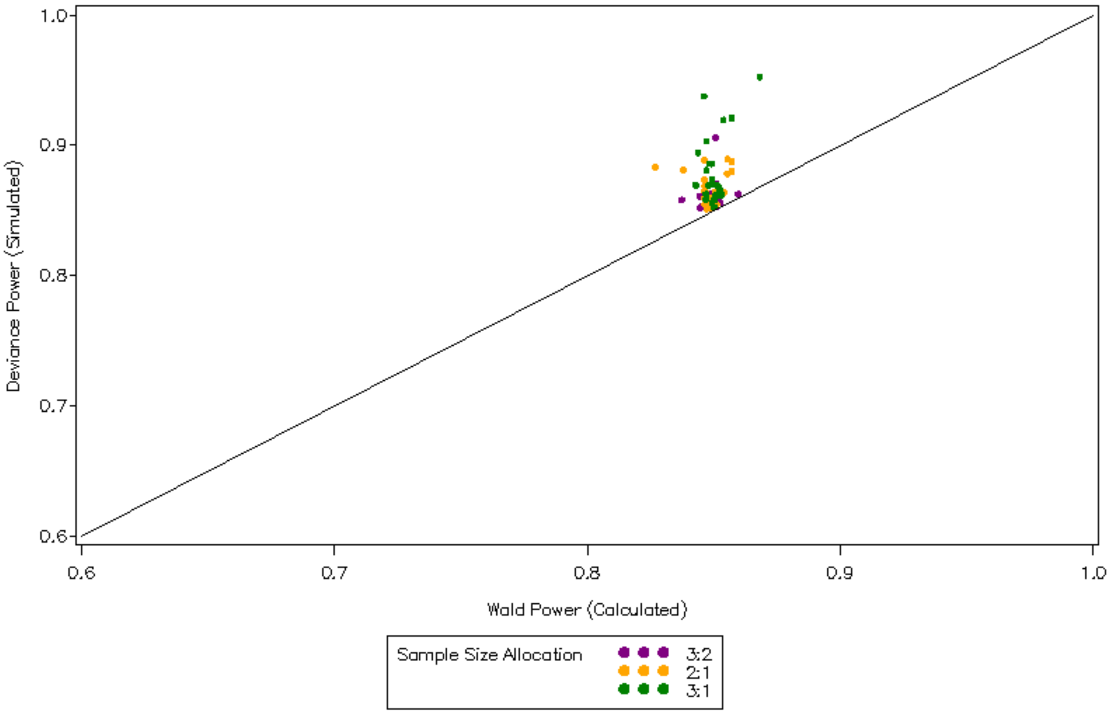
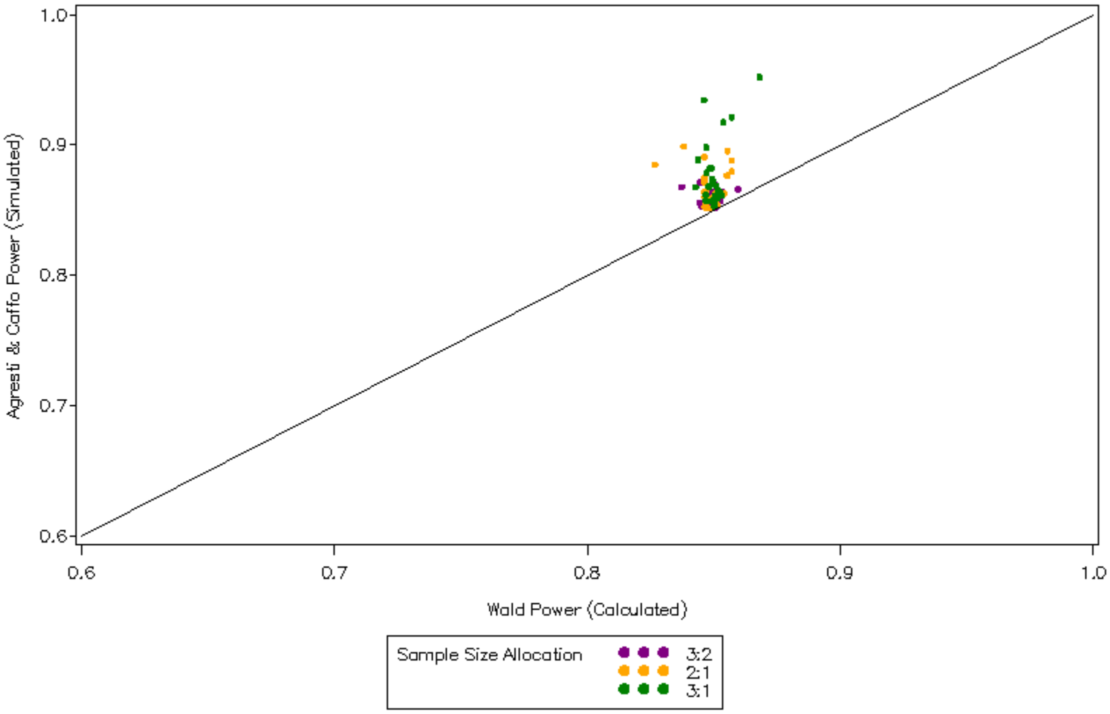


Figure 2.22 Comparison of Simulated and Calculated Power

Agresti & Caffo Simulated Method & Wald Calculated Method

By Sample Size Allocation

Alpha=0.025, Delta=0, Allocation 3:2, 2:1, 3:1



Chapter 3

Methods for Analyzing Three-arm Trials with Binomial Proportions as the Primary Endpoint

I. Introduction

Non-inferiority clinical trials are used in a setting where the new experimental medication, test treatment, must be not unacceptably worse than the current active control treatment by a specified amount related to a condition of interest. Their importance in the pharmaceutical industry is becoming more widespread[1]. In many settings, it is also important to include a placebo arm in these trials for reasons discussed by Koch & Röhmel including situations where the active control may lack compelling proof of efficacy, the effect of the active control is small compared to the placebo effect, the active control effect over placebo is widely variable among trials, and the understanding of the condition of interest is not complete[1]. Three-arm trials are then used to conclude non-inferiority of the test treatment to the active control by showing that the test treatment preserves a certain pre-determined percentage of effect over placebo that the active control treatment preserves over placebo.

Frequently the primary endpoint of these three-arm trials is a dichotomous outcome resulting in a proportion for each treatment. Pigeot et. al.[2] and Schwartz[3] have discussed methodology for assessing the percentage of effect preserved related to the non-inferiority hypothesis for a continuous outcome. Tang and Tang[4] have modified these methods to use proportions from binomial endpoints. Additionally, sample size and power formulas are of interest for this setting to aid in the planning of three-arm trials with proportions as the primary endpoint.

II. Methods for Assessing Non-inferiority in a Three-arm Trial

The null hypothesis of inferiority in these three-arm trials is created to perform a statistical test of the percentage of effect that the test treatment to placebo preserves over the effect of the active control treatment to placebo, and can be written as H_0: $(\pi_T-\pi_P)/(\pi_C-\pi_P) = \lambda \leq \lambda_0$ with the alternative hypothesis as H_A: $(\pi_T-\pi_P)/(\pi_C-\pi_P) = \lambda > \lambda_0$ with π_T, π_C, and π_P representing the population proportion of patients having the outcome of interest in the test, active control, and placebo groups, respectively. The pre-determined percentage of effect that the test treatment must preserve is $100\lambda_0$.

The first of these methods creates a Wald statistic[4] using heterogeneous variances for the treatment groups as seen in (3.1).

$$Z_W = \frac{\left\{\hat{p}_T - \lambda_0\hat{p}_C - (1-\lambda_0)\hat{p}_P\right\}}{\sqrt{\frac{1}{n_T}\left\{\hat{p}_T(1-\hat{p}_T) + \frac{\lambda_0^2\hat{p}_C(1-\hat{p}_C)}{C_C} + \frac{(1-\lambda_0)^2\hat{p}_P(1-\hat{p}_P)}{C_P}\right\}}} \tag{3.1}$$

where $\hat{p}_T$, $\hat{p}_C$, and $\hat{p}_P$ are the observed proportions of the outcome of interest in the test, active control, and placebo groups, respectively. Additionally, the sample size in the test group is represented as n_T with the sample sizes in the active control and placebo groups represented as a proportion of the sample size in the test group where $n_C = C_C n_T$ and $n_P = C_P n_T$. This statistic can be compared to a standard normal distribution yielding a p-value for the test of the null hypothesis.

A confidence interval for λ can be computed from this statistic based on Fieller's method to yield a lower and upper limit which contains all possible values of λ which would not be rejected using the Wald statistic, Z_W in (3.1). These confidence limits are the solutions to the equation in (3.2)

$$\frac{n_T\left\{\hat{p}_T - \lambda\hat{p}_C - (1-\lambda)\hat{p}_P\right\}^2}{\left\{\hat{p}_T(1-\hat{p}_T) + \frac{\lambda^2\hat{p}_C(1-\hat{p}_C)}{C_C} + \frac{(1-\lambda)^2\hat{p}_P(1-\hat{p}_P)}{C_P}\right\}} = z_\alpha^2 \qquad (3.2)$$

where z_α is the 100(1-α) percentile for a standard normal distribution. The equation in (3.2) is then solved for λ to produce the upper and lower confidence limits (λ_{WL}, λ_{WU}). These limits are computed as in (3.3). The lower limit, λ_{WL}, is the focus of the current discussion.

$$\lambda_{WL} = \frac{-B_W - \sqrt{B_W^2 - 4A_W C_W}}{2A_W}, \lambda_{WU} = \frac{-B_W + \sqrt{B_W^2 - 4A_W C_W}}{2A_W} \qquad (3.3)$$

$$A_W = \left\{(\hat{p}_C - \hat{p}_P)^2 - \frac{z_\alpha^2\hat{p}_C(1-\hat{p}_C)}{C_C n_T} - \frac{z_\alpha^2\hat{p}_P(1-\hat{p}_P)}{C_P n_T}\right\}$$

$$B_W = \left\{-2(\hat{p}_C - \hat{p}_P)(\hat{p}_T - \hat{p}_P) + \frac{2z_\alpha^2\hat{p}_P(1-\hat{p}_P)}{C_P n_T}\right\}$$

$$C_W = \left\{(\hat{p}_T - \hat{p}_P)^2 - \frac{z_\alpha^2\hat{p}_T(1-\hat{p}_T)}{n_T} - \frac{z_\alpha^2\hat{p}_P(1-\hat{p}_P)}{C_P n_T}\right\}$$

Another method for assessing the non-inferiority hypothesis is one that is modified from that for the difference in proportions as proposed by Agresti and Caffo[5]. This method adds one success and one failure to each group, thereby adding three successes and three failures in total. This Modified Agresti & Caffo statistic as seen in (3.4) where

$\tilde{p}_T = (y_T + 1)/(n_T + 2)$, $\tilde{p}_C = (y_C + 1)/(n_C + 2)$, and $\tilde{p}_P = (y_P + 1)/(n_P + 2)$ can also be compared to a standard normal distribution yielding a p-value for the test of the null hypothesis of inferiority.

$$Z_{AC} = \frac{\left\{\tilde{p}_T - \lambda_0 \tilde{p}_C - (1-\lambda_0)\tilde{p}_P\right\}}{\sqrt{\dfrac{\tilde{p}_T(1-\tilde{p}_T)}{n_T+2} + \dfrac{\lambda_0^2 \tilde{p}_C(1-\tilde{p}_C)}{n_C+2} + \dfrac{(1-\lambda_0)^2 \tilde{p}_P(1-\tilde{p}_P)}{n_P+2}}} \tag{3.4}$$

A confidence interval as similar to that computed for the Wald statistic can be computed for the Agresti & Caffo method by solving the equation in (3.5) for λ to produce upper and lower confidence limits (λ_{ACL}, λ_{ACU}), with the current discussion focusing on λ_{ACL}, the lower confidence limit.

$$\frac{\left\{\tilde{p}_T - \lambda_0 \tilde{p}_C - (1-\lambda_0)\tilde{p}_P\right\}^2}{\left\{\dfrac{\tilde{p}_T(1-\tilde{p}_T)}{n_T+2} + \dfrac{\lambda_0^2 \tilde{p}_C(1-\tilde{p}_C)}{n_C+2} + \dfrac{(1-\lambda_0)^2 \tilde{p}_P(1-\tilde{p}_P)}{n_P+2}\right\}} = z_\alpha^2 \tag{3.5}$$

These limits are computed as in (3.6).

$$\lambda_{ACL} = \frac{-B_{AC} - \sqrt{B_{AC}^2 - 4A_{AC}C_{AC}}}{2A_{AC}}, \lambda_{ACU} = \frac{-B_{AC} + \sqrt{B_{AC}^2 - 4A_{AC}C_{AC}}}{2A_{AC}} \tag{3.6}$$

$$A_{AC} = \left\{(\tilde{p}_C - \tilde{p}_P)^2 - \frac{z_\alpha^2 \tilde{p}_C(1-\tilde{p}_C)}{n_C} - \frac{z_\alpha^2 \tilde{p}_P(1-\tilde{p}_P)}{n_p}\right\}$$

$$B_{AC} = \left\{-2(\tilde{p}_C - \tilde{p}_P)(\tilde{p}_T - \tilde{p}_P) + \frac{2z_\alpha^2 \tilde{p}_P(1-\tilde{p}_P)}{n_P}\right\}$$

$$C_{AC} = \left\{(\tilde{p}_T - \tilde{p}_P)^2 - \frac{z_\alpha^2 \tilde{p}_T(1-\tilde{p}_T)}{n_T} - \frac{z_\alpha^2 \tilde{p}_P(1-\tilde{p}_P)}{n_P}\right\}$$

Non-inferiority can also be assessed through the use of estimators obtained from maximum likelihood methods for the proportion of events in the test, active control, and placebo groups based on the joint distribution of the events as the product of three independent binomial distributions for the three groups. The first is based on the Deviance statistic which is computed as the -2 times the difference in the natural logarithms of the likelihood using the proportions computed from the maximum likelihood estimators under the alternative hypothesis and the likelihood using the proportions computed from the maximum likelihood estimators under the null hypothesis. The second of these is based on a Pearson statistic in the form of [(observed – expected)^2 / expected] using the maximum likelihood estimators under the null hypothesis. These methods produce test statistics for which p-values can be obtained by using the chi-square distribution under one degree of freedom.

Additionally, Tang and Tang[4] present results for a test statistic based on maximum likelihood estimates (RMLE) of the proportions restricted under the null hypothesis. These RMLE estimates are used in the denominator for (3.1) to replace the observed proportions. This statistic is also compared to a standard normal distribution to produce a corresponding p-value. These RMLE estimates do not have a closed-form solution and therefore require additional resources for their computation. Software such as SAS[6] can be used to obtain these estimates through PROC GENMOD, a procedure used to fit generalized linear models, with a specification of a binomial distribution and an identity link. The model statement fits one parameter for the control group, one parameter for the placebo group, with the test group being restricted by the null hypothesis λ_0 for the parameter for the control group and (1- λ_0) for the parameter for the placebo group.

Another method can be used which replaces the observed proportions used in the denominator in (3.1) with proportions restricted by the null hypothesis, but using weighted least squares for estimation instead of maximum likelihood estimation. These estimates can also be obtained using software such as SAS[6] through PROC CATMOD, a procedure used to fit categorical models, with a similar specification as in the above model using weighted least squares to estimate the means.

The methods previously discussed, including the Deviance method, Pearson method, RMLE method, and the weighted least squares (WLS) method are described in the form of a test statistic. It is possible to produce corresponding confidence limits through an iterative process of computing an interval of all possible values of the null hypothesis which the test statistic does not reject at the specified alpha level.

III. Performance of Methods based on Simulations for Assessing Non-inferiority

Simulations were used to study the properties of these methods in various scenarios to assess type I error and power. Scenarios included varying the following parameters:

1. π_C, the population proportion of events in the control group: 0.6, 0.7, 0.8, 0.9
2. π_P, the population proportion of events in the placebo group: 0.2, 0.3, 0.4, 0.5, 0.6, 0.7
3. λ_0, the percentage of effect that the test treatment must preserve under the null hypothesis: 0.6, 0.7, 0.8, 0.9
4. $\lambda=(\pi_T - \pi_P)/(\pi_C - \pi_P)$, the population percentage of effect that the test treatment preserves: λ_0, $(1+ \lambda_0)/2$, 1, 1.1
5. π_T, the population proportion of events in the test group: $\pi_T = \lambda\pi_C + (1- \lambda)\pi_P$
6. α, the one-sided alpha level: 0.000625, 0.005, 0.01, 0.025

7. n_T, the sample size in the test group is calculated to have 85% power to contradict the null hypothesis for the specified placebo proportion and no difference between the test and active control treatment groups with $C_C=n_C/n_T$ and $C_P=n_P/n_T$ as $n_T = \frac{(z_\alpha + z_{1-\beta})^2 \left\{(1+\lambda_0^2)\pi_C(1-\pi_C) + (1-\lambda_0)^2 \pi_P (1-\pi_P)\right\}}{(1-\lambda_0)^2 (\pi_C - \pi_P)^2}$
8. n_C, the sample size in the active control group: $n_C=C_C n_T$
9. n_P, the sample size in the placebo group: $n_P=C_P n_T$
10. Sample size allocation for test:active control:placebo as 1:1:1, 2:1:1, 2:2:1, 3:1:1, 3:2:1, 3:3:1

For each of the 10,000 replications performed, a sample from the specified binomial distribution was drawn separately for the test, active-control, and placebo groups. All of the test statistics were computed for the same replication and a conclusion of non-inferiority or not was determined for the applicable one-sided test according to whether or not the p-value from the corresponding test statistic was smaller than the nominal alpha level. The average of these indicator variables for the demonstration of non-inferiority produced a simulated power for the methods when the true percentage of effect preserved for test over active control exceeded the null hypothesis and a type I error rate when this true percentage of effect was equal to the null hypothesis.

When the number of events in any of the treatment groups was zero or if a method failed to produce a logical test statistic, because of estimated proportions being outside of the (0,1) range, then the Agresti and Caffo method was used as a replacement.

IV. Results of Simulations

A brief summary of the simulation results can be found in Table 3.1. Discussion will include type I error considerations, power considerations, and sample size calculations for the design of these non-inferiority trials.

A. Type I Error Considerations

The Wald method generally yielded the highest type I errors as compared to all other methods. The WLS method also tended to yield higher type I error rates than other methods. The RMLE, Deviance, and Pearson methods tended to produce type I errors closest to the nominal level for most scenarios, as displayed in Figures 3.1 – 3.6 by the parameters varied in the simulations.

The type I error performance of the methods was similar across alpha levels, although these results are not shown. The type I error performance was closer to the nominal level for the Wald method when allocations included 2:1 or 3:1 for test:control. The opposite was seen for the WLS method with type I errors further from the nominal level for allocations of 2:1 or 3:1 for test:control as seen in Figure 3.1. The performance of the Deviance, Pearson, and RMLE methods seem to be unaffected by the choice of sample size allocation.

The simulated type I error rates were much closer to the nominal level and less variable with larger non-inferiority margin (λ_0) in Figure 3.2, smaller event rates in the control group (π_C) in Figure 3.3, and larger event rates in the placebo group (π_P) in Figure 3.4; with these specific scenarios having smaller sample size as well. The Wald, Agresti & Caffo, and WLS methods tended to have type I errors higher than the nominal level when

π_C=0.8, 0.9 and π_P=0.2, 0.3 where event rates may have been small. These patterns are seen in Figure 3.5 which displays type I errors by π_C- π_P. The larger the difference between the event rates in the control and placebo groups, the more type I error violations occur for the Wald, Agresti & Caffo, and WLS methods. Type I error violations also occur for smaller sample sizes as seen in Figure 3.6.

B. Power Considerations

Discussion of power will focus only on situations where the methods maintained the approximate nominal type I error levels. The RMLE, Deviance, and Pearson methods tend to maintain nominal type I error levels for all scenarios. The RMLE and Pearson methods yield almost identical power results as seen in Figure 3.7. Therefore, further discussions will include only the RMLE method. The Deviance method tends to produce slightly higher power than the RMLE method as seen in Figure 3.8, but this difference is not very large for most cases with the largest discrepancies for larger values of π_C, specifically π_C =0.9.

In the cases where the Wald, Agresti & Caffo, and WLS methods maintained appropriate nominal type I error levels, the RMLE method yields similar power as seen in Figure 3.9 for the Wald method, Figure 3.10 for the Agresti & Caffo method, and Figure 3.11 for the WLS method.

C. Sample Size Allocations

In three-arm non-inferiority trials, the sample size should be allocated according to economic feasibility, power related to the non-inferiority hypothesis, and ethical arguments when assigning patients to the placebo arm. These considerations may result in allocations other than a 1:1:1 balanced allocation for the three groups, usually with fewer patients randomized to the placebo group. These may include allocating sample size for test:active-control: placebo as 2:1:1, 2:2:1, 3:1:1, 3:2:1, and 3:3:1. Tang and Tang[4] suggest that the 3:2:1 allocation is most powerful out of the 1:1:1 and the 2:2:1 that they reviewed. The current simulations also show that power is higher when, using the same total sample size, more subjects are allocated to the test and the active control arms than the placebo arm. Figure 3.12 shows that the equal allocation scenario (1:1:1) yields the lowest power for the same overall sample size using the RMLE method.

V. Sample Size Formulas

Calculation of sample size for a specified level of power is an important aspect in designing three-arm trials to assess non-inferiority. Koch and Tangen[7] present a formula for the calculation of sample size as seen in (3.7) for n_T, where $n_C = C_C n_T$ and $n_P = C_P n_T$.

$$n_T = \frac{(z_\alpha + z_\beta)^2 \left\{ \pi_T(1-\pi_T) + \frac{\lambda_0^2 \pi_C (1-\pi_C)}{C_C} + \frac{(1-\lambda_0)^2 \pi_P (1-\pi_P)}{C_P} \right\}}{(\lambda - \lambda_0)^2 (\pi_C - \pi_p)^2} \tag{3.7}$$

This formula can also be solved to obtain the power for a specified sample size as seen in (3.8) where power is obtained as the probability $(1-\beta)$ from $z_{1-\beta}$ as the $(1-\beta)$ quantile of the standard normal distribution.

$$z_{1-\beta} = \frac{\sqrt{n_T}\,(\lambda-\lambda_0)(\pi_C-\pi_P)}{\left\{\pi_T(1-\pi_T)+\dfrac{\lambda_0^2\pi_C(1-\pi_C)}{C_C}+\dfrac{(1-\lambda_0)^2\pi_P(1-\pi_P)}{C_P}\right\}^{1/2}} - z_\alpha \qquad (3.8)$$

Another sample size formula can be used to design a three-arm trial based on the RMLE method. The calculation of n_T in (3.9) uses maximum likelihood estimates of the proportions restrained by the null hypothesis in its computation for $\tilde{\pi}_T$, $\tilde{\pi}_C$, and $\tilde{\pi}_P$. Again, these estimates can be obtained using PROC GENMOD in SAS. A specification of a binomial distribution and an identity link are specified and the model statement is fit as described previously. However, the events/trials syntax requires an observed number of events out of a sample size for each treatment group. These can be specified by assigning arbitrary n's for each treatment group, as long as the appropriate allocation is maintained. The number of events in each treatment group is simply the population proportion in the treatment group multiplied by this arbitrary n. The maximum likelihood estimates under the null hypothesis will be calculated for each treatment group, and then can be implemented in (3.9). These estimates will be the same, regardless of the arbitrary n chosen as long as the sample size allocation is maintained for the treatment groups.

An analogous sample size formula using (3.9) is based on the weighted least squares estimates of the proportions restrained by the null hypothesis for $\tilde{\pi}_T$, $\tilde{\pi}_C$, and $\tilde{\pi}_P$. These estimates can also be obtained using SAS, through PROC CATMOD and then implemented in (3.9), following the same process, as described above using PROC GENMOD, of using arbitrary n's to obtain the weighted least squares estimates under the null hypothesis.

$$n_T = \left\{ z_\alpha \sqrt{\tilde{\pi}_T(1-\tilde{\pi}_T) + \frac{\lambda_0^2 \tilde{\pi}_C(1-\tilde{\pi}_C)}{C_C} + \frac{(1-\lambda_0)^2 \tilde{\pi}_P(1-\tilde{\pi}_P)}{C_P}} + z_{1-\beta} \sqrt{\pi_T(1-\pi_T) + \frac{\lambda_0^2 \pi_C(1-\pi_C)}{C_C} + \frac{(1-\lambda_0)^2 \pi_P(1-\pi_P)}{C_P}} \right\}^2 \Big/ (\lambda-\lambda_0)^2(\pi_C-\pi_P)^2 \tag{3.9}$$

This formula can also be written to yield power as the probability (1-β) from a standard normal distribution for a specified sample size as seen in (3.10).

$$z_\beta = \frac{\sqrt{n_T}\left(\lambda-\lambda_0\right)\left(\pi_C-\pi_P\right) - z_\alpha \sqrt{\tilde{\pi}_T(1-\tilde{\pi}_T) + \frac{\lambda_0^2 \tilde{\pi}_C(1-\tilde{\pi}_C)}{C_C} + \frac{(1-\lambda_0)^2 \tilde{\pi}_P(1-\tilde{\pi}_P)}{C_P}}}{\sqrt{\pi_T(1-\pi_T) + \frac{\lambda_0^2 \pi_C(1-\pi_C)}{C_C} + \frac{(1-\lambda_0)^2 \pi_P(1-\pi_P)}{C_P}}} \tag{3.10}$$

The calculated power based on the methods previously described is compared with the simulated power for the methods and scenarios performed in the simulations. The simulated power and calculated power based on the RMLE method are fairly similar, with the simulated power being slightly higher in certain scenarios as displayed in Figure 3.13 – 3.18 by the parameters varied in the simulations.

Figure 3.19 compares the Wald simulated and calculated power for all scenarios and additionally only for those cases where the type I error is controlled at the nominal level in Figure 3.20. The Wald simulated power is slightly greater then the calculated power, but only for those cases where type I error is controlled at the nominal level.

Figure 3.21 summarizes the WLS calculated and simulated powers. These are similar, but the calculated power is slightly higher than the simulated power, even in cases where the nominal type I error is achieved as seen in Figure 3.22.

The RMLE calculated power is a good method to use when comparing it to the simulated power of Agresti & Caffo. The simulated power is slightly higher than this RMLE calculated power (Figure 3.23), especially in cases where the nominal type I error is maintained as seen in Figure 3.24.

Additionally, the RMLE sample size method is also appropriate for use with the Deviance method. Figure 3.25 shows that the simulated Deviance power is similar or slightly higher than the RMLE calculated power.

VI. Assessing Non-inferiority in a Three-arm Trial: 1 vs 2 Trials Paradigm

In a regulatory setting, it is often the standard to require two confirmatory trials for efficacy in order to obtain approval[8]. There are compelling reasons for this convention, however in many cases these two separate trials are run under very similar protocols and are run separately simply to adhere to this convention. Maca et. al.[9] discuss this scenario and include reasons in some circumstances why these two separate trials could be combined to yield a larger base of knowledge regarding the efficacy of the drug of interest. However, there may still be interest in ensuring that the two separate trials meet at least some minimum level of efficacy so that the combined data is not driven entirely by only one of the two trials.

In the two trials setting, a one-sided p-value of 0.025 (generally) would be required for each of the two trials separately. This would result in an overall alpha level of 0.000625 for the combined project (both trials together). Maintaining this overall alpha level at 0.000625 can also be easily done by simply combining the two trials and analyzing this combined data at an alpha level of 0.000625. Maca et. al.[9] show that the overall project alpha level can also be maintained at the 0.000625 level by requiring each single trial to meet a criteria, say 0.10, and requiring the alpha level for the combined data to meet a 0.0007005 significance level. As the individual alpha level for the studies increases, the combined alpha level decreases in order to maintain this project level alpha at 0.000625. Other selected cases are summarized in Table 3.2.

The benefits of implementing these modified alpha levels are evident in discussions related to the overall power of the project. In scenario 1, using the simple two separate trials approach, if each trial is designed with 80% power then the overall project power is 64%, by relying on the independence of the trials and therefore multiplying the powers together. Maca et. al.[9] have shown that Scenario 2, which only makes a requirement on the combined data, yields a much higher power than this. However, to ensure that each separate trial is also providing adequate signals for efficacy, implementing Scenarios 3, 4, or 5 yields lower but similar power to Scenario 2 while resulting in a much higher power than the traditional separate trials in Scenario 1.

The performance of the methods under current discussion for the non-inferiority hypothesis in a three-arm trial is of interest for the five scenarios discussed by Maca et. al. Simulations were designed in an identical manner to those used in Section III for the one-trial scenario for assessing non-inferiority with 10,000 replications. The sample size was also calculated in a similar fashion and then split in half for each of the separate trials so that total overall sample sizes remain the same. The type I error and power will be discussed for each of the five scenarios summarized above, with methods for assessing non-inferiority including the Wald method, the Agresti & Caffo method, and the RMLE method.

The simulated type I errors for these methods are summarized by sample size allocation in Figures 3.26 – 3.28, by non-inferiority margin in Figures 3.29 – 3.31, by the population event rate in the control group in Figures 3.32 – 3.34, and by the population event rate in the placebo group in Figures 3.35 – 3.37. These figures have bands around the nominal 0.000625 level which show the precision of the 10,000 simulations as approximately ±0.0005. Each of the five scenarios have fairly similar type I error rates, within each method or parameter of interest. The type I error rates are dependent on the non-inferiority margin λ_0, π_C, and π_P as was the case in the simulations discussed previously for a single three-arm non-inferiority trial. However, the RMLE method has type I error levels closer to the nominal level for all settings.

The more obvious distinctions in the five scenarios relate to the power of the test for non-inferiority. In all situations, the power for Scenario 1 is much lower than that for Scenarios 2-5. The highest power is seen in Scenario 2 for the simple combined analysis with only an overall alpha level of 0.000625 specified and then drops slightly for Scenarios 3, 4, and 5 as shown for each of the methods in Figures 3.38 – 3.40 by the sample size allocation, in Figures 3.41 – 3.43 by the non-inferiority margin, in Figures 3.44 – 3.46 by the population event rate in the control group, and in Figures 3.47 – 3.49 by the population event rate in the placebo group. The power for the RMLE method is lower than that for the Wald and Agresti & Caffo methods, but with the benefit of more closely maintaining type I error at the nominal level.

The results of these simulations affirm the discussion by Maca et. al. and further confirm the use of these scenarios in this specific application for three-arm non-inferiority trials. The implications of choosing to design a trial using scenarios 3, 4, or 5 are a higher project power using the same sample size, and therefore a potential reduction in total number of subjects and cost in implementing the trial.

VII. Performance of Methods based on Simulations for Assessing Dual Endpoints of Superiority and Non-inferiority

In addition to assessing the non-inferiority hypothesis of the percentage of effect that the test treatment preserves compared to the active control treatment over placebo, regulatory agencies may also require proof that the test treatment is superior to the placebo treatment. In most settings, assessment of superiority of test over placebo is performed in the first step of the analyses. If superiority is shown, then analysis proceeds to the non-inferiority assessment of the test treatment compared to active control treatment[1]. If superiority is not shown in the first step of the analyses, then testing ends and does not proceed to the non-inferiority hypothesis.

In a simple setting with one trial, superiority can be assessed at a specified alpha level and if significant, then testing can proceed to the non-inferiority hypothesis at this same alpha level. This approach controls the type I error for multiple testing of the superiority and non-inferiority hypotheses through the use of hierarchical testing. Testing of the superiority hypothesis has little effect on type I error or power because the sample size required for the non-inferiority hypothesis makes the power for the superiority hypothesis very large and close to 1 as evidenced in Table 3.3 where the sample size needed for superiority at α^2 is much smaller than the sample size needed for non-inferiority at α.

Although the setting described above requires confirmation of both superiority of the test treatment to placebo and non-inferiority of the test treatment to the active control (relative to placebo), regulatory agencies may desire a stronger degree of comfort surrounding the superiority hypothesis. This is especially the case because the superiority hypothesis is greatly overpowered in a trial designed to show non-inferiority as previously discussed.

The results obtained by Maca et. al.[9] can be used to place more stringent requirements on the superiority hypothesis. In this setting, the overall trial can be divided into two smaller trials for only the test of superiority. The five different scenarios from Maca et. al. seen in Table 3.2 are used for the test of superiority. If this is significant, then testing proceeds to non-inferiority on the combined trials at the usual 0.025 alpha level.

The simulations were again implemented for similar scenarios as described in Section III with 10,000 replications. The F-M 3[19] method described in Chapter 2 for the difference in proportions was used to assess the superiority hypothesis and the RMLE method was used to assess the non-inferiority hypothesis. The more stringent requirements placed on the alpha level when implementing the Maca et. al. scenarios for the superiority test do not seem to change the results because, again, even reducing the alpha level still makes the superiority test sufficiently powered for the sample sizes required for the non-inferiority hypothesis. Simulated type I error is summarized in Figures 3.50 – 3.53 where the observed percentage of effect is the specified non-inferiority margin for the test of non-inferiority but with superiority maintained for the test treatment over the placebo treatment. Simulated power is summarized in Figures 3.54 – 3.57 by the sample size allocation, non-inferiority margin, and the population event rates in the control and placebo groups.

Sample size calculations are provided in Table 3.3 for the non-inferiority and superiority hypotheses. The non-inferiority calculations are provided for the Wald, F-M 3, and WLS sample size methods. The superiority calculation for test versus placebo is provided using the Wald sample size method for the difference in proportions. The sample size needed for the test of superiority at α^2 is, in most cases, less than half the sample size needed for the non-inferiority test at α.

References

1. Koch A., and Röhmel J. Hypothesis Testing in the “Gold Standard” Design for Proving the Efficacy of an Experimental Treatment Relative to Placebo and a Reference. Journal of Biopharmaceutical Statistics **2004**, 14(2): 315-325.

2. Pigeot I., Schafer J., Röhmel J., and Hauschke D. Assessing Non-inferiority of a New Treatment in a Three-arm Clinical Trial Including a Placebo. Statistics in Medicine **2003**, 22: 883-899.

3. Schwartz T. A Two-stage Sample Size Recalculation Procedure for Placebo- and Active-controlled Non-inferiority Trials, Chapter II. *DrPH Dissertation at Univeristy of North Carolina at Chapel Hill*. **2003**.

4. Tang M. L., and Tang N. S. Tests of Noninferiority via Rate Difference for Three-arm Clinical Trials with Placebo. Journal of Biopharmaceutical Statistics **2004**, 14: 337-347.

5. Agresti A., and Caffo B. Simple and Effective Confidence Intervals for Proportions and Difference of Proportions Results from Adding Two Successes and Two Failures. The American Statistician **2000**, 54(4): 280-288.

6. SAS®, Version 8.02. SAS Institute Inc.: Cary, NC, 1999.

7. Koch G., and Tangen C. Nonparametric Analysis of Covariance and its Role in Noninferiority Clinical Trials. Drug Information Journal **1999**, 33(4): 1145-1159.

8. U.S. Food and Drug Administration, U.S. Department of Health and Human Services, Center for Drug Evaluation and Research (CDER). Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products, **1998**.

9. Maca J., Gallo P., Branson M., and Maurer W. Reconsidering Some Aspects of the Two-Trials Paradigm. Journal of Biopharmaceutical Statistics **2002**, 12(2): 107-119.

10. Farrington, C. P., and Manning G. Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-zero Risk Difference or Non-unity Relative Risk. Statistics in Medicine **1990**, 9: 1447-1454.

Table 3.1 Summary of Simulation and Sample Size Calculation Results

Method	**Type I Error Violations**	**Power Considerations***	**Sample Size Calculations**
Wald	- Sample size allocations T:C = 1:1, 3:2 - π_C=0.8, 0.9 - π_P=0.2, 0.3	RMLE has higher power	Wald
Agresti & Caffo	- π_C=0.8, 0.9 - π_P=0.2, 0.3	RMLE has similar power	RMLE
RMLE		Good power	RMLE
Deviance		Good power	RMLE
Pearson		Good power	RMLE
WLS	- Sample size allocations T:C = 2:1, 3:1 - π_C=0.8, 0.9 - π_P=0.2, 0.3	RMLE has similar power	WLS

*For scenarios where type I error is appropriately controlled at the nominal level

Table 3.2 Summary of Scenarios which maintain Project Level α=0.0006250

from Maca et. al.[9]

Scenario	Separate Trials α	Combined Trials α
1	0.025	None
2	None	0.0006250
3	0.15	0.0006574
4	0.10	0.0007005
5	0.05	0.0008905

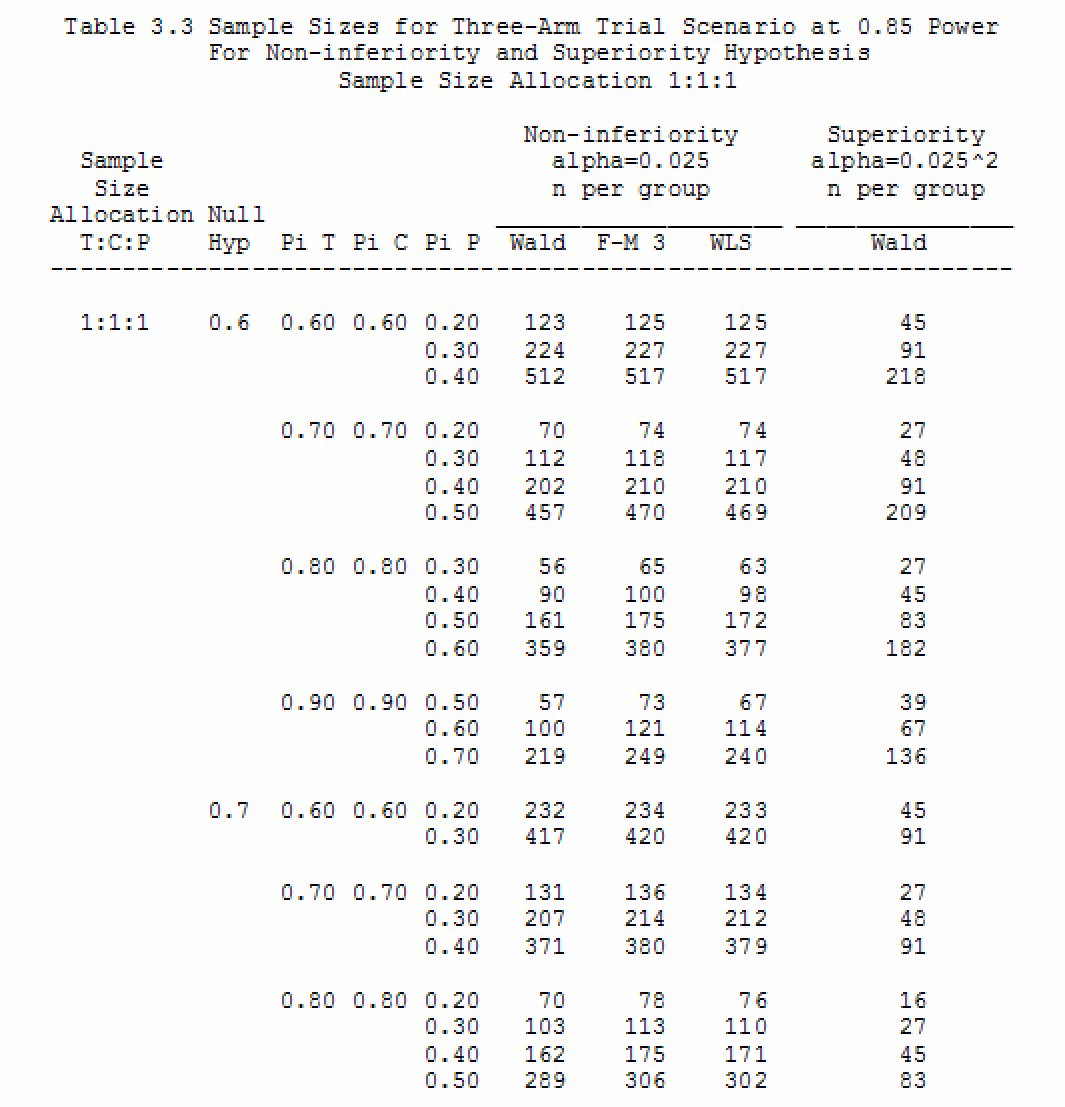


Table 3.3 Sample Sizes for Three-Arm Trial Scenario at 0.85 Power
For Non-inferiority and Superiority Hypothesis
Sample Size Allocation 1:1:1

Sample Size Allocation	Null				Non-inferiority alpha=0.025 n per group			Superiority alpha=0.025^2 n per group
T:C:P	Hyp	Pi T	Pi C	Pi P	Wald	F-M 3	WLS	Wald
1:1:1	0.6	0.60	0.60	0.20	123	125	125	45
				0.30	224	227	227	91
				0.40	512	517	517	218
		0.70	0.70	0.20	70	74	74	27
				0.30	112	118	117	48
				0.40	202	210	210	91
				0.50	457	470	469	209
		0.80	0.80	0.30	56	65	63	27
				0.40	90	100	98	45
				0.50	161	175	172	83
				0.60	359	380	377	182
		0.90	0.90	0.50	57	73	67	39
				0.60	100	121	114	67
				0.70	219	249	240	136
	0.7	0.60	0.60	0.20	232	234	233	45
				0.30	417	420	420	91
		0.70	0.70	0.20	131	136	134	27
				0.30	207	214	212	48
				0.40	371	380	379	91
		0.80	0.80	0.20	70	78	76	16
				0.30	103	113	110	27
				0.40	162	175	171	45
				0.50	289	306	302	83

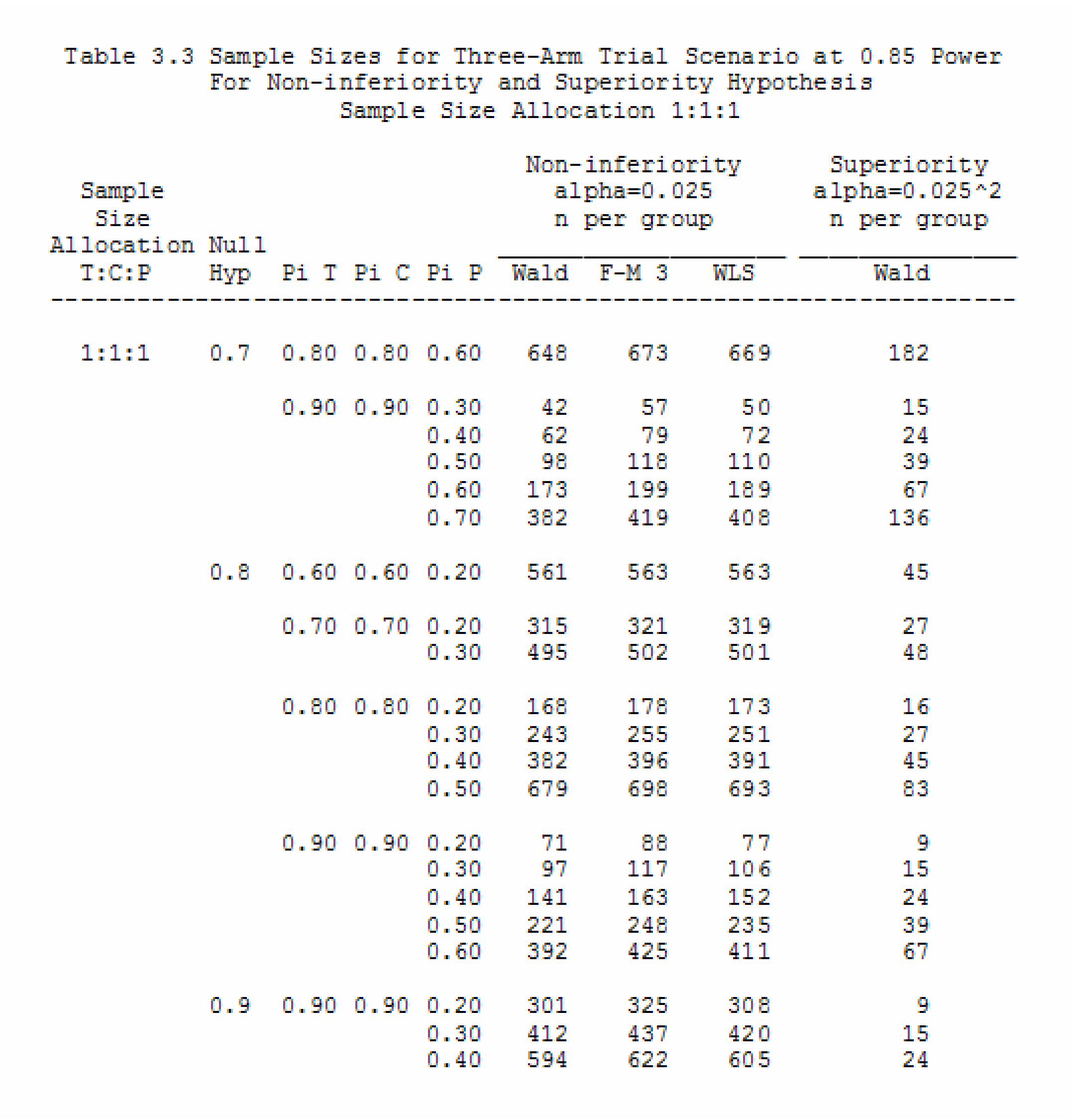


Table 3.3 Sample Sizes for Three-Arm Trial Scenario at 0.85 Power
For Non-inferiority and Superiority Hypothesis
Sample Size Allocation 1:1:1

Sample Size Allocation T:C:P	Null Hyp	Pi T	Pi C	Pi P	Non-inferiority alpha=0.025 n per group Wald	F-M 3	WLS	Superiority alpha=0.025^2 n per group Wald
1:1:1	0.7	0.80	0.80	0.60	648	673	669	182
		0.90	0.90	0.30	42	57	50	15
				0.40	62	79	72	24
				0.50	98	118	110	39
				0.60	173	199	189	67
				0.70	382	419	408	136
	0.8	0.60	0.60	0.20	561	563	563	45
		0.70	0.70	0.20	315	321	319	27
				0.30	495	502	501	48
		0.80	0.80	0.20	168	178	173	16
				0.30	243	255	251	27
				0.40	382	396	391	45
				0.50	679	698	693	83
		0.90	0.90	0.20	71	88	77	9
				0.30	97	117	106	15
				0.40	141	163	152	24
				0.50	221	248	235	39
				0.60	392	425	411	67
	0.9	0.90	0.90	0.20	301	325	308	9
				0.30	412	437	420	15
				0.40	594	622	605	24

Figure 3.1 Summary of Simulated Type I Error
By Sample Size Allocation
Alpha=0.025

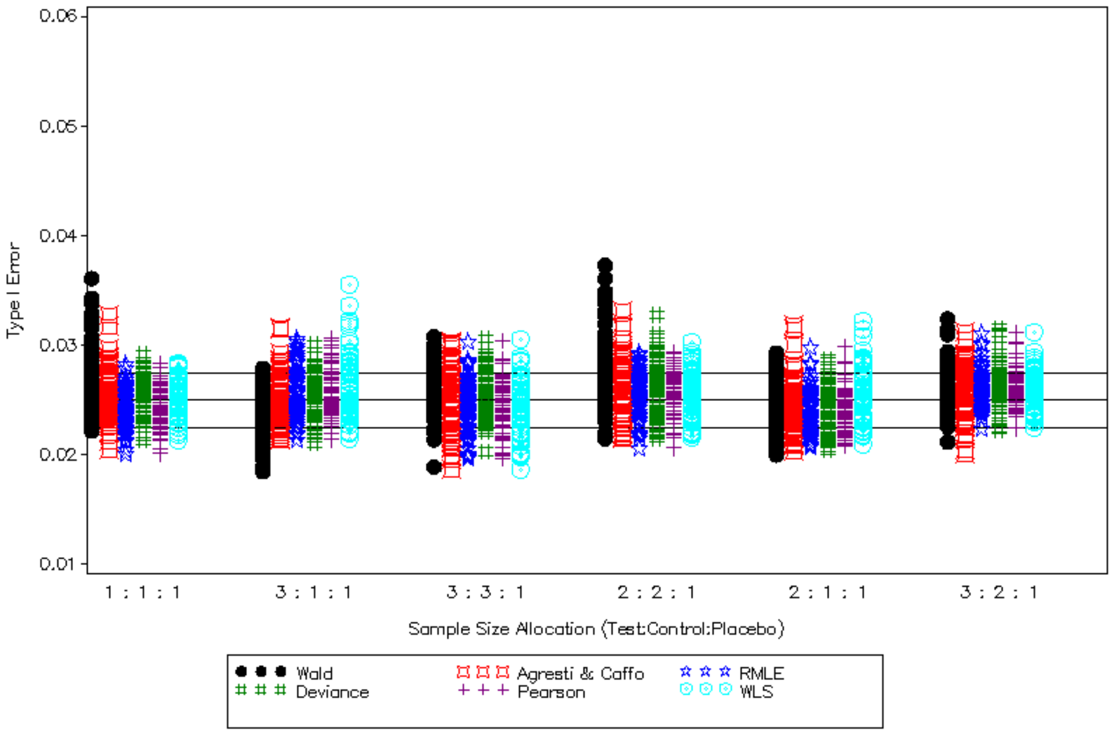


Figure 3.2 Summary of Simulated Type I Error
By Non-inferiority Margin
Alpha=0.025

