MULTIPLE TESTING PROCEDURES UNDER GROUP SEQUENTIAL DESIGN

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Aiying Chen
May, 2016

Examining Committee Members:

Sanat K. Sarkar, Advisory Chair, Statistics
Subhadeep Mukhopadhyay, Statistics
Zhigen Zhao, Statistics
Dror Rom, External Reader, Ptosoft Clinical

ABSTRACT

MULTIPLE TESTING PROCEDURES UNDER
GROUP SEQUENTIAL DESIGN

Aiying Chen
DOCTOR OF PHILOSOPHY

Temple University, 2016

Sanat K. Sarkar, Chair

This dissertation is focused on multiple hypotheses testing procedures under group sequential design, in which the data are accrued sequentially or periodically in time. We propose two stepwise procedures using the error spending function approach. The first procedure controls the Family-Wise Error Rate (FWER), under the assumption that the test statistics follow normal distributions with known correlations. This procedure involves repeated application of a step-down procedure at each stage on the hypotheses that are not rejected in the previous stages. The second proposed procedure is a group sequential BH procedure (GSBH) controlling the False Discovery Rate (FDR), which is a natural extension of the original BH method from single to multiple stages under a group sequential design. Similar to the proposed step-down procedure controlling the FWER, a step-up procedure is applied on the active hypotheses at each stage in the GSBH procedure. This GSBH procedure is theoretically proved to control the FDR under some positive dependence condition. An adaptive version of GSBH procedure (ad.GSBH) is also introduced, which is proved to control the FDR under independence. Simulation studies are performed to investigate the performance of these three procedures. The simulation results show that these procedures are often powerful and provide more reduction of the expected sample sizes compared to their relevant competitors.

ACKNOWLEDGEMENTS

This dissertation could not have been completed without the support and help from so many people.

First and foremost, I would like to express my sincerest and deepest gratitude to my advisor, Dr. Sanat K. Sarkar, for guiding me and supporting me through the course of this dissertation research. He gave me insightful ideas and comments, and his extensive knowledge, rigorous teaching, and dedication to research will certainly be an inspiration to me throughout my career. It is truly my fortune to work with him.

I would also like to thank my other committee members, Dr. Zhigeng Zhao, Dr. Subhadeep Mukhopadhyay, and Dr. Dror Rom, for their time in reviewing this work, and their valuable suggestions and comments throughout this research. In addition, I would like to thank Dr. Jagbir Singh, Dr. William Wei for their thoughtful help and great support in the past several years.

I also want to show my grateful thanks to two dear friends and mentors at Merck, Li and Sammy, for their time in discussing questions with me, and reviewing my work. They also gave me lots of helpful suggestions in my research and work. Without their help, I would not be where I am right now.

Finally, I would like to thank my family for their endless support and love to complete this Ph.D degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In many fields, such as clinical trials, disease prevention etc., it is natural to monitor the results sequentially or periodically to take actions such as early termination or modification of the trial. Compared to a fixed-sample test, a group sequential design has the attractive feature of being generally more flexible and more efficient. For example, a group sequential design in clinical trial allows for interim analyses of the accumulating data which can lead to early stopping of the trial, resulting in smaller sample sizes, less patient exposure to unnecessary harmful effect and cost reduction. When performing repeated significance tests across interim looks, the probability of making a type I error is inflated. As a result, when deriving a group sequential test, the determination of the rejection boundaries is driven by the goal that the type I error is controlled at some specific nominal level. Since data are accumulated, certain dependence structure exists among the test statistics at different stages and this can be taken into account in the determination of rejection boundaries.

Instead of testing one single hypothesis during the course of data accumulation, it is common in clinical trials that multiple hypotheses are tested at each interim look. For example, in comparison to a placebo, often multiple endpoints of a treatment need to be evaluated or multiple treatments (multiple doses of a treatment) are considered. While procedures have been proposed to sequentially test an intersection of null hypotheses which can lead to an overall statement about the endpoints or treatments considered (Jennison and Turnbull (1993)), it is often desirable to make inference for each individual hypothesis, i.e., testing multiple hypotheses simultaneously. In this case, it is necessary that critical boundaries should be adjusted not only for repeated testing of the same hypothesis at different stages, but also for multiplicity due to testing multiple hypotheses simultaneously at each stage.

Two types of error rate occur in hypothesis testing problems, type I error and type II error. Type I error occurs when a true null hypothesis is falsely rejected, or falsely discovered; whereas, type II error occurs when a null hypothesis is false but accepted. Unlike controlling type I error rate in single hypothesis testing, a multiple testing procedure guards against some appropriate joint type I error rate. Several measures of the joint error rate are considered in the literature. The most common one is the family-wise error rate (FWER), the probability of making at least one false rejection, or false discovery. Alternative error rate considered more in recent research is the false discover rate (FDR), the expected proportion of false rejections among all rejections. When the number of hypotheses being tested is large, FWER controlling is extremely conservative; whereas, controlling the FDR is less conservative.

Considerable research has been conducted in the framework of group sequential multiple testing in terms of controlling the FWER. For example, Tang and Geller (1999) extended the closed testing of multiple hypotheses to the sequential setting. De and Baron (2012a, b), Bartroff and Song (2014) proposed sequential tests for multiple hypotheses controlling type I and type II error rates. Bartroff and Lai (2010) proposed a general method of constructing a multi-stage analog of Holm's procedure, hereafter to be refereed to as the multistage Holm procedure (MTholm), and also an alternative multi-stage step-down procedure assuming closedness and a nested structure among the set of tested hypotheses.

Often, at each interim stage, the test statistics corresponding to different hypotheses are correlated. For example, the multiple endpoints under investigation often exhibit similar or opposite features with changing treatment conditions; and the features of the same endpoint with increasing doses are of course related. Despite those dependence information, many group sequential multiple testing procedures were developed under the assumption that the test statistics are independent. Even though some procedures were later extended to positive dependence condition, the correlation information among test statistics were not used to construct the procedures and hence improving the performance of the procedure is more desirable.

The effect of correlation on multiple testing procedures has been investigated in the literature. It has been shown that correlation can substantially increase the number of the false positives and hence make testing procedures very unreliable (Owen (2005), Qiu et al. (2005)). Efron(2007) argued that though the individual distributions of the test statistics under null hypotheses are known, the presence of correlation among them can narrow or widen their distribution. It has been shown that when correlation information has been properly incorporated, a more powerful testing procedure can be obtained. In Chapter 4, we develop an FWER controlling procedure that has the following two features: 1) The rejection boundaries take into account dependence of the test statistics that correspond to the same hypothesis at different stages and also to the different hypotheses within the same stage, under multivariate normal distributional setting. 2) Similar to the multistage Holm procedure in Bartroff and Lai (2010), at each stage, our proposed multiple testing procedure is applied to the hypotheses that are not rejected in the previous stage.

Benjamini and Hochberg (1995) proposed a powerful and easy-to-use procedure, called the BH procedure, to control the FDR under independence. Later, Benjamini and Yekutieli (2001) proved that the BH procedure also controls the FDR under some positive dependence condition. In addition, they proposed a modification of the BH procedure, called the BY procedure, that controls the FDR under arbitrary dependence. The BH procedure has gained much popularity because of its wide applicability. Although methods for controlling the FDR in a fixed-sample design have been well developed in the literature, limited research has been conducted to develop methods for controlling the

FDR under a multistage group sequential design. Recently, Bartroff and Song (2013) proposed a sequential BH (SBH) procedure controlling the FDR and the FNR by extending the original BH procedure to a sequential or group sequential setting with appropriately adjusting the BH critical values in each stage. However, their general method does not give the exact expression of the critical boundaries. Besides, their procedure was only theoretically proved to control the FDR under independence condition.

In many studies where a group sequential design is used, the test statistics at each stage are often positively dependent, just as in the fixed sample design. It is then desired to develop a procedure which controls the FDR under positive dependence and effectively identifies the false null hypotheses in a group sequential framework. In Chapter 5, we propose an FDR controlling procedure, which we call the group sequential BH (GSBH) procedure, extending the original BH procedure from one stage to multiple stages under a group sequential design using the error spending function approach. In addition to giving explicit expressions for the critical boundaries to be easily used in real applications, the proposed GSBH procedure is theoretically proved to control the FDR under independence and some positive dependence conditions. Unlike having the common critical constants between stages as in the multistage Holm and the SBH procedures, the use of error spending function approach gives different boundaries between stages, making our two procedures flexible and attractive. Depending on the choice of error spending function, one can make rejections at earlier or later stages of the design, achieving more expected sample size saving or more power.

We carry out extensive simulation studies to investigate how our proposed procedures perform in terms of the FWER or the FDR control, power performance, as well as the expected sample size saving. Our simulation results show that the proposed procedures are more powerful than their relevant competitors.

The rest of the dissertation is organized as follows: in Chapter 2, we provide some preliminaries in terms of notations, notions of error control, such as

FWER and FDR controls, and procedures controlling those errors. Chapter 3 is a review of sequential and group sequential methods. We propose our two step-wise multiple testing procedures under group sequential design and show the simulation results in Chapters 4 and 5. Chapter 6 briefly states our future research.

CHAPTER 2

PRELIMINARIES

The purpose of this chapter is to review some background concepts for the discussions in the succeeding chapters.

Hypothesis testing involves making decisions concerning two competing statements about some population parameters, refereed to as null and alternative hypotheses, by using the observed data from a population. Sometimes, people may make erroneous decisions when accepting or rejecting the null and alternative hypotheses. Two types of error can result from any testing procedure: rejecting a true null hypothesis (type I error or false positive) and accepting a false null hypothesis (type II error or false negative). Hypothesis testing procedures are then developed to minimize the probability of making erroneous decisions about the true null (type I error), and at the same time maximizing the power of detecting false null. It is, however, impossible to simultaneously minimize the two types of error in a fixed-sample setting. Hence, minimizing the type II error while controlling the type I error at some prefixed significance level α is then what a hypothesis testing procedure does by finding a rejection region (for the null hypothesis) based on the distribution of some properly chosen test statistic.

Multiple hypotheses testing was motivated by the need to test a set of hypotheses simultaneously in scientific investigations. Here, each hypothesis testing has its own type I and type II errors, and the probability of making at

Table 2.1: Outcomes of m tested hypotheses

Hypotheses	Accept H_0	Reject H_0	Total
True Null	U	V	m_0
Non-true Null	T	S	m_1
Total	A	R	m

least one of those type I errors gets inflated when they are tested simultaneously. Hence, some appropriate joint or compound measure of type I error rate is needed. To guarantee at a certain level of significance that can be attached to the overall statistical findings, a multiple hypotheses testing procedure is then developed to guard against a suitably defined joint type I error rate.

The outcomes of a multiple testing procedure can be described as in Table 2.1, where A is the total number of accepted hypotheses; R is the total number of rejected hypotheses; V is the number of true hypotheses that are falsely rejected (type I errors); S is the number of true positives; T is the number of false negative (type II errors), and U is the number of true negatives. Among these quantities, A and R are observable once a multiple testing procedure is applied, but others are unobservable random variables. Different error measures considered in multiple testing can be defined in terms of the quantities in Table 2.1.

2.1 Overall Measure of Type I Errors

We now describe the most commonly used measures of type I error rate in multiple testing problems, although we will primarily focus on controlling the FWER and FDR in this dissertation.

2.1.1 Family-wise Error Rate (FWER)

1. The family-wise error rate (FWER) is the probability of making at least one type I error,

$$\text{FWER} = P(V \geq 1).$$

2. The generalized k-FWER is the probability of making at least k type I errors, for some pre-specified integer $k \geq 1$,

$$k\text{-FWER} = P(V \geq k).$$

3. The per-family error rate (PFER) is the expected value of the number of type I errors,

$$\text{PFER} = E(V).$$

4. The per-comparison error rate (PCER) is the expected value of the proportion of type I errors among the m tests,

$$\text{PCER} = \tfrac{1}{m}E(V).$$

Classical approaches to multiple hypotheses testing usually call for the control of the FWER, such as in the well-known Bonferroni and Holm procedures. On the other hand, when controlling the FWER at some prefixed level α, each individual hypothesis would be tested at lower levels, which leads to extremely conservative results when m is large.

2.1.2 False Discovery Rate (FDR)

False Discovery Proportion (FDP) is defined as the proportion of type I errors among the rejected hypotheses, i.e.,

$$\text{FDP} = \tfrac{V}{R \vee 1},$$

where $R \vee 1 = \max(R, 1)$. By this definition, FDP $= 0$ when $R = 0$. The false discovery rate (FDR) is the expected value of FDP, that is,

$$FDR = E(FDP) = E(\tfrac{V}{R \vee 1}) = E(\tfrac{V}{R} | R > 0)P(R > 0).$$

Because of the fact that the number of false rejection increases as the number of rejection increases, it maybe more reasonable to control the proportion of errors rather than to control the probability of making at least one error. In fact, since its proposition in Benjamini and Hochberg (1995), the FDR has gained continuous attention in practice and the literature. Two properties of the FDR are noticeable. First, when $m = m_0$, that is, all null hypotheses are true, the FDR reduces to the FWER; second, when $m_0 < m$, i.e., some of the null hypotheses are not true, the FDR is less than the FWER. Therefore, the FDR is a more conservative error rate, and so controlling it can allow more rejections.

Note that we say a multiple testing procedure controls an error rate weakly if the control is only offered when all the null hypotheses are true. If a multiple testing procedure controls the corresponding error rate under any configuration of the true and false null hypotheses, then we say such control is a strong control. Unless otherwise pointed out in this work, we always refer to the strong control.

To compare the performance of multiple testing procedures, we usually compare the power performance of the procedures under the precondition of controlling the joint type I error rate. The most commonly used in the literature is average power, which is the expected proportion of false null hypotheses that are correctly rejected among all false null hypotheses, i.e, $\frac{1}{m_1}E(S)$. Another one is false non-discovery rate (FNR), the expected proportion of the falsely accepted null hypotheses among those that are accepted, FNR$=$ $E\{\frac{U}{A \vee 1}\}$ (see Genovese and Wasserman (2002); Sarkar (2004)).

2.2 Global Testing and Multiple Testing

When testing multiple hypotheses simultaneously, either a global test or a multiple test can be designed. A global test is a single hypothesis test in which the null hypothesis H_0 is the intersection of m individual null hypotheses, $H_0 =$

$\bigcap_{i=1}^{m} H_i$, and the underlying test is designed to control the corresponding type I error rate. Whereas, in a multiple hypotheses test, there are m single null hypotheses, and the underlying test is designed to guard against some overall measure of their type I errors. Further, if the intersection null hypothesis H_0 is rejected in a global test, one can not make further decisions on which hypotheses are true, and which are false. This kind of questions are what we want to answer using a multiple testing approach.

2.3 Stepwise and Single Step Procedures

Multiple testing procedures can be categorized into two kinds: single step and stepwise. In a single step procedure, we reject any H_i when its corresponding $P_i < t$, for some threshold $t \in (0, 1)$ that guarantees a control over an overall error rate at some significance level α. Here, the decision on H_i does not depend on that regarding any other hypothesis H_j. On the other hand, in a stepwise procedure, a threshold is determined by comparing the ordered p-values, $P_{(1)} \leq \cdots \leq P_{(m)}$, with a set of nondecreasing critical constants $c_1 \leq \cdots \leq c_m$. Depending on whether to start with the minimum $P_{(1)}$ or the maximum $P_{(m)}$, stepwise procedures can be categorized into step-down and step-up procedures.

1. Step-up procedure: Given these critical constants $c_1 \leq c_2 \leq \cdots \leq c_m$, it reject $H_{(i)}$ (the hypothesis corresponding to $P_{(i)}$) for all $i \leq \hat{k}$, where $\hat{k} = max\{i, \quad P_{(i)} \leq c_i\}$, if the maximum exists; otherwise, accepts all hypotheses.
   Some commonly used step-up procedures are: Hochberg's procedure and BH procedure.

2. Step-down procedure: Given the critical constants $c_1 \leq c_2 \leq \cdots \leq c_m$, it rejects $H_{(i)}$ for all $i \leq \hat{k}$, where $\hat{k} = max\{i, \quad P_{(j)} \leq c_i, \quad \forall j = 1, 2, \ldots, i\}$, if the maximum exists; otherwise, accepts all hypotheses.
   A commonly used step-down procedure is: Holm's procedure.

It can be seen that, given the same set of critical values and p-values, a step-up procedure is generally more powerful than its step-down counterpart.

A multiple hypotheses testing procedure includes finding a rejection region based on some suitably chosen test statistics. In single hypothesis testing, the distribution of the test statistic is known under the null hypothesis, hence it is possible to derive the rejection boundary; however, in multiple hypotheses testing, the joint distribution or the dependence structure of the test statistic is usually unknown. In many cases, the multiple response variables are correlated with each other, but we do not know the configuration of the true null and untrue null, nor the dependence structure among each other. In order to derive rejection regions to control overall type I error rate, often assumptions about the dependence structure are made, for example, independence, or certain type of positive dependence.

2.4 Positive Dependence

Benjamini and Yekutieli (2001) introduced the following positive dependence condition called regression dependence on the subset (PRDS), with the subset referring to the test statistics or the p-values corresponding to the true null hypotheses. In the following definition, X_i is assumed to be the underlying test statistic or the p-value associated with testing H_i.

Definition 2.1. *The random variables* $(X_1, \ldots, X_m)$ *are said to have positive regression dependence(on subset of null test statistics or p-values, PRDS) if* $E(\phi(X_1, X_2, \ldots, X_m) \mid X_i = c)$ *is non-decreasing function of c for each* $i \in I_0$*, with* I_0 *being the set of indices corresponding to the null test statistics or p-values, and for any non-decreasing function* ϕ *of the* X_i*'s.*

A slightly weaker condition of positive dependence has often been used later, see Sarkar (2008a).

Definition 2.2. *A multivariate distribution is said to have the PRDS property if* $E(\phi(X_1, X_2, \ldots, X_m) \mid X_i \geq c)$ *is non-decreasing function of c for each*

$i \in I_0$, with I_0 being the set of indices corresponding to the null test statistics or p-values, and for any non-decreasing function ϕ of the X_i's.

Many multivariate statistics satisfy those two definitions listed above, for example, conditionally independent statistics, multivariate normal distribution with positive correlation coefficients, as well as multivariate t distribution with the associated normals having positive correlations. See Sarkar and Chang (1997), Sarkar (1998) for more details.

2.5 Testing Principles

The general principles for multiple testing procedures controlling FWER include union-intersection principle and closed testing principle.

2.5.1 Union-Intersection Principle

Roy (1953) proposed the Union-Intersection principle to construct a global test. Suppose that a test exists for each hypothesis H_i, according to the Union-Intersection principle, the rejection region for H_0 is the union of the rejection regions for each H_i. Hence, H_0 is rejected if and only of at least one H_i is rejected. Roy and Bose (1953) also showed that if a global test based on the Union-Intersection principle controls the type I error at level α, then this test is also a multiple test that strongly controls the FWER at level α. FWER controlling procedure based on this principle includes the Bonferroni test, Sidak test, Fisher's combination test, as well as the Simes test.

