Adjusting for multiple test when endpoints are correlated

By

Ernest BIZIMANA (ernest@aims.edu.gh)

June 2019

AN ESSAY PRESENTED TO AIMS-GHANA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN MATHEMATICAL SCIENCES

DECLARATION

This work was carried out at AIMS-Ghana in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS-Ghana or any other University.

Student: Ernest BIZIMANA

Supervisor: Prof. Arne Ring

ACKNOWLEDGEMENTS

First of all, I want to thank all the students, tutors and staff involved in my AIMS journey. My sincere gratitude to my supervisor Prof. Arne Ring for his continuous guidance, patient, motivation and invaluable ideas and knowledge throughout this project. Special thanks to my Tutor Evelyn Arthur Anto for her patience and comments that greatly improved this project. I am thankful to my family and friends for their encouragement and support.

THANK YOU to anyone who has donated in any way shape or form!

Abstract

In clinical trial multiple test involves the control of family wise error rate at pre-specified significance level(α). By definition, the family wise error rate is the probability of observing at least one false rejection in m performed test. Different methods can be implemented to control that probability but we have to account for the correlation between the data. Those methods could work better or worse when the data are correlated. So we investigate different methods of controlling family wise error rate based on correlation. The methods to be investigated for family wise error rate control are; Adjusting for fixed assumed correlation, Plug-in estimated correlation and Improved Berger and Boos method. Based on the results of each method, the investigation help us to know which approach can adjust the family wise error rate boundaries more precisely to the significance level. The research found suggest to use Berber and Boos by considering the sample with larger sample size and larger confidence level to compute the lower bound of the nuisance parameter which leads to the accurate control of family wise error rate boundaries.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background of study

Nowadays, many people make unmanageable statements on different aspect of their daily activities by the choice of idea or claim. Researchers, scientific analysts often have hypothesis about particular points of their work areas. However, these hypotheses are not always true and can be tested for affirmation or rejection according to different motives. During this realistic process, you take your ideas, and do a reality check against it via data analysis, the same way researchers or analysts collect data that enable them to support or doubt the hypothesis. Statistical testing helps researchers or analysts to draw a conclusion with evidence based on data.

The objective in all tests is to set the type I error level at lower value and then to use a statistical test which minimizes the type II error level for a given sample size. If the hypothesis are more than one then multiple test is expected and when several tests are performed, the probability for wrong decisions can increase. Avoiding this, some methods are required to restrict the chances of deciding wrongly. These methods could work better or worse when the data are correlated. In this project, the key point for multiple test is the probability of observing at least one false rejection in m performed test, and this probability is call family wise error rate (FWER) [3]. Most of multiple test correction methods minimize this probability to be less than the significance level (alpha). In other words, controlling type I error rate in multiple testing is frequently done by splitting the accepted and pre-specified type I error rate alpha (α), and then testing the various null hypotheses at fractions of alpha.

Multiplicity control is an important statistical issue in clinical trials, where the family wise error rate should be much controlled. A number of primary endpoints can be investigated with the aim of providing convincing evidence [4]. In these situations planning of the sample size becomes more complex due to the different alternative hypotheses related to the different endpoints and due to the assumed correlation between endpoints. The primary endpoint is the outcome by which the effectiveness of treatments in a clinical trial is evaluated, the fairness of endpoints is that the primary endpoints should be defined identically for both groups.

The presence of more than one primary endpoints in a clinical trial usually means that some adjustments of the observed p-values for multiplicity of tests may be required for the control of family wise error rate at a pre-specified significance level. In this study we discuss statistical concerns associated with some commonly used multiple endpoint adjustment procedures (Bonferroni). We also present some Monte Carlo simulation results to demonstrate the performance of correlation between endpoints in controlling the type I error rate.

This study will be focussing on two groups (A & B) with endpoint for each. The endpoints data are bivariate normal distributed with unknown correlation between them.

1.2 Problem statement

During multiple test when the number of hypotheses increase the probability of observing at least one false rejection also increases. This is the big challenge because it is needed that this probability remains below the desired significance level(α). This probability is also called family wise error rate[5].

$$FWER = 1 - (1 - \alpha)^m,$$

where α is the significance level and m is the number of performed test.

There are different methods to deal with this issue like Bonferroni[6] in case of independent tests, Sidak SS[7] and others, these methods can be applied to control the family wise error rate. In this study, we will work with two independent tests with correlated endpoints.

The purpose is to control the family wise error rate when the endpoints are correlated. In other words, we need to keep the family wise error rate boundaries at fixed level means around significance level. Various methods will be implemented based on correlation between endpoints.

1.3 Objective of the study

The main objective of this study is to investigate different methods of adjusting for multiple test. A valid adjustment for multiple test must account for the correlation between data. In this study the family wise error rate control requirement is expressed in the following Equation,

$$\text{FWER} = P\{\text{Reject at least one true null hypothesis}\} \leq \alpha.$$

1.3.1 Specific objectives.

a. Adjusting family wise error rate boundaries using three methods:

Adjust for fixed assumed correlation: This method involves assuming the correlation leading to the highest possible type I error rate to adjust the boundaries of the FWER around significance level.

Plug-in estimated correlation: Instead of using the assumed correlation, this method uses the estimated correlation between endpoints to control the FWER boundaries around significance level.

Improved Berger and Boos method: Instead of using the estimated correlation between the endpoints, this method uses the lower boundary limit of estimated correlation.

b. Investigate among these methods which is more conservative than others in adjusting or controlling FWER boundaries when endpoints are correlated.

1.4 Research questions

Q1. What is the issue of multiplicity?

Q2. which correlation estimator can control better the FWER ?

Q3. What is the best method for multiplicity adjustment with correlated endpoints ?

1.5 Significance of the study

One can say that the issue in multiplicity is to find a method that treats multiple test wisely, which means finding a method that increases the power substantially while at the same time tightly controlling the false rejection error. It is necessary to revise the theory of multiple test by considering the attentiveness of the critical boundary underlying problem. Since this study is relevant to clinical trials (health-related interventions used to evaluate the effects on health outcomes) we will focus on FWER adjustment by taking into account the correlation between endpoints to avoid the rejection of true null hypothesis.

This study investigates the performance of those methods that correct the critical boundary underlying case. The underlying case can be expressed by $(\rho, \alpha) < \max(FWER)$, where (ρ, α) is a critical value.

1.6 Organization of the study

In Chapter 2 will discuss some basic definitions and preliminaries that are more useful in this study. Chapter 3 discusses the issues of multiplicity when endpoints are correlated and the methods to handle it. In Chapter 4 we will discuss the results for all methods used in chapter 3. Chapter 5 will summarize the research findings and recommendations for future study as an outcome of this research.

2. Basic definitions and preliminaries

In this chapter we define some terminologies which are useful for us to understand this essay.

2.1 Statistical testing

Statistical testing is the process involved in hypothesis testing. It is also called confirmatory data analysis.

This section discusses some terms involved in statistical testing procedure like; statistical hypothesis (null hypothesis, alternative), level of confidence, level of significance, P-value and critical value then statistical examples. This process is done through taking sample in a population.

2.1.1 Statistical hypothesis. Statistical hypothesis is an idea/claim or promise with limited evidence to start investigation or further study. It can also be defined as a statement or declaration about the distribution of one or more random variables.

2.1.2 Null hypothesis(H_o) and Alternative hypothesis(H_a). Null hypothesis is the hypothesis of no difference. Alternative hypothesis is the opposite of null hypothesis since they are mutually exclusive which means if one occur another can not.

We can take physical example to express hypothesis;

It is mutual that Ernest eat food in average of $2kg$. A cooker claim that Ernest will no longer eat $2kg$ after finishing his studies. Then hypothesis will be

$$H_o: \quad \mu = \quad \mu_o, \quad \text{against} \begin{cases} H_a : \mu \neq \mu_o, & \text{(two tailed)} \\ H_a : \mu > \mu_o \\ H_a : \mu < \mu_o \end{cases} \tag{2.1.1}$$

in this case $\mu_o = 2kg$.

2.1.3 Confidence Level. How sure are we to set our null hypothesis before we start the study. The confidence level is expressed in percentages. Assume the confidence level to be $c = 95\%$.

2.1.4 Significance level (α). Significance involves knowing the limitation which will help us to make our decision easily.

Mathematically, is given by

$$\alpha = 1 - c,$$

so, if we take our confidence interval $c = 95\%$, then $\alpha = 0.05$ which can help us to determine critical value according to the chosen statistical test, as all hypothesis testing are liable to errors, here are two basic types of error;

Type I error is the error due to the rejection of null hypothesis while it is true. The probability of committing type I error is denoted by alpha (α).

Type II error is the error due to the rejection of alternative hypothesis while it is true. It can be defined as incorrect acceptance null hypothesis. The probability of committing the type II error is denoted by beta (β).

2.1.5 P-value and Critical value. P-value is the probability of test statistic. Both P-value and Critical value help to take a decision during our statistical test.

If the P-value is small (i.e less than the significance level), we reject the null hypothesis which implies that our sample gives reasonable evidence to support the alternative hypothesis. If the P-value is not small (i.e greater than the significance level), we fail to reject null hypothesis which means that, we are right with the null hypothesis .

In hypothesis test, Critical values depend on a test statistic, which is specific to the type of test. Critical values are essentially cut-off values that define regions where the test statistic is unlikely to lie. This means if the value of test statistic exceed the critical value then the null hypothesis is rejected (see Subsection (2.1.8)).

Again, if the value of test statistic is found to be equal to the critical value, we reject the null hypothesis.

2.1.6 Applications of Hypothesis test. The hypothesis testing have many different test [8, 9], some concern with; averages, different between means, sample proportional and different between proportions, variances.

In this part, we will see some example of the test concerning averages. We suppose that the variable is approximately normally distributed or consider that the sample are from the normal populations or that they are large enough to justify normal approximations.

2.1.7 Test Concerning Averages. Like in interval estimation for population mean, there are three cases to consider; population variance (known or unknown σ^2), sample drawn from population approximately normal or large and the data are independent identical distributed which means all random variables have the same probability distribution.

Usually the hypothesis is set up as in Equation (2.1.1), and the test statistic depends on whether the variance σ^2 is known or unknown.