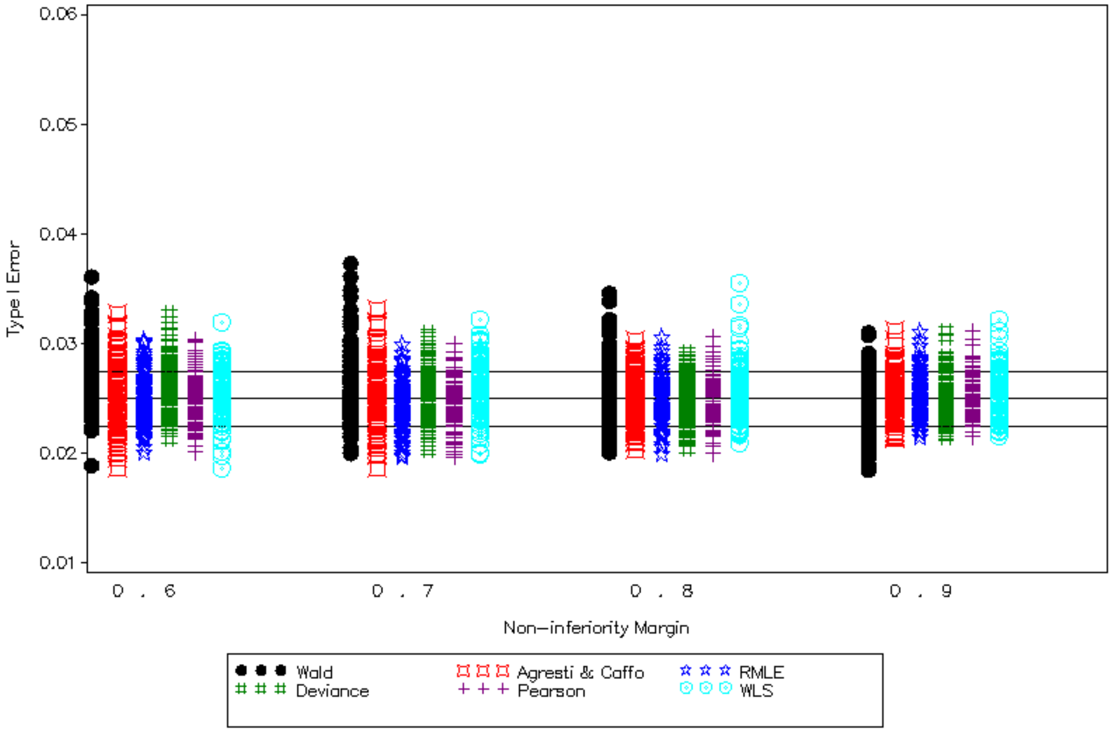


Figure 3.3 Summary of Simulated Type I Error
By Pi C
Alpha=0.025

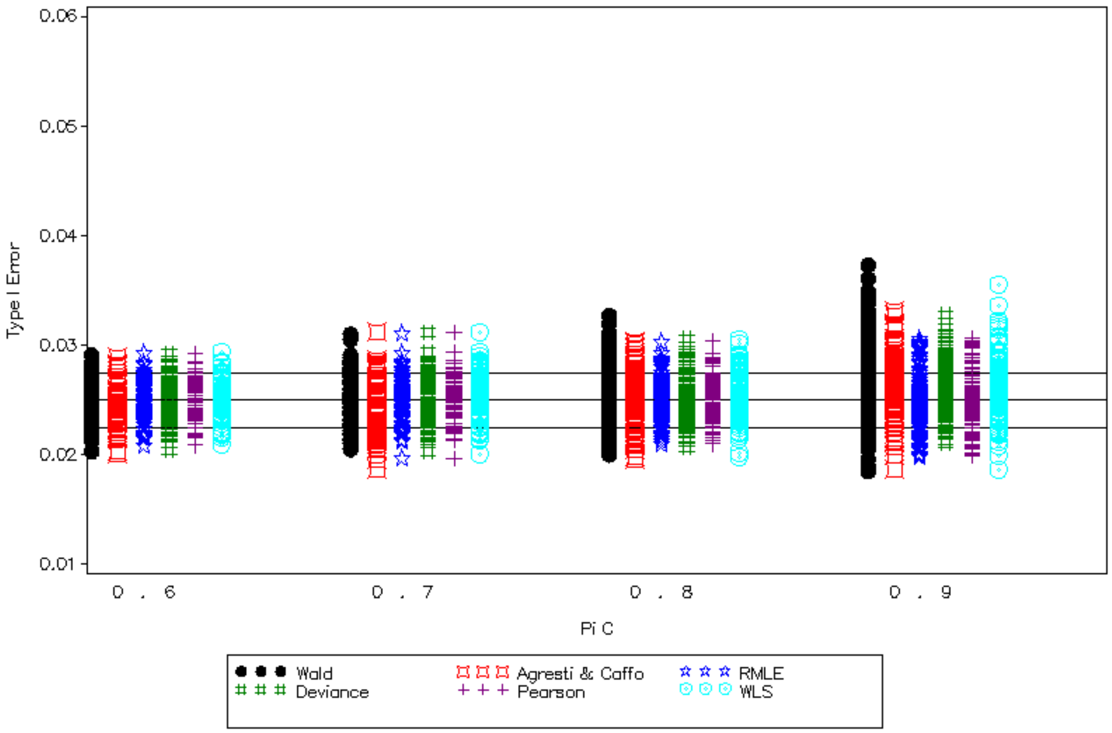


Figure 3.4 Summary of Simulated Type I Error
By Pi P
Alpha=0.025

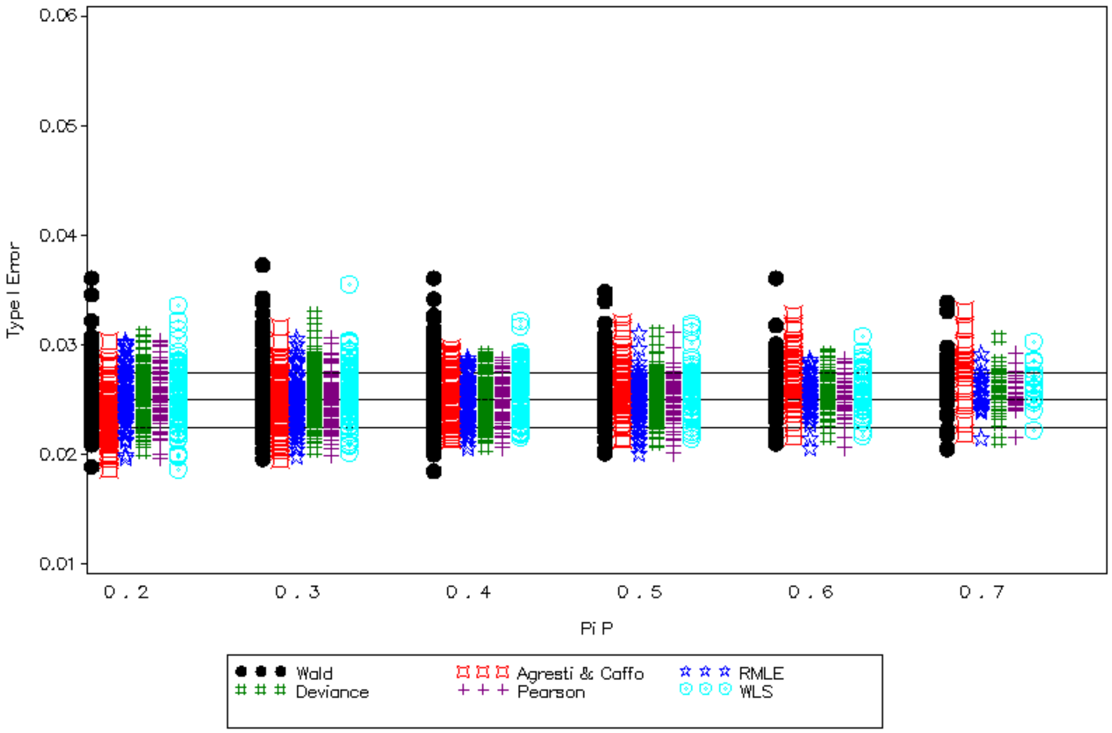


Figure 3.5 Summary of Simulated Type I Error
By (Pi C – Pi P)
Alpha=0.025

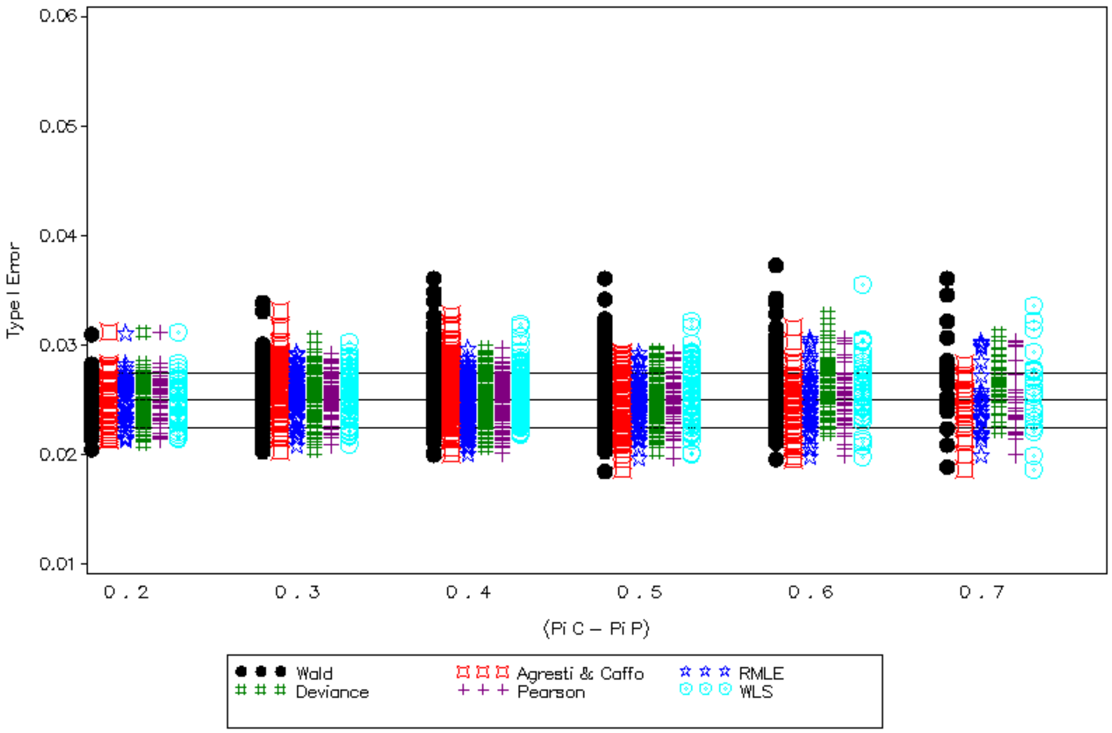


Figure 3.6 Summary of Type I Error
By Total Sample Size
Alpha=0.025

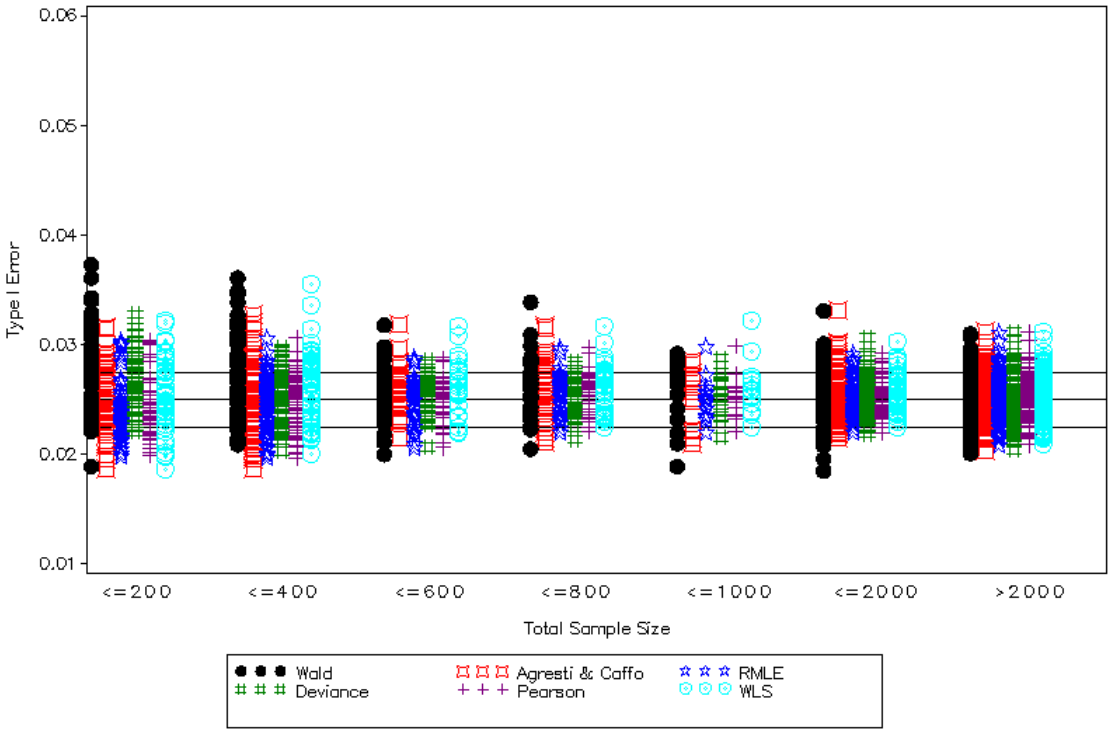


Figure 3.7 Comparison of Simulated Power

RMLE & Pearson Simulated Power

By Sample Size Allocation

Alpha=0.025, Lamda=1

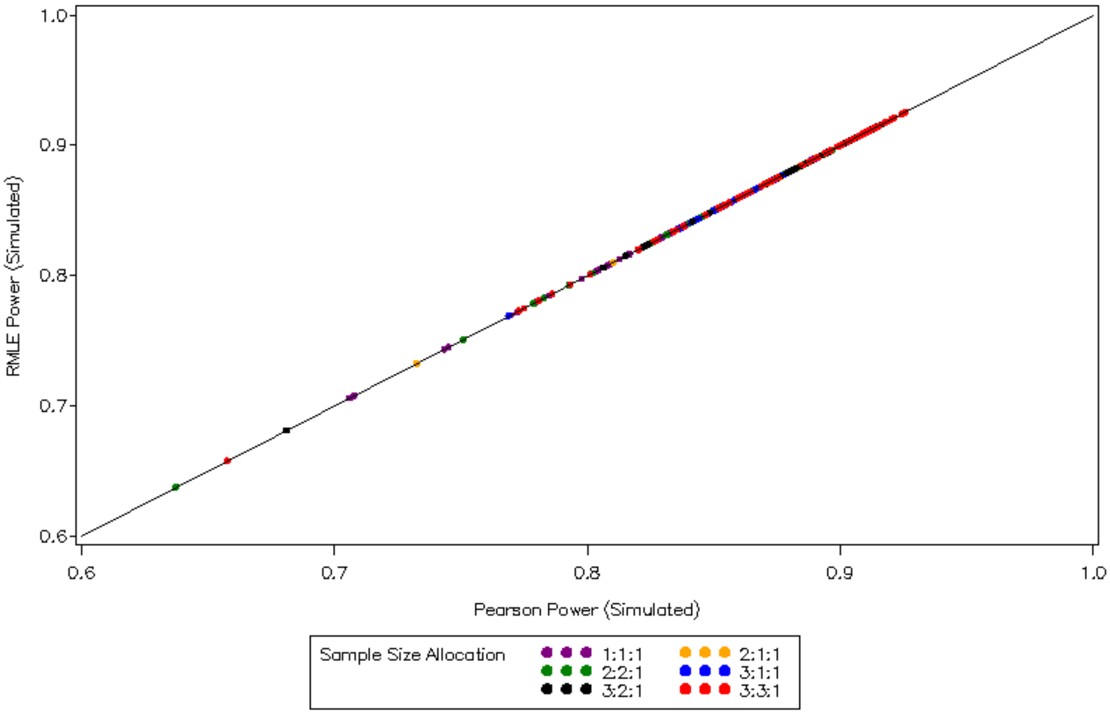


Figure 3.8 Comparison of Simulated Power

RMLE & Deviance Simulated Power

By Pi C

Alpha=0.025, Lamda=1

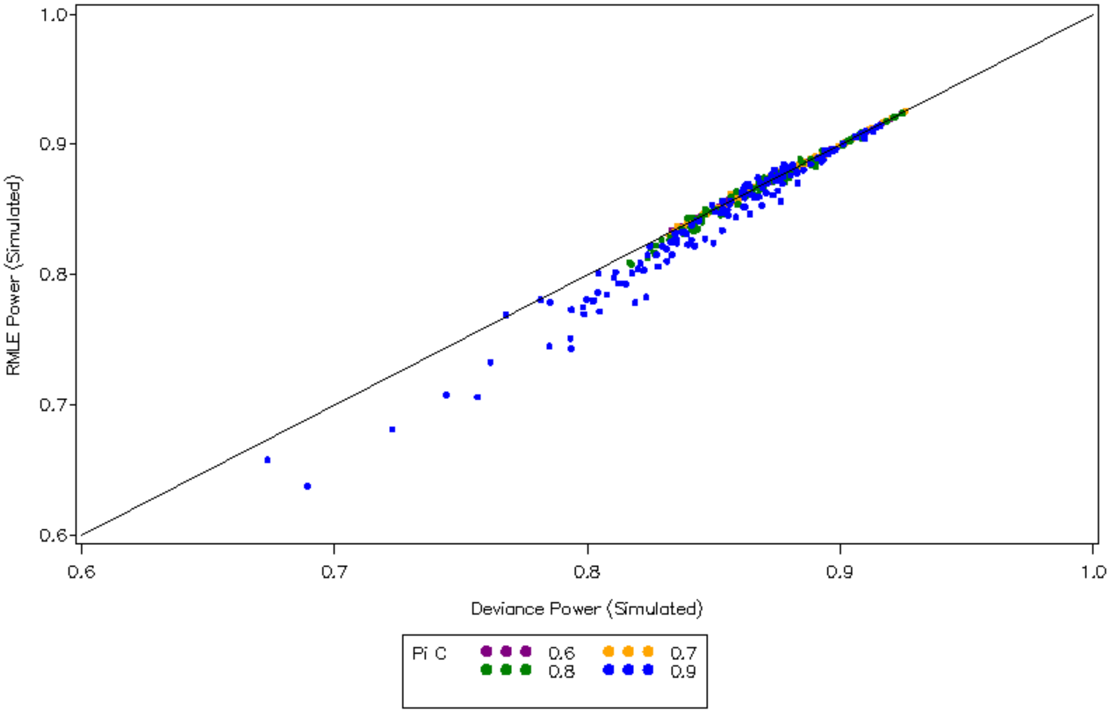


Figure 3.9 Comparison of Simulated Power

RMLE & Wald Simulated Power

By Non-inferiority Margin

Alpha=0.025, Lamda=1, Pi C<=0.7, Pi P>=0.4, Allocation=2:1:1,3:1:1

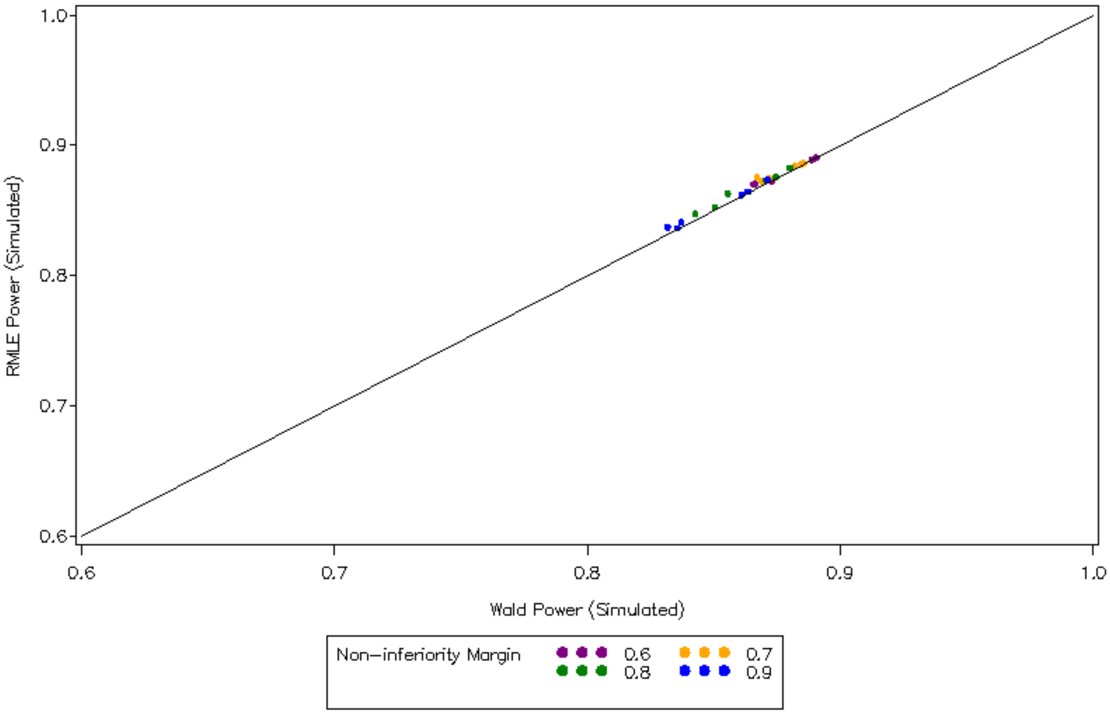


Figure 3.10 Comparison of Simulated Power

RMLE & Agresti & Caffo Simulated Power

By Non-inferiority Margin

Alpha=0.025, Lamda=1, Pi C<=0.7, Pi P>=0.4

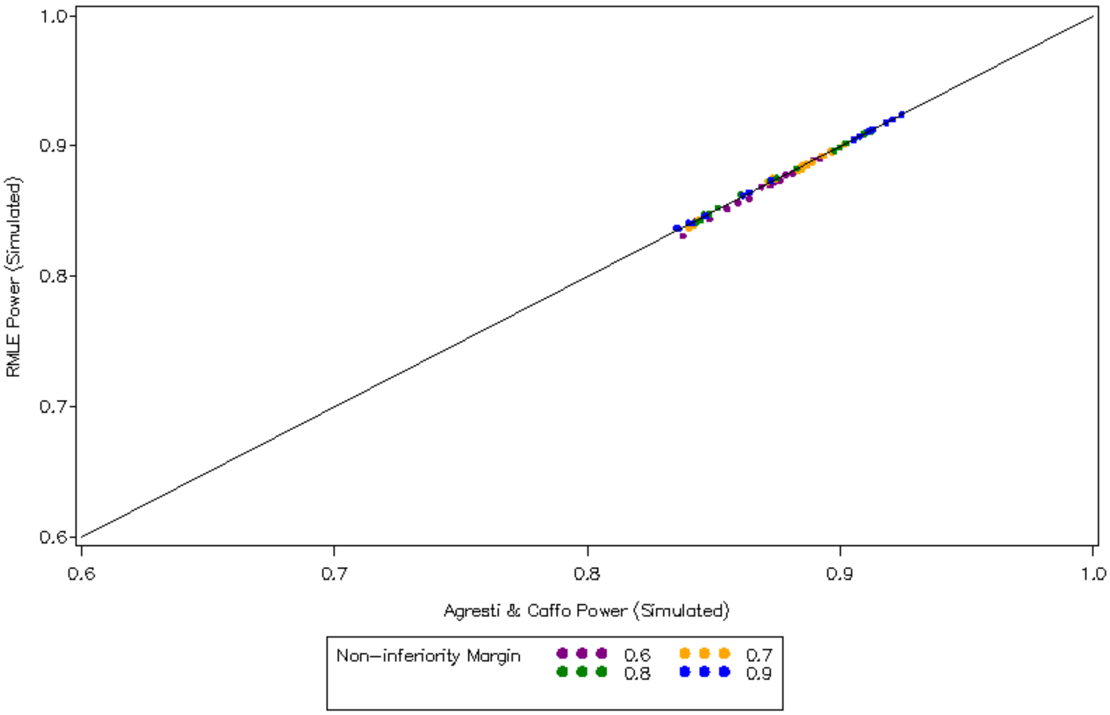


Figure 3.11 Comparison of Simulated Power

RMLE & WLS Simulated Power

By Non-inferiority Margin

Alpha=0.025, Lamda=1, Pi C<=0.7, Pi P>=0.4, Allocation=1:1:1,3:3:1,2:2:1,3:2:1

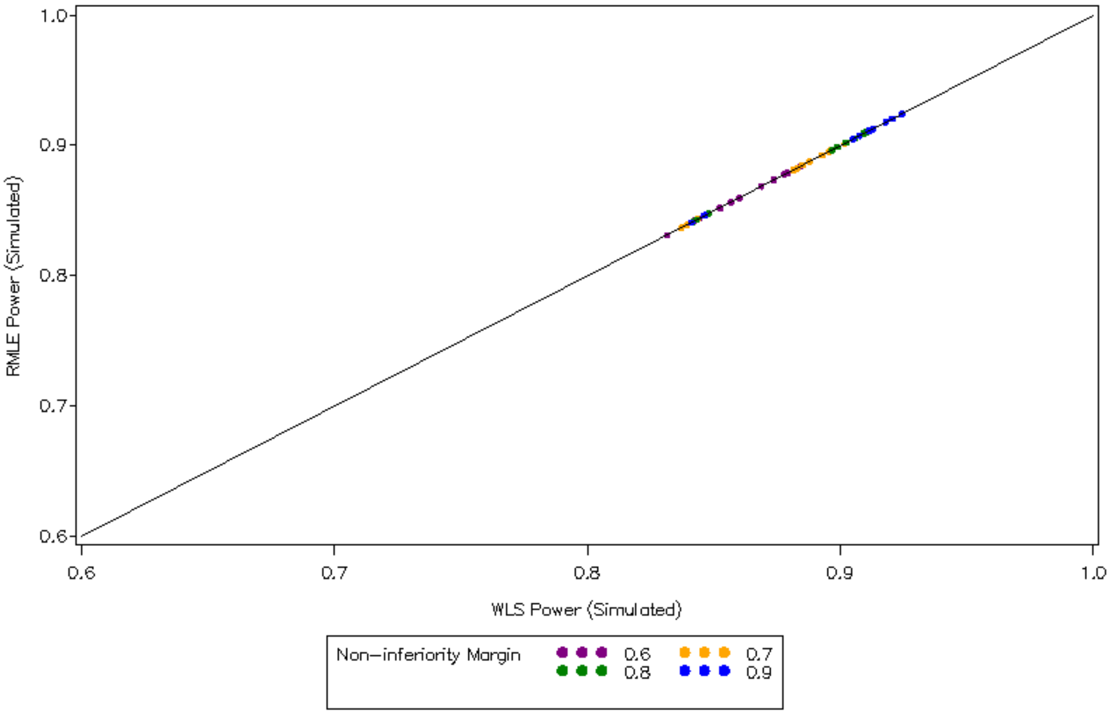


Figure 3.12 Summary of Simulated Power – RMLE Method

By Total Sample Size & Sample Size Allocation

Alpha=0.025, Lambda=1

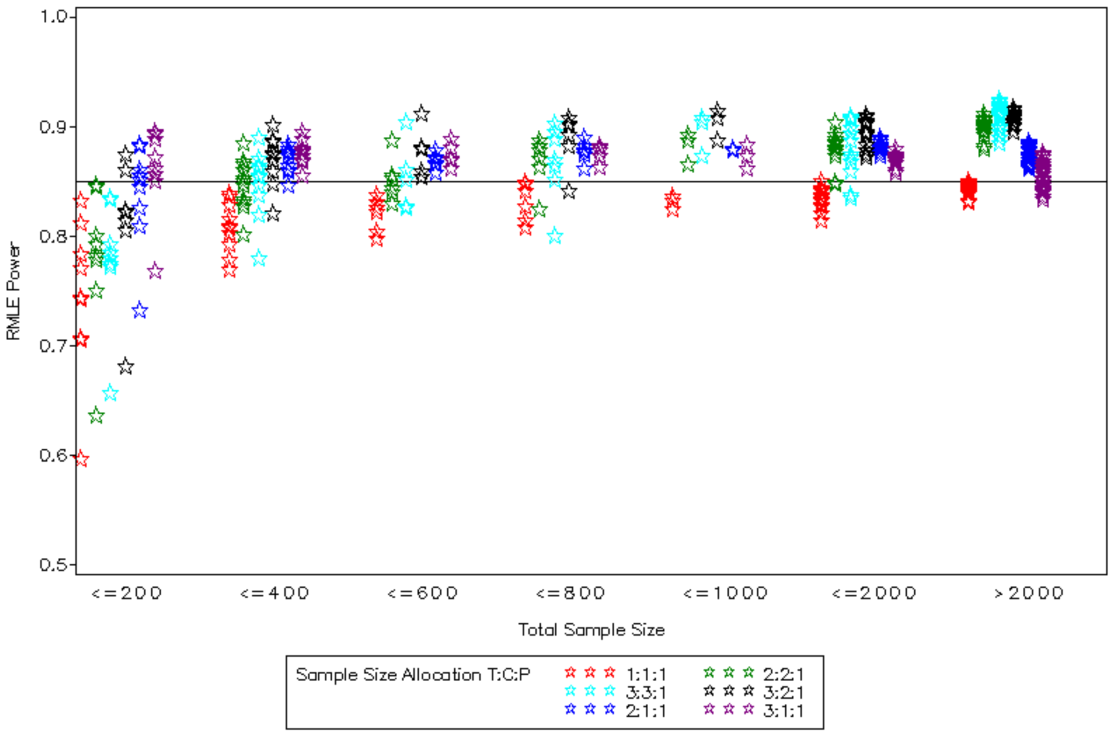


Figure 3.13 Comparison of Simulated and Calculated Power

RMLE Method
By Sample Size Allocation
Alpha=0.025, Lambda=1

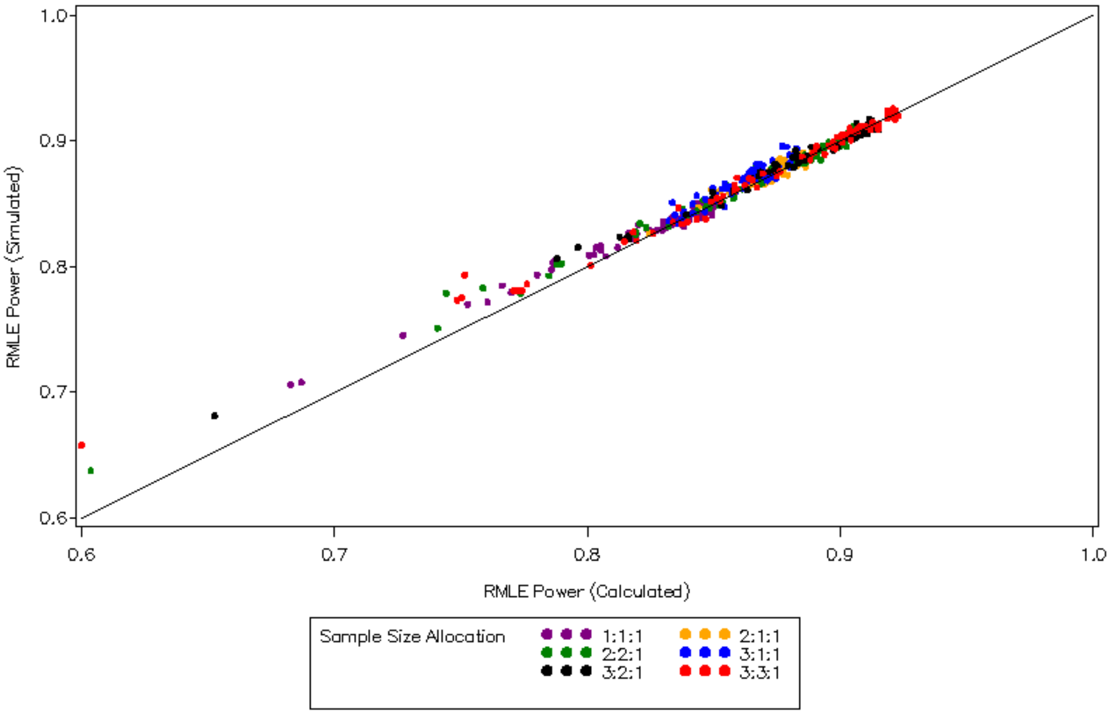


Figure 3.14 Comparison of Simulated and Calculated Power

RMLE Method
By Non-inferiority Margin
Alpha=0.025, Lambda=1

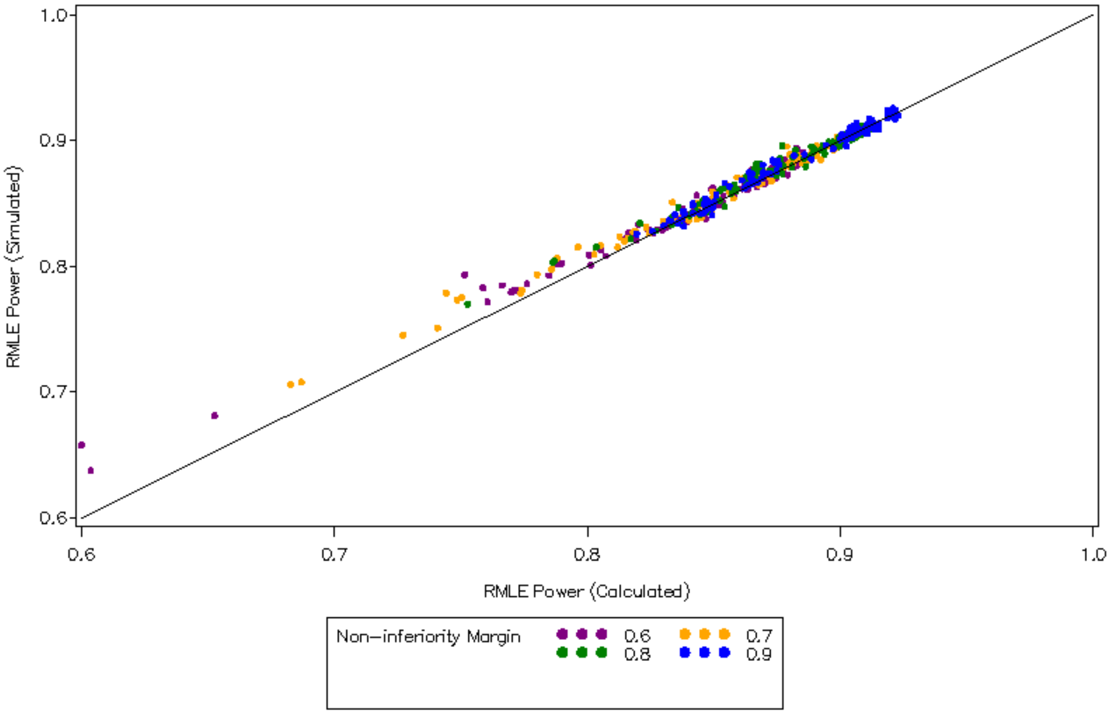


Figure 3.15 Comparison of Simulated and Calculated Power

RMLE Method
By Pi C
Alpha=0.025, Lambda=1

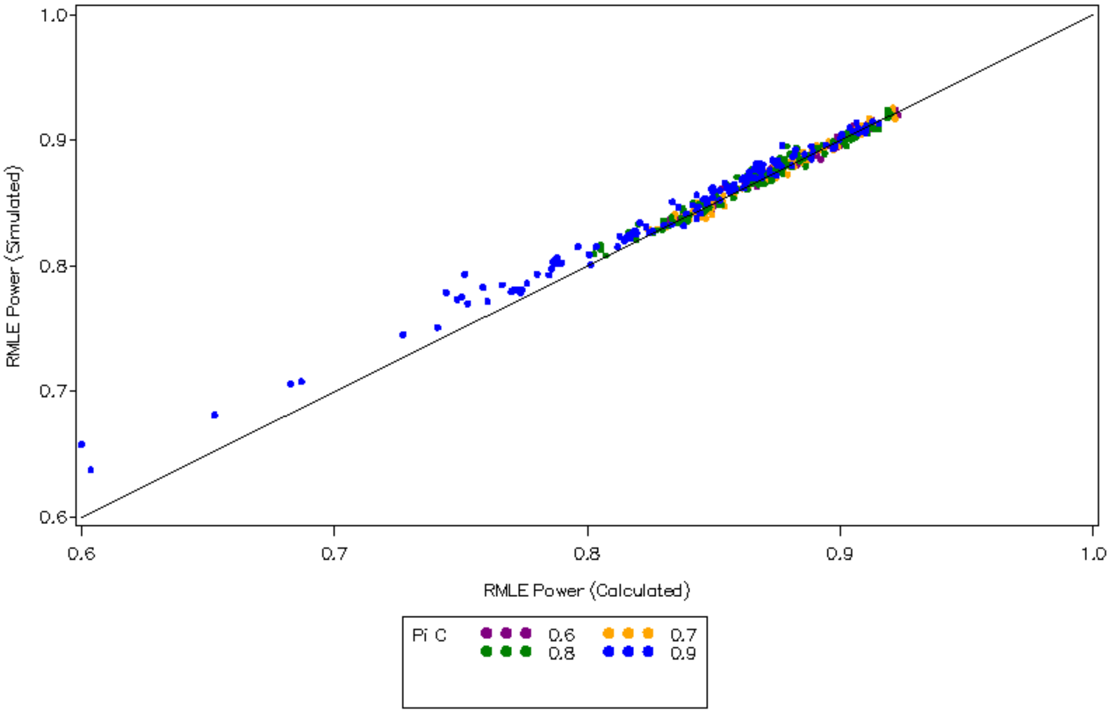


Figure 3.16 Comparison of Simulated and Calculated Power

RMLE Method
By Pi P
Alpha=0.025, Lambda=1

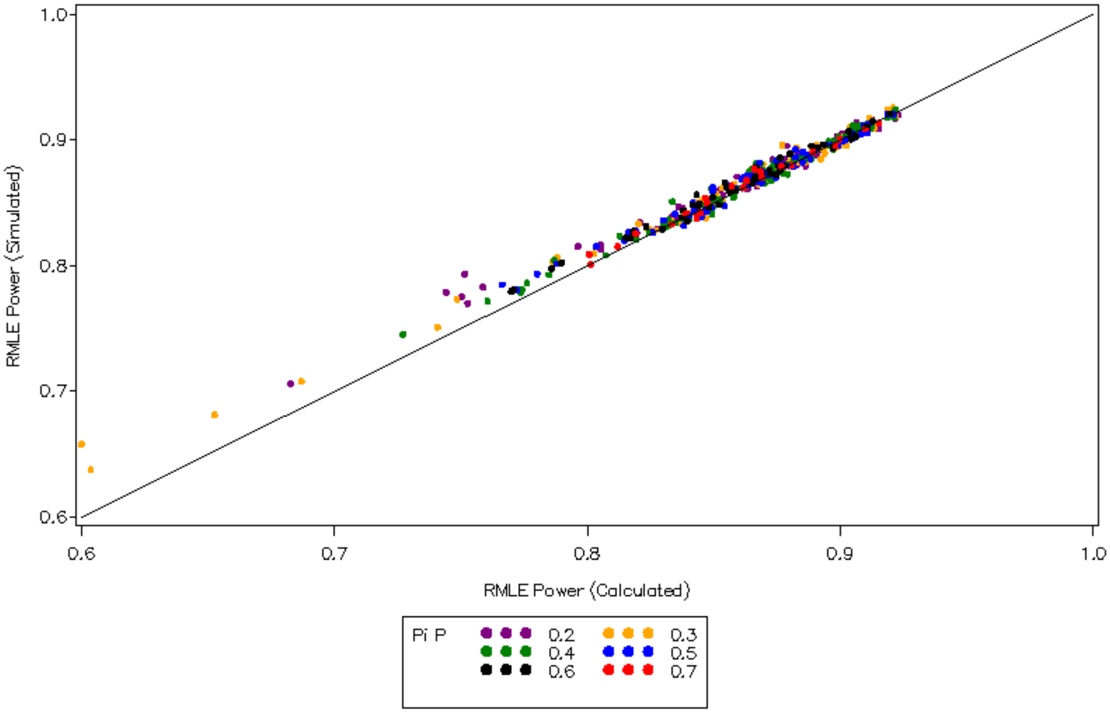


Figure 3.17 Comparison of Simulated and Calculated Power
RMLE Method
By (Pi C – Pi P)
Alpha=0.025, Lambda=1

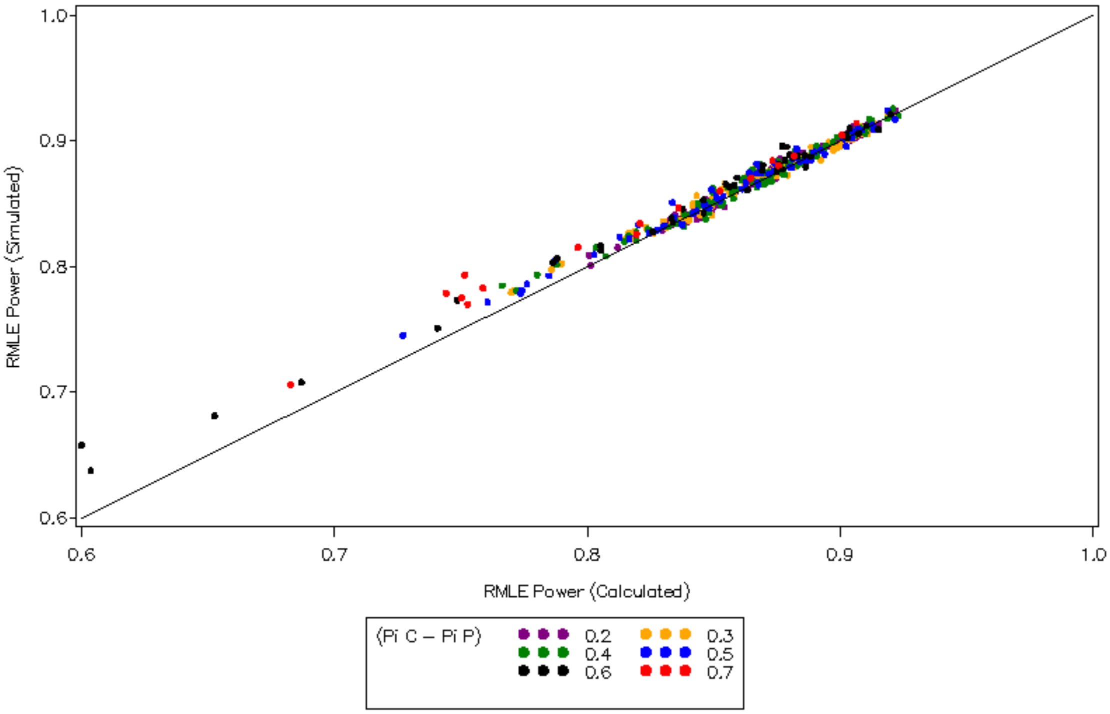


Figure 3.18 Comparison of Simulated and Calculated Power
RMLE Method
By Total Sample Size
Alpha=0.025, Lambda=1

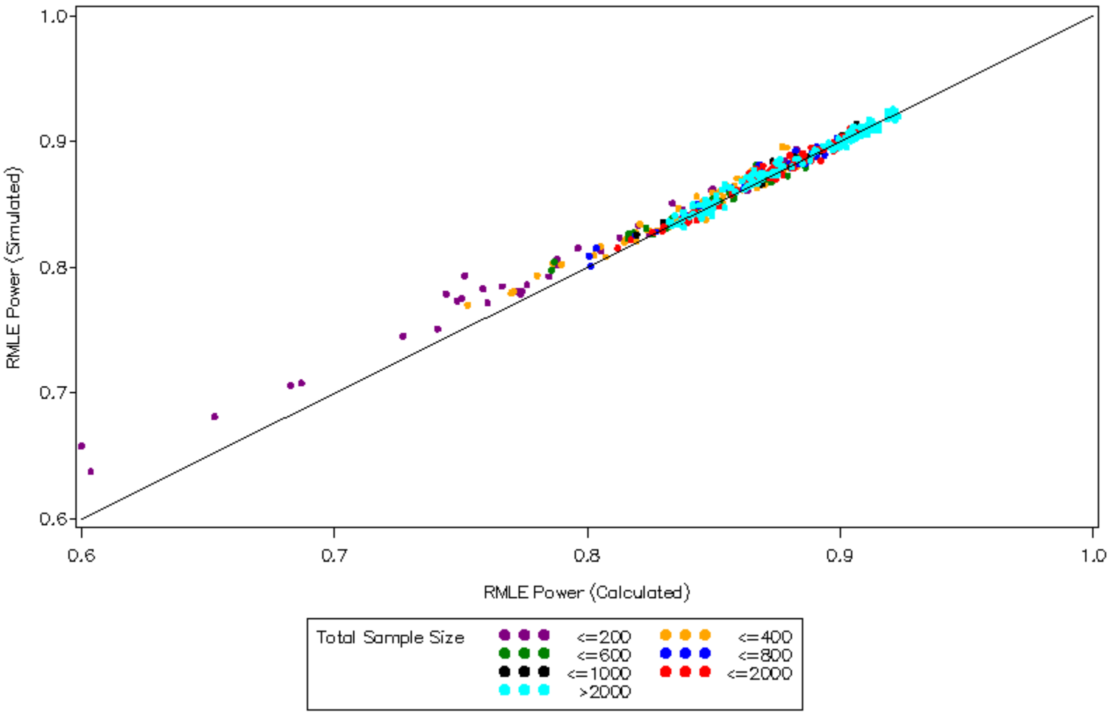


Figure 3.19 Comparison of Simulated and Calculated Power

Wald Method
By Sample Size Allocation
Alpha=0.025, Lambda=1

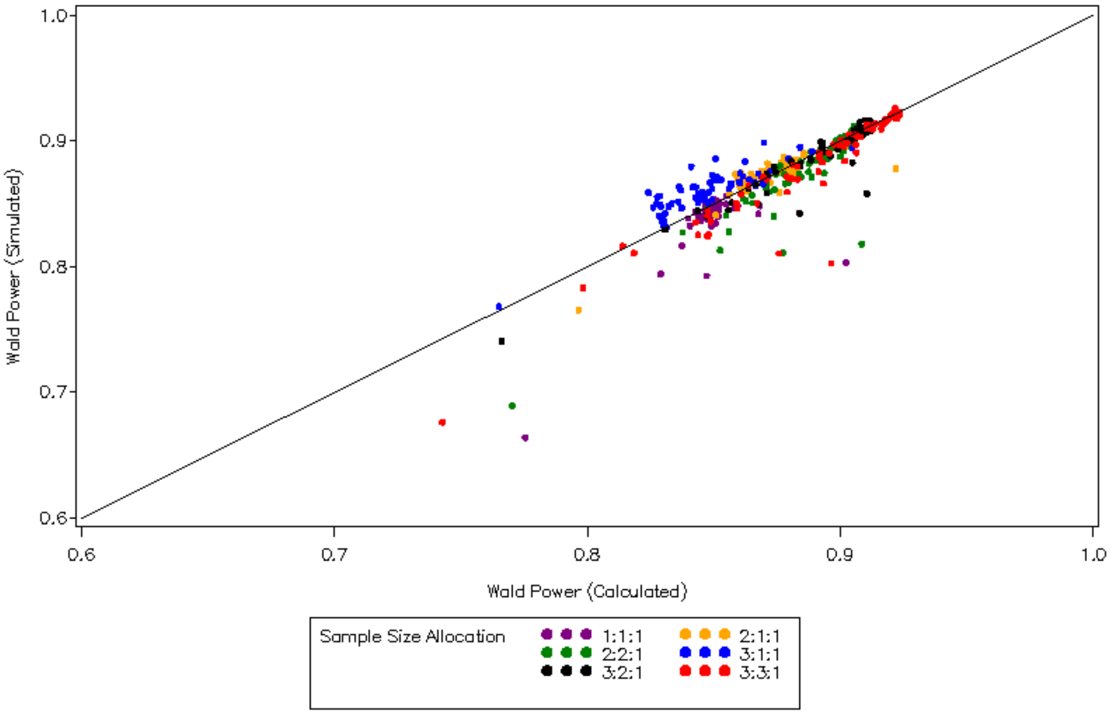


Figure 3.20 Comparison of Simulated and Calculated Power

Wald Method
By Sample Size Allocation
Alpha=0.025, Lambda=1, Pi C<=0.7, Pi P>=0.4, Allocation=2:1:1,3:1:1

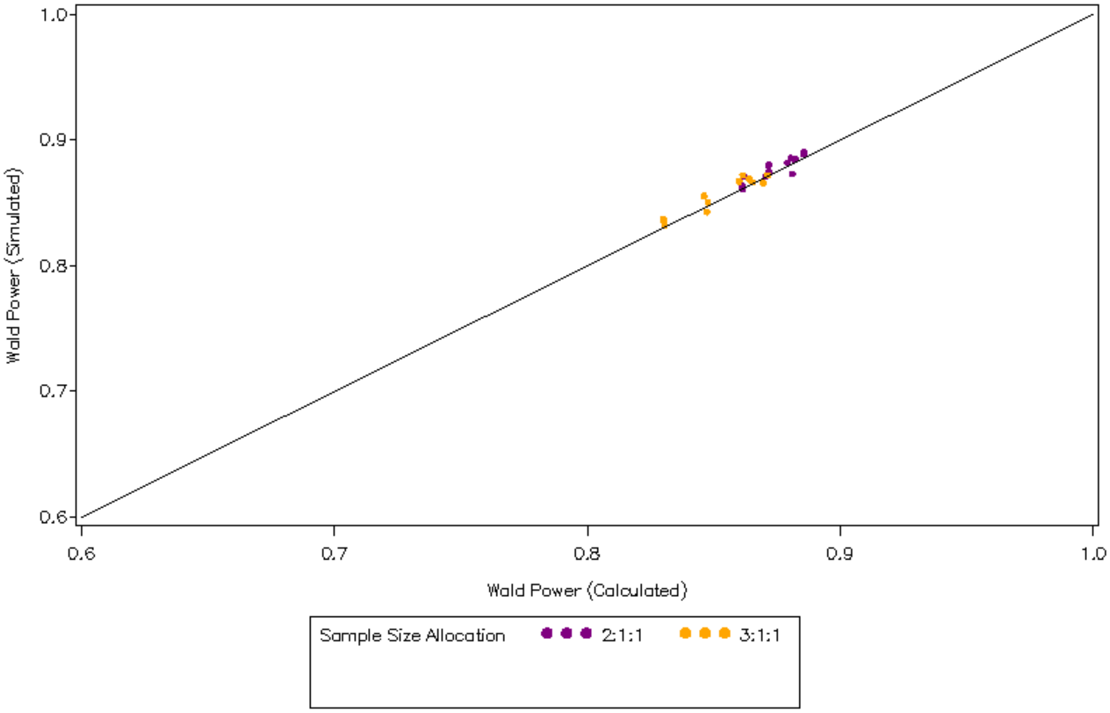


Figure 3.21 Comparison of Simulated and Calculated Power

WLS Method

By Sample Size Allocation

Alpha=0.025, Lambda=1

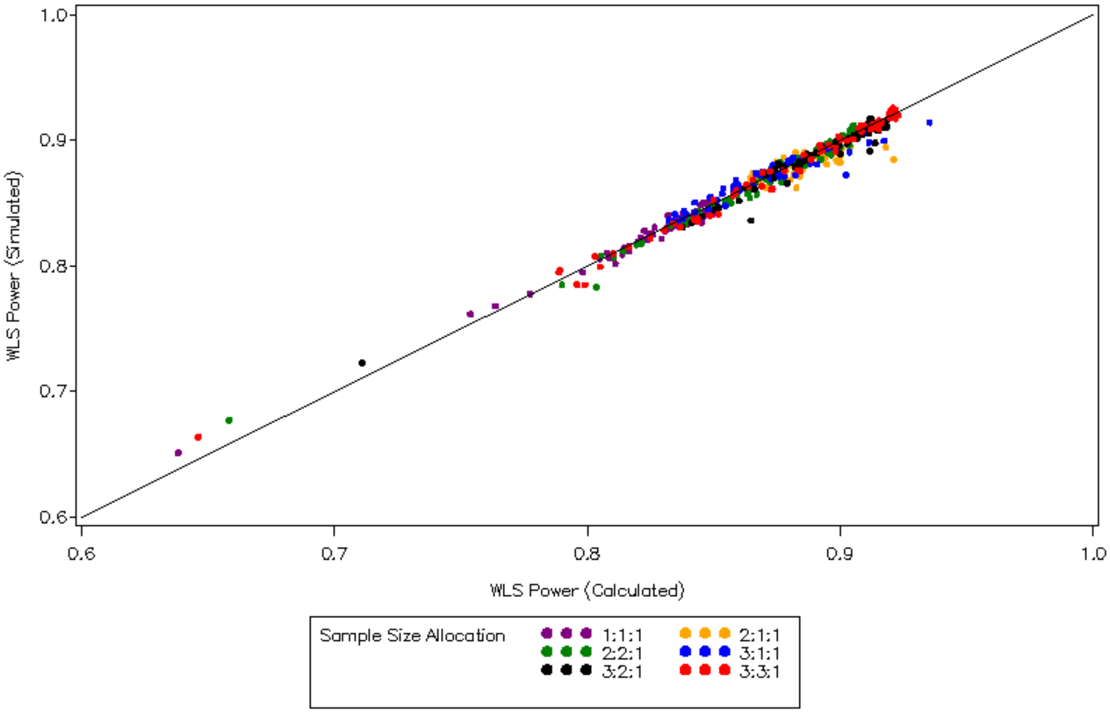


Figure 3.22 Comparison of Simulated and Calculated Power

WLS Method

By Sample Size Allocation

Alpha=0.025, Lambda=1, Pi C<=0.7, Pi P>=0.4, Allocation=1:1:1,3:3:1,2:2:1,3:2:1

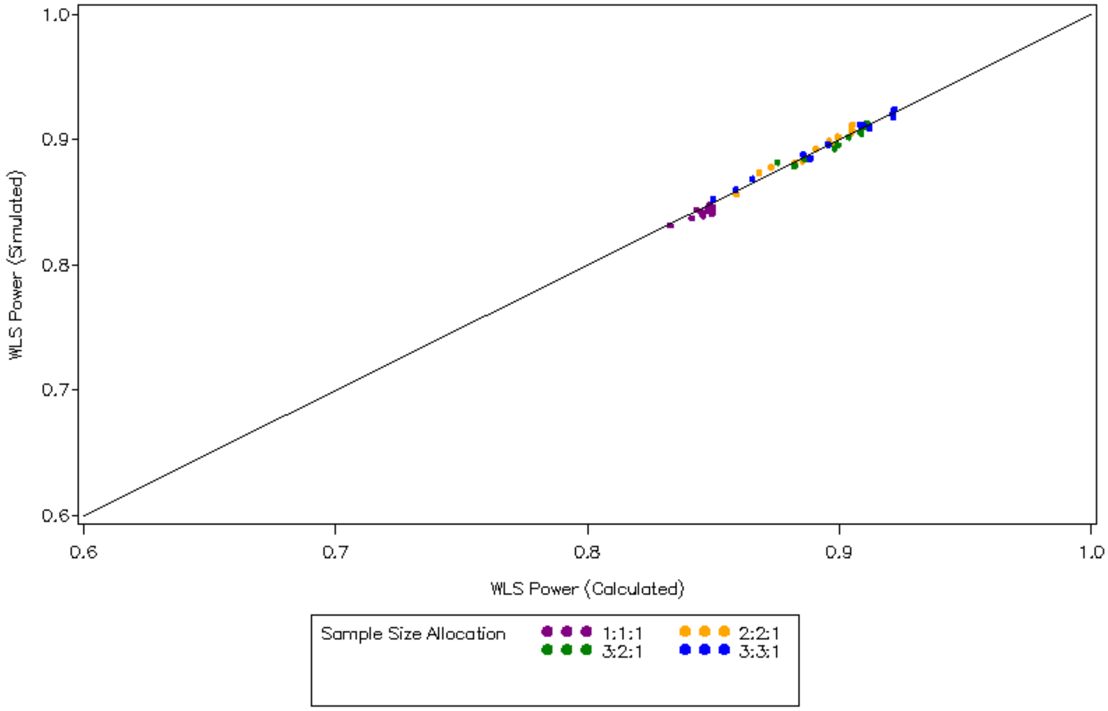


Figure 3.23 Comparison of Simulated and Calculated Power

Agresti & Caffo Simulated Method & RMLE Calculated Method
By Sample Size Allocation
Alpha=0.025, Lambda=1

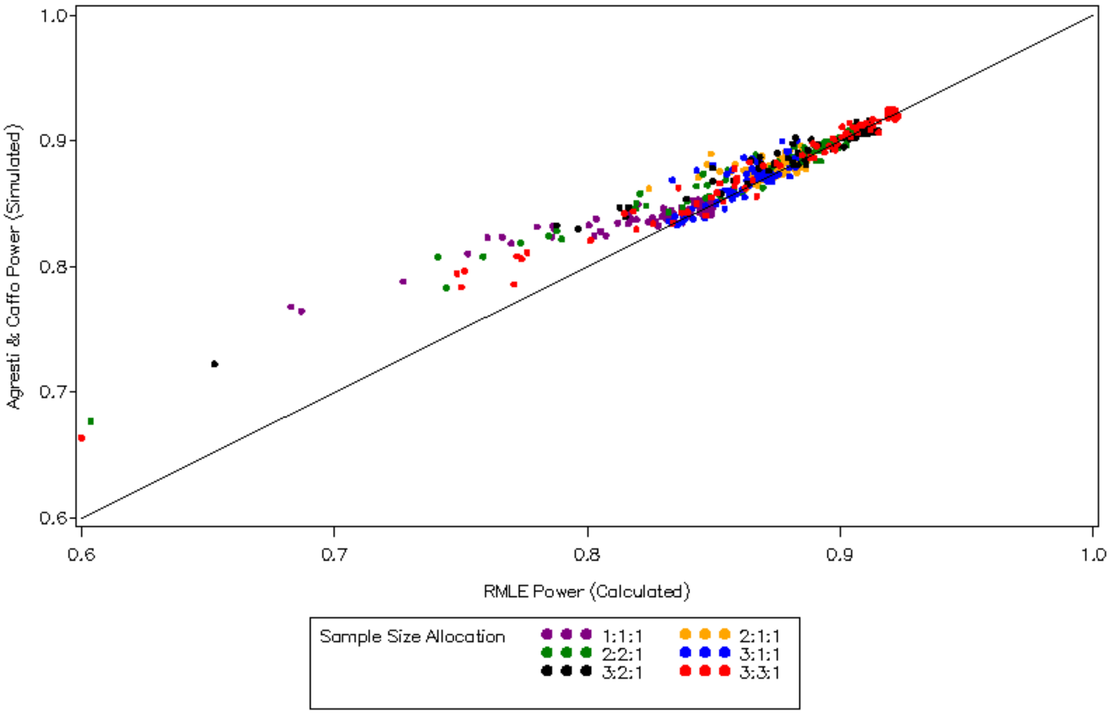


Figure 3.24 Comparison of Simulated and Calculated Power

Agresti & Caffo Simulated Method & RMLE Calculated Method
By Sample Size Allocation
Alpha=0.025, Lambda=1, Pi C<=0.7, Pi P>=0.4

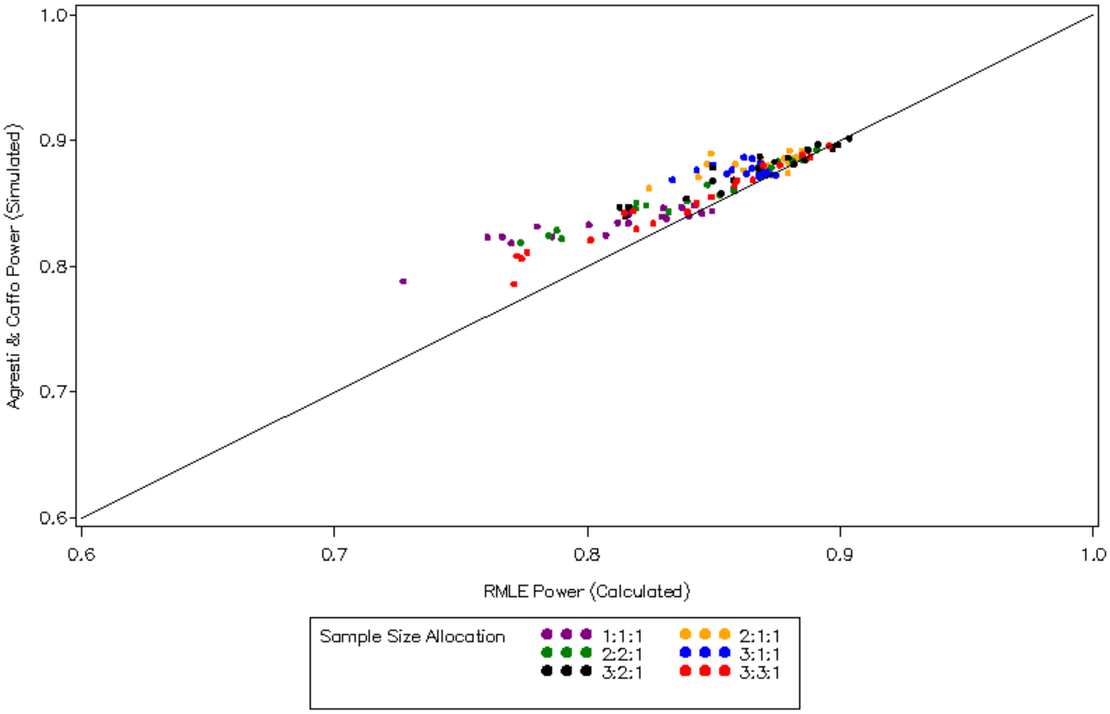


Figure 3.25 Comparison of Simulated and Calculated Power

Deviance Simulated Method & RMLE Calculated Method
By Sample Size Allocation
Alpha=0.025, Lambda=1