2.5.2 Closed Testing Principle

The closure method was proposed by Marcus, Peritz and Gabriel (1976). For a family of hypotheses $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$, consider all possible intersections of subsets of hypotheses $H_I = \bigcap_{i \in I} H_i$, where I is nonempty subset of $I_1 = \{1, 2, \ldots, m\}$. If there exists an α level test for each H_I, then a closed testing method rejects H_i if and only if every H_J containing H_I is rejected at level α for all $J \supseteq I$. It can be shown that the closed testing procedure strongly controls FWER at level α.

However, as the number of hypotheses m increases, the number of intersection tests in a closed testing procedure also increases exponentially. This is a serious burden and sometimes it is not necessary to test all nonempty intersection hypotheses H_I's containing H_i in order to make a decision on H_i. In fact, it is possible to form a step-up or step-down test as a possible shortcut version of the corresponding closed testing procedure, such as Holm's procedure (1979), Hommel's procedure (1988), and Hochberg's procedure (1988).

Before we review the most commonly used multiple testing procedures that control the FWER or FDR, we look at some well-known global testing procedures.

Now we review procedures controlling the FWER and FDR.

2.6 Procedures Controlling FWER

2.6.1 Bonferroni Procedure

The Bonferroni method is a classical and simple method that allows multiple comparison statements to be made while strongly controlling the FWER at level $\pi_0\alpha$, with π_0 being the proportion of true null hypotheses. This is a single-step test which rejects all H_i with $P_i \leq t$, with $t = \alpha/m$.

The advantage of this method is its simplicity and also the flexibility in the sense that it makes no assumptions about the joint distribution of the test statistics (strongly control FEWR under any arbitrary dependence condition). However, it can be seen that Bonferroni procedure is extremely conservative when m is large. Sidak (1967) made a modification of this procedure by setting $t^* = 1-(1-\alpha)^{\frac{1}{m}}$. Sidak's procedure is a single step test too, and more powerful than Bonferroni procedure under independence and positive dependence, since that $t^* \geq t$. Many other results have been obtained to improve the Bonferroni procedure, including Holm (1979), Hommel (1988), and Hocheberg (1988).

2.6.2 Holm's Procedure

Based on the closed testing principle, Holm (1979) derived a step-down procedure with critical values $c_i = \alpha/(m-i+1), \quad i = 1, 2, \ldots, m$. It is a shortcut version of the closed testing procedure in which the Bonferroni test is applied sequentially to test each intersection hypothesis. Holm's procedure is also distribution-free, but it is uniformly more powerful than Bonferroni procedure. Under independence and positive dependence condition, Holland and Copenhaver (1987) improved this procedure by letting $c_i^* = 1-(1-\alpha)^{\frac{1}{m-i+1}}$ to control the FWER at level α.

2.6.3 Hochberg Procedure

Hochberg (1988) derived a procedure based on Sime's global test, which is also a shortcut version of closed testing procedure. It is a step-up procedure with the same set of critical values as those in Holm's procedure, i.e., $c_i = \frac{\alpha}{m-i+1}$. Obviously, it is more powerful than Holm's procedure. However, it controls the FWER at α when the p-values are independent or positively dependent in some sense, which was proved in Sarkar and Chang (1997) and Sarkar (1998).

2.7 Procedures Controlling the FDR

2.7.1 BH and BY Procedures

The well-known Benjamini and Hochberg method, also known as the BH method was introduced by Benjamini and Hochberg (1995). The BH method is a step-up procedure, with the same critical constants as those in Simes test, i.e., $c_i = \frac{i}{m}\alpha, \quad i = 1, 2, \ldots, m$. Benjamini and Hochberg (1995) first proved that the FDR of their method can be controlled conservatively at $\pi_0\alpha$ when the p-values are independent, although later Benjamini and Yekutieli (2001) proved that this control is in fact exact (see also Sarkar, 2002). This procedure is also referred to as linear step-up procedure (LSU) since the set of critical constants are linear.

Benjamini and Yekutieli (2001) extended the use of the BH method to dependent p-values by proving that when the p-values satisfy the PRDS condition, the BH method controls the FDR at level less than or equal to $m_0\alpha/m$. Under arbitrary dependence, they made a conservative modification of the BH procedure, called the BY procedure, to control the FDR at level less than or equal to $m_0\alpha/m$. Specifically, they proposed using a step-up procedure with critical constants $c_i = \frac{i\alpha}{m\sum_{i=1}^{m} 1/i}$. Later, some other procedures were proposed to control the FDR under arbitrary dependence; for instance, Sarkar (2008b) proposed a stepwise multiple testing procedure by using the critical constants $c_i = \frac{i(i+1)\alpha}{2m^2}$; Blanchard and Roquain (2009) constructed a step-up procedure with $c_i = \frac{i(i+1)(2i+1)}{3m(m+1)}$.

2.7.2 Adaptive BH Procedures

Since the BH method controls the FDR at level less than or equal to $m_0\alpha/m$, with $m_0 < m$, and m_0 is usually unknown, it is conservative and may be improved by making use of information about m_0 from the data.

Storey (2002) advocated an estimation based approach to control the FDR. It is based on fixing the rejection region for each hypothesis, rather than finding it to control the FDR at a given level as in the BH procedure, and then estimating the FDR before controlling it. Storey considered a single-step test rejecting H_i if $P_i \leq t$ to develop his approach under a two-class mixture model, defined below, for the data.

Definition 2.3 (Two-class mixture model). *Let θ_i be a binary random variable indications that the hypothesis H_i is true or false, i.e., $\theta_i = 1$ if the hypothesis H_i is nonnull and $\theta_i = 0$ otherwise, for $i = 1, \ldots, m$. Assume that $\theta_i, i = 1, 2, \ldots, m$, are independent Bernoulli random variables with $P(\theta_i = 0) = \pi_0$*

and $P(\theta_i = 1) = \pi_1$. *Let* $X_i, i = 1, 2, \ldots, m$, *be generated as*

$$X_i|\theta_i \sim (1 - \theta_i)F_0 + \theta_i F_1.$$

Then the marginal cumulative distribution function of X *is the mixture distribution* $F(x) = \pi_0 F_0(x) + \pi_1 F_1(x)$, *and the probability density function is* $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$.

Under the mixture model, the FDR of this single step test is then given by

$$FDR(t) = \frac{\pi_0 t}{F(t)} P(R(t) > 0),$$

where, $R(t) = \sum_{i=1}^{m} I(P_i < t)$, and $F(t) = \pi_0 t + (1 - \pi_0)F_1(t)$ (see Storey 2002; Liu and Sarkar (2010)). The following point estimator of FDR was considered in Storey's work,

$$\widehat{FDR}_\lambda(t) = \frac{m\hat{\pi}_0(\lambda)t}{R(t) \vee 1}.$$

Here, $\hat{\pi}_0(\lambda) = \frac{m-R(\lambda)}{m(1-\lambda)}, \quad \lambda \in [0, 1)$. He suggested thresholding p-values with $\hat{t}_\alpha = P_{(\hat{l}(\lambda))}$, where $\hat{l}(\lambda) = \max\{1 \leq j \leq m : \quad \widehat{FDR}_\lambda(P_{(j)}) \leq \alpha\}$, then reject $H_{(1)}, H_{(2)}, \ldots, H_{(\hat{l}_\lambda)}$ in a step-up test.

By including an estimator for π_0 in $\widehat{FDR}(t)$, this estimation approach is more direct and simpler than the BH method while strongly controls the FDR in the meantime. Storey also showed that when $\pi_0 = 1$ or $\lambda = 0$, this adaptive approach under the above model is equivalent to the BH approach under the most conservative case.

Storey's idea of incorporating an estimator of π_0 into FDR opens up the possibility of developing methods other than the BH method that can potentially improve the control of FDR. These methods are generally refereed to as adaptive BH methods. Various adaptive methods are proposed in the literature after Storey's paper, each corresponding to this type of estimated FDR

$$\widehat{FDR}(t) = \frac{\hat{m}_0 t}{R(t) \vee 1},$$

here, $\hat{m}_0 = m\hat{\pi}_0$. Storey et al. (2004) proposed a slightly modified version of the estimator mentioned in Storey (2002), with $\hat{m}_0 = \frac{m-R(\lambda)+1}{1-\lambda}$ for any fixed $\lambda \in [0, 1)$. Benjamini, Krieger and Yekutieli (2006) suggested using $\hat{m}_0 = \frac{m-k+1}{1-P(k)}$ for any fixed $1 \leq k \leq m$. Other forms of estimators can be seen in Gavrilov, Benjamini and Sarkar (2009), Blanchard and Roquain (2008), Sarkar (2008b) and so on. These adaptive methods have been shown to control the FDR under independence of the p-values. Theoretical proof of the validity of these adaptive BH procedures under dependence is generally difficult and not available so far, but simulations results are encouraging.

CHAPTER 3

SEQUENTIAL METHODS AND GROUP SEQUENTIAL METHODS

In any long-term follow up study, such as a large-scale clinical trial, it is natural to make interim analyses of the data and, when necessary, take actions such as early termination or modification of the design for economic benefits and safety reasons. Such multiple looks at the data would lead to a type I error well in excess of α. Sequential statistical methods are formal statistical procedures for valid interim analyses of accruing data that were proposed to avoid such multiplicity problem. Initially, monitoring plans were fully sequential where an interim analysis is made after every observation is made. However, such continuous monitoring of the data can be a serious practical burden in reality. Instead, rather than performing analysis after every new observation, periodic analyses can be made only at a small number of intervals after groups of observations are made. Such schemes are called group sequential analysis. It has been shown that such groups sequential procedures still enjoy the many benefits of fully sequential tests in terms of lower expected sample sizes and shorter average study lengths. In this chapter, we review some of the important sequential and group sequential methods in the literature.

3.1 Sequential Methods

Wald (1947) introduced the sequential probability ratio test (SPRT), which is mainly concerned with the problems of selecting one of the two competing hypotheses. Specifically, consider testing a simple null against a simple alternative hypothesis regarding some parameter θ, say, $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$. The test statistic of the SPRT is the log of the ratio of the likelihood under the model specified by the alternative hypothesis to the likelihood under the model specified by the null hypothesis. The test chooses the constant rejection boundaries a and b such that the probability of type I and type II error are approximately equal to α and β, respectively. Then successive observation is taken as long as the log likelihood ratio is within the boundary (a, b) and the test stops once an observation is found to cross the boundary and the null hypothesis is rejected or accepted depending on whether the log-likelihood ratio is greater than b or less than a.

It was shown that this procedure leads to lower sample size comparing with the corresponding fixed sample tests. In fact, Wald and Wolfowitz (1948) proved that among all tests with type I error not exceeding α and type II error not exceeding β, the SPRT attains the smallest expected sample size when either H_0 or H_1 is true. However, despite of this optimality result, the average sample size of this test can be large when true θ is not equal to θ_0 or θ_1. Also, as we can see, the sample size of this test is not bounded since the sampling continues as long as the boundaries are not crossed. To overcome these disadvantages, instead of considering the parallel boundaries for which the critical values a and b stay constant irrespective of the cumulative sample size, non-parallel boundaries were derived where the critical boundary values depend on the sample size. Wald (1947) made a simple modification of the SPRT to truncate at a certain sample size N, which is called truncated sequential probability ratio test (TSPRT); Armitage (1957) proposed a restricted sequential procedures at a truncated sample size by replacing the parallel straight lines with two convergent straight line boundaries which are symmetric about line $x = 0$; Anderson (1960) studied the SPRT with possible truncation at N by replacing the parallel straight lines with two arbitrary straight lines. Then the intersection of the two lines is the upper bounds of the sample size N.

Another strand of sequential methods is the sequential experimental design which includes the topics such as adaptive sampling rules, for example, data dependent treatment allocation, with the purpose of reducing the number of subjects assigned to the inferior treatment. For instance, if the variance of a sample is higher than expected, then it is natural to take more observations to reach a desired power. Stein (1945) proposed a two stage sequential procedure, where the variance of the first stage is used for planning second stage; Hayre (1985) proposed two adaptive methods for choosing sample size, if the observed significance level is close to the specified significance level, then a small number of observations is planned for next stage, since a small number of observations is sufficed enough to get a desirable test result, and vice versa. In the work of Bauer and Köhne (1994), they proposed an adaptive procedure based on not the pooling data, but the combination of the p-values from the two separate stages trough Fisher's combination test.

3.2 Group Sequential Methods

In a seminar held by Culter et. al (1966), Shaw first proposed the idea of group sequential methods for a clinical trial by using a term "block sequential analysis" for the periodic data inspection and decision making. The term "group sequential design" was specifically used by Elfring and Schultz (1973) to compare two treatments with binary response. In their developed group sequential procedures, they assumed the equal number of sample size on each stage, and a maximum number of stages or groups was specified before the experiment.

Several important works that form a major impetus for group sequential methods include Pocock (1977), O'Brien and Fleming (1979), Slud and Wei (1982), Lan and DeMets (1983), Kim and DeMets (1987). Before we present a review of these methods, we first illustrate a general formulation of a group sequential procedure by taking the example of the basic two-treatment comparison problem in the simplest case where observations are independently and normally distributed with common known variance. Denote the observations for treatment A and B by X_{Ai} and X_{Bi} respectively. Assume that the $X_{Ai} \overset{\text{iid}}{\sim} N(\theta_A, \sigma^2)$ and $X_{Bi} \overset{\text{iid}}{\sim} N(\theta_B, \sigma^2)$, $i = 1, \ldots$. We want to test $H_0 : \quad \theta = 0$ against its one-sided alternative $H_A : \quad \theta > 0$ at level α, with $\theta = \theta_A - \theta_B$, the difference in means between the two populations represented by the treatment group. Assume that a maximum number of groups K and a group size n are chosen and accumulating data are analyzed after each group of $2n$ observations with n observations on each treatment group. Here we assume the equality of group sizes, that is n is constant with respect to different k's. For each $k = 1, \ldots, K$, a standardized test statistics Z_k is computed from the first k groups of observations as follows,

$$Z_k = \frac{1}{\sqrt{2nk\sigma^2}} \left(\sum_{i=1}^{nk} X_{Ai} - \sum_{i=1}^{nk} X_{Bi} \right).$$

Then the sequence $\{Z_1, \ldots, Z_k\}$ follows the multivariate normal distribution with marginal distribution $N\left(\frac{\sqrt{nk}}{\sqrt{2\sigma^2}} (\mu_A - \mu_B), 1 \right)$, $k = 1, \ldots, K$. A group sequential test then specifies a sequence of critical values, $\{c_1, \ldots, c_K\}$, which is computed numerically to achieve a specified type I error α based on the joint distribution of $\{Z_1, \ldots, Z_K\}$. The summarizing statistic Z_k is monitored by comparing it with c_k. If $Z_k \geq c_k$, the test terminates and H_0 is rejected, $k = 1, \ldots, K$. In other words, the sequence of critical values $\{c_1, \ldots, c_K\}$ forms a boundary for the sequence of test statistics $\{Z_1, \ldots, Z_K\}$ and H_0 is rejected if the boundary is crossed. If the sequence $\{Z_1, \ldots, Z_K\}$ stays within the boundary until the planned termination, then the null hypothesis is accepted at the final stage K. The group size n for the test is then determined by attaining a given power β at some specified value of interest for θ, say $\pm\delta$ that represents the clinical significant difference. Different group sequential test gives rise to different sequence of critical constants and different distributions for sample size.

The Pocock's test has the constant rejection boundaries, $c_k = C_P(K, \alpha), \quad k = 1, \ldots, K$, with $C_P(K, \alpha)$ satisfying $P_{H_0}\{\cup_{k=1}^{K}(Z_k \geq C_P(K, \alpha))\} = \alpha$. Note that $C_P(K, \alpha)$ does not depend on the group size n. It is the power requirement

$$P_{\mu_A - \mu_B = \pm\delta}\{\cup_{k=1}^{K}(Z_k \geq C_P(K, \alpha))\} = 1 - \beta,$$

that determines the appropriate group size for some δ. Again the calculation of the maximum sample size depends on the joint distribution of $Z_1, \ldots, Z_K$.

As an alternative to constant rejection boundary in Pocock test, O'Brien and Fleming (1979) proposed a test in which the critical constants c_k decrease as the study progresses, so that at the earliest analyses it is more difficult to reject H_0 and easier later on. The O'Brien and Fleming's test has a sequence of critical values $c_1, \ldots, c_K$ with $c_k = C_B(K, \alpha)\sqrt{(K/k)}$. Again, $C_B(K, \alpha)$ ensures an overall type I error probability α.

Both the Pocock (1977) and the O'Brien and Fleming (1979) tests are easy to use. The Pocock test has narrower boundaries initially, offering a greater opportunity for early stopping. The O'Brien and Fleming test has wide early boundaries which make it unlikely that a study will terminate at very early stage. As shown in their numerical studies, both tests offer reductions in expected sample size over the fixed sample test when $\mid \theta \mid$ is sufficiently large and the Pocock test has lower expected sample sizes than the O'Brien and Fleming test when $\mid \delta \mid$ is large, because it offers the opportunity of early stopping in such cases, but they have rather high maximum sample sizes and expected sample size when $\mid \delta \mid$ is small.

Wang and Tsiatis (1987) proposed a family of two-sided group sequential tests indexed by a parameter Δ with rejection boundaries

$$c_k = C_{WT}(K, \alpha, \Delta)(k/K)^{\Delta - 1/2}, \quad k = 1, \ldots, K.$$

For different Δ, the test offers different shapes of boundaries, including Pocock and O'Brien and Fleming test as special cases (respectively when $\Delta = 0.5$ and $\Delta = 0$). For various Δ values between 0 and 0.5, Wang and Tsiatis test gives intermediate boundaries, more conservative than Pocock's test, but less conservative than O'Brien and Flemming's. They also provide comparisons with Pocock (1977) and O'Brien, Flemming (1979) in terms of maximum sample size and expected sample size at certain power levels. These three sequential designs show obvious sample size reduction compared with the corresponding non-sequential design. The optimal Δ-class test provides almost the same maximum and expected sample size as Pocock's test when power is 0.9 and higher; approximately the same as Pocock and O'Brien and Flemming when power is 0.6 and lower; however when power is 0.7 and 0.8, their test produces more desirable results than the other two tests. Jennison and Turnbull (2000) tabulate more comparison results in terms of maximum sample size with different $\mid \delta \mid$ values.

Although we have introduced the above procedures in a setting of normal distribution, as shown in the corresponding papers, these tests can be easily adaptable to a variety of response distributions.