When σ^2 is known, we use z-test [8, 9] with the following test statistic

$$z = \frac{\bar{X} - \mu_o}{\sqrt{\frac{\sigma^2}{n}}}, \quad n > 30.$$

Where; $\bar{X}$ is sample mean, μ_o population mean, and n the sample size. In this test, population may be normal or not. Referring to Equation (2.1.1), the critical regions for the alternatives within z-test are given by;

$$\mid z \mid \geq z_{\frac{\alpha}{2}}, \quad z \geq z_\alpha, \quad \text{and} \quad z \leq -z_\alpha, \quad \text{respectively.}$$

If the variance σ^2 is unknown and the sample size is lees than thirty $(n < 30)$, we can use t-test[9]

with the following test statistic

$$t = \frac{\bar{X} - \mu_o}{\sqrt{\frac{s^2}{n}}} \sim t_{(n-1)}.$$

Where; $(n-1)$ is the degree of freedom, s the standard deviation of sample and in this case the z-test cannot be used. Referring to Equation (2.1.1), the critical regions for the alternatives within t-test are given by,

$$| t | \geq t_{\frac{\alpha}{2},n-1}, \quad t \geq t_{\alpha,n-1}, \quad \text{and} \quad t \leq -t_{\alpha,n-1}, \text{ respectively.}$$

2.1.8 Examples for test statistic .

Example 2.1.8.1. It is claimed that the height of a certain species of plant is normally distributed with mean $110cm$. In order to test this claim, a sample of 21 plants of this species is observed and found that $\bar{X} = 105$ cm, and $s = 15cm$. Is there evidence at 5% to reject this declaration?

Solution

Step 1: We test $H_o : \mu = 110cm$ against $H_a : \mu \neq 110cm$.

Step 2: Since the population variance is unknown and sample size $(n < 30)$ the test statistic is ;

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t(20).$$

Step 3: Critical region: Two tailed test. We reject H_o if $| t | \geq t_{0.05,20} = 2.086$

Step 4: From the sample we have

$$t = \frac{150 - 110}{\frac{15}{\sqrt{21}}} = -1.528.$$

Step 5: Conclusion: since $| t | < 2.086$, we do not have enough evidence to reject our claim that the mean height is 110cm.

Example 2.1.8.2. The hourly wages in a particular industry are normally distributed with mean of $\$1.320$ and σ of $\$250$. A company in this industry employs 40 workers paying them an average of $\$1.220$ per hour. Can this company be accused of paying substandard wages? the significance level is 0.1.

Solution

The hypotheses are

$$H_o : \mu = 1.320 \quad \text{against} \quad H_a : \mu < 1.320.$$

Since the standard deviation is known and n>30, we use z-test the

$$z = \frac{1.220 - 1.320}{\frac{250}{\sqrt{40}}} = -2.53.$$

We reject H_o if $z < z_{0.1}$ = -1.28. Therefore we reject the null hypothesis and conclude that the company is underpaying its employees.

2.2 Multiple test

Generally, type I error occurs when a null hypothesis is true. The group of hypothesis from the same sample in a given population can be defined as a family and the probability of observing a type I error for a single test is given by alpha (α). Multiple test involves running different independent statistical test within a sample in a given population by considering different hypothesis to enable researchers to discover the strangeness and conclude that they have significance to the whole population.

From the group of hypothesis (family), we can compute the combined error measure which is family wise error rate (FWER).

Definition 2.2.0.1. Family-wise Error rate

The family-wise error rate can be defined as the probability of committing at least one type I error in two or more statistical studies with the same sample data[3, 5]. For the family of m independent null hypothesis with significance level(α), the probability of observing at least one false rejection is expressed by

$$FWER = 1 - (1 - \alpha)^m \tag{2.2.1}$$

with;

$$\begin{aligned} \alpha &: \text{Probability of making error.} \\ 1 - \alpha &: \text{The probability of not making error.} \\ (1 - \alpha)^m &: \text{Probability of not making error in } m \text{ hypothesis test.} \\ FWER &: \text{The probability of making at least one error in } m \text{ hypothesis test.} \end{aligned}$$

Example 2.2.0.2. Assume we have 10 hypothesis to test with significance level of 0.05. We can now find the probability of observing at least one significance result just due to chance.

$$\begin{aligned} FWER &= 1 - (1 - 0.05)^{10} \\ &= 0.40 \end{aligned}$$

So, by considering 10 tests, we have a 40% chance of observing at least one type I error.

Example 2.2.0.3. A test with four treatment (A,B,C,D) and five endpoints can be analysed with 30 independent null hypothesis i.e with six pairwise test(A-B, A-C, A-D, B-C, B-D, C-D) for each one of the five endpoints. When a family of 30 independent null hypothesis is test at $\alpha = 0.05$, what is the probability of observing at least on type I error (FWER)?

Answer;

$$\begin{aligned} FWER &= 1 - (1 - 0.05)^{30} \\ &= 0.78. \end{aligned} \tag{2.2.2}$$

The above calculated probability means that the risk of observing at least one false rejection in a family of 30 independent hypothesis is 78%.

Example 2.2.0.4. Consider the case of 40 hypothesis to test with significance level of 0.05. We can now find the probability of observing at least one significance result just due to chance.

$$\begin{aligned} FWER &= 1-(1-0.05)^{40} \\ &= 0.87. \end{aligned}$$

By considering 40 tests, we have a 87% chance of observing at least one type I error.

It is clear that within the same significance level, as the number of hypothesis increases in a sample, the probability of observing at least one type I error increases, this is a big issue with multiple test! [10], Why? because we want to reduce the probability of getting at least one type I error due to the chance, so that it remains below our desired significance level (α).

To control the family wise error rate means that make it less or equal to the pre-specified alpha in m independent test

$$\lim_{m\longrightarrow\infty} \sup[1-(1-\alpha)^m] \leq \alpha.$$

Since the idea is to minimize the probability of rejecting the null hypothesis H_o, we look for methods that can help us to achieve the reasonable level of family wise error rate. One of them is Bonferroni correction.

Bonferroni correction involved by dividing significance level by number of statistical analyses performed. Instead of using the significance level (α), it uses $\frac{\alpha}{m}$. Where m is the number of independent test. The idea is due to the first order Taylor expansion as alpha approaches zero;

$$\begin{aligned} 1-(1-\alpha)^m &= 1-(1-m\alpha) \\ &= m\alpha. \end{aligned}$$

So, the Bonferroni method will be

$$FWER_B = 1-\left(1-\frac{\alpha}{m}\right)^m. \tag{2.2.3}$$

If we use the Bonferroni method in the above Example (2.2.2) the probability of observing at least one type I error will be

$$FWER_B = 1-\left(1-\frac{0.05}{30}\right)^{30} = 0.048, \tag{2.2.4}$$

this means that the probability of observing at least one type I error in a family is 4.8% which is smaller that significance level ($\alpha = 5\%$), therefore no rejection of null hypothesis.

Now the possible way to reject the null hypothesis is when the p-value is less than $\frac{\alpha}{m}$. Instead of adjusting downward the significance level Benferroni can also adjust for P-value of each hypothesis so that they will become bigger than significance level to not reject the null hypothesis, like if we have $P_1 < P_2 < \cdots < P_m$ we multiply each p-value by m Ludbrook [3]. But not all p-values we multiply exactly with m , we follow; $p_1' = mp_1$, $p_2' = (m-1)p_2, \ldots, p_m' = p_m$. In this case any multiplied P-value less than alpha is determined to be statistically significance.

3. Testing endpoints with unknown correlation

In this chapter we will address the objective of test performance in clinical trial with the aim of controlling the family wise error rate resulted from endpoints testing. A randomized controlled trial was conducted with 200 persons from overweight and obese populations with serious mental illness. The purpose is to show the effectiveness of treatments and much about testing the difference in each endpoint from two different groups A and B. Each group contains two treatments. We stop the trial if we get an adverse effects (negative results) and if one of treatments shows statistically significance benefits the trial is continued.

Treatment 1 presents fitness improvement and treatment 2 presents activity to facilitate weight loss. We do a group sequential design where each group is composed of 100 participants and each participant will have two treatments, see clinical example [11]. Endpoint is representing the outcomes of treatments in each group, the endpoints we are testing are primary endpoints because primary endpoint is endpoint which shows the effectiveness between treatments. For further details about endpoint see [12, 13].

Hypothesis:

In both groups, the null hypothesis is that the difference in means for both treatments is greater or equal to zero.

$$\text{Endpoint1;} \qquad H_o : \mu \geq 0, \quad H_a : \mu < 0$$

$$\text{Endpoint2;} \qquad H_o : \mu' \geq 0, \quad H_a : \mu' < 0$$

The global null hypothesis being that one or both null hypotheses are true and alternative hypothesis is that none of the null hypothesis is true.Then we repeat the process 10^6 times to have different results. More about clinical trial hypothesis setting see [14].

The observation

$$Y_i^j = (Y_{i1}^j, Y_{i2}^j), \;\; j \in \{A, B\}, i = 1, 2, ..., 100, \tag{3.0.1}$$

from each of two treatments in every group, are independently normal distributed with mean (μ_1^j, μ_2^j) and corr$(Y_{i1}^j, Y_{i2}^j) = \rho \geq 0$ with covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

All parameter are unknown except the $\sigma 1$ and $\sigma 2$. Further more check Tamhane et al. [15].

3.1 Problem with unknown correlation

Multiple endpoints analysis in clinical trials use tests based on joint multivariate normal distribution that assumes correlations to be known or estimated based on interim analysis (sample data).

Since the erroneous specification of the true correlation ρ can have a large impact on FWER, it is advisable to use the most conservative adjustment across all possible correlations. However with an observed data, we cannot use the correlations which are not supported by the data.

3.1.1 Control the family wise error rate at significance level ($\alpha = 0.05$).

Before multiplicity correction with significance level ($\alpha = 0.05$), from endpoints data the maximum family wise error rate was 0.1, see Figure 3.1 which shows that the maximum FWER is larger than the significance level. The exact correlation ρ between endpoints is usually unknown, but often assumed to be known or estimated based on interim analysis (sample data) [1, 15].