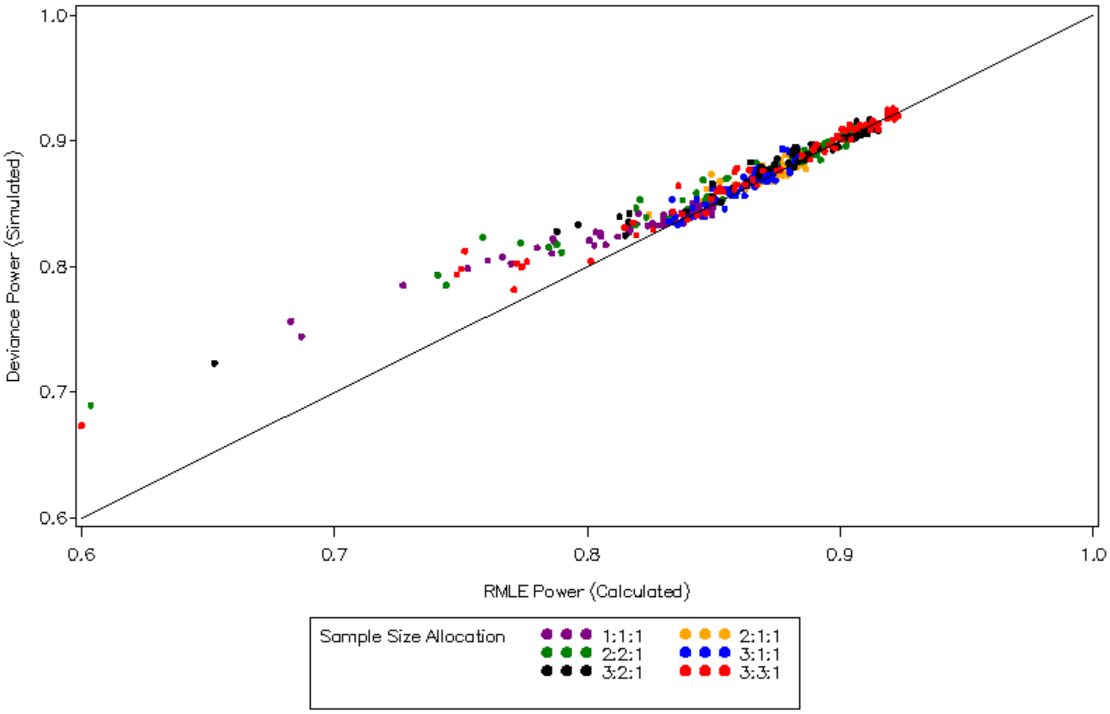


Figure 3.26 Summary of Simulated Type I Error

Non-inferiority in Two Separate Trials
Wald Method
By Sample Size Allocation

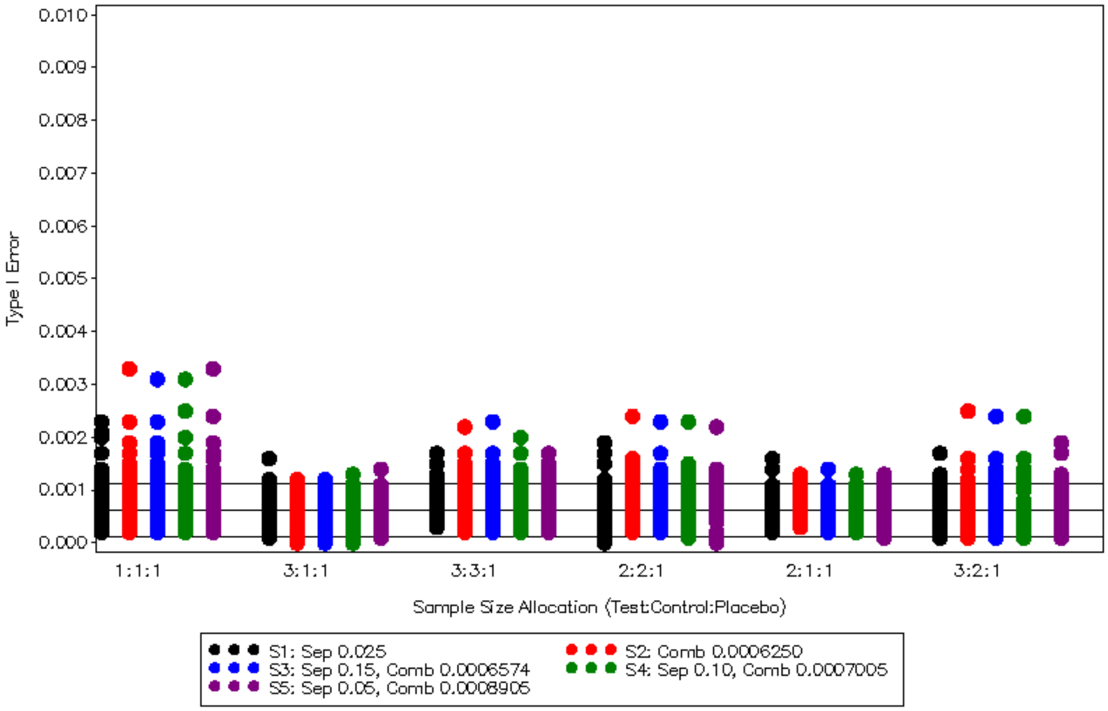


Figure 3.27 Summary of Simulated Type I Error

Non–inferiority in Two Separate Trials
Agresti & Caffo Method
By Sample Size Allocation

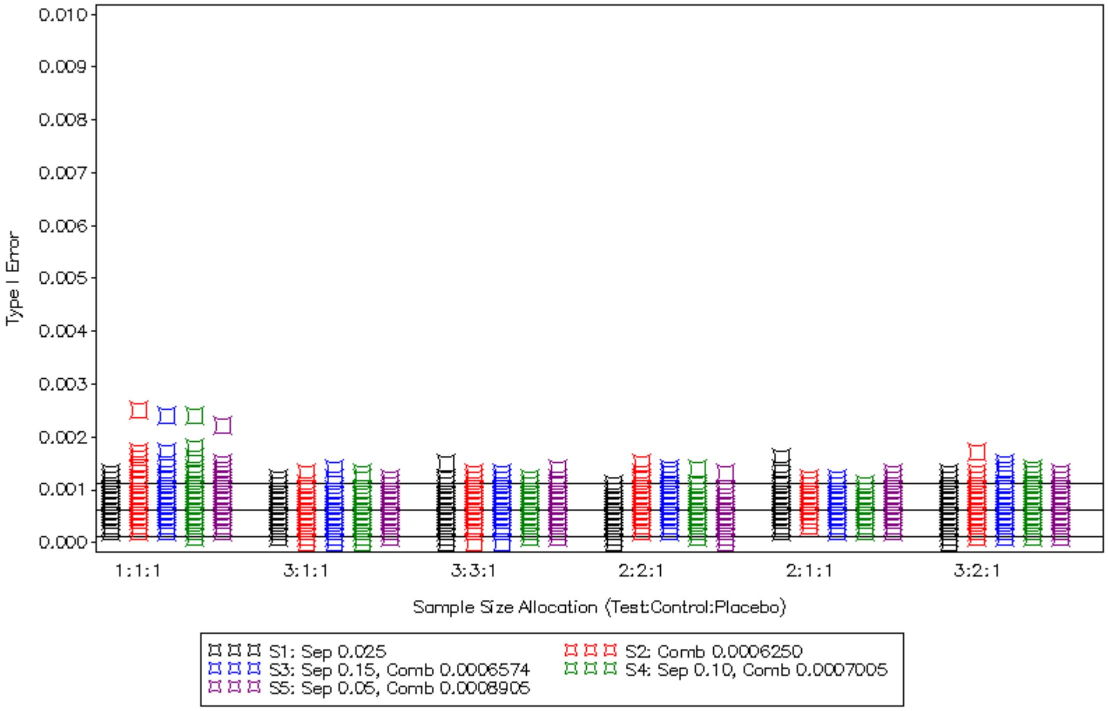


Figure 3.28 Summary of Simulated Type I Error

Non–inferiority in Two Separate Trials
RMLE Method
By Sample Size Allocation

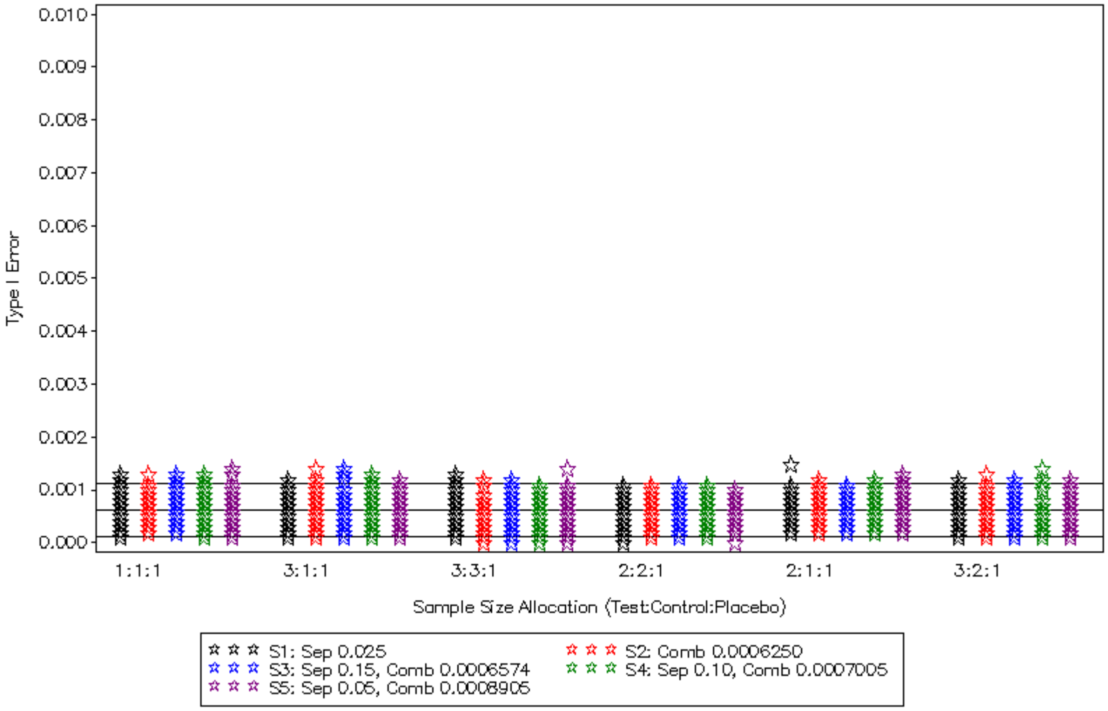


Figure 3.29 Summary of Simulated Type I Error

Non−inferiority in Two Separate Trials
Wald Method
By Non−inferiority Margin

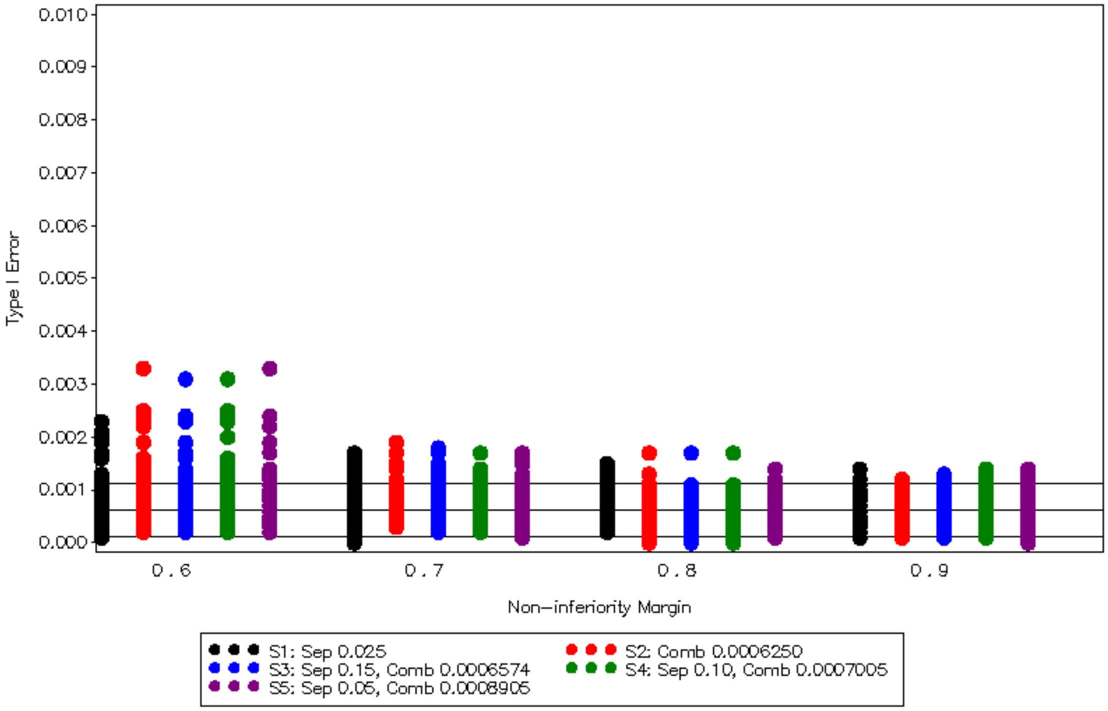


Figure 3.30 Summary of Simulated Type I Error

Non−inferiority in Two Separate Trials
Agresti & Caffo Method
By Non−inferiority Margin

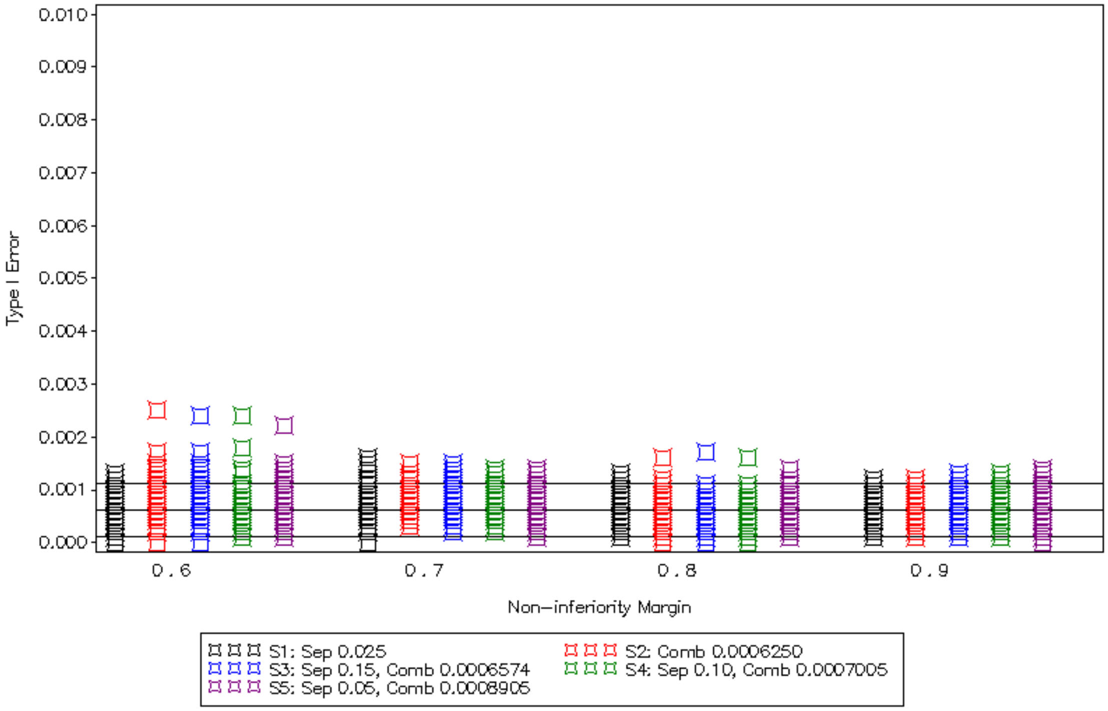


Figure 3.31 Summary of Simulated Type I Error

Non–inferiority in Two Separate Trials
RMLE Method
By Non–inferiority Margin

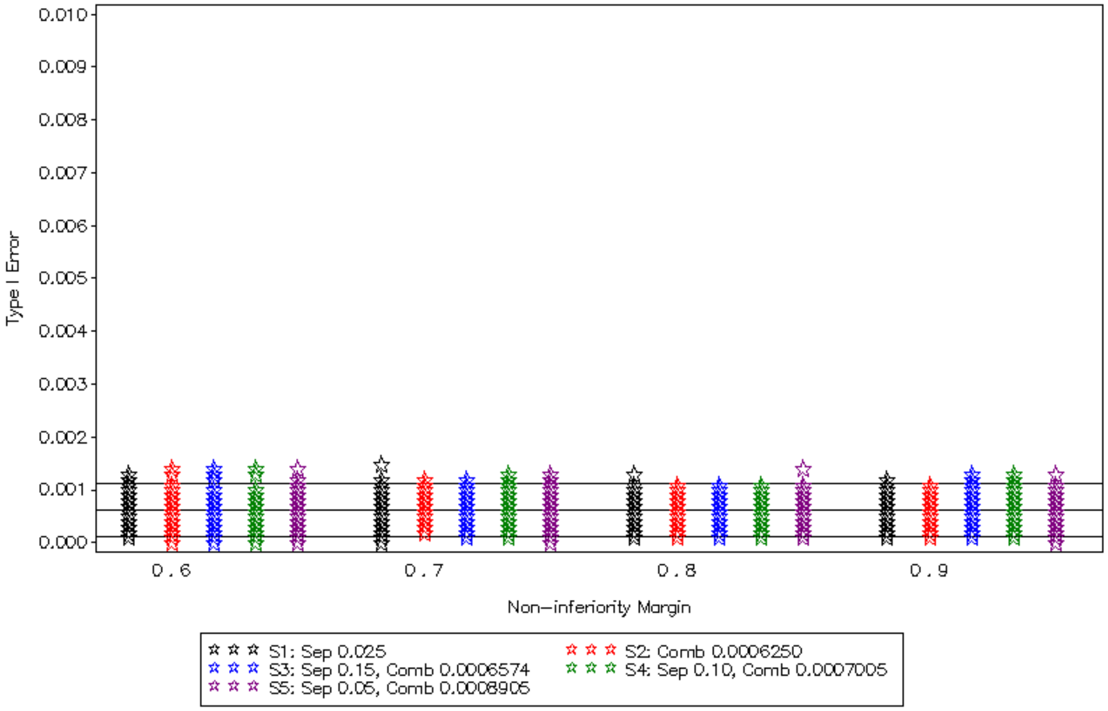


Figure 3.32 Summary of Simulated Type I Error

Non–inferiority in Two Separate Trials
Wald Method
By Pi C

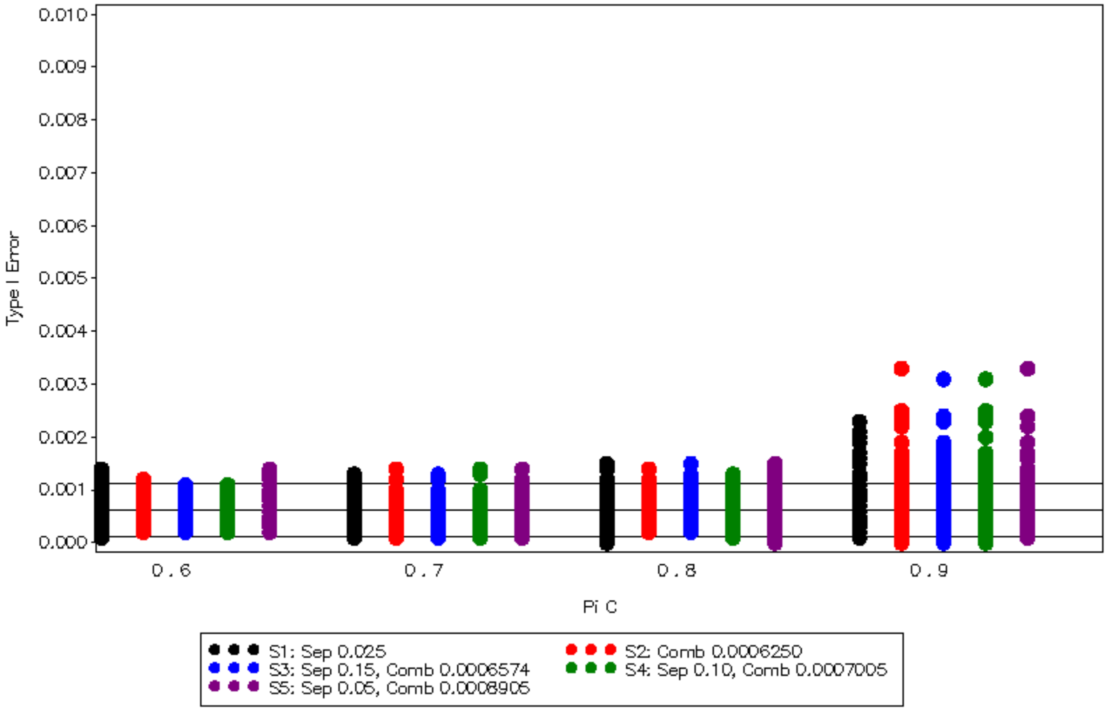


Figure 3.33 Summary of Simulated Type I Error
Non−inferiority in Two Separate Trials
Agresti & Caffo Method
By Pi C

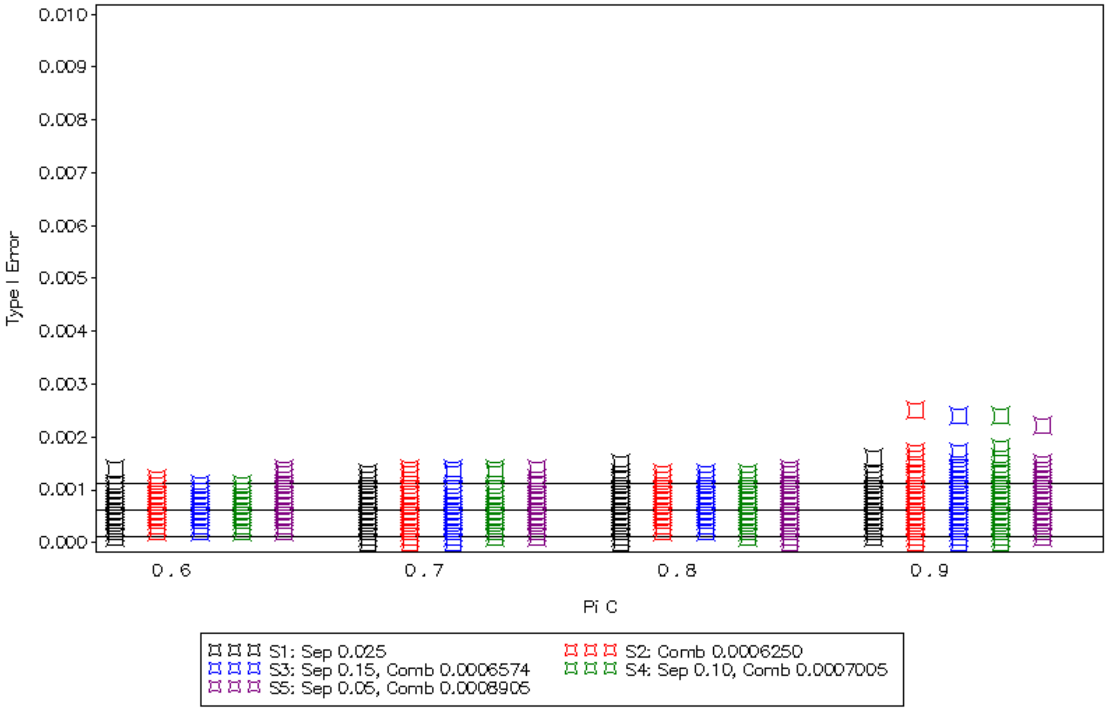


Figure 3.34 Summary of Simulated Type I Error
Non−inferiority in Two Separate Trials
RMLE Method
By Pi C

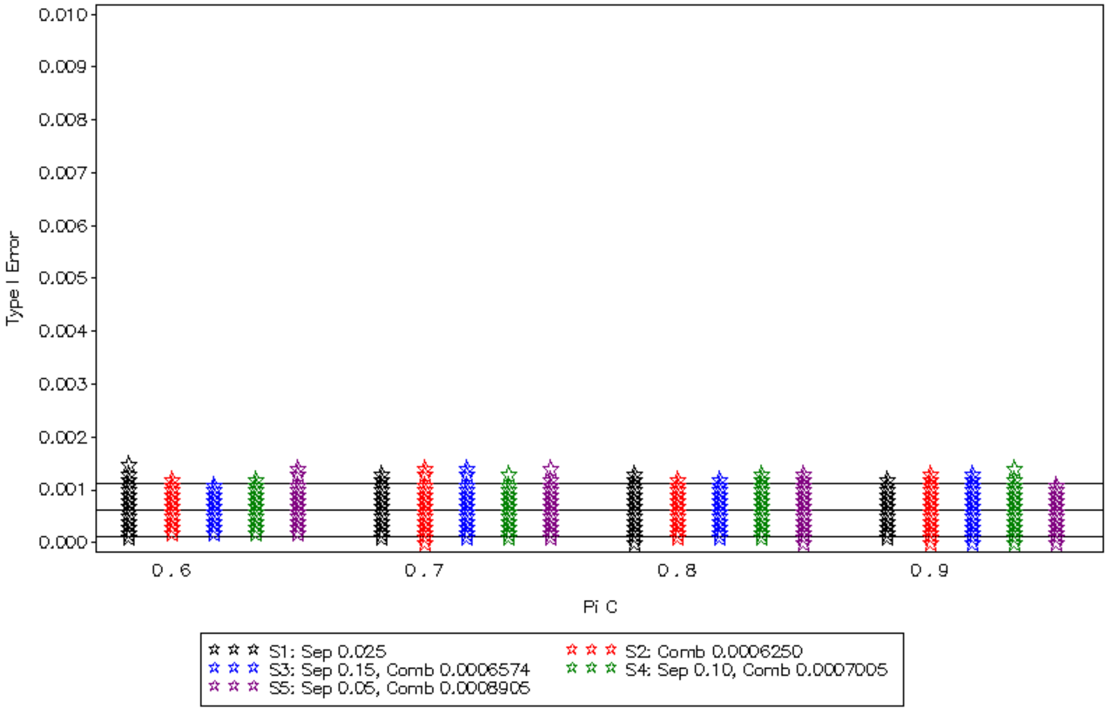


Figure 3.35 Summary of Simulated Type I Error

Non-inferiority in Two Separate Trials

Wald Method

By Pi P

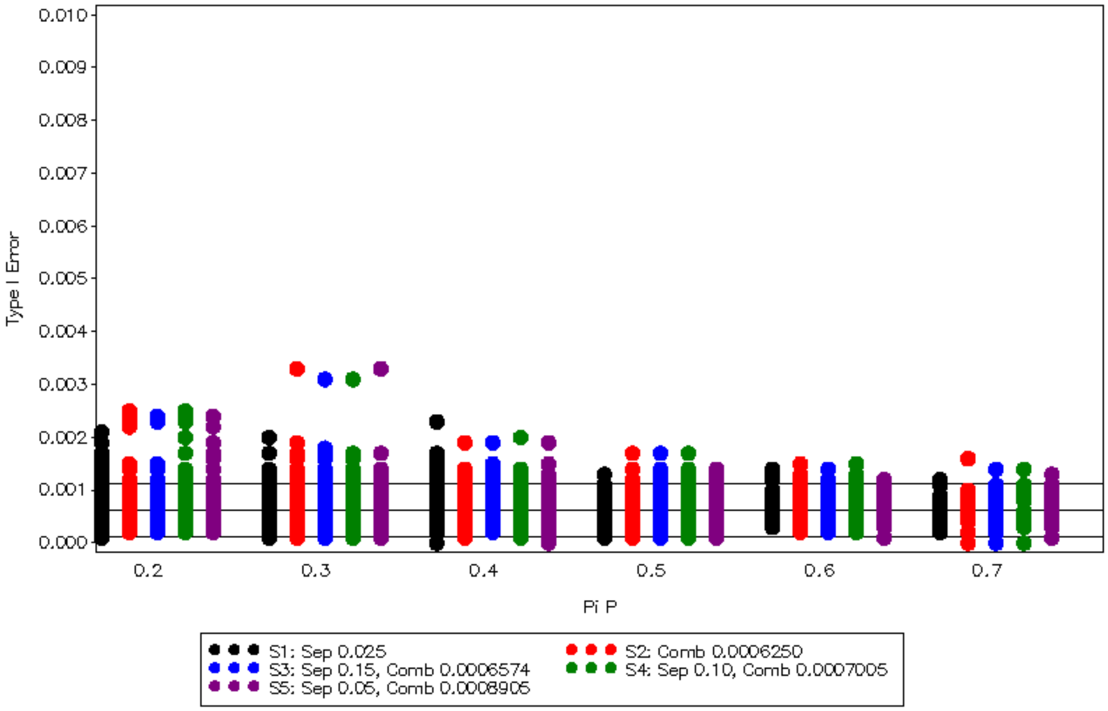


Figure 3.36 Summary of Simulated Type I Error

Non-inferiority in Two Separate Trials

Agresti & Caffo Method

By Pi P

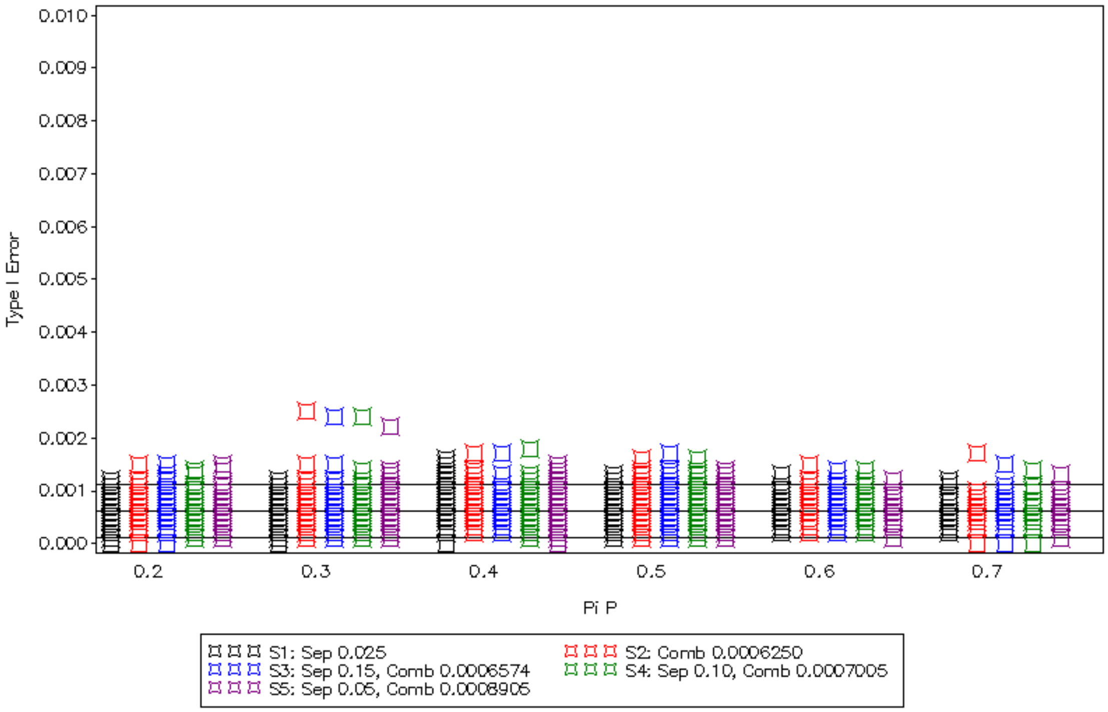


Figure 3.37 Summary of Simulated Type I Error

Non–inferiority in Two Separate Trials
RMLE Method
By Pi P

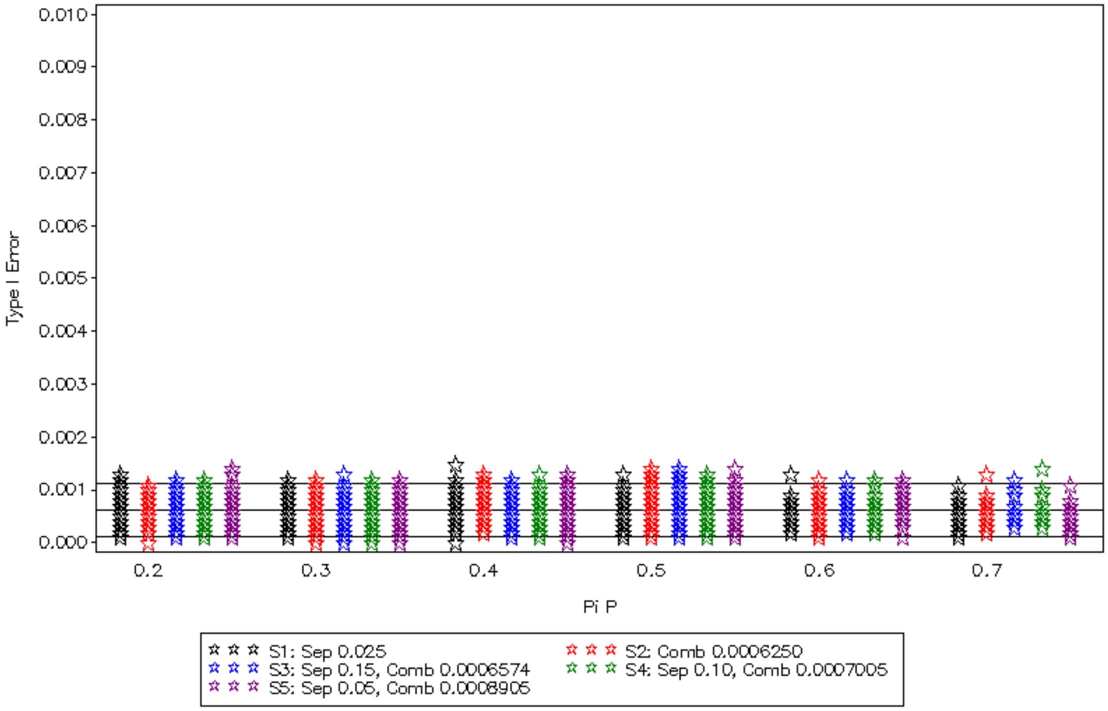


Figure 3.38 Summary of Simulated Power

Non–inferiority in Two Separate Trials
Wald Method
By Sample Size Allocation

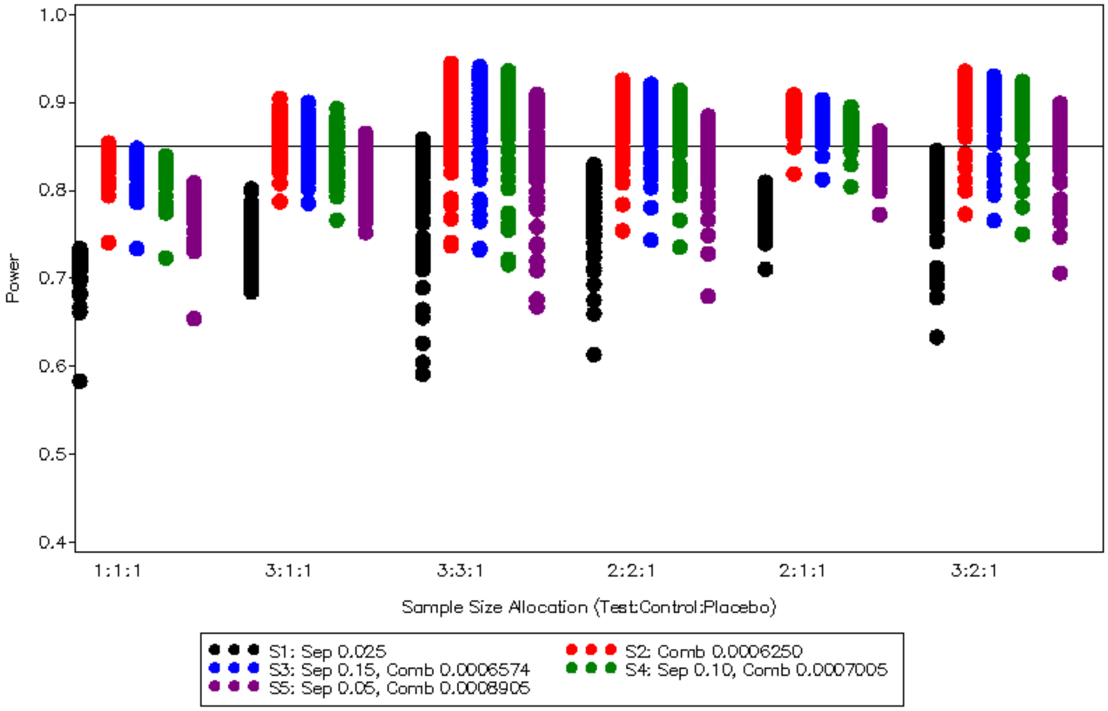


Figure 3.39 Summary of Simulated Power

Non−inferiority in Two Separate Trials
Agresti & Caffo Method
By Sample Size Allocation

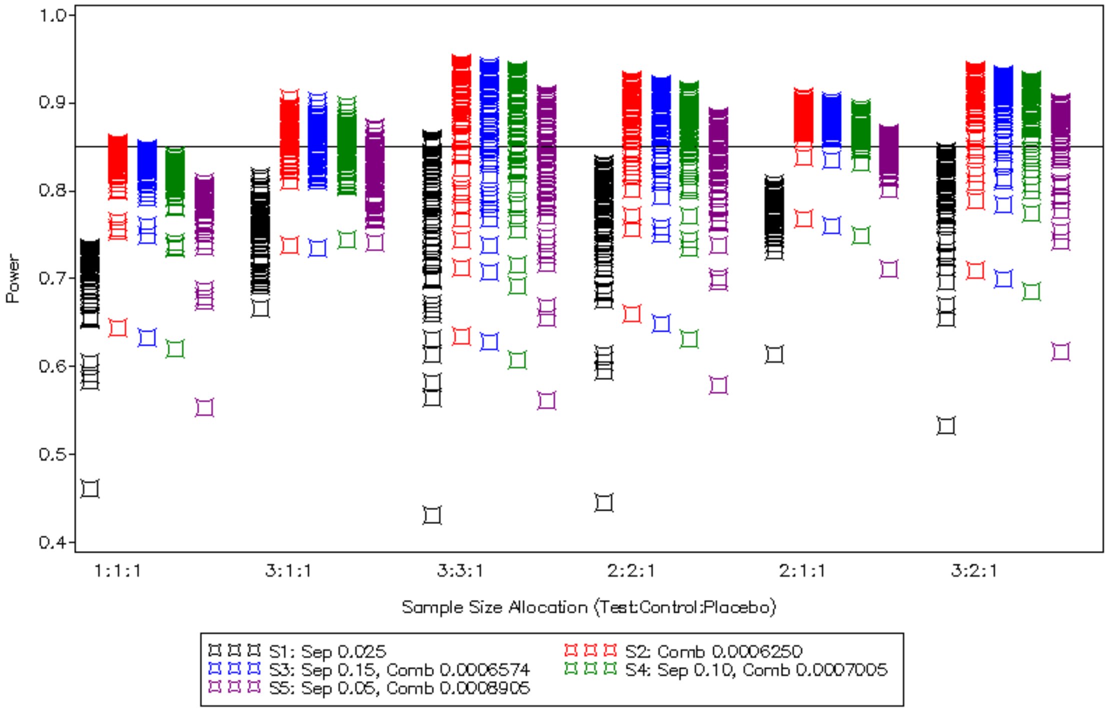


Figure 3.40 Summary of Simulated Power

Non−inferiority in Two Separate Trials
RMLE Method
By Sample Size Allocation

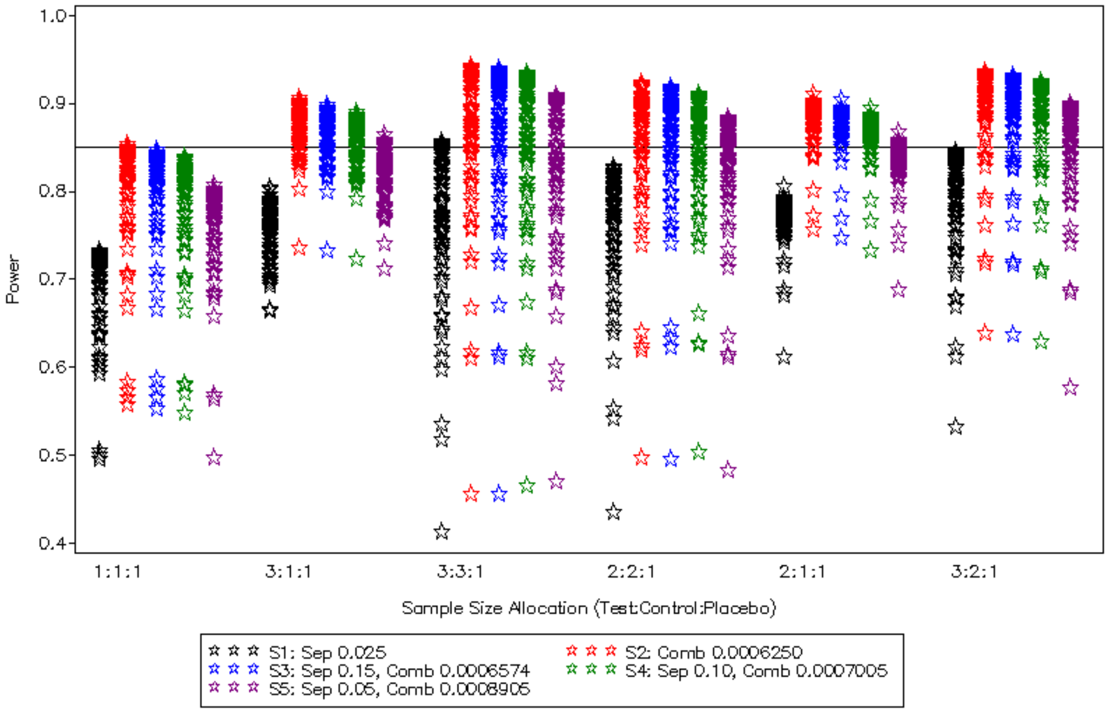


Figure 3.41 Summary of Simulated Power

Non−inferiority in Two Separate Trials
Wald Method
By Non−inferiority Margin

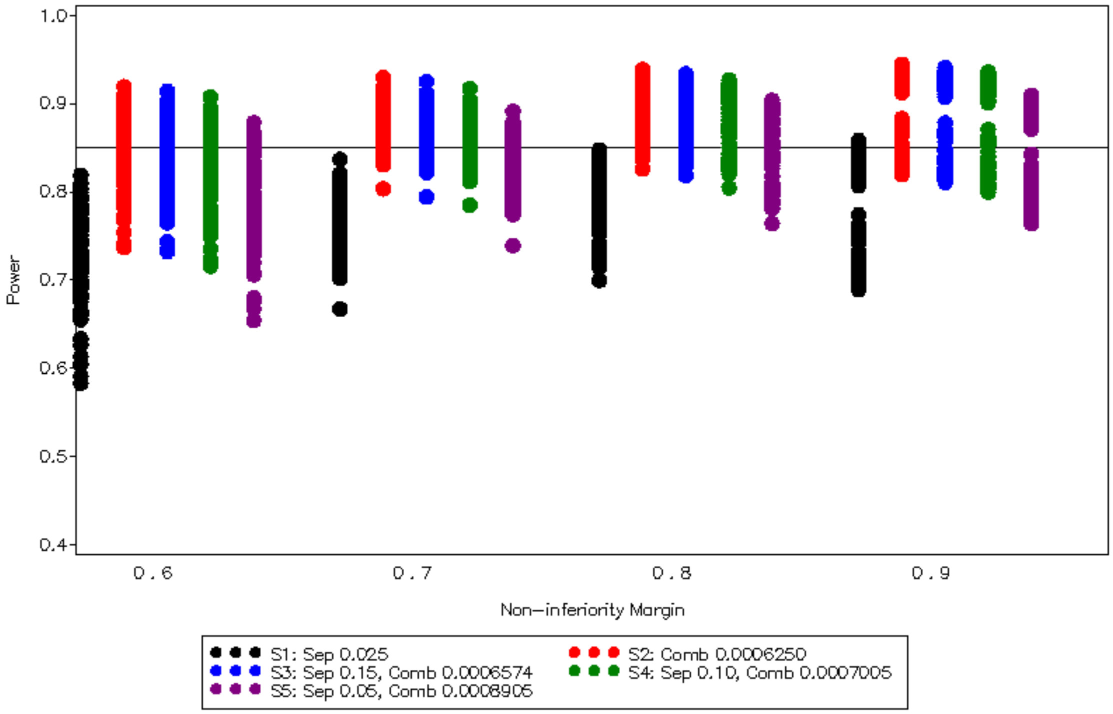


Figure 3.42 Summary of Simulated Power

Non−inferiority in Two Separate Trials
Agresti & Caffo Method
By Non−inferiority Margin

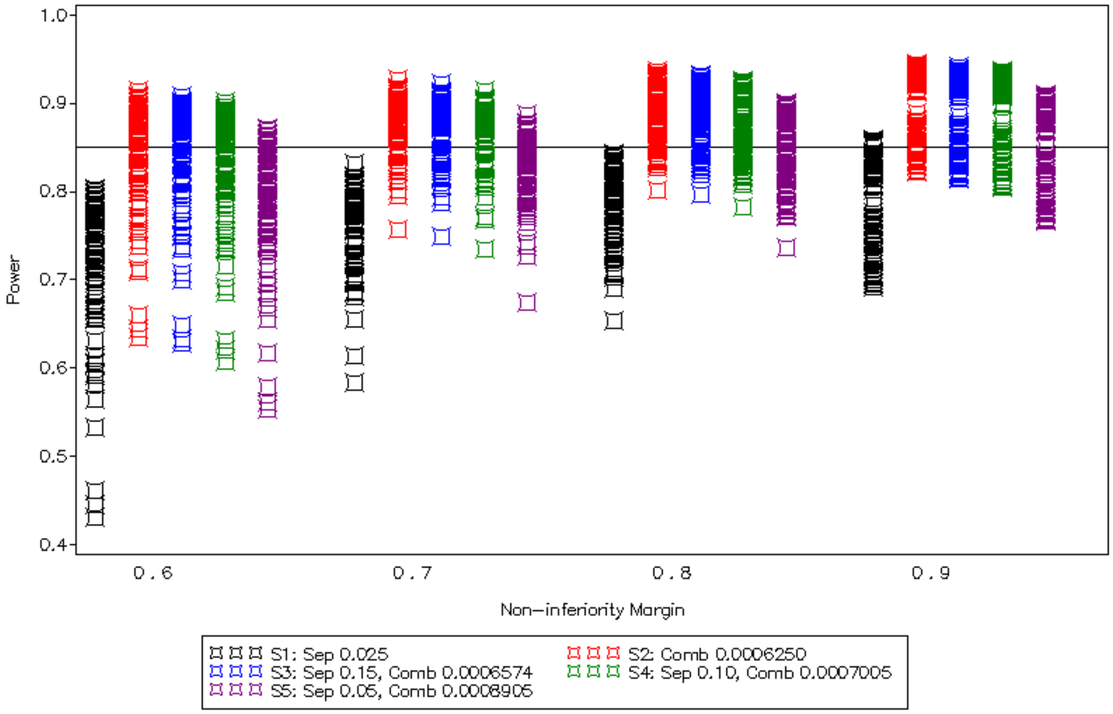


Figure 3.43 Summary of Simulated Power
Non−inferiority in Two Separate Trials
RMLE Method
By Non−inferiority Margin

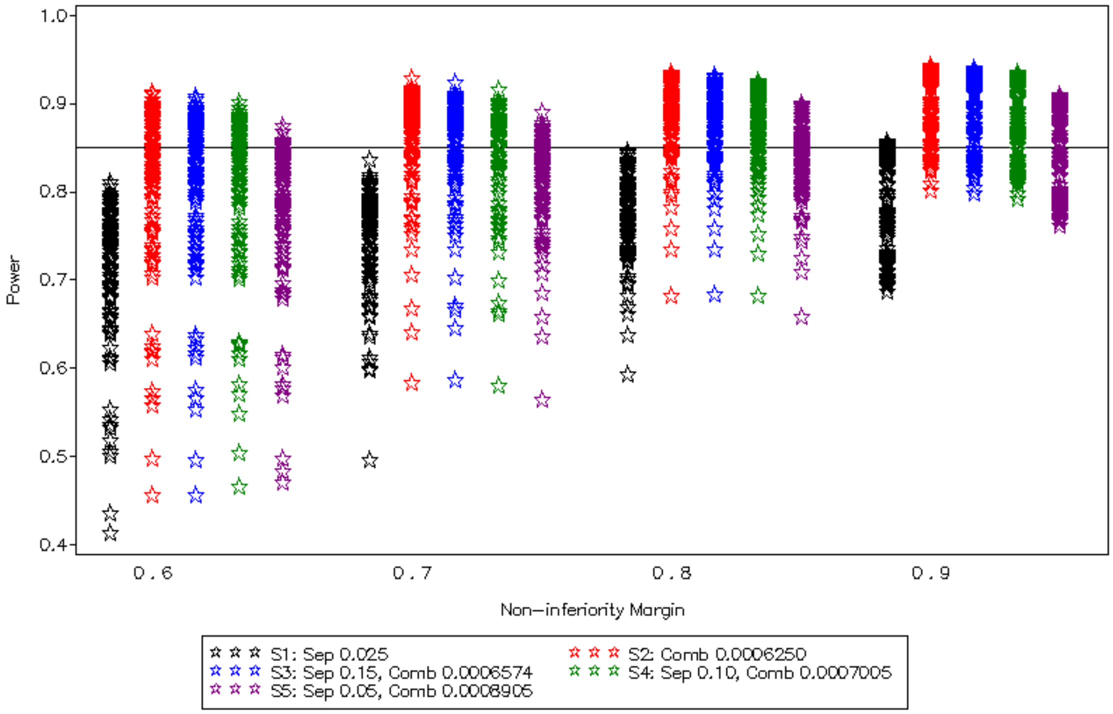


Figure 3.44 Summary of Simulated Power
Non−inferiority in Two Separate Trials
Wald Method
By Pi C

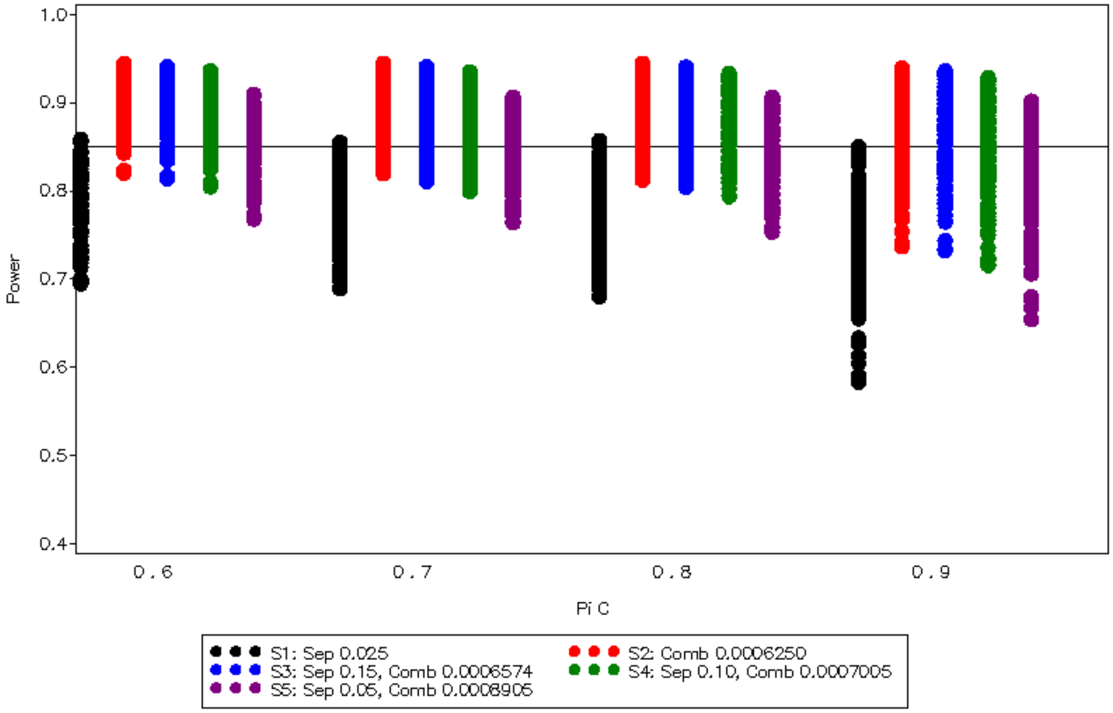


Figure 3.45 Summary of Simulated Power
Non–inferiority in Two Separate Trials
Agresti & Caffo Method
By Pi C

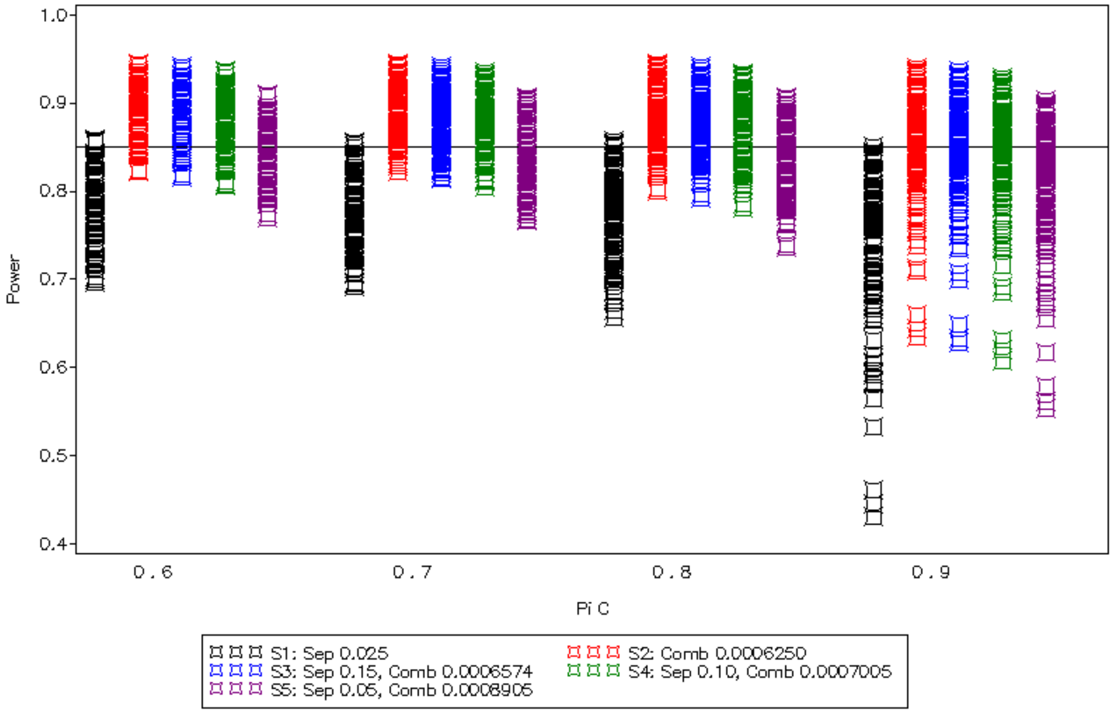


Figure 3.46 Summary of Simulated Power
Non–inferiority in Two Separate Trials
RMLE Method
By Pi C