The other assumption we have made for the above group sequential procedures is the requirement of equal group sizes and predefined maximum number of interim analysis K . Slud and Wei (1982) introduced the key idea of error spending, using the exact method partitioning the total type I error rate α between interim looks for guaranteeing the overall type I error rate. The sequential tests are performed at pre-specified time points $t_1 < t_2 < \cdots < t_K$. The overall type I error α is partitioned into $\alpha_1, \alpha_2, ..., \alpha_K$,with $\sum\limits_{i=1}^{K} \alpha_i = \alpha$ where α_k is the type I error spent at interim stage k; once α_k is specified, the critical boundaries could be numerically calculated. However, in their methods, K is fixed and α_k is pre-specified. Lan and DeMets (1983), and, Kim and DeMets (1987) extended this error spending method and proposed more flexible tests that only require a maximum sample size N to be fixed. In their proposed method, a non-decreasing function $\alpha^*(t_k)$ satisfying $\alpha^*(0) = 0$, and $\alpha^*(1) = \alpha$ is used characterizing the rate of spending type I error at each analysis stage; t_k is the information level up to the k-th analysis stage which is given by $t_k = \sum_{\tilde{k}=1}^{k} \frac{n_{\tilde{k}}}{N}$; $\alpha^*(t_k)$ specifies the cumulative type I error rate spent up to time point t_k. Then the one-side critical boundaries satisfying $P_{\delta=0}(Z_1 > c_1) = \alpha^*(t_1)$, where $t_1 = n_1/N$. It can be shown that $c_1 = \Phi^{-1}(1 - \alpha^*(t_1))$. The remaining critical boundaries could be calculated successively through $P_{\theta=0}(\bigcap_{i=1}^{k-1}(Z_i < c_i), Z_k > c_k) = \alpha^*(t_k) - \alpha^*(t_{k-1})$ for $k = 2, 3, ..., K$. They also provided multiple functions as the error spending functions.

Two special α-spending functions $\alpha^*(t_k) = \alpha \cdot \log(1+(e-1)\cdot t_k)$ and $\alpha^*(t_k) = 2(1-\Phi(z_{\alpha/2}/\sqrt{t_k}))$ approximate Pocock's and O'Brien and Flemming's critical boundary values, respectively when applied to cases with equal sample size allocation. Later on, more α-spending functions were proposed, Kim and DeMets (1987) with the one-parameter family, $\alpha^*(\mid \delta \mid, t_k) = \alpha \cdot t_k^{|\delta|}$ for some positive value $\mid \delta \mid> 0$ and Hwang et al. (1990) with more generally one-parameter family

$$\alpha^*(\gamma, t_k) = \begin{cases} \alpha \cdot \frac{1-e^{-\gamma t_k}}{1-e^{-\gamma}} & \gamma \neq 0 \\ \alpha \cdot t_k & \gamma = 0 \end{cases}.$$

This spending function yields similar results to Wang and Tsiatis' (1987) Δ-class group sequential method.

Group sequential inference can be made in Bayesian paradigm by considering the parameter of interest θ as a random variable with a known prior probability distribution $P(\theta)$. The posterior distribution is updated as the data accumulating, $P(\theta|data) \propto L(data|\theta)P(\theta)$. Statistical inference concerning θ can be made at each stage. For example, a one-side or two-side interval estimate for θ can be made by setting the posterior probability equal to some pre-specified level. Also, stopping rules for early termination of the group sequential design can be determined based on the posterior distribution. For example, we can terminate the trial early at some intermediate stage k if

$$P(\theta \in \mathscr{A} \mid data) < \epsilon,$$

where ϵ is a pre-specified value. Berry (1985) proposed fully sequential procedures with the posterior probability $P(\delta > 0 \,|\, data) \geq 0.90$ or $P(\delta < 0 \,|\, data) \geq 0.90$, δ is the parameter. Freedman and Spiegelhalter (1989) presented similar stopping rules to construct their classic sequential procedures.

3.3 Group Sequential Methods with Multiple Endpoints

Because of the large expense involved in conducting a large-scale clinical trial, it is common for trials to be conducted to evaluate multiple endpoints simultaneously.

One way to handle multiple endpoints is Bonferroni correction, that is, running separate univariate group sequential procedure on each variable. For example, with m response variables, each group sequential test will be conducted at significance level α/m to guard against the overall type I error at α. The advantage of this method is its distribution-free and simplicity, although it is too conservative, and does not take into account the relationship between endpoints.

The other approach is reduction to a univariate or global test statistics, these include Hotelling's T^2 statistics, χ^2 and F statistics. Pocock et al.(1987) and Tang et al. (1989) described group sequential tests based on O'Brien's (1984) general least squared statistic (GLS); Jennison and Turnbull (1991) proposed group sequential χ^2 and F tests; Tang et al. (1993) considered group sequential tests based on the approximate likelihood ratio (ALR) statistic of Tang, Gnecco and Geller (1989). However, it is not always appropriate to combine multiple response variables into a single summary statistic. For instance, sometimes, people are more interested in different aspects of a new drug. From such procedures, one can not make a valid statistical conclusion pointing towards each individual endpoint. Therefore, considering these properties separately is more desirable, which leads to the group sequential multiple testing approaches.

3.3.1 Group Sequential Methods for Multiple Primary Endpoints

Jennison and Turnbull (1993) presented a formulation of the testing problem for a bivariate response in a fixed sample design, which rejects the single null hypothesis all at once. The null hypothesis is $H_0 : \theta_i > c$ for at least one $i, \quad i = 1, 2$; the alternative hypothesis is $H_1 : \quad \theta_i < c$ for all i. They also generalized this formulation to group sequential procedures that have both upper and lower boundaries. Separate univariate group sequential tests are employed on each of the two responses. If both univariate tests simultaneously indicate the acceptable treatment results at the same stage, then the trial is stopped to accept the treatment. Otherwise, the treatment is rejected.

Tang and Geller (1999) extended the closed testing procedure for multiple hypothesis in fixed sample setting (Lehmacher et al. 1991) to group sequential setting. The proposed closed testing procedure for multiple endpoints proceeds as: conduct interim analyses to test the global null hypothesis $\cap_{i=1}^{m} H_i$. Once this global null hypothesis is rejected at time t^*, stop the trial and apply the closed testing procedure to all the sub-hypothesis H_F based on the corresponding test statistics and critical boundaries If no hypothesis is rejected, continue the trial to next stage and repeat the closed testing procedure, until all hypotheses are rejected or the final stage is reached. Note that previously rejected hypotheses are rejected without retesting.

De and Baron (2012a) proposed sequential tests for multiple hypotheses with appropriately adjusted stopping boundaries in terms of controlling the type I and type II error rates. Bartroff and Lai (2010), Bartroff and Song (2015) introduced a general method of extending the Holm procedure, that controls the FWER, to the sequential data. In particular, their procedure assumes the critical constants $C_s^{(j)}$ satisfying the inequality

$$P(X_n^{(j)} \geq C_s^{(j)} \quad \text{for some} \quad n \in \mathbf{N}) \leq \frac{\alpha}{(m-s+1)},$$

for all $s = 1, \ldots, m$. Here $\mathbf{N}$ can be a set of possible sample sizes. Let $I_1 = \{1, 2, \ldots, m\}$ and I_j be the index set for the hypotheses that have not been rejected up to stage $j-1$, (the hypotheses under testing at stage j). By letting $n_0 = 0$ and $r_0 = 0$, then the multistage step-down procedure works as follows. Keep sampling and conduct interim analysis until at least one rejection happens at j-th stage, then apply a step down procedure and do the following:

1. (a) Define $n_j = \inf\left\{n \in \mathbf{N} : n > n_{j-1} \quad \text{and} \quad X_n^{(i)} \geq C_{r_{j-1}+1}^{(j)} \quad \text{for some} \quad i \in I_j\right\}$.
2. Denote the ordered test statistics at stage j as $X_{(1)}^{((j),n_j)} \geq \cdots \geq X_{(|I_j|)}^{((j),n_j)}$ and the corresponding hypotheses as $H_{(1)}^{((j),n_j)}, \ldots, H_{(|I_j|)}^{((j),n_j)}$.
3. (a) Reject $H_{(1)}^{((j),n_j)}, \ldots, H_{(r_j)}^{((j),n_j)}$ with r_j defined as follows:
$$r_j = \min\left\{r \geq 1 : X_{(r+1)}^{((j),n_j)} < C_{r+r_{j-1}+1}^{(j)}\right\}.$$
4. Continue on testing until reaching the final stage $j = K$ or $n_j = \max \mathbf{N}$. Update I_{j+1} be the index set of the remaining hypotheses, and $r_{j+1} = r_{j-1} + r_j$.

The procedures mentioned above do not take into account the correlation structure among the test statistics, and they can be conservative when this dependence structure is known or can be easily modeled. Therefore, in Chapter 4, we propose a new procedure which accounts for the dependence by using the error spending function approach.

Although the methods for controlling the FDR in a fixed-sample design have been well developed in the literature, limited research has been conducted to develop methods for controlling the FDR in a multistage group sequential design. Recently, a general method of extending the original BH method to sequential data controlling the FDR was introduced by Bartroff and Song (2013). Similar to the multistage Holm procedure, their rejective sequential BH procedure, by requiring the critical values satisfying the inequality $P\{X_n^{(k)} \geq C_s^{(k)} \quad \text{for some} \quad n \in \mathbf{N}\} \leq \left(\frac{s}{m}\right)\alpha$ for all s, except that they repeatedly applied a step-up procedure at each stage, and in steps 1 and 3,

1. (b) Define $n_j = \inf\left\{n \in \mathbf{N} : n > n_{j-1} \quad \text{and} \quad X_n^{(i)} \geq C_i^{(j)} \quad \text{for some} \quad i \in I_j\right\}$.

3. (b) Reject $H_{(r_j)}^{((j),n_j)}, \ldots, H_{(|I_j|)}^{((j),n_j)}$ with r_j defined as follows:

$$r_j = \min\left\{r \leq |I_j| : X_{(r)}^{((j),n_j)} \geq C_{m-r+1}^{(j)}\right\}.$$

They proved that the above procedure controls the FDR at level $\dfrac{m_0}{m}\alpha$ under independence, and at level $\dfrac{m_0}{m}\sum_{k=1}^{m}\frac{1}{k}\alpha$ under arbitrary dependence. As seen from their simulation studies, their proposed sequential BH procedure achieves a sizable reduction in expected sample size compared to the fixed sample BH procedure applied to the data cumulated up till the final stage. Even though their simulation results show the FDR is controlled at level $\frac{m_0}{m}\alpha$ under positive dependence, they did not give a theoretical proof. Also, despite the generality of their procedure, no exact expression for the calculation of the critical boundaries were given. In Chapter 5, we present our proposed group sequential BH procedure with explicit critical boundaries using the α spending function approach.

Recycling is a more powerful tool by reallocating the significance levels from the rejected hypotheses to the unrejected hypotheses. Bretz et al. (2009) and Burman et al. (2009) proposed graphical approaches to construct multiple testing procedures with recycling based on weighted Bonferroni tests. We illustrate the graphic approach of the Holm procedure with two hypotheses. The initial graph is to allocate the significance level $\alpha/2$ to each hypothesis H_1 and H_2; while in the graphic method, after H_1 is rejected, the significance level allocated to H_2 is α, instead of $\alpha/2$. Maurer and Bretz (2013) and Ye et al. (2013) extended this method to more stages to construct group sequential procedures with recycling based on weighted Holm tests. In Ye et al. (2013)'s work, they proposed two recycling group sequential procedures, group sequential Holm variable procedure (GSHv) and group sequential Holm fixed procedure (GSHf); GSHv procedure allocates the recycled significance level to all stages for the unrejected hypotheses; GSHf allocating the recycled significance level only to the final stage for the unrejected hypotheses.

3.3.2 Group Sequential Method for Primary and Secondary Endpoints

For hierarchically ordered endpoints, for example, a primary and a secondary endpoint, or a primary and several secondary endpoints, it is generally required that the primary endpoint acts as a gatekeeper for the secondary endpoint, which means the secondary endpoint can be tested if and only if the primary endpoint is tested statistically significant, otherwise the trial is stopped and null hypothesis is accepted (O'Neill 1997). Gatekeeping procedure is applied for this kind of problems.

Dmitrienko and Tamhane (2007, 2010) reviewed the gatekeeping procedures for such endpoints in the context of non-sequential settings. Glimm, Maurer, and Bretz (2010) and Tamhane et al. (2010) studied the same problem of testing a primary and a secondary endpoint in a two-stage group sequential setting. While Tamhane et al. (2010) listed the explicit primary and secondary critical boundaries strongly controlling the family-wise error rate at nominal level α.

Here is a illustration of this method for one primary and one secondary endpoint. Assume the observations for the primary endpoint are i.i.d. $N(\mu_1, \sigma_1^2)$, observations for the secondary endpoint are i.i.d. $N(\mu_2, \sigma_2^2)$, the two endpoints are correlated with coefficient $\rho \geq 0$. The null hypotheses to be tested are $H_1 : \mu_1 = 0$ and $H_2 : \mu_2 = 0$. Let the test statistics (X_i, Y_i), $i = 1, 2$ for the primary and secondary endpoint respectively, is the standardized cumulative sample means at the at i-th stage; denote the critical boundaries (c_1, c_2) for (X_1, X_2) and (d_1, d_2) for (Y_1, Y_2). The group sequential procedure for this problem operates as follows.

Stage 1: If $X_1 \leq c_1$, then continues to stage 2. If $X_1 > c_1$, reject H_1 and test H_2. If $Y_1 > d_1$, reject H_2; otherwise, accept H_2. In either case, stop the hypotheses testing.

Stage 2: If $X_2 \leq c_2$, accept H_1 and stop testing; otherwise, reject H_1 and test H_2. if $Y_2 > d_2$, reject H_2; otherwise, accept H_2.

The critical boundaries (c_1, c_2) and (d_1, d_2) satisfy controlling the FWER at level α.

According to the closedness principle, if the type I error rates for the primary and secondary endpoint are all controlled at level α, then the overall FWER is controlled at α, since the intersection $H_1 \cap H_2$ is the subset of H_1 according to the procedure, controlling type I error rate for the primary endpoint at local α also control the type I error rate for this intersection at local α.

From next chapter, we propose our group sequential procedures for multiple primary endpoints controlling the type I error rate.

CHAPTER 4

A STEP-DOWN GROUP SEQUENTIAL PROCEDURE CONTROLLING THE FWER

In this chapter, we consider testing a set of m null hypotheses $\{H_i : \mu_i = 0, \quad i = 1, \ldots, m\}$ against its one-sided alternative hypotheses $\{K_i : \mu_i > 0, \quad i = 1, \ldots, m\}$ under a $K \geq 2$ stage group sequential design. We construct a step-down procedure under group sequential design that takes the correlations between endpoints into account using error spending function approach.

Let $n_1 < n_2 < \cdots < n_K$ be the cumulative sample size up to each of K stages. Let I_k be the index set for the active hypotheses at the beginning of the k-th stage (those not rejected in the previous stages), with $I_1 = \{1, 2, \ldots, m\}$. Let I_0 be the index set of the true null hypotheses. For $k = 1, \ldots, K$, denote the test statistics corresponding to the active hypotheses at stage k as $\mathbf{X}^{((k),I_k)} = \{X_j^{((k),I_k)}, j \in I_k\}$ their ordered version as $\{X_{(1)}^{((k),I_k)} \leq \cdots \leq X_{(|I_k|)}^{((k),I_k)}\}$, with their corresponding hypotheses being $\{H_{(1)}^{((k),I_k)}, \ldots, H_{(|I_k|)}^{((k),I_k)}\}$. Here $|\cdot|$ is the cardinality of a set.

4.1 Determination of the Boundaries

In a step-down procedure, we solve for the $c_i^{(k)}$ recursively from the following set of equations for $1 \leq k \leq K, 1 \leq i \leq m$,

$$P\left\{\bigcap_{j=1}^{k}\bigcap_{s=1}^{i}(X_s^{(j)} \leq c_i^{(j)})\right\} = 1 - \alpha(t_k) \tag{4.1}$$

where t_k is the information fraction up to stage k, which is defined as $t_k = n_k/n_K$.

When deriving the boundaries $c_i^{(k)}$ from equation (4.1), we make the following assumptions. We assume the observation $\mathbf{X}_i = \{X_{i1}, X_{i2}, \cdots, X_{im}\}, i = 1, \ldots, n_K$ follows some m-variate distribution. Let the standardized cumulative sample mean, $\mathbf{X}^{(k)} = \{X_i^{(k)}, i = 1, \ldots, m\}$ with $X_i^{(k)} = \frac{1}{\sqrt{n_k}} \sum_{j=1}^{n_k} X_{ij}$, be the test statistics at stage k, s.t. under the null hypothesis H_i,

$$\begin{aligned}
&E(X_i^{(k)}) = \mu_i = 0, \quad var(X_i^{(k)}) = 1 \\
&corr(X_i^{(k)}, X_j^{(k)}) = \rho_{ij} \\
&corr(X_i^{(k)}, X_i^{(l)}) = a_{kl} \\
&corr(X_i^{(k)}, X_j^{(l)}) = \rho_{ij} a_{kl}
\end{aligned}$$

with $a_{kl} = \sqrt{n_k/n_l}$ for $i, j = 1, 2, \ldots, m; k, l = 1, 2, \ldots, K$.

4.2 The Proposed Step-down Procedure

Having recursively solved the critical boundary matrix $(c_j^{(k)})_{K\times m}$ from (4.1),

$$(c_j^{(k)})_{K\times m} = \begin{pmatrix} c_1^{(1)} & c_2^{(1)} & \cdots & c_m^{(1)} \\ c_1^{(2)} & c_2^{(2)} & \cdots & c_m^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ c_1^{(K)} & c_2^{(K)} & \cdots & c_m^{(K)} \end{pmatrix}.$$

Our proposed procedure proceeds as follows:

- Stage 1: Let $r^{(1)} = \max\{r \in I_1 : X_{(m)}^{((1),I_1)} > c_m^{(1)}, \ldots, X_{(m-r+1)}^{((1),I_1)} > c_{m-r+1}^{(1)}\}$, based on the critical boundary $\{c_1^{(1)}, c_2^{(1)}, \ldots, c_m^{(1)}\}$. Reject the hypotheses $H_{(m)}^{((1),I_1)}, \ldots, H_{(m-r^{(1)}+1)}^{((1),I_1)}$. If $r^{(1)} = m$, then stop and terminate the test; otherwise, continue to stage 2, and test the hypotheses that are not rejected at this stage.

  $\vdots$

- Stage k: In general, let $r^{(k)} = max\{r \in I_k : X_{(|I_k|)}^{((k),I_k)} > c_{|I_k|}^{(k)}, \ldots, X_{(|I_k-r+1|)}^{((k),I_k)} > c_{|I_k-r+1|}^{(k)}\}$ based on critical boundary at the k stage $\{c_1^{(k)}, \ldots, c_{|I_k|}^{(k)}\}$. Reject the hypotheses $H_{|I_k|}^{((k),I_k)}, \ldots, H_{(|I_k|-r^{(k)}+1)}^{((k),I_k)}$. If $r^{(k)} = |I_k|$, then stop and terminate the test. Otherwise, continue to stage $k+1$.

- Continue to test until all the hypotheses are rejected or reach the final stage K.

Note that the above multistage step-down procedure keeps screening out the largest statistics and rejecting their corresponding hypotheses at each stage. We assume that the rejected hypotheses are automatically rejected at consequent stages. Therefore, these rejected hypotheses won't be tested again at later stages.