Family wise error rate control requirement is given by

$$\text{FWER} = P\{\text{Reject at least one true null hypothesis}\} \leq \alpha. \tag{3.1.1}$$

From observation in Equation (3.0.1), let denote (c_1, c_2) as the critical boundaries (values) of (Y_i^A, Y_i^B), $i = 1, \ldots, n_1$, for n_1 observation.

Let consider the distribution of the z statistics $Z = (Z_1, Z_2) \sim N((\theta_1, \theta_2), \Sigma)$.

Stage 1:

If $Z_1 < c_1$ we continue to stage two. If $Z_1 \geq c_1$, reject H_1 and test H_2.

Stage 2:

If $Z_2 < c_2$, accept H_2 and stop testing. If $Z_2 \geq c_2$, reject H_2.

The critical boundaries (c_1, c_2) must be determined to satisfy Equation (3.1.1).

If (c_1, c_2) is an α-level boundary under H_1, Tamhane et al. [15] showed that FWER is controlled at pre-specified alpha (α). That is,

$$P_{H_1}(Z_1 > c_1) + P_{H_1}(Z_1 \leq c_1, Z_2 > c_2) = \alpha, \tag{3.1.2}$$

therefore any couple (c_1, c_2) which satisfies the above equation is called alpha level boundary. Let us take some example of Pocock [16] boundary which uses $c_1 = c_2 = c$, where $c > 0$ and must be determined in each case to satisfy Equation (3.1.2).

Both numerical simulation and analytical function show that the FWER increases as critical boundary (value) increases and the overall maximum family wise error rate occurs at critical value $(1, 0.05)$, see Figure 3.2.

Following the result of Monte Carlo simulation in Section 3.3.1, we find an analytical function to approximate the family wise error rate

$$FWER(\rho) = -\left(\frac{\gamma}{k}\right) \underbrace{\left(\frac{\alpha}{m}\right)^{-\rho} \left(1 - \frac{\alpha}{m}\right)^{(1+\rho)}}_{f(\rho)} + \gamma \tag{3.1.3}$$

The function can be deduced into

$$FWER(\rho) = \gamma\left(1 - \frac{f(\rho)}{k}\right) \tag{3.1.4}$$

where $-f(\rho)$ is the main function but non probabilistic with two parameters α and m.

The α is the significance level 0.05 and m the number of statistical test performed $(2$ test$)$. In general these parameters play the role in generating the family wise error rate particularly m in case of Bonferroni correction as shown in Equation (2.2.3).

Other parameters are:

ρ is the correlation which is varying between $[-1, 1]$.

γ and k play a great role of changing the main function $-f(\rho)$ into a probabilistic function noted by $FWER$ in Equation (3.1.3). Specifically γ and k help to determine the boundaries of FWER. The parameters γ and k work in the following manner;

Before we multiply γ in Equation (3.1.4) we first formed a probabilistic function using parameter k which is $(1 - \frac{f(\rho)}{k})$ in Equation (3.1.4) above. This means that parameter k help us to set the probability function related to the family wise error rate in general. In our case $k = 76$ is fixed.

So, γ is a parameter which determines the image of $\rho = -1$, it means for each curve the value of γ determines the true maximum FWER for its corresponding function. The value of gamma cannot exceed 0.1. see Figure 3.2.

Generally parameters γ and k help us to set a range in which the FWER will lie.

From Equation (3.1.3), we can find the value of γ at each critical value $(r, 0.05)$ by just substituting the value of r and α in the following function.

$$FWER(r) = -\left(\frac{\gamma}{76}\right)(0.025)^{-r}(1 - 0.025)^{(1+r)} + \gamma = \alpha. \tag{3.1.5}$$

We can now look at different examples in which γ is computed within critical boundary which is $c = (r, 0.05)$. This involves assuming any correlation r to find the value of γ which is the maximum FWER at that assumed correlation. Within the formula in Equation (3.1.5) we can consider $r = 1$ then $\gamma = 0.1$ and $r = -1$ then $\gamma = 0.05$ in the examples below.

Example 3.1.1.1. Assume the critical point $(1, 0.05)$

$$\begin{aligned} -\left(\frac{\gamma}{76}\right)\left(\frac{1}{0.025}\right)(1 - 0.025)^2 + \gamma &= 0.05 \\ \gamma &= 0.1. \end{aligned}$$

Example 3.1.1.2. Assume the critical point $(-1, 0.05)$

$$\begin{aligned} -\left(\frac{\gamma}{76}\right)(0.025) + \gamma &= 0.05 \\ \gamma &= 0.05. \end{aligned}$$

Example 3.1.1.3. Consider $(0.7, 0.05)$ then we replace that value in Equation (3.1.5) to have

$$\begin{aligned} -\left(\frac{\gamma}{76}\right)(0.025^{-0.7})(1-0.025)^{1.7}+\gamma &= 0.05 \\ \gamma &= 0.0602. \end{aligned}$$

Example 3.1.1.4. Consider $(0.5, 0.05)$ then we replace that value in Equation (3.1.5) to have

$$\begin{aligned} -\left(\frac{\gamma}{76}\right)(0.025^{-0.5})(1-0.025)^{1.5}+\gamma &= 0.05 \\ \gamma &= 0.05481. \end{aligned}$$

Example 3.1.1.5. Consider $(0.9, 0.05)$ then we replace that value in Equation (3.1.5) to have

$$\begin{aligned} -\left(\frac{\gamma}{76}\right)(0.025^{-0.5})(1-0.025)^{1.5}+\gamma &= 0.05 \\ \gamma &= 0.07658. \end{aligned}$$

Below is the graph of family wise error rate against correlation ρ showing the maximum family wise error rate with two performed test at significance level ($\alpha = 0.05$). Analytically the below figure can be found at $c = (1, 0.05)$ and $\rho \in [-1, 1]$.

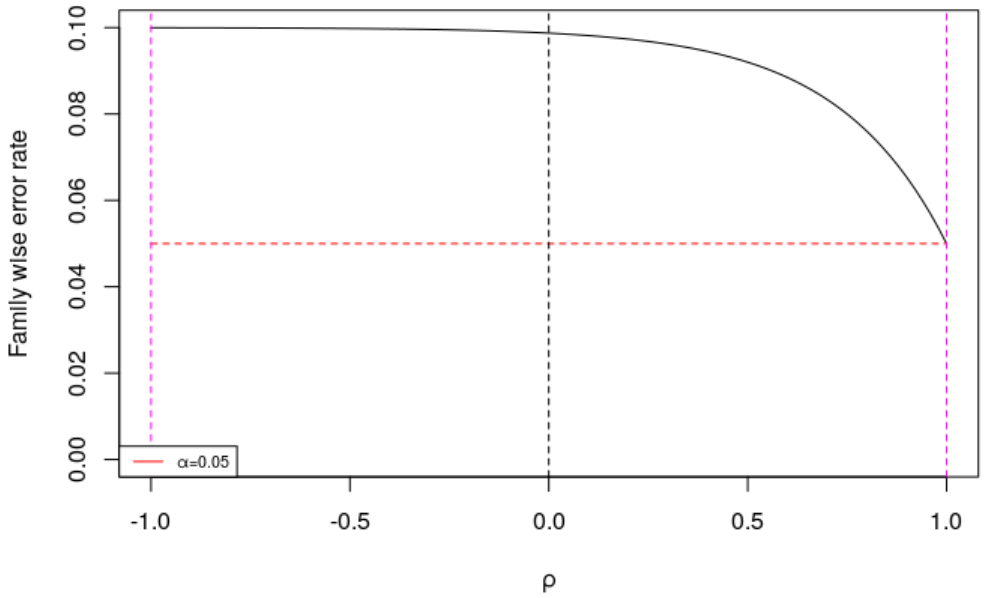

Figure 3.1: Family wise error rate before any adjustment.

Bonferroni correction

Drawing inferences in multiple test is performed by adjusting decision rules for each hypothesis. This can be achieved by adjusting significance level (α) downward or adjusting P-value upward.

Bonferroni correction is now adjusting significance level downward which means dividing the significance level ($\alpha = 0.05$) by the number of performed tests (2 tests), the FWER changes within interval with the maximum value of 0.05. Analytically the result of Bonferroni correction of the above Figure 3.1 can be found by setting critical boundary $c = (-1, 0.05)$ and $\rho \in [-1, 1]$ which represented by the violet curve in Figure 3.2.

3.2 Multiplicity techniques when endpoints are correlated

In this section we discuss the different techniques of controlling the family wise error rate with the unknown correlation between endpoints.

3.2.1 Adjust for fixed assumed correlation.

The method is also called worst case approach. The aim is to adjust the family wise boundary with assumed correlation r between endpoints. This method assumes correlation between endpoints which might lead to the worse results. This is equivalent to not improving the boundaries at all. In this method we use $c_1 = c_2 = (r, 0.05)$ boundary where r is the assumed correlation for the unknown correlation (ρ) then;

FWER $> \alpha$, when the (c_1, c_2) boundary is underestimated [1]. FWER $< \alpha$, when the (c_1, c_2) boundary is overestimated.

Following the example computed above especially Example 3.1.1.3 and Example 3.1.1.4, one can construct a graph to show how any assumed correlation r that satisfies the condition $(r, 0.05)$ can generate the maximum value of the family wise error rate within $\rho \in [-1, 1]$ as shown in Figure 3.2 and Figure 3.3.

Assuming $r = 1$, we get the black curve in Figure 3.2 with the maximum family error rate of 0.1.

Assuming $r = 0.7$, we get the blue curve in Figure 3.2 with the maximum family wise error rate of 0.0602 .

Assuming $r = -1$, we get the violet curve in Figure 3.2 with the maximum family wise error rate of 0.05.

The points where the function in Equation (3.1.3) hits the line $\alpha = 0.05$ are the critical values or critical boundaries.

By plotting the FWER against correlation ρ, we observed that when the assumed correlation r increases the FWER also increases within interval, see Figure 3.2 with different critical points $(1, 0.05)$, $(0.9, 0.05)$, $(0.8, 0.05)$, $(0.7, 0.05)$ and $(0.5, 0.05)$. Bonferroni method shows how the reduction of significance level can reduce the probability of observing at least one false rejection

in m performed test and is presented by the violet curve with critical boundary $(-1, 0.05)$ in Figure 3.2.

Figure(3.2) shows different maximum family wise error rate based on the assumed correlations.