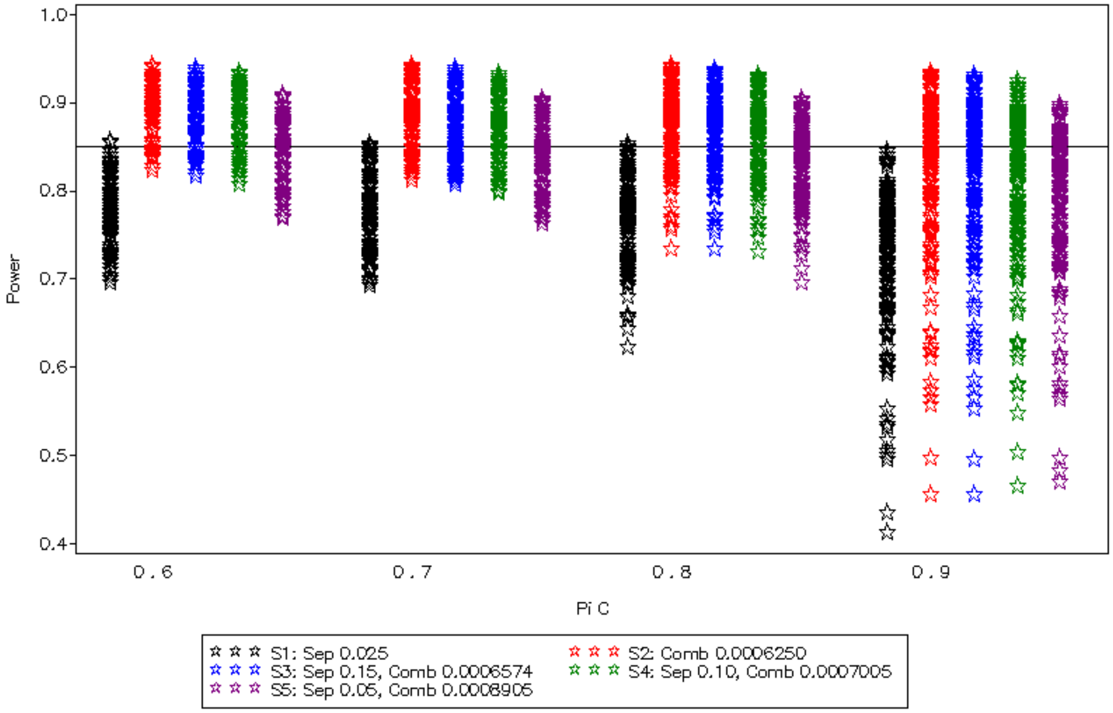


Figure 3.47 Summary of Simulated Power

Non–inferiority in Two Separate Trials
Wald Method
By Pi P

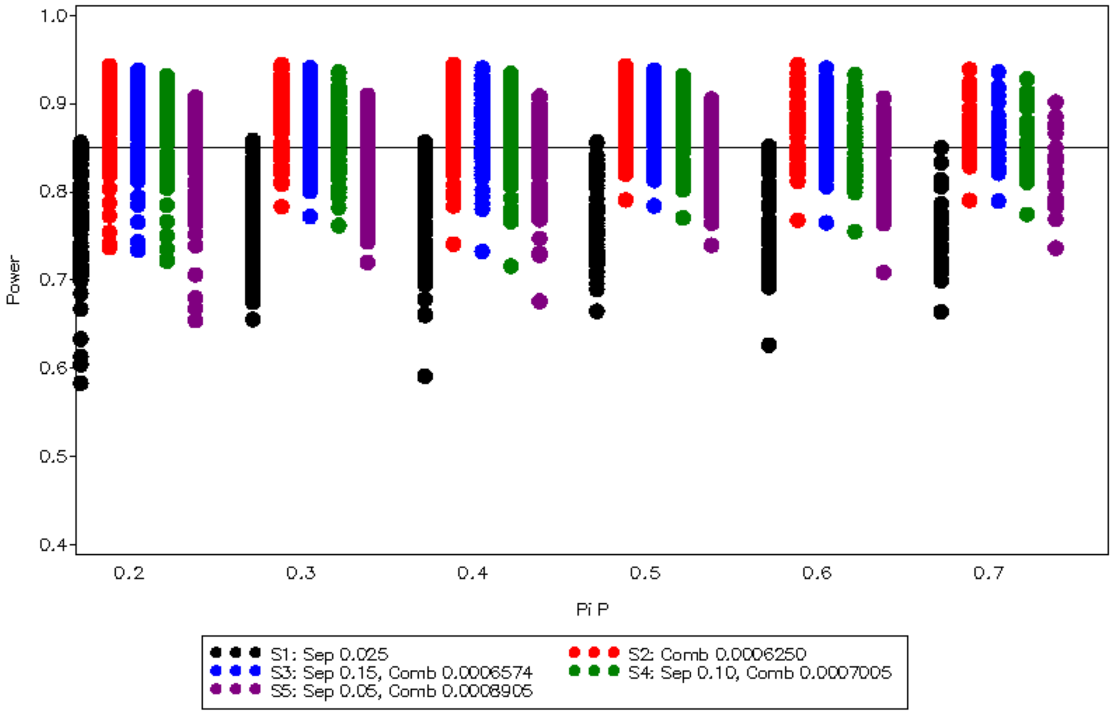


Figure 3.48 Summary of Simulated Power

Non–inferiority in Two Separate Trials
Agresti & Caffo Method
By Pi P

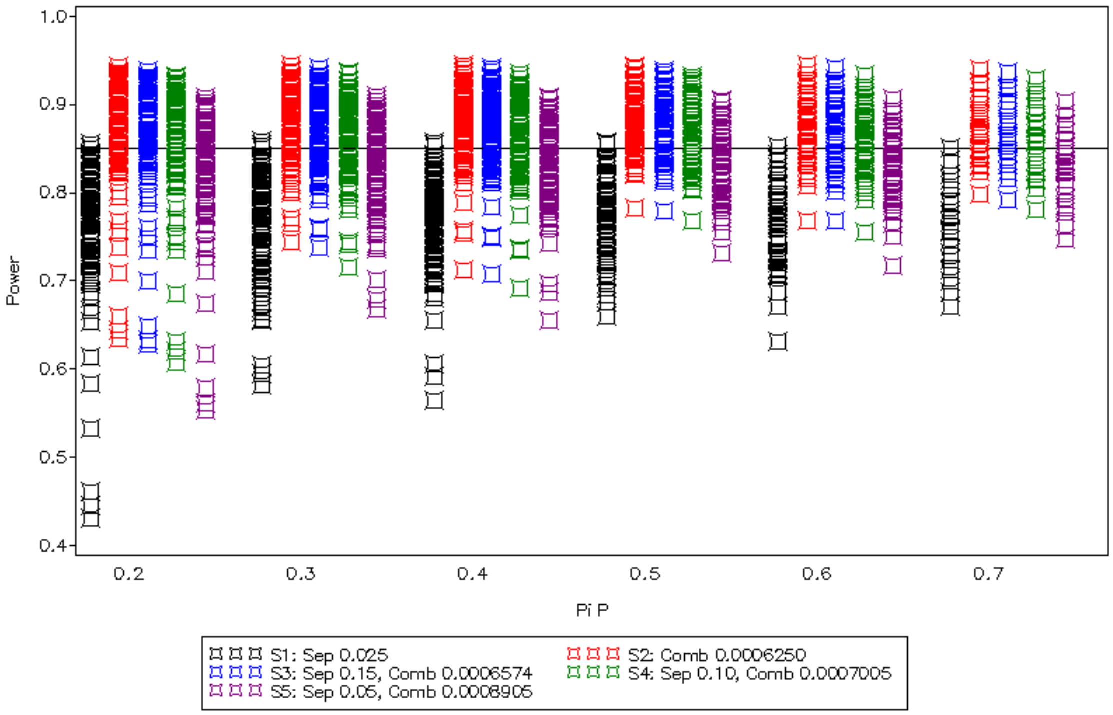


Figure 3.49 Summary of Simulated Power

Non–inferiority in Two Separate Trials
RMLE Method
By Pi P

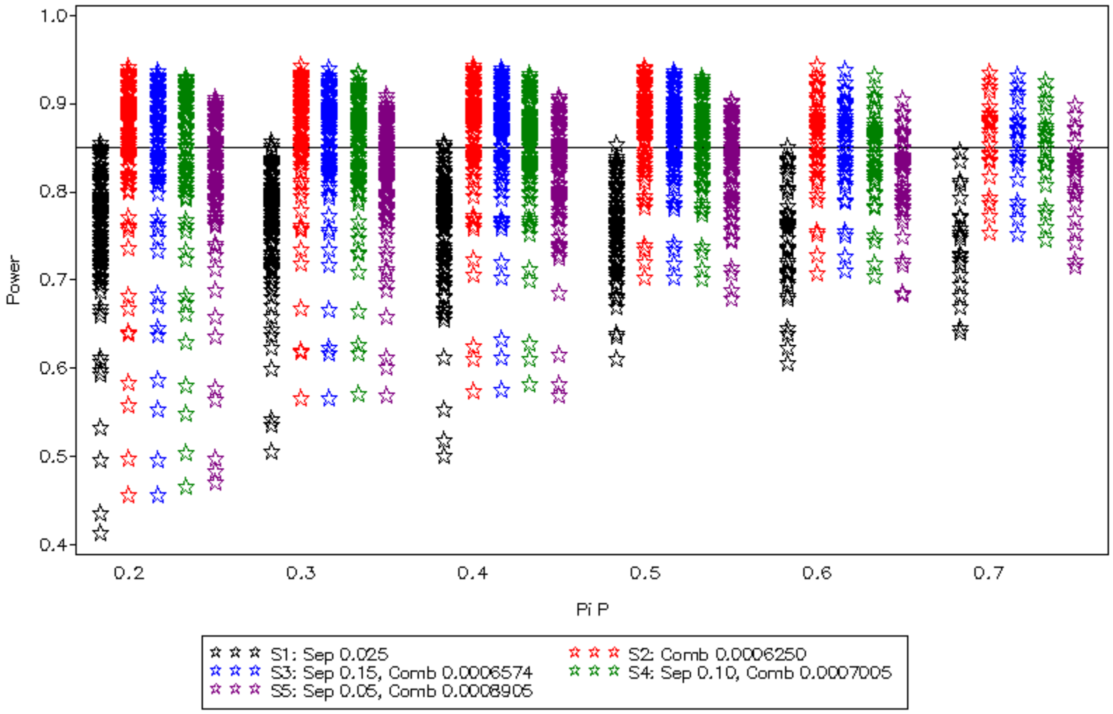


Figure 3.50 Summary of Type I Error

Non–inferiority:Combined Data using RMLE Method
Superiority:Two Separate Trials using F–M 3 Method
By Sample Size Allocation

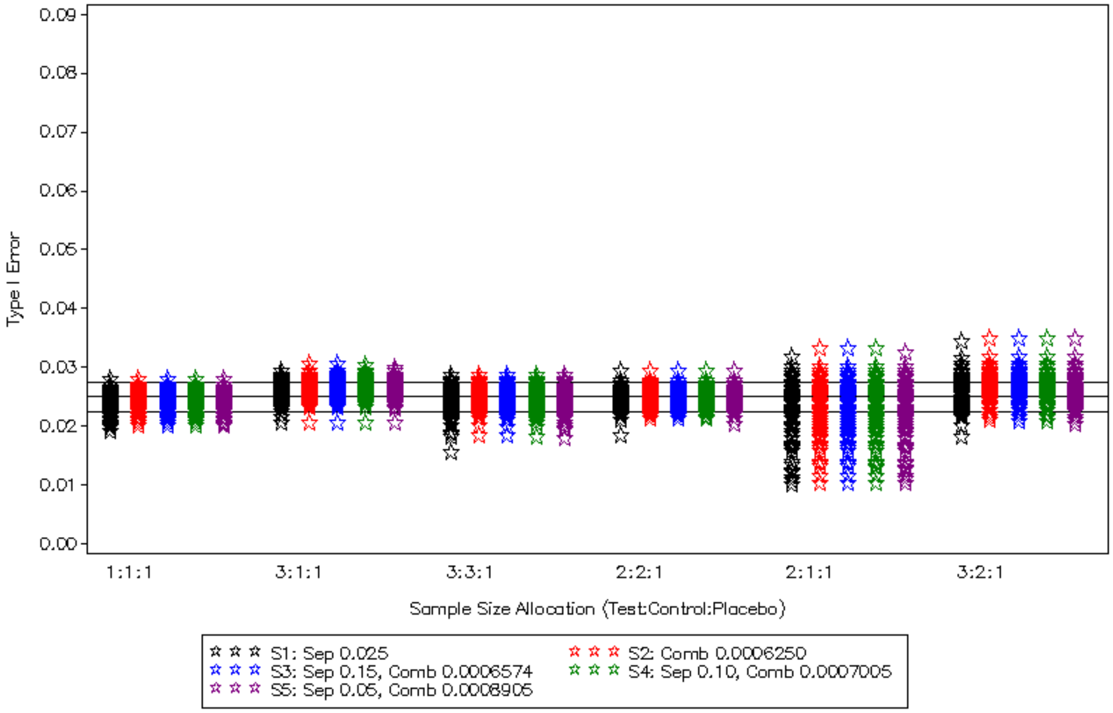


Figure 3.51 Summary of Type I Error
Non−inferiority:Combined Data using RMLE Method
Superiority:Two Separate Trials using F−M 3 Method
By Non−inferiority Margin

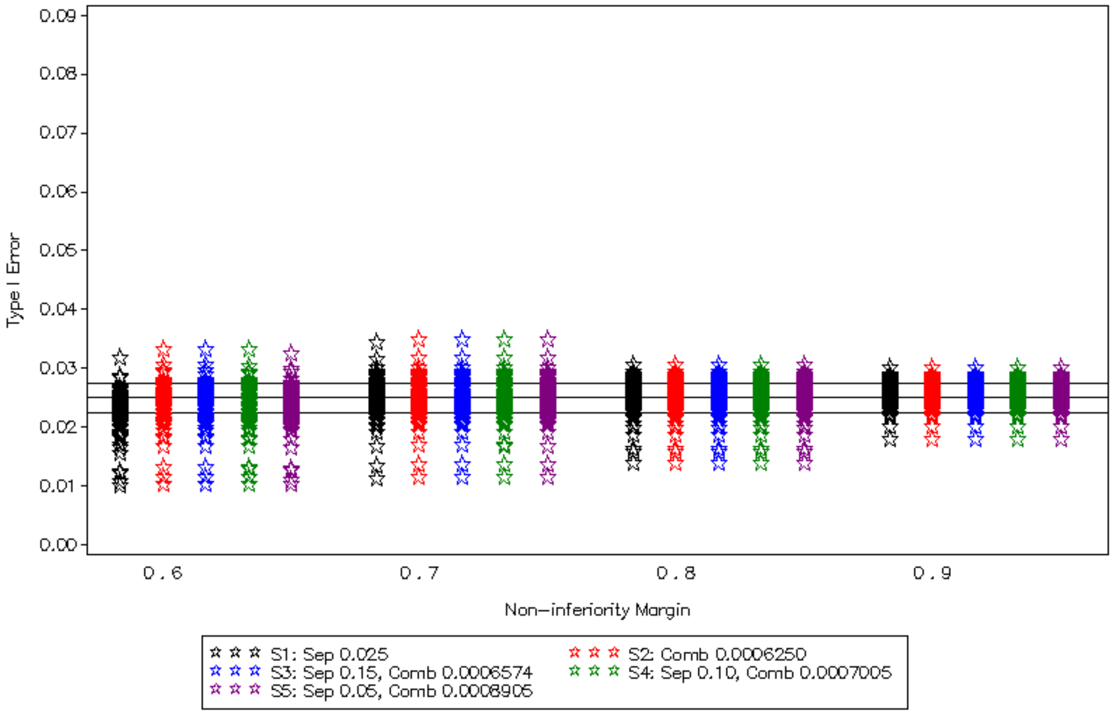


Figure 3.52 Summary of Type I Error
Non−inferiority:Combined Data using RMLE Method
Superiority:Two Separate Trials using F−M 3 Method
By Pi C

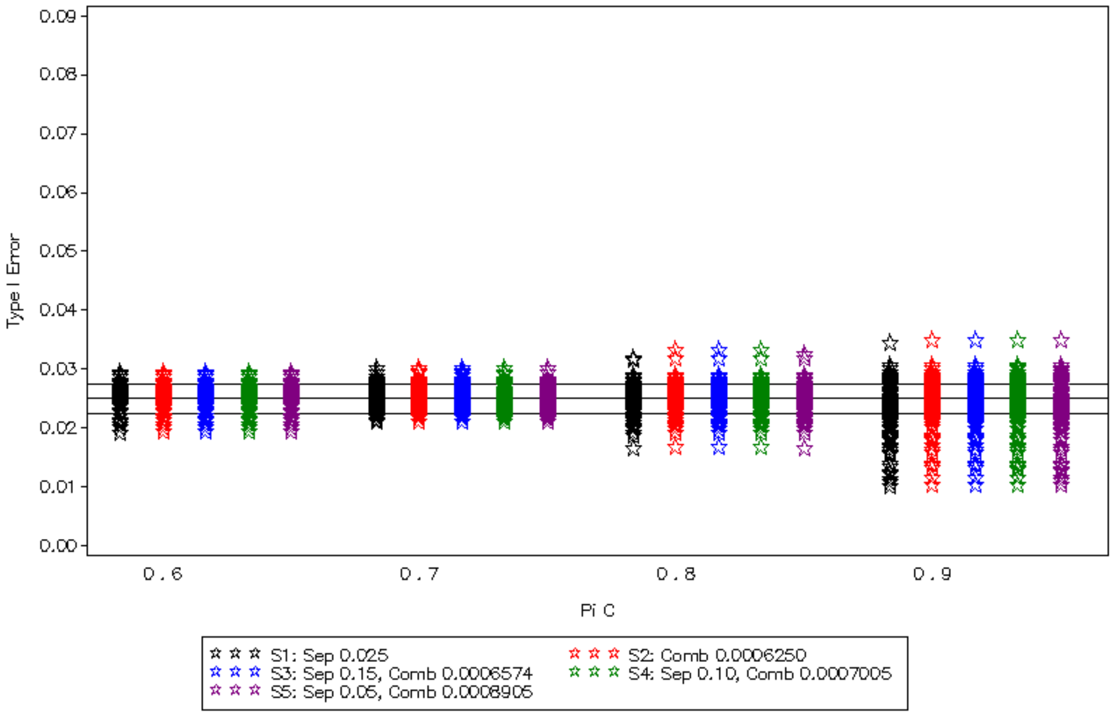


Figure 3.53 Summary of Type I Error
Non–inferiority:Combined Data using RMLE Method
Superiority:Two Separate Trials using F–M 3 Method
By Pi P

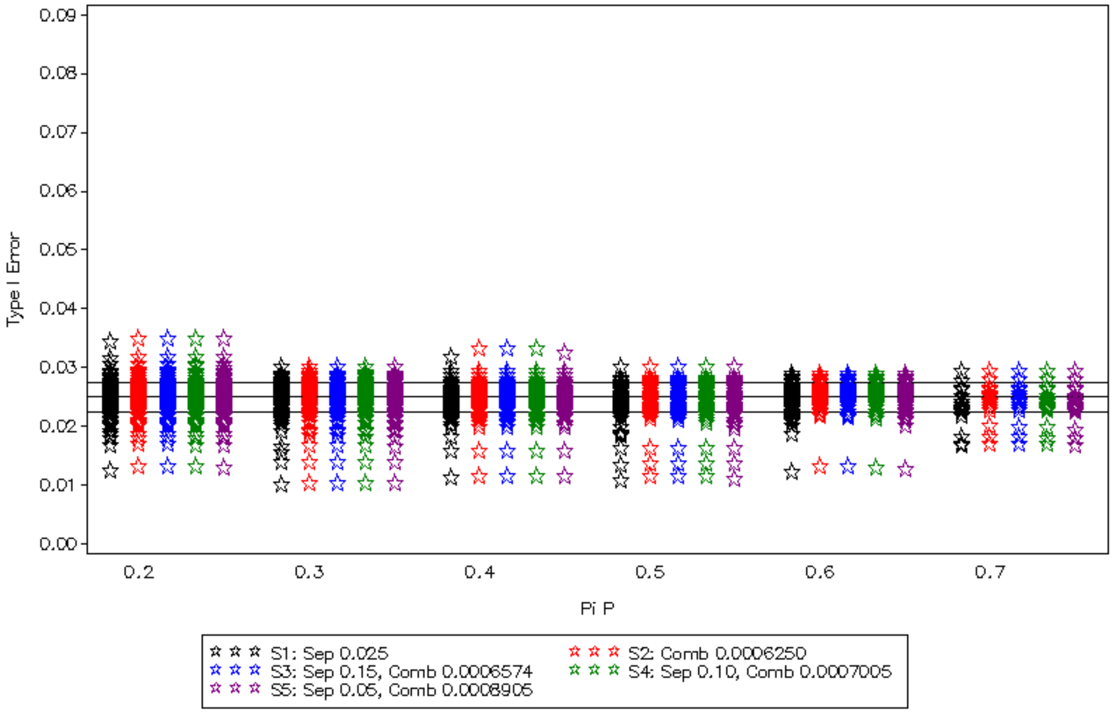


Figure 3.54 Summary of Simulated Power
Non–inferiority:Combined Data using RMLE Method
Superiority:Two Separate Trials using F–M 3 Method
By Sample Size Allocation

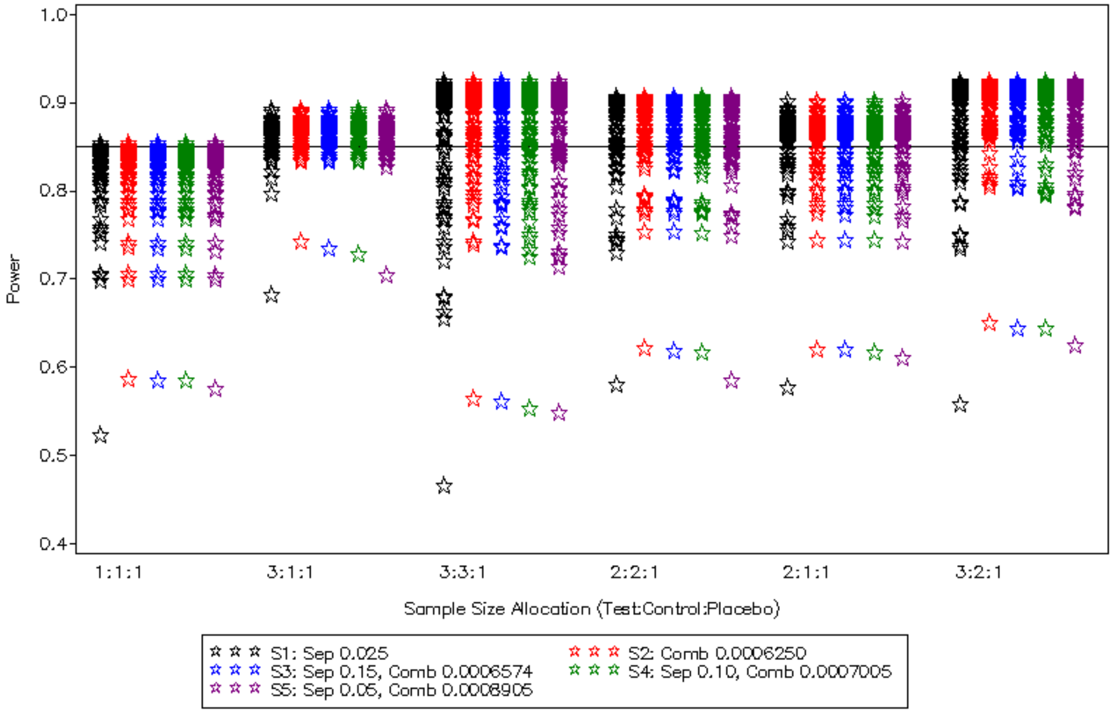


Figure 3.55 Summary of Simulated Power
Non−inferiority:Combined Data using RMLE Method
Superiority:Two Separate Trials using F−M 3 Method
By Non−inferiority Margin

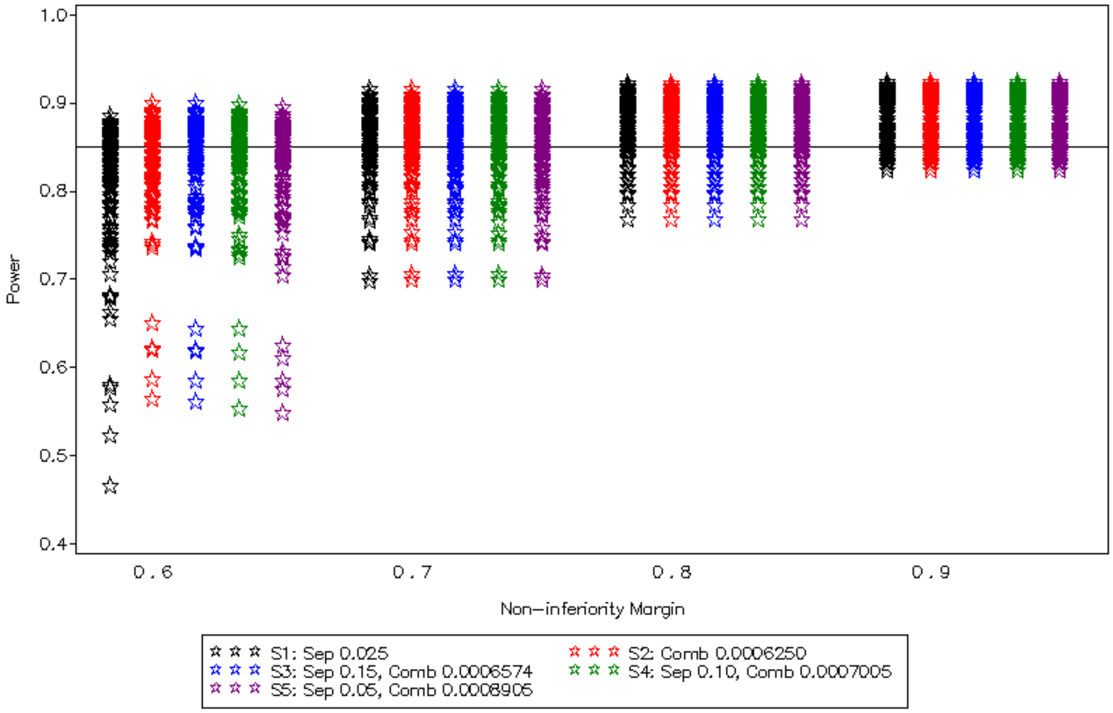


Figure 3.56 Summary of Simulated Power
Non−inferiority:Combined Data using RMLE Method
Superiority:Two Separate Trials using F−M 3 Method
By Pi C

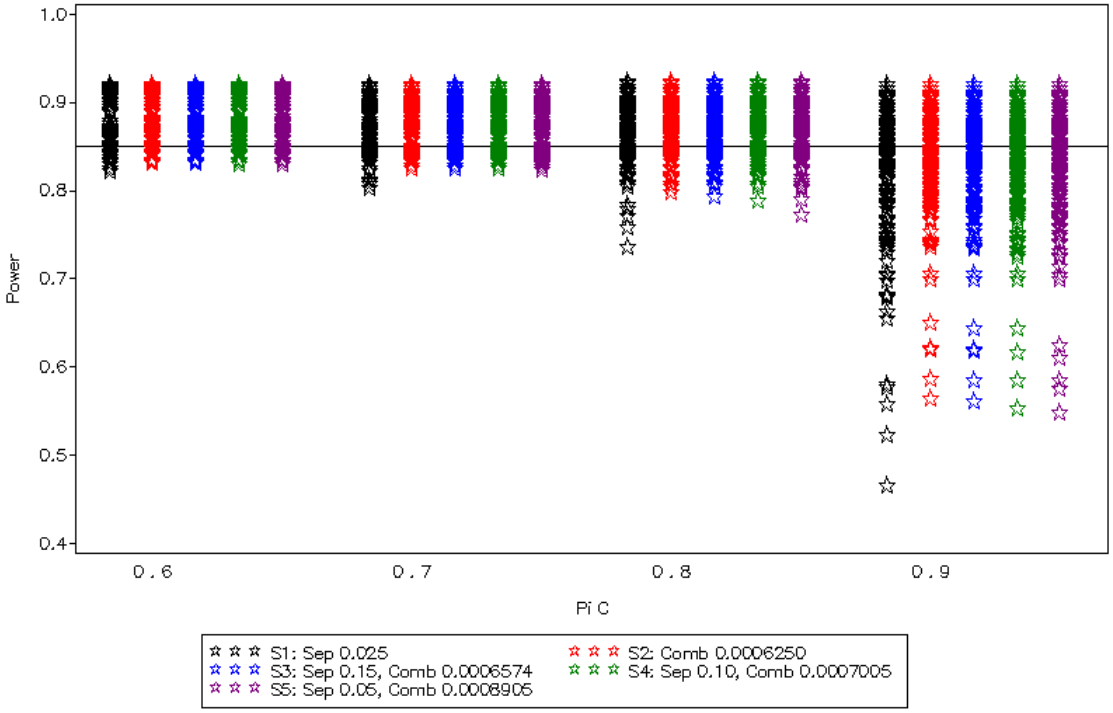
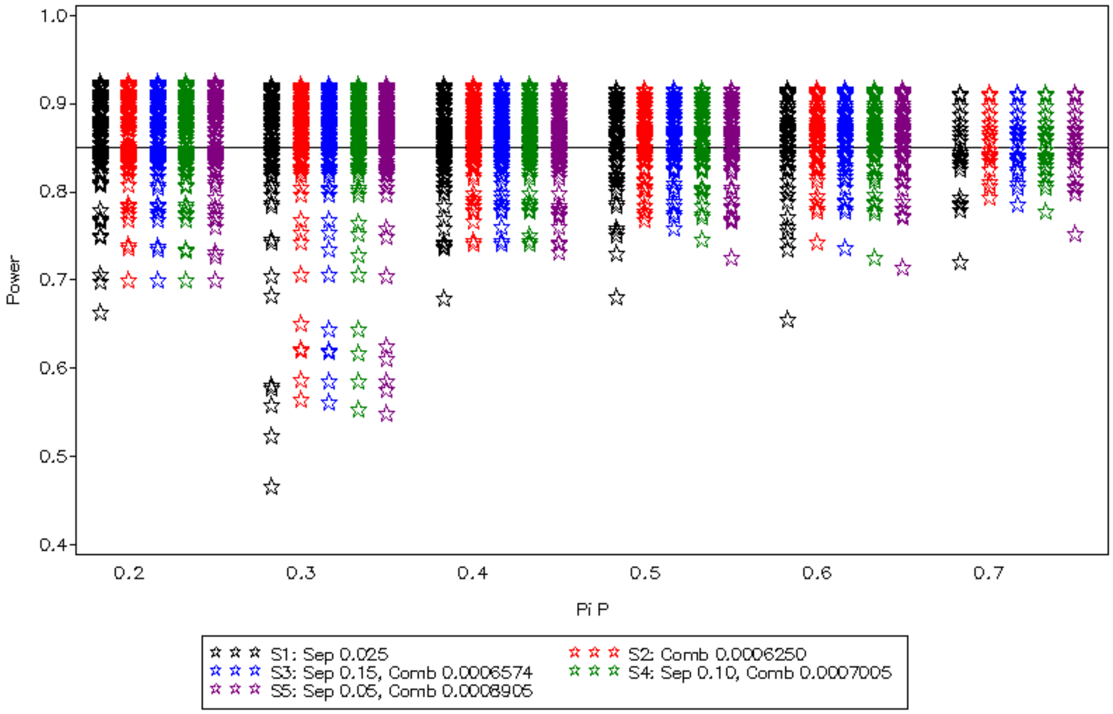


Figure 3.57 Summary of Simulated Power

Non−inferiority:Combined Data using RMLE Method

Superiority:Two Separate Trials using F−M 3 Method

By Pi P



Chapter 4

Methods for Analyzing Stratified Non-inferiority Trials with Binomial Proportions as the Primary Endpoint with Criteria for the Risk Ratio or the Risk Difference

I. Introduction

Non-inferiority trials are designed for settings where the objective is showing the new experimental test treatment is not unacceptably worse than the current active control treatment by a specified amount. The test treatment may be equivalent in efficacy, but have less severe adverse events or a better dosing regimen for patient compliance. In the design of these trials, there may be strata that have to be accounted for in the initial planning and also in the final analyses. These strata could be different geographic regions of patient recruitment or they could be based on covariates thought to have some differential effect related to the efficacy outcome of interest such as gender or age groups.

The focus of the present discussion includes dichotomous primary endpoints with criteria for analyzing the risk ratio or the risk difference for the test and active control treatments, while accounting for the stratification variable. A review of the current methodology and modifications of them will be presented. Simulations will be used to investigate the performance of these methods for various situations related to type I error and power. Additionally, sample size and power formulas will be discussed for the planning of these non-inferiority trials when taking into account stratification.

II. Assessing Non-inferiority of a Risk Ratio in a Stratified Trial

A. Methods for the Stratified Risk Ratio

The hypotheses for testing non-inferiority of the test treatment compared to the active control treatment for the risk ratio rely on a pre-determined non-inferiority limit, θ_0, which is seen in $H_0 : \theta_h = \pi_{Th}/\pi_{Ch} \geq \theta_0$ and H_A: $\theta_h = \pi_{Th}/\pi_{Ch} < \theta_0$ where $\theta_h = \pi_{Th}/\pi_{Ch}$ is the population risk ratio for the test group versus the active control group in the *h*th stratum, with *h*=1, 2, …, H and with π_{Th} and π_{Ch} representing the population proportion of patients with the unfavorable outcome of interest in the test and active control groups, respectively, for the *h*th stratum.

A test of the null hypothesis of inferiority can be performed using a test statistic and comparing the subsequent p-value to the specified alpha level or through the computation of a confidence interval and the evaluation of its inclusion of the null value, θ_0. Both of these approaches to testing the null hypothesis are seen in the methods presented.

Gart[1] proposed a method based on the test statistic in (4.1) which follows a standard normal distribution and produces a corresponding p-value

$$z_G(\theta_0) = \frac{\sum_h \left\{ \frac{(y_{Th} - n_{Th}\tilde{p}_{Th})}{(1-\tilde{p}_{Th})} \right\}}{\left\{ \sum_h v_h(\theta_0, \tilde{p}_{Ch}) \right\}^{1/2}} \tag{4.1}$$

where $v_h(\theta_0, \tilde{p}_{Ch}) = \dfrac{n_{Ch} n_{Th} \tilde{p}_{Th}}{n_{Th}(\theta_0 - \tilde{p}_{Th}) + n_{Ch}(1-\tilde{p}_{Th})}$

In (4.1), y_{Th} is the observed number of events in the test group for the *h*th stratum out of n_{Th} total subjects in the test group and y_{Ch} is the observed number of events in the active control group for the *h*th stratum out of n_{Ch} total subjects in the active control group. The proportions $\tilde{p}_{Ch}$ and $\tilde{p}_{Th} = \theta_0 \tilde{p}_{Ch}$ are maximum likelihood estimates under the null hypothesis and can be computed in a closed-form manner by solving the equation $a_h \tilde{p}_{Ch}^2 + b_h \tilde{p}_{Ch} + c_h = 0$ where $a_h = (n_{Th} + n_{Ch})\theta_0$, $b_h = -\{(y_{Ch} + n_{Th})\theta_0 + y_{Th} + n_{Ch}\}$, and $c_h = (y_{Th} + y_{Ch})$. These estimates can also be obtain using SAS[2] through PROC GENMOD, a procedure used to fit generalized linear models, where there would be specification of a binomial distribution with a log link. The model statement fits only the intercept and includes an offset term where the offset for the control group is zero and the offset for the test group is the natural logarithm of the non-inferiority margin θ_0. This is done separately for each stratum to obtain the maximum likelihood estimates for the test and control groups. Gart[1] first presented this test statistic as a form from which to iteratively compute a confidence interval for the risk ratio. Gart also suggested that when the sample sizes in the treatment groups were unequal or when the observed proportions in the treatment groups were near 0 or 1, then the statistic may be skewed.

In 1988, Gart and Nam[3] extended this method to add a skewness correction to the test statistic seen in (4.2).

$$z_{GC}(\theta_0) = z_G(\theta_0) - \frac{\tilde{\gamma}(\theta_0)(z_\alpha^2 - 1)}{6} \tag{4.2}$$

where

$$\tilde{\gamma}(\theta_0) = \frac{\sum_h \left[(1-\tilde{p}_{Th})(1-2\tilde{p}_{Th})/(n_{Th}\tilde{p}_{Th})^2 - (1-\tilde{p}_{Ch})(1-2\tilde{p}_{Ch})/(n_{Ch}\tilde{p}_{Ch})^2\right]\left[v_h(\theta_0, \tilde{p}_{Ch})\right]^3}{\left[\sum_h v_h(\theta_0, \tilde{p}_{Ch})\right]^{3/2}}$$

This method was also designed to yield a confidence interval through iterative solutions using the test statistic in (4.2), and for this confidence interval, Gart and Nam suggest that it appropriately provides (1-2α)% coverage when the minimum cell count is ≥ 2.5.

Yanagawa, Tango, and Hiejima[4] presented a method to calculate a test statistic for the hypothesis of non-inferiority for the risk ratio in a stratified setting as seen in (4.3). This method is an extension to that presented by Farrington and Manning[5] for the unstratified setting.

$$z_{YTH}(\theta_0) = \frac{\sum_h \left(y_{Th} - n_{Th}\tilde{p}_{Th}\right)}{\left\{\sum_h \frac{n_{Th}n_{Ch}\tilde{p}_{Th}(1-\tilde{p}_{Th})^2}{\theta_0 n_{Th}(1-\tilde{p}_{Ch}) + n_{Ch}(1-\tilde{p}_{Th})}\right\}^{1/2}} \tag{4.3}$$

where $\tilde{p}_{Ch}$ and $\tilde{p}_{Th}$ are the maximum likelihood estimates under the null hypothesis as described previously. Yanagawa et. al. provide simulations of this test statistic for the three strata scenario which yields approximately nominal type I error rates, except when the sample sizes are small.

Although not evaluated as part of the current discussion, Miettinen and Nurminen[6] proposed a method that calculates a test statistic for the stratified risk ratio which follows a chi-square distribution. This method is iterative in nature and updates initial weights with better estimates as the iterations progress.

All of the methods described in (4.1) - (4.3) can be iteratively solved for values of θ that do not reject the null hypothesis, thus resulting in a lower and upper bound on θ which yields a confidence interval. Calculating these confidence intervals can be computationally intensive, especially as the number of strata increases.

Methods for the test of non-inferiority that produce confidence intervals directly include those produced by SAS in the FREQ procedure[2]. The first of these is based on a Mantel-Haenszel combined risk ratio across strata with a point estimate of $\theta_{MH} = \frac{\sum_h (y_{Th} n_{Ch})/(n_{Th} + n_{Ch})}{\sum_h (y_{Ch} n_{Th})/(n_{Th} + n_{Ch})}$ and the corresponding confidence interval seen in (4.4).

$$\{\theta_{MH} \exp(-z_\alpha \hat{\sigma}), \theta_{MH} \exp(z_\alpha \hat{\sigma})\} \tag{4.4}$$

where $\hat{\sigma}^2 = \hat{v}[\ln(\theta_{MH})] = \frac{\sum_h [n_{Th} n_{Ch} (y_{Th} + y_{Ch}) - y_{Th} y_{Ch} (n_{Th} + n_{Ch})]/(n_{Th} + n_{Ch})^2}{\left\{\sum_h \frac{y_{Th} n_{Ch}}{(n_{Th} + n_{Ch})}\right\}\left\{\sum_h \frac{y_{Ch} n_{Th}}{(n_{Th} + n_{Ch})}\right\}}$

The second of these methods is based on a Logit combined risk ratio across strata with a point estimate of $\theta_L = \exp\left\{\frac{\sum_h w_h \ln(\theta_h)}{\sum_h w_h}\right\}$ and the corresponding confidence interval seen in (4.5).

$$\left\{\theta_L \exp\left(\frac{-z_\alpha}{\sqrt{\sum_h w_h}}\right), \theta_L \exp\left(\frac{z_\alpha}{\sqrt{\sum_h w_h}}\right)\right\} \tag{4.5}$$

where $\theta_h = \frac{y_{Th} / n_{Th}}{y_{Ch} / n_{Ch}}$, $w_h = \frac{1}{v[\ln(\theta_h)]}$ and $v[\ln(\theta_h)] = \frac{1 - \hat{p}_{Th}}{y_{Th}} + \frac{1 - \hat{p}_{Ch}}{y_{Ch}}$

A modification of the Logit interval in (4.5) (termed Agresti method) uses proportions based on the Adapted Agresti method from Chapter 1 seen in (1.3) which adds additional counts to those observed distributed according to the null hypothesis (θ_0) and the allocation of sample size to each treatment group.

An additional modification of the Logit interval in (4.5) (termed ML Logit method) uses proportions obtained as the maximum likelihood estimates under the null hypothesis as described above as $\tilde{p}_{Ch}$ and $\tilde{p}_{Th}$ for the estimates in the variance, using

$$v\left[\ln(\theta_h)\right] = \frac{1-\tilde{p}_{Th}}{y_{Th}} + \frac{1-\tilde{p}_{Ch}}{y_{Ch}}.$$

A method for the stratified risk ratio is based on a Deviance statistic as twice the difference in the log likelihood values for likelihoods under the null and the alternative hypotheses. This can be implemented using PROC GENMOD in SAS[2] and a test statistic computed by subtracting the two deviances. Implementation using SAS includes fitting the model under the null hypothesis by specifying a binomial distribution with a log link. The model would include an intercept parameter and a parameter for strata with an offset for the control group that is zero and an offset for the test group that is the natural logarithm of the non-inferiority margin θ_0. Additionally the alternative hypothesis is fit by also specifying a binomial distribution with a log link. The model includes an intercept parameter and parameters for treatment and strata without any offset values specified. The difference in -2 Log Likelihood values between the two models is the value of the test statistic. This statistic is compared to the chi-square distribution with one degree of freedom to obtain a corresponding p-value. Additionally, a Wald test statistic can be calculated based on the parameter estimate and corresponding standard error produced by fitting the likelihood under the alternative hypothesis, using the parameter for the treatment effect.

B. Performance of Methods for the Stratified Risk Ratio

Simulations were used to study the properties of the methods in various scenarios to assess type I error and power. Scenarios for the stratified risk ratio include varying the following parameters:

1. H, the total number of strata: H=2
2. π_{Ch}, the population proportion of events in the control group: π_{C1}=0.05 – 0.20, π_{C2}=0.15 – 0.30, where $\pi_{C1} \leq \pi_{C2}$
3. $\theta_h=\pi_{Th}/\pi_{Ch}$, the population risk ratio: 0.667, 1.000, 1.500, 2.000
4. π_{Th}, the population proportion of events in the test group: $\pi_{Th}=\theta_h\pi_{Ch}$
5. θ_0, the null hypothesis risk ratio: 1.5, 1.75, 2.0
6. α, the one-sided alpha level: 0.0005, 0.005, 0.025
7. Sample size allocation for test:control = n_{Th}:n_{Ch} = 1:2, 1:1, 2:1
8. Sample size allocation for strata 1:strata 2 = N_1:N_2 = 1:3, 1:2, 2:3, 1:1, 3:2, 2:1, 3:1
9. $n_{C.}=n_{C1}+n_{C2}$, the total sample size across strata for the control group is calculated simplistically to have 85% power to contradict the null hypothesis θ_0 for equivalence of the average of the test and control groups across strata with $s_h=n_{Th}/(\ n_{Th}+n_{Ch})$, $\bar{\pi}_T=(\pi_{T1}+\pi_{T2})/2$, $\bar{\pi}_C=(\pi_{C1}+\pi_{C2})/2$ as

$$n_{C.}=\frac{(z_\alpha+z_\beta)^2\left\{\frac{1}{(s_h/(1-s_h))\bar{\pi}_T}+\frac{1}{\bar{\pi}_C}\right\}}{\{\ln(1/\theta_0)\}^2}$$

10. n_{Ch}, the sample size in the control group for the *h*th stratum: $n_{C1}=t_1n_{C.}$, $n_{C2}=(1-t_1)n_{C.}$ where $t_1=N_1/(N_1+N_2)$

11. n_{Th}, the sample size in the test group for the *h*th stratum: $n_{T1}=t_1 n_{T.}$, $n_{T2}=(1-t_1)n_{T.}$

For each combination of the parameters, 10,000 replications were generated using a random sample from the binomial distributions of $y_{Th} \sim bin(n_{Th}, \pi_{Th})$ and $y_{Ch} \sim bin(n_{Ch}, \pi_{Ch})$ for each of the h=1, 2 strata. For each replication, upper confidence limits or test statistics with corresponding p-values for the stratified risk ratio methods were calculated. If any of the event counts were equal to zero or the method failed to produce a valid result, then the Agresti method was used as the default because this method yields an upper confidence limit in all scenarios.

For each method, an indicator variable was created for each replication that is set equal to 1 if the upper confidence limit for the stratified risk ratio was less than the null hypothesis, θ_0, (or the p-value for the test was less than alpha) and set equal to 0 otherwise. This indicator was then averaged across the 10,000 replications to produce a probability. For scenarios where $\theta<\theta_0$, this probability is the power for the test of non-inferiority as the probability of rejecting the null hypothesis when it is false. For scenarios where $\theta=\theta_0$, this probability is the type I error rate for the test of non-inferiority as the probability of rejecting the null hypothesis when it is true. The power and type I error results are also summarized with respect to other parameters that were varied in the simulations.

Figures 4.1 – 4.6 summarize the simulated type I error rates for each of the methods where the sample size is allocated equally across the two strata for an alpha level of 0.025. Results are similar for other values of alpha, although these are not graphically summarized. All of the methods maintain the approximate nominal type I error level for the allocation of sample size to the treatment groups of Test:Control as 1:2. However, only the Gart-SC method and the Deviance method maintain the type I error for the 1:1 and 2:1 allocations as seen in Figure 4.1. The other methods have higher than nominal type I error rates in these situations. The Gart method and YTH method have higher than nominal type I errors in the 1:1 and 2:1 allocations, but these methods perform fairly well with slightly higher type I errors than the Gart-SC and Deivance methods.

The methods have type I errors closer to the nominal level for smaller null hypothesis risk ratios (i.e., θ_0=1.5) seen in Figure 4.2. As the total sample size increases, the type I error is closer to the nominal level (Figure 4.6). Due to this factor, the effect of the null hypothesis may be connected to the total sample size as less stringent null hypotheses require smaller sample sizes.

There is no effect of the control proportions on the type I error in either strata 1 (Figure 4.3) or strata 2 (Figure 4.4). However, as the difference in the control proportions across the strata increases, the type I error is closer to the nominal level (Figure 4.5).

Discussions of power will focus on situations where the type I error is controlled at the nominal level. Figure 4.7 compares the Gart-SC and Deviance methods where the population proportion in the treatment groups for each strata is equal (θ=1). These methods are very similar with respect to simulated power, with the Gart-SC method having slightly higher power for the 2:1 treatment allocation scenario.

Comparison of the Gart and YTH methods shows that the Gart method yields slightly higher power in all situations (Figure 4.8). The Gart method also yields higher power than the Gart-SC method (Figure 4.9), but only for the 1:1 and 2:1 allocation settings where the type I error of the Gart method is not quite controlled at the nominal level. For the 1:2 treatment allocation setting, the power for these methods is very similar.

The ML Logit method and the Mantel-Haenszel (MH) method have similar simulated power for the treatment allocation scenario of T:C as 1:2 (Figure 4.10) where the type I error is controlled at the nominal level. The Logit method yields higher power in this scenario compared to the Agresti method (Figure 4.11). However, the MH method has higher power than the Logit method (Figure 4.12). The Wald and MH methods are very similar, with the Wald method having slightly higher power for the 1:2 treatment allocation setting (Figure 4.13). Although these methods may be appropriate for the treatment allocation setting of 1:2, the Gart-SC method yields higher power than these methods even in this setting as is shown compared to the Wald method in Figure 4.14.

In many trials, it may not be feasible to enroll subjects equally among the strata. There may be smaller populations of subjects for one strata compared to another or there may be economical reasons for differential allocations. Even without any of the constraints previously discussed, differential allocations to the strata may have impacts on type I error or power for the stratified risk ratio methods. Figures 4.15 – 4.17 summarize the simulated type I error for the better methods including Gart, Gart-SC, YTH, and Deviance, for scenarios where the sample size is allocated differently across strata. These figures display this summary for treatment allocations of Test: Control as 1:2 (Figure 4.15), 1:1 (Figure 4.16), and 2:1 (Figure 4.17). The type I error is similar across strata allocation scenarios.

While the type I error is unaffected by strata allocation, the power depends directly on the allocation of sample size to the strata. The simulation scenarios were chosen so that the control proportion in the first strata was always equal to or smaller than the control proportion in the second strata ($\pi_{C1} \leq \pi_{C2}$). In this setting, it is obvious that as more sample size is placed in the strata with the larger control proportion, strata 2, the power increases.

Figures 4.18 – 4.20 display this graphically, separately for the treatment allocations of test:control as 1:2 (Figure 4.18), 1:1 (Figure 4.19), and 2:1 (Figure 4.20). When designing these non-inferiority trials, particular attention should be made to ensure that the sample size is allocated appropriately to the strata so as to maximize the power for a fixed number of subjects.

C. Sample Size Formulas for the Stratified Risk Ratio

Calculation of sample size for a specified level of power is an important aspect in designing non-inferiority trials. In the setting of a stratified analysis, it is important to understand the implications of adjusting for the strata when calculating sample sizes. Nam[7] discusses a sample size formula for the stratified risk ratio based on the score test as described by Gart[1] seen in (4.6).

$$N = \left[z_{\alpha} \left\{ \sum_{h=1}^{H} \frac{t_h s_h (1-s_h)\tilde{\pi}_{Th}}{s_h(\theta_0 - 1) + (1-\tilde{\pi}_{Th})} \right\}^{1/2} + z_{1-\beta} \left\{ \sum_{h=1}^{H} t_h s_h (1-s_h) \left(\frac{(1-\pi_{Ch})(1-\pi_{Th})}{E_h} \right)^2 \left(\frac{s_h \theta_0^2 \pi_{Ch}}{(1-\tilde{\pi}_{Th})^2 (1-\pi_{Ch})} + \frac{(1-s_h)\pi_{Th}}{(1-\tilde{\pi}_{Ch})^2 (1-\pi_{Th})} \right) \right\}^{1/2} \right]^2 \Bigg/ \left[\sum_{h=1}^{H} \left\{ t_h s_h (\pi_{Th} - \tilde{\pi}_{Th}) / (1-\tilde{\pi}_{Th}) \right\} \right]^2$$

(4.6)

where $E_h = 2\tilde{\pi}_{Th} - (1-\theta_0 \pi_{Ch}) + s_h(1-\pi_{Th}) - \theta_0(1-\pi_{Ch})$, $N = \sum_{h=1}^{H} (n_{Ch} + n_{Th})$,

$t_h = \frac{n_{.h}}{\sum_{h=1}^{H} n_{.h}}$, $n_{Th} = s_h t_h N$, and $n_{Ch} = (1-s_h)t_h N$. The proportions $\tilde{\pi}_{Th}$ and $\tilde{\pi}_{Ch}$ are the maximum likelihood estimates under the null hypothesis. These proportions can be obtained using PROC GENMOD in SAS^2 as similar to that described previously. The Nam sample size formula in (4.6) can also be solved for power as seen in (4.7).

$$z_{1-\beta} = \left[\sqrt{N} \left\{ \sum_{h=1}^{H} \left\{ t_h s_h (\pi_{Th} - \tilde{\pi}_{Th}) / (1 - \tilde{\pi}_{Th}) \right\} \right\} - z_\alpha \left\{ \sum_{h=1}^{H} \frac{t_h s_h (1 - s_h) \tilde{\pi}_{Th}}{s_h (\theta_0 - 1) + (1 - \tilde{\pi}_{Th})} \right\}^{1/2} \right] \Bigg/ \left\{ \sum_{h=1}^{H} t_h s_h (1 - s_h) \left(\frac{(1 - \pi_{Ch})(1 - \pi_{Th})}{E_h} \right)^2 \left(\frac{s_h \theta_0^2 \pi_{Ch}}{(1 - \tilde{\pi}_{Th})^2 (1 - \pi_{Ch})} + \frac{(1 - s_h) \pi_{Th}}{(1 - \tilde{\pi}_{Ch})^2 (1 - \pi_{Th})} \right) \right\}^{1/2} \tag{4.7}$$

The sample size formula from chapter 1 (1.15), based on the Taylor Series method is modified for stratified analysis seen in (4.8) where $n_{C.} = \sum_{h=1}^{H} n_{Ch}$, and $s_h = \dfrac{n_{Th}}{n_{Th} + n_{Ch}}$.

$$n_{C.} = \frac{(z_\alpha + z_{1-\beta})^2 \left\{ \sum_{h=1}^{H} t_h \left(\frac{1}{(s_h / (1 - s_h)) \pi_{Th}} + \frac{1}{\pi_{Ch}} \right) \right\}}{\left\{ \sum_{h=1}^{H} t_h^{\,2} \ln\left(\frac{\theta_h}{\theta_0} \right) \right\}^2} \tag{4.8}$$

The sample size for each treatment group in each strata can be calculated in the following manner: $n_{Th} = t_h R n_{C.}$ and $n_{Ch} = t_h n_{C.}$ This Taylor Series formula in (4.8) can be solved for power as seen in (4.9) where power=$\Phi(z_{1-\beta})$ and $z_{1-\beta}$ is the (1-β) quantile of the standard normal distribution.

$$z_{1-\beta} = \frac{\sqrt{n_{C.}} \left\{ \sum_{h=1}^{H} t_h^{\,2} \ln\left(\frac{\theta_h}{\theta_0} \right) \right\}}{\left\{ \sum_{h=1}^{H} t_h \left(\frac{1}{(s_h / (1 - s_h)) \pi_{Th}} + \frac{1}{\pi_{Ch}} \right) \right\}^{1/2}} - z_\alpha \tag{4.9}$$

It is important that the sample size formulas used to design the stratified risk ratio trials to assess non-inferiority are operating at the specified power level. The Nam formula and the Taylor Series formula will be compared to the simulated powers of the methods to assess the appropriateness of their use in designing these trials.

The Gart-SC method yields nominal type I error rates and fairly high power compared to the other methods. A comparison of the simulated power from the Gart-SC method and the Nam calculated power is presented for treatment allocation of test:control as 1:2 (Figure 4.21), 1:1 (Figure 4.22), and 2:1 (Figure 4.23). The Gart-SC simulated power is higher than the Nam calculated power for the 1:2 scenario and approximately equal for the 1:1 scenario. However, the Gart-SC simulated power is slightly lower than the Nam calculated power for the 2:1 situation, and more so as more sample size is allocated to strata 1 with the smaller control proportion. This difference in calculated and simulated power is not large, especially for situations designed to have fairly high power.

The Gart simulated power is always slightly larger than the Nam calculated power for all treatment allocation scenarios (Figures 4.24 – 4.26). However, the Gart method may not maintain the nominal type I error as well for the 1:1 and 2:1 settings.

Calculated power for the Taylor Series method is displayed in Figures 4.27 – 4.29 compared to the Gart-SC simulated power. The Taylor Series sample size formula is much simpler than the Nam formula. However, it is conservative and requires more sample size than necessary for the specified power level when the difference in control proportions for the strata is larger (>0.20). When (π_{C2}-π_{C1}) is smaller, this method has a lower calculated power than the Nam simulated power and may not provide enough sample size for the necessary power.