Theorem 1. *The multistage step-down procedure defined above strongly controls the FEWR at the desired level* α*.*

Proof. Let m_0 be the number of the true hypotheses, $X_{(m_0)}^{((k),I_k)}$ is the maximum of the ordered test statistics that corresponds to a true hypothesis at the k-th stage, then

$$
\begin{aligned}
P\{V \geq 1\} &\leq P\{\bigcup_{k=1}^{K}(X_{(m_0)}^{((k),I_k)} > c_{m_0}^{(k)})\} \\
&= 1 - P\{\bigcap_{k=1}^{K}(X_{(m_0)}^{((k),I_k)} \leq c_{m_0}^{(k)})\} \\
&= 1 - (1 - \alpha(t_K)) \\
&= 1 - (1 - \alpha) \\
&= \alpha. \qquad (4.2)
\end{aligned}
$$

$\square$

We illustrate our procedure using the special case of two primary endpoints ($m = 2$) and two stages ($K = 2$). Assume they are distributed as bivariate normal with mean $\mu = (\mu_1, \mu_2)$ and equal correlation ρ. Consider testing $H_1 : \mu_1 = 0$ and $H_2 : \mu_2 = 0$ against their right-sided alternatives. In general, under $H_i, \mu_i = 0 (i = 1, 2), X_i \sim N(0, 1)$ and $Y_i \sim N(0, 1)$. Here, $X_i = \dfrac{1}{\sqrt{n_1}} \sum_{j=1}^{n_1} X_{ij}$ and $Y_i = \dfrac{1}{\sqrt{n_2}} \sum_{j=1}^{n_2} X_{ij}$ are the standardized cumulative test statistics for testing H_i at stage 1 and stage 2 respectively. The covariance matrix Σ between (X_1, X_2, Y_1, Y_2) are given by

$$
\begin{bmatrix}
1 & \rho & a & a\rho \\
 & 1 & a\rho & a \\
 & & 1 & \rho \\
 & & & 1
\end{bmatrix} \qquad (4.3)
$$

where $a = \sqrt{\frac{n_1}{n_2}}$. The critical boundaries, $c_1^{(1)}, c_2^{(1)}, c_1^{(2)}, c_2^{(2)}$ should satisfy the following conditions:

$$
\begin{aligned}
&P\{X_1 \leq c_1^{(1)}\} = 1 - \alpha(t_1), \\
&P\{X_1 \leq c_2^{(1)}, X_2 \leq c_2^{(1)}\} = 1 - \alpha(t_1), \\
&P\{X_1 \leq c_1^{(1)}, Y_1 \leq c_1^{(2)}\} = 1 - \alpha, \\
&P\{X_1 \leq c_2^{(1)}, X_2 \leq c_2^{(1)}, Y_1 \leq c_2^{(2)}, Y_2 \leq c_2^{(2)}\} = 1 - \alpha. \qquad (4.4)
\end{aligned}
$$

Given the covariance matrix Σ , we can use the package "mvtnorm" in R to solve for $c_1^{(1)}, c_2^{(1)}, c_1^{(2)}, c_2^{(2)}$ according to (4.4).

Remark 1. In our proposed procedure, we assume that the correlation coefficient ρ_{ij} between the test statistics at the same stage is known, and the critical boundaries were computed based on the known value of ρ_{ij}. In reality, this assumption may not hold. We now show that, when test statistics satisfy certain positive dependence condition, our procedure with critical boundaries derived based on $\rho = 0$ is the most conservative one in the sense that the FWER is still controlled at the prefixed level regardless of the true value of the ρ. Specifically, in the following proposal, we show that when $m = 2$ and $K = 2$, the boundaries based on $\rho = 0$ is the smallest when such dependence condition holds.

Proposal 1. *Suppose $m = 2$ and $K = 2$. Assume the test statistics in two stages (X_1, X_2, Y_1, Y_2) satisfy the positive orthant dependence condition ($P(X > x, Y > y) \geq P(X > x)P(Y > y)$, for all x, y) Let the boundaries $c_1^{(1)}, c_2^{(1)}, c_1^{(2)}, c_2^{(2)}$ based on the true value of ρ be computed according to (4.4). Denote the boundaries based on the $\rho = 0$ as $c_1^{(1)*}, c_2^{(1)*}, c_1^{(2)*}, c_2^{(2)*}$. Then we have $c_i^{(j)*} \leq c_i^{(j)}$ for $i = 1, 2$ and $j = 1, 2$.*

Proof: When $\rho = 0$, we have

$$\begin{aligned} c_1^{(1)*} &= \Phi_1^{-1}(1 - \alpha(t_1)), \\ c_2^{(1)*} &= \Phi_2^{-1}(\sqrt{1 - \alpha(t_1)}), \\ c_1^{(2)*} &= \Phi_2^{-1}(1 - \alpha, a), \\ c_2^{(2)*} &= \Phi_2^{-1}(\sqrt{1 - \alpha}, a, c_2^{(1)}), \end{aligned} \tag{4.5}$$

Since $c_1^{(1)} = c_1^{(1)*}$ and $c_1^{(2)} = c_1^{(1)*}$, we only need to show $c_1^{(2)} \geq c_1^{(2)*}$ and $c_2^{(2)} \geq c_2^{(2)*}$. By definition, we have,

$$\begin{aligned}
&P\{X_1 \leq c_2^{(1)*}, X_2 \leq c_2^{(1)*}\}\\
=&2\int_{-\infty}^{c_2^{(1)*}} \Phi\left(x\sqrt{\frac{1-\rho}{1+\rho}}\right)\phi(x)dx\\
\geq&\Phi(c_2^{(1)*})\\
\geq&1-\alpha(t_1), \qquad (4.6)
\end{aligned}$$

and,

$$\begin{aligned}
&P\{X_1 \leq c_2^{(1)*}, X_2 \leq c_2^{(1)*}, Y_1 \leq c_2^{(2)*}, Y_2 \leq c_2^{(2)*}\}\\
\geq&(1-\alpha(t_1))\int_{-\infty}^{c_2^{(1)*}}\int_{-\infty}^{c_2^{(1)*}} \Phi_2\left(\frac{c_2^{(2)*}-ax_1}{\sqrt{1-a^2}}, \frac{c_2^{(2)*}-ax_2}{\sqrt{1-a^2}}, \rho\right)\phi_2(x_1, x_2, \rho)dx_1dx_2\\
\geq&(1-\alpha_{(t_1)})\left\{\int_{-\infty}^{c_2^{(1)*}} \Phi\left(\frac{c_2^{(2)*}-ax}{\sqrt{1-a^2}}\right)\phi(x)dx\right\}^2\\
=&(1-\alpha_{(t_1)})\left\{\sqrt{\frac{1-\alpha}{1-\alpha(t_1)}}\right\}^2\\
=&1-\alpha. \qquad (4.7)
\end{aligned}$$

Based on (4.6) and (4.7) and the definition of $c_1^{(2)}$ and $c_2^{(2)}$, we have $c_1^{(2)} \geq c_1^{(2)*}$ and $c_2^{(2)} \geq c_2^{(2)*}$.

4.3 Numerical Studies

To evaluate the performance of the proposed procedure, we perform a simulation study to compare the FWER, the average power and the expected sample size saving of the following two procedures: the proposed procedure and the multistage Holm procedure (MTholm) in Bartroff and Lai (2010). As has been noted, like the fixed-sample Holm procedure, the multistage Holm procedure (MTholm) in Bartroff and Lai (2010) does not require the test statistics (at each stage) to satisfy any dependence assumption, as the multiplicity adjustment was Bonferroni-type adjustment. However, this generality is at the cost of the conservativeness of the procedure. When the dependence structure is known or can be easily modeled, more powerful procedures should be developed and applied. Our procedure takes advantage of this specific dependence structure and hence it is expected to be more powerful than the multistage Holm procedure. Also, it is known that multistage procedures often have a reduction in the expected sample size than their fixed-sample counterparts even though they are often less powerful. When the loss of power is not substantial and the sample size reduction is important such as those cases in clinical trials, the multistage version is then preferred.

We consider testing $m = 4$ hypotheses using the function α_{PO}, which is the case in a four-endpoint clinical trial, respectively, in 2, 3 and 4 stages by controlling the FWER at level $\alpha = 0.05$. For the 2 stage test, we generate $N = 240$ multivariate normal random variables with mean $\boldsymbol{\mu} = (0, 0, 0.2, 0.5)$ and covariance matrix $\Sigma = (1-\rho)diag(m) + \rho \mathbf{1}_m \mathbf{1}_m'$, reflecting our assumption that the correlations between the test statistics corresponding to the tested hypotheses are equal. We then calculate the test statistics at stage 1 as the standardized sample mean vector of the first $n_1 = 120$ generated observations and the test statistics at stage 2 as the standardized sample mean vector of the total $n_2 = 240$ observations. We then apply our procedure, for each fixed ρ in $(0.2, 0.3, \ldots, 0.7)$, and the multistage Holm procedure (MTholm) to the data in both stages. We also compute the expected total number of sample size for both of the sequential procedures. We repeat the above process 10000 times to compute the FWER, average power and the average sample size saving for each of the procedures out of these 10000 repetitions. Here the average power is defined to be the expected proportion of correctly rejected hypotheses out of all false null hypotheses. The expected sample size saving is the percent decrease of the total expected sample size relative the total maximum sample size, which is defined as $1 - E(N)/N$. We perform the 3 stage and 4 stage tests in the same way except that, for the 3 stage test, the test statistics at stage 1, 2 and 3 are calculated to be the standardized sample mean vector of

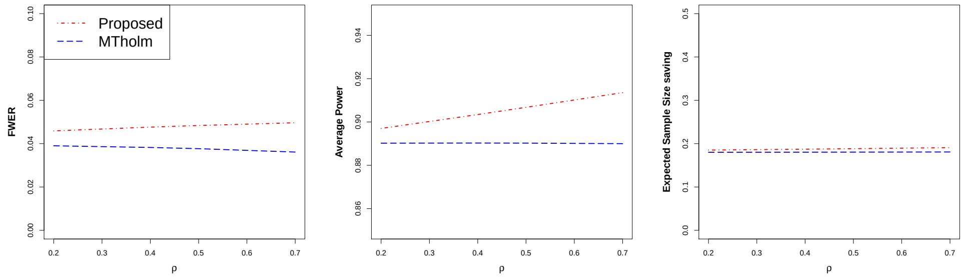


Figure 4.1: Under $K = 2$ stage group sequential design, with $\mu = (0, 0, 0.2, 0.5)$, and $(t_1, t_2) = (1/2, 1)$.

the first $n_1 = 80, n_2 = 160$ and $n_3 = 240$ generated observations, and that, for the 4 stage test, the test statistics at stage 1, 2, 3 and 4 are calculated to be the standardized sample mean vector of the first $n_1 = 60, n_2 = 120$ and $n_3 = 180$ and $n_4 = 240$ generated observations.

The simulation results are shown in Figures 4.1-4.3. We see that all procedures control the FWER in all cases and generally, the FWER of the proposed procedure is close to $\alpha = 0.05$ as ρ increases. Our procedure is more powerful than the multistage Holm procedure in terms of the average power, and saves slightly more expected sample size. In addition, we see that this power gain increases with increasing correlation ρ and increasing number of interim looks (stages). Further, our procedure achieves more sample size reduction in a four stage design than in two stage design. In addition, we also did simulation using the derived boundaries based on $\rho = 0$ in 2, 3 and 4 stages. It showed that FWERs are all controlled at level α, and this proposed procedure is more powerful than multistage Holm procedure, and less powerful than procedures using known correlation ρ, which makes sense to us, considering multistage Holm procedure made no assumption in distribution of the test statistics. Therefore, when the correlation is unknown, we can use the boundaries corresponding to the independent case.

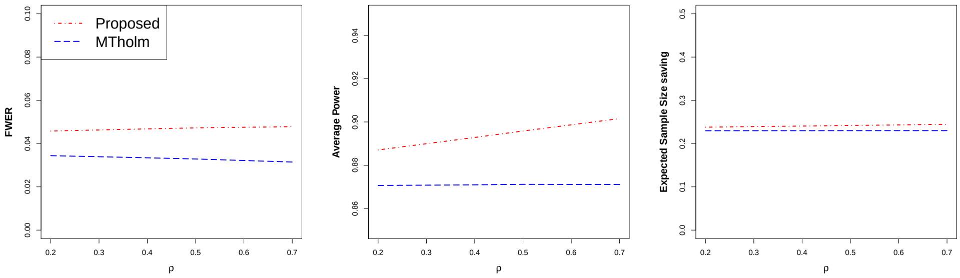


Figure 4.2: Under $K = 3$ stage group sequential design, with $\mu = (0, 0, 0.2, 0.5)$, and $(t_1, t_2, t_3) = (1/3, 2/3, 1)$.

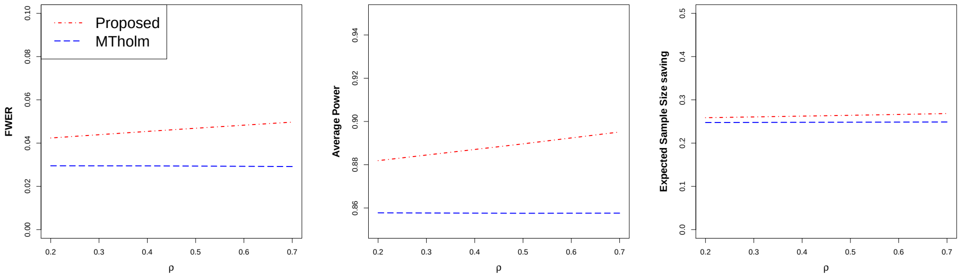


Figure 4.3: Under $K = 4$ stage group sequential design, with $\mu = (0, 0, 0.2, 0.5)$, and $(t_1, t_2, t_3, t_4) = (1/4, 2/4, 3/4, 1)$.

4.4 Conclusion

In this chapter, we derive a step-down group sequential multiple testing procedure that not only takes into account the dependence structure among the test statistics due to repeated testing of the same hypotheses but also accounts for the specific dependence structure among the test statistics corresponding to the tested hypotheses at each of the interim looks. The use of the error spending function method makes the procedure very flexible. Depending on the choose of the error spending function, one can make rejections at earlier or later stages. Through our simulation studies, we see that our proposed procedure is more powerful and also achieves more sample size reduction compared to the multistage Holm procedure which does not take into account the specific correlations among the test statistics corresponding to the hypotheses simultaneously tested at each of the interim looks. Further, it seems that more improvement of our procedure can be achieved, over the multistage Holm procedure, with increasing amount of correlations and increasing number of interim looks. On the other hand, we need to recognize that our procedure does require some specific dependence structure among the test statistics corresponding to the tested hypotheses. Hence, under this specific dependence setting, our procedure is more powerful and efficient.

CHAPTER 5

GROUP SEQUENTIAL BH PROCEDURE CONTROLLING THE FDR

In many studies where a group sequential design is used, the test statistics at each stage are often positively correlated, just as in the fixed sample design. It is then desired to develop a procedure which controls the FDR under positive dependence and effectively identifies the false null hypotheses in a group sequential framework. In this chapter, we propose such a procedure, which we call the GSBH procedure, extending the original BH procedure from one stage to multiple stages under a group sequential design by using the error spending function approach. Specifically, consider testing m null hypotheses $H_1, \ldots, H_m$ in a K-stage group sequential design. Let $\mathbf{P}^{(k)} = \{P_1^{(k)}, \ldots, P_m^{(k)}\}$ be the p-values corresponding to the m null hypotheses at stage k, $k = 1, \ldots, K$. Assume $P_i^{(k)}$, for $i = 1, \ldots, m$ and $k = 1, \ldots, K$, satisfy Assumptions 1 and 2 in Section 5.2. Our procedure first allocates to each analysis stage a nominal error level based on some α spending function. Then, at each stage, we apply, at the corresponding nominal level, a step-up procedure based on the p-values corresponding to the active hypotheses (those not rejected in the previous stages) and the BH critical values. We show that the proposed GSBH procedure controls the FDR at level $m_0\alpha/m$ under Assumptions 1 and 2, where m_0 is the proportion of true null hypotheses out of the m tested hypotheses. A new adaptive procedure in group sequential setting is introduced, by incorporating an estimate of the number of true null hypotheses, which we call ad.GSBH procedure. We theoretically prove the FDR of this ad.GSBH procedure is controlled under independence among each hypotheses.

We carry out extensive simulation studies to investigate how our two proposed procedures perform in terms of the FDR control, average power, the expected sample size saving, and the FNR, which is the expected proportion of false hypotheses that are accepted out of the total number of accepted hypotheses. We perform these simulation studies using the following two α spending function, $\alpha_{PO}(t) = \alpha \log(1+(e-1)t)$ and $\alpha_{OF}(t) = 2(1-\Phi(z_{\alpha/2}/\sqrt{t}))$, which respectively approximates the rejection boundaries of Pocock (1977) and O'Brien Flemming (OF) group sequential tests with equal group sizes. A real data set is also used to demonstrate our proposed procedure.

5.1 Notations

Suppose that the m null hypotheses $H_i, i = 1, 2, \ldots, m$, are to be simultaneously tested in a K-stage group sequential design. Let $n_1 < n_2 < \cdots < n_K$ be the cumulative sample sizes up to each of the K stages. Denote the test statistics corresponding to the m null hypotheses at stage k as $\mathbf{T}^{(k)} = \{T_j^{(k)}, j = 1, \ldots, m\}$ and their corresponding p-values as $\mathbf{P}^{(k)} = \{P_j^{(k)}, j = 1, \ldots, m\}$. Write the ordered p-values at stage k as $\{P_{(1)}^{(k)} \le \cdots \le P_{(m)}^{(k)}\}$ and their corresponding hypotheses as $\{H_{(1)}^{(k)}, \ldots, H_{(m)}^{(k)}\}$. Let I_k be the index set for the active hypotheses at the beginning of the k-th stage (those not rejected in the previous stages), $k = 1, \ldots, K$, with $I_1 = \{1, 2, \ldots, m\}$. Let I_0 be the index set of the true null hypotheses. For $k = 1, \ldots, K$, denote the p-values corresponding to the active hypotheses at stage k as $\mathbf{P}^{((k),(I_k))} = \{P_j^{((k),(I_k))}, j \in I_k\}$, their ordered versions as $\{P_{(1)}^{((k),(I_k))} \le \cdots \le P_{(|I_k|)}^{((k),(I_k))}\}$, with their corresponding hypotheses being $\{H_{(1)}^{((k),(I_k))}, \ldots, H_{(|I_k|)}^{((k),(I_k))}\}$. Here $|\cdot|$ is the cardinality of a set with $|I_0| = m_0$.

5.2 Group Sequential BH Procedure

We show that the proposed procedure controls the FDR at the desired level under the following two assumptions:

Assumption 1. *For any coordinatewise nondecreasing function* φ, $E(\varphi(\mathbf{P}^{(k)}) \mid P_i^{(k)} \leq t)$ *is nondecreasing in* t, *for each* $i \in I_0$ *and* $k = 1, \ldots, K$.

Assumption 2. *For any coordinatewise nondecreasing function* φ, $E\{\varphi(P_i^{(k)}) \mid P_i^{(k)} \leq t, \mathbf{P}^{((1),I_1)}, \ldots, \mathbf{P}^{((k-1),I_k)}\}$ *is nondecreasing in* t, *for each* $i \in I_0$ *and* $k = 1, \ldots, K$.