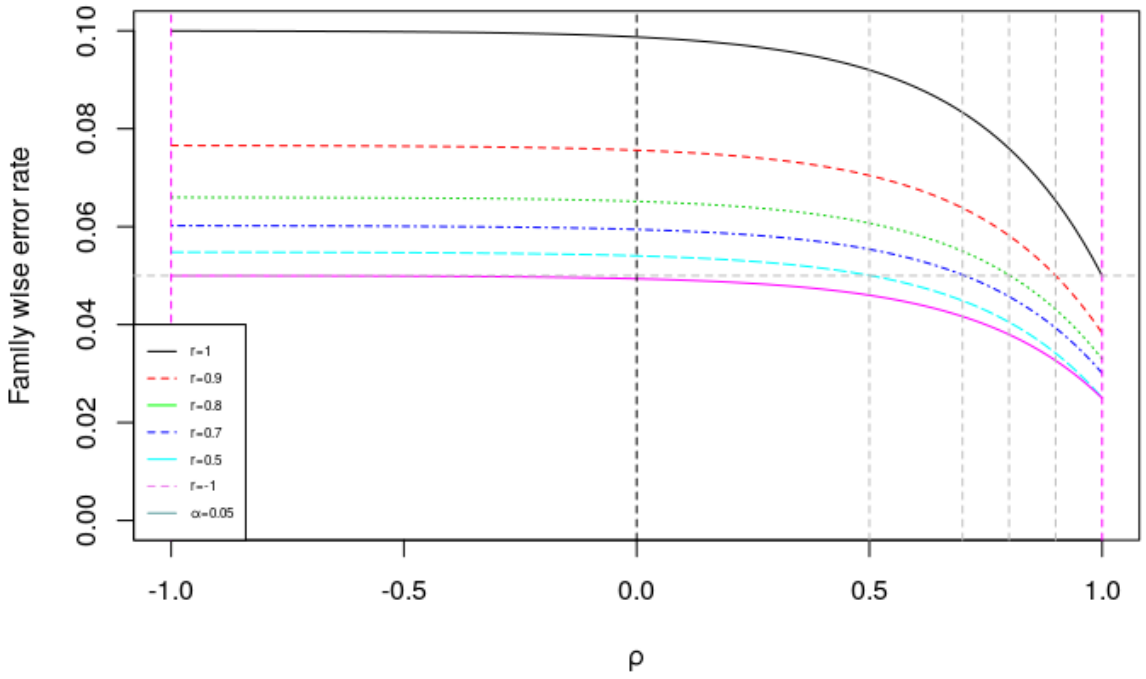

Figure 3.2: Family wise error rate with assumed correlation.

The family wise error rate in the above Figure3.2 is described as follow:

For $\rho > r$, the family wise error rate increases below 0.05 and its minimum is at $\rho = 1$.

For $\rho < r$, the family wise error rate decreases above 0.05 and its maximum is at $\rho = -1$.

One can also assume different correlations to further investigate the effect of assumed correlation on the FWER. This can be done to check the behavior of the FWER in the case of assuming small value and large value of correlation.

Figure 3.2 shows that higher assumed correlation between endpoints leads to larger maximum level of FWER.

From Monte Carlo simulation of multivariate normal distributed data in Section 3.3.1, which presents the data of two endpoints, Figure 3.3 is the plot of the FWER with different assumed correlations.

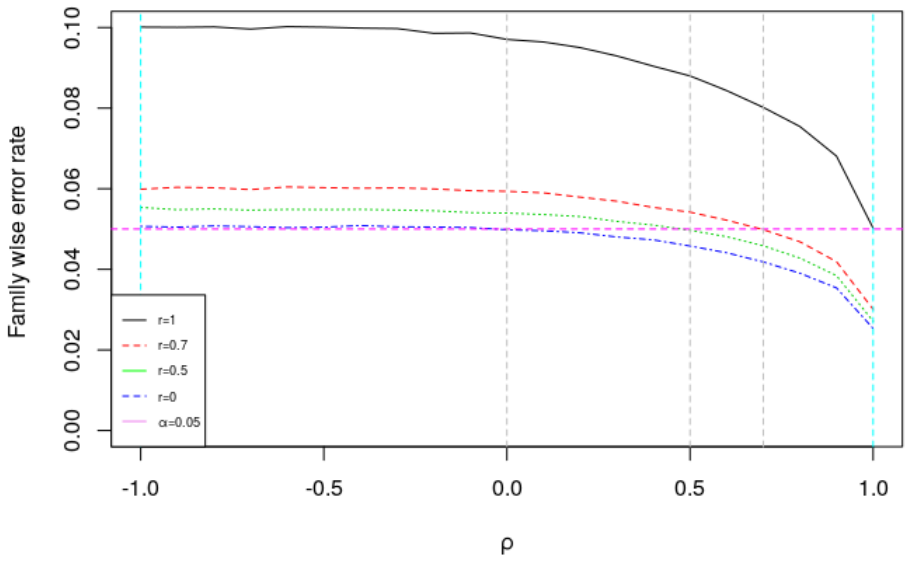

Figure 3.3: Family wise error rate with assumed correlation.

3.2.2 Plug-in estimated correlation.

The aim is to adjust the family wise boundary with estimated correlation $\hat{r}$ between endpoints using sample data. We estimate the correlation $\hat{r}$ between endpoints using sample estimate for correlations between the outcomes of two treatments in each group. Using Monte Carlo simulation we simulate different correlation based on each estimator formula in the Table 3.1. Since we will use the consistent estimator of ρ, we prefer to use different sample size to see if the control of FWER boundaries does not depend on the sample size. In this method we will try different cases with sample size of $n = 10$ and another with sample size $n = 50$.

Calculation of sample estimate for correlation. The method is based on averaging across Pearson correlation estimates from two groups $\hat{r}_A$, $\hat{r}_B$. The estimated correlation at each group will be,

$$\hat{r}_j \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(Y_{i1}^{j}-\bar{Y}_{1}^{j})(Y_{i2}^{j}-\bar{Y}_{2}^{j}),$$

Where

$$\bar{Y}_{k}^{j} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_{ik}^{j}\,,\;\; k\in\{1,2\}.$$

Table 3.1: **Different estimators for endpoints correlation estimation.[1, 2]**

Correlation	Formula	Name
$\hat{r}_{\text{pool}}$	$\hat{r}_{\text{pool}} = \frac{\hat{r}_A + \hat{r}_B}{2}$	pooled
$\hat{r}_{\text{blind}}$	$\hat{r}_{\text{blind}} = \frac{1}{2(n-1)} \sum_{i=1}^{n} (Y_{i1}^j - \bar{Y}_1^j)(Y_{i2}^j - \bar{Y}_2^j)$	blinded
$\hat{r}_{\text{Fisher}}$	$z_j = \frac{1}{2} \ln\left(\frac{1+\hat{r}_j}{1-\hat{r}_j}\right) \rightarrow \bar{z} = \frac{z_A + z_B}{2} \rightarrow \hat{r}_{\text{Fisher}} = \frac{e^{2\bar{z}}-1}{e^{2\bar{z}}+1}$	Fisher
$\hat{r}_{\text{OP}}$	$\hat{r}_{\text{OP}} = \frac{1}{2} \sum_{j \in A,B} \left(\hat{r}_j + \frac{\hat{r}_j(1-\hat{r}_j^2)}{2(n-3)}\right)$, n is sample size	Olkin and Pratt

The above table summarize the formula used to estimate the correlation $\hat{r}$ between two endpoints. The method involved in determining the estimated correlation using different estimators in Table 3.1 then use the estimated value to find critical values $(\hat{r}, \alpha)$. Since the formula in the table above will not give the same estimate we would like to know which among them can estimate precisely the value of ρ to further control the family wise error rate boundaries.

The critical boundary is given by $c_1 = c_2 = (\hat{r}, 0.05)$. Where $\hat{r}$ is the estimated correlation for the unknown correlation.

From the sample with sample size, $n = 10$, we find Figure 4.1 with different estimated correlations from different estimators as summarised in Table 3.1.

From the sample with sample size $n = 50$, we find Figure 4.2 with different estimated correlations from different estimators as summarised in Table 3.1.

Plug-in estimated correlation method uses an interim estimate of correlation to control the FWER boundaries. The results of these method show that there is a family wise inflation which decreases as the sample sizes increases as shown in Figure 4.1 and Figure 4.2.

Distribution of Estimators.

According to the sample size or the size of the piloting data, we can estimate the correlation by doing multiple simulation say 10^5 simulations run for each sample size then we find the distribution for each estimator in Table3.1 at a given correlation ρ. The estimated correlations can help us to determine the expectation of the squared deviation from their mean (variance).

Consider different sample sizes; $n = 10$, $n = 50$ and $n = 1000$ where $\sigma_1 = \sigma_2 = 1$ with 10^5 simulation runs to find the estimated value of ρ with different estimators.

Figure 3.4 shows how the estimated correlation are distributed. With $n = 10$, the distribution of correlation estimated by pooled estimator is the same as the distribution estimated by Fisher which means that the red curve is overlapped by the green curve.