III. Assessing Non-inferiority of a Risk Difference in a Stratified Trial

A. Methods for the Stratified Risk Difference

The hypotheses for testing non-inferiority of the test treatment compared to the active control treatment for the risk difference rely on a pre-determined non-inferiority limit, Δ_0 which is seen in $H_0 : \Delta_h = \pi_{Th} - \pi_{Ch} \leq \Delta_0$ and H_A: $\Delta_h = \pi_{Th} - \pi_{Ch} > \Delta_0$ where $\Delta_h = \pi_{Th} - \pi_{Ch}$ is the population risk difference for the test group versus the active control group in the *h*th stratum, with *h*=1, 2, …, H and π_{Th} and π_{Ch} representing the population proportion of patients with the favorable outcome of interest in the test and active control groups, respectively, for the *h*th stratum.

Similar to the setting where the risk ratio is of interest, when the risk difference is of primary interest for analysis, a test statistic can be formulated to produce a p-value for the null hypothesis of inferiority. Iterative methods can be used to compute a corresponding confidence interval which includes all values for which the null hypothesis would not be rejected at the specified alpha level.

Gart and Nam[8] present a method based on a standard normal statistic that is computed as in (4.10) where y_{Th} is the observed number of events in the test group for the *h*th stratum out of n_{Th} total subjects in the test group and y_{Ch} is the observed number of events in the active control group for the *h*th stratum out of n_{Ch} total subjects in the active control group

$$z_{GN}(\Delta_0) = \frac{\sum_h (y_{Th} - n_{Th}\tilde{p}_{Th}) / \tilde{v}_{Th}}{\tilde{V}^{1/2}} \tag{4.10}$$

where $\tilde{V} = \sum_{h} \tilde{V}_{h}$, $\tilde{V}_{h} = \left\{ \frac{\tilde{v}_{Th}}{n_{Th}} + \frac{\tilde{v}_{Ch}}{n_{Ch}} \right\}^{-1}$, $\tilde{v}_{Th} = \tilde{p}_{Th}(1-\tilde{p}_{Th})$, and $\tilde{v}_{Ch} = \tilde{p}_{Ch}(1-\tilde{p}_{Ch})$

with the proportions $\tilde{p}_{Ch}$ and $\tilde{p}_{Th}$ computed as maximum likelihood estimates of the proportions under the null hypothesis with closed-form solutions defined as

$\tilde{p}_{Th} = 2u_h \cos(w_h) - \frac{b_h}{3a_h}$ and $\tilde{p}_{Ch} = \tilde{p}_{Th} - \Delta_0$ with the following components:

$$a_h = 1 + \frac{n_{Ch}}{n_{Th}}, \quad b_h = -\left\{ 1 + \frac{n_{Ch}}{n_{Th}} + \hat{p}_{Th} + \frac{n_{Ch}}{n_{Th}} \hat{p}_{Ch} + \Delta_0 \left(\frac{n_{Ch}}{n_{Th}} + 2 \right) \right\},$$

$$c_h = \Delta_0{}^2 + \Delta_0 \left(2\hat{p}_{Th} + \frac{n_{Ch}}{n_{Th}} + 1 \right) + \hat{p}_{Th} + \frac{n_{Ch}}{n_{Th}} \hat{p}_{Ch}, \quad d_h = -\hat{p}_{Th}\Delta_0(1+\Delta_0),$$

$$u_h = sign(v_h) \left\{ \frac{b_h{}^2}{(3a_h)^2} - \frac{c_h}{3a_h} \right\}^{1/2}, \quad v_h = \frac{b_h{}^3}{(3a_h)^3} - \frac{b_h c_h}{6a_h{}^2} + \frac{d_h}{2a_h}, \quad w_h = \frac{1}{3}\left\{ \Pi + \cos^{-1}\left(\frac{v_h}{u_h{}^3} \right) \right\}$$

These estimates can also be obtained using PROC GENMOD in SAS[2] with specification of a binomial distribution with a identity link. The model statement fits only the intercept and includes an offset term where the offset for the control group is zero and for the test group is set equal to the specified non-inferiority margin Δ_0. This model is fit separately for each stratum to produce maximum likelihood estimates for the test and control groups.

Gart and Nam[8] propose a skewness-corrected version of the test statistic in (4.10) as seen in (4.11), which should reduce the skewed nature of the corresponding confidence interval from (4.10).

$$z_{GNC}(\Delta_0) = z_{GN}(\Delta_0) - \tilde{\gamma}(\Delta_0)(z_\alpha^2 - 1)/6 \tag{4.11}$$

where $\tilde{\gamma}(\Delta_0) = \frac{\sum_h \tilde{\mu}_{3h}(\Delta_0)}{\tilde{V}^{3/2}}$ and $\tilde{\mu}_{3h}(\Delta_0) = \tilde{V}_h^3 \left\{ \tilde{v}_{Th}(1-2\tilde{p}_{Th})/n_{Th}^2 - \tilde{v}_{Ch}(1-2\tilde{p}_{Ch})/n_{Ch}^2 \right\}$

Yanagawa, Tango, and Hiejima[4] provide a test statistic for the risk difference seen in (4.12) which can be compared to a standard normal distribution to obtain a p-value for comparison against the specified alpha level.

$$z_{YTH} = \frac{\sum_h \left[y_{Th} - n_{Th}\tilde{p}_{Th}\right]}{\left\{\sum_h \frac{n_{Th}n_{Ch}\tilde{p}_{Th}{}^2(1-\tilde{p}_{Th})^2}{n_{Th}\tilde{p}_{Ch}(1-\tilde{p}_{Ch}) + n_{Ch}\tilde{p}_{Th}(1-\tilde{p}_{Th})}\right\}^{1/2}} \tag{4.12}$$

This test statistic also uses the proportions $\tilde{p}_{Th}$ and $\tilde{p}_{Ch}$ calculated from score equations as described previously for the method by Gart and Nam in (4.10). This method is a stratified extension to that proposed by Farrington and Manning[5] and studied in chapter 2 on the risk difference. Yanagawa et. al.[4] provide limited simulation results which suggest this method controls the type I error at approximately the nominal level except for scenarios with small sample sizes.

A method for the stratified risk difference is based on a Deviance statistic as twice the difference in the log likelihood values for likelihoods under the null and alternative hypotheses. This can be implemented using PROC GENMOD in SAS[2] and a test statistic computed by subtracting the two deviances. Implementation using SAS includes fitting the model under the null hypothesis by specifying a binomial distribution with an identity link. The model would include an intercept parameter and a parameter for strata with an offset for the control group that is zero and an offset for the test group that is the non-inferiority margin Δ_0. Additionally the alternative hypothesis is fit by also specifying a binomial distribution with an identity link. The model includes an intercept parameter and parameters for treatment and strata without any offset values specified. The difference in -2 Log Likelihood values between the two models is the value of the test statistic. This statistic is compared to the chi-square distribution with one degree of freedom to obtain a corresponding p-value. An additional method uses the parameter estimate for treatment from the model fit under the alternative hypothesis to create a Wald statistic.

While not implemented in the current discussion, Miettinen and Nurminen[6] propose an iterative method for producing a test statistic for the risk difference which is similar to that mentioned for the risk ratio and follows a chi-square distribution. This process is computer intensive and requires more resources for simply computing the test statistic than those mentioned above.

The methods in (4.10) – (4.12) also require iterative methods if a confidence interval is desired in addition to the test statistic. A confidence interval can be computed for each of these methods by finding values of Δ for which the test statistic fails to reject the null hypothesis of inferiority for the specified alpha level.

O'Gorman et. al.[9] compare two methods of producing confidence intervals for the stratified setting. The first of these is based on a weighted least squares methodology for computing the weights as seen in (4.13), originally described by Kleinbaum, Kupper, and Morganstern[10].

$$d_{WLS} \pm \frac{z_\alpha}{\left\{\sum_h W_{WLS,h}\right\}^{1/2}} \tag{4.13}$$

where $\mathrm{d_{WLS}} = \dfrac{\sum\limits_{\mathrm{h}} \mathrm{W_{WLS,h} d_h}}{\sum\limits_{\mathrm{h}} \mathrm{W_{WLS,h}}}$, $\mathrm{d_h} = \mathrm{\hat{p}_{Th}} - \mathrm{\hat{p}_{Ch}}$, and $W_{WLS,h} = \left\{\dfrac{\hat{p}_{Th}(1-\hat{p}_{Th})}{n_{Th}} + \dfrac{\hat{p}_{Ch}(1-\hat{p}_{Ch})}{n_{Ch}}\right\}^{-1}$.

O'Gorman et. al.[9] also present a confidence interval using Cochran-Mantel-Haenszel weights as seen in (4.14), which was presented by Cochran[11].

$$d_{CMH} \pm z_{\alpha} \frac{\left\{\sum_{h} L_h\right\}^{1/2}}{\left\{\sum_{h} W_{CMH,h}\right\}} \tag{4.14}$$

where $d_{CMH} = \frac{\sum_{h} W_{CMH,h} d_h}{\sum_{h} W_{CMH,h}}$, $W_{CMH,h} = \frac{n_{Th} n_{Ch}}{n_{Th} + n_{Ch}}$, and

$$L_h = \frac{\left\{y_{Th}(n_{Th} - y_{Th}) n_{Ch}^3 + y_{Ch}(n_{Ch} - y_{Ch}) n_{Th}^3\right\}}{\left\{n_{Th} n_{Ch}(n_{Th} + n_{Ch})^2\right\}}$$

O'Gorman et. al. recommended the CMH weights over the WLS weights for computing the confidence interval for a stratified risk difference because the CMH method showed approximate nominal coverage while the WLS method varied widely in its coverage probabilities for the entire confidence interval. This recommendation is based on simulation results for scenarios using at least 8 strata and small proportions less than 0.10.

The intervals in (4.13) for the WLS method can be modified to extend the Agresti and Caffo method[12] for the unstratified setting for the risk difference seen in chapter 2 (2.3). This interval replaces the observed event rates in (4.13) with rates which add one success and one failure to each treatment group with $p`_{Th}=(y_{Th}+1)/(n_{Th}+2)$ and $p`_{Ch}=(y_{Ch}+1)/(n_{Ch}+2)$.

Sato[13] proposes a method based on the CMH interval in (4.14) which yields a confidence interval by using a Fieller-type method to obtain the lower and upper limits as seen in (4.15). There is no assessment of the performance of this interval within Sato's discussion.

$$\left(\frac{2C_-}{B_- + \sqrt{B_-^2 - 4AC_-}}, \frac{B_+ + \sqrt{B_+^2 - 4AC_+}}{2A} \right) \tag{4.15}$$

where $A = W_{CMH}^2 + z_\alpha^2 \sum_h \frac{n_{Th}^3 n_{Ch}}{(n_{Th}+n_{Ch})^2(n_{Th}+n_{Ch}-1)}$,

$$B_\pm = 2\left[W_{CMH} d_{CMH} \pm \frac{1}{2}\right] W_{CMH} - z_\alpha^2 \sum_h \frac{n_{Th}^2 n_{Ch}\left[n_{Th}+n_{Ch}-2(y_{Th}+y_{Ch})\right]}{(n_{Th}+n_{Ch})^2(n_{Th}+n_{Ch}-1)},$$

$C_\pm = \left\{W_{CMH} d_{CMH} \pm \frac{1}{2}\right\}^2 - z_\alpha^2 V_{CMH}$, $\mathrm{W}_{\mathrm{CMH}} = \sum_{\mathrm{h}} \frac{\mathrm{n}_{\mathrm{Th}} \mathrm{n}_{\mathrm{Ch}}}{(\mathrm{n}_{\mathrm{Th}} + \mathrm{n}_{\mathrm{Ch}})}$, and

$$\mathrm{V}_{\mathrm{CMH}} = \sum_{\mathrm{h}} \frac{\mathrm{n}_{\mathrm{Th}} \mathrm{n}_{\mathrm{Ch}} (\mathrm{y}_{\mathrm{Th}} + \mathrm{y}_{\mathrm{Ch}})(\mathrm{n}_{\mathrm{Th}} + \mathrm{n}_{\mathrm{Ch}} - \mathrm{y}_{\mathrm{Th}} - \mathrm{y}_{\mathrm{Ch}})}{(\mathrm{n}_{\mathrm{Th}} + \mathrm{n}_{\mathrm{Ch}})^2 (\mathrm{n}_{\mathrm{Th}} + \mathrm{n}_{\mathrm{Ch}} - 1)}$$

B. Performance of Methods for the Stratified Risk Difference

Simulations were used to study the properties of the methods in various scenarios to assess type I error and power. Scenarios for the stratified risk difference include varying the following parameters:

1. H, the total number of strata: H=2
2. π_{Ch}, the population proportion of events in the control group: π_{C1}=0.60 – 0.75, π_{C2}=0.70 – 0.95, where $\pi_{C1} \leq \pi_{C2}$
3. $\Delta_h = \pi_{Th} - \pi_{Ch}$, the population risk difference: Δ_0, $\Delta_0/2$, 0, 0.025, 0.05
4. π_{Th}, the population proportion of events in the test group: $\pi_{Th} = \pi_{Ch} + \Delta_h$
5. Δ_0, the null hypothesis risk difference: -0.10, -0.075, -0.05
6. α, the one-sided alpha level: 0.0005, 0.005, 0.025
7. Sample size allocation for test:control = $n_{Th}:n_{Ch}$ = 1:2, 1:1, 2:1
8. Sample size allocation for strata 1:strata 2 = $N_1:N_2$ = 1:3, 1:2, 2:3, 1:1, 3:2, 2:1, 3:1

9. $n_{C.}=n_{C1}+n_{C2}$, the total sample size across strata for the control group is calculated to have 85% power to contradict the null hypothesis Δ_0 for equivalence of the weighted average of the test and control groups across strata with $s_h=n_{Th}/(\ n_{Th}+n_{Ch})$, $\bar{\pi}_T = (\pi_{T1} + \pi_{T2})/2$, $\bar{\pi}_C = (\pi_{C1} + \pi_{C2})/2$ as

$$n_{C.} = \frac{(z_\alpha + z_\beta)^2([s_h/(1-s_h)]+1)\bar{\pi}_C(1-\bar{\pi}_C)}{[s_h/(1-s_h)]\Delta_0^2}$$

10. n_{Ch}, the sample size in the control group for the *h*th stratum: $n_{C1}=t_1n_{C.}$, $n_{C2}=(1\text{-}t_1)n_{C.}$ where $t_1=N_1/(N_1+N_2)$
11. n_{Th}, the sample size in the test group for the *h*th stratum: $n_{T1}=t_1n_{T.}$, $n_{T2}=(1\text{-}t_1)n_{T.}$

For each of the 10,000 replications performed, a sample from the specified binomial distribution as $y_{Th} \sim \text{bin}(n_{Th}, \pi_{Th})$ and $y_{Ch} \sim \text{bin}(n_{Ch}, \pi_{Ch})$ for each of the h=1, 2 strata was generated separately. The confidence limits and p-values for the methods were computed for the same replication and a conclusion of non-inferiority or not was determined according to whether the one-sided lower confidence limit exceeded the specified non-inferiority margin or similarly if the p-value was below the alpha level. The average of the zero or one indicator variables for demonstration of non-inferiority or not resulted in a probability. This probability corresponds to the power of the methods when the population difference in proportions is better than the non-inferiority margin. The probability results in a type I error when the specified population difference in proportions is equal to or below the non-inferiority margin. If the number of events in any of the groups was equal to zero or the method failed to produce a logical result, then the Agresti method was implemented because this method yields a result in all scenarios. The simulated type I error and power will be summarized, specifically to reflect the effect of varying parameters in the different scenarios of the simulations.

The simulated type I error for an alpha level of 0.025 is summarized in Figures 4.30 – 4.35. Results are similar for other alpha levels, although these are not graphically displayed. The performance of the methods for the stratified risk difference does vary according to the treatment allocation as seen in Figure 4.30. The Gart & Nam, Gart & Nam-SC, YTH, and Deviance methods perform at approximately the nominal type I error level for the 1:2 treatment allocation setting. These methods also perform fairly well in the 1:1 setting, but the other methods also have closer to nominal simulated type I error rates. For the 2:1 treatment allocation setting, the WLS, Deviance, and Wald methods have type I errors closer to the nominal level.

The simulated type I error of the methods becomes less variable and closer to the nominal level for smaller (more stringent) non-inferiority difference margins as seen in Figure 4.31. This may be a consequence of the larger sample sizes required for testing smaller non-inferiority differences. As seen in Figure 4.35, as the total sample size increases, the methods are better at achieving the nominal type I error.

Although there do not appear to be differences in performance of the methods for the control proportions in strata 1 (Figure 4.32), it appears that smaller control proportions in strata 2 (Figure 4.33) have closer to nominal type I error rates. There do appear to be performance discrepancies related to the difference in the control proportions in the strata (π_{C2}-π_{C1}) seen in Figure 4.34. As the difference between control proportions in the two strata decreases, the type I error of the methods is closer to the nominal level.

Discussions of power will focus on situations where the type I error is controlled at the nominal level. Overall, the Gart & Nam, Gart & Nam-SC, YTH, and Deviance methods tend to control type I error fairly well across all scenarios. Comparison of the Gart & Nam method to the Gart & Nam-SC method in Figure 4.36 shows very similar simulated power for the 1:1 treatment allocation setting. However, the Gart & Nam method has slightly higher power for the 2:1 setting and the Gart & Nam-SC method has slightly higher power in the 1:2 setting. The Gart & Nam method has consistently higher power than the YTH method (Figure 4.37), especially as the control proportion in strata 2 (π_{C2}) increases. The Gart & Nam method compared to the Deviance method yields similar results (Figure 4.38) as with the Gart & Nam-SC method, where the Deviance method has higher power in the 1:2 treatment allocation setting. Figure 4.39 compares the Gart & Nam-SC method to the Deviance method, and suggests that the Deviance method has slightly higher power in the 1:2 treatment allocation setting.

The CMH method tends to have appropriate type I error in the 1:1 and 2:1 treatment allocation settings. This method is compared to the Gart & Nam method in this setting in Figure 4.40. The CMH power is lower than the Gart & Nam power, especially as π_{C2} increases.

In the 2:1 treatment allocation setting, the Gart & Nam method yields higher power compared to the WLS method (Figure 4.41) and the Wald method (Figure 4.42), but the Gart & Nam method has type I errors that are slightly higher than nominal level in this situation. The Wald method yields slightly higher power than the WLS method in this 2:1 setting (Figure 4.43).

As has just been shown, the methods have different performance depending on the treatment allocation. In addition, the methods may be affected by the sample size allocation to each of the strata. Figures 4.44 – 4.46 summarize the simulated type I error for the methods which perform better for the treatment allocation scenarios, with the 1:2 allocation in Figure 4.44 for the Gart & Nam, Gart & Nam-SC, and Deviance methods; the 1:1 allocation in Figure 4.45 for the Gart & Nam, Gart & Nam-SC, CMH, and Deviance methods; and the 2:1 allocation in Figure 4.46 for the Gart & Nam, WLS, CMH, and Wald methods. The type I error is fairly similar across strata allocation scenarios and is similar to that already summarized for the 1:1 strata allocation scenario.

However, the power does depend on the allocation of sample size to the strata. The simulation scenarios were chosen so that the control proportion in the first strata was always equal to or smaller than the control proportion in the second strata ($\pi_{C1} \leq \pi_{C2}$). As more sample size is placed in this second strata with the higher proportion, the power increases although this increase is not as marked as in the stratified risk ratio setting, except for the CMH method. Figures 4.47 – 4.49 summarize these scenarios as similar to that summarized for the type I error rates across the strata allocation settings. Yet, the power does not seem to be as dependent on the allocation to the strata as in the risk ratio setting. This difference in power is more obvious between settings such as 1:2 versus 2:1 for the strata allocation, but is not quite as distinct for small increases in sample size to the second strata as in going from 1:2 to 1:3 allocation settings. The power displayed in these figures has a fairly wide range and is dependent on the proportion in the second strata. As this proportion increases, the power increases.

C. Sample Size Formulas for the Stratified Risk Difference

Nam[14] developed a stratified sample size formula based on the risk difference from the score test described by Gart and Nam[8] in (4.10). This sample size formula is seen in (4.16) for $N = \sum_{h=1}^{H}(n_{Th} + n_{Ch})$ where $n_{Th} = s_h t_h N$ and $n_{Ch} = (1 - s_h)t_h N$.

$$N = \frac{\left\{z_\alpha v_0^{1/2} + z_{1-\beta} v_1^{1/2}\right\}^2}{r^2} \tag{4.16}$$

where $r = \sum_{h=1}^{H}\left\{\frac{t_h s_h(\pi_{Th} - \tilde{\pi}_{Th})}{\tilde{\pi}_{Th}(1 - \tilde{\pi}_{Th})}\right\}$, $v_0 = \sum_{h=1}^{H}\left\{\frac{t_h s_h(1 - s_h)}{s_h \tilde{\pi}_{Ch}(1 - \tilde{\pi}_{Ch}) + (1 - s_h)\tilde{\pi}_{Th}(1 - \tilde{\pi}_{Th})}\right\}$,

$$v_1 = \sum_{h=1}^{H}\left\{\frac{t_h s_h(1 - s_h)}{\frac{s_h\left[\tilde{\pi}_{Ch}(1 - \tilde{\pi}_{Ch})\right]^2}{\pi_{Ch}(1 - \pi_{Ch})} + \frac{(1 - s_h)\left[\tilde{\pi}_{Th}(1 - \tilde{\pi}_{Th})\right]^2}{\pi_{Th}(1 - \pi_{Th})}}\right\},\ s_h = \frac{n_{Th}}{n_{Th} + n_{Ch}},\text{ and } t_h = \frac{n_{Th} + n_{Ch}}{N}$$

Nam compares this sample size formula to an analogous formula ignoring stratification. The results suggest that ignoring strata in the design of a trial results in an overestimate of sample size. The sample size formula in (4.16) can be solved for power as seen in (4.17) where power = $\Phi(z_{1-\beta})$ is the (1-β) quantile of the standard normal distribution.

$$z_{1-\beta} = \frac{\sqrt{N}r - z_\alpha v_0^{1/2}}{v_1^{1/2}} \tag{4.17}$$

A sample size formula can be obtained as an extension to the Wald sample size formula (2.26) presented in chapter 2 on the risk difference. This sample size formula is seen in (4.18) for $n_{T.} = \sum_{h=1}^{H} n_{Th}$.

$$n_{T.} = \frac{\left(z_\alpha + z_{1-\beta}\right)^2 \left\{ \sum_{h=1}^{H} t_h \pi_{Th}(1-\pi_{Th}) + t_h[s_h/(1-s_h)]\pi_{Ch}(1-\pi_{Ch}) \right\}}{\left\{ \sum_{h=1}^{H} t_h(\pi_{Th} - \pi_{Ch} - \Delta_0) \right\}^2} \tag{4.18}$$

The sample size formula in (4.18) can be solved for power seen in (4.19) where power = $\Phi(z_{1-\beta})$ is the (1-β) quantile of the standard normal distribution.

$$z_{1-\beta} = \frac{\sqrt{n_{T.}} \left\{ \sum_{h=1}^{H} t_h(\pi_{Th} - \pi_{Ch} - \Delta_0) \right\}}{\left\{ \sum_{h=1}^{H} t_h \pi_{Th}(1-\pi_{Th}) + t_h[s_h/(1-s_h)]\pi_{Ch}(1-\pi_{Ch}) \right\}^{1/2}} - z_\alpha \tag{4.19}$$

These sample size formulas can only be useful in the design of a non-inferiority stratified risk difference trial if the resulting sample size yields a power similar to that specified in the calculations. The Nam formula and the Wald formula for sample size calculations are compared to the simulated powers of select methods to assess their appropriateness for use in the design of these trials.

The Gart & Nam method for assessing the stratified risk difference is compared to the Nam calculated power obtained from (4.17) in Figure 4.50 by treatment allocation and in Figure 4.51 by the control proportion in strata 2. The simulated and calculated power of these methods agrees very closely, with slight variation but of small magnitude with increasing power for increasing π_{C2}. The Deviance simulated power is also compared to the Nam calculated power in Figure 4.52 by treatment allocation and in Figure 4.53 for values of π_{C2}. These methods also agree very closely.

The Wald method controls the type I error slightly better in the 2:1 treatment allocation setting than the Gart & Nam method. The Wald simulated power is compared to the Nam calculated power in Figure 4.54, but in the 2:1 treatment allocation setting the Wald power is slightly lower than this Nam calculated power. This suggests that sample size may need to be slightly increased in a 2:1 allocation scenario in order to yield the appropriate power when using the Wald method for assessing the stratified risk difference.

The simpler Wald sample size calculation in (4.18) with the power calculation in (4.19) is compared to the Gart & Nam method in Figure 4.55 and to the Wald method in Figure 4.56 by the control proportion in strata 2. The Gart & Nam power is generally higher than the Wald calculated power, especially as π_{C2} decreases. This Wald formula can yield large discrepancies when compared to the simulated power and my not be operating at the desired power specified in the calculations.

IV. Implications for Overall Significance with Conditions on Individual Strata

The methods for the stratified risk ratio and stratified risk difference do not include any verification of homogeneity of effects across the strata. This homogeneity is an important consideration, especially if the strata have differing treatment proportions. Trial design may include two strata for males and females or in an anti-infective setting the strata may be two different strains of the bacteria, one which has developed resistance and the other which has not. In these settings, it is important to show non-inferiority overall but also to show that each strata is also trending in the correct direction. Additionally, regulatory agencies may require these conditions on the strata to ensure the overall effect is appropriate. This issue is addressed by requiring the test of non-inferiority in the individual strata to meet a larger, but trending alpha level along with the stratified test of non-inferiority across strata meeting a smaller alpha level. The effects of adding these additional criteria for each of the strata will be assessed related to the type I error and power overall including the individual and stratified tests.

Simulations for the stratified risk ratio are similar to those already presented, but with more limited scenarios using 10,000 replications. Specifically, strata allocations include only the 1:3, 1:1, and 3:1 setting. Additionally, the population risk ratio is set at equality (θ=1). The stratified alpha level is set at 0.025 with the individual strata alpha levels at 0.05, 0.10, and 0.15. Focus of this discussion will include assessment of the stratified risk ratio using the Gart-SC method and the individual strata assessments made using the power divergence method, λ=0.5 as explained in chapter 1 (1.10).

Figures 4.57 – 4.59 summarize the type I error for these scenarios for each of the treatment allocation settings. The figures display the type I error for the Gart-SC method without requiring that each individual strata meet the specified additional alpha level. These type I errors are similar to those already presented and are maintained at approximately the nominal level (in this scenario α=0.025). However, with the additional conditions placed on the individual strata, the overall alpha level drastically decreases below the nominal α=0.025. As is expected, as the alpha level on the individual strata becomes more stringent, the type I error is lower. This suggests that when these side conditions are required for the individual strata, the alpha level for the stratified test could be increased over the nominal level to still maintain control of the type I error. For example, the stratified test alpha level could be set slightly greater than α=0.025 with the individual strata side conditions, and the overall type I error would still be maintained at α=0.025.

Placing these additional criteria on the individual strata also results in a reduction in overall power, which is smaller as the alpha level for the individual strata decreases. The overall power with the additional criteria is compared to the power without these criteria for the Gart-SC method in Figure 4.60 for α=0.05 for the individual strata, Figure 4.61 for α=0.10 for the individual strata, and Figure 4.62 for α=0.15 for the individual strata by the strata allocation and differences between the control proportions in each of the strata. As the difference in proportions between control groups for the strata increases, the power with the side conditions is much smaller than the power without these conditions regardless of choice of strata allocation. However, it is much worse for the strata allocation which places fewer subjects in the strata with the larger proportion (and therefore the larger variance) as seen for strata allocations of 3:1 (because in the simulations $\pi_{C1} \leq \pi_{C2}$). The effect on power of requiring these additional criteria can be somewhat mitigated by allocating more sample size to the strata with the larger control proportion as seen for the strata allocation of 1:3. In this setting, power for the setting which requires this additional criteria on the individual strata is similar for even larger differences in the control groups to the power with the additional criteria for a 1:1 strata allocation setting with small differences between the control groups. As is expected, the power with the additional criteria becomes much smaller than the setting without these criteria as the alpha level for the individual strata decreases.

The overall power with the side conditions on the individual strata may be increased by allowing the test of the stratified risk ratio to be performed at a slightly higher alpha level because the method is operating at a much lower overall type I error than the nominal α=0.025, when the side conditions on the strata are added.

These additional criteria for the individual strata were also studied for the stratified risk difference. Simulations used were similar to those previously described for the stratified risk difference, but with fewer scenarios including only strata allocations of 1:3, 1:1, and 3:1, an overall alpha level of 0.025, and with the criteria for the individual strata of α=0.05, 0.10, and 0.15 for the 10,000 replications. The Gart & Nam-SC method was used for the test of the overall stratified risk difference. The Farrington-Manning 3 method (2.12) from chapter 3 on the risk difference was used for the test of non-inferiority for the individual strata.

Results for the stratified risk difference with the additional criteria are summarized in Figures 4.63 – 4.65 for the type I error for the Gart & Nam-SC method without these criteria and with the criteria for the different alpha levels. Again, the overall type I error with these criteria is much lower than the 0.025 level.

The simulated power is compared for the setting with and without the side conditions on the individual strata in Figure 4.66 for α=0.05 for the individual strata, Figure 4.67 for α=0.10 for the individual strata, and Figures 4.68 for α=0.15 for the individual strata. These figures are summarized by the strata allocation scenarios and the differences between control proportions in each of the strata. Again, it is seen that with lower alpha levels required for the individual strata the power decreases drastically. However, allocation of sample size to the strata yields similar power for the 1:3 and 1:1 settings with lower power for the 3:1 setting. The power with the additional criteria on the strata is not as affected (i.e., more closely agrees with the stratified power without the conditions) when more sample size has allocation to strata 1 but is much more affected when more sample size has allocation to strata 2 as in the 1:3 allocation.

V. Discussion

Methods for the analysis of the risk ratio and the risk difference in a stratified setting including two strata have been reviewed. The performance of these methods is dependent on the treatment allocation, the control proportion, and the overall sample size. Performance of these methods can be improved if more sample size is allocated to the strata with the larger influence on the applicable variance.

Sample size formulas for these stratified settings have also been identified and assessed for similarities with the simulated power of the proposed methods Such evaluation addresses the need for statisticians to be able to appropriately plan and power these stratified non-inferiority trials.

The issue of confirming homogeneity across the strata is addressed by adding side conditions that the test of non-inferiority in the individual strata also be significant at a trending alpha level. Adding these additional criteria reduces the overall type I error below the nominal level used to perform the stratified test. Increasing the alpha level for the stratified test could be allowable while still maintaining an overall type I error at the nominal level.

References

1. Gart J. Approximate Tests and Interval Estimation of the Common Relative Risk in the Combination of 2x2 Tables. Biometrika **1985**, 72(3): 673-677.

2. *SAS OnlineDoc®, Version 8*. SAS Institute Inc.: Cary, NC, 1999.

3. Gart J., and Nam J. Approximate Interval Estimation of the Ratio of Binomial Parameters: A Review and Corrections for Skewness. Biometrics **1988**, 44: 323-338.

4. Yanagawa T., Tango T., and Hiejima Y. Mantel-Haenszel-Type Tests for Testing Equivalence or More than Equivalence in Comparative Clinical Trials. Biometrics **1994**, 50: 859-864.

5. Farrington, C. P., and Manning G. Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-zero Risk Difference or Non-unity Relative Risk. Statistics in Medicine **1990**, 9: 1447-1454.

6. Miettinen O., and Nurminen M. Comparative Analysis of Two Rates. Statistics in Medicine **1985**, 4: 213-226.

7. Nam J. Sample Size Requirements for Stratified Prospective Studies with Null Hypothesis of Non-unity Relative Risk using the Score Test. Statistics in Medicine **1994**, 13: 79-86.

8. Gart J., and Nam J. Approximate Interval Estimation of the Difference in Binomial Parameters: Correction for Skewness and Extension to Multiple Tables. Biometrics **1990**, 46: 637-643.

9. O'Gorman T., Woolson R., and Jones M. A Comparison of Two Methods of Estimating a Common Risk Difference in a Stratified Analysis of a Multicenter Clinical Trial. Controlled Clinical Trials **1994**, 15: 135-153.

10. Kleinbaum DG, Kupper LL, and Morganstern H. Epidemiologic Research. New York, Van Nostrand Reinhold, **1982**.

11. Cochran WG. Some Methods of Strengthening the Common χ^2 Tests. Biometrics **1954**, 10: 417-451.

12. Agresti A., and Caffo B. Simple and Effective Confidence Intervals for Proportions and Difference of Proportions Results from Adding Two Successes and Two Failures. The American Statistician **2000**, 54(4): 280-288.

13. Sato T. A Further Look at the Cochran-Mantel-Haenszel Risk Difference. Controlled Clinical Trials **1995**, 16: 359-361.

14. Nam J. Sample Size Determination in Stratified Trials to Establish the Equivalence of Two Treatments. Statistics in Medicine **1995**, 14: 2037-2049.

Figure 4.1 Summary of Simulated Type I Error for Stratified Risk Ratio

By Treatment Allocation

Alpha=0.025, Strata Allocation=1:1

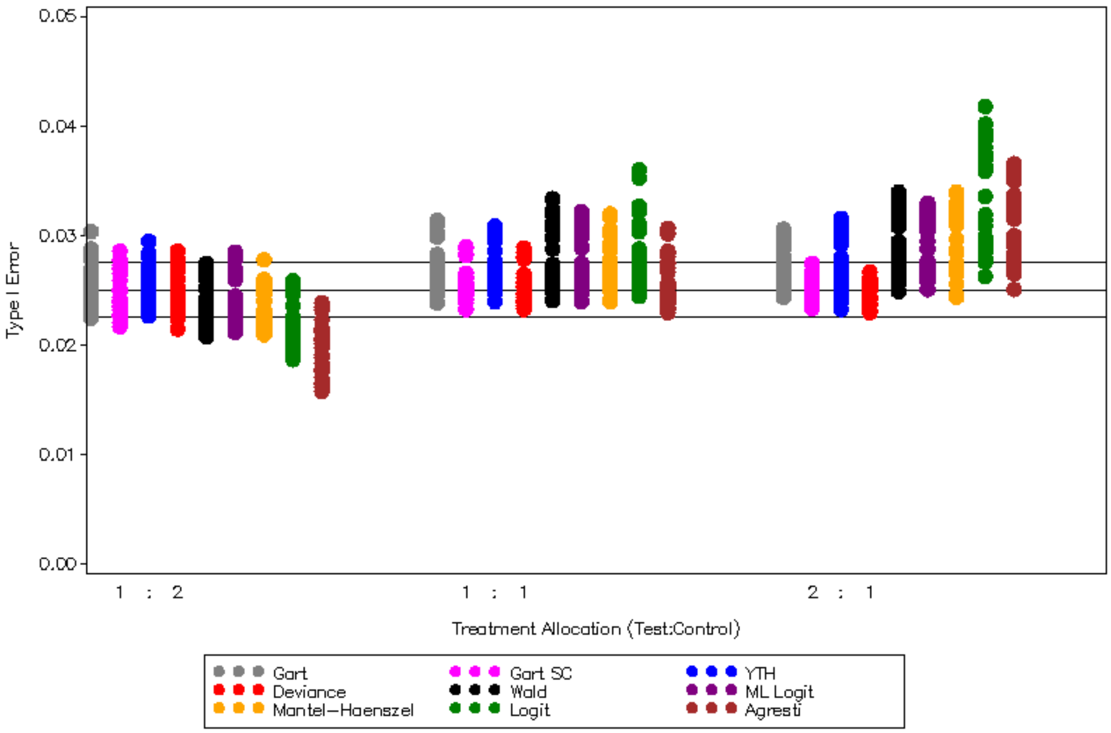


Figure 4.2 Summary of Simulated Type I Error for Stratified Risk Ratio

By Null Hypothesis Risk Ratio

Alpha=0.025, Strata Allocation=1:1

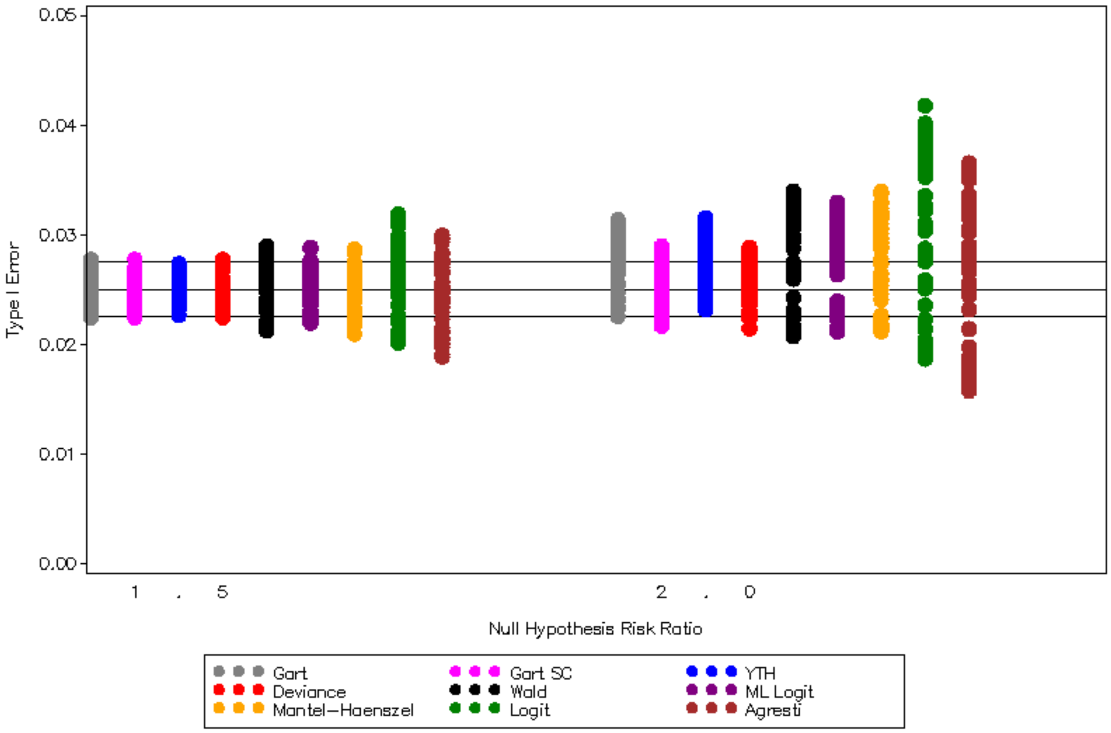


Figure 4.3 Summary of Simulated Type I Error for Stratified Risk Ratio
By Pi C1
Alpha=0.025, Strata Allocation=1:1

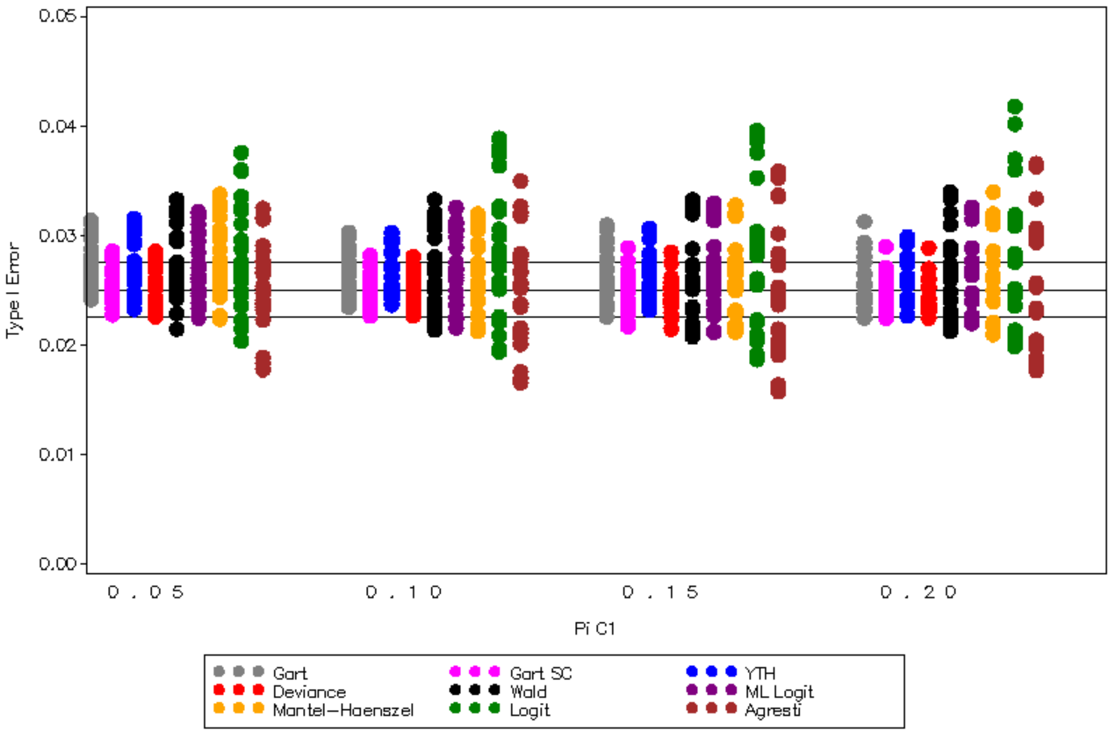


Figure 4.4 Summary of Simulated Type I Error for Stratified Risk Ratio
By Pi C2
Alpha=0.025, Strata Allocation=1:1

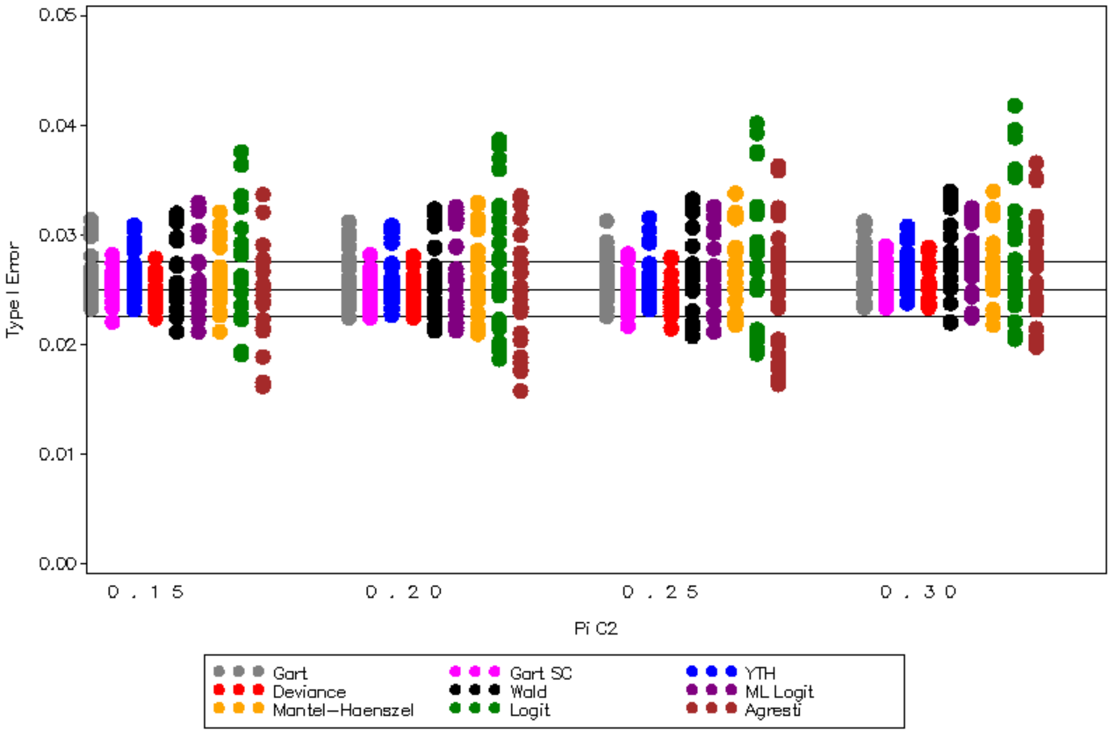


Figure 4.5 Summary of Simulated Type I Error for Stratified Risk Ratio

By (Pi C2 – Pi C1)

Alpha=0.025, Strata Allocation=1:1

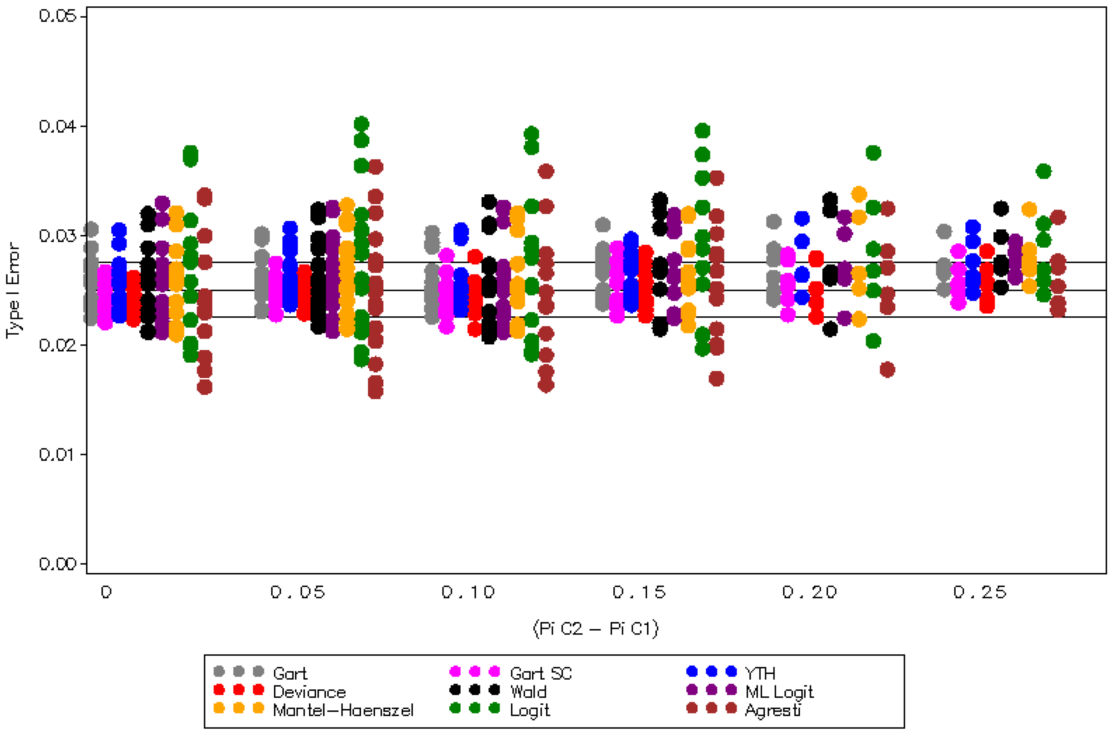


Figure 4.6 Summary of Simulated Type I Error for Stratified Risk Ratio

By Total Sample Size

Alpha=0.025, Strata Allocation=1:1

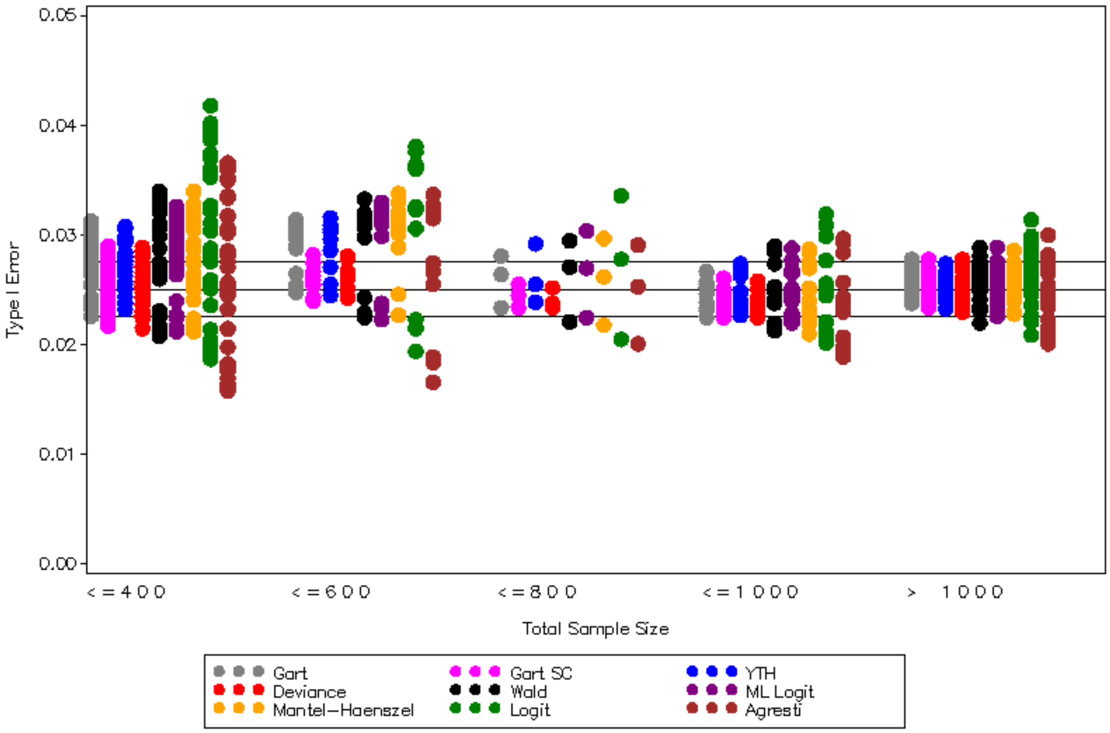


Figure 4.7 Comparison of Simulated Power for Stratified Risk Ratio

Gart–SC & Deviance Simulated Power
By Treatment Allocation
Strata allocation=1:1, Alpha=0.025, Theta=1

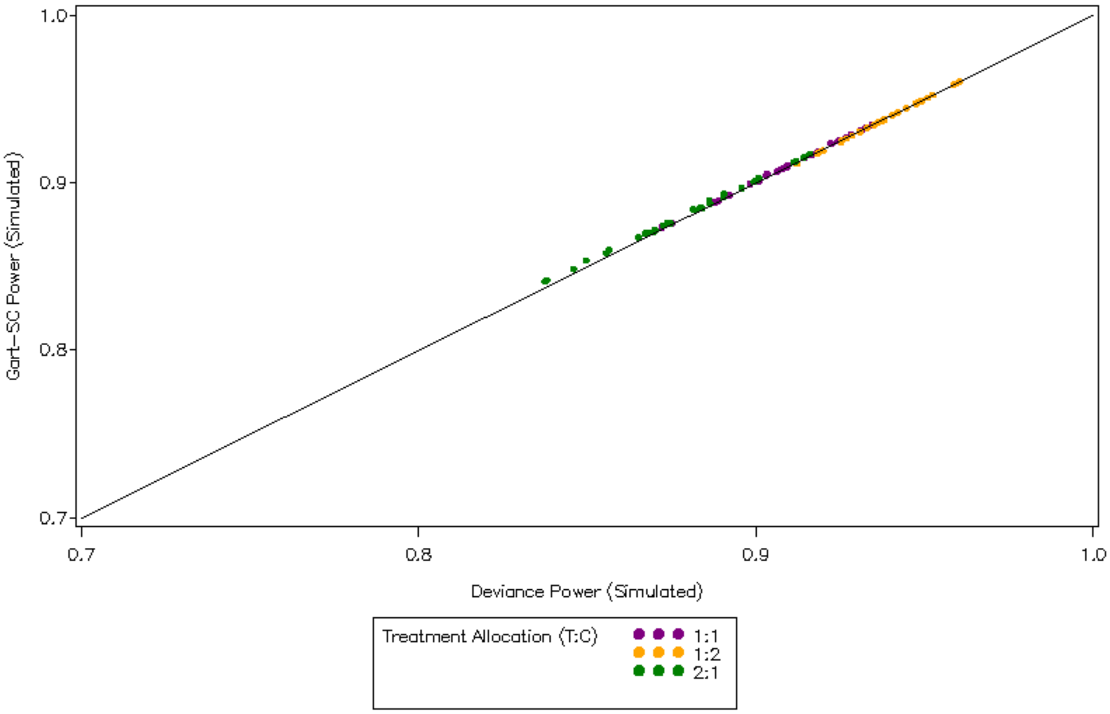


Figure 4.8 Comparison of Simulated Power for Stratified Risk Ratio

Gart–SC & YTH Simulated Power
By Treatment Allocation
Strata allocation=1:1, Alpha=0.025, Theta=1