Assumption 1 is a common positive dependence assumption that is assumed when controlling the FDR (see Benjamin and Yekutieli (2001), also Sarkar (2002)). Assumption 2 is satisfied when the test statistics at each stage follow some common multivariate distributions, such as, multivariate normal or multivariate t distributions. See Karlin and Rinott (1980), Sarkar and Chang (1997) and Sarkar (1998) for more details.

Example 5.1. *Assume* $K = 2$ *group sequential design. Let* $\overrightarrow{\mathbf{X}}_1$ *and* $\overrightarrow{\mathbf{X}}_2$, *the test statistics at stages 1 and 2, respectively, be jointly distributed as multivariate normal, with* $\overrightarrow{\mathbf{X}}_1 \sim \overrightarrow{\mathbf{X}}_2 \sim N(\vec{\mu}, \Sigma)$ *and,*

$$cov \begin{pmatrix} \overrightarrow{\mathbf{X}}_1 \\ \overrightarrow{\mathbf{X}}_2 \end{pmatrix} = \begin{pmatrix} \Sigma & a^2\Sigma \\ a^2\Sigma & \Sigma \end{pmatrix},$$

with $a = \sqrt{n_1/n_2}$. *The correlation coefficients in* Σ *are non-negative. Then the conditional distribution of* $\overrightarrow{\mathbf{X}}_2|\overrightarrow{\mathbf{X}}_1 = \vec{x}_1 \sim N(\vec{\mu}_{2|1}, \Sigma_{2|1})$, *with* $\vec{\mu}_{2|1} = \vec{\mu} + a^2\Sigma\Sigma^{-1}(\vec{x}_1 - \vec{\mu})$, *which is increasing in* $\vec{x}_1$, *and* $\Sigma_{2|1} = \Sigma - a^2\Sigma\Sigma^{-1}\Sigma = (1 - a^2)\Sigma$. *Thus,* $\overrightarrow{\mathbf{X}}_2|\overrightarrow{\mathbf{X}}_1 = \vec{x}_1$ *follows multivariate normal distribution with positive correlation coefficients, and satisfy the PRDS condition,* $E(\phi(\overrightarrow{\mathbf{X}}_2)|\overrightarrow{\mathbf{X}}_1 =$

$\vec{x}_1, x_{2i} < t$) is an increasing function in t for any increasing function of ϕ. Therefore, Assumption 2 is satisfied.

Given an α spending function $\alpha(t)$ and K sequences of critical constants $\lambda_i^{(k)} = \frac{i}{m}\alpha_k$, $i = 1, \ldots, m$, where $\alpha_k = \alpha(t_k) - \alpha(t_{k-1}), \quad k = 1, \ldots, K$, the K stage group sequential BH procedure proceeds as follows:

- Stage 1: Let $R_1 = \max\{1 \leq i \leq m : p_{(i)}^{((1),(I_1))} \leq \lambda_i^{(1)}\}$ be the number of rejections of a step-up procedure based on $\mathbf{P}^{((1),(I_1))}$ and critical boundaries $\{\lambda_j^{(1)}, \quad j = 1, \ldots, m\}$. Reject the hypotheses $H_{(i)}^{((1),(I_1))}$ for $i \leq R_1$. If $R_1 = m$, then reject all hypotheses and stop. Otherwise, all the remaining hypotheses are tested in stage 2.

  ⋮
- Stage k: Let $R_k = \max\{j \leq |I_k| : p_{(j)}^{((k),(I_k))} \leq \lambda_{\sum_{i=1}^{k-1} R_i + j}^{(k)}\}$ be the number of rejections of a step-up procedure based on $\mathbf{P}^{((k),(I_k))}$ and the critical boundaries $\{\lambda_j^{(k)}, j = \sum_{i=1}^{k-1} R_i + 1, \ldots, m\}$. Reject the hypotheses $H_{(j)}^{((k),(I_k))}$ with all $j \leq R_k$. If $R_k = |I_k|$, then reject all hypotheses and stop testing. Otherwise, continue to the next stage.
- Continue until all the hypotheses are rejected or the final stage is reached.

Note that the above K-stage group sequential procedure successively rejects the hypotheses with small p-values by applying the above step-up procedure at each stage. Let V_k and R_k denote the number of false rejections and total rejections at stage k, respectively, for $k = 1, \ldots, K$. Then, the overall FDR of this K stage group sequential procedure is given by

$$FDR = E\left\{\frac{V_1 + V_2 + \cdots + V_K}{(R_1 + R_2 + \cdots + R_K) \vee 1}\right\}.$$

Theorem 2. *The FDR of the above defined K-stage group sequential BH procedure satisfies*

$$FDR \leq \frac{m_0}{m}\alpha,$$

under Assumptions 1 and 2.

The proof of this Theorem is given in the Appendix.

5.3 Adaptive GSBH method

Since the BH method controls the FDR at level less than or equal to $\pi_0\alpha/m$, with $\pi_0 < 1$ under most of cases and being unknown, it is conservative and may be improved by making use of information about π_0 from the data. Various adaptive versions of the BH method incorporating an estimate of π_0 into the original BH method were proposed in the literature, for example, Storey (2002), Storey, et al. (2004), Benjamini, Krieger and Yekutieli (2006), Gavrilov, Benjamini and Sarkar (2009) and Blanchard and Roquain (2008). These adaptive methods have been shown to control the FDR under independence of the p-values.

Similarly, the GSBH procedure proposed in this dissertation can also be improved through an appropriate estimator for π_0. There are many different ways of estimating π_0 in the literature. Such as, the estimator $\hat{\pi}_0 = \dfrac{m - R_0}{(1-\lambda)m}$ in Bejamini, Krieger and Yekutieli (2006), with $\lambda = \alpha/(1+\alpha)$, and R_0 is the number of rejections of BH procedure at level λ. In this work, we consider the type of estimator from Storey, et al. (2004),

$$\hat{\pi}_0 = \frac{m - R + 1}{m(1-\alpha)}.$$

In a K stages group sequential design, we consider making an estimation until the final stage based on the information from all the previous stages, which is defined as

$$\hat{\pi}_0 = \frac{m - R_1 - R_2 - \cdots - R_{k-1} + 1}{m\Pi_{i=1}^{(K-1)}(1-\alpha_i)^{(K-i)}}, \tag{5.1}$$

and then plug it into the procedure while choosing the thresholds at the final stage. In adaptive BH procedure for a single stage design, the critical values are adjusted to $\lambda_i^* = i\alpha/(m\hat{\pi}_0)$. However, since the FDR control at the fist $K-1$ stages is at level $\pi_0(\alpha_1 + \cdots + \alpha_{K-1}) = \pi_0(\alpha - \alpha_K)$, instead of setting the critical values at the K-th stage as $\lambda_i^{(K)} = i\alpha_K/(m\hat{\pi}_0)$, we consider the new adaptive procedure with critical values at the final stage as

$$\lambda_i^{*(K)} = i\frac{\alpha_K + (1-\hat{\pi}_0)(\alpha_1 + \cdots + \alpha_{K-1})}{m\hat{\pi}_0}, \tag{5.2}$$

which is obviously larger than $\lambda_i^{(K)}$.

Definition 5.1 (Adaptive BH method in Multi-stage GSPs). *Consider the GSBH procedure in a K stage group sequential design, with R_i, $i = 1, \ldots, K-1$, based on the $K-1$ sequences of critical constants $\lambda_i^{(k)} = i\alpha_k/m, k = 1, \ldots, K-1; i = 1, \ldots, m$, and the K-th stage critical constants given by $\lambda_i^{*(K)}$ in (5.2), and $\hat{\pi}_0$ defined in (5.1).*

Theorem 3. *The adaptive BH method defined above for multistage group sequential design controls the FDR control under independence.*

A proof of this theorem is given in Appendix.

5.4 Simulation Studies

In this section, we investigate how well our proposed GSBH procedure performs relative to the fixed sample BH procedure applied on the full data. It is also important to investigate how the performance of the above adaptive BH method for multistage GSPs changes between using $\lambda_i^{*(K)}$ and $\lambda_i^{(K)} = \alpha_K/(m\hat{\pi}_0)$ as the last stage critical values. We denote the adaptive procedures with $\lambda_i^{*(K)}$ and $\lambda_i^{(K)}$ by ad.GSBH1 and ad.GSBH2 respectively. We (1) generate $N = n_K = 240$ multivariate normal random variables $\mathbf{Z} \sim N(\boldsymbol{\mu}, \Sigma)$, with $\boldsymbol{\mu} = (\mu_i, i = 1, \ldots, m)$, and $\Sigma = (\rho)\mathbf{1_m}\mathbf{1_m}' + \text{diag}(1-\rho)$, where m_0 of these μ_i's are set at zero and the rest at 0.3; (2) allocate n_K/K samples to each of the K stages and consider testing $H_i : \mu_i = 0$ against $K_i : \mu_i > 0$ simultaneously for $i = 1, \ldots, m$; (3) calculate the cumulative standardized sample mean of each stage and then convert them to p-values based on the null distribution; (4) apply four procedures, the GSBH, ad.GSBH1, ad.GSBH2, and the BH procedure on the full data to the generated data; and (5) note the false discovery proportion, the proportion of false nulls that are rejected, the proportion of false nulls that are accepted, and the used sample size for all of these procedures. The critical values in the ad.GSBH1 procedure is

$$\lambda_i^{*(k)(1)} = \begin{cases} i\alpha_k/m & k = 1, \ldots, K-1. \\ i\frac{\alpha_K + (1-\hat{\pi}_0)(\alpha_1 + \cdots + \alpha_{K-1})}{m\hat{\pi}_0} & k = K. \end{cases}$$

While the critical values in the ad.GSBH2 procedure is

$$\lambda_i^{*(k)(2)} = \begin{cases} i\alpha_k/m & k = 1, \ldots, K-1. \\ i\frac{\alpha_K}{m\hat{\pi}_0} & k = K. \end{cases}$$

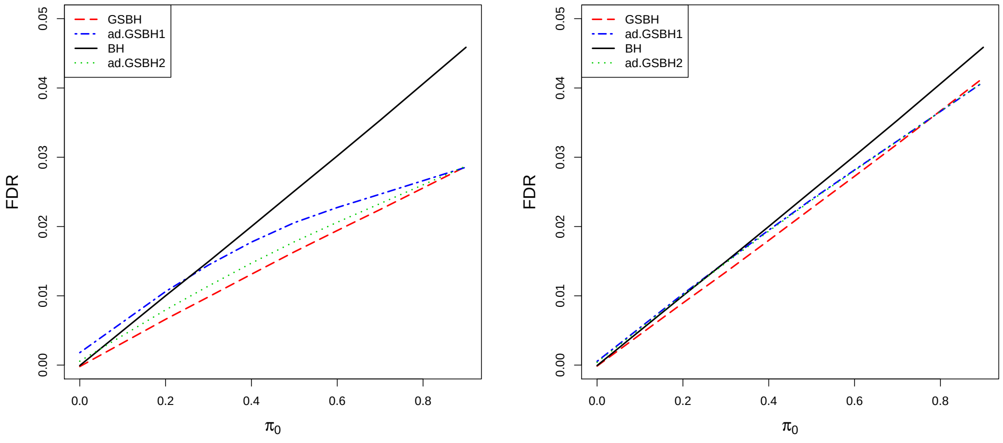
For $m = 50$, we repeat steps (1)-(5) 500 times to obtain the simulated values of FDR, average power, FNR, and expected sample size saving, defined as $1 - E(N)/(N \times m)$, with $E(N)$ being the expected used sample size. Since the BH procedure makes use of the data at all stages, it is expected to perform the best.

Since the adaptive procedures are only theoretically proven under independence, we first fix $\rho = 0$ and for each $\pi_0 = m_0/m$ in $\{0, 0.1, \ldots, 0.9\}$, computed the FDR, FNR, average power and expected sample size saving for the four procedures, using two different α spending functions: the O'Brien-Flemming (OF) α_{OF} and the Pocock (PO) α_{PO} spending function, for $K = 2$ and $K = 3$ respectively. Figure 5.1 to 5.3 show these results. The upper row shows the results for two stage $K = 2$ group sequential design, the bottom shows that for three stage $K = 3$ design. All four procedures control the FDR under all cases in Figure 5.1. Figure 5.2 and 5.3 show the difference of these three methods, ad.GSBH1, ad.GSBH2 and GSBH relative to the BH procedure in terms of power and FNR performance under equal allocation, and the power difference of these three procedures comparing with that of the BH procedure is larger when using α_{OF} than that using α_{PO} spending function. This is due to the property of these two spending functions, α_{PO} equally spends the significance level to each stage, while α_{OF} saves more to the final stage. Among these three methods, the two adaptive methods provide sharper FDR control than GSBH, and are more powerful when π_0 is not large enough, whether using α_{OF}, or using α_{PO}. It is also noted that, while if π_0 is very large and close to 1, GSBH is more powerful than the other two adaptive procedures, and seems to be a better choice. Between these two adaptive procedures, the ad.GSBH1 procedure gains more power improvements relative to the BH procedure, than ad.GSBH2. The smaller π_0 is, the more improvements. This improvements appears to be larger in the three stage design than that in two stage design.

We show that the GSBH procedure controls the FDR not only under independence, but also under some positive dependence condition. It is also interesting to investigate how the two adaptive procedures perform under positive dependence condition. We then perform the same simulation experiments except now we fix $\pi_0 = 0.5$ and for each ρ in $\{0, 0.1, \ldots, 0.9\}$, and get similar conclusion as in Figure 5.1-5.3. Figure 5.4 displays the overall FDR control for these two adaptive methods, which are still maintained at the prefixed level. It can be seen from Figure 5.5 and 5.6 that, the two adaptive methods achieve more power improvements relative to the BH procedure when the correlation among each hypothesis is not high. We also report the sample size saving of GSBH procedure in Figure 5.7-5.8 comparing with the full data BH procedure, for different values of π_0 and ρ. It is noticed the three stage design saves more than two stage design, which is the advantage of group sequential method; using the α_{PO} type spending function saves more than using α_{OF} type spending function, under equal allocation, because of α_{OF} type function allocating smaller significance level in the earlier stages.

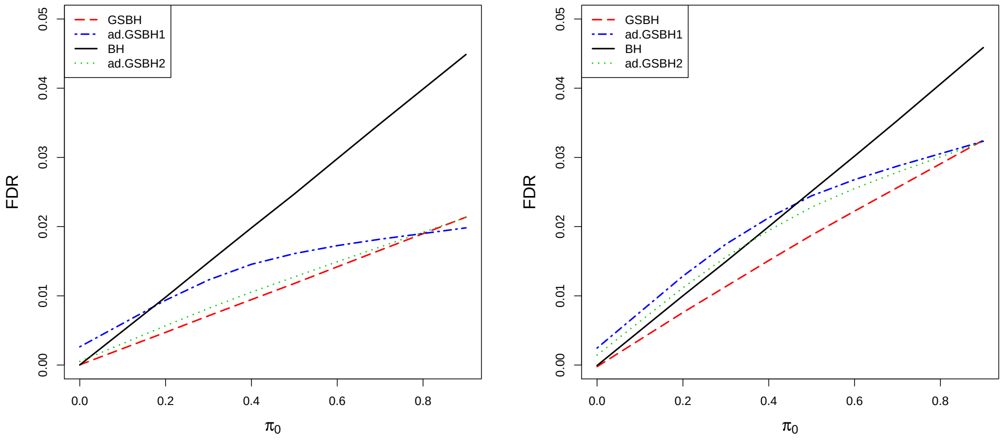
5.5 Real data application

Gene expression in multiple myeloma was generated with Affymetrix Human U95A chips, each consisting of 12,625 probe sets, in 36 patients without and 137 patients with bone lytic lesions (Tian et al., 2003; Jeffery et al., 2006; Zehetmayer et al., 2008; and Sarkar et al., 2013). As in Sarkar et al. (2013), we



(a) Adaptive PO 2 Stage

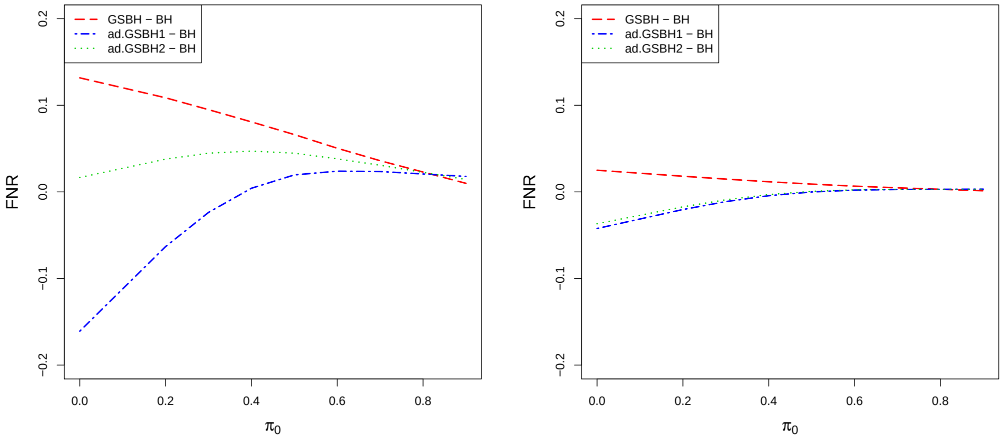
(b) Adaptive OF 2 Stage



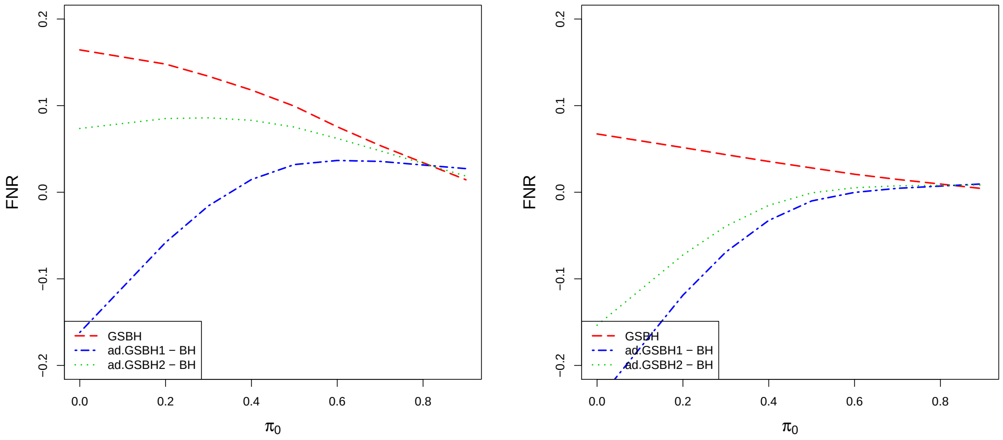
(c) Adaptive PO 3 Stage

(d) Adaptive OF 3 Stage

Figure 5.1: Comparisons of ad.GSBH1 and ad.GSBH2, with simulated FDRs of the GSBH and BH procedure.