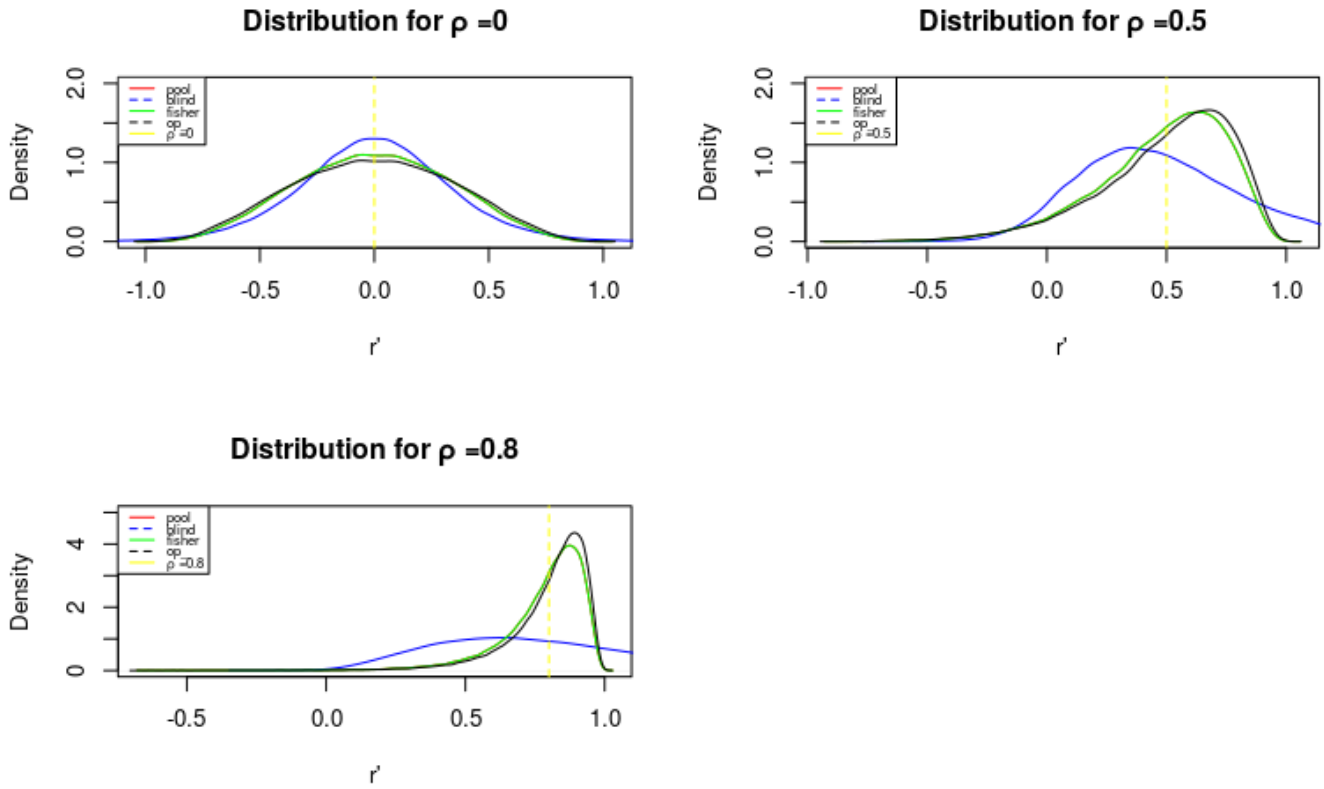

Figure 3.4: Distribution of estimators for $n = 10$.

Generally the distribution of Pearson correlation is not Normal because Pearson's correlation follows student t-distribution with degrees of freedom $n - 2$ especially in the null case ($\rho = 0$) [17]. Therefore the distribution of estimated correlation $\hat{r}$ is a student t-distribution with degrees of freedom $n - 2$, when the true correlation is assumed to be zero.

In case of $\rho > 0$, the distribution is more complex [18]. Therefore the case of $\rho = 0.5$ and $\rho = 0.8$ have complex distributions.

The red line represents the pooled estimate in the Figure 3.4 overlapped by green line which is representing fisher estimate.

Figure 3.5 shows how the estimated correlation are distributed. With $n = 50$, the distribution of correlation estimated by pooled estimator is overlapped by the distribution estimated by Fisher which means that their curves are confounded.

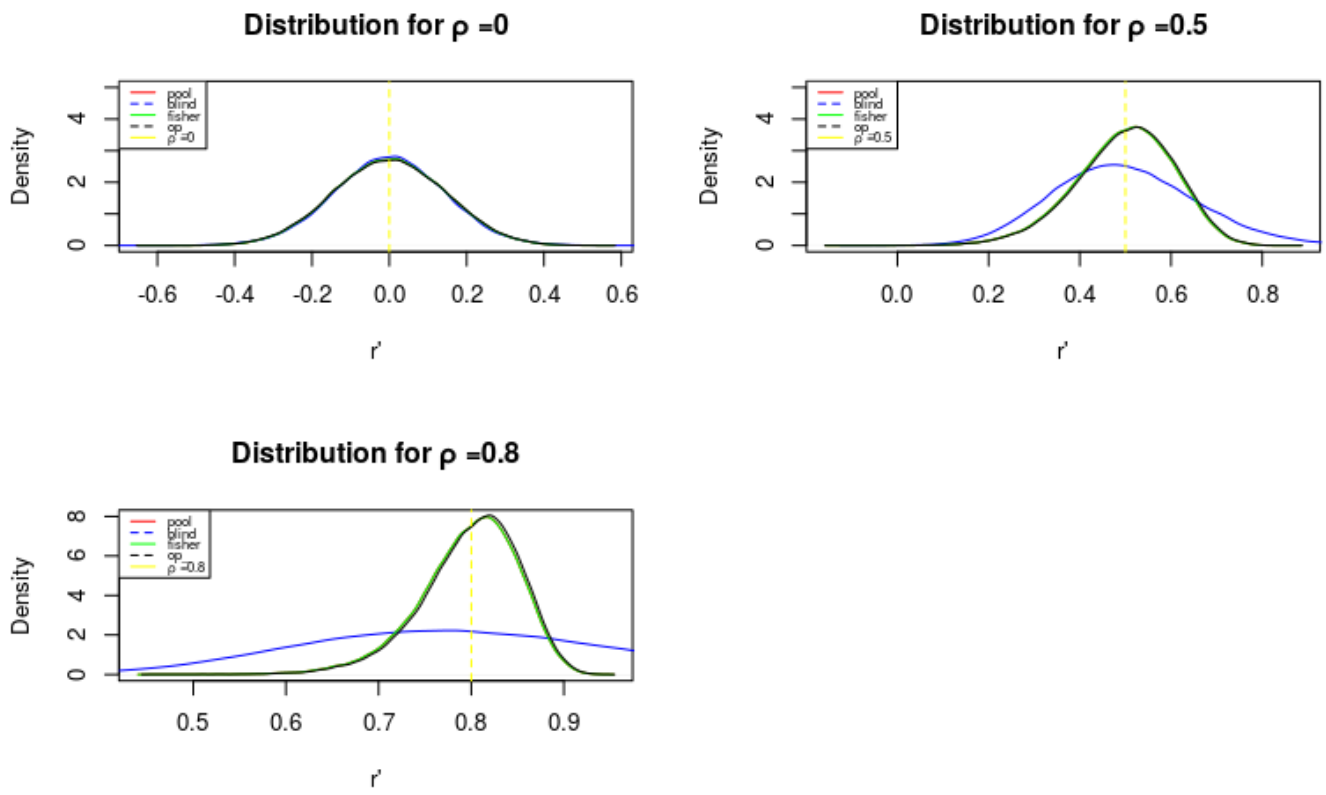

Figure 3.5: Distribution of estimators for $n = 50$.

The same as in Figure 3.4 the distribution at $\rho = 0$ is a student's t-distribution not normal distribution.

The estimated correlation $\hat{r}$ is likely to follow a complex distribution when the true correlation is assumed to be 0.5. The same as $\rho = 0.8$.

The red line represents the pooled estimate in the Figure 3.5 overlapped by green line which is representing fisher estimate.

Below figures show how the estimated correlation are distributed. With $n = 1000$, the distribution of correlation estimated by pooled estimator overlapped by the distribution estimated by fisher which means that the red curve is overlapped by green curve.

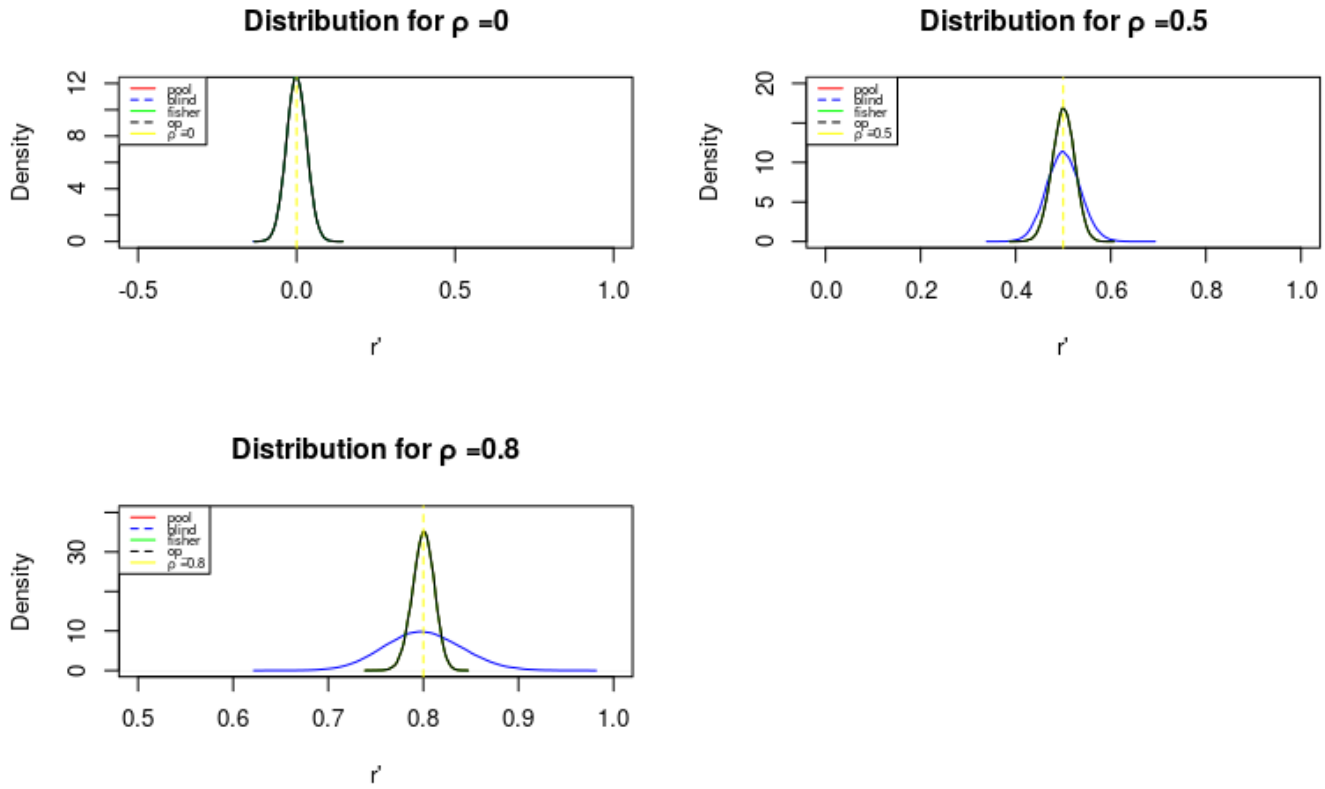

Figure 3.6: Distribution of estimators for $n = 1000$.