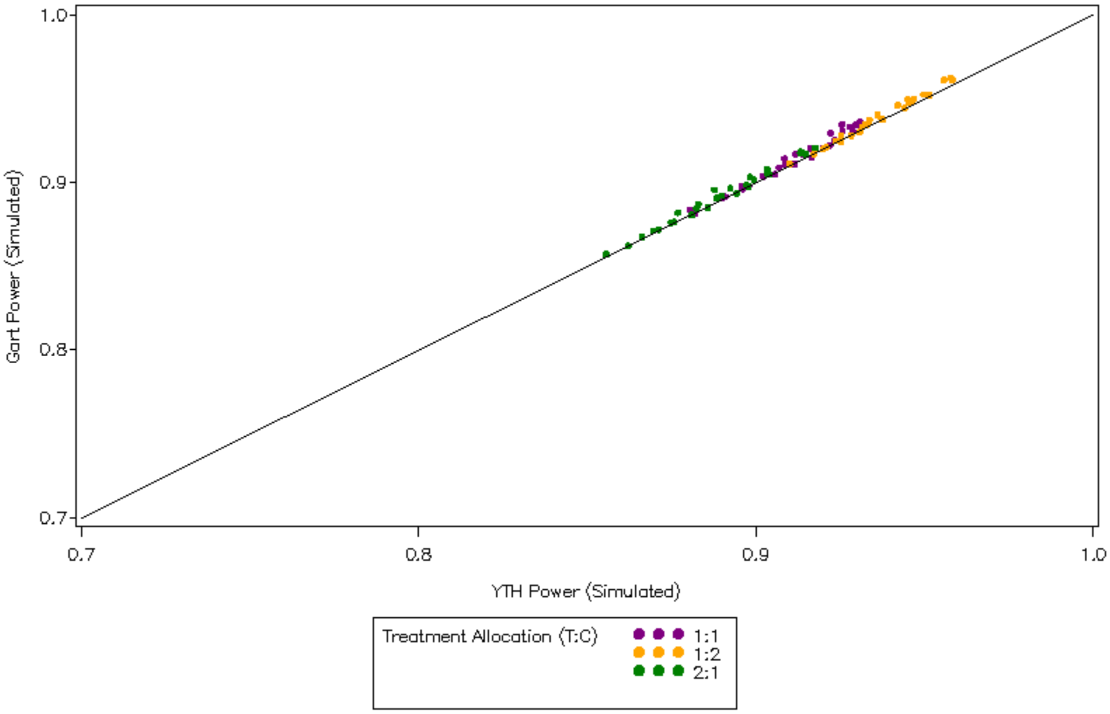


Figure 4.9 Comparison of Simulated Power for Stratified Risk Ratio

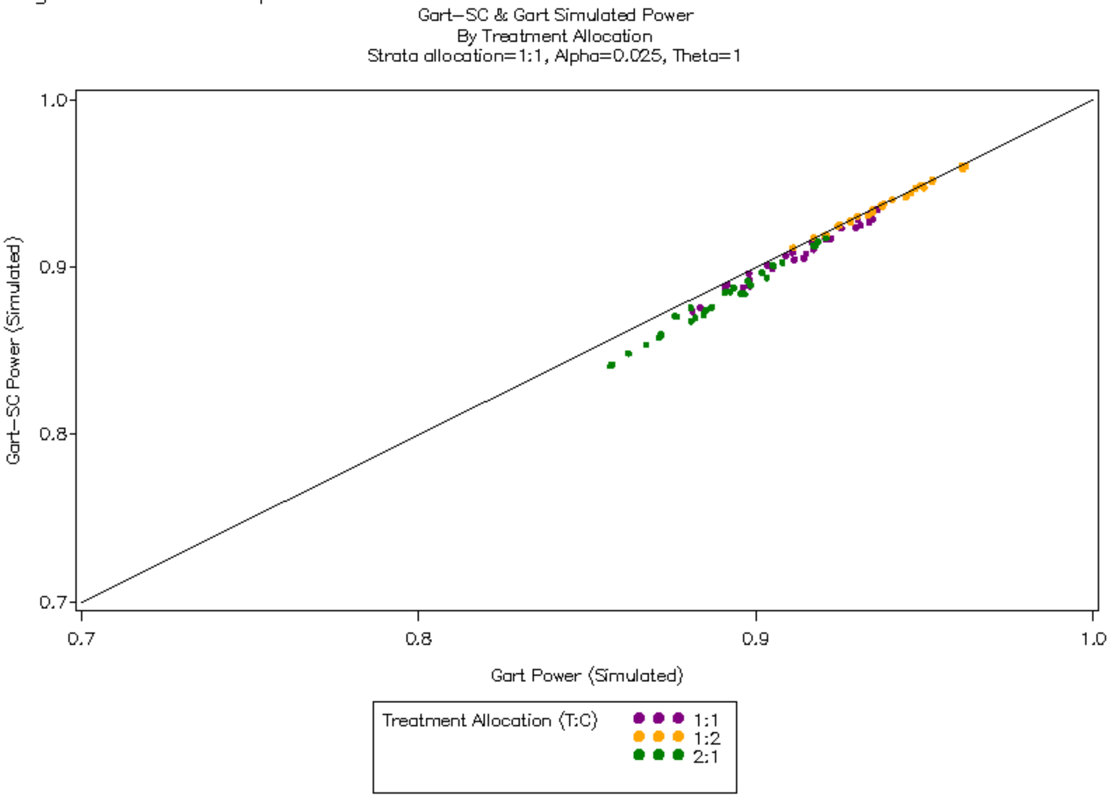


Figure 4.10 Comparison of Simulated Power for Stratified Risk Ratio

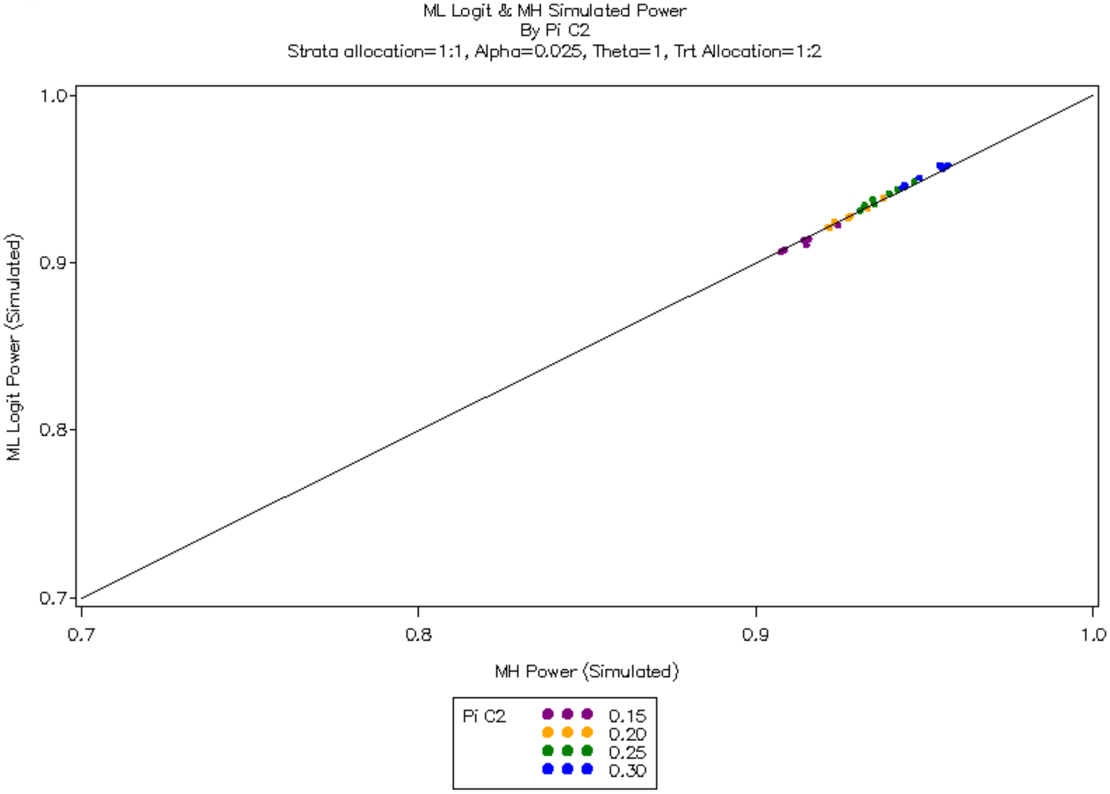


Figure 4.11 Comparison of Simulated Power for Stratified Risk Ratio

Logit & Agresti Simulated Power
By Pi C2
Strata allocation=1:1, Alpha=0.025, Theta=1, Trt Allocation=1:2

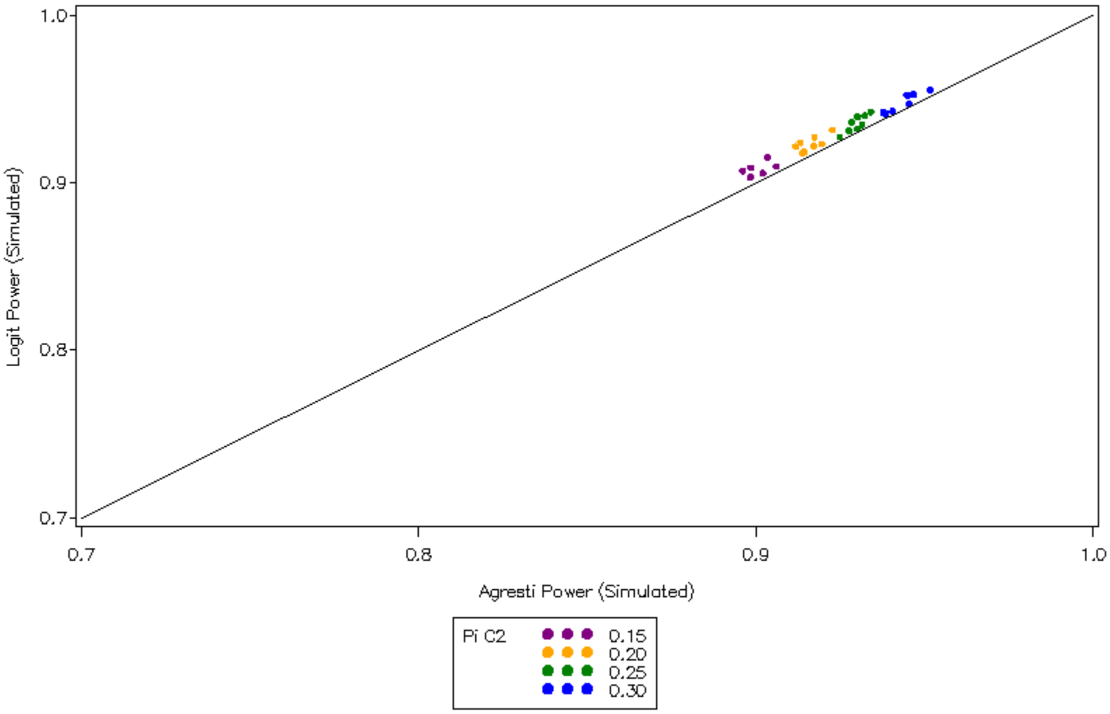


Figure 4.12 Comparison of Simulated Power for Stratified Risk Ratio

MH & Logit Simulated Power
By Pi C2
Strata allocation=1:1, Alpha=0.025, Theta=1, Trt Allocation=1:2

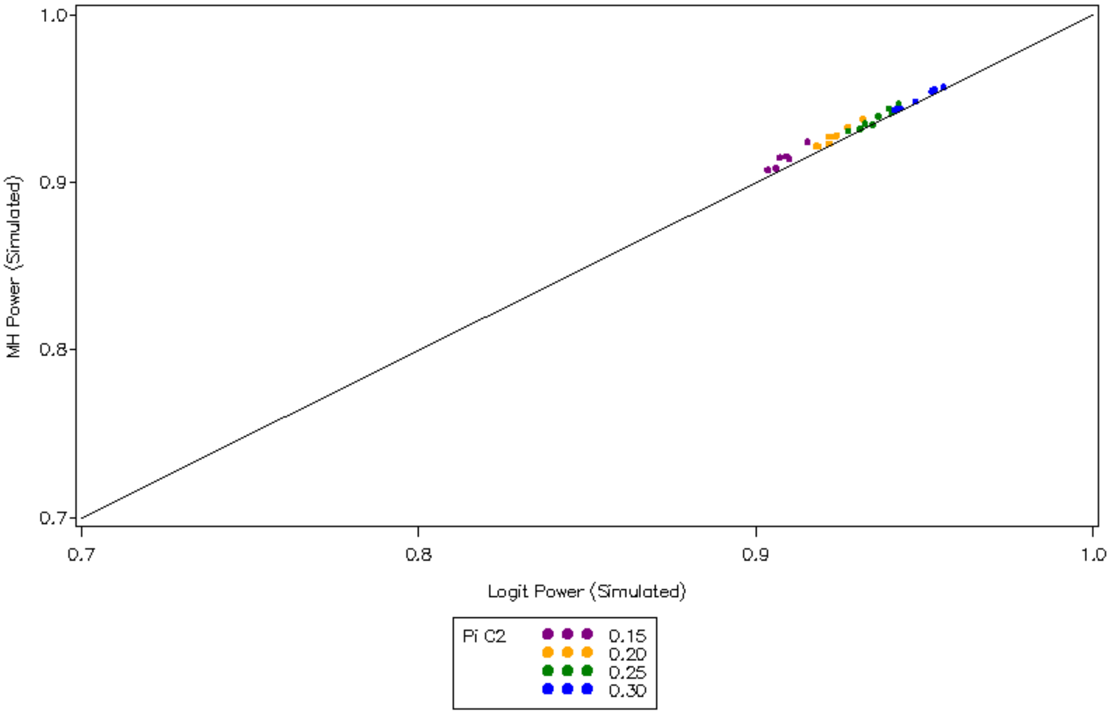
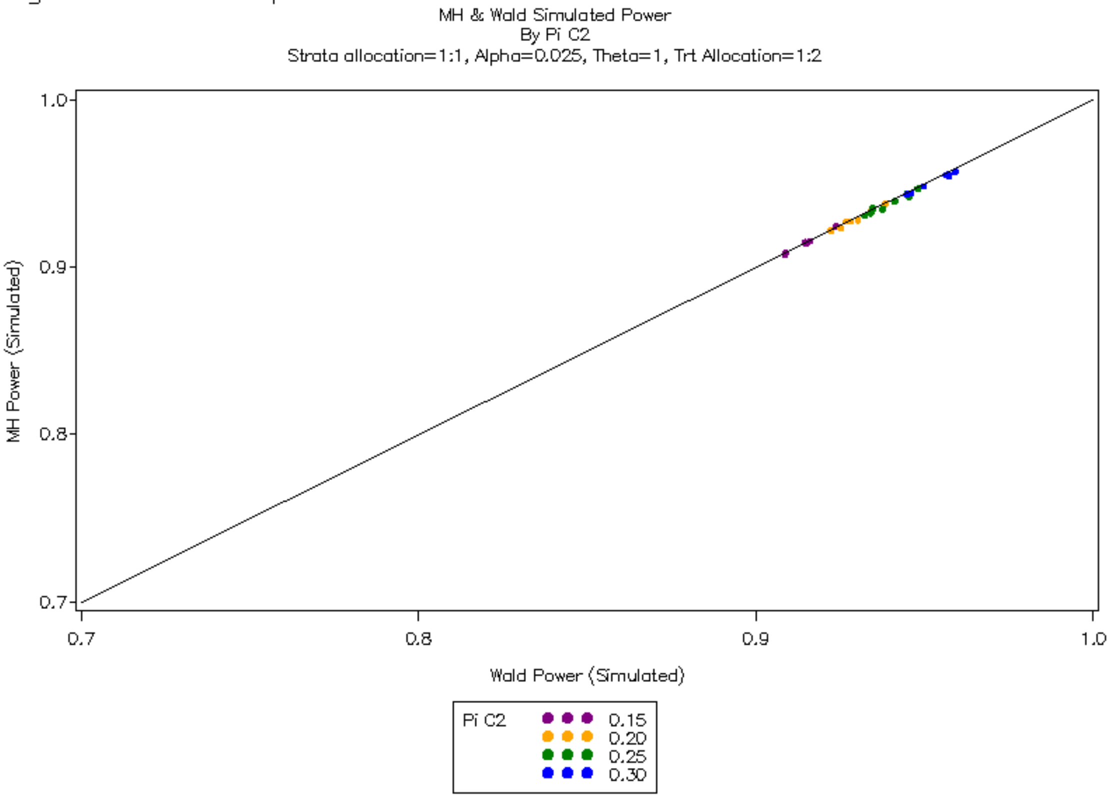
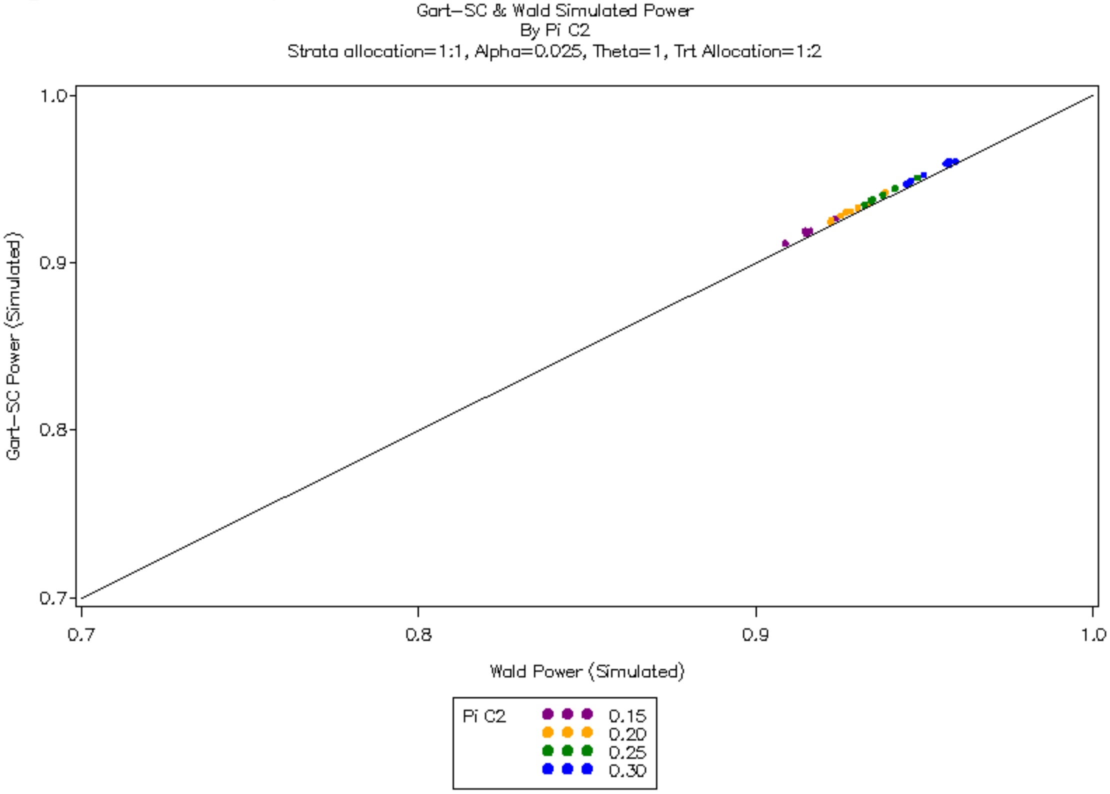


Figure 4.13 Comparison of Simulated Power for Stratified Risk Ratio



MH & Wald Simulated Power
By Pi C2
Strata allocation=1:1, Alpha=0.025, Theta=1, Trt Allocation=1:2

Figure 4.14 Comparison of Simulated Power for Stratified Risk Ratio



Gart-SC & Wald Simulated Power
By Pi C2
Strata allocation=1:1, Alpha=0.025, Theta=1, Trt Allocation=1:2

Figure 4.15 Summary of Simulated Type I Error for Stratified Risk Ratio
By Strata Allocation
Alpha=0.025, Treatment Allocation=1:2

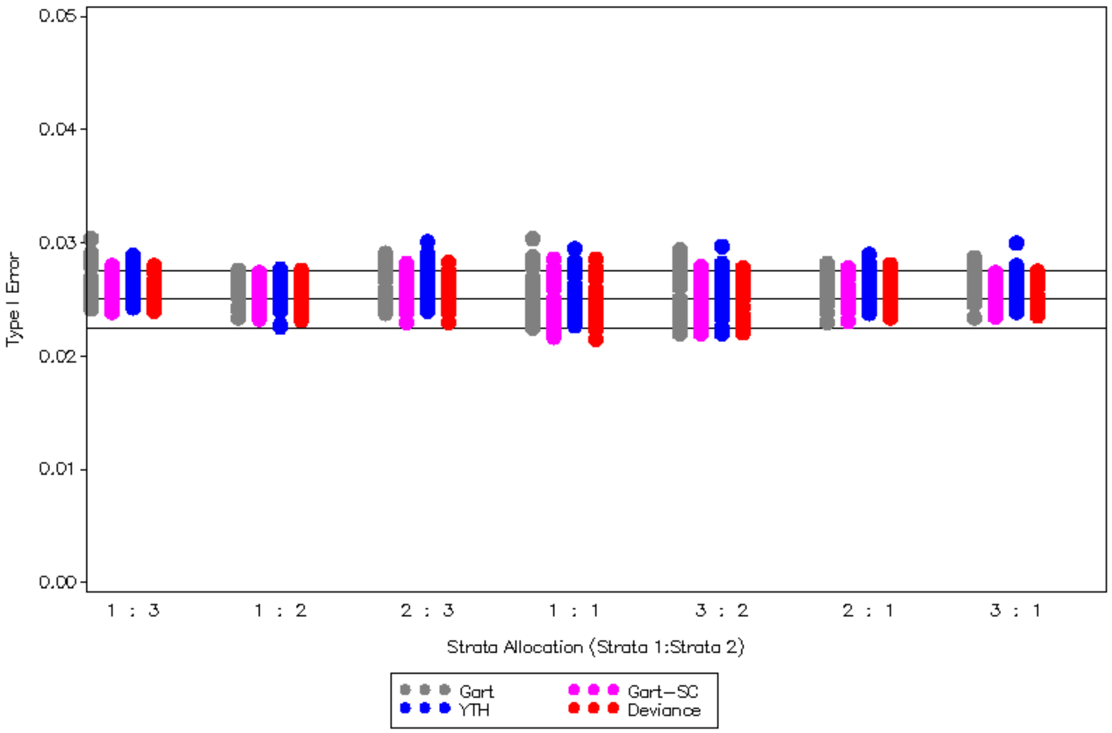


Figure 4.16 Summary of Simulated Type I Error for Stratified Risk Ratio
By Strata Allocation
Alpha=0.025, Treatment Allocation 1:1

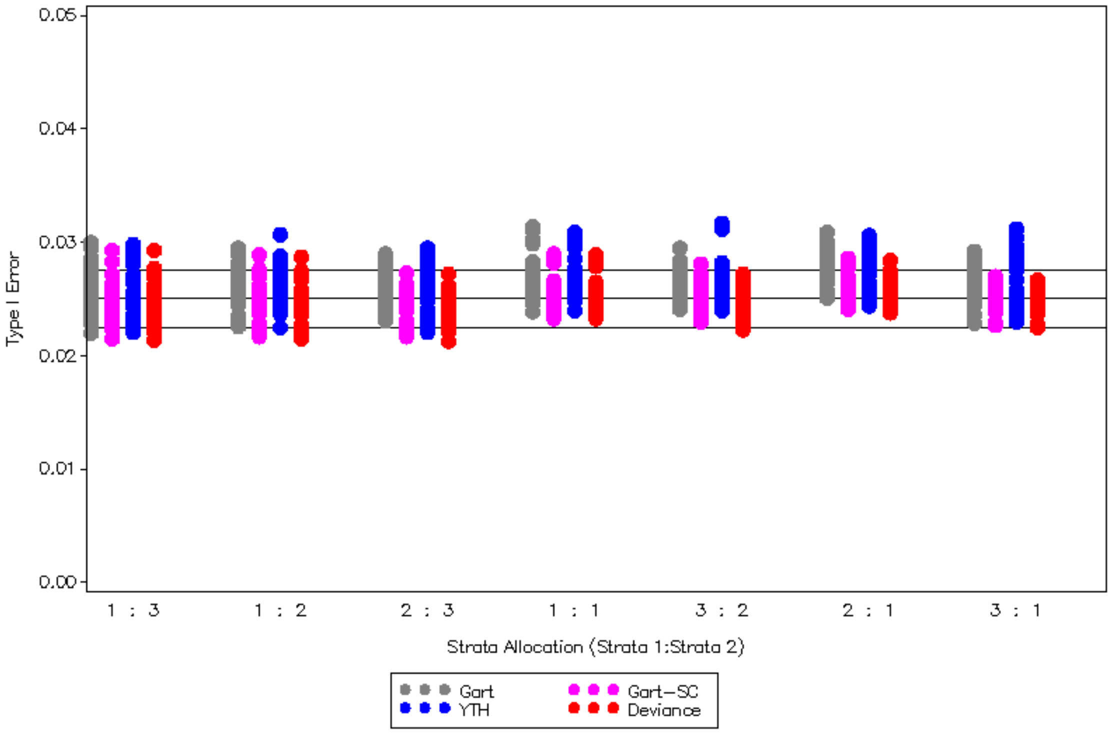


Figure 4.17 Summary of Simulated Type I Error for Stratified Risk Ratio
By Strata Allocation
Alpha=0.025, Treatment Allocation 2:1

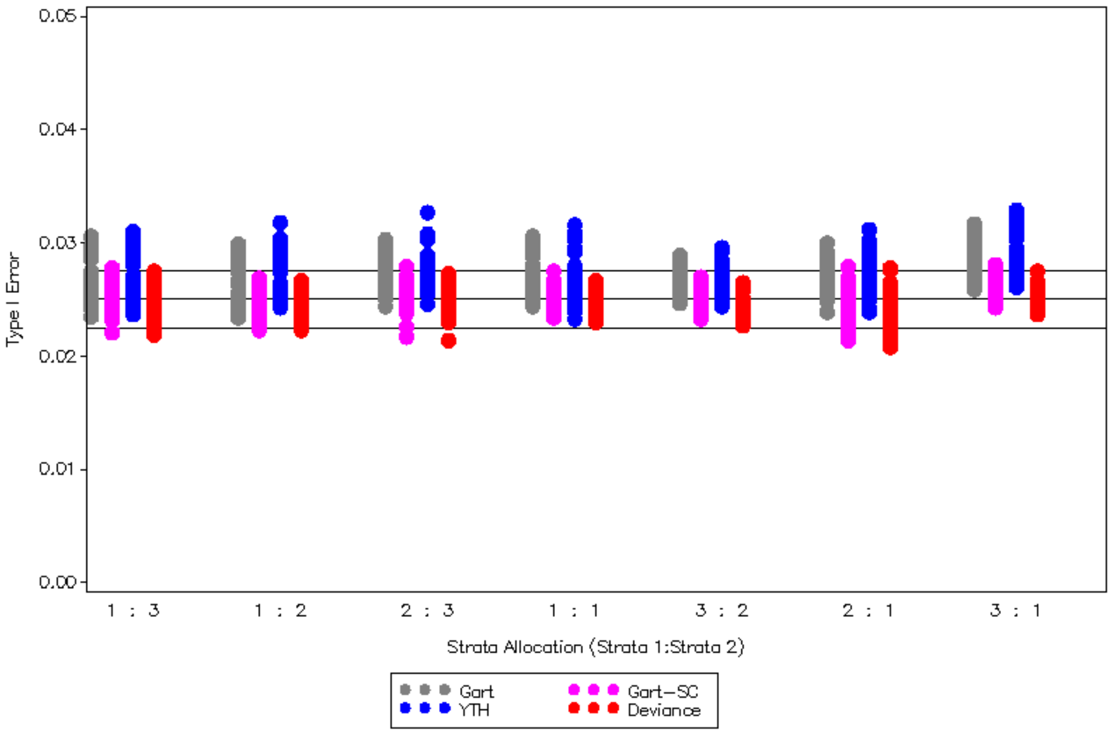


Figure 4.18 Summary of Simulated Power for Stratified Risk Ratio
By Strata Allocation
Alpha=0.025, Treatment Allocation=1:2

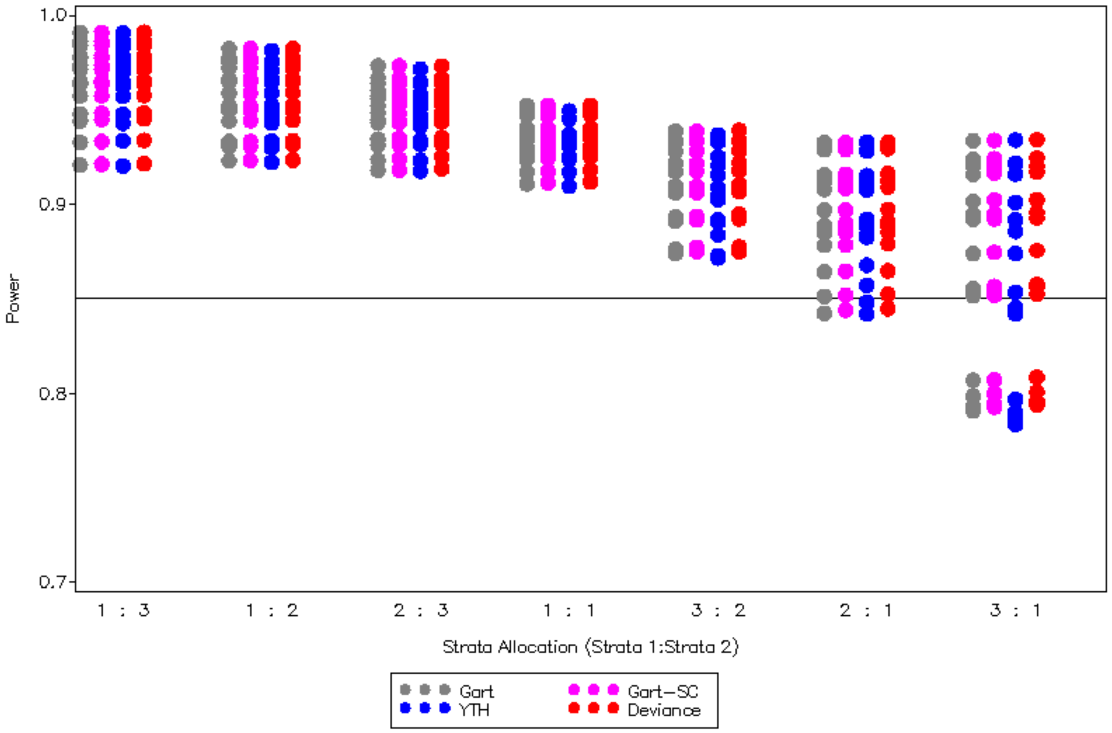


Figure 4.19 Summary of Simulated Power for Stratified Risk Ratio
By Strata Allocation
Alpha=0.025, Treatment Allocation=1:1

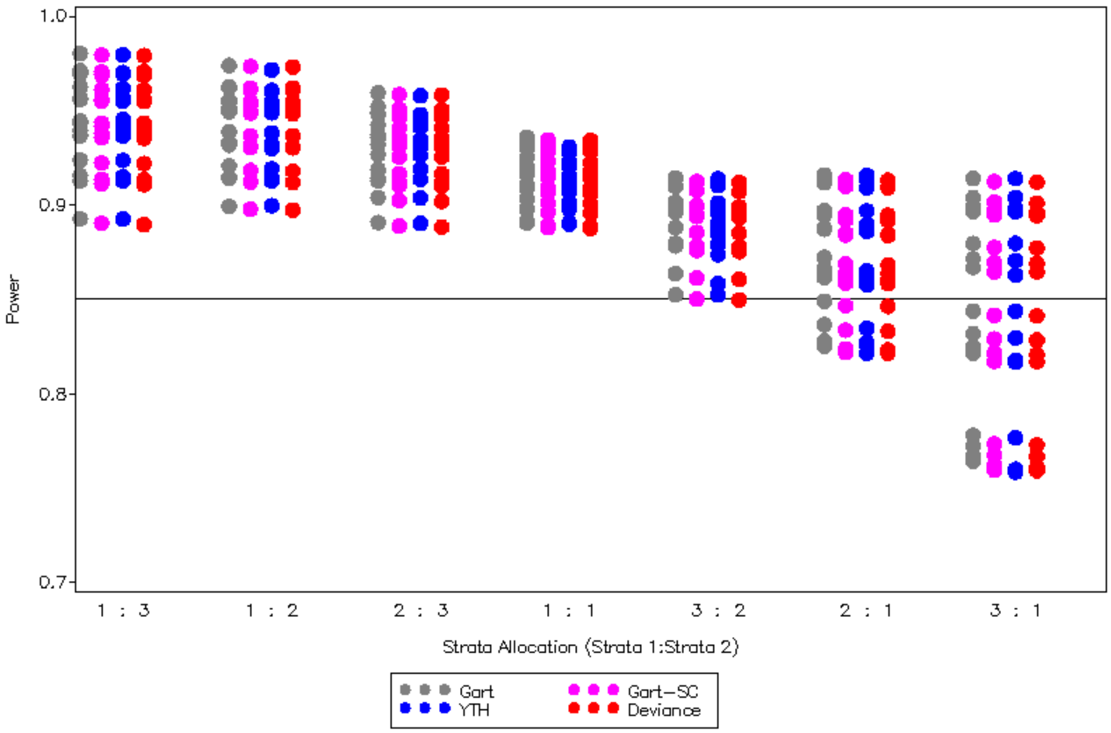


Figure 4.20 Summary of Simulated Power for Stratified Risk Ratio
By Strata Allocation
Alpha=0.025, Treatment Allocation=2:1

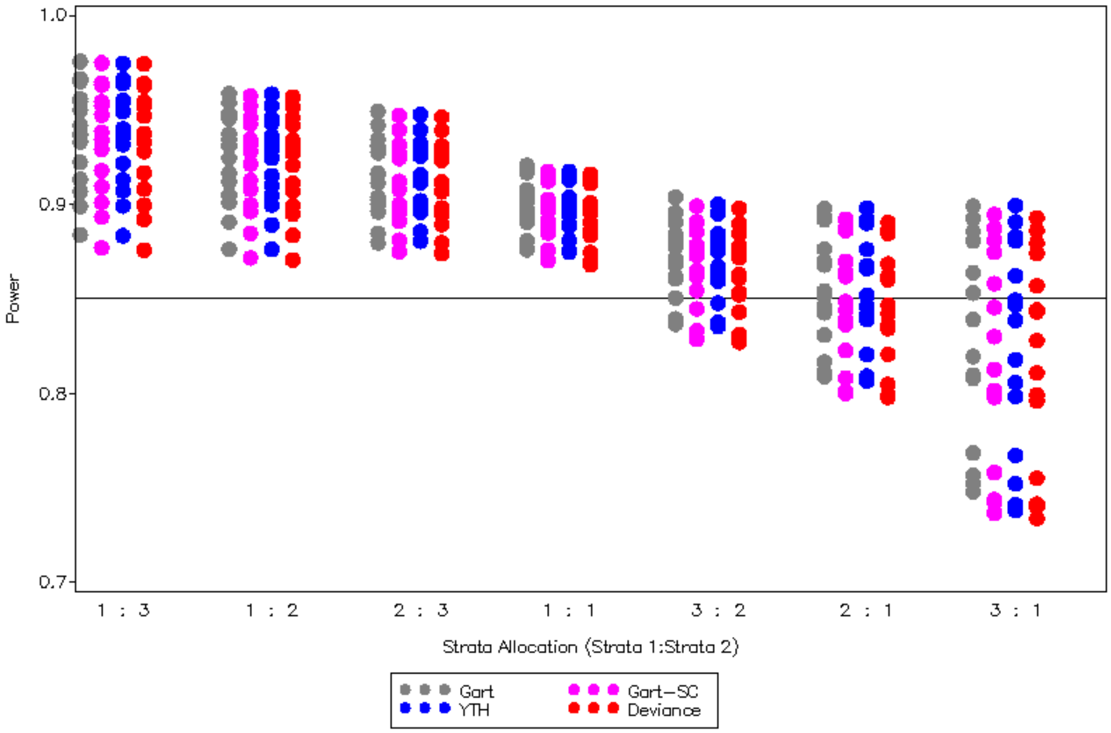


Figure 4.21 Comparison of Simulated & Calculated Power for Stratified Risk Ratio
Gart–SC Simulated & Nam Calculated Power
By Strata Allocation
Alpha=0.025, Theta=1, Treatment Allocation=1:2

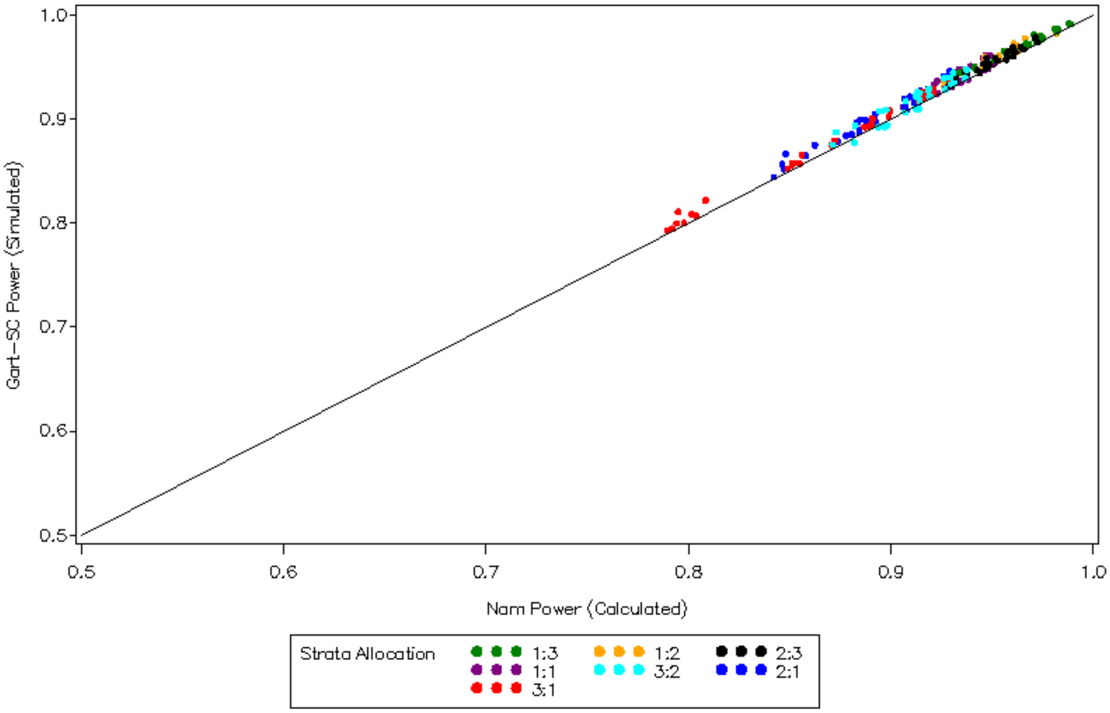


Figure 4.22 Comparison of Simulated & Calculated Power for Stratified Risk Ratio
Gart–SC Simulated & Nam Calculated Power
By Strata Allocation
Alpha=0.025, Theta=1, Treatment Allocation=1:1

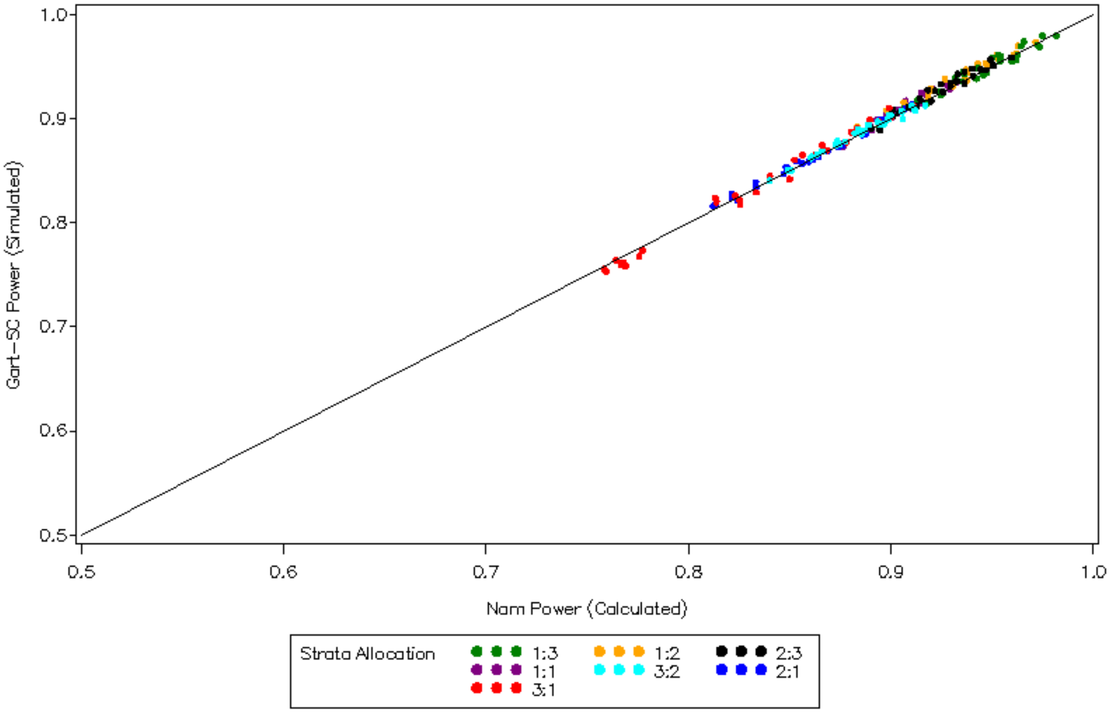


Figure 4.23 Comparison of Simulated & Calculated Power for Stratified Risk Ratio
Gart–SC Simulated & Nam Calculated Power
By Strata Allocation
Alpha=0.025, Theta=1, Treatment Allocation=2:1

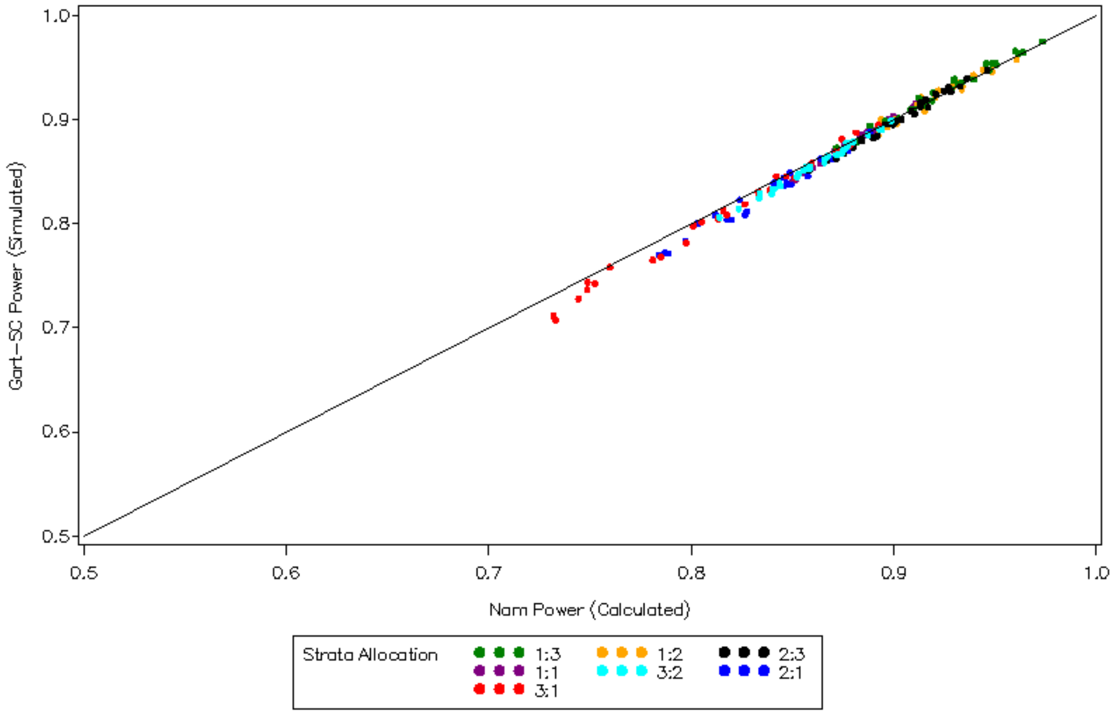


Figure 4.24 Comparison of Simulated & Calculated Power for Stratified Risk Ratio
Gart Simulated & Nam Calculated Power
By Strata Allocation
Alpha=0.025, Theta=1, Treatment Allocation=1:2

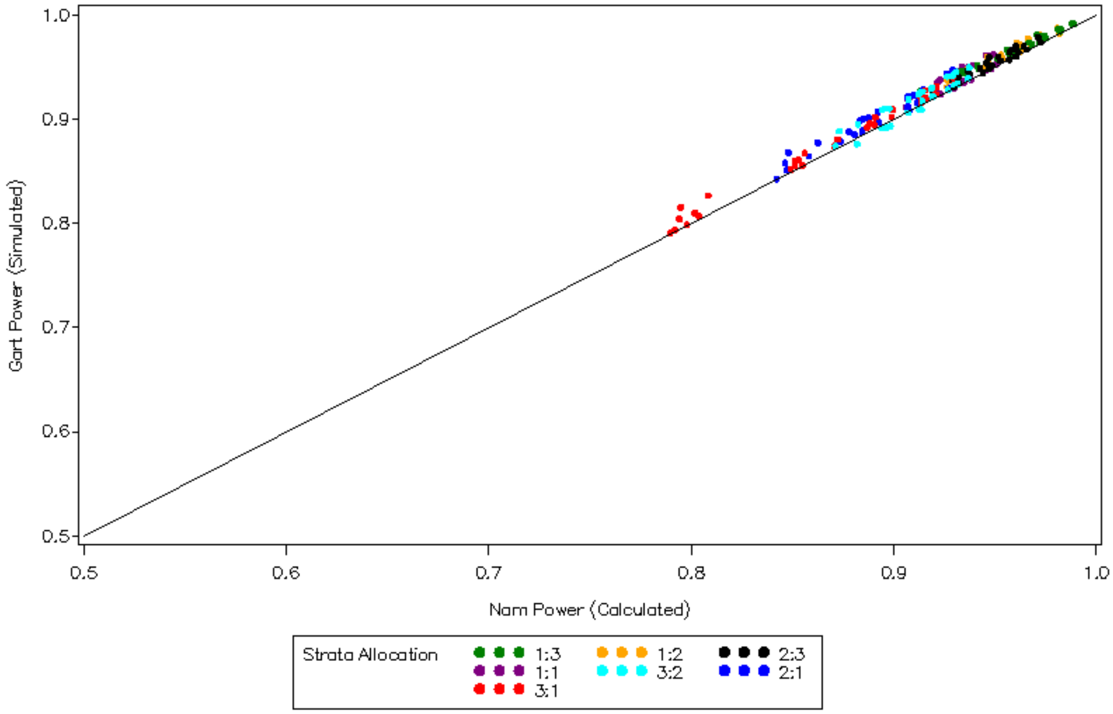


Figure 4.25 Comparison of Simulated & Calculated Power for Stratified Risk Ratio
Gart Simulated & Nam Calculated Power
By Strata Allocation
Alpha=0.025, Theta=1, Treatment Allocation=1:1

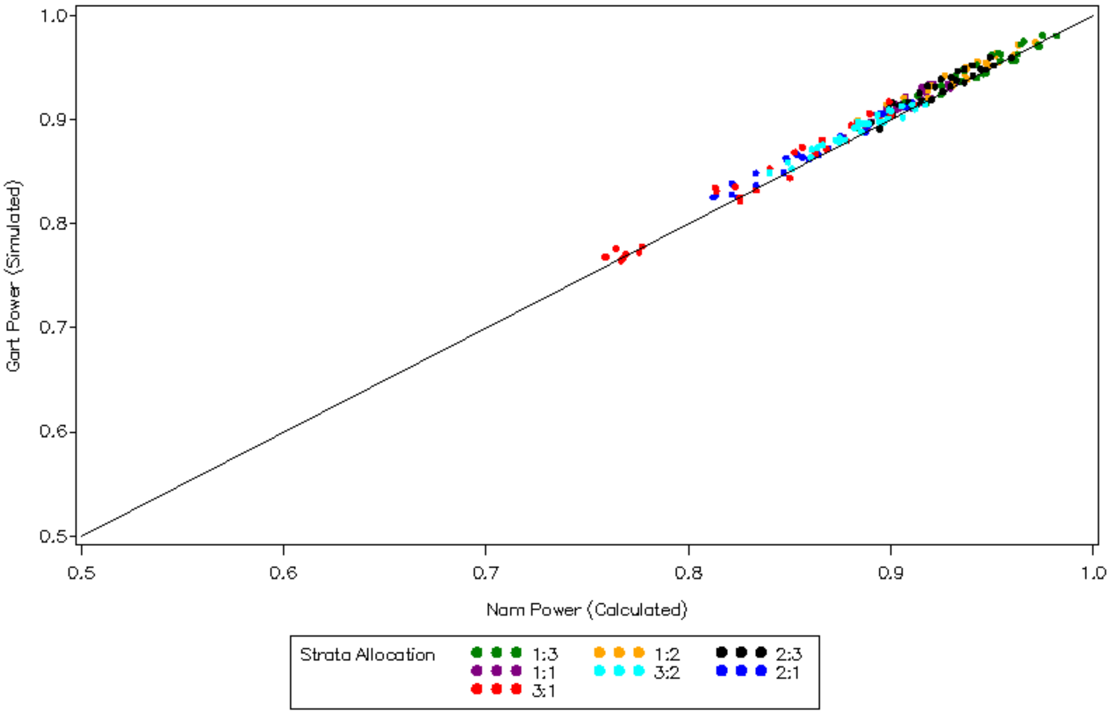


Figure 4.26 Comparison of Simulated & Calculated Power for Stratified Risk Ratio
Gart Simulated & Nam Calculated Power
By Strata Allocation
Alpha=0.025, Theta=1, Treatment Allocation=2:1

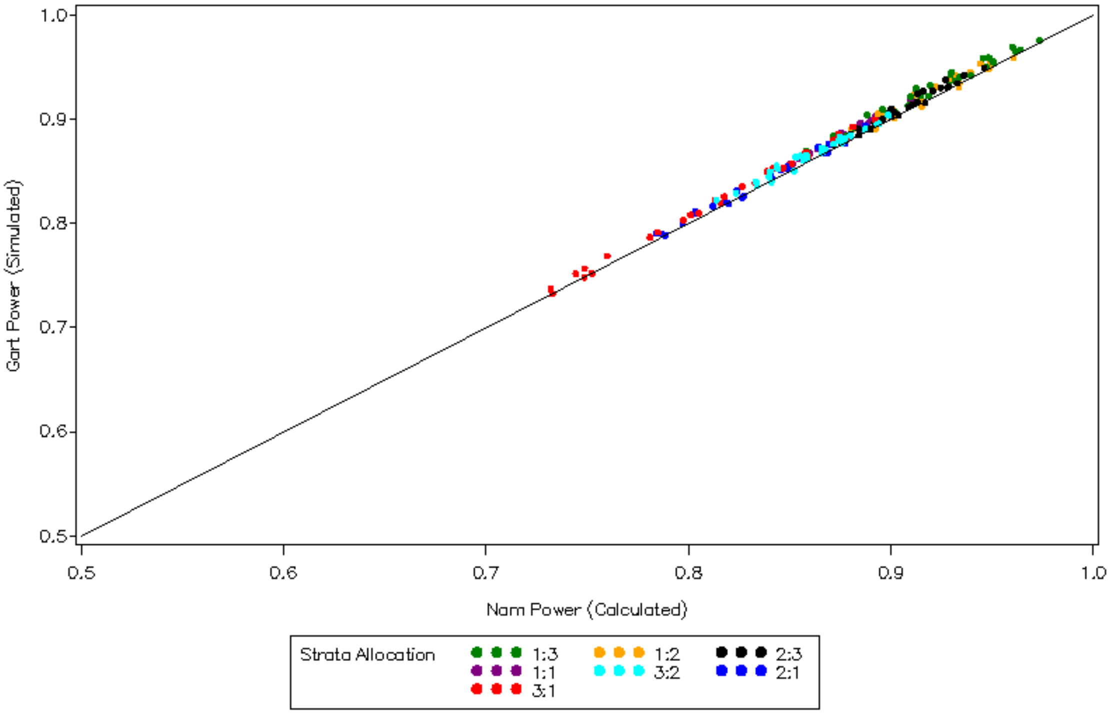


Figure 4.27 Comparison of Simulated & Calculated Power for Stratified Risk Ratio

Gart–SC Simulated & TS Calculated Power

By (Pi C2 – Pi C1)

Alpha=0.025, Theta=1, Treatment Allocation=1:2

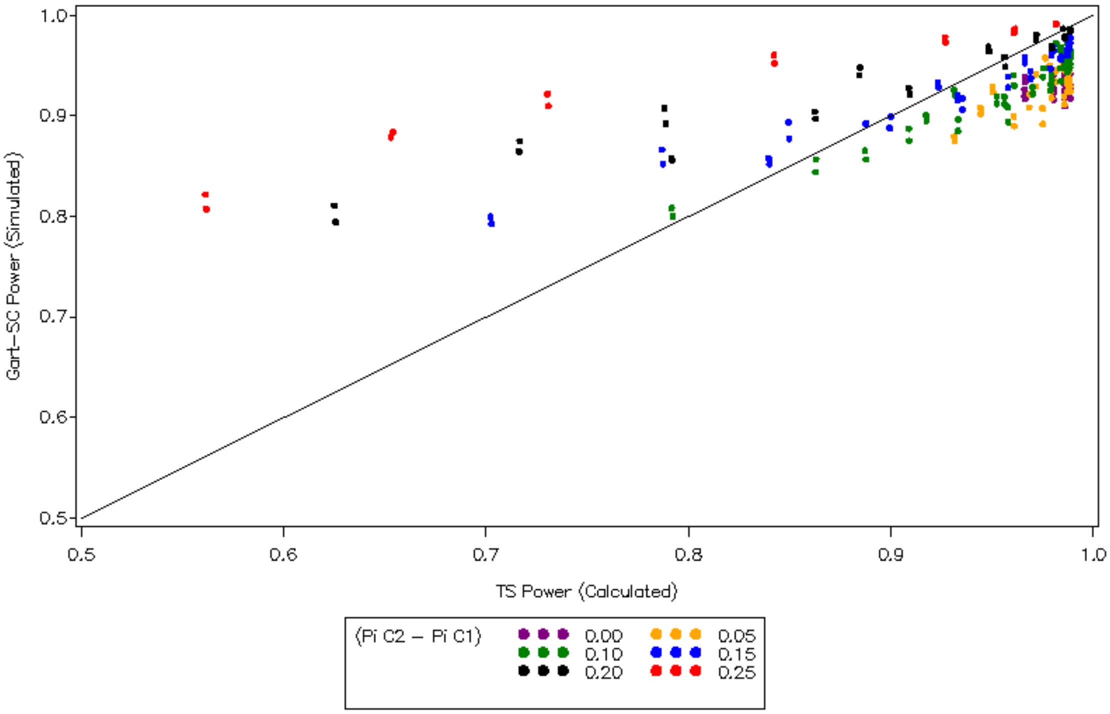


Figure 4.28 Comparison of Simulated & Calculated Power for Stratified Risk Ratio

Gart–SC Simulated & TS Calculated Power

By (Pi C2 – Pi C1)

Alpha=0.025, Theta=1, Treatment Allocation=1:1

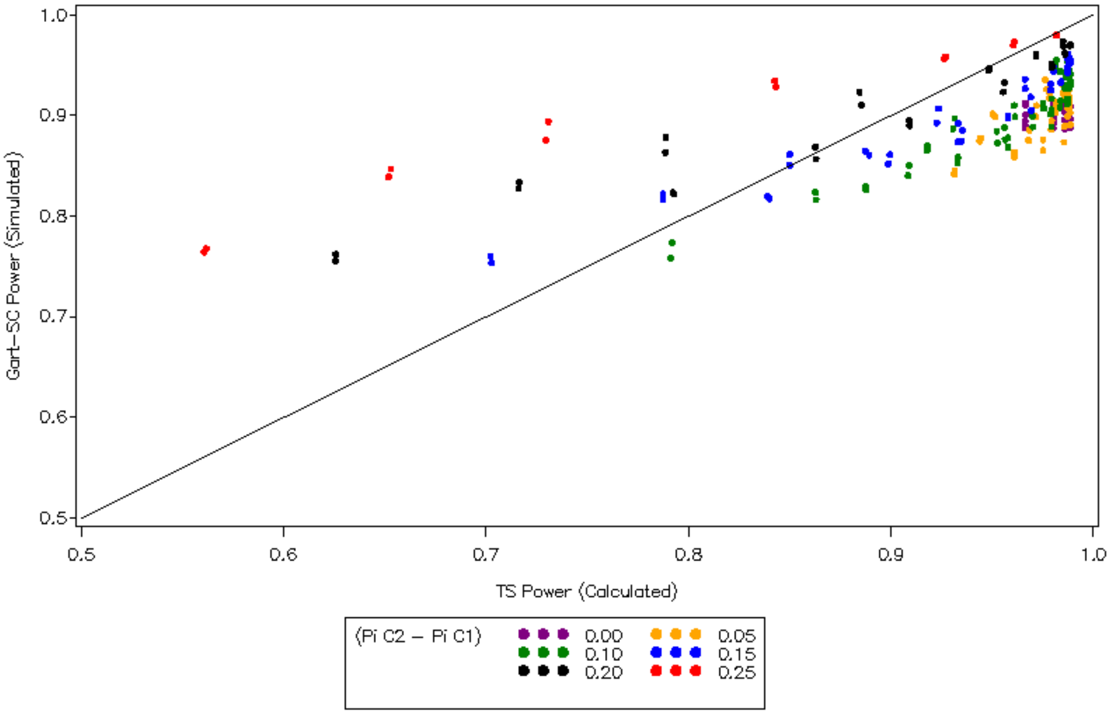


Figure 4.29 Comparison of Simulated & Calculated Power for Stratified Risk Ratio

Gart–SC Simulated & TS Calculated Power

By (Pi C2 – Pi C1)