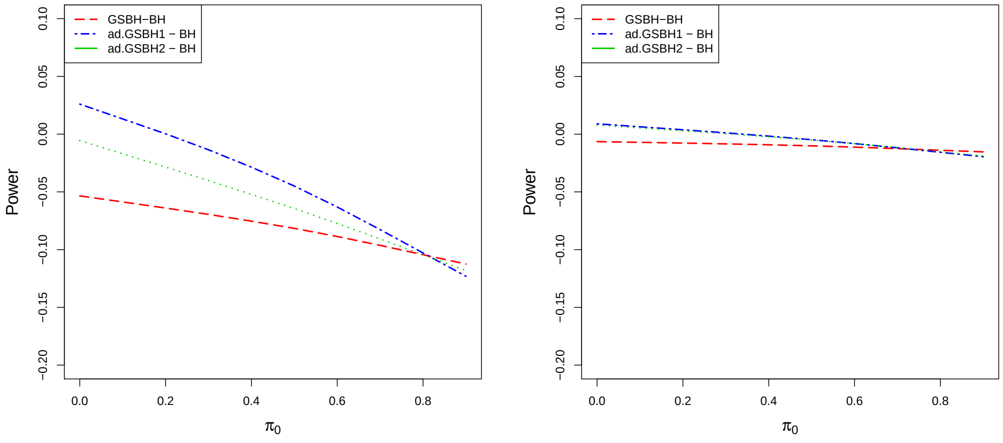


(a) Adaptive PO 2 Stage (b) Adaptive OF 2 Stage



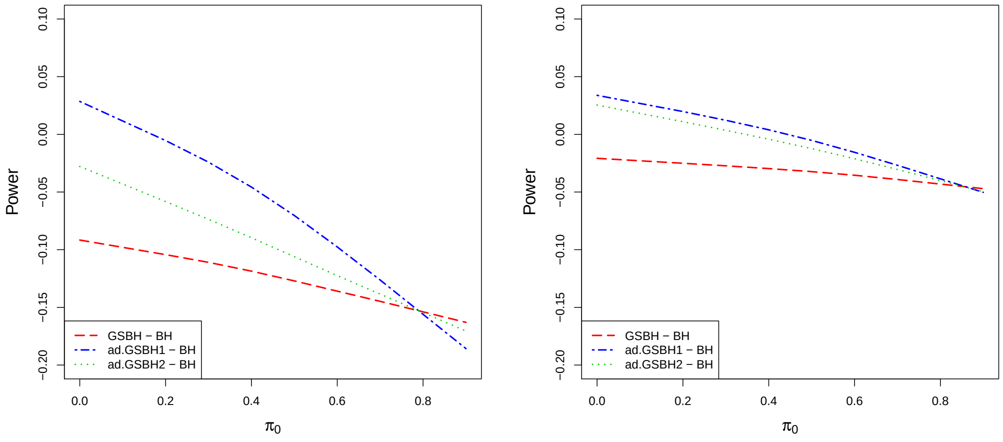
(c) Adaptive PO 3 Stage (d) Adaptive OF 3 Stage

Figure 5.2: Comparisons of ad.GSBH1 and ad.GSBH2, with simulated FNRs of the GSBH relative to the BH procedure.



(a) Adaptive PO 2 Stage

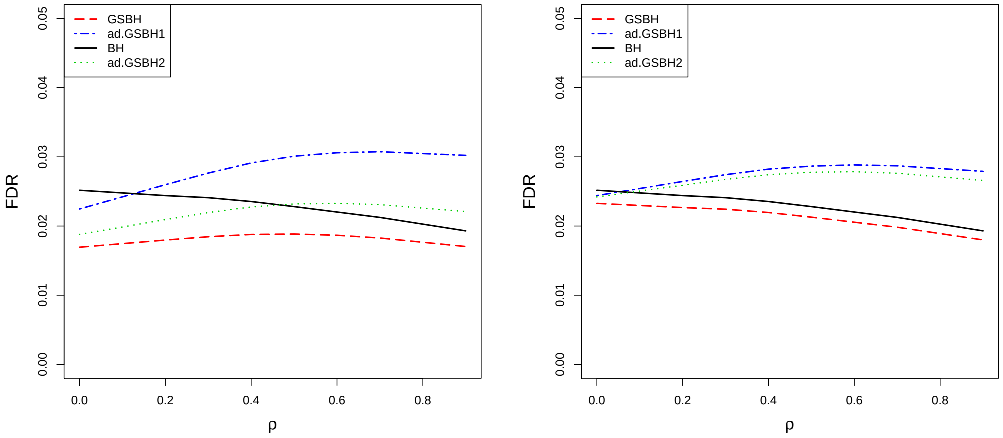
(b) Adaptive OF 2 Stage



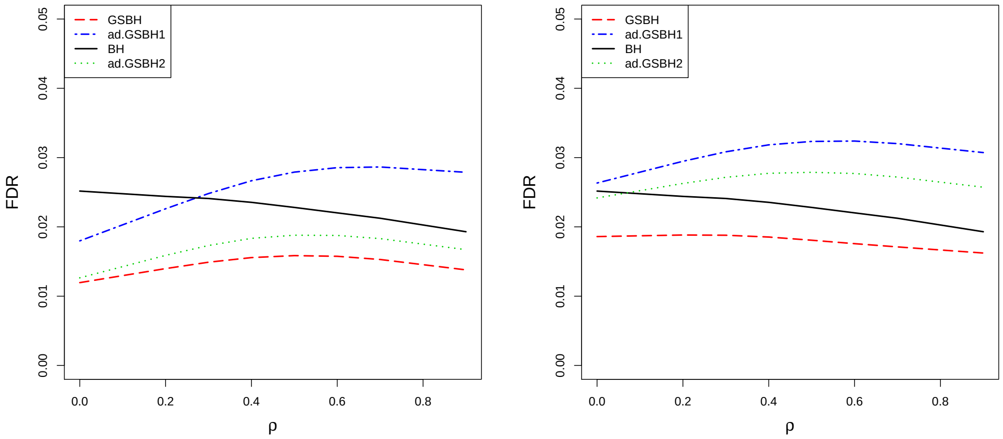
(c) Adaptive PO 3 Stage

(d) Adaptive OF 3 Stage

Figure 5.3: Comparisons of ad.GSBH1 and ad.GSBH2, with simulated average power of the GSBH relative to the BH procedure.

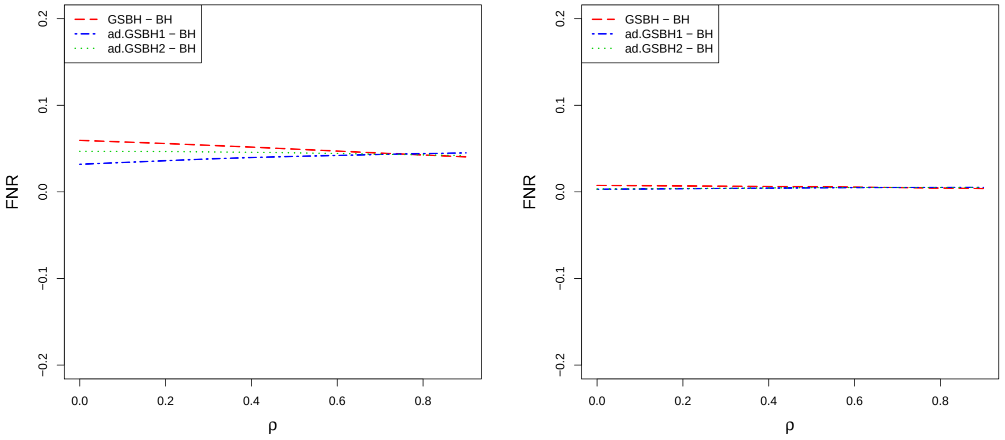


(a) PO 2 Stage (b) OF 2 Stage



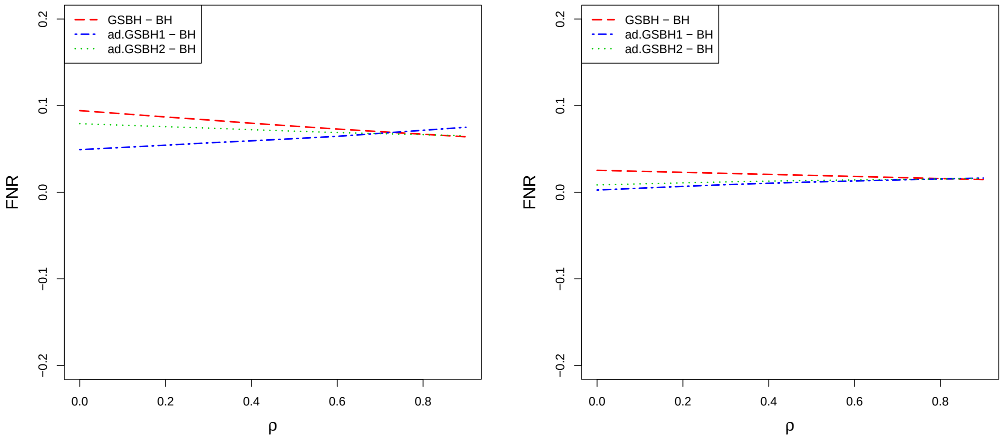
(c) PO 3 Stage (d) OF 3 Stage

Figure 5.4: Comparisons of ad.GSBH1 and ad.GSBH2, with simulated FDRs of the GSBH relative to the BH procedure.



(a) PO 2 Stage

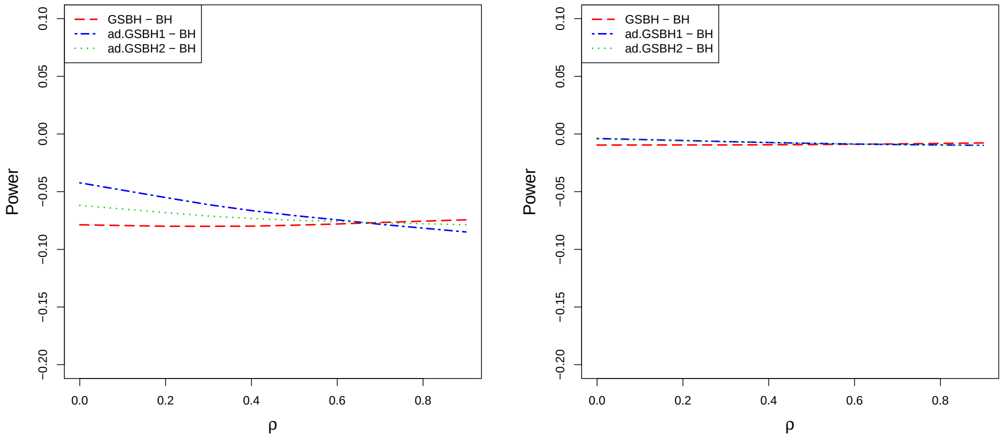
(b) OF 2 Stage



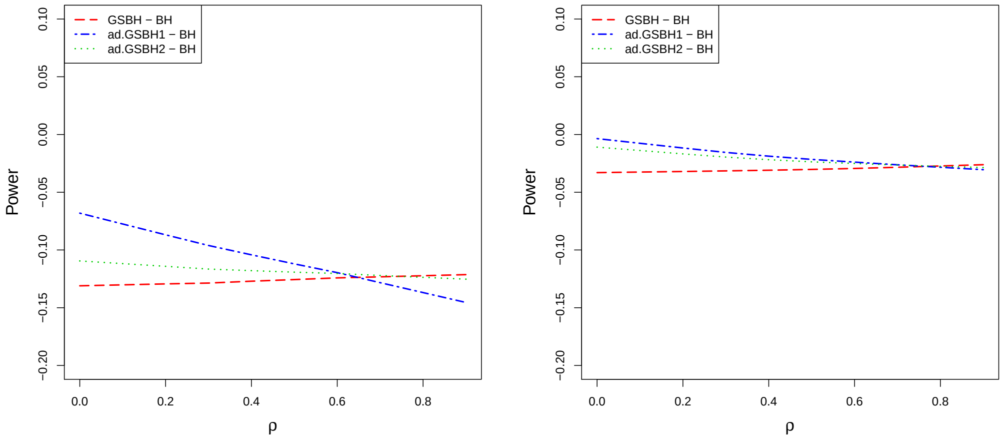
(c) PO 3 Stage

(d) OF 3 Stage

Figure 5.5: Comparisons of ad.GSBH1 and ad.GSBH2, with simulated FNRs of the GSBH relative to the BH procedure.

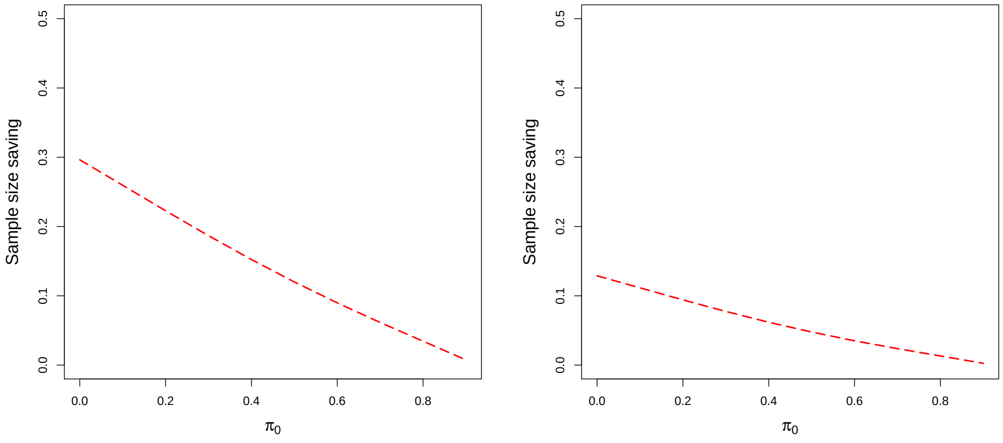


(a) PO 2 Stage (b) OF 2 Stage



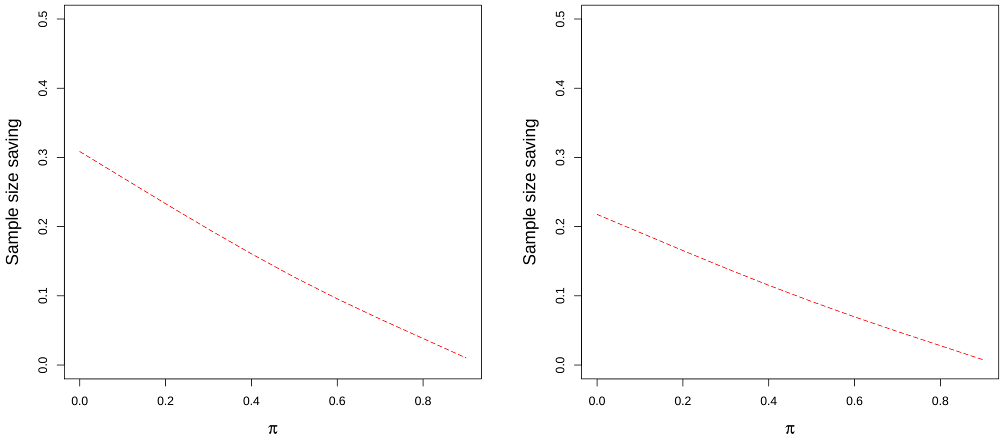
(c) PO 3 Stage (d) OF 3 Stage

Figure 5.6: Comparisons of ad.GSBH1 and ad.GSBH2, with simulated average power of the GSBH relative to the BH procedure.



(a) PO 2 Stage

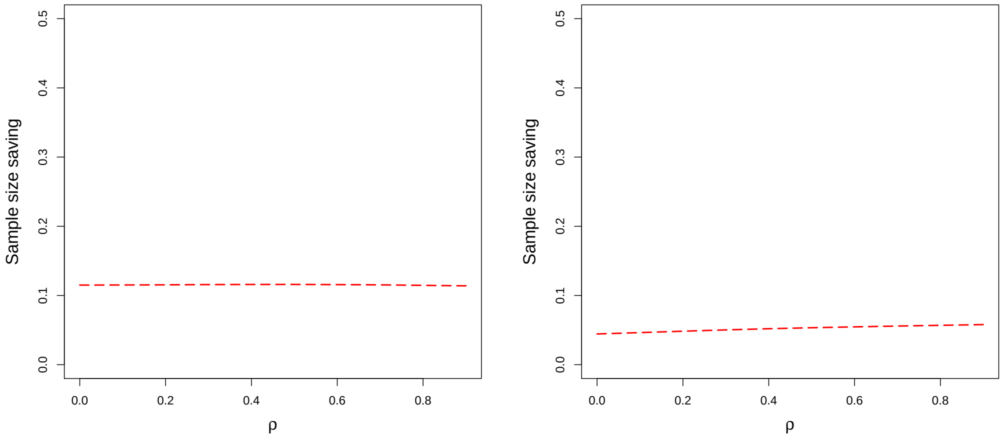
(b) OF 2 Stage



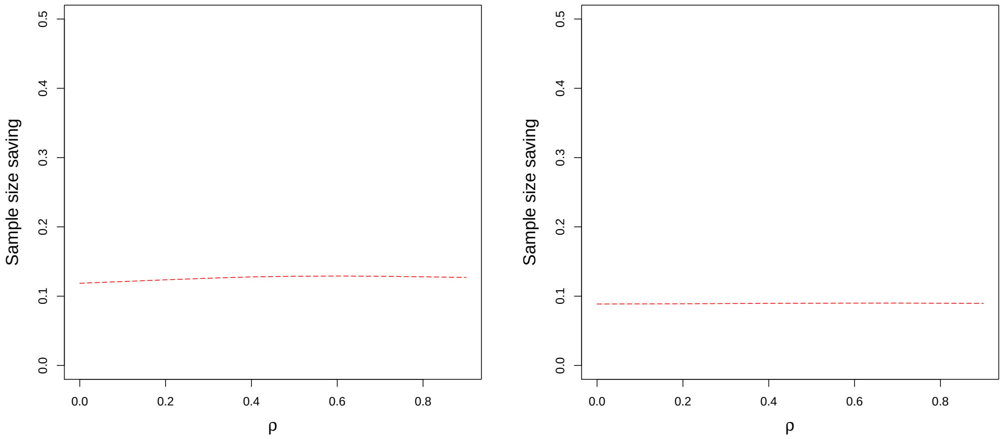
(c) PO 3 Stage

(d) OF 3 Stage

Figure 5.7: Comparisons of GSBH with simulated sample size saving with BH procedure.



(a) PO 2 Stage (b) OF 2 Stage



(c) PO 3 Stage (d) OF 3 Stage

Figure 5.8: Comparisons of GSBH with simulated sample size saving with BH procedure.

Table 5.1: The number of discoveries out of 12,265 probe sets in a 2-stage and 3-stage group sequential design for the GSBH, ad.GSBH1, ad.GSBH2 and BH procedure controlling the FDR at level $\alpha = 0.025$, each using α_{PO} and α_{OF} spending function

# Stages	α_{PO} Spending Func.			α_{OF} Spending Func.			
	GSBH	ad.GSBH1	ad.GSBH2	GSBH	ad.GSBH1	ad.GSBH2	BH
2	21	20	21	84	83	84	85
3	11	11	11	56	56	56	85

compare the gene expression measurements of 36 patients with and 36 control patients without bone lytic lesions to see whether the expressions levels are significantly higher in treatment group.