The distribution of estimated correlation $\hat{r}$ is likely to be normal but still follow student t-distribution when the true correlation is assumed to be zero.

The estimated correlation $\hat{r}$ is likely to be normal distributed but still a complex distribution when the true correlation is assumed to be 0.5. The same as $\rho = 0.8$.

The red line represents the pooled estimate in the Figure 3.6 overlapped by green line which is representing Fisher estimate.

In fact as the sample size increases the correlation estimators tend to be normally distributed. In other words, the distribution approaches a normal distribution when n goes to infinity. The distributions become narrower with more subjects. The R code generates the distribution of estimators based on our endpoints data can be found by following the links in appendix 5.6.

3.2.3 Berger and Boos method.

The method involves, determining the lower bound of nuisance parameter (s^2) from estimated correlation and uses it to calculate the critical boundaries.

Assume $c = (\rho, \alpha)$ to be the real boundary based on true correlation ρ which is unknown. We can set the confidence width based on the nuisance parameter s^2 (variance of estimated correlations). Let $X = 1 - \epsilon$, be a confidence level with confidence width of $(0$ toward $1 - \epsilon)$ for the nuisance

parameter when H_o is true. The value of ϵ and the confidence level X should be specified before looking at the data and we restrict that P_ϵ never be smaller than ϵ. Therefore practically, ϵ is chosen to be small, such as 0.001 or 0.0001 as suggested by Berger and Boos [19]. Taking the value of $\epsilon = 0.001$, we find the confidence level to be

$$\begin{aligned} X = 1 - 0.0001 &= 0.9999 \\ &= 99.99\% \end{aligned}$$

Upper Critical and Lower Critical Limits Approach

Computing the upper bound and the lower bound of our nuisance parameter (variance of estimated correlation) using the confidence level $1 - \epsilon$ we can use the upper critical limit (UCL) and lower critical limit (LCL) approach. The anticipated distribution of the pilot variation of estimated correlation is a chi-squared distribution [20].

In order to implement the UCL and LCL approaches, a variance s^2 (variance of estimated correlation) from the pilot data is obtained and the one-sided $100X\%$ UCL,s^2_{UCL} and $100X\%$ LCL,s^2_{LCL} are calculated where $X = 1 - \epsilon$. A one-sided $100X\%$ UCL and LCL for the variance can be calculated from

$$\begin{aligned} s^2_{UCL} &= \frac{(2n-2)s^2}{\chi^2_{(1-X),2n-2}} \\ s^2_{LCL} &= \frac{(2n-2)s^2}{\chi^2_{X,2n-2}} \end{aligned} \tag{3.2.1}$$

where n is the sample size, $2(n-1)$ is the degree of freedom from two groups with equal participants for the variance estimate, and $\chi^2_{X,k}$ denotes the $(X = 1 - \epsilon)$ percentile of the chi-squared distribution.

Since we are interested in determining the lower bound we can choose different values of $100X\%$, by taking 70%, 80%, 90%, and 99%, to compute the lower critical limit for each estimator. Whitehead et al. [20] suggested the confidence level of 80% or 95%.

With $n = 10$, the variance of estimated correlation and their corresponding lower critical limit with consistent estimator of ρ are as follow,

Table 3.2: **Lower bound of different estimators for endpoints correlation** $n = 10$

Estimator	Variance(s^2)	LCL $X = 0.7$	LCL $X = 0.9$	LCL $X = 0.99$	LCL $X = 0.8$
pooled	0.02132770	0.01840624	0.01417517	0.010246230	0.01645042
blind	0.18124420	0.15641740	0.12046150	0.087073130	0.13979680
fisher	0.02132770	0.01840624	0.01417517	0.010246230	0.01645042
OP	0.02015499	0.01739417	0.01339574	0.009682837	0.01554589

With $n = 50$, the variance of estimated correlation and their corresponding lower critical limit with consistent estimator ρ are as follow,

Table 3.3: **Lower bound of different estimators for endpoints correlation $n = 50$.**

Estimator	Variance(s^2)	LCL $X = 0.7$	LCL $X = 0.9$	LCL $X = 0.99$	LCL $X = 0.8$
pooled	0.002851771	0.002662586	0.002394644	0.002081209	0.002545672
blind	0.033608970	0.031379360	0.028221600	0.024527670	0.030001500
fisher	0.002851771	0.002662586	0.002394644	0.002081209	0.002545672
OP	0.002801886	0.002616010	0.002352756	0.002044803	0.002501141

We can use the calculated LCL at confidence level $X = 0.99$ and $X = 0.8$ to compute the maximum family wise error rate by substituting the corresponding LCL of each estimator in Equation (3.1.5).

The use of lower critical limit of estimated correlation with confidence level can control the boundaries of family wise error rate in two cases:

Case 1: Same confidence level $100X\%$ but different sample size. See Figure 4.5 and Figure 4.6.

Case 2: Same sample size but different confidence level. See Figure 4.4 and Figure 4.6

3.3 Simulations of family wise error rate control

Objective of the simulation was to control the family wise error rate boundaries using correlation between endpoints in different cases:

Case 1: Incorrect family wise error rate against correlation ρ.

Case 2: Bonferroni correction.

Case 3: Adjusting for fixed assumed correlation.

Case 4: Plug-in estimated correlation.

Case 5: Improved Berger and Boos.

3.3.1 Monte Carlo simulation for bivariate normal distribution with incorrect FWER.

Using MASS package with $mvrnorm$ function in R, we simulate bivariate normally distributed data with mean (0,0) and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

The data presents the outcomes of treatments in each group, A and B. Each group contains 100 participants. We performed tests in both endpoints (outcomes of treatments in each group). We checked if the P-values are less than alpha (0.05) for both endpoints with the global null hypothesis to follow that both or one hypothesis is true. In other words, both or one of P-values are greater than α. 10^6 simulations run at significance level of 0.05. So, we plot the FWER against correlation ρ which varies between -1 and 1 to find Figure 3.1, which shows that the

maximum FWER is 0.10 which needs to be adjusted downwards as done in the following correction methods.

3.3.2 Bonferroni correction simulation.

After finding that the maximum FWER is larger than significance level, we performed a Monte Carlo simulation with 10^6 simulations run at the significance level of 0.025, in other words we applied Bonferroni correction by dividing the significance level by the number of test performed test ($\frac{\alpha}{2} = 0.025$). The analytical Bonferroni results is represented by the violet coloured curve in Figure 3.2. The simulated Bonferroni graph from endpoints data is shown in the figure below:

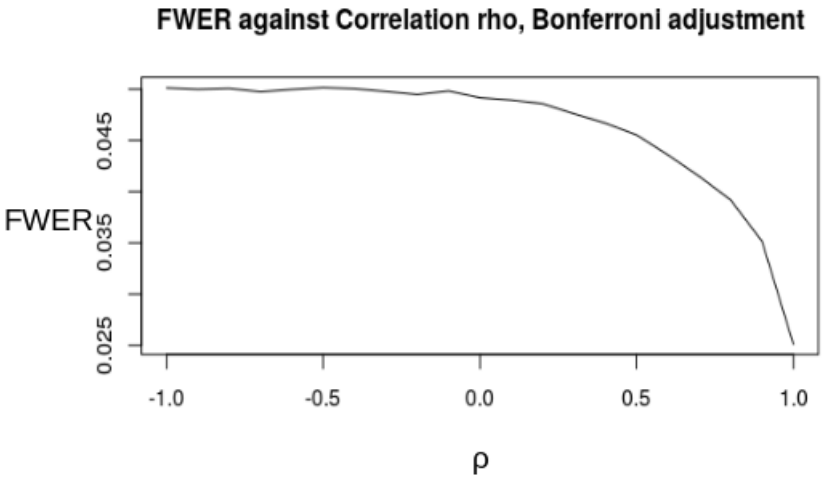

Figure 3.7: Corrected Family wise error rate by Bonferroni approach.

Since we needed to control the FWER boundaries around the significance level $(\alpha = 0.05)$ using the correlation between endpoints, we simulated for the following methods with different correlation between endpoints. The R code, we simulated for Bonferroni correction based on our endpoints data can be found by following the link in appendix 5.1.

3.3.3 Adjusting for fixed correlation simulation.

The outcome of Bonferroni correction shows that after correction the maximum family wise error rate is at 0.05 with the critical value $(-1, 0.05)$, where $r = -1$ is the assumed correlation. Making different assumption of correlation in our experiment like $r = 1$,$r = 0.9$, $r = 0.8$, $r = 0.7$ and $r = 0.5$ as shown in Figure 3.2 with critical values $(1, 0.1)$,$(0.9, 0.05)$, $(0.8, 0.05)$, $(0.7, 0.05)$ and $(0.5, 0.05)$ receptively. We get different maximum FWER base on each assumed correlation. The same is shown numerically in Figure 3.3.

It is clear that the variation of assumed correlation between endpoints affects the FWER. The method of fixed assumed correlation leads to the determination of different maximum FWER. In this experiment higher assumed correlation between endpoints leads to larger FWER as shown in Figure 3.2.

The R code for analytical function and numerical simulation for fixed assumed correlation method are found by following the links in appendix 5.2 and appendix 5.3 respectively.

3.3.4 Plug-in estimated correlation.

In this simulation instead of using the assumed correlation to control the FWER boundaries around the significance level $(\alpha = 0.05)$, we use the estimated correlation between the outcomes of treatments in each group with different estimators in Table 3.1 .

We performed Monte Carlo simulation for each plug-in method in Table 3.1 with different samples ($n = 10$ and $n = 50$) from the outcomes of treatments in each group A and B. Then estimate the correlation between the outcomes in each group to get $\hat{r}_A$ and $\hat{r}_B$. Using the values of $\hat{r}_A$ and $\hat{r}_B$, we estimate the value of correlation between endpoints for each plug-in method.