Alpha=0.025, Theta=1, Treatment Allocation=2:1

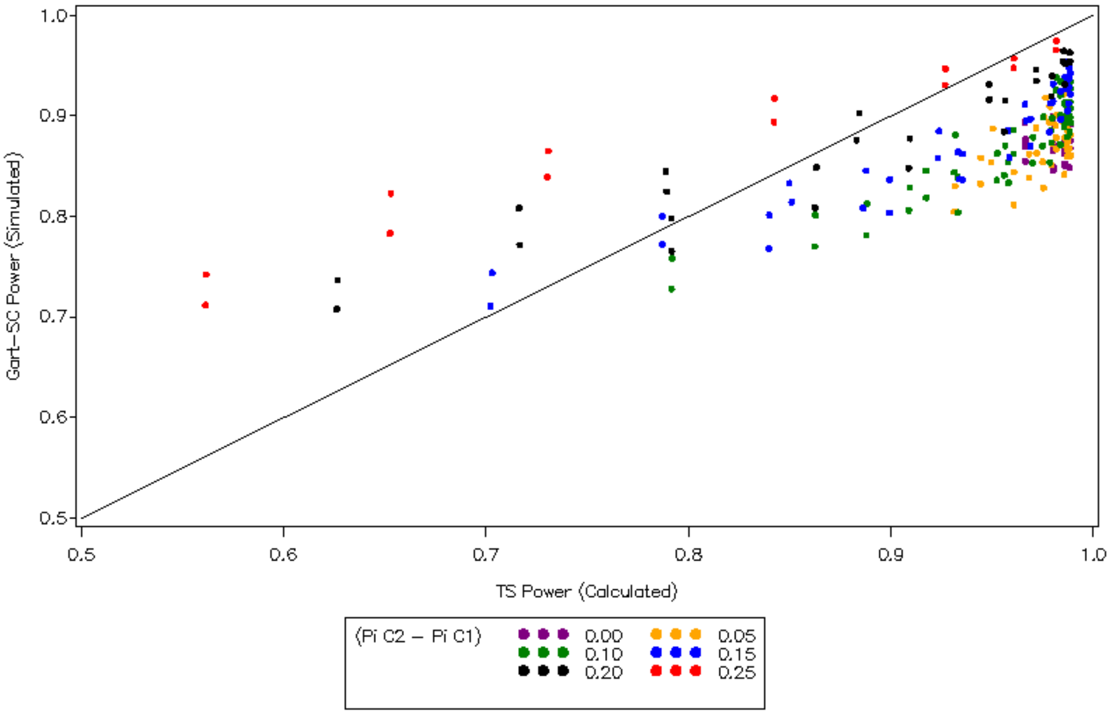


Figure 4.30 Summary of Simulated Type I Error for Stratified Risk Difference

By Treatment Allocation

Alpha=0.025, Strata Allocation=1:1

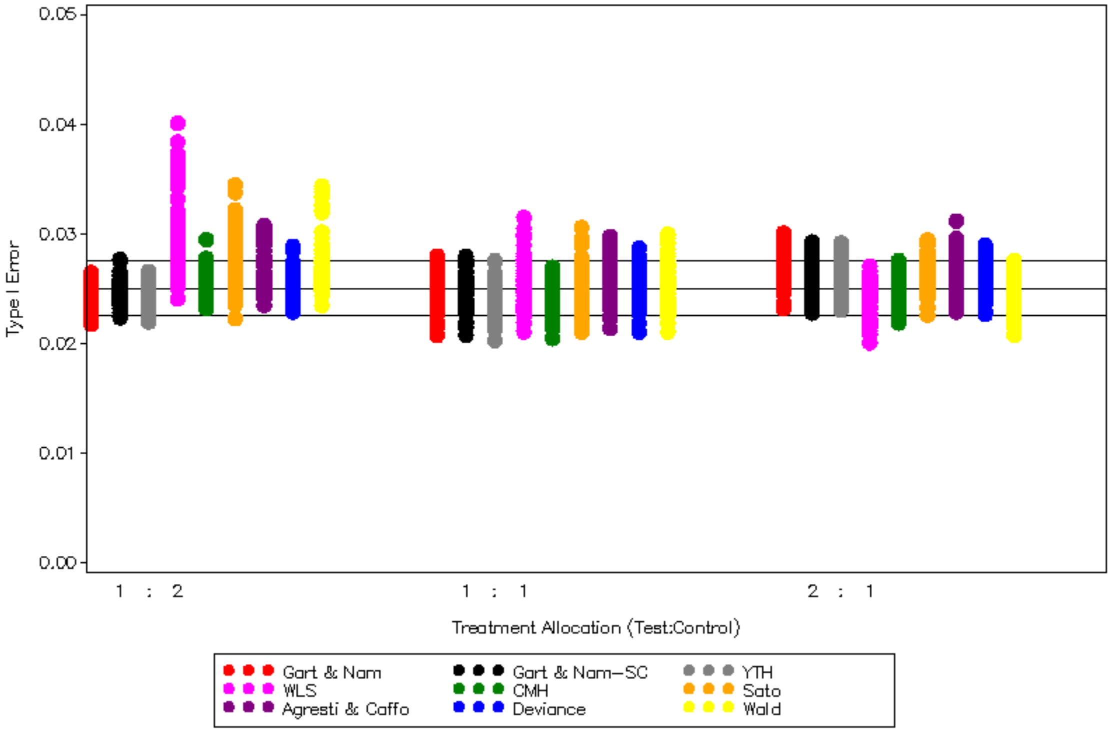


Figure 4.31 Summary of Simulated Type I Error for Stratified Risk Difference
By Null Hypothesis Risk Difference
Alpha=0.025, Strata Allocation=1:1

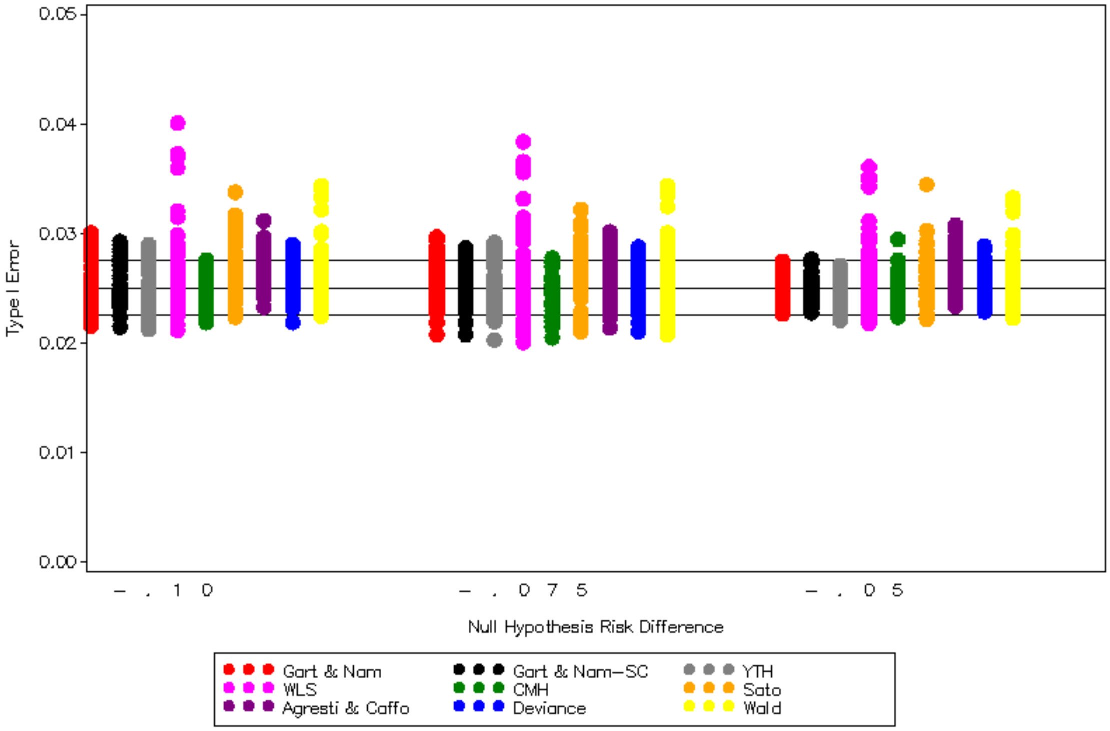


Figure 4.32 Summary of Simulated Type I Error for Stratified Risk Difference
By Pi C1
Alpha=0.025, Strata Allocation=1:1

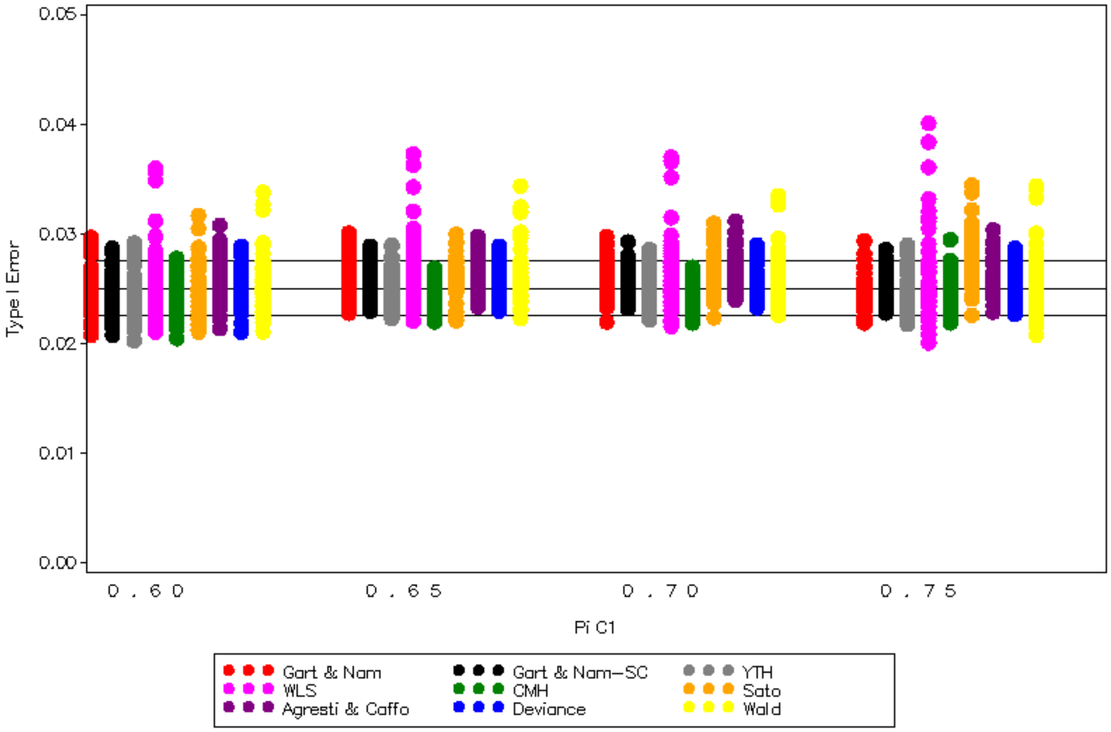


Figure 4.33 Summary of Simulated Type I Error for Stratified Risk Difference
By Pi C2
Alpha=0.025, Strata Allocation=1:1

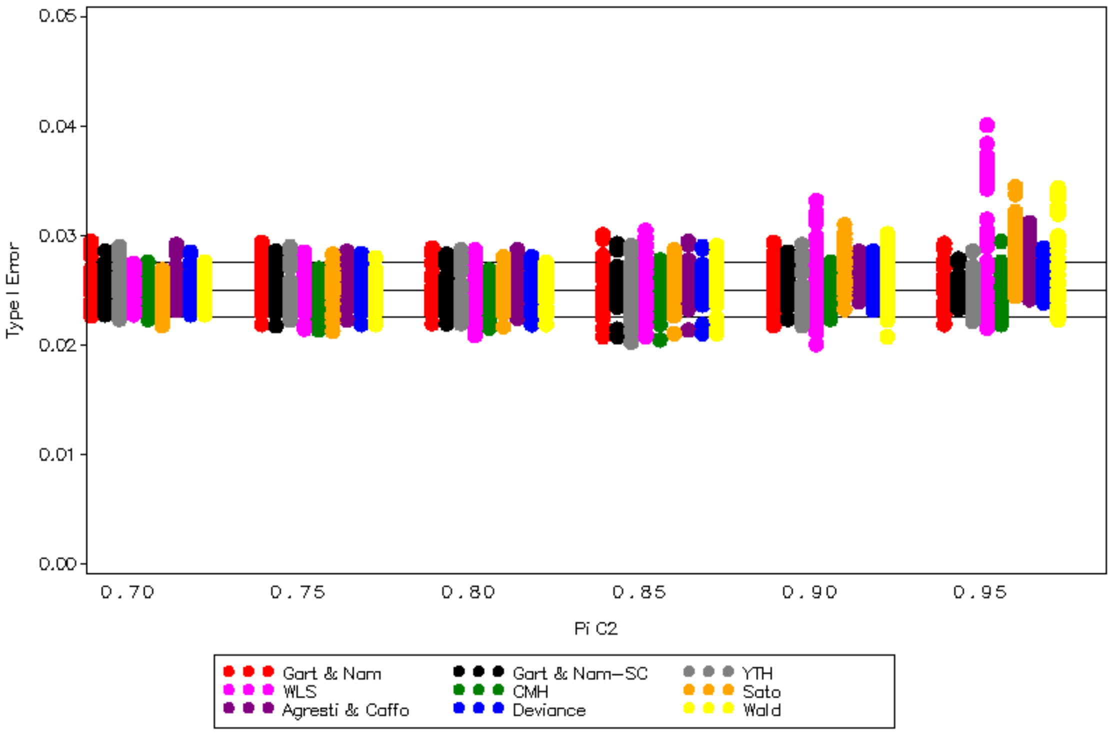


Figure 4.34 Summary of Simulated Type I Error for Stratified Risk Difference
By (Pi C2 – Pi C1)
Alpha=0.025, Strata Allocation=1:1

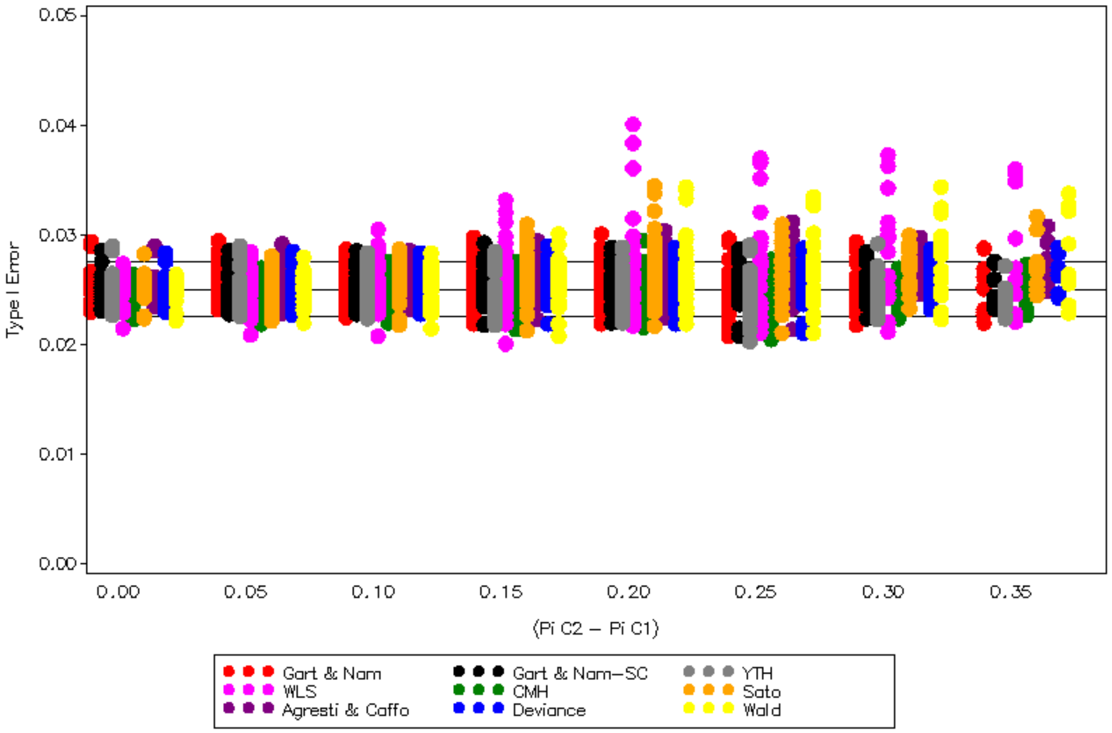


Figure 4.35 Summary of Simulated Type I Error for Stratified Risk Difference
By Total Sample Size
Alpha=0.025, Strata Allocation=1:1

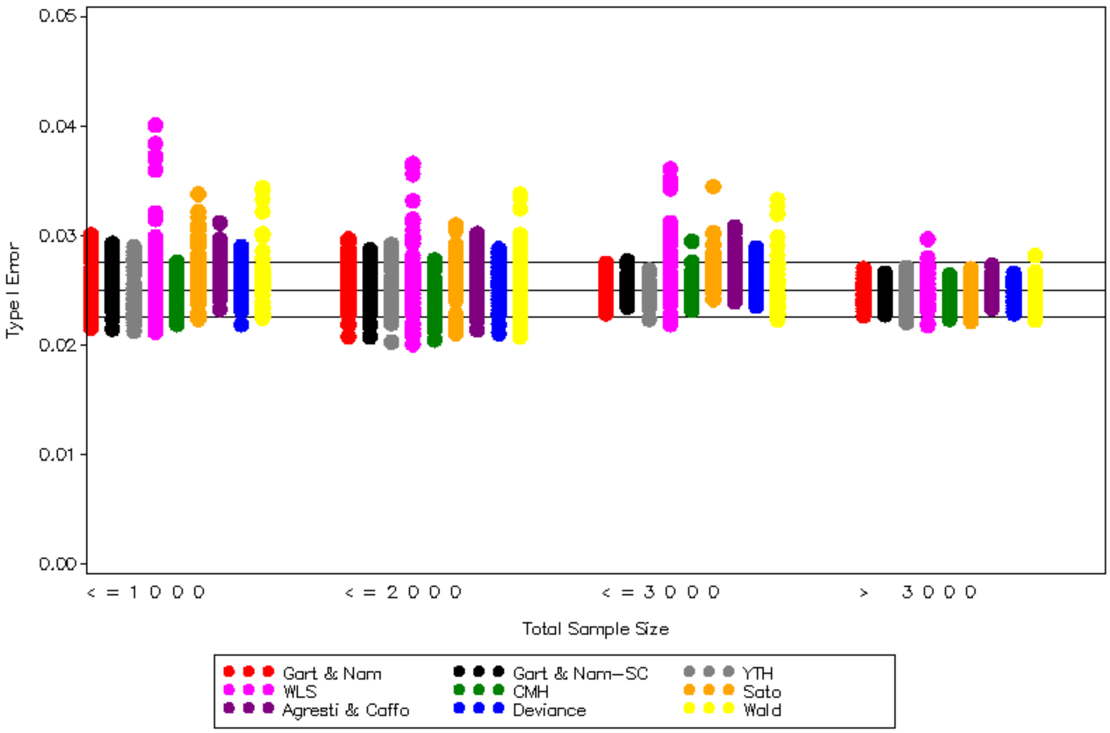


Figure 4.36 Comparison of Simulated Power for Stratified Risk Difference
Gart & Nam and Gart & Nam-SC Simulated Power
By Treatment Allocation
Alpha=0.025, Strata Allocation=1:1, Delta=0

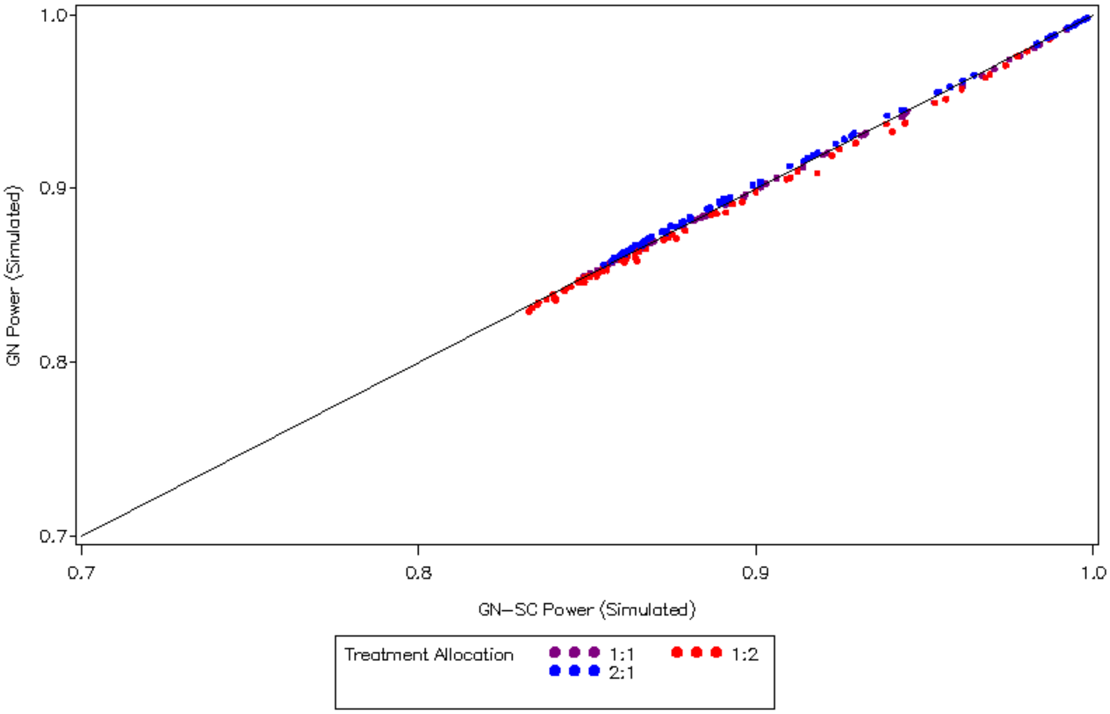


Figure 4.37 Comparison of Simulated Power for Stratified Risk Difference

Gart & Nam and YTH Simulated Power

By Pi C2

Alpha=0.025, Strata Allocation=1:1, Delta=0

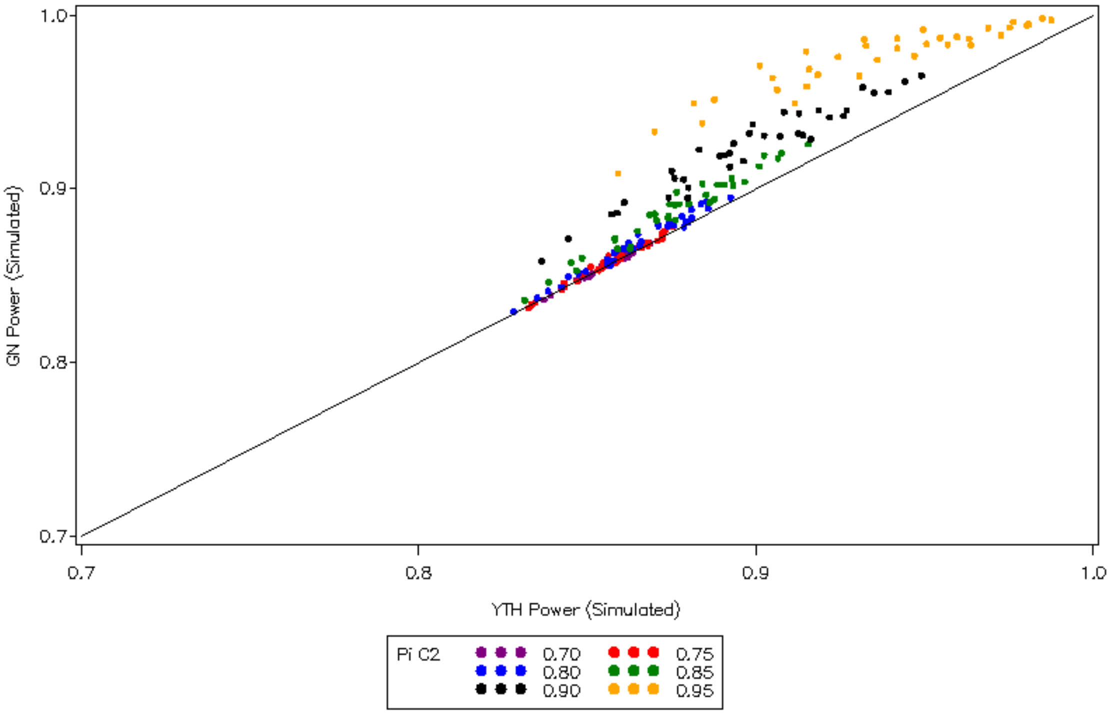


Figure 4.38 Comparison of Simulated Power for Stratified Risk Difference

Gart & Nam and Deviance Simulated Power

By Treatment Allocation

Alpha=0.025, Strata Allocation=1:1, Delta=0

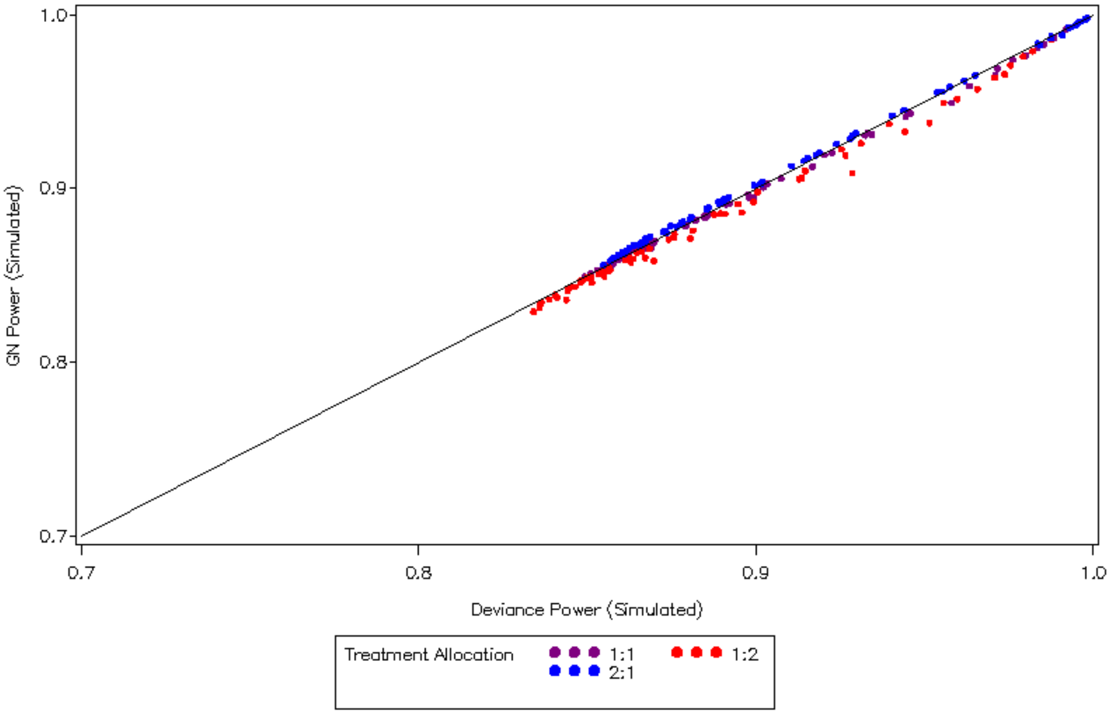


Figure 4.39 Comparison of Simulated Power for Stratified Risk Difference

Gart & Nam-SC and Deviance Simulated Power

By Treatment Allocation

Alpha=0.025, Strata Allocation=1:1, Delta=0

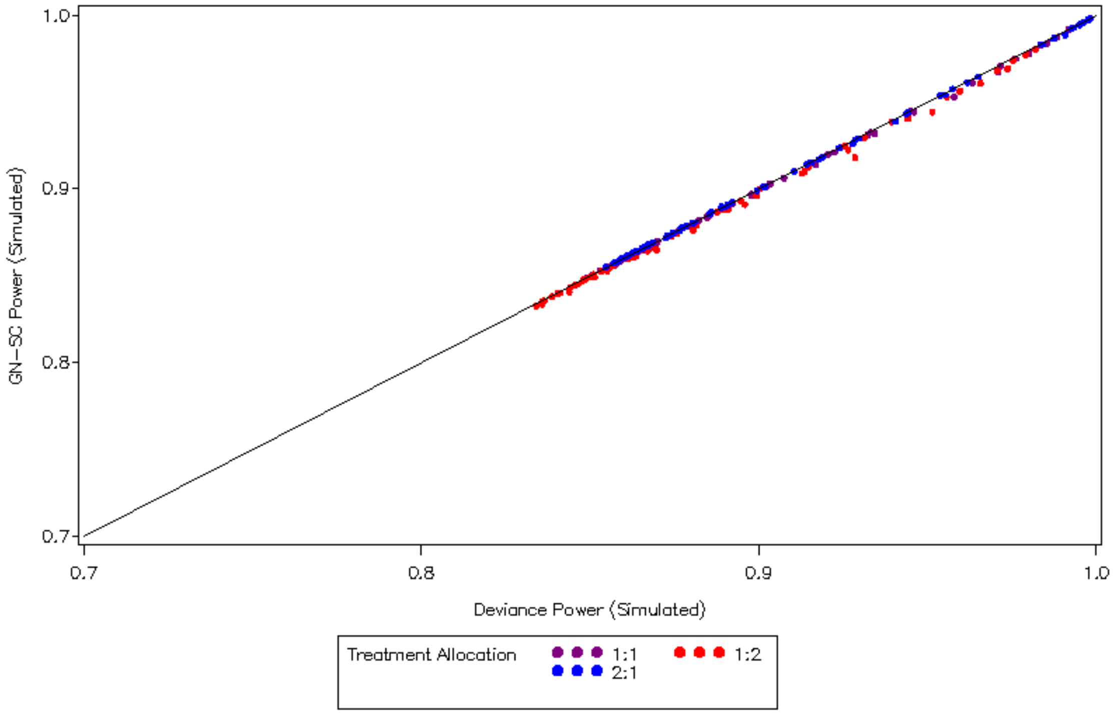


Figure 4.40 Comparison of Simulated Power for Stratified Risk Difference

Gart & Nam and CMH Simulated Power

By Pi C2

Alpha=0.025, Strata Allocation=1:1, Delta=0, Treatment Allocation=1:1, 2:1

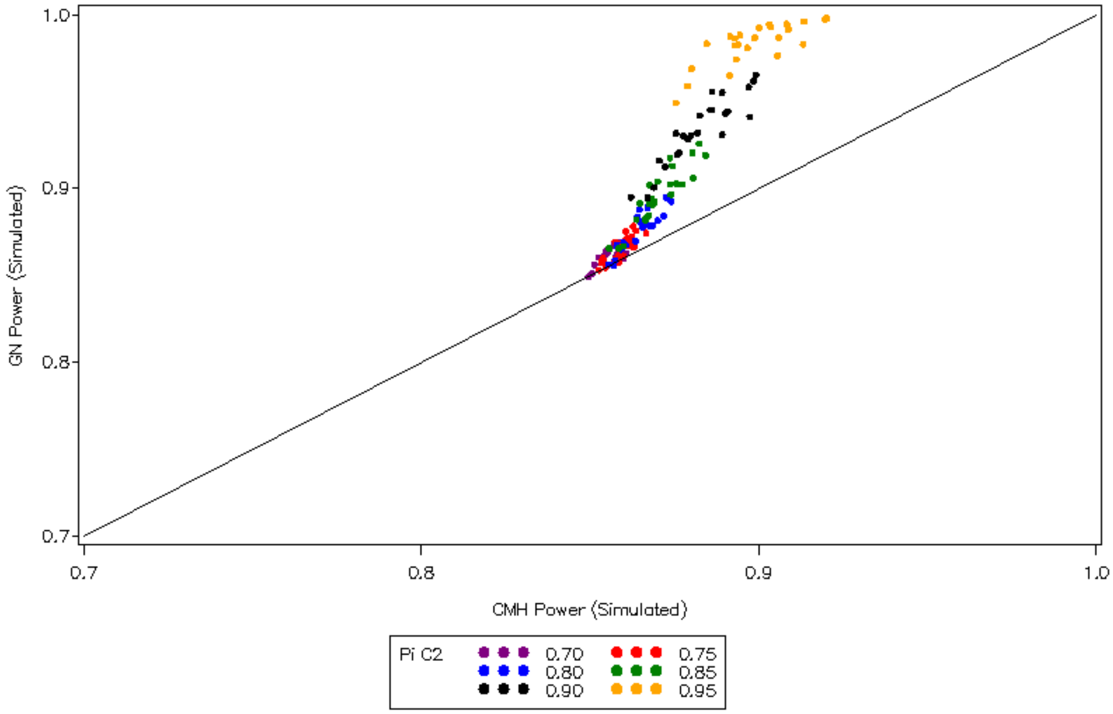


Figure 4.41 Comparison of Simulated Power for Stratified Risk Difference

Gart & Nam and WLS Simulated Power

By Pi C2

Alpha=0.025, Strata Allocation=1:1, Delta=0, Treatment Allocation=2:1

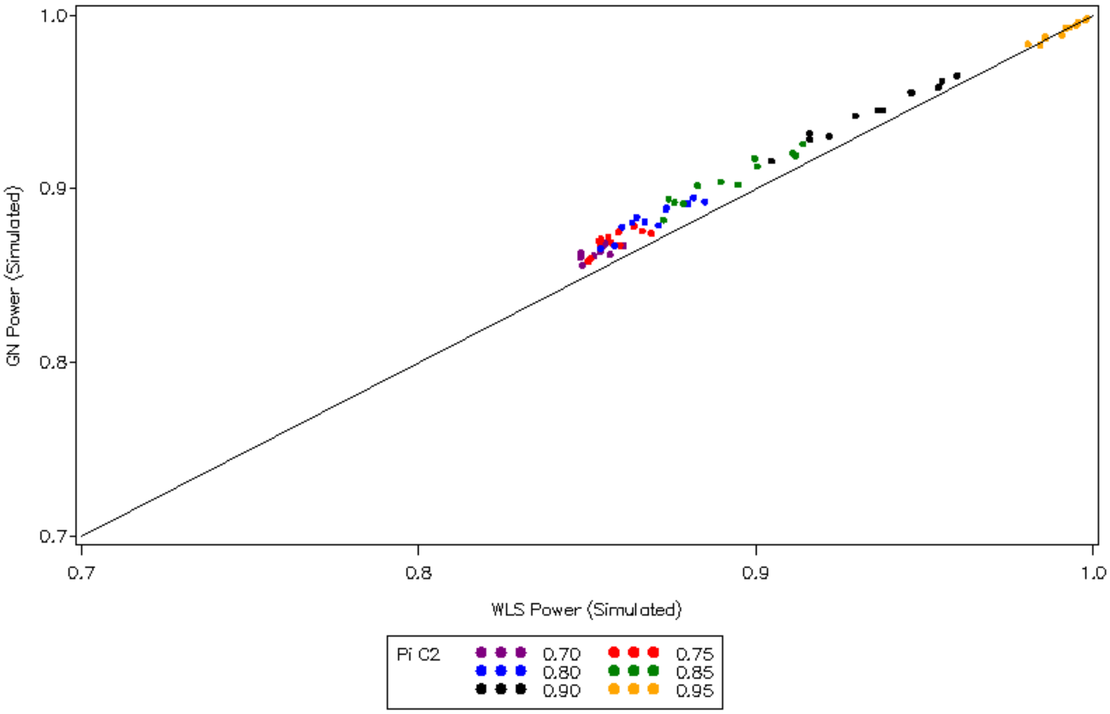


Figure 4.42 Comparison of Simulated Power for Stratified Risk Difference

Gart & Nam and Wald Simulated Power

By Pi C2

Alpha=0.025, Strata Allocation=1:1, Delta=0, Treatment Allocation=2:1

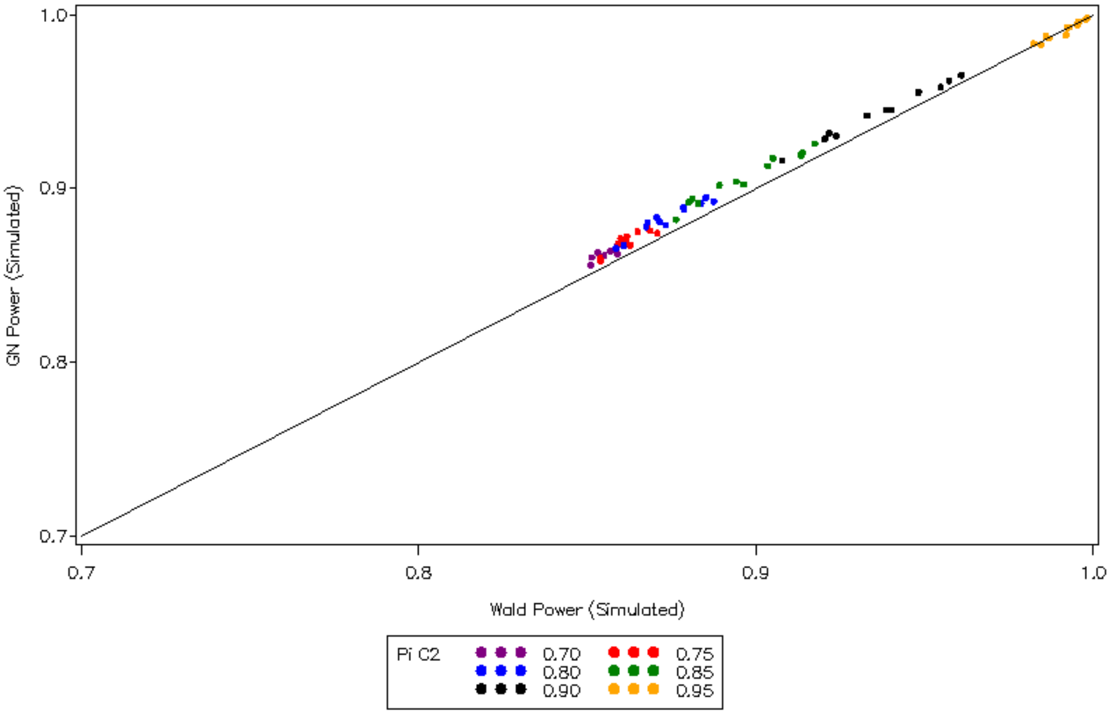


Figure 4.43 Comparison of Simulated Power for Stratified Risk Difference

WLS and Wald Simulated Power
By Pi C2
Alpha=0.025, Strata Allocation=1:1, Delta=0, Treatment Allocation=2:1

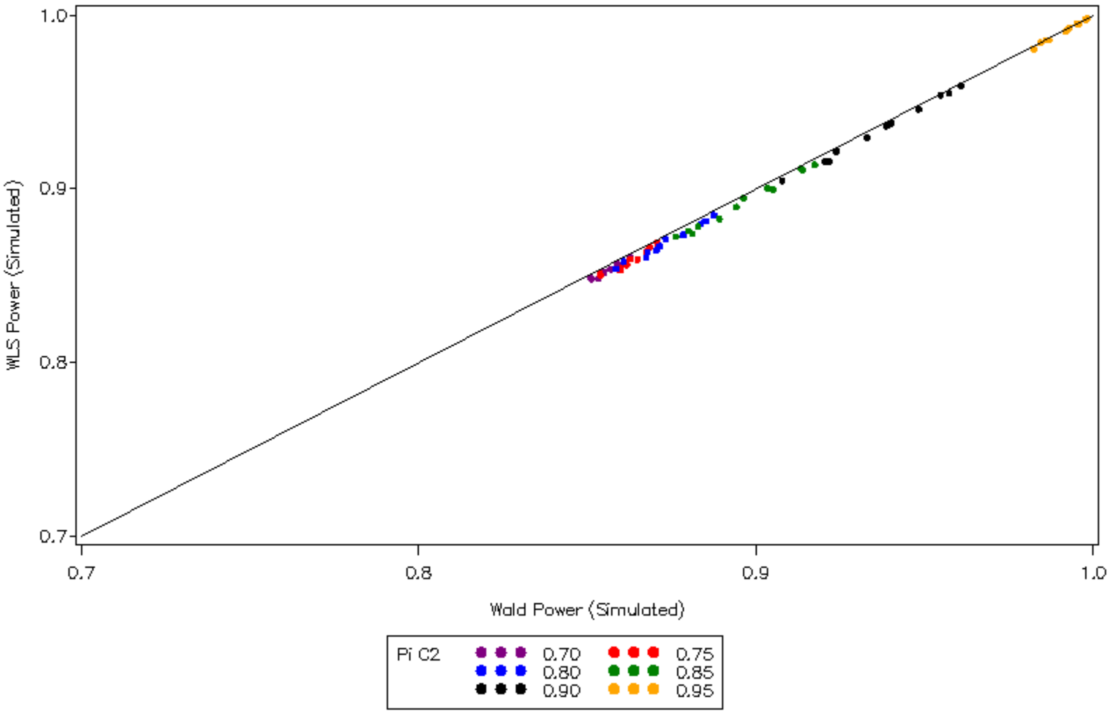


Figure 4.44 Summary of Simulated Type I Error for Stratified Risk Difference

By Strata Allocation
Alpha=0.025, Treatment Allocation=1:2

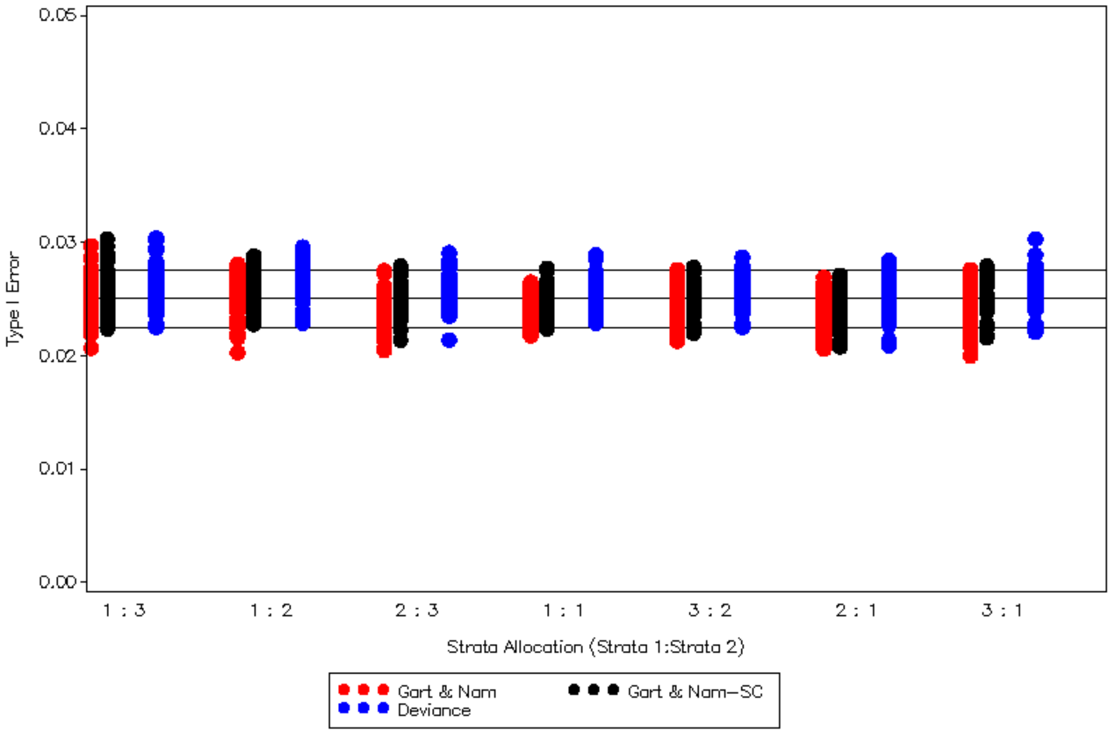


Figure 4.45 Summary of Simulated Type I Error for Stratified Risk Difference
By Strata Allocation
Alpha=0.025, Treatment Allocation 1:1

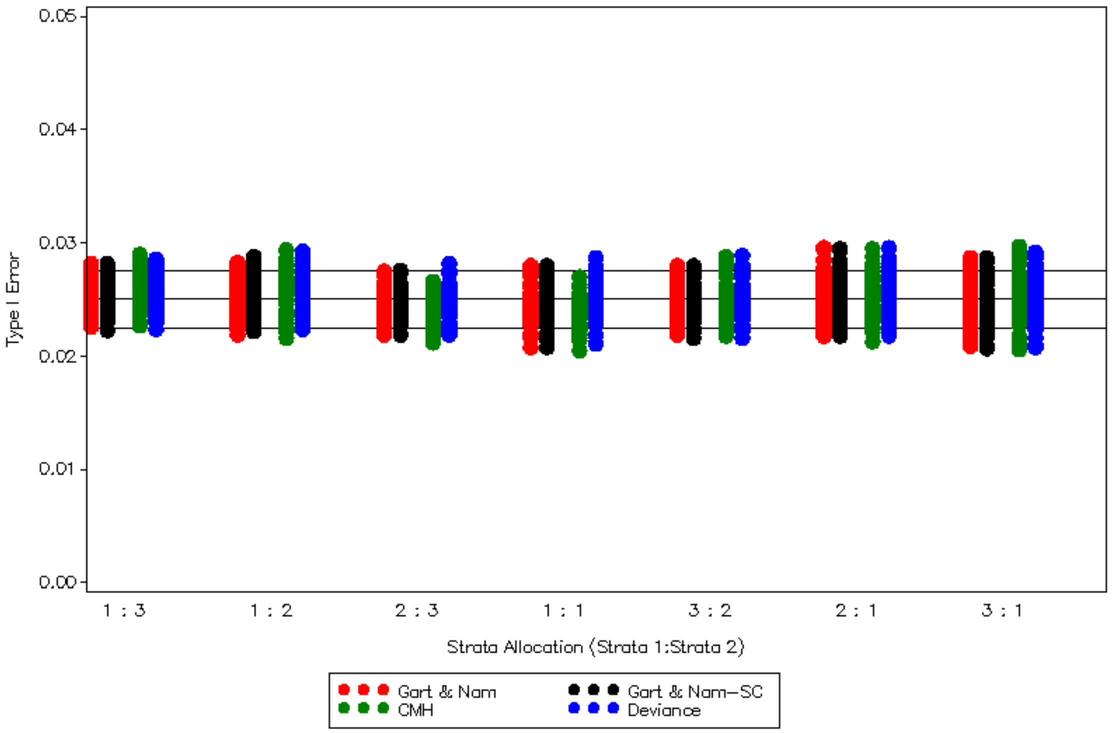


Figure 4.46 Summary of Simulated Type I Error for Stratified Risk Difference
By Strata Allocation
Alpha=0.025, Treatment Allocation 2:1

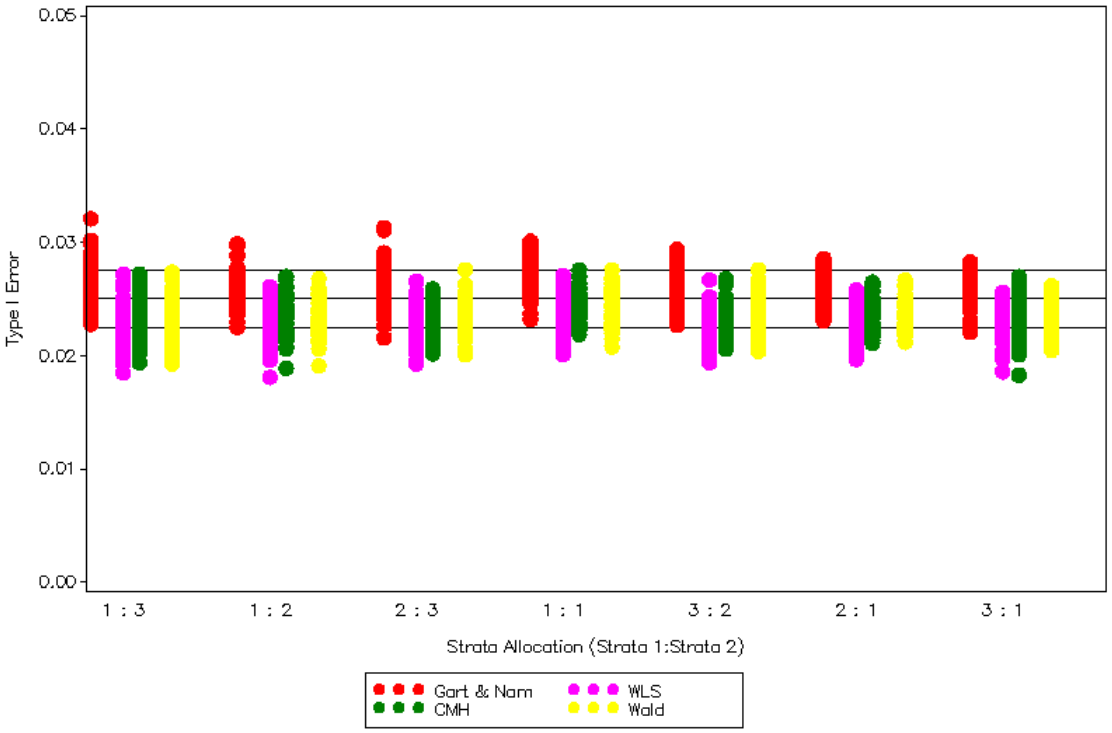


Figure 4.47 Summary of Simulated Power for Stratified Risk Difference
By Strata Allocation
Alpha=0.025, Treatment Allocation=1:2, Delta=0

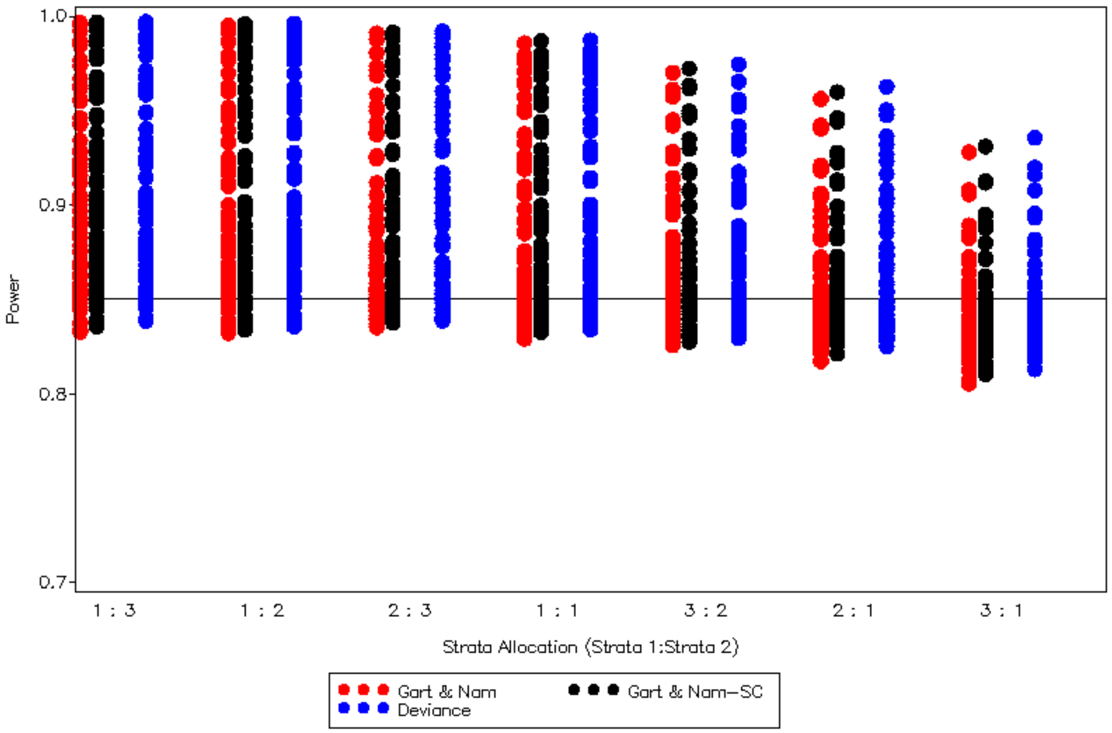


Figure 4.48 Summary of Simulated Power for Stratified Risk Difference
By Strata Allocation
Alpha=0.025, Treatment Allocation 1:1, Delta=0

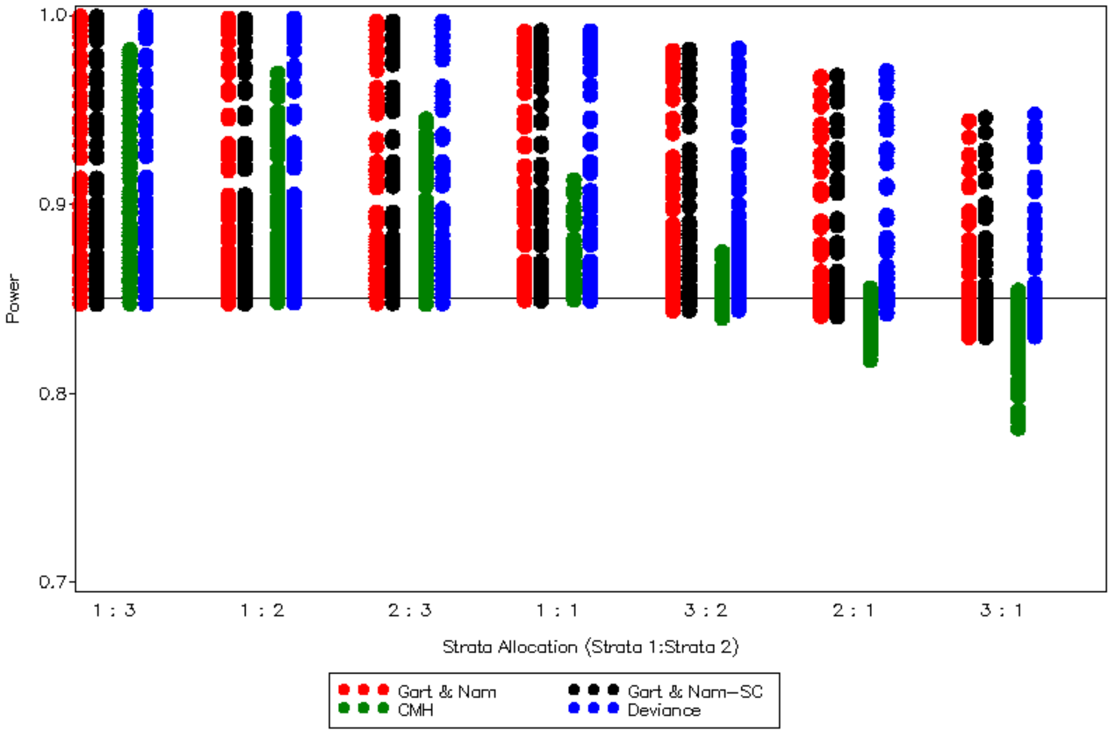


Figure 4.49 Summary of Simulated Power for Stratified Risk Difference
By Strata Allocation
Alpha=0.025, Treatment Allocation 2:1, Delta=0