We treat the data as the patients are recruited according to a two-stage or three-stage group sequential design with equal group size. For example, with a 2-stage group sequential design, we include the first 18 patients for both the treatment and control groups for stage 1 and the next 18 patients per group for stage 2. We calculate the first and second stage p-values based on the corresponding one-sided t-test applied to the data cumulated up till the respective stages. We then apply our two proposed procedures, GSBH and ad.GSBH1 using two error spending functions α_{OF} and α_{PO}, the ad.GSBH2 and the BH procedure to the data to test all the $m = 12,625$ probe set gene expression measurements with the FDR controlled at level 0.025.

The results are reported in Table 5.1 for both two-stage and three-stage design. It can be seen that the GSBH and ad.GSBH2 procedure using α_{OF} is more powerful in the two stage designs than using α_{PO}. Particularly, it rejects almost as many as the BH procedure applied all cumulated data in the two stage design.

5.6 Discussion

In this chapter, we propose a group sequential FDR controlling procedure (GSBH), which is a natural extension of the classical BH procedure, by using the error spending function approach. Like the BH procedure, our procedure is easy to use and, importantly, our procedure controls the FDR under positively dependence, which is often the case in reality. We also produce an general version of the adaptive BH procedure from single stage to multiple stage group sequential setting. We have proved the FDR control of this adaptive GSBH procedure (ad.GSBH1) under independence among these hypotheses. Simulation studies showing the power improvement of the ad.GSBH1 procedure relative to the general form of multistage adaptive BH procedure (ad.GSBH2) is also provided. The use of error spending function makes the proposed procedure very flexible. For example, by choosing the appropriate error spending function, one can achieve early rejection at earlier or later stages of the design. Further, from the simulation studies and real data application, we see that choice of α spending function has impact on the sample size saving and power of the proposed procedure. In applications, depending on whether sample size reduction or rejection power is more critical, one can make choice of the appropriate error spending function to meet the goal.

5.7 Appendix

5.7.1 Proof of Theorem 1.

Proof.

$$\begin{aligned} FDR &= E\left\{\frac{V_1+V_2+\ldots+V_K}{R_1+R_2+\ldots+R_K}\right\} \\ &\le E\left\{\frac{V_1}{R_1}\right\}+E\left\{\frac{V_2}{R_1+R_2}\right\}+\ldots+E\left\{\frac{V_K}{R_1+R_2+\ldots+R_K}\right\} \qquad (5.3) \end{aligned}$$

Note that, in a step up procedure on every stages,

$$\begin{aligned}
E\left\{\frac{V_1}{R_1 \vee 1}\right\} &= \sum_{i\in I_0}\sum_{r_1=1}^{m}\frac{1}{r_1}Pr\left\{P_i^{(1)} \le \lambda_{r_1}^{(1)}, R_1 = r_1\right\}\\
&\le \sum_{i\in I_0}\sum_{r_1=1}^{m}\frac{1}{r_1}Pr\left\{P_i^{(1)} \le \lambda_{r_1}^{(1)}, R_1^{(-i)} = r_1 - 1\right\}\\
&\le \sum_{i\in I_0}\frac{\alpha(t_1)}{m}\sum_{r_1=0}^{m-1}Pr\left\{R_1^{(-i)} = r_1 \mid P_i^{(1)} \le \alpha_{r_1+1}^{(1)}\right\}\\
&\le \frac{m_0}{m}\alpha(t_1).
\end{aligned}$$

The last inequality is true under independence or positive dependence among the first stage p-values $\mathbf{P}^{(1)}$.

Then we have that, for $k = 2, \ldots, K$, with $r_0 = 0$,

$$\begin{aligned}
&E\left\{\frac{V_k}{R_1 + \ldots + R_k}\right\}\\
=&\sum_{i\in I_0}\sum_{j=1}^{k-1}\sum_{r_j=0}^{m-\sum_{s=0}^{j-1}r_s-1}\sum_{r_k=1}^{m-\sum_{j=0}^{k-1}r_j}\frac{1}{\sum_{i=1}^{k}r_i}Pr\left\{\bigcap_{j=1}^{k-1}P_i^{(j)} > \lambda_{r_j+1}^{(j)}, P_i^{(k)} \le \lambda_{\sum_{j=1}^{k}r_j}^{(k)}, \bigcap_{j=1}^{k}R_j = r_j\right\}\\
\le&\sum_{i\in I_0}\sum_{j=1}^{k-1}\sum_{r_j=1}^{m-\sum_{s=0}^{j-1}r_s-1}\sum_{r_k=1}^{m-\sum_{j=0}^{k-1}r_j}\frac{1}{\sum_{i=1}^{k}r_i}\\
&Pr\left\{\bigcap_{j=1}^{k-1}P_i^{(j)} > \lambda_{r_j+1}^{(j)}, P_i^{(k)} \le \lambda_{\sum_{j=1}^{k}r_j}^{(k)}, \bigcap_{j=1}^{k-1}R_j^{(-i)} = r_j, R_k^{(-i)} = r_k - 1\right\}\\
=&\sum_{i\in I_0}\sum_{j=1}^{k-1}\sum_{r_j=0}^{m-\sum_{s=0}^{j-1}r_s-1}\sum_{r_k=1}^{m-\sum_{j=0}^{k-1}r_j}\frac{1}{\sum_{i=1}^{k}r_i}Pr\left\{P_i^{(k)} < \lambda_{\sum_{j=1}^{k}r_j}^{(k)}\right\}\\
&Pr\left\{\bigcap_{j=1}^{k-1}R_j^{(-i)} = r_j, R_k^{(-i)} = r_k - 1 \mid P_i^{(k)} < \lambda_{\sum_{j=1}^{k}r_j}^{(k)}\right\}\\
\le&\sum_{i\in I_0}\frac{\alpha(t_k)-\alpha(t_{k-1})}{m}\sum_{j=1}^{k}\sum_{r_j=0}^{m-\sum_{s=0}^{j-1}r_s-1}Pr\left\{\bigcap_{j=1}^{k}R_j^{(-i)} = r_j \mid P_i^{(k)} < \lambda_{\sum_{j=1}^{k}r_j+1}^{(k)}\right\}\\
\le&\frac{m_0}{m}\{\alpha(t_k)-\alpha(t_{k-1})\}.
\end{aligned}$$

$R_1^{(-i)}$ is the number of rejections of a step-down procedure based on $\mathbf{P}^{(1)(-i)}$ and critical values $\lambda_1^{(1)}, \ldots, \lambda_{m-1}^{(1)}$. In general, for $j = 2, \ldots, k-1$, $R_j^{(-i)}$ is the number of rejections of a step-down procedure based on those j stage p-values corresponding to the active hypotheses, except for $P_i^{(j)}$, and the critical values $\lambda^{(j)}_{m-\sum_{s=1}^{j-1} R_s^{k(-i)}+1}, \ldots, \lambda^{(j)}_{m-1}$. $R_k^{(-i)}$ is the number of rejections of a step-up procedure based on k stage p-values corresponding to the active hypotheses, except for $P_i^{(k)}$, and critical values $\lambda^{(j)}_{m-\sum_{s=1}^{k-1} R_s^{(-i)}+2}, \ldots, \lambda^{(j)}_{m}$.

The last inequality comes from below:

$$
\begin{aligned}
&\sum_{i=1}^{k} \sum_{r_i=0}^{m-\sum_{j=0}^{i-1} r_j-1} Pr\left\{\bigcap_{j=1}^{k} R_j^{(-i)} = r_j \mid P_i^{(k)} < \lambda^{(k)}_{\sum_{j=1}^{k} r_j+1}\right\} \\
\leq &\int_0^1 \cdots \int_0^1 \sum_{i=1}^{k} \sum_{r_i=0}^{m-\sum_{j=0}^{i-1} r_j-1} Pr\left\{\bigcap_{j=1}^{k} R_j^{(-i)} = r_j \mid P_i^{(k)} < \lambda^{(k)}_{\sum_{j=1}^{k} r_j+1}, \mathbf{P}^{((1),I_1)}, \cdots, \mathbf{P}^{((k-1),I_{k-1})}\right\} \Pi_{l=1}^{k-1} d\mathbf{P}^{((l),I_l)} \\
= &\int_0^1 \cdots \int_0^1 \left\{\sum_{i=1}^{k} \sum_{r_i=0}^{m-\sum_{j=0}^{i-1} r_j-1} Pr\left\{\bigcap_{j=1}^{k-1} R_j^{(-i)} = r_j, R_k^{(-i)} \geq r_k \mid P_i^{(k)} < \lambda^{(k)}_{\sum_{j=1}^{k} r_j+1}, \mathbf{P}^{((1),I_1)}, \cdots, \mathbf{P}^{((k-1),I_{k-1})}\right\} - \right. \\
&\left. \sum_{i=1}^{k} \sum_{r_i=0}^{m-\sum_{j=0}^{i-1} r_j-1} Pr\left\{\bigcap_{j=1}^{k-1} R_j^{(-i)} = r_j, R_k^{(-i)} \geq r_k+1 \mid P_i^{(k)} < \lambda^{(k)}_{\sum_{j=1}^{k} r_j+1}, \mathbf{P}^{((1),I_1)}, \cdots, \mathbf{P}^{((k-1),I_{k-1})}\right\}\right\} \Pi_{l=1}^{k-1} d\mathbf{P}^{((l),I_l)} \\
\leq &\int_0^1 \cdots \int_0^1 \sum_{i=1}^{k-1} \sum_{r_i=0}^{m-\sum_{j=0}^{i-1} r_j-1} Pr\left\{\bigcap_{j=1}^{k-1} R_j^{(-i)} = r_j, R_k^{(-i)} \geq 0 \mid P_i^{(k)} < \lambda^{(k)}_{\sum_{j=1}^{k} r_j+1}, \mathbf{P}^{((1),I_1)}, \cdots, \mathbf{P}^{((k-1),I_{k-1})}\right\} \Pi_{l=1}^{k-1} d\mathbf{P}^{((l),I_l)} \\
\leq &1.
\end{aligned}
$$

Thus, the theorem is proved. □

5.7.2 The Proof of Theorem 2

Proof. When $K = 2$, the FDR in this adaptive procedure, which we denote as FDR^* satisfies the following,

$$
\begin{aligned}
FDR^* \leq &E\left\{\frac{V_1}{R_1}\right\} + E\left\{\frac{V_2^*}{R_1 + R_2^*}\right\} \\
= &\sum_{i \in I_0} E\left\{\frac{I(P_i^{(1)} \leq \lambda^{(1)}_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right\} + \sum_{i \in I_0} E\left\{\frac{I(P_i^{(1)} > \lambda^{(1)}_{R_1+1}, P_i^{(2)} \leq \lambda^{*(2)}_{R_1+R_2^*})}{R_1 + R_2^*}\right\},
\end{aligned}
\tag{5.4}
$$

V_2^* and R_2^* are the number of false rejections and total rejections respectively based on $\lambda_j^{(1)} = i\alpha_1/m$, and $\lambda_j^{*(2)}$.

$$\begin{aligned} \lambda_j^{*(2)} &= j\left\{\frac{\alpha_2 + (1-\hat{\pi}_0)\alpha_1}{m\hat{\pi}_0}\right\} \\ &= j\left\{\frac{\alpha(1-\alpha_1)}{m - R_1 + 1} - \frac{\alpha_1}{m}\right\}, \end{aligned}$$

with $\hat{\pi}_0 = (m - R_1 + 1)/(1-\alpha_1)$. The first summation in (4) is less than or equal to $\pi_0\alpha_1$. The second summation in (4) is,

$$\begin{aligned} &\sum_{i \in I_0} E\left\{\frac{I(P_i^{(1)} > \lambda_{R_1+1}^{(1)},\, P_i^{(2)} \leq \lambda_{R_1+R_2^*}^{*(2)})}{R_1 + R_2^*}\right\} \\ =&\sum_{i\in I_0}\sum_{r_1=0}^{m-1}\sum_{r_2=1}^{m-r_1} E\left\{\frac{I(P_i^{(1)} > \lambda_{r_1+1}^{(2)},\, P_i^{(2)} \leq \lambda_{r_1+r_2}^{*(2)},\, R_1 = r_1, R_2^* = r_2)}{r_1 + r_2}\right\} \\ =&\sum_{i\in I_0}\sum_{r_1=0}^{m-1}\sum_{r_2=1}^{m-r_1} E\left\{\frac{I(P_i^{(1)} > \lambda_{r_1+1}^{(1)},\, P_i^{(2)} \leq \lambda_{r_1+r_2}^{*(2)},\, \tilde{R}_1^{(-i)} = r_1, R_2^{*(-i)} = r_2 - 1)}{r_1 + r_2}\right\} \\ \leq&\sum_{i\in I_0}\sum_{r_1=0}^{m-1}\sum_{r_2=1}^{m-r_1} E\left\{\frac{I(P_i^{(2)} \leq \lambda_{r_1+r_2}^{*(2)},\, \tilde{R}_1^{(-i)} = r_1, R_2^{*(-i)} = r_2 - 1)}{r_1 + r_2}\right\} \\ =&\sum_{i\in I_0} E\left\{\frac{I(P_i^{(2)} \leq \lambda_{\tilde{R}_1^{(-i)} + R_2^{*(-i)} + 1}^{*(2)})}{\tilde{R}_1^{(-i)} + R_2^{*(-i)} + 1}\right\}, \end{aligned} \tag{5.5}$$

here, $\tilde{R}_1^{(-i)} = \tilde{R}_1^{(-i)}(\lambda_1^{(1)}, \ldots, \lambda_{m-1}^{(1)})$,similarly defined as $R_1^{(-i)}$, based on $\{P_1^{(1)}, \ldots, P_m^{(1)}\}/\{P_i^{(1)}\}$ and critical constants $\lambda_1^{(1)} \leq \cdots \leq \lambda_{m-1}^{(1)}$, and

$$\begin{aligned} R_2^{*(-i)} &= R_2^{*(-i)}(\tilde{R}_1^{(-i)}) \\ &= \max\{1 \leq j \leq m - 1 - \tilde{R}_1^{(-i)},\, P_{(j)}^{(2)(-i)} \leq \lambda_{\tilde{R}_1^{(-i)}+1+j}^{*(2)}\}. \end{aligned}$$

$\lambda_j^{*(2)}$ is less than or equal to $\lambda_j^{**(2)}$, which satisfies,

$$\lambda_j^{**(2)} = j\left\{\frac{\alpha(1-\alpha_1)}{m - \tilde{R}_1^{(-i)}} - \frac{\alpha_1}{m}\right\}.$$

Then, (5) is less than or equal to

$$
\begin{aligned}
&\sum_{i\in I_0} E\left\{\frac{I(P_i^{(2)} \le \lambda^{**(2)}_{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1})}{\tilde{R}_1^{(-i)} + R_2^{*(-i)} + 1}\right\} \\
\le&\sum_{i\in I_0} E\left\{\frac{\lambda^{**(2)}_{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1}}{\tilde{R}_1^{(-i)} + R_2^{*(-i)} + 1}\right\} \\
=&\sum_{i\in I_0} E\left\{\frac{\alpha(1-\alpha_1)}{m-\tilde{R}_1^{(-i)}} - \frac{\alpha_1}{m}\right\} \\
=&\alpha\sum_{i\in I_0} E\left\{\frac{1-\alpha_1}{m-\tilde{R}_1^{(-i)}}\right\} - \alpha_1\sum_{i\in I_0} E\left\{\frac{1}{m}\right\} \\
\le&\alpha - \pi_0\alpha_1. \qquad (5.6)
\end{aligned}
$$

The inequality in (6) comes from the following fact:

$$
\begin{aligned}
\{P_i > \lambda_{r_1+1}, R_1 = r_1\} =&\{P_i > \lambda_{r_1+1},\ P_{(r_1)} \le \lambda_{r_1}, P(r_1+1) > \lambda_{r_1+1}, \cdots, P_{(m)} > \lambda_m\} \\
\supseteq&\{P_i > \lambda_m,\ P_{(r_1)} \le \lambda_{r_1}, P(r_1+1) > \lambda_{r_1+1}, \cdots, P_{(m)} > \lambda_m\} \\
\equiv&\{P_i > \lambda_m,\ P^{(-i)}_{(r_1)} \le \lambda_{r_1}, P^{(-i)}_{(r_1+1)} > \lambda_{r_1+1}, \cdots, P^{(-i)}_{(m-1)} > \lambda_{m-1}\}.
\end{aligned}
$$

If any false negative happens at the first stage, then we have

$$
\begin{aligned}
FNR = Pr\{A>0\} - \sum_{i\in I_0}\sum_{r_1} E&\left\{\frac{I(P_i^{(1)} > \lambda^{(1)}_{r_1+1},\ P^{(1)}_{(r_1)} \le \lambda^{(1)}_{r_1}, P^{(1)}_{(r_1+1)} > \lambda^{(1)}_{r_1+1}, \cdots, P^{(1)}_{(m)} > \lambda^{(1)}_m)}{m-r_1}\right\} \\
\le Pr\{A>0\} - \sum_{i\in I_0}\sum_{r_1} E&\left\{\frac{I(P_i^{(1)} > \lambda^{(1)}_{m},\ P^{(1)(-i)}_{(r_1)} \le \lambda^{(1)}_{r_1}, P^{(1)(-i)}_{(r_1+1)} > \lambda^{(1)}_{r_1+1}, \cdots, P^{(-i)}_{(m-1)} > \lambda^{(1)}_{m-1})}{m-r_1}\right\} \\
\le 1 - \sum_{i\in I_0} E&\left\{\frac{I(P_i^{(1)} > \lambda^{(1)}_m)}{m-R_1^{(-i)}}\right\} \\
\le 1 - \sum_{i\in I_0} E&\left\{\frac{1-\alpha_1}{m-\tilde{R}_1^{(-i)}}\right\}.
\end{aligned}
$$

Therefore,

$$
\sum_{i\in I_0} E\left\{\frac{1-\alpha_1}{m-\tilde{R}_1^{(-i)}}\right\} \le 1.
$$

Thus, the FDR^* in (4) is less than or equal to $\pi_0\alpha_1 + \alpha - \pi_0\alpha_1 = \alpha$. In general, for any $K \ge 2$, with the first $K-1$ sequence of critical values

$\lambda_i^{(k)} = i\alpha_k/m, \quad k = 1, \cdots, K-1$, and

$$\begin{aligned}\lambda_j^{*(K)} &= j * \left\{ \frac{\alpha_K + (1-\hat{\pi}_0)(\alpha_1 + \cdots + \alpha_{K-1})}{m\hat{\pi}_0} \right\} \\ &= j * \left\{ \frac{\alpha}{m\hat{\pi}_0} - \frac{\alpha_1 + \cdots + \alpha_{K-1}}{m} \right\}.\end{aligned}$$