From the sample whose sample size is $n = 10$, we get the following correlations values; $\hat{r}_{\text{pool}} = 0.1508447$, $\hat{r}_{\text{blind}} = 0.1594903$, $\hat{r}_{\text{Fisher}} = 0.1555490$ and $\hat{r}_{\text{OP}} = 0.1604195$. Since the critical value is presented by $(\hat{r}, 0.05)$, we replace each critical value in Equation (3.1.5) to get the corresponding maximum FWER. As Bonferroni does, after finding the maximum FWER corresponding to each critical value, we divide each maximum FWER by number of test 2 to get the boundaries, then perform a Monte Carlo simulation for each value to get Figure 4.1

From the sample whose sample size is $n = 50$, we get the following correlations values; $\hat{r}_{\text{pool}} = 0.09125962$, $\hat{r}_{\text{blind}} = 0.08387398$, $\hat{r}_{\text{Fisher}} = 0.09126035$ and $\hat{r}_{\text{OP}} = 0.09222236$.

Since we already have the estimated correlation, we plug each critical value $(\hat{r}, 0.05)$ in Equation (3.1.5) to find the maximum FWER corresponding to each estimated correlation $\hat{r}$. To follow Bonferroni correction we divide each and every maximum family wise error rate corresponding to its correlation by 2 to know the exact boundaries of the FWER as Bonferroni does. The value we get we used them in numerical simulation to get Figure 4.2.

The R code we simulated to be used in this plug-in estimated correlation method of correction is found by following the links in appendix 5.5. The same for the simulation of estimated correlation from different estimators in appendix 5.4.

3.3.5 Improved Berger and Boos.

In this method instead of using the estimated correlation to control the FWER boundaries around significance level $(\alpha = 0.05)$, it uses the lower critical limit of the nuisance parameter s^2 from estimated correlation which can be determined using the formula in Equation (3.2.1). We perform Monte Carlo simulation for both $n = 10$ and $n = 50$, with 10^5 simulation runs to estimate correlation with different estimators then we compute the variance of estimated correlation to help as to get the lower critical limit (UCL) using the formula in Equation (3.2.1) with confidence level 80% and 99%. After getting the lower bound we substitute the critical value $(LCL, 0.05)$ in Equation (3.1.5) to get the maximum FWER for each LCL. The computed maximum FWER is divided by 2 to know the exact boundaries of the FWER, then use the value to numerical simulation to get Figure 4.3 and Figure 4.4 when the confidence level is $X = 0.8$, then Figure 4.5 and Figure 4.6 when confidence level is $X = 0.99$.

The R code we simulated to be used in this Berger and Boos methods of correction is found by following the links in appendix 5.7.

In the next chapter, we will discuss the result by interpreting the meaning of the graph based on the family wise error rate boundaries and the significance level $(\alpha = 0.05)$.

4. Analysing results from different methods for adjustment

In this chapter we discuss the outcome of the three methods discussed in Chapter 2. We will also compare the results to see which method is doing better in our study.

With the multivariate simulated data for the two endpoints , we performed two independent tests at significance level ($\alpha = 0.05$) and the maximum family wise error before multiple test correction was 0.1 which is larger than pre-specified significance level as Figure 3.1 shows. Since the maximum family wise error rate is larger than alpha, we did different correction based on correlation to get different results.

4.1 Results for fixed assumed correlation:

The method concerned with assuming the correlation leading to the highest possible type I error rate to adjust the boundaries. The results of these methods show that with the assumed correlation the FWER can take different maximum level between 0.05 and 0.1. Based on correlation assumption in Figure 3.1, the method shows a little change in FWER control because the correlation we are assuming is not related to our data. We found out that if the endpoints have the assumed correlation, the assumptions might lead to the worst control of FWER. If the Z_1 and Z_2 are greater than the boundary (c_1, c_2) the FWER will not be controlled under the significance level because the method can underestimate the boundaries.

This method can lead to the reduction of statistical power which is the probability that the test will reject a false null hypothesis because the endpoints are taking the correlation under assumption without referring to the data. As the power decreases, the chance of type II error increases.

4.2 Results for correlation estimator:

The method uses an interim estimate of the correlation to calculate the adjusted boundaries for the family wise error rate. Observing both Figure 4.1 and Figure 4.2, we find that as the sample size increases the family wise error rate inflation reduces. This means that as the sample increases the estimation get more precisely.

Figure 4.1 is the result for four different estimators with the sample size $n = 10$.

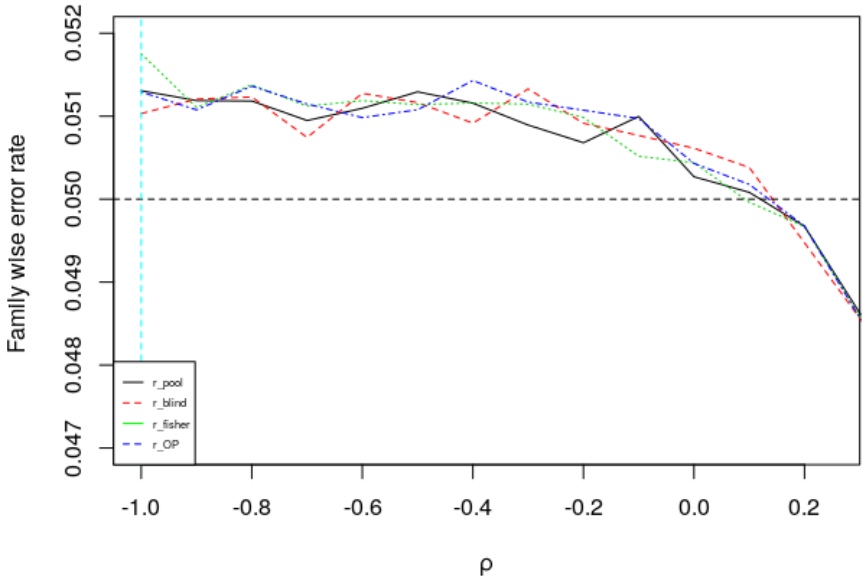

Figure 4.1: Family wise error rate with different estimated correlation at $n = 10$.

Below is the result for four different estimators with the sample size $n = 50$.

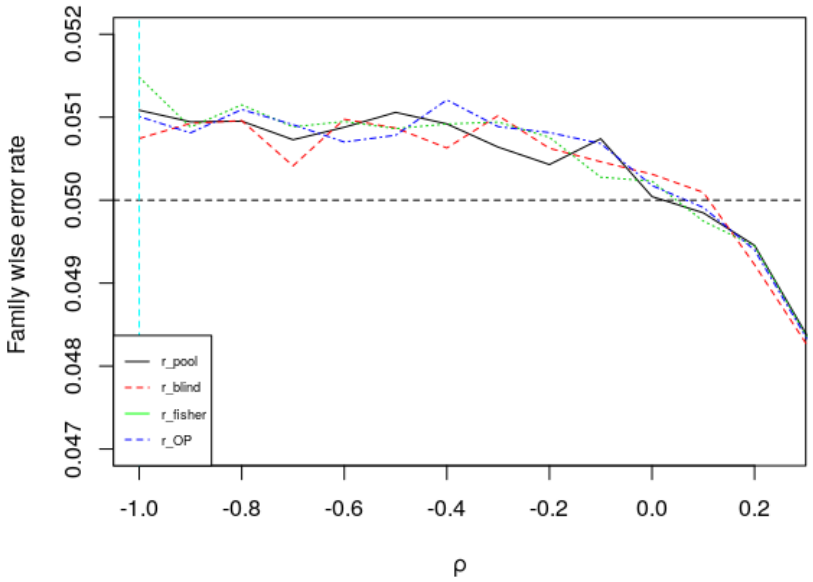

Figure 4.2: Family wise error rate with estimated correlation at $n = 50$.

In both Figure 4.1 and Figure 4.2 according to the results of different correlation estimators discussed in Table 3.1, we found that $\hat{r}_{\textbf{blind}}$ estimator is the best due to the FWER boundary near the significance level ($\alpha = 0.05$) with small inflation. The estimation of correlation leads to the better control of family wise error rate with small inflation because the determination of the critical boundary is based on the correlation from the data. This method shows that as the sample size increases we get more precise estimation of ρ and the large sample size reduce overestimation.

4.3 Result for improved Berger and Boos:

This method uses the lower bound of nuisance parameter from the estimated correlation to control the FWER boundaries. Figure 4.5 and Figure 4.6 show that to control the FWER depend on the sample size and the chosen confidence level. For fixed confidence level the large sample size leads to the better control of the FWER boundaries.

In this method, we will discuss the result according to the sample size and confidence level. This study used two samples with different sample size (n=10 & n=50) and two different confidence levels (80% & 99%) for each sample.

4.3.1 Results for 80% confidence level.

Below is the plot presenting the results of Berger and Boos for four different estimators within the sample size $n = 10$.

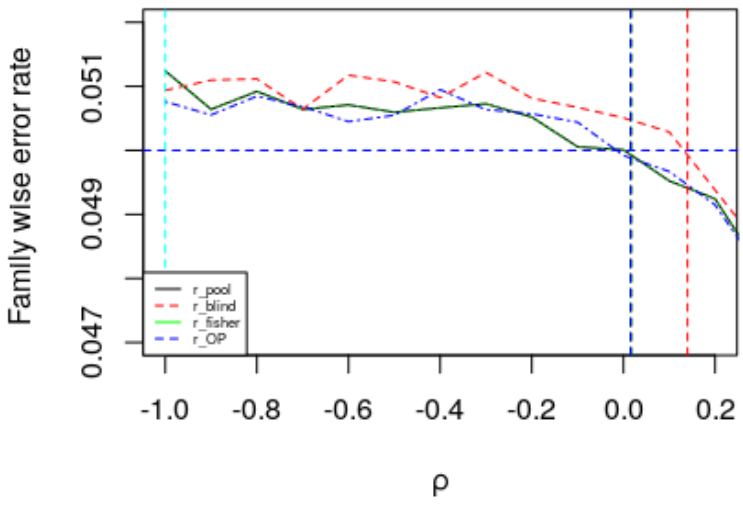

Figure 4.3: Family wise error rate with lower critical limit of estimated correlation.

Below is the plot presenting the results of Berger and Boos for four different estimators within the sample size $n = 50$.