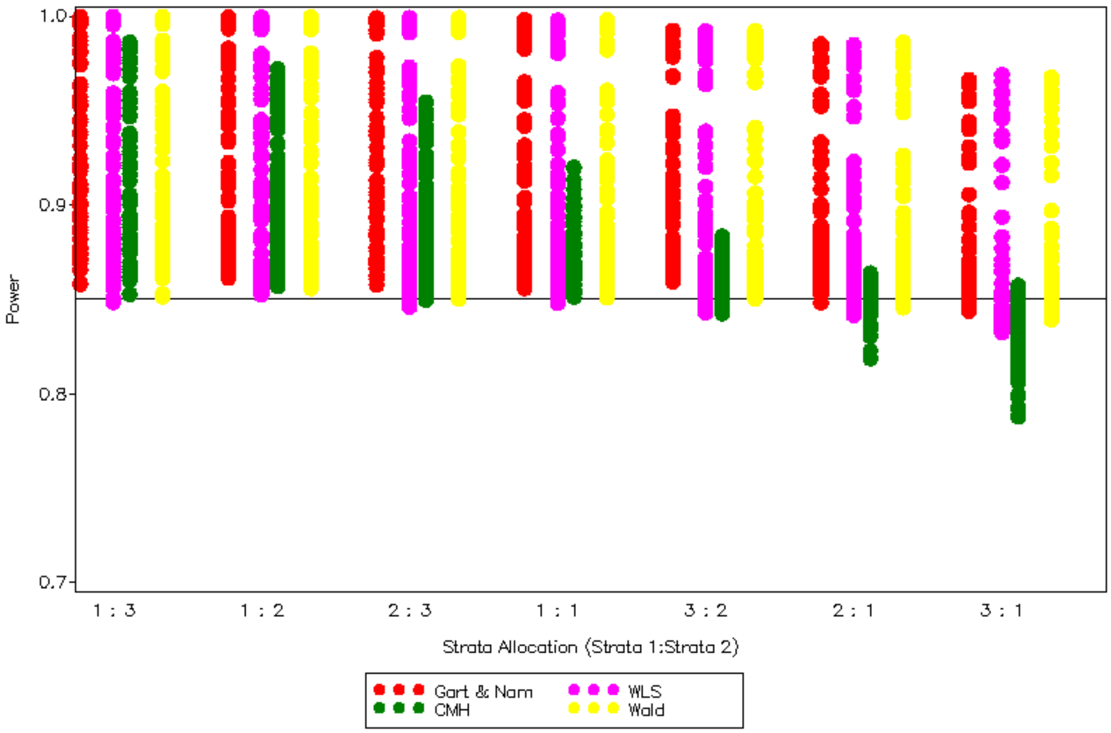


Figure 4.50 Comparison of Simulated & Calculated Power for Stratified Risk Difference
Gart & Nam Simulated and Nam Calculated Power
By Treatment Allocation
Alpha=0.025, Delta=0

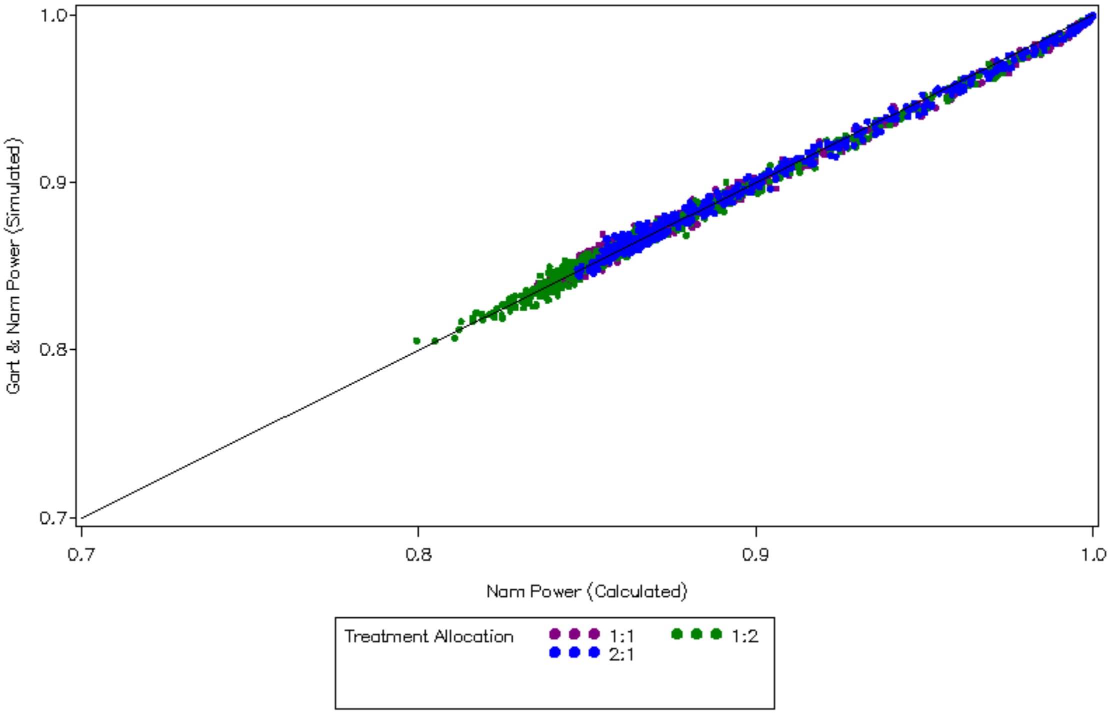


Figure 4.51 Comparison of Simulated & Calculated Power for Stratified Risk Difference
Gart & Nam Simulated and Nam Calculated Power
By Pi C2
Alpha=0.025, Delta=0

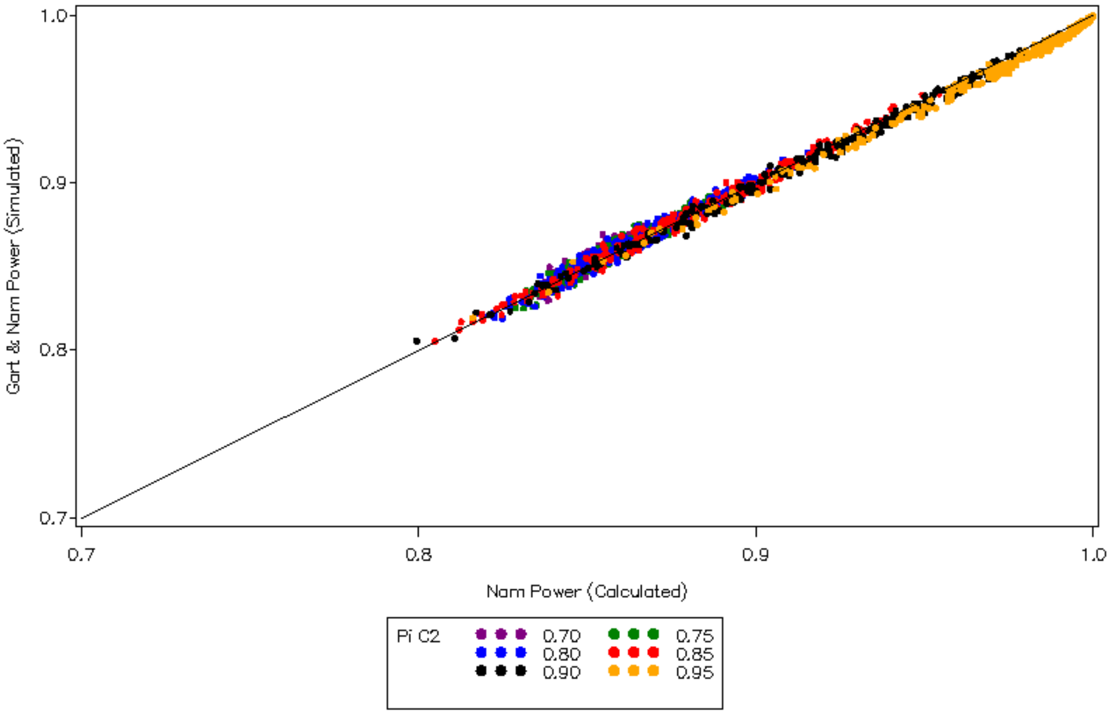


Figure 4.52 Comparison of Simulated & Calculated Power for Stratified Risk Difference
Deviance Simulated and Nam Calculated Power
By Treatment Allocation
Alpha=0.025, Delta=0

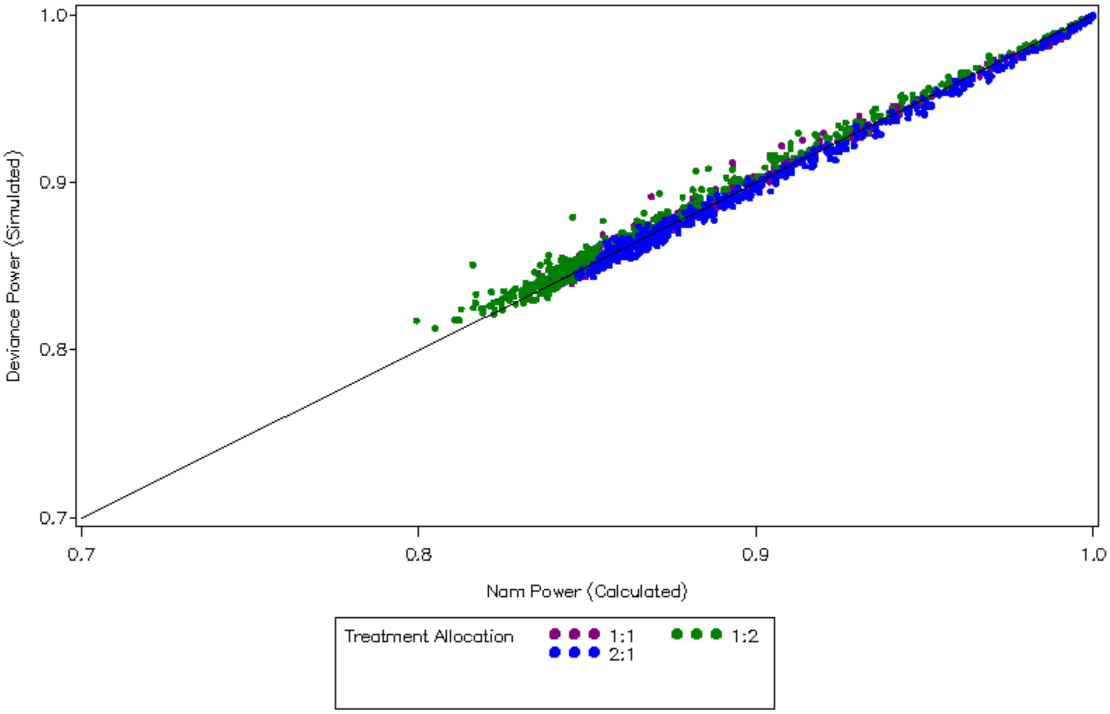


Figure 4.53 Comparison of Simulated & Calculated Power for Stratified Risk Difference

Deviance Simulated and Nam Calculated Power
By Pi C2
Alpha=0.025, Delta=0

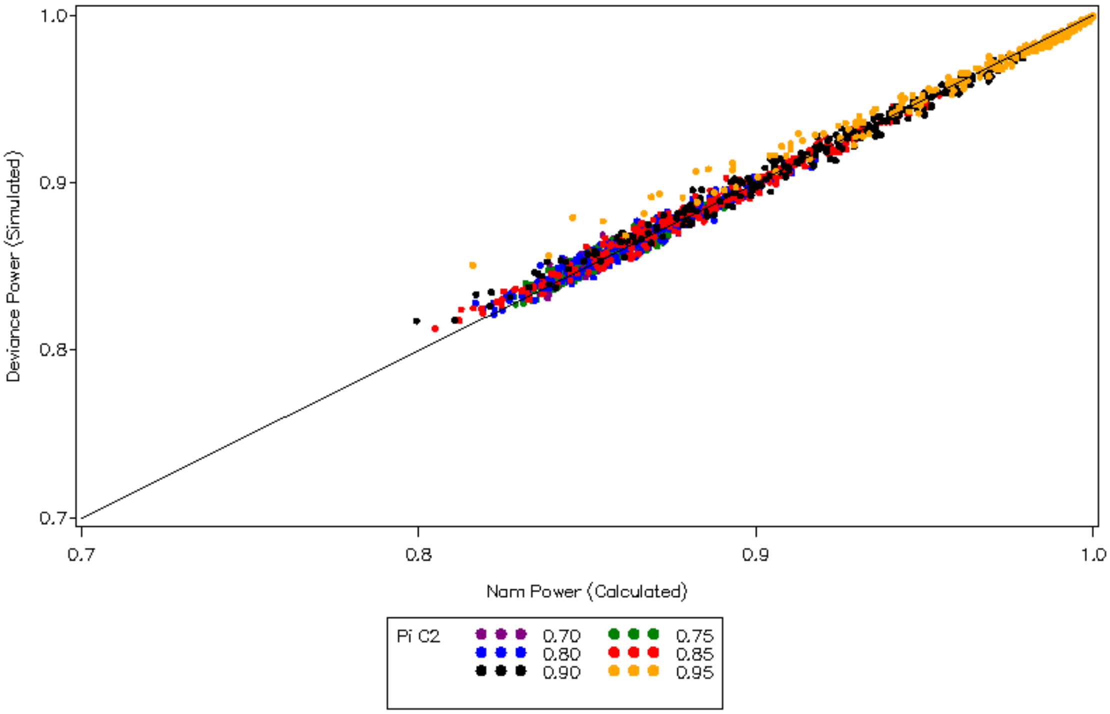


Figure 4.54 Comparison of Simulated & Calculated Power for Stratified Risk Difference

Wald Simulated and Nam Calculated Power
By Treatment Allocation
Alpha=0.025, Delta=0

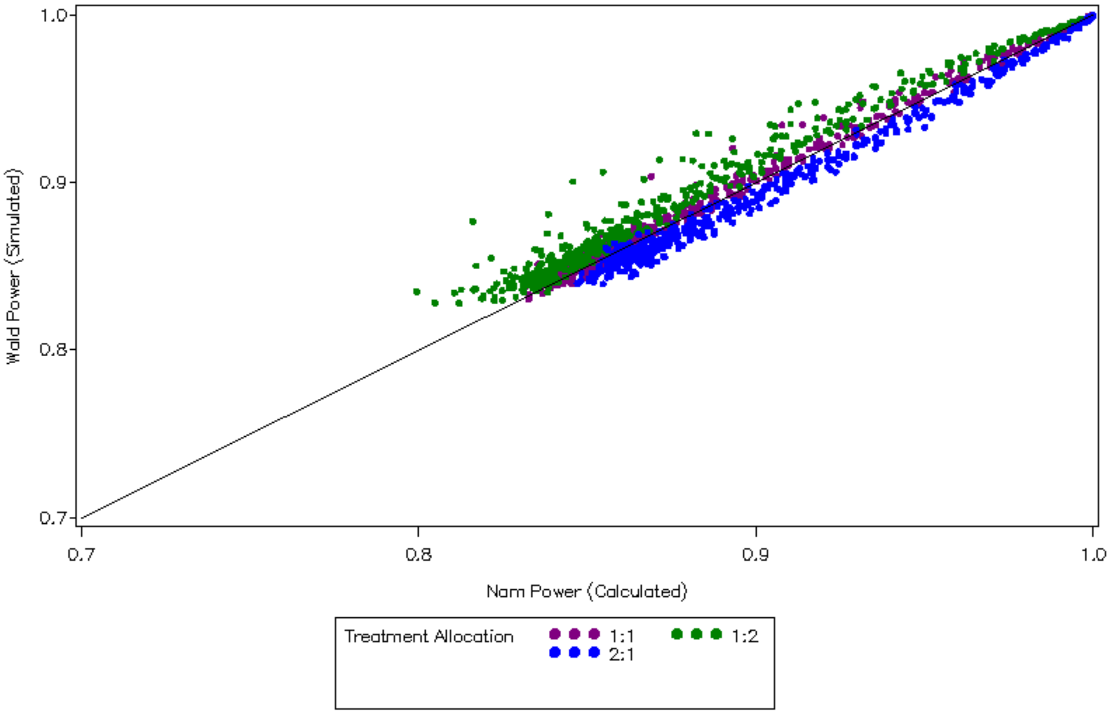


Figure 4.55 Comparison of Simulated & Calculated Power for Stratified Risk Difference

Gart & Nam Simulated and Wald Calculated Power

By Pi C2

Alpha=0.025, Delta=0

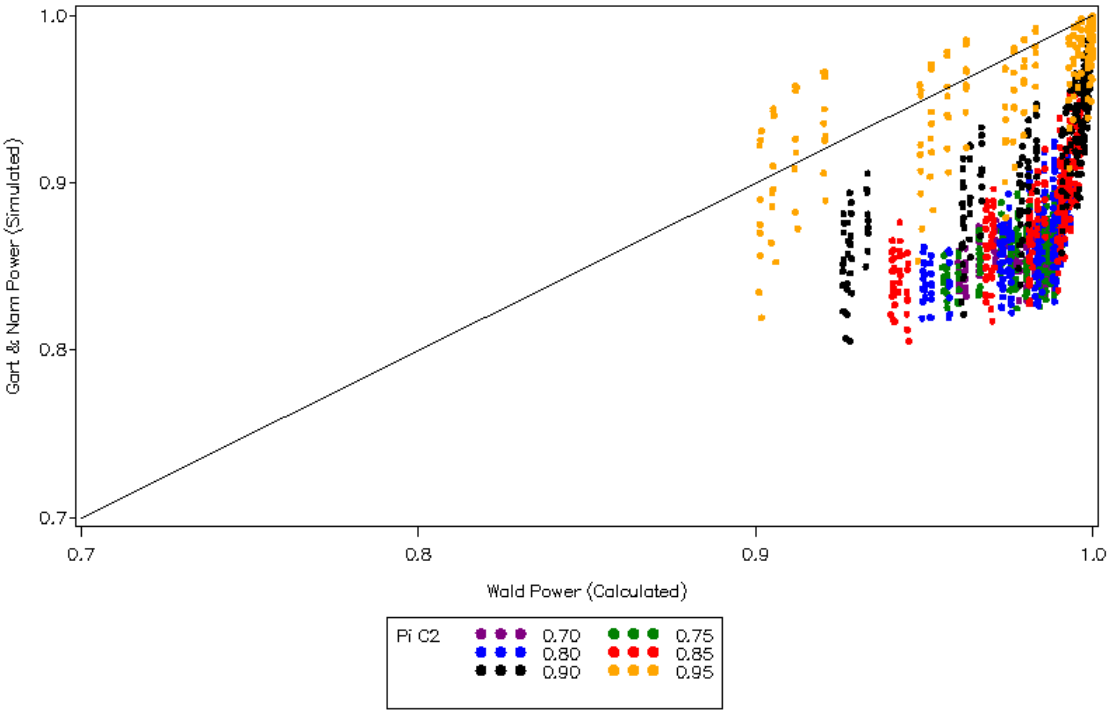


Figure 4.56 Comparison of Simulated & Calculated Power for Stratified Risk Difference

Wald Simulated and Wald Calculated Power

By Pi C2

Alpha=0.025, Delta=0

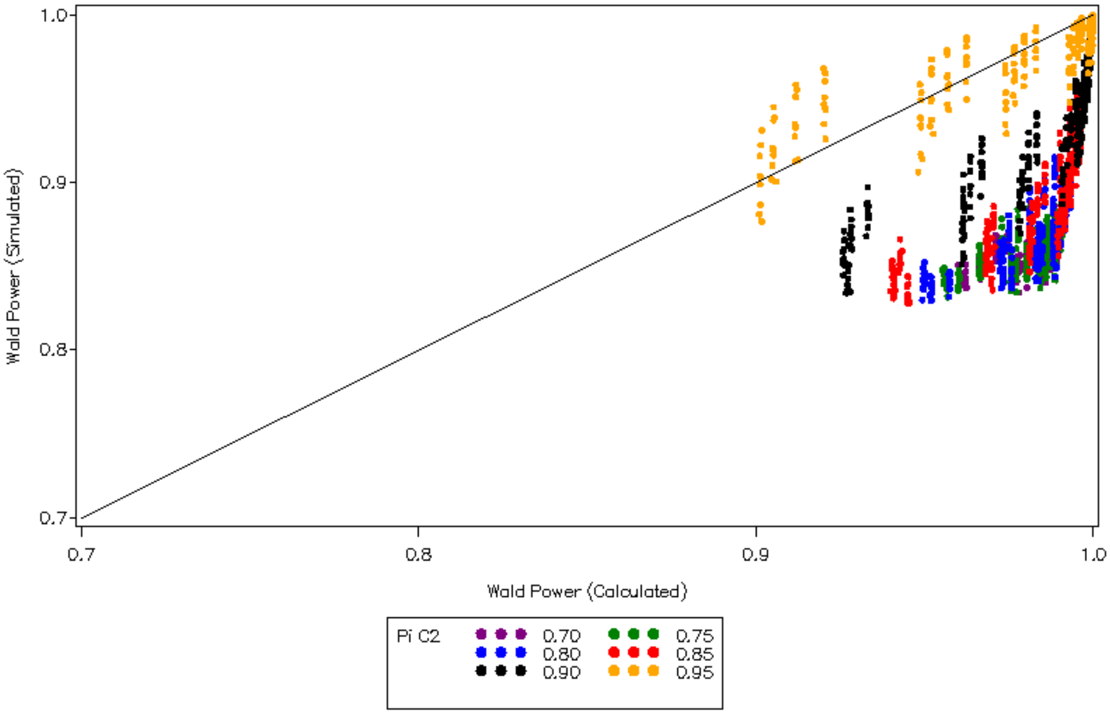


Figure 4.57 Summary of Simulated Type I Error for Stratified Risk Ratio
with Side Conditions on Individual Strata
Gart–SC Method
Treatment Allocation=1:2

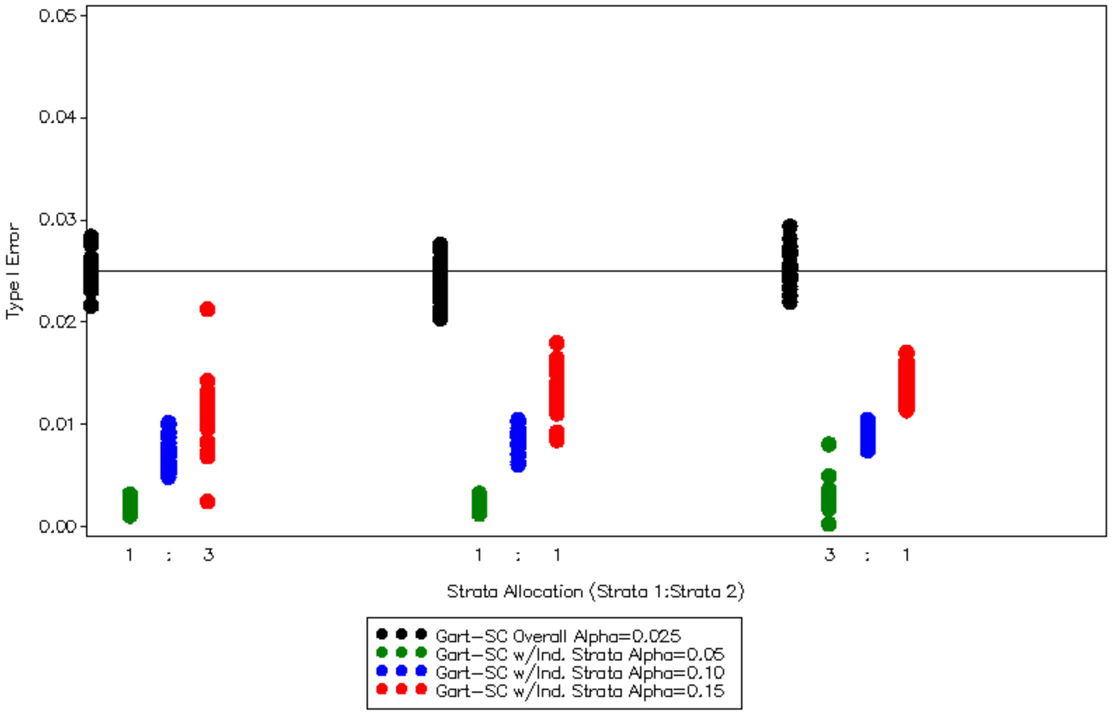


Figure 4.58 Summary of Simulated Type I Error for Stratified Risk Ratio
with Side Conditions on Individual Strata
Gart–SC Method
Treatment Allocation=1:1

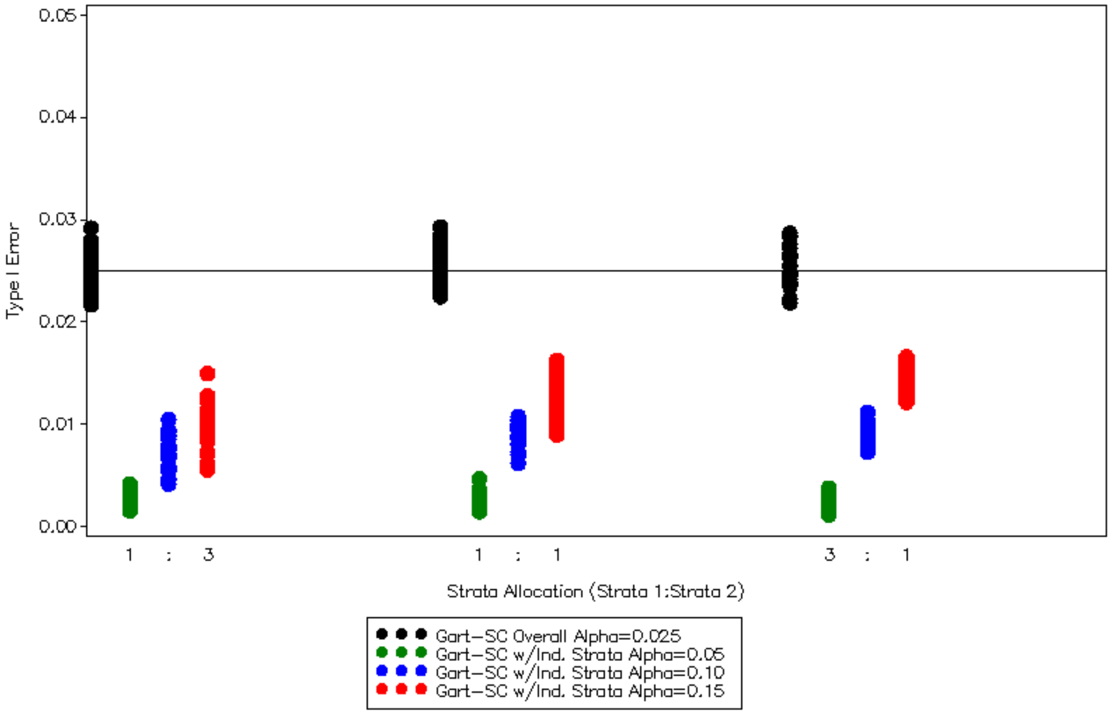


Figure 4.59 Summary of Simulated Type I Error for Stratified Risk Ratio
with Side Conditions on Individual Strata
Gart–SC Method
Treatment Allocation=2:1

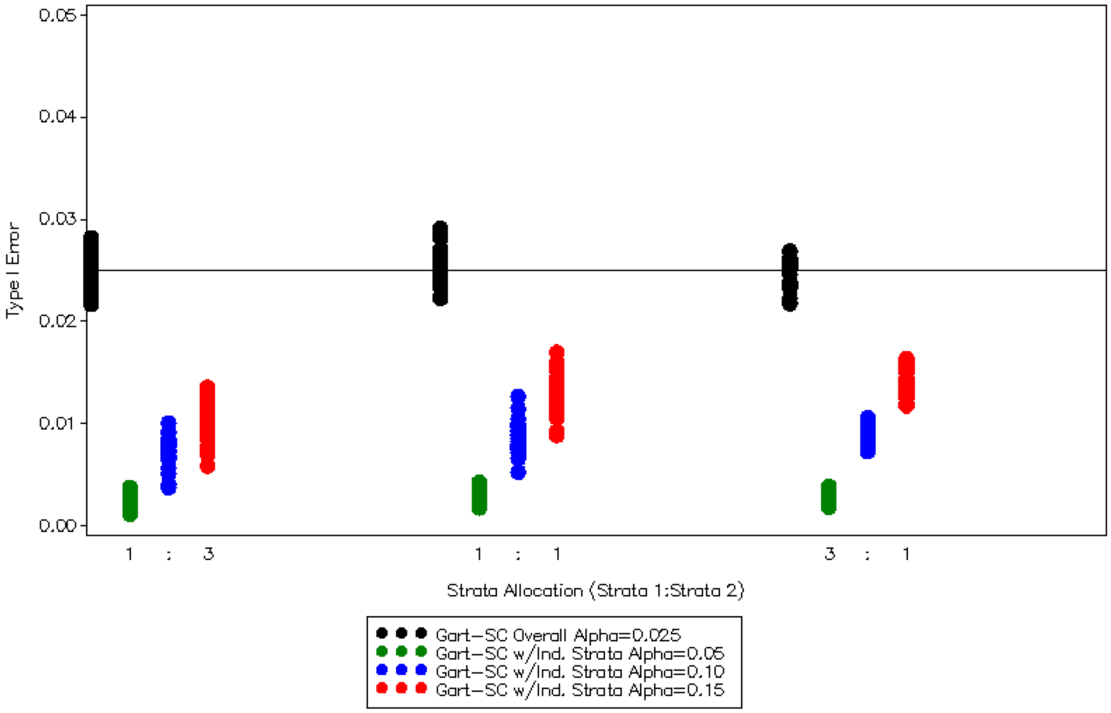


Figure 4.60 Comparison of Simulated Power for Stratified Risk Ratio
with Side Conditions on Individual Strata
Gart–SC Method
By Strata Allocation & (Pi C2 – Pi C1)
Theta=1, Overall Strata Alpha=0.025, Individual Strata Alpha=0.05

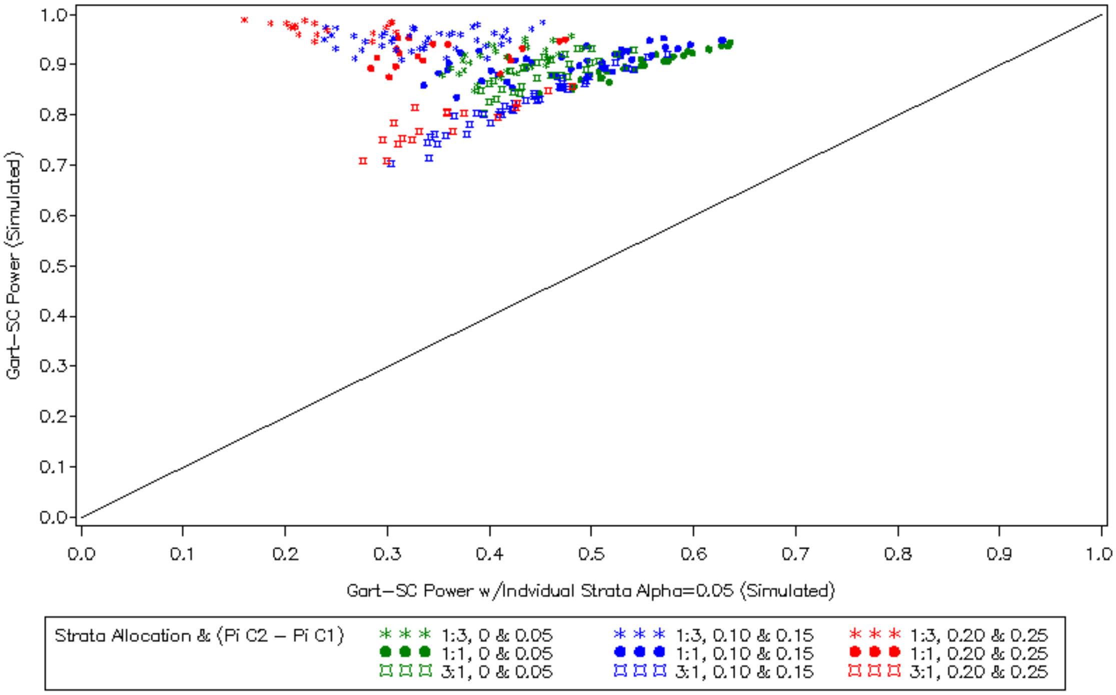


Figure 4.61 Comparison of Simulated Power for Stratified Risk Ratio
with Side Conditions on Individual Strata
Gart–SC Method
By Strata Allocation & (Pi C2 – Pi C1)
Theta=1, Overall Strata Alpha=0.025, Individual Strata Alpha=0.10

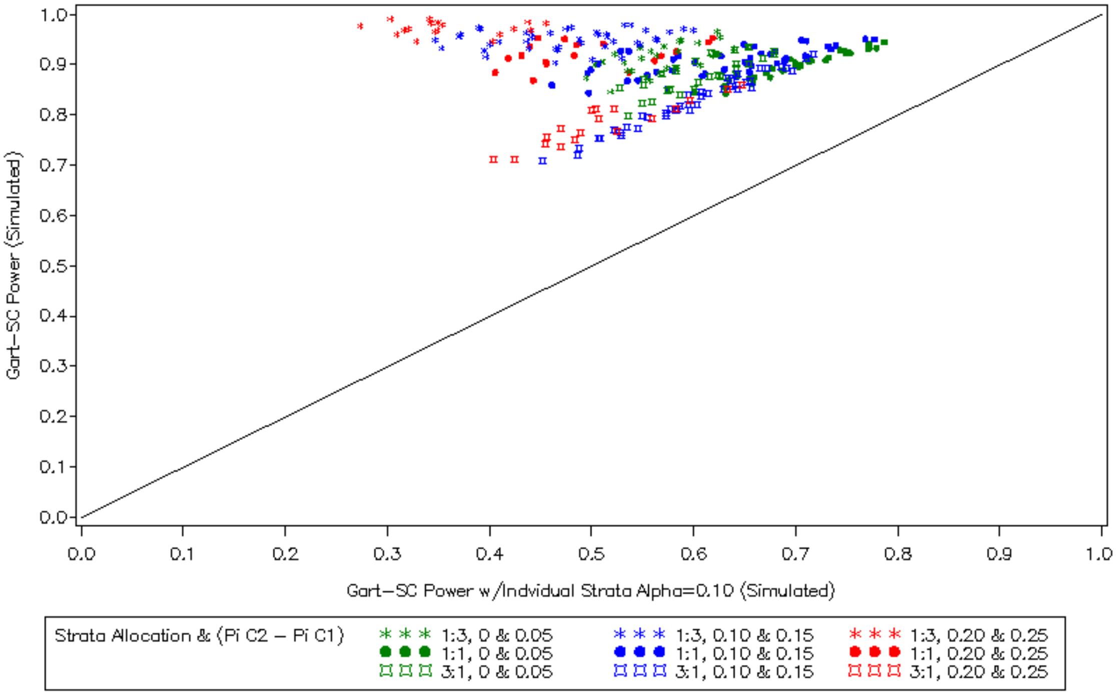


Figure 4.62 Comparison of Simulated Power for Stratified Risk Ratio
with Side Conditions on Individual Strata
Gart–SC Method
By Strata Allocation & (Pi C2 – Pi C1)
Theta=1, Overall Strata Alpha=0.025, Individual Strata Alpha=0.15

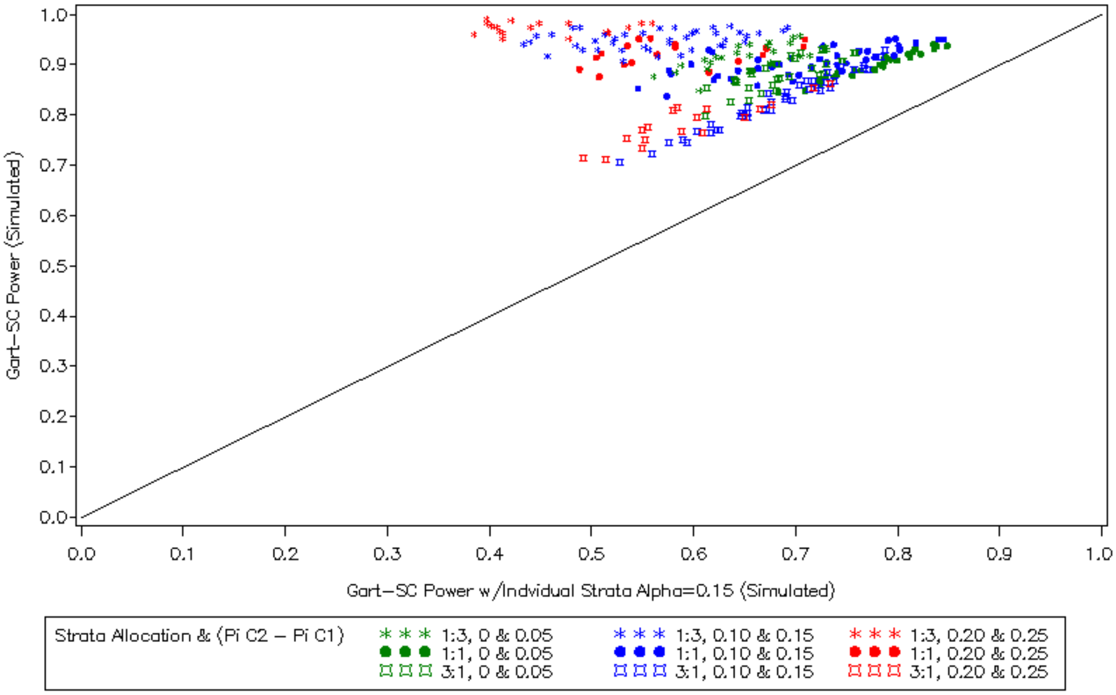


Figure 4.63 Summary of Simulated Type I Error for Stratified Risk Difference
with Side Conditions on Individual Strata
Gart & Nam–SC Method
Treatment Allocation=1:2

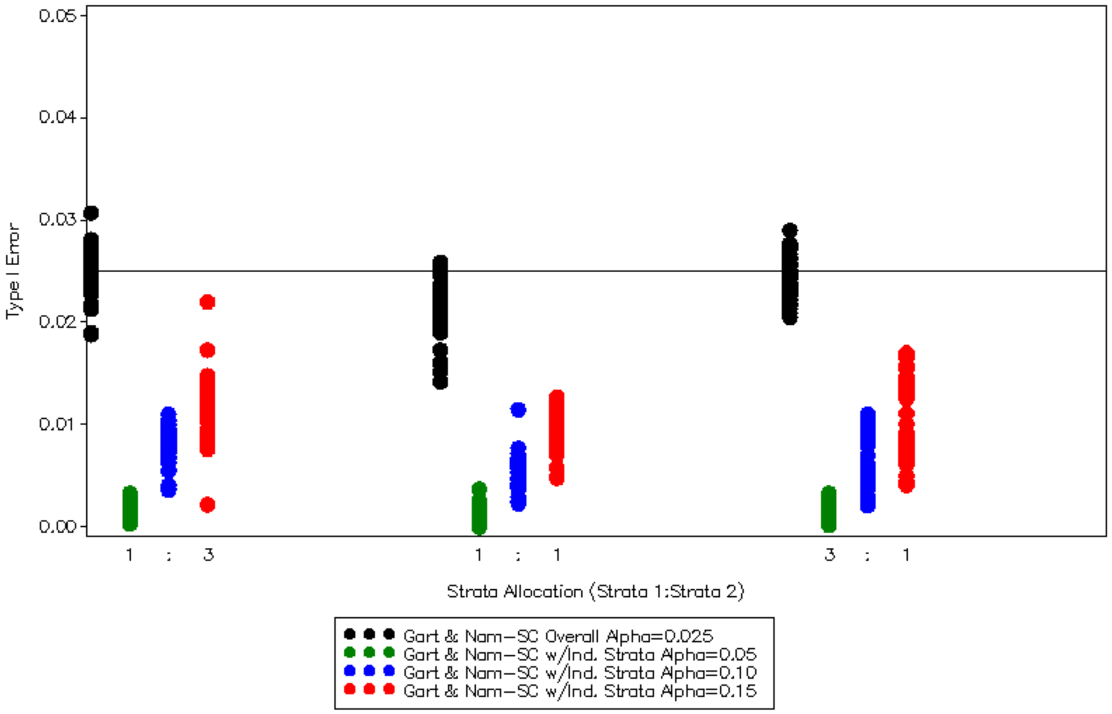


Figure 4.64 Summary of Simulated Type I Error for Stratified Risk Difference
with Side Conditions on Individual Strata
Gart & Nam–SC Method
Treatment Allocation=1:1

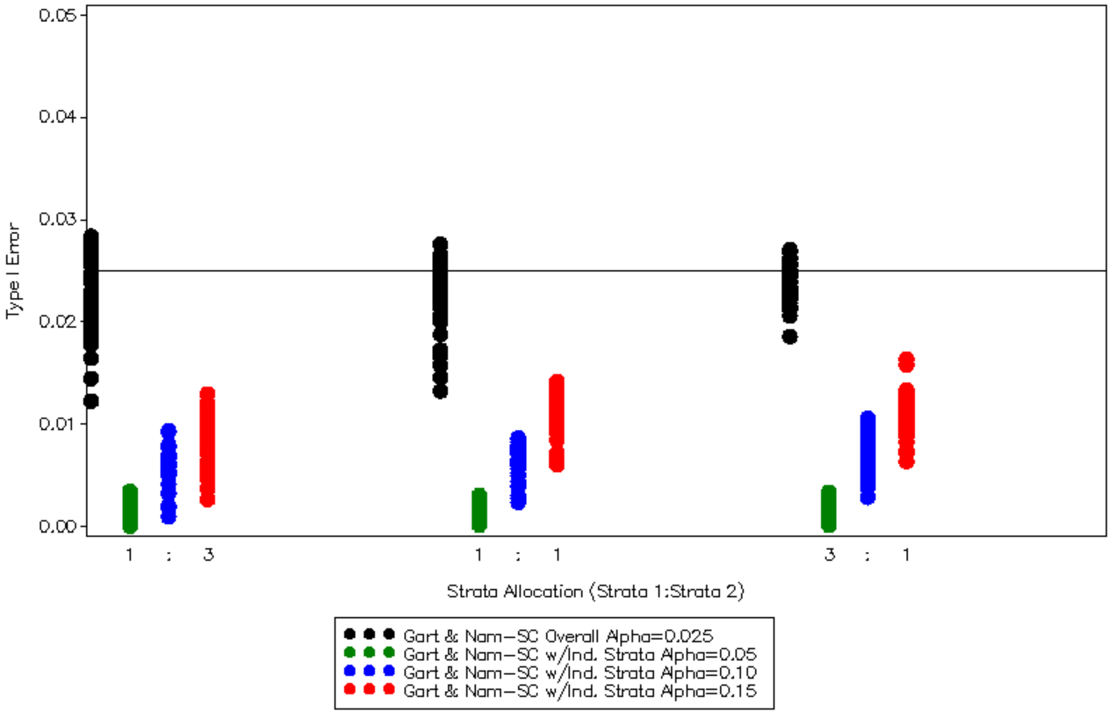


Figure 4.65 Summary of Simulated Type I Error for Stratified Risk Difference
with Side Conditions on Individual Strata
Gart & Nam-SC Method
Treatment Allocation=2:1

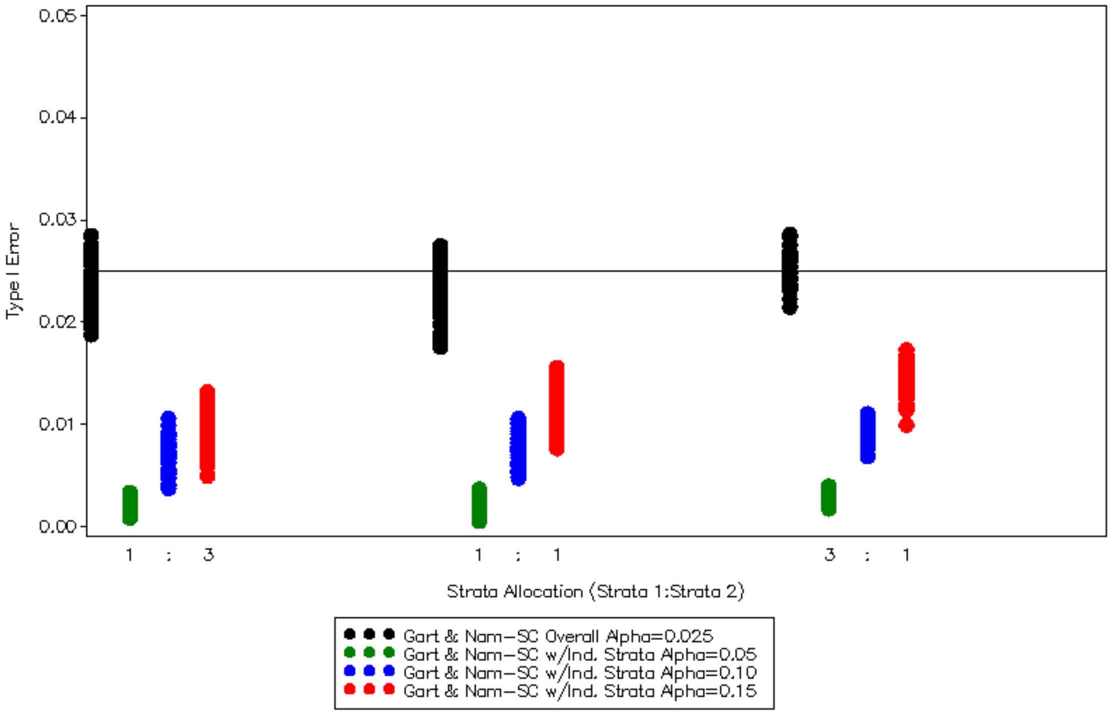


Figure 4.66 Comparison of Simulated Power for Stratified Risk Difference
with Side Conditions on Individual Strata
Gart & Nam-SC Method
By Strata Allocation & (Pi C2 − Pi C1)
Delta=0, Overall Strata Alpha=0.025, Individual Strata Alpha=0.05

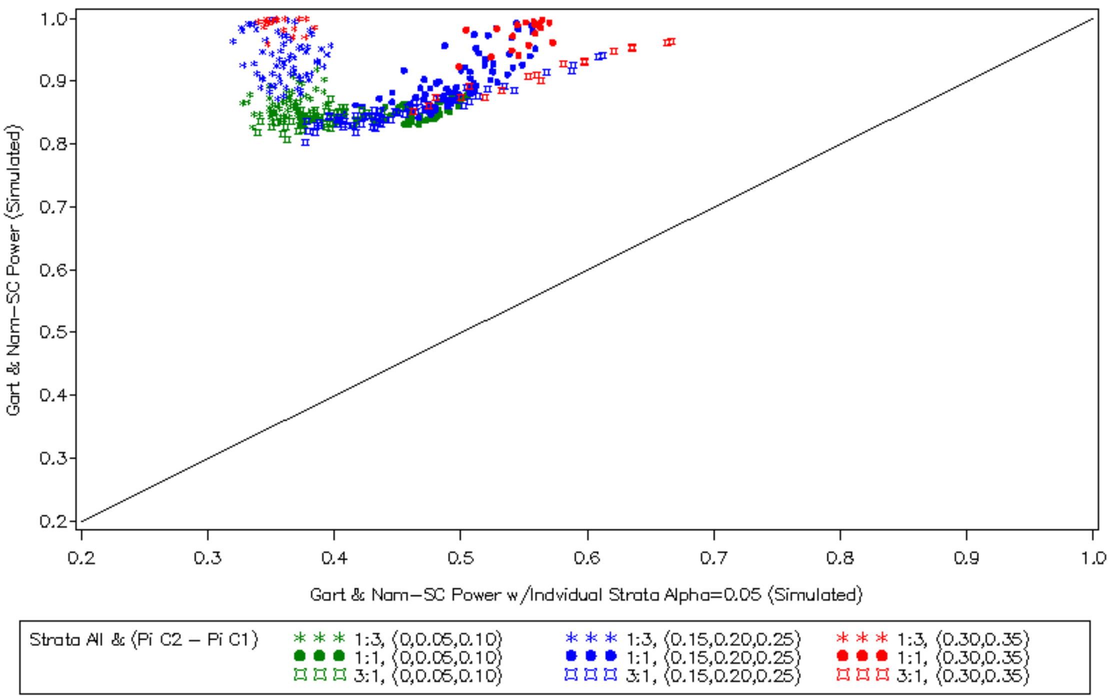


Figure 4.67 Comparison of Simulated Power for Stratified Risk Difference
with Side Conditions on Individual Strata
Gart & Nam–SC Method
By Strata Allocation & (Pi C2 – Pi C1)
Delta=0, Overall Strata Alpha=0.025, Individual Strata Alpha=0.10

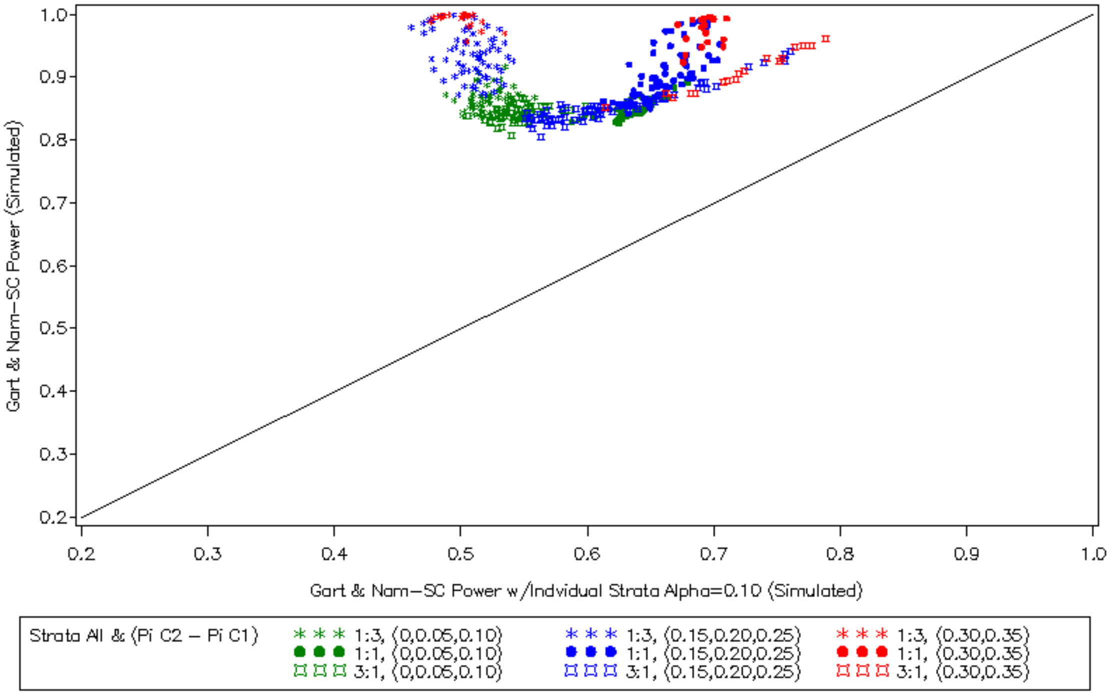
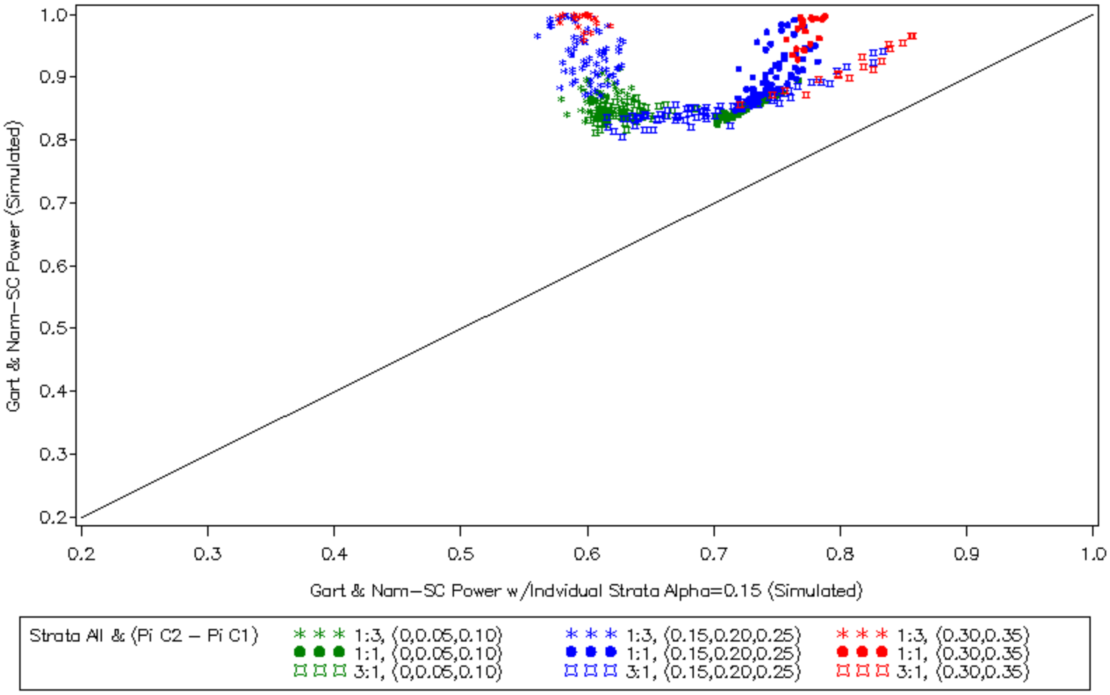


Figure 4.68 Comparison of Simulated Power for Stratified Risk Difference
with Side Conditions on Individual Strata
Gart & Nam–SC Method
By Strata Allocation & (Pi C2 – Pi C1)
Delta=0, Overall Strata Alpha=0.025, Individual Strata Alpha=0.15



DISCUSSION AND FUTURE RESEARCH

This body of work is meant to provide practicing statisticians with clarity around the design and implementation of non-inferiority clinical trials with dichotomous endpoints. Methodology for the risk ratio has been assessed and new methods have been developed including the Adapted Agresti method to address the non-inferiority hypothesis. Existing methods for the risk difference as well as the Deviance and Weighted Least Squares methods have been developed to address the non-inferiority hypothesis. Performance of these methods for type I error and power were considered as related to changing various population parameters of interest. Specifically, the sample size allocation to the treatment groups is influential in the performance of these methods. The treatment allocation and the other parameters specified in the simulations were sparsely addressed in the existing literature through more limited simulations, but these parameters were directly addressed within this research.

The methods for both the risk ratio and the risk difference were assessed in all situations specified, even if the counts were small or if the methods failed to produce an appropriate solution. In these cases substitutions were made to these methods with the exact odds ratio used for the risk ratio and the Agresti & Caffo method used for the risk difference because these replacements yield solutions in all scenarios. These modifications have not been previously considered in the literature, but this pattern of substituting with use of an alternate method would be used in practice if assumptions for the standard methodology were not met. Therefore, these simulations are more similar to performance of the methods in practical situations.

The need for cohesive methods to calculate sample size for the design of the trial and also analyze the resulting data has been addressed for the risk ratio and the risk difference. Existing calculations were assessed and new formulas for sample size were developed, including the Taylor Series method for the risk ratio. Discussion included comparison of the planned power versus the simulated power.

Additionally, the non-inferiority trial which has a placebo arm as well as an active-control arm has been discussed as related to methodology for analyzing the percentage of effect maintained by the test group over the control group, relative to the placebo group. Performance of these methods has been extensively assessed and corresponding sample size calculations related back to these methods for appropriateness of use. The effect of sample size allocation to the treatment groups and additionally the other parameters varied in the simulations were assessed as to the effect on type I error and power of the methods. This setting also presents an opportunity to understand the implications of requiring proof of non-inferiority using two separate but similar trials compared to using one larger trial. This one larger trial setting may require fewer subjects for the same power. Proof may also be required that the test treatment is superior to the placebo treatment group in addition to the non-inferiority of test to active-control. This also presents scenarios where the type I error and power for overall testing are maintained at appropriate levels.

Extensions of the methodology for the risk ratio and risk difference were developed and reviewed to address analysis using strata. Focus included two strata which may represent relevant sub-populations within the larger trial such as gender or disease severity. Sample size formulas were also included within this discussion to understand how the planned power relates to the resulting power at the end of the trial. Homogeneity of effects across strata is addressed by requiring the strata to reject the null inferiority hypothesis in addition to the overall stratified test having to reject the null hypothesis. This may be a regulatory requirement to ensure consistency of effect across the entire population of subjects and within the relevant subgroups in the trial. Additional research may include defining the necessary individual strata alpha levels and the stratified alpha level necessary to achieve a specified overall alpha level for the tests.

This assessment of the null hypothesis using stratified methods only included cases where the treatment effect was consistent across the strata and the null hypothesis of interest was also the same for each of the strata. The methodology presented should be able to address scenarios where the treatment effect and the null hypothesis of interest is not the same for the strata. Additional research would be needed to ensure appropriate performance of the methods in these scenarios. Also, sample size formulas should be assessed for appropriateness of use in these situations.