Then,

$$\begin{aligned}& E\left\{ \frac{V_K^*}{R_1 + \cdots + R_K^*} \right\} \\ =& \sum_{i \in I_0} \sum_{l=1}^{K} \sum_{r_l} E \left\{ \frac{I(\cap_{j=1}^{K-1}(P_i^{(j)} > \lambda^{(j)}_{\sum_{s=1}^{j-1} r_s + 1}),\ P_i^{(K)} \le \lambda^{*(K)}_{\sum_{s=1}^{K} r_s}, \cap_{s=1}^{K-1}(R_s = r_s), R_K^* = r_K^*)}{R_1 + \cdots + R_K} \right\} \\ \le& \sum_{i \in I_0} E \left\{ \frac{I(P_i^{(K)} \le \lambda^{*(K)}_{\sum_{j=1}^{K-1} \tilde{R}_j^{(-i)} + 1 + R_K^{*(-i)}})}{\sum_{j=1}^{K-1} \tilde{R}_j^{(-i)} + R_K^{*(-i)} + 1} \right\}. \qquad (5.7)\end{aligned}$$

V_K^* and R_K^* are respectively the false rejections and total rejections at the Kth stage in this adaptive procedure based on $\lambda_j^{*(K)}$ with

$$\begin{aligned}\tilde{R}_k^{(-i)} &= \tilde{R}_k^{(-i)}(\tilde{R}_1^{(-i)}, \cdots, \tilde{R}_{k-1}^{(-i)}) \\ &= \max\{1 \le j \le m - 1 - \sum_{s=1}^{k-1} R_s^{(-i)}, P_{(j)}^{(k)} \le \lambda^{(k)}_{\sum_{s=1}^{k-1} R_s^{(-i)} + j + 1}\}, k = 1, \cdots, K-1,\end{aligned}$$

and

$$\begin{aligned}R_K^{*(-i)} &= R_K^{*(-i)}(\tilde{R}_1^{(-i)}, \cdots, \tilde{R}_{K-1}^{(-i)}) \\ &= \max\{1 \le j\prime \le m - 1 - \sum_{s=1}^{K-1} \tilde{R}_s^{(-i)}, P_{(j)}^{(K)} \le \lambda^{*(K)}_{\sum_{s=1}^{K-1} \tilde{R}_s^{(-i)} + j\prime + 1}\}.\end{aligned}$$

Note that,

$$\lambda_j^{*(K)} = j * \left\{ \frac{\alpha \Pi_{k=1}^{K-1}(1-\alpha_k)^{K-k}}{m - \sum_{k=1}^{K-1} R_k + 1} - \frac{\alpha_1 + \cdots + \alpha_{K-1}}{m} \right\},$$

is less than or equal to

$$\lambda_j^{**(K)} = j * \left\{ \frac{\alpha \Pi_{k=1}^{K-1}(1-\alpha_k)^{K-k}}{m - \sum_{k=1}^{K-1} \tilde{R}_k^{(-i)}} - \frac{\alpha_1 + \cdots + \alpha_{K-1}}{m} \right\}.$$

Then, (4) is equal to or less than

$$
\begin{aligned}
&\sum_{i\in I_0} E\left\{\frac{I(P_i^{(K)} \leq \lambda^{**(K)}_{\sum_{j=1}^{K-1}\tilde{R}_j^{(-i)}+R_K^{*(-i)}+1})}{\sum_{j=1}^{K-1}\tilde{R}_j^{(-i)}+R_K^{*(-i)}+1}\right\}\\
=&\sum_{i\in I_0} E\left\{\frac{Pr(P_i^{(K)} \leq \lambda^{**(K)}_{\sum_{j=1}^{K-1}\tilde{R}_j^{(-i)}+1+R_K^{*(-i)})}}{\sum_{j=1}^{K-1}\tilde{R}_j^{(-i)}+R_K^{*(-i)}+1}\right\}\\
\leq&\sum_{i\in I_0} E\left\{\frac{\lambda^{**(K)}_{\sum_{j=1}^{K-1}\tilde{R}_j^{(-i)}+R_K^{*(-i)}+1}}{\sum_{j=1}^{K-1}\tilde{R}_j^{(-i)}+R_K^{*(-i)}+1}\right\}\\
=&\alpha\sum_{i\in I_0} E\left\{\frac{\Pi_{k=1}^{K-1}(1-\alpha_k)^{K-k}}{m-\sum_{k=1}^{K-1}\tilde{R}_k^{(-i)}}\right\}-(\alpha_1+\cdots+\alpha_{K-1})*\sum_{i\in I_0} E\left\{\frac{1}{m}\right\}\\
\leq&\alpha\sum_{i\in I_0} E\left\{\frac{Pr(P_i^{(1)}>\alpha_1,\cdots,P_i^{(K-1)}>\alpha_{K-1})}{m-\sum_{k=1}^{K-1}\tilde{R}_k^{(-i)}}\right\}-\pi_0(\alpha_1+\cdots+\alpha_{K-1})\\
\leq&\alpha-\pi_0(\alpha_1+\cdots+\alpha_{K-1}).
\end{aligned}
$$

The last inequality comes form the fact that, if any false negative happens, then

$$
\begin{aligned}
&\{P_i^{(1)}>\lambda^{(1)}_{r_1+1},\cdots,P_i^{(K-1)}>\lambda^{(K-1)}_{\sum_{k=1}^{K-1}r_k+1},\cap_{k=1}^{K-1}(R_k=r_k)\}\\
\supseteq&\{P_i^{(1)}>\alpha_1,\cdots,P_i^{(K-1)}>\alpha_{K-1},\cap_{k=1}^{K-1}(\tilde{R}_k^{(-i)}=r_k)\},
\end{aligned}
$$

and

$$
\begin{aligned}
FNR=&Pr\{A>0\}-\\
&\sum_{i\in I_0}\sum_{j=1}^{K-1}\sum_{r_j}\frac{1}{m-\sum_{j=1}^{K-1}r_j}Pr\{P_i^{(1)}>\lambda^{(1)}_{r_1+1},\cdots,P_i^{(K-1)}>\lambda^{(K-1)}_{\sum_{k=1}^{K-1}r_k+1},\cap_{k=1}^{K-1}(R_k=r_k)\\
&\leq Pr\{A>0\}-\\
&\sum_{i\in I_0}\sum_{j=1}^{K-1}\sum_{r_j}\frac{1}{m-\sum_{j=1}^{K-1}r_j}Pr\{P_i^{(1)}>\alpha_1,\cdots,P_i^{(K-1)}>\alpha_{K-1},\cap_{k=1}^{K-1}(\tilde{R}_k^{(-i)}=r_k)\}\\
&\leq 1-\sum_{i\in I_0}E\left\{\frac{Pr\{P_i^{(1)}>\alpha_1,\cdots,P_i^{(K-1)}>\alpha_{K-1}\}}{m-\sum_{k=1}^{K-1}\tilde{R}_k^{(-i)}}\right\}.
\end{aligned}
$$

Therefore,

$$\sum_{i \in I_0} E \left\{ \frac{Pr\{P_i^{(1)} > \alpha_1, \cdots, P_i^{(K-1)} > \alpha_{K-1}\}}{m - \sum_{k=1}^{K-1} \tilde{R}_k^{(-i)}} \right\} \leq 1.$$

Thus, the overall FDR of this adaptive procedure is less than or equal to $\pi_0(\alpha_1 + \cdots + \alpha_{K-1}) + \alpha - \pi_0(\alpha_+ \cdots + \alpha_{K-1}) = \alpha$.

The theorem is proved. $\square$

CHAPTER 6

SUMMARY AND FUTURE WORK

The focus of this dissertation has been deriving procedures that control the commonly used error rates in testing multiple hypotheses under group sequential designs. In Chapter 4, we propose a group sequential step-down procedure controlling the FWER which takes into account the correlations among the hypotheses. This procedure makes the assumption that the test statistics corresponding to the null hypotheses follow multivariate normal distribution with known correlation structure. In Chapter 5, we extend the BH procedure to a multistage procedure that controls the FDR under group sequential designs as well as an adaptive version of this procedure. We theoretically prove that while the proposed procedure in its non-adaptive form controls the FDR under independence or some type of positive dependence condition, its adaptive version controls the FDR under independence. In both of the above procedures, we make use of the α-spending function to allocate the total error rates across different stages of the design and provide the explicit boundary values, which make our procedures flexible and easy to use.

Other error rates that scientists seek to control in large scale and multistage experiments include k-FWER and γ-FDP. For example, genes in a microarry experiment usually exhibit strong dependency on each other, making it unlikely to detect exactly one significant gene. In such cases, allowing more false discoveries is more desirable. The k-FWER, defined as the probability of rejecting at lease $k \geq 1$ true null hypotheses was proposed by Lehmann and Romano (2005) as a less stringent error rate than the FWER. A generalization of the FDR, γ-FDP, which is a probability of that the false discovery proportion exceeds some fixed number $\gamma \in [0, 1)$, was also proposed in Lehmann and Romano (2005), where a step-down procedure controlling this error rate was developed. Later, Guo, He, and Sarkar (2014) further generalized the methodology of γ-FDP under dependence assumption. As a future research direction, we will consider exploring the possibility of developing procedures that control these alternative error rates in the framework of group sequential design where interim analyses are arranged.

Benjamini and Yekutieli (2001) proposed a step-up procedure, called the BY procedure, that controls the FDR under arbitrary dependence. This procedure makes use of the Hommel inequality and is very conservative in general. However, when the positive dependence assumption is not supported and hence the BH procedure not applicable, this is an alternative that could be used. One other future direction of ours is to investigate the possibility of developing a multi-stage procedures that will control the FDR under arbitrary dependence under group sequential designs.

Besides the frequentist approach on controlling the relevant error rates in multiple testing, in recent years, much research has been done to advocate the Bayesian multiple testing procedures. For example, Sun and Cai (2007), He, Sarkar, and Zhao (2014), et al.. Not much research has been conducted in the direction of deriving multi-stage procedures that controls the relevant error rates from Bayesian point of view. This is another direction we want to explore.

REFERENCES

[1] R Core Team (2013). R: A Language and Environment for Statistical Computing. http://www.R-project.org/.

[2] Anderson,T.W. (1960). A modification of sequential probability ratio test to reduce the sample size. *Ann. Math. Statist.* **31**, 165-197.

[3] Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **44**, 9-26.

[4] Armitage, P. and Schneiderman, M. (1958). Statistical problems in a mass screening program. *Ann. NY. Acad. Sci.* **76**, 896-908.

[5] Bartroff, J. and Lai, T.L. (2010). Multistage tests of multiple hypotheses. *Communications in Statistics - Theory & Methods* **39**, 1597-1607.

[6] Bartroff, J. and Song, J. (2013). Sequential tests of multiple hypotheses controlling false discovery and non-discovery rates.(Submitted Online)

[7] Bartroff, J. and Song, J. (2014). Sequential tests of multiple hypotheses controlling type I and II family-wise error rates. *J. Statist. Plann. Inf.* **14**, 100-114.

[8] Bartroff, J. and Song, J. (2015). A rejection principle for sequential tests of multiple hypotheses controlling family-wise error rates. *Scandinavian J. of Statist.* **43**, 3-19.

[9] Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie* **20**, 130-148.

[10] Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statist. Med.* **18**, 1833-1848.

[11] Bauer, P. and Kohne, K. (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50**, 1029-1041.

[12] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

[13] Benjamini, Y., Krieger, A.M. and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika* **93**, 491-507.

[14] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188.

[15] Berry, D.A. (1985). Interim analysis in clinical trials: Classical vs. Bayesian approaches. *Statist. Med.* **4**, 521-526.

[16] Blanchard, G. and Roquain, E. (2009). Adaptive FDR control under dependence and independence. *J. Mach. Learn. Res.* **10**,2837-2871.

[17] Bretz, F., Maurer, W., Brannath W. and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statist. Med.* **28**, 586-604.

[18] Burman, C.F., Sonesson, C. and Guilbaud, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Statist. Med.* **28**, 739-761.

[19] Culter, S.J., Greenhouse, S.W., Cornifield, J. and Schneiderman, M.A. (1966). The role of hypothesis testing in clinical trials. *J. Chron. Diseases* **19**, 857-892.

[20] De, S. and Baron, M. (2012a). Sequential Bonferroni methods for multiple hypothesis testing with strong control of familywise error rates I and II. *Sequential Analysis* **31**, 238-262.

[21] De, S. and Baron, M. (2012b). Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *J. Statist. Plann. Inf.* **142**, 2059-2070.

[22] Dmitrienko A. and Tamhane A.C. (2007). Gatekeeping procedures with clinical trial applications. *J. Pharma. Statist.* **6**, 171-180.

[23] Dmitrienko, A. and Tamhane, A.C. (2010). Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Statist. Med.* **30**, 1473-1488.

[24] Eales, J. and Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13-24.

[25] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102**, 93-103.

[26] Elfring, G.L. and Schultz, J.R. (1973). Group sequential designs for clinical trials. *Biometrics* **29**, 471-477.

[27] Freedman, L.S. and Spiegelhalter, D.J. (1989). Comparison of bayesian with group sequential for monitoring clinical trials. *Contr. Clin. Trials* **10**, 357-367.

[28] Gavrilov, Y., Benjamini, Y. and Sarkar, S. (2009). An adaptive step-down procedure with proven FDR control. *Ann. Statist.* **37**, 619-629.

[29] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Res. Statist. Soc. B* **64**, 499-517.

[30] Glimm, E., Maurer, W., and Bretz, F. (2010). Hierarchical testing of multiple endpoints in group sequential trials. *Statist. Med.* **29**, 219-228.

[31] Guo W., He L. and Sarkar S. (2014). Further results on controlling the false discovery proportion. *Ann. Statist.* **42**, 1070-1101.

[32] Hayre, L.S. (1985). Group sequential sampling with variable group sizes. *J. Res. Statist. Soc. B.* **47**, 90-97.

[33] He, L., Sarkar, S. and Zhao, Z. (2014). Capturing the severity of type II errors in high-dimensional multiple testing. *J. Multi. Analy.* **142**, 106-116.

[34] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika* **75**, 800-802.

[35] Holland, B. and Copenhaver, M. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* **43**, 417-423.

[36] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65-70.

[37] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383-386.

[38] Hwang, I.K., Shih, W.J. and DeCani, J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statist. Med.* **9**, 1439-1445.

[39] Jeffery, I., Higgins, D. and Culhance, A. (2006). Comparison and evaluation of methods for generating differential expressed genes lists from micro-array data. *BMC Bioinfo.* **7**, 359-375.

[40] Jennison, C. and Turnbull, B.W. (1991). Exact calculations for the sequential t, χ^2 and F tests. *Biometrika* **78**, 133-141.

[41] Jennison, C. and Turnbull, B.W. (1993). Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**, 741-752.

[42] Jennison, C. and Turnbull, B.W. (2000). Group sequential methods with application to clinical trials. *Chapman & Hall/CRC, New York.*

[43] Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities I: multivariate totally positive distributions. *J. Multi. Anal.* **10**, 467-498.

[44] Kiefer, J. and Weiss, L. (1957). Some properties of generalized sequential probability ratio tests. *Ann. Math. Statist.* **28**, 57-74.

[45] Kim, K. and DeMets, D.L. (1987). Design and analysis of group sequential tests based on type I error spending rate function. *Biometrika* **74**, 149-154

[46] Lai, T.L. (1973). Optimal stopping and sequential tests with minimize the maximum sample size. *Ann. Statist.* **1**, 659-673.

[47] Lan, K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

[48] Lehmacher, W., Wassmer, G. and Reitmeir, P. (1991). Procedures for two sample comparisons with multiple endpoints controlling the experiment-wise error rate. *Biometrics* **47**, 511-521.

[49] Lehmann, E.L. and Romano, J.P. (2005). Generalizations of the family-wise error rate. *Ann. Statist.* **33**, 1138-1154.

[50] Liu, F. and Sarkar, S. (2010). A note on estimating the false discovery rate under mixture model. *J. Statist. Plann. Inf.*, **140**, 1601-1609.

[51] Marcus, R., Peritz, E. and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655-660.

[52] Maurer, W. and Bretz, F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statist. Biopharm. Res.* **5**, 311-320.

[53] O'Brein, P.C. and Flemming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-556.

[54] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**,1079-1087.

[55] O' Neill, R.T. (1997). Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials* **18**, 550-556.

[56] Owen, A.B. (2005). Variance of the number of false discoveries. *J. Res. Statist. Soc. B* **67**, 411-426.

[57] Pocock, S.J.(1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

[58] Pocock, S.J., Geller, N.L. and Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487-498.

[59] Qiu, X., Klebanov, L. and Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statist. Appl. Genet. Mol. Biol.*, 4, article 34.

[60] Rom, D. (1990). A sequential rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663-665.

[61] Roy, S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24**, 220-238.

[62] Roy, S.N. and Bose, R.C. (1953). Simultaneous confidence interval estimation. *Ann. Math. Statist.* **24**, 513-536.

[63] Sarkar, T.K. (1969). Some lower bounds of reliability. Technical report, 124, Department of operation research and statistics, Stanford Univ.

[64] Sarkar, S. (1998). Some probability inequalities for the ordered MTP2 random variables: A proof of the Simes conjecture. *Ann. Statist.* **26**, 494-504.

[65] Sarkar, S. and Chang, C. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.* **92**, 1601-1608.

[66] Sarkar, S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30**, 239-257.

[67] Sarkar, S. (2004). FDR-controlling stepwise procedures and their false negatives rates. *J. Statist. Plann. Inf.*, **125**, 119-137.

[68] Sarkar, S. (2008a). Generalizing Simes′ test and Hochberg′s step-up procedure. *Ann. Statist.* **36**, 337-363.

[69] Sarkar, S. (2008b). On methods controlling the false discovery rate. *Sankhyā, Series A* **70**, 135-168.

[70] Sarkar, S., Chen, J. and Guo, W. (2013). Multiple testing in a two-stage adaptive design with combination tests controlling FDR. *J. Amer. Statist. Assoc.* **108**, 1385-1401.

[71] Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62**, 626-633.

[72] Slud, E.V. and Wei, L.J. (1982). Two sample repeated significance test based on the modified Wilcoxon statistics. *J. Amer. Statist. Assoc.* **77**, 862-868.

[73] Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16**, 243-258.

[74] Storey, J.D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. B* **64**, 479-498.

[75] Storey, J.D., Taylor, J.E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Soc. B* **66**, 187-205.

[76] Sun, W. and Cai, T.T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, **102**, 901-912.

[77] Tamhane, A.C., Mehta, C.R. and Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics* **66**, 1166-1176.

[78] Tang, D.I., Geller, N.L. and Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**, 23-30.

[79] Tang, D.I. and Geller, N.L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* **55**, 1188-1192.

[80] Tang, D.I., Gnecco, C. and Geller, N.L. (1989). Design of group sequential clinical trials with multiple endpoints. *J. Amer. Statist. Assoc.* **84**, 776-779.

[81] Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y. and Barlogie, B. (2003). The role of the WNT-signaling antagonist DKKI in the development of osteolytic lesions in multiple myeloma.*NewEngland J. Med.* **349**, 2438-2494.

[82] Wald, A. (1947). Sequential analysis. *New York:Wiley.*

[83] Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19**, 326-339.

[84] Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193-200.

[85] Westfall, P.H and Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *J. Statist. Plann. and Inf.* **99**, 25-40.

[86] Ye, Y., Li, A., Liu, L. and Yao B. (2012). A group sequential Holm procedure with multiple primary endpoints. *Statist. Med.* **32**, 1112-1124.

[87] Zehetmayer, S., Bauer, P. and Posch, M. (2008). Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statist. Med.* **27**, 4145-4160.