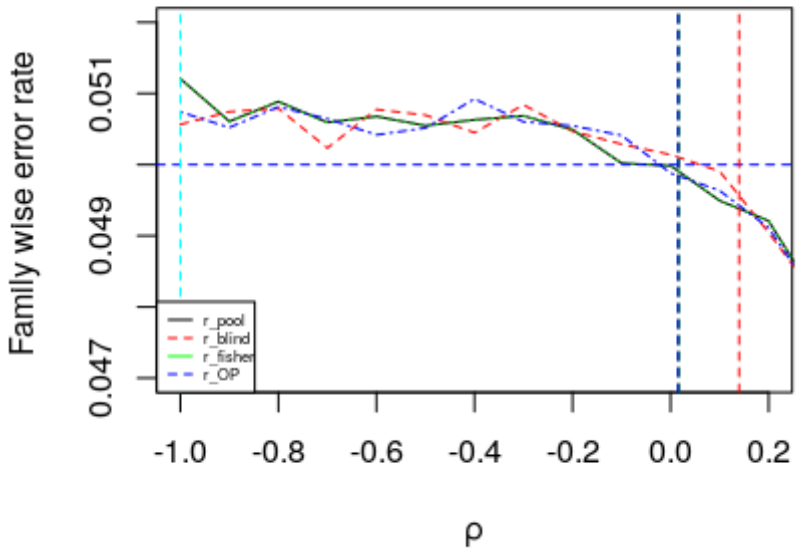

Figure 4.4: Family wise error rate with lower critical limit of estimated correlation.

4.3.2 Results for 99% confidence level.

Below is the plot presenting the results of Berger and Boos for four different estimators within the sample size $n = 10$.

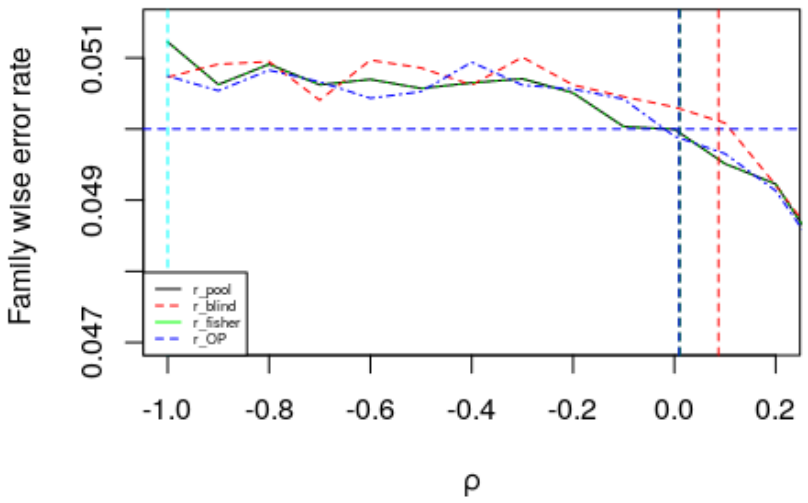

Figure 4.5: Family wise error rate with lower critical limit of estimated correlation.

Below is the plot presenting the results of Berger and Boos for four different estimators within the sample size $n = 50$.

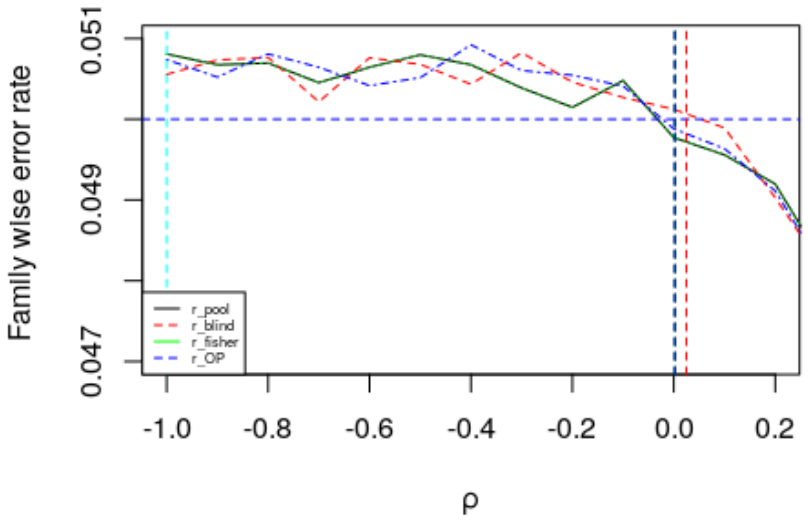

Figure 4.6: Family wise error rate with lower critical limit of estimated correlation.

The results of Berger and Boos method show that, within the same confidence level, as the

sample size increases the overestimation reduces which leads the boundaries of FWER to be more closer to the significance level. See Figure 4.5 and Figure 4.6 with the same confidence level 99%.

Berger and Boos results also show that, within the same sample size, as ϵ become small the estimation of ρ get more precise which leads the boundaries of FWER to be more closer to the significance level. See Figure 4.4 and Figure 4.6.

5. Conclusion

The structure of correlation between endpoints offers an easy way to describe different techniques of testing endpoints with unknown correlation. The correlation between endpoints must be taken into account in reducing the probability of observing at least one false rejection during multiple endpoints testing.

The essay summarises the issues of multiplicity in clinical trial and the way of correcting it when endpoints are correlated. The results show that controlling the family wise error rate boundaries by considering the correlation which is not related to the data leads to the worse control. Considering the estimated correlation from piloting data leads to the precise estimation of true correlation as the sample size increases. The results suggest to use the lower critical limit of the nuisance parameter from estimated correlation (Berger and Boos method) by considering the larger sample size to reduce overestimation and small value of ϵ to have an accurate and more precise control of the family wise error rate boundaries at the significance level in clinical trial.

In future study we can extend the work by increasing the number of endpoints within a sample. We can also discuss about multiple test with primary and secondary endpoints.

References

[1] Ajit C Tamhane, Yi Wu, and Cyrus R Mehta. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (i): unknown correlation between the endpoints. *Statistics in Medicine*, 31(19):2027–2040, 2012.

[2] Ralph A Alexander. A note on averaging correlations. *Bulletin of the Psychonomic Society*, 28(4):335–336, 1990.

[3] John Ludbrook. Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology*, 25(12):1032–1037, 1998.

[4] Dennis C Turk, Robert H Dworkin, Michael P McDermott, Nicholas Bellamy, Laurie B Burke, Julie M Chandler, Charles S Cleeland, Penney Cowan, Rozalina Dimitrova, John T Farrar, et al. Analyzing multiple endpoints in clinical trials of pain treatments: Immpact recommendations. *Pain*, 139(3):485–493, 2008.

[5] Josef Hochberg and A Coautor TAMHANE. Multiple comparison procedures. Technical report, John Wiley & Sons,, 1987.

[6] Eric W Weisstein. Bonferroni correction. 2004.

[7] Peter H Westfall and S Stanley Young. P value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, 84(407):780–786, 1989.

[8] Gopal K Kanji. *100 statistical tests*. Sage, 2006.

[9] Rosalind LP Phang. Basic concepts in hypothesis testing, 2018.

[10] Kristin L Sainani. The problem of multiple testing. *PM&R*, 1(12):1098–1103, 2009.

[11] Stephen J Bartels, Sarah I Pratt, Kelly A Aschbrenner, Laura K Barre, Kenneth Jue, Rosemarie S Wolfe, Haiyi Xie, Gregory McHugo, Meghan Santos, Gail E Williams, et al. Clinically significant improved fitness and weight loss among overweight persons with serious mental illness. *Psychiatric Services*, 64(8):729–736, 2013.

[12] Dean A Follmann. *Primary Efficacy Endpoint*. American Cancer Society, 2007. ISBN 9780471462422.

[13] Janet Wittes, Edward Lakatos, and Jeffrey Probstfield. Surrogate endpoints in clinical trials: cardiovascular diseases. *Statistics in medicine*, 8(4):415–425, 1989.

[14] Frank Konietschke. Multiple testing problems in pharmaceutical statistics. a. dmitrienko, ac tamhane, and f. bretz (2010). boca raton: Chapman & hall/crc biostatistics series. isbn: 978-1-584 88984-7. *Biometrical Journal*, 54(4):579–580, 2012.

[15] Ajit C Tamhane, Cyrus R Mehta, and Lingyun Liu. Testing a primary and a secondary endpoint in a group sequential design. *Biometrics*, 66(4):1174–1184, 2010.

[16] Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[17] David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. Chapman and Hall/CRC, 1979.

[18] Jorge M Lobo. More complex distribution models or more representative data? *Biodiversity informatics*, 5, 2008.

[19] Roger L Berger and Dennis D Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

[20] Amy L Whitehead, Steven A Julious, Cindy L Cooper, and Michael J Campbell. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical methods in medical research*, 25(3):1057–1073, 2016.

[21] Peter C O'Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

APPENDICES

This section contains the links related to all methods used in this essay.

5.1 Monte Carlo simulation for Bonferroni correction

Below is the link that can lead you to Monte Carlo simulation for bivariate normal distributed data with incorrect family wise error rate correction and then Bonferroni correction.

https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/15

5.2 Analytical function for fixed assumed correlation

The link below is for analytical function to approximate the result of Monte Carlo simulation for fixed assumed correlation method.

https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/4

5.3 Monte Carlo simulation for fixed assumed correlation

The below link is for Monte Carlo simulation to adjust family wise error rate boundaries with fixed assumed correlation method.

https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/5

5.4 Monte Carlo simulation for correlation estimators

Below is the link for estimating correlation using different estimators (Pooled, Blinded, Fisher, Olkin Pratt) between the outcome of treatments in each group.The code combines both estimations with $n = 50$ and $n = 10$.

https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/3

5.5 Monte Carlo simulation for plug in estimated correlation

$n = 10$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/1

$n = 50$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/2

5.6 Monte Carlo simulation for Distribution of estimators

$n = 10$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/13

$n = 50$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/14

$n = 1000$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/9

5.7 Monte Carlo simulation for Berger and Boos

For confidence level 99%:

$n = 10$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/16

$n = 50$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/17

For confidence level 80%:

$n = 10$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/19

$n = 50$: https://github.com/ErnestMupenziBIZI/Presention-and-talk/pull/18